Uncertainty Quantification and Error Propagation in the Enthalpy and Entropy of Surface Reactions Arising from a Single DFT Functional

Gerhard R. Wittreich[a], Geun Ho Gu[b], Daniel J. Robinson[a], Markos A. Katsoulakis[c], Dionisios G. Vlachos[a,*]

[a] Department of Chemical and Biomolecular Engineering, Catalysis Center for Energy Innovation, RAPID Manufacturing Institute, and Delaware Energy Institute (DEI), Newark, DE 19716, USA

[b] Graduate School of EEWS, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, 305-335, South Korea

[c] Department of Mathematics and Statistics, University of Massachusetts, Amherst, Massachusetts 01003, USA

Abstract

Accounting for parametric uncertainty in models is essential for quantifying the models' predictive ability. Recently, approaches have been introduced to estimate parametric uncertainty in kinetic models while accounting for correlations among energy parameters. However, correlations have been estimated indirectly and correlations in entropies have not been accounted for. For surface catalyzed microkinetic models of >C2 (more than two carbon-containing) molecules, which consist of thousands of reaction steps and intermediate surface species, first-principles density functional theory (DFT) is costly, and thus, estimation of thermochemistry and reaction barriers requires surrogate methods of DFT, such as group additivity and Brønsted–Evans–Polanyi relationships, respectively. For such parametrization, model uncertainty is unclear. This work develops a framework to overcome these gaps using group additivity and a single DFT functional. We estimate correlations in parameters of kinetic models and quantify uncertainty for thermochemistry, reaction barriers, reaction paths, and ultimately reaction rates accounting also for the contribution of entropic uncertainty. The approach is illustrated on propane combustion and ethane oxidative dehydrogenation reactions.

Keywords

Uncertainty quantification, error propagation, kinetics, catalysis, group additivity

Introduction

First-principles microkinetic models (MKM) link *ab initio* density functional theory (DFT) data to reactor simulation and design[1–7]. MKMs can inform design of experiments, identify catalyst properties that optimize quantities-of-interest (QoI) for discovery of new catalysts, and link catalyst properties to reactor design[5–11]. MKMs depend on both thermodynamic and kinetic parameters (called collectively hereafter as model parameters or simply parameters) for all species and reactions in a reaction network.[3,12–17] Uncertainty in these parameters and its propagation to QoI's become essential in assessing the predictive ability of a model and finding ways to reduce uncertainty.

MKM parameters of surface catalyzed models are difficult to determine experimentally. Instead, DFT or semi-empirical group additivity (GA) schemes are employed to estimate thermodynamic properties. DFT combined with transition state theory (TST) or semi-empirical Bronsted-Evans-Polanyi (BEP) relationships are used to determine reaction barriers and eventually kinetic parameters[18,19]. DFT error affects model accuracy. Additionally, semi-empirical methods, which are often linear fits of DFT data, add error due to linearization.

It has recently been realized that energies of species and reactions are correlated and these correlations affect model prediction.[20–23] Medford et al.[22] examined the ammonia chemistry using the BEEFvdW (Bayesian error estimation functional with van der Waals correlation) functional, which offers an estimate of the variance of the relative electronic energies calculated over a family of functionals. The one-sigma variance among functionals in the adsorption energy was found to be 3.2 kcal/mol. Sutton et al.[21] injected results from 5 functionals into a Bayesian prior estimate postulating a one-sigma energy uncertainty of 5 kcal/mol. Correlations altered the most influential parameters and key reaction intermediates and reactions. Walker et al.[23] studied the water-gas shift reaction also using four functionals and found that the reaction path shifts depending on functional. These earlier works underscored the fact that failure to account for correlations misstates the variability of the reaction rates and can misidentify the rate-controlling reaction step(s). Work by others[24,25], following a similar approach, reached similar conclusions.

The variance in parameters in prior work using the BEEF-vdW functional or multiple functionals represents systematic differences between functionals, and the estimated correlations do not represent the actual correlations between species on the same catalyst with the same functional. This is because the species energies of a reaction network form a one-dimensional vector, and a method for estimating correlations (typical for 2D matrices) is lacking. Additionally, previous studies accounted only for correlations in electronic energies and excluded entropy. Finally, estimation of model parameters with multiple functionals requires significant resources that become impractical for large reaction networks.

The purpose of this work is to develop a scalable statistical framework to estimate correlations in both enthalpic and entropic parameters and propagate them in a MKM model. Specifically, this work enables one to perform uncertainty quantification for the same functional without invoking multiple functionals (a task that is expensive and limited in data due to the small ensemble of functionals available) or interpolative schemes, such as BEEF-vdW. Notably, it does so for both energy and entropy terms. Furthermore, it enables uncertainty quantification for large reaction mechanisms, which has been unattainable by previous works relying on expensive DFT calculations. The framework developed can be extended to include parametric uncertainty

arising from other relationships, such as linear scaling relations, BEP's, and adsorbate-adsorbate interactions.

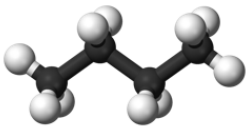
Methods

DFT enthalpic and entropic correlations cannot be estimated on a catalyst using a single functional. We propose to overcome this limitation using GA as explained below. The GA method can parameterize the thermochemistry of large molecules and/or reaction networks. GA was introduced by Benson [26] for gas molecules and was only recently extended to adsorbed species on surfaces [27]. GA relies on graph theory defining each molecule as a collection of sub-graphs (hereafter referred to as groups) with a frequency of occurrence for each. The values assigned to GA groups ($\boldsymbol{\beta}$) are determined from DFT-calculated thermodynamic properties ($\mathbf{Y}_{\mathrm{T}}$) of a (training) set of molecules via linear regression by minimizing the difference between the GA- and the DFT-estimated parameters. Entropic contributions due to finite temperature are included using vibrational frequencies and statistical mechanics, i.e., species electronic energy and frequency data are converted into species enthalpy, entropy, and free energy datasets. The entropic uncertainty contributes ~12% of the model root mean squared error (RMSE) to the Gibb's free energy. This impacts equilibrium calculations and reverse reaction rates (Table 1). We expect the entropic contribution to uncertainty to be larger for reactions with lower barriers (DFT work under revision in J. Phys. Chem. C, to be cited here) and at higher temperatures.

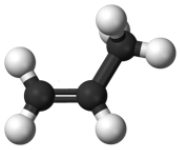
Table 1: GA model uncertainty (pGrAdd:GRWSurface2018 database).

Thermodynamic Property (298.15 K)	RMSE
Enthalpy	4.53 [kcal/mol]
Entropy	2.07 [cal/mol K]
Free energy	5.15 [kcal/mol]

Importantly, we propose that GA provides a natural framework to estimate correlations among groups from the training molecule-group configuration matrix $\mathbf{X}_{\mathrm{T}}$ using standard statistics and linear algebra. $\mathbf{X}_{\mathrm{T}}$ describes the frequency of occurrence of groups in molecules, and the subscript T stands for training, as demonstrated in Figure 1.



Butane	
Group	Frequency
C(C)(H)3	2
C(C)2(H)2	2
CC	3



Propene	
Group	Frequency
C(C)(H)3	1
CC	1
C[d](C[d])(H)2	1
C[d](C)(C[d])(H)	1

Configuration Matrix

Molecule \ Group	C(C)(H)3	C(C)2(H)2	CC	C[d](C[d])(H)2	C[d](C)(C[d])(H)
Butane	2	2	3	0	0
Propene	1	0	1	1	1

Figure 1: Example of groups and their frequency of occurrence in propene and butane (note that the group additivity scheme employed here is slightly different from that of Benson's) shown in the top tables and the molecule-group configuration matrix of butane and propene shown at the bottom. The notation follows the typical group additivity one, e.g., C(C)(H)3 implies a C connected to a C and 3 Hs. [d] stands for double bonded C.

The configuration matrix is usually an over-determined matrix with many more molecules (samples) than groups (parameters). The thermodynamic estimated group contribution vector $\left(\hat{\boldsymbol{\beta}}\right)$ is determined via ordinary least-squares (OLS) by the dot-product of the pseudo-inverse of the configuration matrix $\left(\mathbf{X}_{\mathrm{T}}^{+}\right)$ with the vector of thermodynamic values $\left(\mathbf{Y}_{\mathrm{T}}\right)$ (Eq. [1]):

$$\left.\begin{aligned} &\mathbf{X}_{\mathrm{T}} \cdot \mathbf{Y}_{\mathrm{T}} = \boldsymbol{\beta} \\ &\hat{\boldsymbol{\beta}} = \mathbf{X}_{\mathrm{T}}^{+} \cdot \mathbf{Y}_{\mathrm{T}} \end{aligned}\right\} \mathbf{X}_{\mathrm{T}}^{+} \Rightarrow \text{Pseudoinverse of } \mathbf{X}_{\mathrm{T}}$$
$$\left.\mathbf{X}_{\mathrm{T}}^{+} = \left(\mathbf{X}_{\mathrm{T}}^{'} \cdot \mathbf{X}_{\mathrm{T}}\right)^{-1} \cdot \mathbf{X}_{\mathrm{T}}^{'}\right\} \text{Ordinary least squares (OLS)} \qquad [1]$$

The model variance for each thermodynamic property is determined by comparing the predicted thermodynamic property to the DFT calculated property of each training molecule and dividing the mean squared difference by the degrees of freedom (Eq. [2]):

$$\hat{\sigma}^2 = \frac{\left\| \left(\mathbf{I}_{\mathrm{N_T}} - \mathbf{H}_{\mathrm{T}}\right) \mathbf{Y}_{\mathrm{T}} \right\|^2}{m - g - 1}; \; \mathbf{H}_{\mathrm{T}} = \mathbf{X}_{\mathrm{T}} \left(\mathbf{X}_{\mathrm{T}}^{'} \mathbf{X}_{T}\right)^{-1} \mathbf{X}_{\mathrm{T}}^{'} \qquad [2]$$

where $\hat{\sigma}^2$ is the unbiased estimator of model variance, $\hat{\mathbf{Y}}_{\mathrm{T}}$ is the estimated thermodynamic properties of the training molecules, (m - g - 1) is the training model degrees of freedom where m is the number of molecules (samples), g the number of groups (parameters), -1 accounts for computing the intercept, and $\mathbf{H}_{\mathrm{T}}$ is the "hat" matrix (the "hat" or projection matrix maps the

vectors of observed values to vectors of predicted values $\mathbf{Y}_T = \mathbf{H}_T \mathbf{Y}_T$) as defined above. The squared L2 norm of the numerator gives the sum of the squared estimate of errors (SSE).

Eqs. 1-2 are basic statistics. Properly accounting for correlations in the training data allows us to assign a portion of the model uncorrelated variance to each group, avoiding overstating the variance for any group. The impact of correlated data can be easily demonstrated in a simple bivariate example (see SI: Correlated data: Bivariate example).

The correlations in GA define a multi-variate problem for the distribution of thermodynamic input parameters. The distribution of thermodynamic properties $\mathbf{Y}_P$ for a given set of molecules can be stated as the conditional probability of $\mathbf{Y}_P$ given $\mathbf{X}_P$, $\mathbf{X}_T$, and $\mathbf{Y}_T$:

$$P(\mathbf{Y}_P|\mathbf{X}_P, \mathbf{X}_T, \mathbf{Y}_T) \quad [3]$$

where $\mathbf{X}_P$ is the molecule-group configuration matrix of a set of molecules that one wishes to predict and $\mathbf{X}_T$, $\mathbf{Y}_T$ are the molecule-group configuration matrix and thermodynamic values of the training set molecules. This distribution can be expanded to:

$$P\left(\mathbf{Y}_P \middle| \mathbf{X}_T, \mathbf{Y}_T, \mathbf{X}_P\right) = P\left(\mathbf{Y}_P \middle| \hat{\mathbf{Y}}_P, \hat{\boldsymbol{\varepsilon}}_{\text{Random}}\right) P\left(\hat{\mathbf{Y}}_P \middle| \mathbf{X}_P, \hat{\boldsymbol{\beta}}\right) P\left(\hat{\boldsymbol{\beta}} \middle| \mathbf{X}_T, \mathbf{Y}_T\right) P\left(\hat{\boldsymbol{\varepsilon}}_{\text{Random}} \middle| \mathbf{X}_T, \mathbf{Y}_T\right) \quad [4]$$

where:

$$\begin{aligned} &\mathbf{Y}_P = \mathbf{X}_P\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{\text{Random}} = \mathbf{X}_P\hat{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}_{\text{Fit}} + \boldsymbol{\varepsilon}_{\text{Random}} = \mathbf{X}_P\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\varepsilon}}_{\text{Random}} + \left(\boldsymbol{\varepsilon}_{\text{Fit}} + \boldsymbol{\varepsilon}_{\text{Random Estimation}}\right) \\ &\hat{\mathbf{Y}}_P = \mathbf{X}_P\hat{\boldsymbol{\beta}} + \left(\boldsymbol{\varepsilon}_{\text{Fit}} + \boldsymbol{\varepsilon}_{\text{Random Estimation}}\right) \therefore \mathbf{Y}_P = \hat{\mathbf{Y}}_P + \hat{\boldsymbol{\varepsilon}}_{\text{Random}} \\ &\hat{\boldsymbol{\beta}} \text{ is the estimated thermodynamic group contribution} \\ &\boldsymbol{\varepsilon}, \hat{\boldsymbol{\varepsilon}} \text{ are the sources of actual and estimated error} \end{aligned} \quad [5]$$

To determine the probability distribution function (PDF) of thermodynamic properties ($\mathbf{Y}_P$), we estimate the error distribution (root mean squared error (RMSE) of a thermodynamic property) resulting from the OLS fit of our training data by using a bootstrap resampling method. We estimate thermodynamic properties using the GA scheme of the pGrAdd software and database (pGrAdd:GRWSurface2018) that was trained using DFT data by Gu et al.[28] This training set comprises 164 molecules containing C (C1 – C5), H and O described by 66 groups. For this published data, the Vienna ab initio Simulation Package (VASP)[29] was used with the Perdew, Burke, and Ernzerhof (PBE) exchange-correlation functional and the dDsC dispersion correction[30,31] to determine electronic energies and vibrational frequencies. The projector augmented-wave function (PAW)[32–34] method was used to treat core electrons and a 400 eV cutoff for valence electrons. A 4 x 4 four-layer unit cell is employed with the bottom two layers remaining fixed, while a Monkhorst−Pack mesh of 3 × 3 × 1 k-points is used to integrate the Brillouin zone[35]. We repeatedly resample $9/10$ of those training molecules, use an OLS fit (described in detail in SI:Group Additivity) to compute group contributions and then determine the RMSE of the predicted thermodynamic property vs. DFT computed using the remaining $1/10$ of the training molecules[36]. This way, we estimate the mean, variance, and distribution of the

RMSE for each thermodynamic property. Results are compared to a normal distribution, using the predicted enthalpy as an example (Figure 2). The error distribution is consistent with a normal distribution.

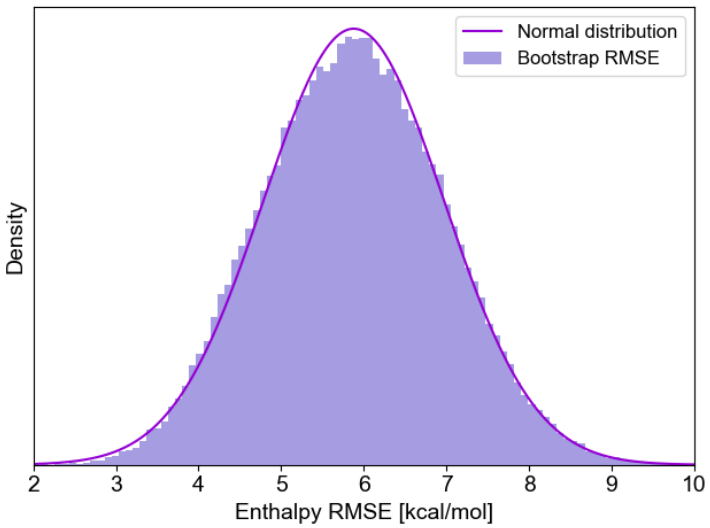


Figure 2: Error distribution of the pGrAdd:GRWSurface2018 database and scheme. Histogram (purple) is the result of a bootstrap resampling to determine the enthalpy RMSE distribution vs. a normal PDF (dark violet), validating that error is normally distributed.

We can also subject the data set to an Anderson-Darling (AD) goodness-of-fit test vs a normal distribution[37,38]. AD and other goodness-of-fit tests fail with very large data sets, like the one we show here, so we need to examine subsets of this data and use the collective results to assess normality. Using a sample size of 100 points (largest sample size that does not expose the sample size weakness of AD) and drawing multiple samples (1000) from the data set results in over 95% of the samples testing as normal with the AD test (p-value of the test ≥ 0.05). This allows us to conclude that the GA fit errors are normally distributed and finalize the multi-variate distribution. Assuming the distribution is Gaussian, the multi-variate normal distribution of $\mathbf{Y}_P$ is then estimated as:

$$\begin{aligned} \mathrm{P}\left(\mathbf{Y}_\mathrm{P}\middle|\mathbf{X}_\mathrm{T},\mathbf{Y}_\mathrm{T},\mathbf{X}_\mathrm{P}\right) &\approx \mathcal{N}\left(\mathrm{E}\left(\mathbf{X}_\mathrm{P}\hat{\boldsymbol{\beta}}+\hat{\boldsymbol{\varepsilon}}_{\mathrm{Random}}+\boldsymbol{\varepsilon}_{\mathrm{Fit,RE}}\right),\mathrm{Var}\left(\mathbf{X}_\mathrm{P}\hat{\boldsymbol{\beta}}+\hat{\boldsymbol{\varepsilon}}_{\mathrm{Random}}+\boldsymbol{\varepsilon}_{\mathrm{Fit,RE}}\right)\right) \\ \mathrm{P}\left(\mathbf{Y}_\mathrm{P}\middle|\mathbf{X}_\mathrm{T},\mathbf{Y}_\mathrm{T},\mathbf{X}_\mathrm{P}\right) &\approx \mathcal{N}\left(\mathbf{X}_\mathrm{P}\hat{\boldsymbol{\beta}},\hat{\sigma}^2\left(\mathbf{I}_{\mathrm{N_P}}+\mathbf{H}_\mathrm{P}\right)\right) \end{aligned} \quad [6]$$

where $\hat{\boldsymbol{\beta}}$ is the estimated thermodynamic group contribution, $\mathbf{X}_\mathrm{T}$, $\mathbf{Y}_\mathrm{T}$ are the configuration and thermodynamic properties of the training molecules, $\mathbf{X}_\mathrm{P}$ is the configuration matrix of the molecules we wish to predict, $\boldsymbol{\varepsilon},\hat{\boldsymbol{\varepsilon}}$ represent the various sources of actual and estimated error, and the "hat" matrix is $\mathbf{H}_\mathrm{P}=\mathbf{X}_\mathrm{P}\left(\mathbf{X}_\mathrm{T}^{'}\mathbf{X}_\mathrm{T}\right)^{-1}\mathbf{X}_\mathrm{P}^{'}$ (see SI: Multivariate normal distribution derivation for GA). The "hat" matrix, $\mathbf{H}_P$, contains the correlation matrix $\mathbf{X}_\mathrm{T}^{'}\mathbf{X}_\mathrm{T}$, a symmetric matrix that represents the variance contribution of each group to the model overall variance in the diagonal elements and the covariance, variance shared between pairs of groups, in the off-diagonal

elements. The correlation matrix ensures that each group is assigned a variance that both represents its contribution to the overall model variance without double-counting variance it shares with another group.

We now use GA to estimate the unperturbed thermodynamic properties of a set of molecules, sample from a multivariate zero-mean distribution representing the variance of each molecule (model variance, $\hat{\sigma}^2_{(Cp,H,S)}$ computed via bootstrap sampling), and sum the two:

$$\mathcal{N}\left(\mathbf{X}_{\mathrm{P}}\hat{\boldsymbol{\beta}},\hat{\sigma}^2\left(\mathbf{I}_{\mathrm{N_P}}+\mathbf{H}_{\mathrm{P}}\right)\right)_{(Cp,H,S)} = \underbrace{\mathbf{X}_{\mathrm{P}}\hat{\boldsymbol{\beta}}_{(Cp,H,S)}}_{\substack{\text{Unperturbed GA}\\ \text{thermodynamicvalues}}} + \underbrace{\mathcal{N}\left(\mathbf{0}_{\mathrm{N_P}},\hat{\sigma}^2_{(Cp,H,S)}\left(\mathbf{I}_{\mathrm{N_P}}+\mathbf{H}_{\mathrm{P}}\right)\right)_{(Cp,H,S)}}_{\substack{\text{Zero-mean multi-variate normal distribution}\\ \text{of molecule thermodynamic value uncertainty}}} \quad [7]$$

Translating a distribution of MKM input parameters into a distribution of MKM QoIs analytically is, in general, not possible as it depends on several nonlinear relationships. Uncertainty in species free energies translates into uncertainty in reaction thermochemistry (enthalpy, entropy, free energy, and equilibrium constant), reaction barriers through BEP relationships, and reaction rate at the MKM level. The following equations illustrate how the distribution of thermodynamic values ($H_i, S_i, G_i + \varepsilon_{i,\mathrm{Perturb}}$) impacts these QoI's:

$$TOF \propto e^{-\frac{E_{a,i}}{RT}} \quad [8]$$

$$E_{a,i} = \varphi_i\left(H_{f,i} + \varepsilon_{i,\text{Later Interactions}} + \varepsilon_{i,\mathrm{Perturb}}\right) + \gamma \qquad \{BEP\} \quad [9]$$

$$K_{eq,i} = e^{-\frac{\Delta G_i + \varepsilon_{i,\text{Later Interactions}} + \varepsilon_{i,\mathrm{Perturb}}}{RT}} \quad [10]$$

In addition, lateral interactions ($\varepsilon_{i,\text{Later Interactions}}$) modify the energy of each adsorbed species based on surface coverages of all species. This introduces a nonlinear recursive relationship. Lateral interactions are a function of surface coverage and the enthalpy change due to the coverage:

$$\varepsilon_{i,\text{Later Interactions}} = f\left(\theta_i, \Delta H_{i,\text{interaction}}\right) \quad [11]$$

Surface coverages are a function of the species binding energy as well as the concentration of species in the gas phase. The species binding energy is also a function of the lateral interaction:

$$\theta_i = f\left(H_i + \varepsilon_{i,H,\text{Later Interactions}} + \varepsilon_{i,\mathrm{Perturb}}, S_i + \varepsilon_{i,S,\text{Later Interactions}} + \varepsilon_{i,\mathrm{Perturb}}, C_{i,gas}\right) \quad [12]$$

Finally, the species-gas concentrations are a function of the reaction rates and equilibrium constants and, in more complex reaction mechanisms, also a function of the reaction path:

$$C_{i,gas} = f\left(\text{TOF}, \text{Reaction Path}\right) \quad [13]$$

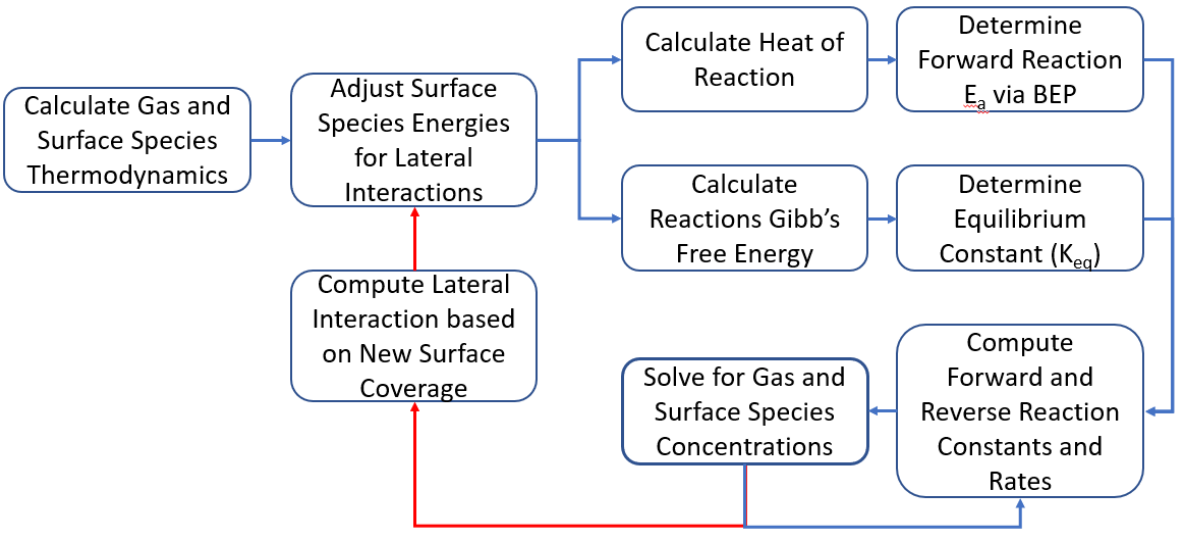


Figure 3: Workflow demonstrating MKM parameterization and solution with the lateral interactions recursive relationship shown in red. When nonlinear algebraic or differential-algebraic solvers are used, all quantities are computed simultaneously rather than recursively.

The workflow of the model is demonstrated in Figure 3. Eqs. 6 and 7 along with S16-S19 in the SI form the base for our sampling and demonstrate how the uncertainty and correlations are accounted for. Monte Carlo stochastic sampling is used to create a family of thermodynamic input parameters built from the vector of expected values $\mathbf{X}_{\mathrm{P}}\hat{\boldsymbol{\beta}}$ summed with repeated samples from the model's zero-mean covariance distribution $\mathcal{N}\left(\mathbf{0}_{\mathrm{N_P}}, \hat{\sigma}^2\left(\mathbf{I}_{\mathrm{N_P}} + \mathbf{H}_{\mathrm{P}}\right)\right)$. Each thermodynamic set determines a set of reaction thermochemistry and rate constants that are input into the MKM to find a solution. The distribution of thermodynamic parameters leads to a distribution of MKM calculations and QoI's. The distribution of QoI's represents the uncertainty of the MKM's.

Demonstrated Examples

Two reaction networks were chosen to demonstrate error propagation. The first is the oxidative dehydrogenation of ethane on a Pt(111) surface. The reaction network consists of 295 elementary reactions generated using RING (a rule-based reaction network generator)[39]. The thermodynamic properties of the 45 surface species are estimated using GA and are the source of our parametric input uncertainty while 9 additional surface species' thermodynamics is estimated using DFT because these are small species that are groups by themselves; using DFT values ensures accuracy. 18 gas species thermodynamics are obtained from NIST[40]. The second is the oxidation of propane which consists of 1,315 elementary reactions obtained using RING[39]. The thermochemistry of the network consists of 147 surface species estimated using GA, 9 surface species using DFT, and 35 gas species using NIST[40]. Reaction barriers are determined from BEP relationships from heats of reaction[41].

GA group contribution estimates were based on the DFT zero Kelvin electronic energy and vibrational data of [28] which was converted to finite temperature thermodynamic quantities using

the statistical mechanics converter in the Python Multiscale Thermochemistry Toolbox (pMuTT). GA group contributions, the group definition scheme, and the automated surface species thermodynamic value estimation were managed using the Python GA software (pGrAdd)[42]. MKM's were executed with the Surface Chemkin software[43] with thermodynamic input files created using pMuTT. Pre-exponential factors for all reactions are a first-order estimate and normalized for the catalyst surface density:

$$A = \frac{k_B}{h \cdot \Gamma^{m-1}} \qquad [14]$$

where k_B is Boltzmann's constant, h is Planck's constant, and Γ is the catalyst site density (see SI: Chemkin input files). The pre-exponential does not include a temperature term as Chemkin includes a separate exponentiated temperature term in the rate constant.

A total of 5,000 perturbed thermodynamic input files were generated for each reaction network and MKM was solved for each at a range of 36 temperatures and a constant reactant feed ratio to assess conversion, turnover frequency (TOF), and apparent activation energy. The same calculations were performed at constant temperature for 9 different reaction feed ratios to assess hydrocarbon and oxygen reaction orders.

Results and Discussion

If the GA groups were independent of each other, the correlation matrix would consist of diagonal elements of value one and off-diagonal elements of value zero. In essence, the correlation matrix (Figure 4a) provides physical insights into the independence of sub-graphs. These correlations are mostly low, as one would anticipate from a robust group additivity scheme where each group contributes maximum information, and the groups are not correlated to each other. Nonetheless, there are islands of higher correlation that they are properly

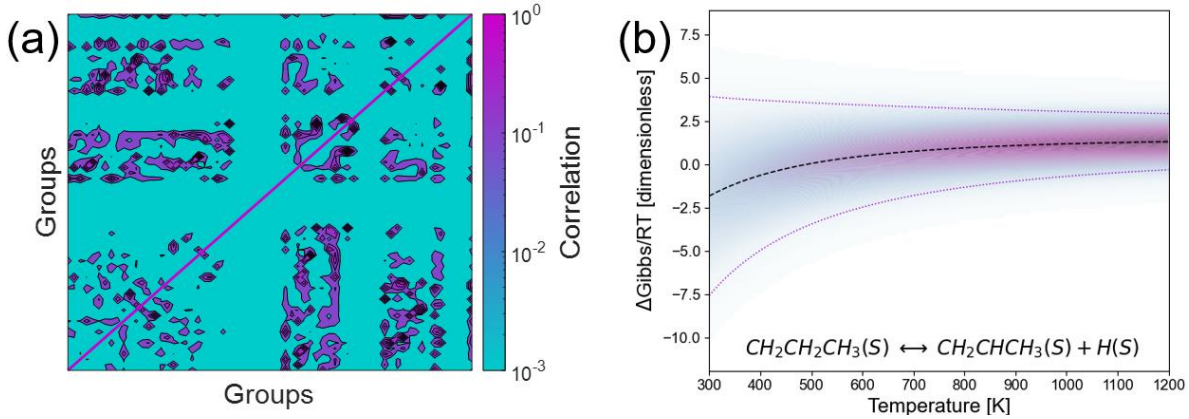


Figure 4: (a) Correlation heatmap between the 66 groups in GA training set used to determine group thermodynamic property contributions. The values are symmetric: corr(X, Y) = corr(Y, X). (b) Distribution of the Gibbs' free energy for a typical elementary surface reaction step. Mean value (dark-dash), 95% Prediction Interval (light dotted) and distribution of values (blue-purple contour map).

accounted for using the covariance matrix described earlier. This matrix, using the GA overall model uncertainties (Figure 4a), also properly assigns both a mean thermodynamic contribution to each group and a variance representing the uncertainty contribution from each group. Surface

species thermodynamic values as a function of temperature are now represented in the model as normal distribution and, consequently, the thermochemistry of elementary reaction steps that includes those surface species is also normally distributed (Figure 4b). The resulting QoI distribution from the MKMs is though unknown. The uncertainty in the dimensionless free energy (ΔG/RT) of a chemical reaction is amplified at low temperatures due to a temperature term in the denominator. Even at high temperatures, where the uncertainty is significantly lower, the range is relatively large, creating significant variation in the equilibrium constant and the backward rate constant. Improving the accuracy of thermochemistry is a critical task. Physics-based[44] and machine learning-based[45] approaches are suitable for this task but outside the scope of this work.

The conversion as a function of temperature shows some of the complexity as a normally distributed thermodynamic input parameter transforms in the model. The complete oxidation of propane (Table 2) is a simpler reaction with propane converted to combustion products with near 100% selectivity (Figure 5a). Many of the solutions are concentrated near zero and near 100% conversion suggesting that once this reaction initiates it completes rapidly (this is typical for combustion reactions). Consequently, the 95% prediction interval[46]

$$PI_h = \hat{y}_h \pm t_{(\alpha/2,n-2)}\sqrt{MSE\left(1+\frac{1}{n}+\frac{\left(x_h-\bar{x}\right)^2}{\sum\left(x_i-\bar{x}\right)^2}\right)} \quad [15]$$

often spans the entire range from 0 to 100%. Here $\hat{y}_h$ is the model predicted value at $h, t_{(\alpha/2,n-2)}$ is the critical value from a 2-sided Student's t- distribution at a total tail area of α (for a 95% prediction interval α = 0.05) and degrees of freedom n-2 (n is the number of observations), MSE is the mean squared error between the predicted $\hat{y}_h$ and the observed y_{observed}, x_i are the observed x-values, $\bar{x}$ is the mean observed x-value, and x_h the x-values at the predicted $\hat{y}_h$.

Table 2: Overall reactions in demonstrated examples of propane total oxidation and ethane oxidative dehydrogenation on a Pt catalyst.

Propane Oxidation

$$C_3H_8 + 5O_2 \rightleftarrows 3CO_2 + 4H_2O \text{ (oxidation)}$$

$$CO + H_2O \rightleftarrows CO_2 + H_2 \text{ (water-gas shift)}$$

Ethane Oxidative Dehydrogenation

$$C_2H_6 + \tfrac{7}{2}O_2 \rightleftarrows 2CO_2 + 3H_2O \text{ (oxidation)}$$

$$C_2H_6 \rightleftarrows C_2H_4 + H_2 \text{ (dehydrogenation)}$$

$$C_2H_6 + H_2 \rightleftarrows 2CH_4 \text{ (hydrogenolysis)}$$

$$\left.\begin{aligned} C_2H_6 + 2H_2O &\rightleftarrows 2CO + 5H_2 \\ C_2H_4 + 2H_2O &\rightleftarrows 2CO + 4H_2 \\ CH_4 + H_2O &\rightleftarrows CO + 3H_2 \end{aligned}\right\} \text{(reforming)}$$

$$CO + H_2O \rightleftarrows CO_2 + H_2 \text{ (water-gas shift)}$$

The oxidative dehydrogenation of ethane (Table 2) is more complex with multiple reaction regimes. The reaction starts with complete oxidation yielding mostly combustion products, followed by a dehydrogenation reaction and a competing hydrogenolysis reaction. If allowed to continue long enough, this mechanism shifts to reforming reactions that will consume the ethylene produced in the dehydrogenation regime. The resulting conversion as *f*(Temperature) plot shows some of the complexity. In the first regime of oxidation, O_2 is quickly consumed, converting 17.5%

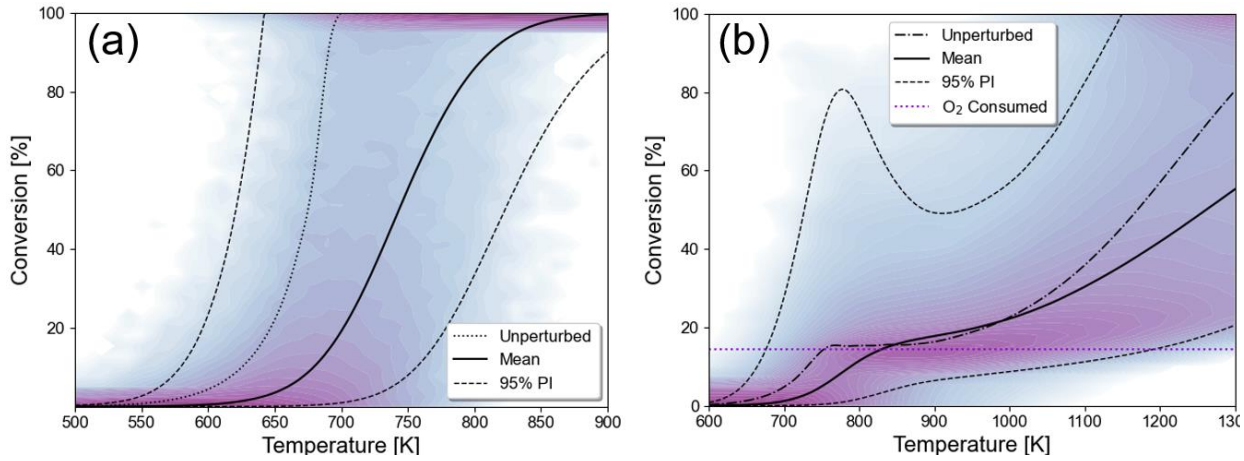


Figure 5: Conversion over a range of temperatures. (a) Propane complete oxidation (C_3H_8:O_2 = 1:3.5). (b) Oxidative dehydrogenation ethane (C_2H_6:O_2 = 1:0.5). Contour shows relative density of solutions from 5,000 simulations.

of the ethane (Figure 5b). The prediction interval and density of solutions at 0 and 17.5% conversion are very similar to propane oxidation. The prediction interval is very wide at the point where most solutions have consumed the oxygen (~750 K). A key QoI is the TOF. Knowing that the reaction energies are normally distributed and exponentiated in the equilibrium constants, the semi-log plot of TOF vs 1000/T (Figure 6) provides an ensemble of solutions in an Arrhenius plot.

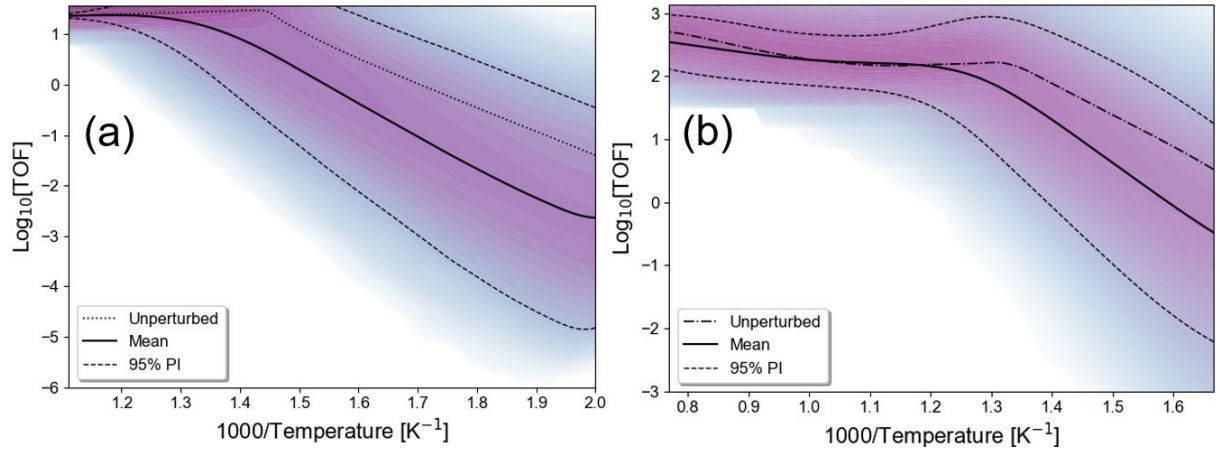


Figure 6: Reaction rate (TOF) analysis. (a) Propane complete oxidation (C_3H_8:O_2 = 1:3.5). (b) Oxidative dehydrogenation ethane (C_2H_6:O_2 = 1:0.5). Contours show relative density of solutions from 5,000 simulations.

The straight-line anticipated from an Arrhenius plot and the symmetric distribution of values around the mean suggest this data is normally distributed at a temperature. The oxidative dehydrogenation of ethane again shows multiple regions of linear relationships consistent with the multiple reaction regimes. It is interesting that the mean result and the unperturbed solution are different. This is a consequence of the nonlinear recursive relationship between parametric uncertainty and distribution of QoI's (for further discussion, see SI: Nonlinear recursive relationships).

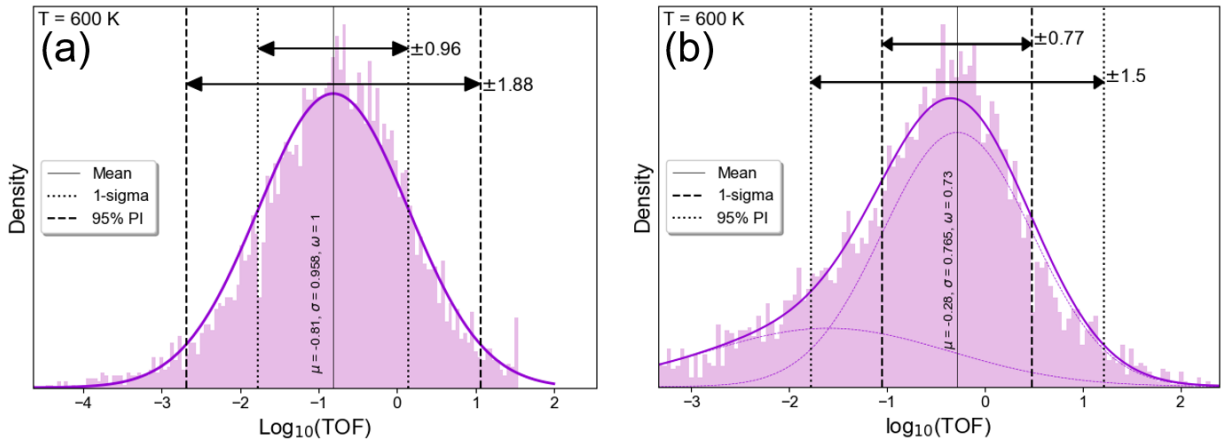


Figure 7: Reaction rate (TOF [s^{-1}]) distribution at a fixed temperature. (a) Propane complete oxidation (C_3H_8:O_2 = 1:3.5). (b) Oxidative dehydrogenation ethane (C_2H_6:O_2 = 1:0.5). Composite distribution (purple-solid) and contributing normal distributions (violet-dotted).

The distribution of log(TOF's) at a fixed temperature (Figure 7) is predicted to be normally distributed given that the apparent activation energy is normally distributed and exponentiated. The propane distribution of log(TOF's) in Figure 7a is normal (by inspection vs an ideal normal distribution and AD normality test) with a 1-sigma uncertainty of about ± 1 order-of-magnitude

$$P(x)_{obs} \approx P(x)_{approx} = \sum_{i=1}^{n_c} \omega_i \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{\frac{-(x-\mu_i)^2}{2\sigma_i^2}} \quad [16]$$

in the reaction rate. The ethane TOF (Figure 7b) is not as clearly normal, having a skew of -0.687 and kurtosis of 1.2, preventing the use of a simple mean and variance applied to normal distributions[47]. We tested a wide range of other distributions, but the goodness-of-fit was unacceptable. The Gaussian mixture model was the best alternative in situations where a single Gaussian or an alternative distribution were inadequate. Alternatively, we can consider a Gaussian mixture model[46,48] (Eq. [16) to fit multiple (n_c) normal distributions within the observed distribution $P(x)_{obs}$. The Gaussian mixture model creates a probability distribution function (pdf) through a combination of Gaussian distributions. This is common practice in, for example, deconvoluting spectra. The idea behind approximating the histogram with a Gaussian mixture is to find the best-fit parameters (weights, means, and covariances) that ensure the best approximation, that is $P(x)_{approx}$ weighted sum of an optimum number of Gaussians (n_c). Any number of distributions can be assumed, but overfitting can occur if too many are used. Evaluating the Bayesian Information Criteria (BIC) loss, which balances fit and model complexity

(n_c), over a range of n_c's and selecting the lowest loss allows us to select the optimum number of Gaussians. We achieve this through the Python scikit-learn machine learning library sklearn.mixture.GaussianMixture[49] using an expectation-maximization (EM) algorithm to accomplish the fit and BIC loss resulting in means (μ_i), standard deviations (σ_i), and weights (ω_i) for the contributing distributions. As the resultant data is now characterized via a combination of normal distributions, we can easily identify the dominant peak and apply mean and variance measures as for any normal distribution.

In Figure 7b, we fit two Gaussians for the ethane ODH, a dominant one with parameters μ_1 = -0.28, σ_1 = 0.77, ω_1=0.75 and a secondary one with μ_2 = -1.69, σ_2 = 1.19, ω_2 = 0.25. The uncertainty in the dominant reaction rate distribution is close to but lower than ± 1 order-of-magnitude. This uncertainty in reaction rate in both reactions considered in this work is slightly lower but consistent with previous works; the specific numbers depend on conditions, e.g., temperature, and probably less so on chemistry, given that comparable level of uncertainty has been seen in several different reactions here and in previous works. The good fit obtained with just two Gaussians indicates that most combinations of parameters lead to a dominant reaction rate, but there are also combinations of parameters that result in a mean reaction rate that is lower by about two orders of magnitude.

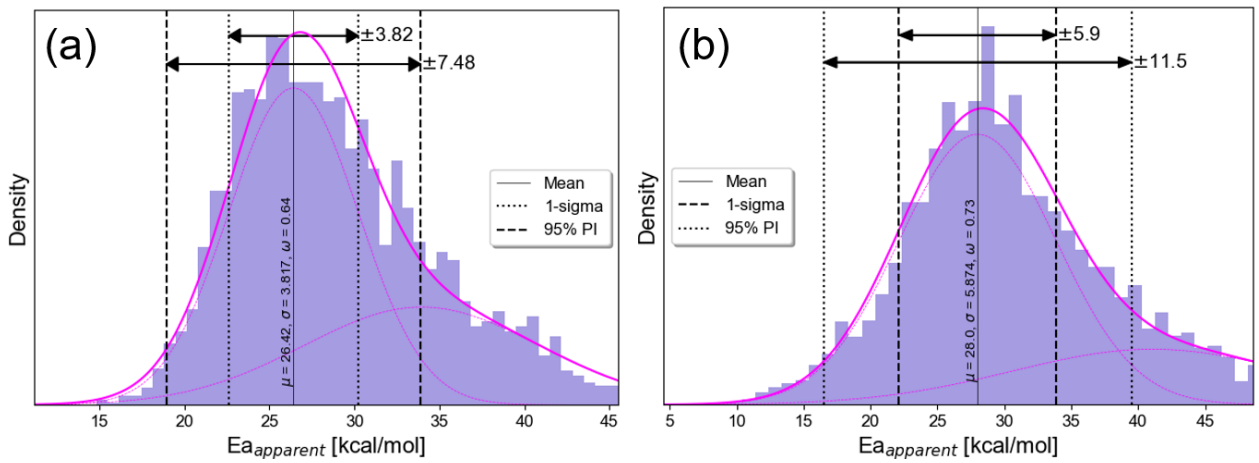


Figure 8: Apparent activation energy distributions. (a) Propane complete oxidation (C_3H_8:O_2 = 1:3.5). (b) Oxidative dehydrogenation ethane (C_2H_6:O_2 = 1:0.5). Composite distribution (violet-solid) and contributing normal distributions (violet-dotted).

The distribution of apparent activation energies for both example mechanisms displays right skew and kurtosis >0 inconsistent with normal distributions. Identical analysis on the apparent activation energy (Figure 8) shows dominant peaks encompassing approximately 70% of the observed data. The one standard deviation uncertainty is 4-6 kcal/mol, with additional normal distribution at higher values resulting in a positive skew to the composite distribution. We believe that the higher mean apparent activation energy corresponds to the lower mean TOF, shown more clearly in Figure 7b for ethane oxidation. The predicted range of uncertainty in activation energies is important for future comparison to experimental data. Furthermore, and surprisingly, the two mean apparent activation energies of ~26 and ~34 kcal/mol for propane oxidation and ~28 and ~42 kcal/mol for oxidation of ethane underscore the widespread in the mean apparent

activation energy predicted by a model and rationalize that combination of parameters may explain the large different often seen from experiments. We hypothesize that small changes in the catalyst could activate different pathways resulting in drastically different apparent activation energies. Further work is needed to exploit this physics deeper.

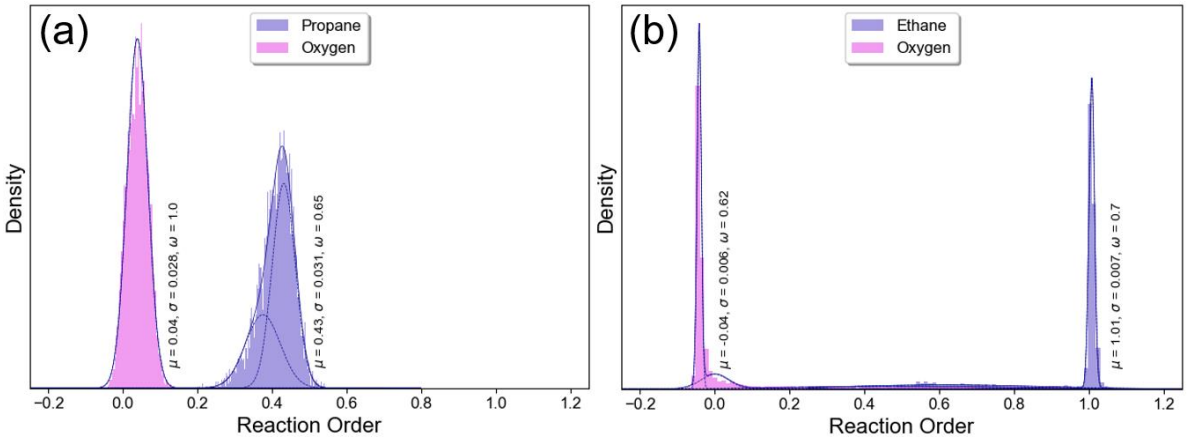


Figure 9: Propane, ethane, and oxygen reaction-order distributions. (a) Propane complete oxidation (C_3H_8:O_2 = 1:3.5). (b) Oxidative dehydrogenation ethane (C_2H_6:O_2 = 1:0.5).

Reaction orders (Figure 9) exhibited much lower variance than other QoI's with one sigma uncertainty ranging from ± 0.006 to 0.03, which is a very low uncertainty. We have advocated in recent years that the reaction order is one of the most sensitive kinetic signatures, e.g., Ref.[50] The tight uncertainty predicted and the model results herein further support this proposal. The dominant peak based on a Gaussian mixture model still encompasses about 70% of the observed data except for the oxygen reaction order in the propane total oxidation mechanism (Figure 9a), which contained all the data. While the reaction orders for the oxidative dehydrogenation of ethane (Figure 9b) have a very small variance for the dominant 70% of the data, the remaining 30% spans a wide range from order 0 to 1 (difficult to discern at the graph's scale). The oxygen reaction order is centered around zero for both hydrocarbons, consistent with most experimental data on hydrocarbon oxidation, e.g., (See Ref. [51] and references therein). The dominant ethane reaction order is ~1, whereas that for propane is ~0.4. The former is consistent with whereas the latter is lower than most experimental data; (See Ref. [51] and references therein) the reason for the lower reaction order for propane is unclear but beyond the scope of interrogating the microkinetic model against experimental data.

The framework we described allows one to assess the uncertainty in each MKM QoI, expressing each as a distribution of solutions. This also allows meaningful comparison to experimental data by comparing their respective 95% prediction intervals to determine if the model and experimental data are statistically indistinguishable or not.

Conclusions

This work has established a framework for assessing error propagation in surface catalyzed microkinetic models (MKMs) by examining error propagation from uncertainty and correlations in model parameters. The framework bypasses limitations of prior work, which relied on multiple functionals or the BEEF functional and in essence, provided correlations among functionals rather

than those between species and reaction parameters. We accomplished this task using the well-established framework of group additivity (GA) and by assigning not only a thermodynamic value to each subgraph or group, as is typical for GA works, but also a variance and the associated correlation among groups (the correlation among groups is small by design; groups should be as independent as possible). Through this framework, entropic correlations are naturally included for the first time. Importantly, since the GA scheme is transferable and explainable, in machine learning terminology, it allows estimation of error for large molecules and large reaction networks from a rather small training set of molecules/species whose thermodynamics is computed via DFT. We demonstrated the framework to ethane oxidative dehydrogenation and propane total oxidation on Pt.

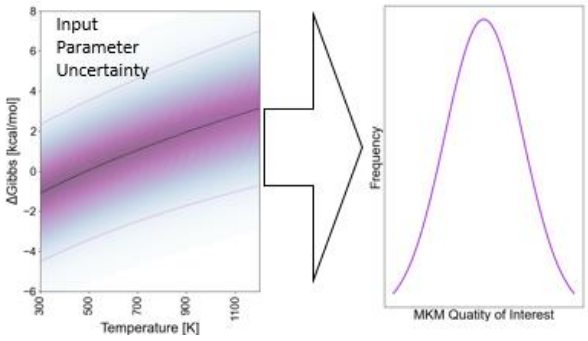
The exponentiation of thermodynamic and kinetic quantities allows us to assume normal distributions for a logarithmic transformation of those quantities. Standard statistical analyses can then be used to assess the distribution of quantities of interest (QoI's) and estimate Gaussian mixed models for more complex composite multiple normal distributions. While the exact transformation of input parameter uncertainty appears reaction-dependent, there are similarities between the two mechanisms. An input uncertainty of 4.5 kcal/mol in enthalpies and 2.1 cal/mol K in entropies (several studies suggest such energy value for the DFT uncertainty) results in ~± 1 (or less) order-of-magnitude uncertainty (one standard deviation) in the TOF. Apparent activation energies possess an uncertainty of 4-6 kcal/mol (15 – 20%), while reaction orders, with very low variance, are a very reliable QoI. These uncertainty intervals allow one to compare models with experimental data and assess if the prediction intervals overlap. Finally, bimodality in reaction rates, apparent activation energies, and reaction orders is possible for certain combinations of parameters. It is important to note that this study examined only the uncertainty in the thermodynamic input parameters as understood through GA. MKM's depend on a variety of parameters as encountered in BEP's, LSR's, and lateral surface interactions as well as model uncertainty regarding the reaction mechanism. A complete assessment of MKM uncertainty would need to consider all these uncertainties but could be performed with the proposed framework.

Supporting Information: Brief description of group additivity**,** bivariate example of correlated data**,** multivariate normal distribution derivation for group additivity, and analysis of nonlinear recursive relationships

Acknowledgments

The work of Gerhard R. Wittreich, Daniel J. Robinson and Dionisios G. Vlachos was supported by the Department of Energy's Office of Energy Efficient and Renewable Energy's Advanced Manufacturing Office under Award Number DE-EE0007888-9.5. The Delaware Energy Institute gratefully acknowledges the support and partnership of the State of Delaware toward the RAPID projects. The work of Markos A. Katsoulakis was supported by NSF TRIPODS CISE- 1934846 and Air Force Office of Scientific Research (AFOSR) FA-9550-18-1-0214.

TOC Graphic


Input
Parameter
Uncertainty
ΔGibbs [kcal/mol]
8
6
4
2
0
-2
-4
-6
300
500
700
900
1100
Temperature [K]
Frequency
MKM Quatity of Interest

References

(1) Salciccioli, M.; Stamatakis, M.; Caratzoulas, S.; Vlachos, D. G. A Review of Multiscale Modeling of Metal-Catalyzed Reactions: Mechanism Development for Complexity and Emergent Behavior. *Chem. Eng. Sci.* **2011**, *66* (19), 4319–4355.

(2) Beretta, A.; Donazzi, A.; Livio, D.; Maestri, M.; Groppi, G.; Tronconi, E.; Forzatti, P. Optimal Design of a CH4 CPO-Reformer with Honeycomb Catalyst: Combined Effect of Catalyst Load and Channel Size on the Surface Temperature Profile. *Catal. Today* **2011**, *171* (1), 79–83.

(3) Alexopoulos, K.; John, M.; Van Der Borght, K.; Galvita, V.; Reyniers, M. F.; Marin, G. B. DFT-Based Microkinetic Modeling of Ethanol Dehydration in H-ZSM-5. *J. Catal.* **2016**, *339*, 173–185.

(4) John, M.; Alexopoulos, K.; Reyniers, M.-F.; Marin, G. B. Reaction Path Analysis for 1-Butanol Dehydration in H-ZSM-5 Zeolite: Ab Initio and Microkinetic Modeling. *J. Catal.* **2015**, *330*, 28–45.

(5) Susnow, R. G.; Dean, A. M.; Green, W. H.; Peczak, P.; Broadbelt, L. J. Rate-Based Construction of Kinetic Models for Complex Systems. *J. Phys. Chem. A* **1997**, *101* (20), 3731–3740.

(6) Motagamwala, A. H.; Ball, M. R.; Dumesic, J. A. Microkinetic Analysis and Scaling Relations for Catalyst Design. *Annu. Rev. Chem. Biomol. Eng.* **2018**, *9*, 413–450.

(7) Motagamwala, A. H.; Dumesic, J. A. Microkinetic Modeling: A Tool for Rational Catalyst Design. *Chem. Rev.* **2021**, *121* (2), 1049–1076.

(8) Thybaut, J. W.; Sun, J.; Olivier, L.; Van Veen, A. C.; Mirodatos, C.; Marin, G. B. Catalyst Design Based on Microkinetic Models: Oxidative Coupling of Methane. *Catal. Today* **2011**, *159* (1), 29–36.

(9) Thybaut, J. W.; Marin, G. B. Single-Event MicroKinetics: Catalyst Design for Complex Reaction Networks. *J. Catal.* **2013**, *308*, 352–362.

(10) Van Speybroeck, V.; Van Der Mynsbrugge, J.; Vandichel, M.; Hemelsoet, K.; Lesthaeghe, D.; Ghysels, A.; Marin, G. B.; Waroquier, M. First Principle Kinetic Studies of Zeolite-Catalyzed Methylation Reactions. *J. Am. Chem. Soc.* **2011**, *133* (4), 888–899.

(11) Sun, J.; Thybaut, J. W.; Marin, G. B. Microkinetics of Methane Oxidative Coupling. *Catal. Today* **2008**, *137* (1), 90–102.

(12) Mhadeshwar, A. B.; Vlachos, D. G. Microkinetic Modeling for Water-Promoted CO Oxidation, Water-Gas Shift, and Preferential Oxidation of CO on Pt. *J. Phys. Chem. B* **2004**, *108* (39), 15246–15258.

(13) Peela, N. R.; Sutton, J. E.; Lee, I. C.; Vlachos, D. G. Microkinetic Modeling of Ethane Total Oxidation on Pt. *Ind. Eng. Chem. Res.* **2014**, *53* (24), 10051–10058.

(14) Deshmukh, S. R.; Mhadeshwar, A. B.; Vlachos, D. G. Microkinetic Modeling of Ammonia Synthesis and Decomposition on Ruthenium and Microreactor Design for Hydrogen Production. *ACS Div. Fuel Chem. Prepr.* **2003**, *48* (2), 936–937.

(15) Gu, G. H.; Mullen, C. A.; Boateng, A. A.; Vlachos, D. G. Mechanism of Dehydration of Phenols on Noble Metals via First-Principles Microkinetic Modeling. *ACS Catal.* **2016**, *6* (5), 3047–3055.

(16) Gu, G. H.; Wittreich, G. R.; Vlachos, D. G. Microkinetic Modeling of Aqueous Phase Biomass Conversion: Application to Ethylene Glycol Reforming. *Chem. Eng. Sci.* **2019**, *197*, 415–418.

(17) Christiansen, M. A.; Vlachos, D. G. Microkinetic Modeling of Pt-Catalyzed Ethylene Glycol Steam Reforming. *Appl. Catal. A Gen.* **2012**, *431–432*, 18–24.
(18) Sutton, J. E.; Vlachos, D. G. Building Large Microkinetic Models with First-Principles[U+05F3] Accuracy at Reduced Computational Cost. *Chem. Eng. Sci.* **2015**, *121*, 190–199.
(19) Mhadeshwar, A. B.; Wang, H.; Vlachos, D. G. Thermodynamic Consistency in Microkinetic Development of Surface Reaction Mechanisms. *J. Phys. Chem. B* **2003**, *107* (46), 12721–12733.
(20) Sutton, J. E.; Vlachos, D. G. Error Estimates in Semi-Empirical Estimation Methods of Surface Reactions. *J. Catal.* **2013**, *297*, 202–216.
(21) Sutton, J. E.; Guo, W.; Katsoulakis, M. A.; Vlachos, D. G. Effects of Correlated Parameters and Uncertainty in Electronic-Structure-Based Chemical Kinetic Modelling. *Nat. Chem.* **2016**, *8* (4), 331–337.
(22) Medford, A. J.; Wellendorff, J.; Vojvodic, A.; Studt, F.; Abild-Pedersen, F.; Jacobsen, K. W.; Bligaard, T.; Norskov, J. K. Assessing the Reliability of Calculated Catalytic Ammonia Synthesis Rates. *Science.* **2014**, *345* (6193), 197–200.
(23) Walker, E.; Ammal, S. C.; Terejanu, G. A.; Heyden, A. Uncertainty Quantification Framework Applied to the Water-Gas Shift Reaction over Pt-Based Catalysts. *J. Phys. Chem. C* **2016**, *120* (19), 10328–10339.
(24) Feng, J.; Lansford, J.; Mironenko, A.; Pourkargar, D. B.; Vlachos, D. G.; Katsoulakis, M. A. Non-Parametric Correlative Uncertainty Quantification and Sensitivity Analysis: Application to a Langmuir Bimolecular Adsorption Model. *AIP Adv.* **2018**, *8* (3), 035021.
(25) Gautier, S.; Steinmann, S. N.; Michel, C.; Fleurat-Lessard, P.; Sautet, P. Molecular Adsorption at Pt(111). How Accurate Are DFT Functionals? *Phys. Chem. Chem. Phys.* **2015**, *17* (43), 28921–28930.
(26) Benson, S. W.; Cruickshank, F. R.; Golden, D. M.; Haugen, G. R.; O'neal, H. E.; Rodgers, A. S.; Shaw, R.; Walsh, R.; Cruickshank, F. R.; Rodgers, A. S.; et al. Additivity Rules for the Estimation of Thermochemical Properties. *Chem. Rev.* **1969**, *69* (3), 279–324.
(27) Kua, J.; Goddard, W. A. Chemisorption of Organics on Platinum. 2. Chemisorption of C_2H_x and CH_x on Pt(111). *J. Phys. Chem. B* **1998**, *102* (47), 9492–9500.
(28) Gu, G. H.; Schweitzer, B.; Michel, C.; Steinmann, S. N.; Sautet, P.; Vlachos, D. G. Group Additivity for Aqueous Phase Thermochemical Properties of Alcohols on Pt(111). *J. Phys. Chem. C* **2017**, *121* (39), 21510–21519.
(29) Kresse, G.; Furthmüller, J. Efficient Iterative Schemes for *Ab Initio* Total-Energy Calculations Using a Plane-Wave Basis Set. *Phys. Rev. B* **1996**, *54* (16), 11169–11186.
(30) Steinmann, S. N.; Corminboeuf, C. A Generalized-Gradient Approximation Exchange Hole Model for Dispersion Coefficients. *J. Chem. Phys.* **2011**, *134* (4), 044117.
(31) Steinmann, S. N.; Corminboeuf, C. Comprehensive Benchmarking of a Density-Dependent Dispersion Correction. *J. Chem. Theory Comput.* **2011**, *7* (11), 3567–3577.
(32) Tang, Y.-H.; Zhang, D.; Karniadakis, G. E. An Atomistic Fingerprint Algorithm for Learning Ab Initio Molecular Force Fields. *J. Chem. Phys.* **2018**, *148* (3), 034101.
(33) Blöchl, P. E. Projector Augmented-Wave Method. *Phys. Rev. B* **1994**, *50* (24), 17953–17979.
(34) Joubert, D. From Ultrasoft Pseudopotentials to the Projector Augmented-Wave Method. *Phys. Rev. B - Condens. Matter Mater. Phys.* **1999**, *59* (3), 1758–1775.

(35) Monkhorst, H. J.; Pack, J. D. Special Points for Brillouin-Zone Integrations. *Phys. Rev. B* **1976**, *13* (12), 5188–5192.
(36) Efron, B.; Gong, G. A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation. *Am. Stat.* **1983**, *37* (1), 36–48.
(37) Stephens, M. A. EDF Statistics for Goodness of Fit and Some Comparisons. *J. Am. Stat. Assoc.* **1974**, *69* (347), 730–737.
(38) D'AGOSTINO, R.; PEARSON, E. S. Tests for Departure from Normality. Empirical Results for the Distributions of B2 and √b1. *Biometrika* **1973**, *60* (3), 613–622.
(39) Rangarajan, S.; Bhan, A.; Daoutidis, P. Language-Oriented Rule-Based Reaction Network Generation and Analysis: Description of RING. *Comput. Chem. Eng.* **2012**, *45*, 114–123.
(40) Linstrom, P. J.; Mallard, W. G. NIST Chemistry webBook, NIST Standard Reference Database Number 69 http://webbook.nist.gov/ (accessed 2019 -04 -01).
(41) Sutton, J. E.; Vlachos, D. G. A Theoretical and Computational Analysis of Linear Free Energy Relations for the Estimation of Activation Energies. *ACS Catal.* **2012**, *2* (8), 1624–1634.
(42) Wittreich, G. R.; Gu, G. H. Python Group Additivity: Database and Libraries https://pypi.org/project/pGrAdd/ (accessed 2021 -06 -01).
(43) Kee, R. J.; Rupley, F. M.; Miller, J. A. Chemkin-II: A Fortran Chemical Kinetics Package for the Analysis of Gas-Phase Chemical Kinetics. *J. Chem. Inf. Model.* **1989**, *53* (9), 1689–1699.
(44) Gu, G. H.; Vlachos, D. G. Group Additivity for Thermochemical Property Estimation of Lignin Monomers on Pt(111). *J. Phys. Chem. C* **2016**, *120* (34), 19234–19241.
(45) Gu, G. H.; Plechac, P.; Vlachos, D. G. Thermochemistry of Gas-Phase and Surface Species via LASSO-Assisted Subgraph Selection. *React. Chem. Eng.* **2018**, *3* (4), 454–466.
(46) Ogunnaike, B. A. *Random Phenomena: Fundamentals of Probability and Statistics for Engineers - CRC Press Book*, 1st ed.; CRC Press, 2009.
(47) PEARSON, K. "DAS FEHLERGESETZ UND SEINE VERALLGEMEINER-UNGEN DURCH FECHNER UND PEARSON*." A REJOINDER. *Biometrika* **1905**, *4* (1–2), 169–212.
(48) Murphy, K. P. *Machine Learning: A Probabilistic Perspective*; Adaptive Computation and Machine Learning series; MIT Press, 2012.
(49) Varoquaux, G.; Buitinck, L.; Louppe, G.; Grisel, O.; Pedregosa, F.; Mueller, A. Scikit-Learn. *GetMobile Mob. Comput. Commun.* **2015**, *19* (1), 29–33.
(50) Su, Y. Q.; Wang, Y.; Liu, J. X.; Filot, I. A. W.; Alexopoulos, K.; Zhang, L.; Muravev, V.; Zijlstra, B.; Vlachos, D. G.; Hensen, E. J. M. Theoretical Approach to Predict the Stability of Supported Single-Atom Catalysts. *ACS Catal.* **2019**, *9* (4), 3289–3297.
(51) Deshmukh, S. R.; Vlachos, D. G. A Reduced Mechanism for Methane and One-Step Rate Expressions for Fuel-Lean Catalytic Combustion of Small Alkanes on Noble Metals. *Combust. Flame* **2007**, *149* (4), 366–383.