RMG-to-KMC with ML, UQ, and Adaptivity

Khachik Sargsyan[1] Habib N. Najm[1] Joy Mueller[4]
Craig Daniels[1] Kyungjoo Kim[2]
Sevy Harris[3] and Richard West[3]

[1]Sandia National Laboratories, Livermore, CA

[2]Sandia National Laboratories, Albuquerque, NM

[3]Northeastern University, Boston, MA

[4]future Sandia National Laboratories, Livermore, CA

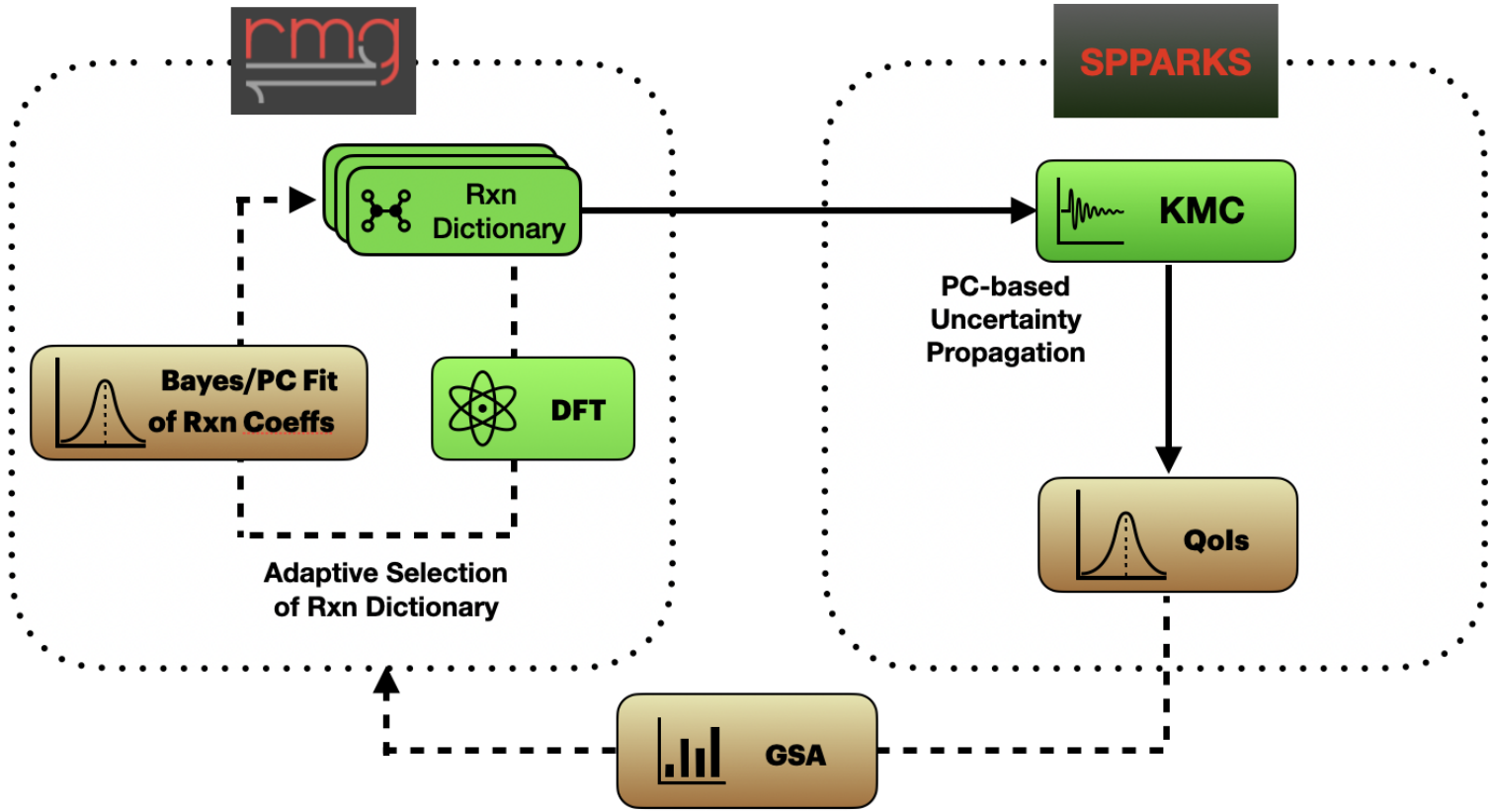
ECC-2021 Annual Meeting
Dec 8, 2021

Outline

outline

Introduction

intro



- First year's plan
  - Add UQ to parameter generation in RMG.
  - RMG API for adaptive KMC.
  - Extend SPPARKS for external online updates of reaction dictionary.
  - Construct and validate PC representation of uncertain rate constants.

Adaptive context

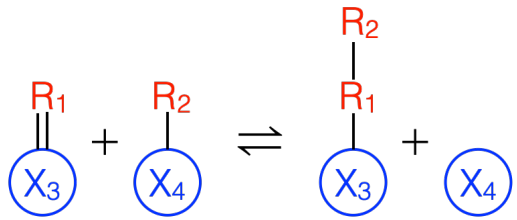
`adaptive`

- We want to build adaptivity in the KMC code framework
- Formulate a search strategy to add reactions to the mechanism
  - Follow a model growth strategy similar to RMG
  - Rely on (global) sensitivity analysis
- An outer loop proposes updating the model depending on computed solution with the previous model
  - Compute solution with current model
  - Propose reactions to be added or refined if an estimated rate of change of the solution over the time history is highly sensitive to them
    - Saddle point search strategies can provide guidance on proposing reactions
    - Transition state theory, DFT, RMG can provide rates
  - Update model and repeat, else terminate per suitable criteria

RMG predicts reactions using templates, and estimates rates using decision trees

rmg

Reaction templates predict possibility of reaction



$$CH_2* + H* \leftrightharpoons CH_3* + *$$

Decision trees estimate rates of reaction

- decision trees hand-crafted (with simple rules) or built automatically (from large database of training data)
- old trees had rates in Brønsted–Evans–Polanyi (or Arrhenius) form
- "new trees" have rate rules in Blowers and Masel form
- fitted to training reactions matching given node

The Blowers and Masel expression for kinetics, like BEP, varies barrier with enthalpy $Ea = f(\Delta H_{rxn})$

The Blowers and Masel expression is:

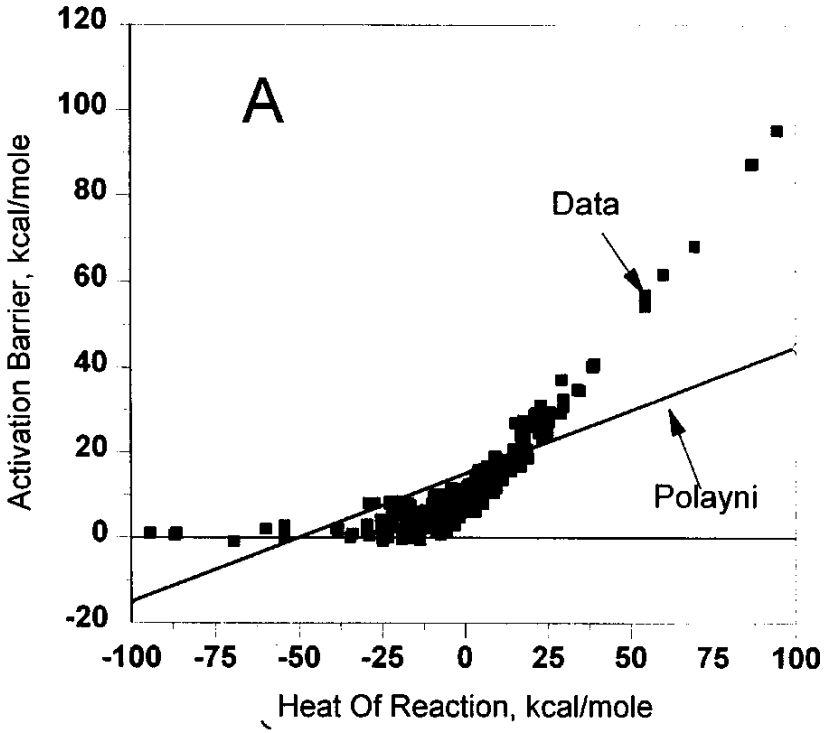
$$Ea = \begin{cases} 0 & \text{for } \Delta H_{rxn} < -4E_a^0 \\ \Delta H_{rxn} & \text{for } \Delta H_{rxn} > 4E_a^0 \\ \frac{\left(w_0 + \frac{\Delta H_{rxn}}{2}\right)(V_P - 2w_0 + \Delta H_{rxn})^2}{V_P^2 - 4w_0^2 + \Delta H_{rxn}^2} & \text{otherwise} \end{cases} \quad (1)$$

where

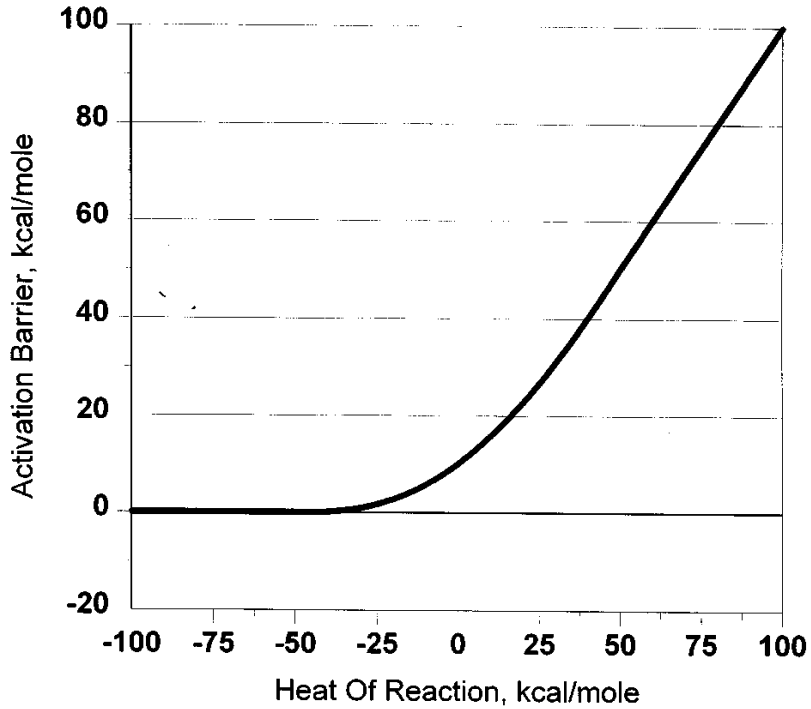
$$V_p = 2w_0 \frac{w_0 + E_a^0}{w_0 - E_a^0} \quad (2)$$

and w_0 is (from the derivation which applies to hydrogen transfer reactions) the average of the bond dissociation energy of the bond breaking and that being formed, *but doesn't matter much!*

The Blowers and Masel expression for kinetics, like BEP, varies barrier with enthalpy $Ea = f(\Delta H_{rxn})$

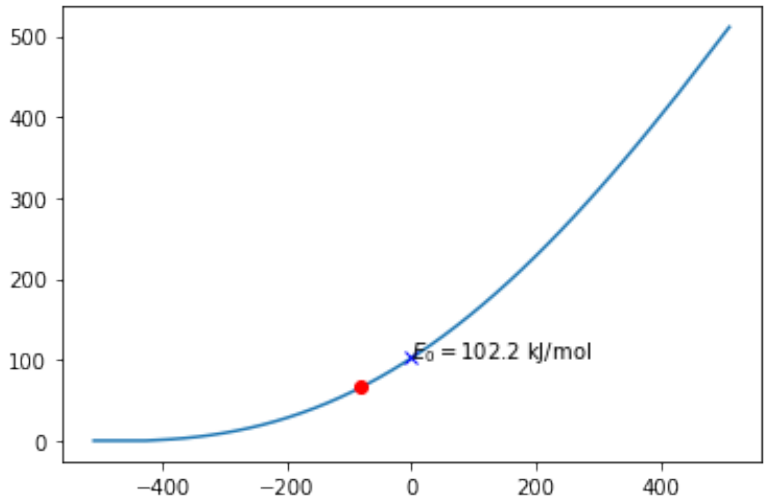


Data (H abstraction reactions) showing BEP is a poor fit.

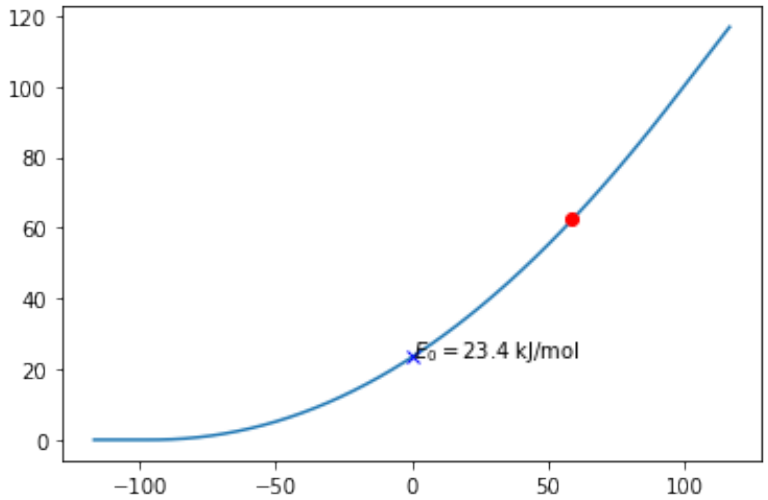


BM expression with $w_0 = 80, 100, 120$ kcal/mol (overlapping), showing it doesn't matter much.

The Blowers and Masel expression could allow barriers to be scaled, automatically, with adsorbate binding energies



For the reaction $HOX + HX \rightleftharpoons H_2OX + X$ with $dH = -80.7$ kJ/mol and $E_a = 66.2$ kJ/mol we derive $E_a^0 = 102.2$ kJ/mol and conclude that a +10 kJ/mol increase in the ΔH_{rxn} (making it less exothermic) would increase the E_a by +4.0 kJ/mol.



For the reaction $H_2O + 2\,X \rightleftharpoons HOX + HX$ with $dH = 59.0$ kJ/mol and $E_a = 62.3$ kJ/mol we derive $E_a^0 = 23.4$ kJ/mol and conclude that a +10 kJ/mol increase in the ΔH_{rxn} (making it more endothermic) would increase the E_a by +8.4 kJ/mol.

Parts of RMG will operate as a service with an API for KMC

- Possible interface
  - in: neighboring adsorbates (eg. HX and CO_2X)
  - out: possible reaction products (eg. CHOOX and X)
  - out: rate of reaction
  - out: enthalpy of reaction (or of species)
- Considerations
  - Time wasted converting strings to graphs, then molecules, etc.
  - Memory footprint
  - parallelization strategies
  - message-passing, server model, API, etc.

UQ intro: Polynomial chaos

pcuq

$$U \simeq \sum_{k=0}^{p} u_k \psi_k(\xi)$$

- Represent input parameters and output QoIs as random variables
- Random variables written as polynomials of standard r.v.'s
- Describes a r.v. U with a vector of *PC modes* $(u_0, u_1, \ldots, u_p)$
- Selection of PC type and order p is a modeling choice
- Standard r.v. ξ, standard orthogonal polynomials $\psi_k(\xi)$, *e.g.*

If ξ Uniform$[-1, 1]$, $\psi_k(\xi)$ are Legendre polynomials,

or

If ξ Normal$(0, 1)$, $\psi_k(\xi)$ are Hermite polynomials

[Wiener, 1938; Ghanem & Spanos, 1991; Xiu & Karniadakis, 2002; Le Maître & Knio, 2010]

Essential use of PC in UQ

$$U \simeq \sum_{k=0}^{K} u_k \Psi_k(\boldsymbol{\xi})$$

pc-use

Strategy:

- Represent model parameters/solution as random variables
- Construct PC for uncertain parameters
- Evaluate PC for model outputs

Advantages:

- Computational efficiency
- Utility
  - Moments: $\mathbb{E}[u] = u_0,\ \mathbb{V}[u] = \sum_{k=1}^{K} u_k^2 ||\Psi_k||^2,\ \ldots$
  - Global Sensitivities – variance decomposition, Sobol' indices
  - Uncertainty propagation
  - Surrogate for forward model

Requirements:

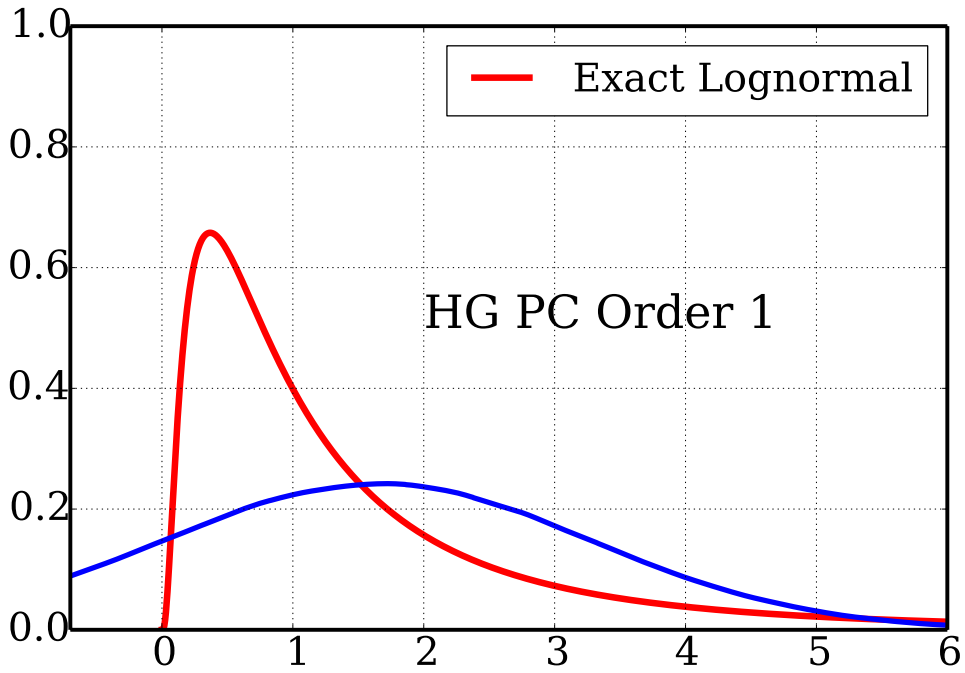
- Finite variances (not a handicap in practice)
- Smooth forward functions

Demo 1D PC

`pc_logn`

$$U \simeq \sum_{k=0}^{p} u_k \psi_k(\xi)$$

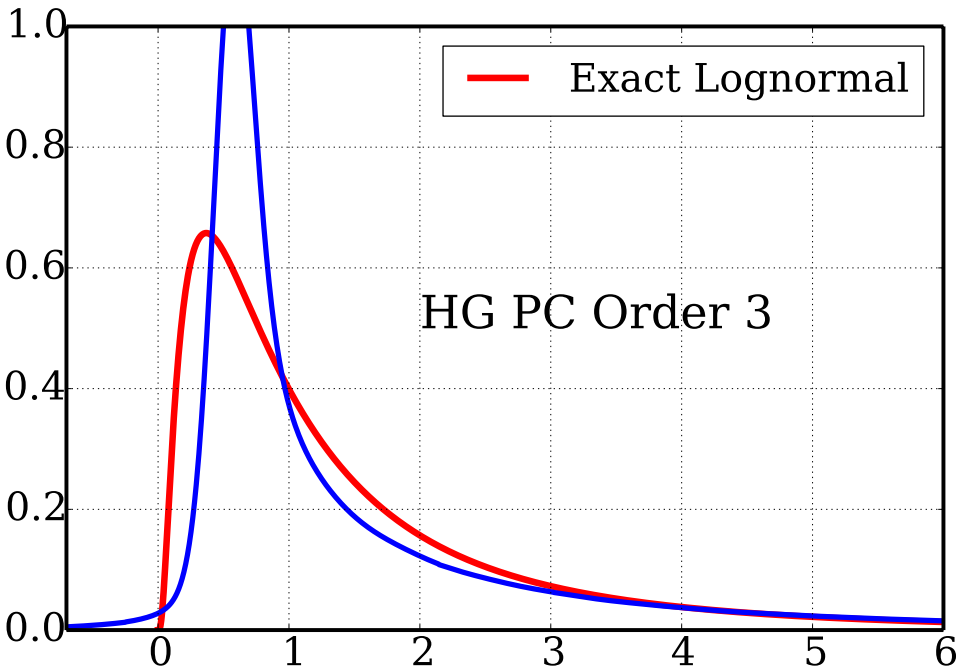
$$U = u_0 + u_1 \underbrace{\xi}_{\psi_1(\xi)}$$



Demo 1D PC

$$U \simeq \sum_{k=0}^{p} u_k \psi_k(\xi)$$

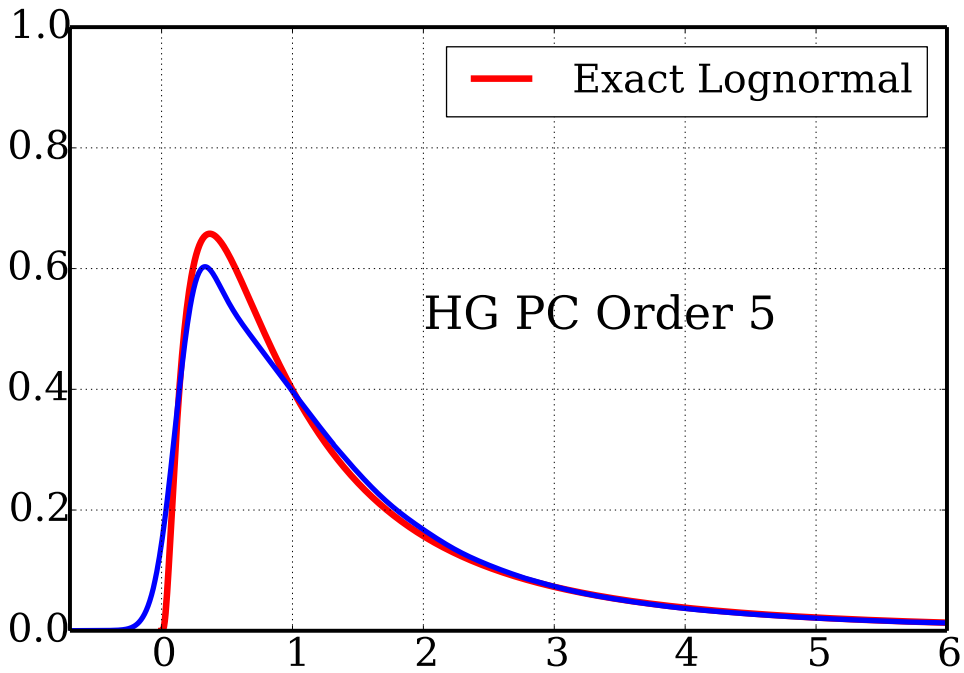
$$U = u_0 + u_1 \underbrace{\xi}_{\psi_1(\xi)} + u_2 \underbrace{(\xi^2 - 1)}_{\psi_2(\xi)} + u_3 \underbrace{(\xi^3 - 3\xi)}_{\psi_3(\xi)}$$



Demo 1D PC

$$U \simeq \sum_{k=0}^{p} u_k \psi_k(\xi)$$

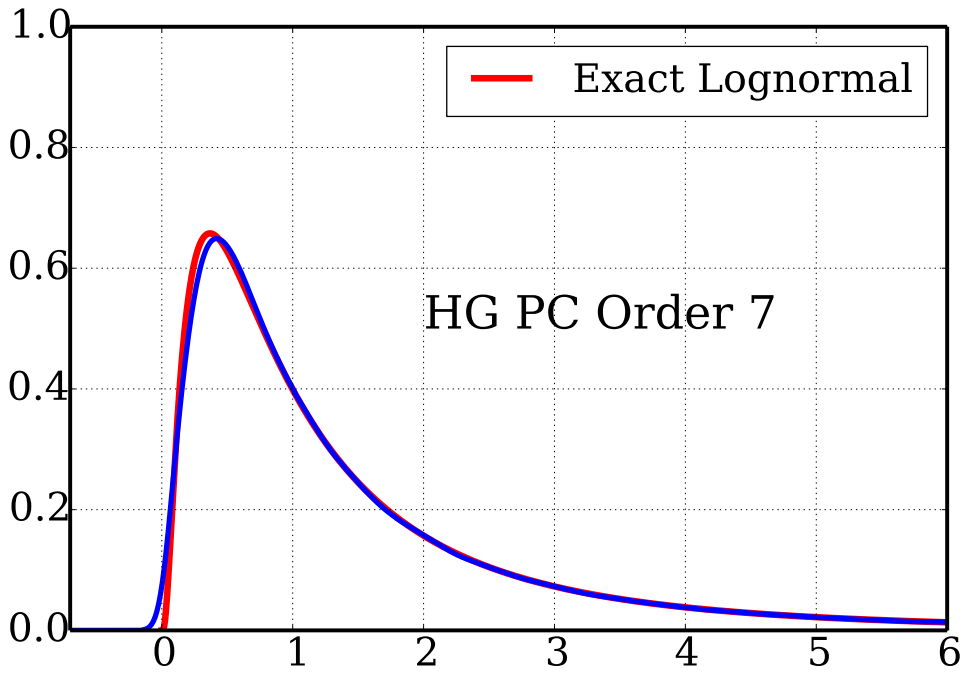
$$U = u_0 + u_1 \underbrace{\xi}_{\psi_1(\xi)} + u_2 \underbrace{(\xi^2 - 1)}_{\psi_2(\xi)} + u_3 \underbrace{(\xi^3 - 3\xi)}_{\psi_3(\xi)} + u_4 \underbrace{(\xi^4 - 6\xi^2 + 3)}_{\psi_4(\xi)} + u_5 \underbrace{(\xi^5 - 10\xi^3 + 15\xi)}_{\psi_5(\xi)}$$



Demo 1D PC

$$U \simeq \sum_{k=0}^{p} u_k \psi_k(\xi)$$

$$U = u_0 + u_1 \underbrace{\xi}_{\psi_1(\xi)} + u_2 \underbrace{(\xi^2 - 1)}_{\psi_2(\xi)} + u_3 \underbrace{(\xi^3 - 3\xi)}_{\psi_3(\xi)} + u_4 \underbrace{(\xi^4 - 6\xi^2 + 3)}_{\psi_4(\xi)} + u_5\psi_5(\xi) + u_6\psi_6(\xi) + u_7\psi_7(\xi)$$



Multivariate Polynomial Chaos is a simple extension

pcuq3

- In general, d physical variables $\boldsymbol{U}$ depending on $\tilde{d}$ stochastic inputs $\boldsymbol{\xi}$

$$U_i = \sum_j u_{ij} \Psi_j(\boldsymbol{\xi}), \text{ for } i = 1, \dots, d.$$

- Multivariate normal is a special case: it is a first order, Gauss-Hermite PC

$$\begin{cases} U_1 = u_{10} + u_{11}\xi_1 \\ U_2 = u_{20} + u_{21}\xi_1 + u_{22}\xi_2 \\ \quad \vdots \qquad\qquad \vdots \\ U_d = u_{d0} + u_{d1}\xi_1 + \dots u_{dd}\xi_d \end{cases}$$

- Jumping ahead: ln k fitting as a linear regression leads to this form

PC features: moment extraction and global sensitivity analysis

`pc-use`

$$U \simeq \sum_{k=0}^{K} u_k \Psi_k(\boldsymbol{\xi})$$

- Moments:
  - Expectation: $\mathbb{E}[U] = u_0$
  - Variance $\mathbb{V}[U] = \sum_{k=1}^{K} u_k^2 ||\Psi_k||^2$
- Global sensitivity analysis (also known as Sobol indices or variance-based decomposition)
  - Main effect sensitivity indices

$$S_i = \frac{\mathbb{V}[\mathbb{E}(U(\boldsymbol{\xi}|\xi_i)]}{\mathbb{V}[U(\boldsymbol{\xi})]} = \frac{\sum_{k \in \mathbb{I}_i} u_k^2 ||\Psi_k||^2}{\sum_{k>0} u_k^2 ||\Psi_k||^2}$$

    ($\mathbb{I}_i$ is the set of bases with only ξ_i involved, and S_i is the uncertainty contribution that is due to i-th parameter only).
  - Total effect sensitivity indices

$$T_i = 1 - \frac{\mathbb{V}[\mathbb{E}(U(\boldsymbol{\xi}|\xi_{-i})]}{\mathbb{V}[U(\boldsymbol{\xi})]} = \frac{\sum_{k \in \mathbb{I}_i^T} u_k^2 ||\Psi_k||^2}{\sum_{k>0} u_k^2 ||\Psi_k||^2}$$

    ($\mathbb{I}_i^T$ is the set of bases with ξ_i involved, including all its interactions).

Uncertainty quantification in RMG

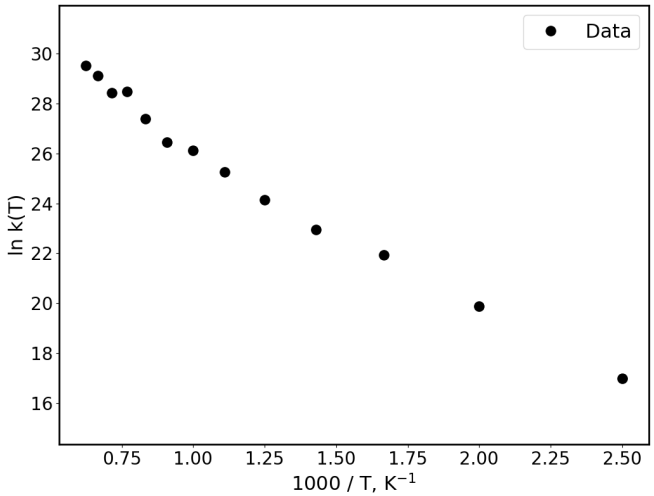
`rmg_kfit`

- For example, consider the modified Arrhenius equation

$$k(T) = A \left(\frac{T}{T_0} \right)^n e^{-\frac{E_a}{RT}},$$

where T_0 and R are known.

- Given $k(T_i)$ at various temperature conditions T_i, need to estimate (A, n, E_a).
- When in doubt, log it!



$$\ln k(T) = \ln(A) + n \ln \left(\frac{T}{T_0} \right) + E_a \frac{-1}{RT}$$

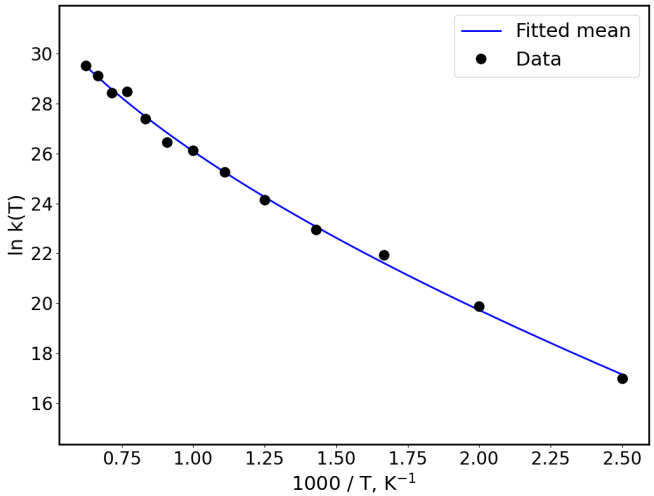
Uncertainty quantification in RMG

- For example, consider the modified Arrhenius equation

$$k(T) = A \left(\frac{T}{T_0} \right)^n e^{-\frac{E_a}{RT}},$$

where T_0 and R are known.

- Given $k(T_i)$ at various temperature conditions T_i, need to estimate (A, n, E_a).
- When in doubt, log it!



$$\ln k(T) = \underbrace{\ln A}_{\alpha} + \underbrace{n}_{\beta} \ln \left(\frac{T}{T_0} \right) + \underbrace{E_a}_{\gamma} \frac{-1}{RT}$$

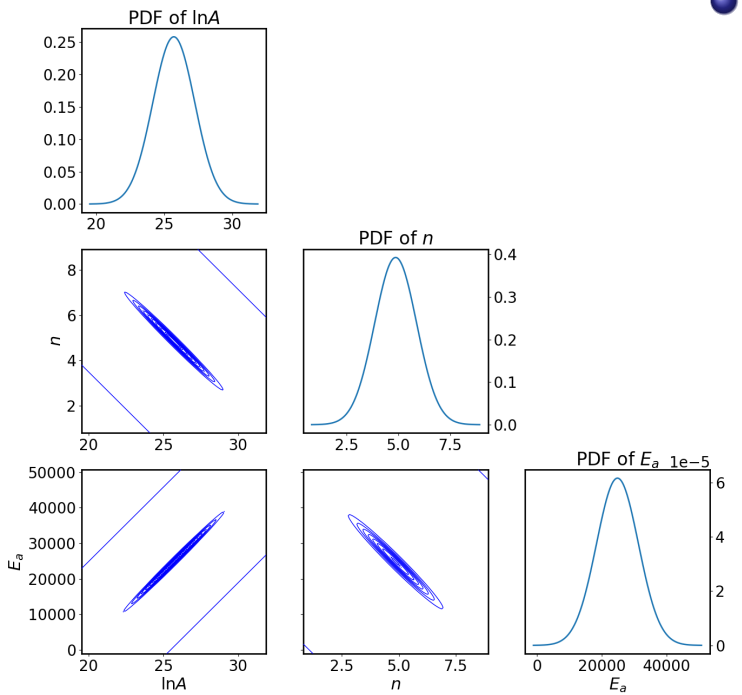
- Linear regression problem to find $(\alpha, \beta, \gamma) = (\ln A, n, E_a)$

Bayesian linear regression (or least-squares)

`bayesfit`

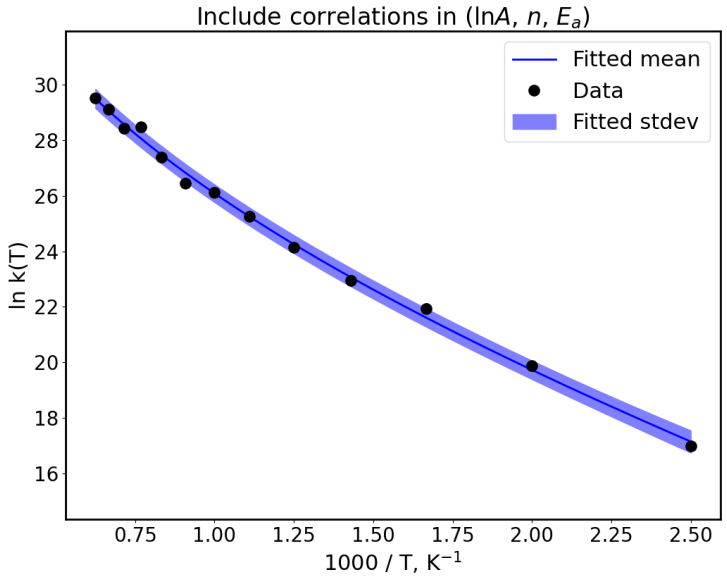
- Linear regression problem to find $\boldsymbol{\theta} = (\alpha, \beta, \gamma) = (\ln A, n, E_a)$

$$\ln k(T) = \alpha + \beta \ln\left(\frac{T}{T_0}\right) + \gamma \frac{-1}{RT}$$



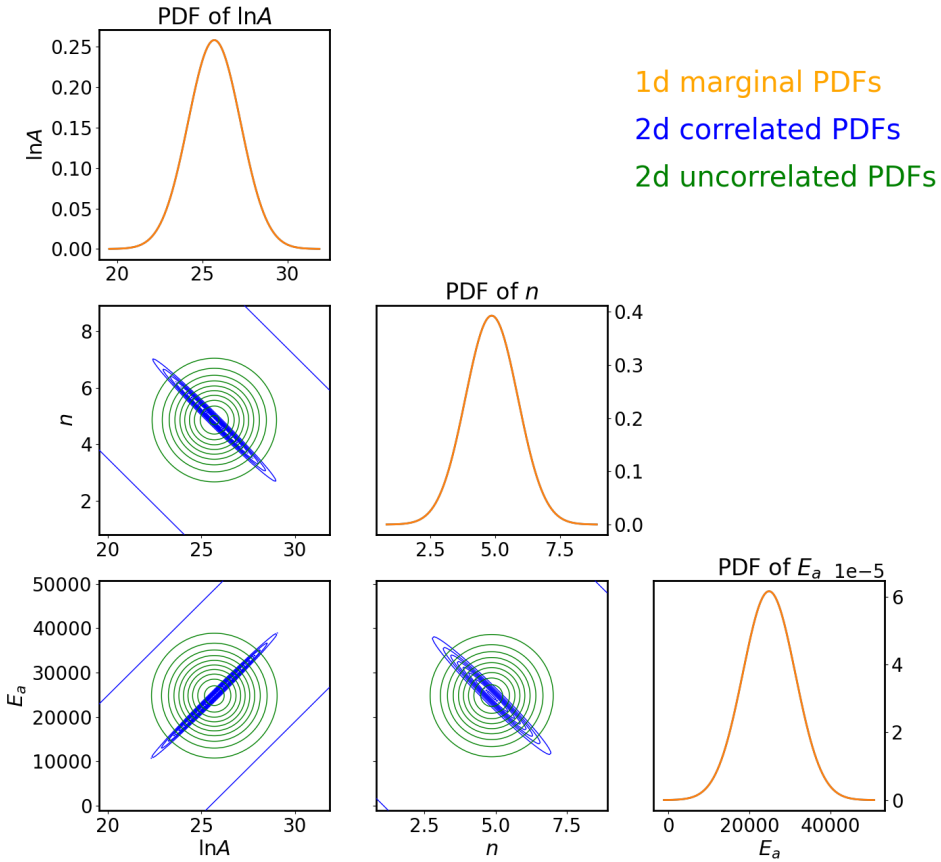
- Closed form solution available:

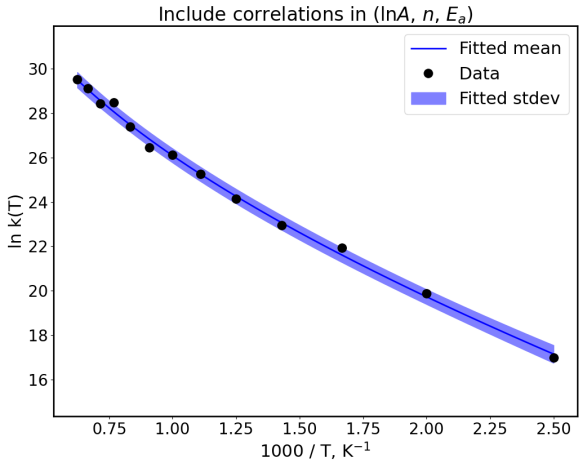
Mean	$\mu_{\boldsymbol{\theta}} = (\ln A^*, n^*, E_a^*)$
Covariance	$\Sigma_{\boldsymbol{\theta}}$

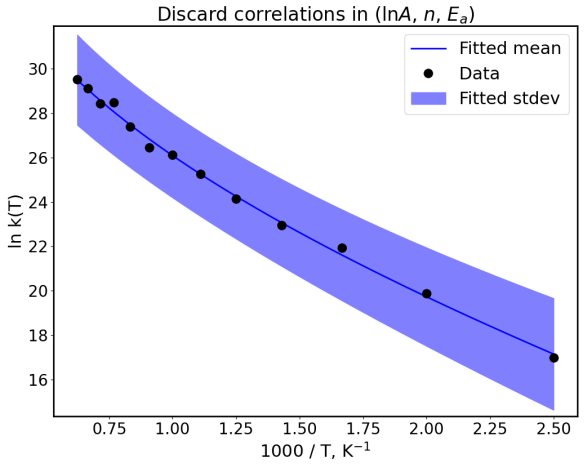


Ignoring correlations leads to overestimation of uncertainties

`corr`







Represent rate coefficients as PC

`pcrmg`

... e.g., in the Arrhenius case:

$$\ln k(T;\boldsymbol{\xi}) = \ln A(\boldsymbol{\xi}) + n(\boldsymbol{\xi}) \ln\left(\frac{T}{T_0}\right) + E_a(\boldsymbol{\xi})\frac{-1}{RT}$$

where $\boldsymbol{\xi} = (\xi_1, \xi_2, \xi_3)$.

$$\begin{cases} \ln A(\boldsymbol{\xi}) = u_{10} + u_{11}\xi_1 \\ n(\boldsymbol{\xi}) = u_{20} + u_{21}\xi_1 + u_{22}\xi_2 \\ E_a(\boldsymbol{\xi}) = u_{30} + u_{31}\xi_1 + u_{32}\xi_2 + u_{33}\xi_3 \end{cases}$$

Leading to

$$\begin{aligned} \ln k(\boldsymbol{\xi};T) &= u_{10} + u_{11}\xi_1 + (u_{20} + u_{21}\xi_1 + u_{22}\xi_2)\ln\left(\frac{T}{T_0}\right) + \\ &+ (u_{30} + u_{31}\xi_1 + u_{32}\xi_2 + u_{33}\xi_3)\frac{-1}{RT} \end{aligned}$$

- Now propagate this through KMC.

Pipe RMG to KMC (with uncertainties)

`kmcuq`

- Recall PC-based uncertainty propagation task: given PC for inputs, find PC for outputs.

$$U \simeq \sum_{k=0}^{K} u_k \Psi_k(\boldsymbol{\xi}) \qquad Z = f(U) \simeq \sum_{k=0}^{K} c_k \Psi_k(\boldsymbol{\xi})$$

- In our case, there are PC expansions for all input $\ln k$'s, e.g.

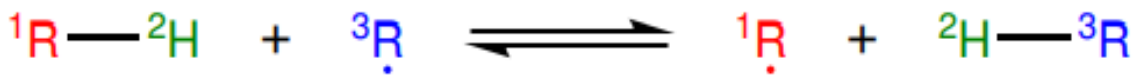
$$\ln k(\boldsymbol{\xi}; T) = u_0(T) + u_1(T)\xi_1 + u_2(T)\xi_2 + u_3(T)\xi_3$$

- Independent k_m's:
  - $U = (\ln k_1, \ldots, \ln k_M)$, i.e. each rate constant is written as a linear PC expansion $\ln k_m = \ln k_m(\xi_{m1}, \xi_{m2}, \xi_{m3})$
  - invoke group GSA to measure sensitivities with respect to k_m's
- Dependent k_m's:
  - Will need PC representation of the correlated ucnertainty, presented next.

RMG Kinetics Estimation (decision trees)

`RMG-old-estimation`

- Database organized into decision trees [Gao, Liu, Green, 2020]: starting from a highest-abstraction root.
- If no exact match found in database, reaction is matched to a family (e.g. H Abstraction) using templates



- Each family has a tree of nodes representing groups, or molecular substructures describing the reacting sites
- Ideally each node also contains an estimate of kinetic parameters corresponding reaction that matches the group structure, but not every node has data
- If a node is matched but has no data, kinetics are estimated from the parent node, or an average of its children if it has no data – this is how uncertainty is estimated empirically.

Uncertainty Correlation Estimation in RMG

`RMG-uncertainty-old`

- Uncertainty of estimate can be broken down into two parts:
  - Estimation error - due to imperfection of applying rate to wrong reaction
  - Intrinsic error - due to the uncertainty of the individual rate rules averaged together to estimate the rate (described earlier)

$$\Delta \ln k_{\text{rate rules}} = \overbrace{\Delta \ln k_{\text{family}} + \log_{10}(N+1)\left(\Delta \ln k_{\text{non-exact}}\right)}^{\text{Estimation}} + \overbrace{\sum_{\text{rule } i} w_i \Delta \ln k_{\text{rule},}}^{\text{Intrinsic}}$$

- Rate estimations may be averages of rates, so correlation comes from overlap of rate sources.
- Uncertainty covariance matrix element:

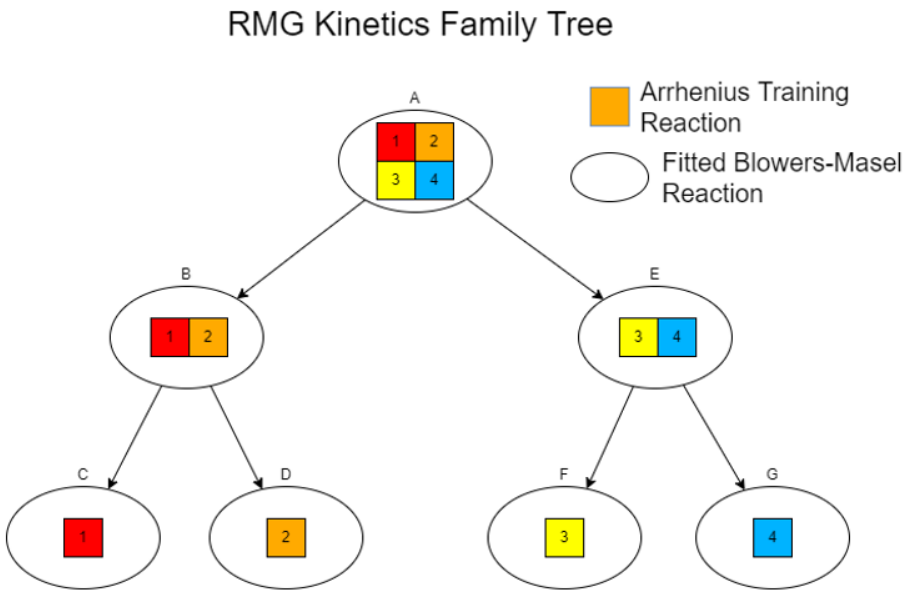
$$\Sigma_{ab} = \text{cov}\left(\ln k_{rate\ rules\ a}, \ln k_{rate\ rules\ b}\right)$$

$$\Sigma_{ab} = \left[\sum_{rule_i}\sum_{rule_j} \omega_{a,i}\omega_{b,j}\text{cov}(\ln(k_{\text{rule,i}}), \ln(k_{\text{rule,j}}))\right] + \text{cov}(\ln k_{\text{family,a}}, \ln k_{\text{family,b}}) + \log_{10}(N_a+1)\log 10(N_b+1)\text{cov}(\ln k_{\text{non-exact,a}}, \ln k_{\text{non-exact,b}})$$

Automated RMG Kinetics Estimation via Blowers-Masel Expression (new trees)

`RMG-kinetics-new`

- RMG family trees are transitioning to automatic tree generation
- Every node/group has data (corresponding Arrhenius training reactions)
- Reactions match a single node instead of averaging

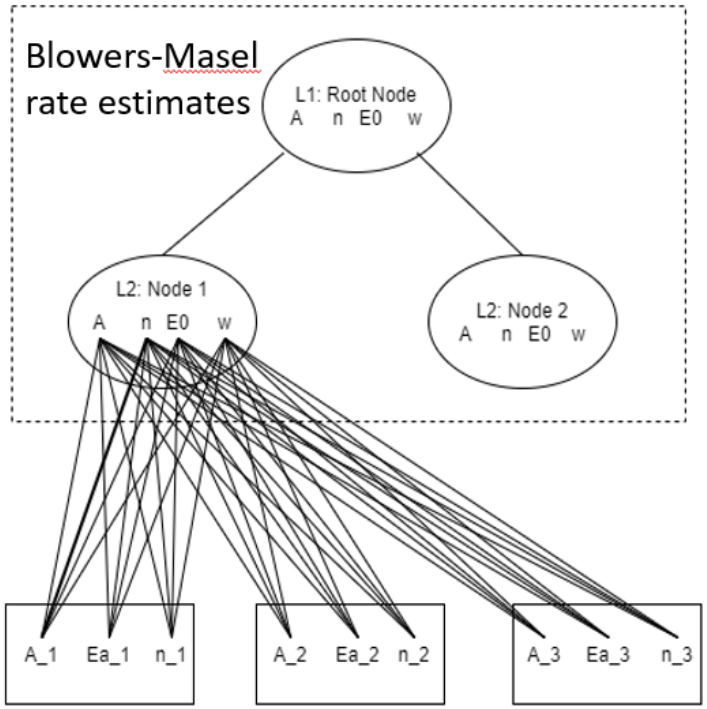


- Kinetics for each node are described by a Blowers-Masel reaction (4 parameter fit similar to Arrhenius) fitted from Arrhenius training reactions
- Uncertainty estimated from variance of leave-one-out estimates

Uncertainty correlations are extracted using local sensitivities

`RMG-correlated-new`

- Correlation of nodes is not straightforward because estimated kinetic parameters come from a curve fit, so it is estimated using local sensitivity analysis
- For each node, sensitivity can be computed for each of the 4 Blowers-Masel kinetic parameters with respect to the 3N training reaction Arrhenius parameters
  - E.g. for the pre-factor:



Arrhenius Training Reactions

$$\text{Sens} = \frac{dA/A}{dA_1/A_1} = \frac{\text{Resulting Perturb. in BM}}{\text{Perturb. in Arrhenius}}$$

PC surrogate construction: steps

non-intrusive

KMC Output $Q(\boldsymbol{k}(\boldsymbol{\xi})) = \sum_j z_j \Psi_j(\boldsymbol{\xi})$

- Construct PC's for reaction rates $\boldsymbol{k}(\boldsymbol{\xi})$
- Sample $\boldsymbol{\xi}$'s
- Compute reaction rates $\boldsymbol{k}(\boldsymbol{\xi})$'s for all $\boldsymbol{\xi}$'s
- Evaluate KMC QoIs $Q(\boldsymbol{k}(\boldsymbol{\xi}))$ for each sample
- Build PC for each QoI: regression/projection to find PC modes z_j's
- *Postprocess:* PDF generation (e.g. to estimate tail probabilities), global sensitivity analysis (e.g. to attribute overall uncertainty to specific reactions), predictive confidence in catalyst performance

Global sensitivity analysis or variance decomposition

`kmcuq_gsa`

KMC Output $Q(\boldsymbol{k}(\boldsymbol{\xi})) = \sum_j z_j \Psi_j(\boldsymbol{\xi})$

- Having constructed the PC expansion above, one can decompose output variance and attribute uncertainties to reactions or groups of reactions

$$\begin{aligned}\text{Var}[Q(k_1, \dots, k_r)] &= V_{k_1} + V_{k_2} + \cdots + V_{k_r} + \\ &+ V_{k_1,k_2} + V_{k_1,k_3} + \cdots + V_{k_{r-1},k_r} + \\ &+ V_{k_1,k_2,k_3} + \cdots + V_{k_{r-2},k_{r-1},k_r} + \dots\end{aligned}$$

- This is a free bi-product of the PC expansion
- Useful feedback to RMG:
  - focus on non-important: safely ignore uncertainties in some reactions, reduces dimensionality
  - focus on important: collect better data for them to reduce uncertainty

Multifidelity PC will help gain efficiency

`kmcuq_mf`

KMC Output $Q(\boldsymbol{k}(\boldsymbol{\xi})) = \sum_j z_j \Psi_j(\boldsymbol{\xi})$

- The key is to construct as accurate surrogate as possible with fewest possible number of $\boldsymbol{\xi}$'s (KMC evaluations)
- Given N evaluations of low-fi $Q_0(\boldsymbol{\xi})$ and $M \ll N$ evaluations of high-fi $Q(\boldsymbol{\xi})$, it is possible to construct accurate PC surrogate

$$Q(\boldsymbol{\xi}) = \underbrace{Q_0(\boldsymbol{\xi})}_{\text{low-fi, lots of evals}} + \underbrace{Q(\boldsymbol{\xi}) - Q_0(\boldsymbol{\xi})}_{\text{high-fi, few evals, but easier to learn}}$$

- Can generalize to > 2 levels/fidelities
- Where to get multiple fidelities?
  - Different lattice sizes and/or reaction/process models
  - Different no. of statistical samples & averaging time-windows
  - Different levels of approximation in parallelization

Timeline

`timeline`

- Year 1:
  - Add UQ to parameter generation in RMG.
  - RMG API for adaptive KMC.
  - Extend SPPARKS for external online updates of reaction dictionary.
  - Construct and validate PC representation of uncertain rate constants.
- Year 2:
  - Develop adaptive KMC with SPPARKS using RMG to update reaction lists on-the-fly.
  - Demonstrate uncertainty propagation through KMC, and GSA for selection of sensitive reactions.

Timeline, cont.

timeline2

- Year 3:
  - Demonstrate effective adaptive KMC coupling of SPPARKS, UQ/GSA, and RMG.
  - Extend UQ with multifidelity approximations with varying lattice sizes.

- Year 4:
  - Demonstrate performance of multifidelity UQ/GSA for adaptive KMC with Kokkos-KMC RMG coupling, and requisite recourse to higher-fidelity pynta-AdTherm rate-constant computations.

Closure and Discussion

closure

- Summary
  - Adaptive framework for reaction selection
  - Propagation of RMG uncertainties through KMC
  - Uncertainty attribution, sensitivity analysis, feedback
- Challenges
  - Adaptive framework will have devil in details
  - Representation of correlated uncertainties in RMG rate coefficient fitting
  - Selection of meaningful KMC QoIs and incorporation of sampling/averaging noise
  - High-dimensionality (large number of species and reactions)
  - Computationally expensive KMC