

**2-2**

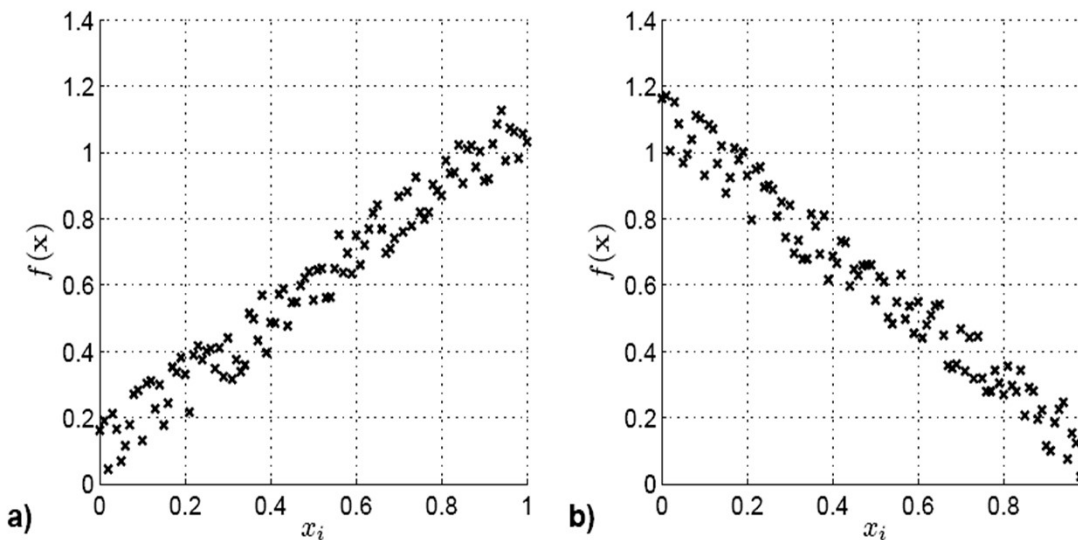
**CALCULATING SENSITIVITY  
INDICES FROM SAMPLING  
METHODS**

# What do global sensitivity indices represent?

- There are different global sensitivity indicators available, but the most commonly used attempts to attribute how much of the overall predicted output variance results from the uncertainties in the most important input parameters.
- This should include individual parameter effects (linear and non-linear) as well as interactive effects between parameters (non-linear responses).
- If we want to include higher order effects (3<sup>rd</sup> and higher) then the sample size required is likely to be huge.
  - Chemical systems rarely exhibit large 3<sup>rd</sup> order and higher responses.

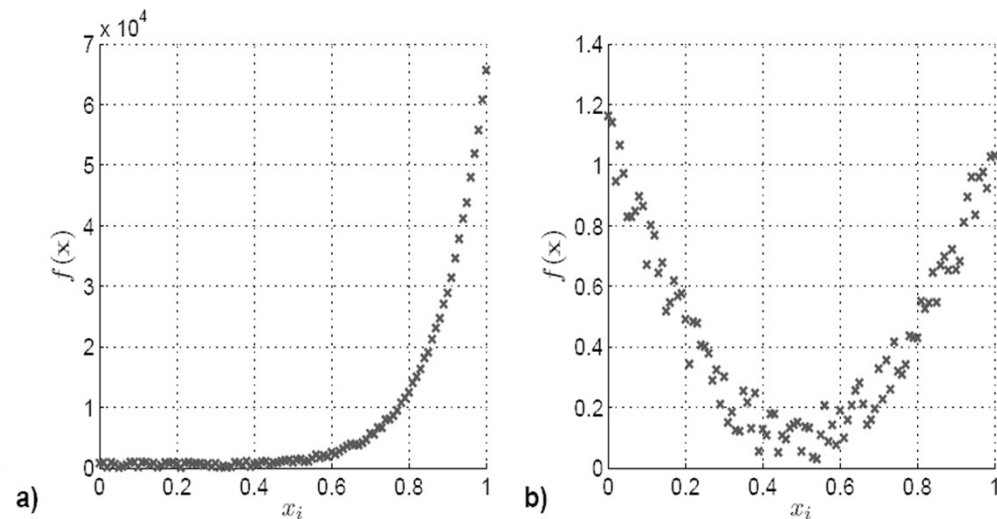
# Scatter plots

- As a simple starting point we can plot the output response to changes in a parameter as a scatter plot.
- Basically a projection from the higher dimensional space to a 2D plot.



Strong positive (left) and negative (right) linear responses to the chosen parameter.

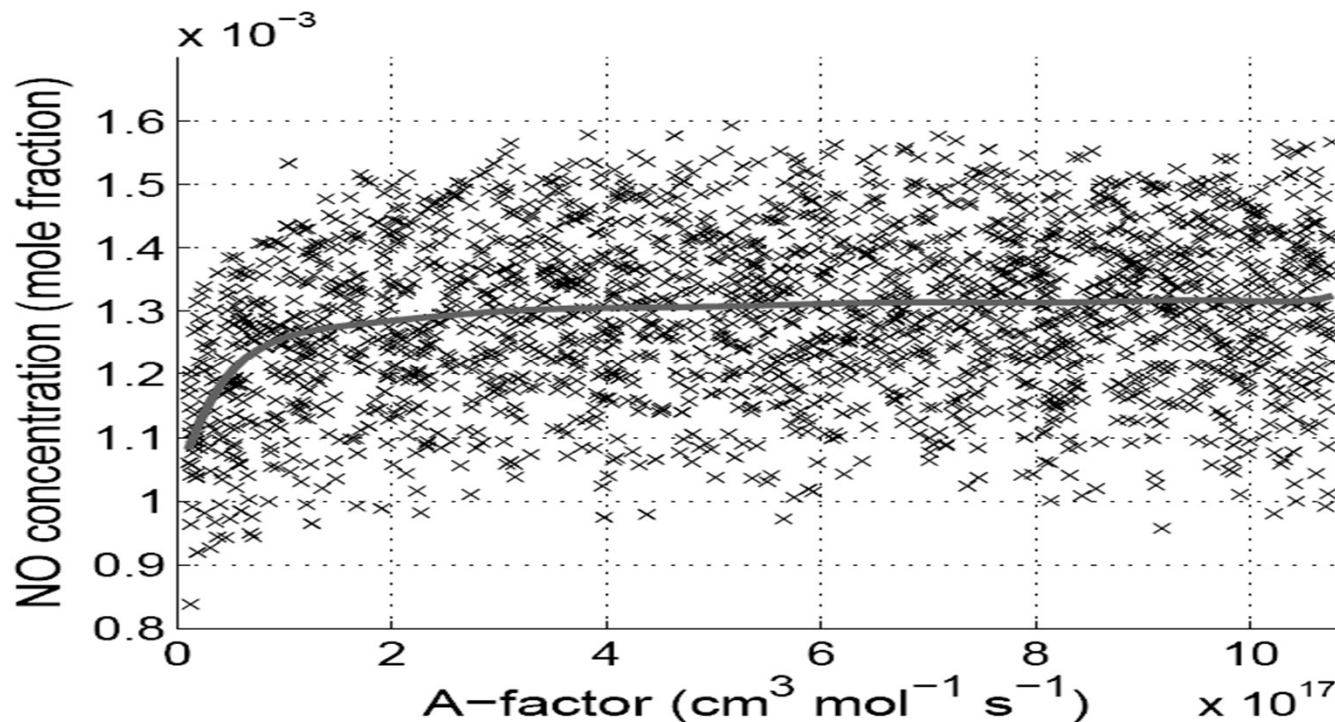
Adapted from ([Ziehn 2008](#))



Strong monotonic (left) and non-monotonic (right) non-linear responses to the chosen parameter. Adapted from ([Ziehn 2008](#))

# Real chemical systems

- Example from NO predictions in a methane flame.
- Effect seems to saturate at higher parameter values - hence sensitivity is clearly not linear.
- Scatter from the effects of other parameter uncertainties clouds the main effect from this A factor.



- Clear to see the potential issues with using linear methods at a single nominal value.
- How do we determine the effect of this A factor from within the scatter?

*(Ziehn, Tomlin 2008)*

# Pearson and Spearman rankings

- Pearson correlation coefficient ( $r$ ) is a measure of the strength of the linear relationship between two variables (e.g. parameter  $x$  and target output  $y$ ), ranging from  $-1$  for a perfect negative correlation to  $+1$  for a perfect positive correlation.
- Calculated by dividing the covariance of the variables by the square root of the product of their variances:

$$r_{xy} = \frac{\sum_{k=1}^m (x_k - \bar{x})(y_k - \bar{y})}{\left[ \sum_{k=1}^m (x_k - \bar{x})^2 \right]^{1/2} \left[ \sum_{k=1}^m (y_k - \bar{y})^2 \right]^{1/2}}$$

- Not very useful for non-linear responses.

# Spearman ranking

- Can be thought of as the Pearson correlation coefficient between ranked variables.
- Data are replaced with their corresponding ranks and then correlation procedures are performed on these ranks.
- The Spearman coefficient therefore assesses how well the relationship between two variables can be described using a monotonic function (Saltelli et al. 2000).
- A Spearman correlation of  $+1$  or  $-1$  therefore occurs when one variable is a perfect monotone function of the other.
- Correlation coefficients should really only be used as a guideline for parameter importance rather than in a strictly quantified way.

# Variance based methods: Sobol's original method

- Is a sampling based method to calculate **fraction of total variance** that can be **attributed to each parameter** in a joint pdf distribution.
- If the model result  $Y_i = f_i(x_1, x_2, \dots, x_N)$  is influenced by independent random parameters, then the joint pdf of the parameters  $P(x_1, x_2, \dots, x_N) = \prod_{j=1}^N p_j(x_j)$ .

- The mean or expected value  $E(Y_i)$  of the calculated result  $Y_i$  is then given by:
$$E(Y_i) = \iiint \dots \int f_i(x_1, x_2, \dots, x_N) \prod_{j=1}^N p_j(x_j) dx_j$$

- while the variance  $V(Y_i)$  of the calculated result  $Y_i$  is specified as:

$$\begin{aligned} V(Y_i) &= \iiint \dots \int (f_i(x_1, x_2, \dots, x_N) - E(Y_i))^2 \prod_{j=1}^N p_j(x_j) dx_j \\ &= \iiint \dots \int f_i^2(x_1, x_2, \dots, x_N) \prod_{j=1}^N p_j(x_j) dx_j - E^2(Y_i) \end{aligned}$$

- If integral calculated with **fixed value of parameter  $x_j$** , variance caused by all other parameters except for  $x_j$ ,  $V(Y_i/x_j)$  obtained.
- If  $V(Y_i/x_j)$  calculated for many values of  $x_j$ , selected according to its pdf, then expected value  $E(V(Y_i/x_j))$  can be calculated.
- Requires integration of  $V(Y_i/x_j)$  over pdf of  $x_j$  (Saltelli et al., 2002).
- $V(Y_i) - E(V(Y_i/x_j))$  equal to reduced variance of  $Y_i$  caused by fixing value of  $x_j$ , and is equal to  $V(E(Y_i/x_j))$ .
- By dividing this **conditional variance** by **unconditional variance**, the first-order sensitivity index for parameter  $x_j$  can be calculated:

$$S_{j(i)} = \frac{V(E(Y_i|x_j))}{V(Y_i)}$$

- Shows **the fraction of the total variance of  $Y_i$  which is reduced when the value of  $x_j$  is held at a fixed value** – a measure of the influence of uncertainty in  $x_j$ .
- The calculation of integrals is non-trivial and the use of a Monte Carlo sampling method (Saltelli et al., 2002) requires **N (2m+1) model runs for first-order indices** where N is the sample size chosen for the Monte Carlo estimates.

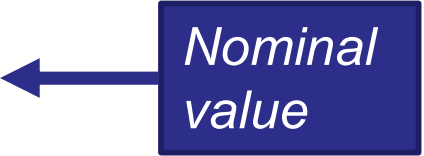


# Response Surface Methods, RSM

- RSM based methods attempt to reduce computational cost of Variance based sensitivity methods by first developing a fitted **meta-model** accurately representing **relationship between model parameters and outputs**.
- If **meta-model** can be fitted with a lower number of model runs then it can be used to calculate variance based indices at lower cost.
- Some similarities with Monte Carlo approaches:
  - first input parameter ranges must be selected
  - then a **suitable sampling approach** taken so that full model runs are obtained across a design suitable for development of accurate meta-model.
- Cost of method driven by cost of providing accurate surrogate model.
- This is not always dependant on size of scheme but is **driven by the complexity of the response surface**.
  - Could be cheaper than Brute Force.

# Polynomial chaos expansion, PCE methods (Najm et al., 2009)

- Here an uncertainty factor  $u_i$  is first assigned to each input variable.
  - Note that this uncertainty parameter  $u_i$  is related to uncertainty parameter  $f$  by  $u_i = 10^f$ .
- Taking the example of rate coefficients, they are then normalised into factorial variables  $\mathbf{x}$  as follows:

$$x_i = \frac{\ln k_i / k_{i,0}}{\ln u_i}$$


The diagram illustrates the normalization process. A blue rectangular box containing the text "Nominal value" has a blue arrow pointing from the box to the numerator of the fraction in the equation  $x_i = \frac{\ln k_i / k_{i,0}}{\ln u_i}$ . This indicates that  $k_{i,0}$  represents the nominal value of the rate coefficient.

- Hence  $x_i = 0$  gives the nominal value of the rate coefficient, and -1 and +1 represent the upper and lower bounds.
- A response surface of the predicted combustion properties is then generated with respect to  $\mathbf{x}$ .

- Often restricted to a 2nd order polynomial expansion which for the  $r$ 'th model response  $\eta_r(\mathbf{x})$  can be written as:

$$\eta_r(\mathbf{x}) = \eta_{r,0} + \sum_{i=1}^m a_{r,i} x_i + \sum_{i=1}^m \sum_{j \geq i}^m b_{r,i,j} x_i x_j$$

- The uncertainty in  $\mathbf{x}$  may be expressed as a polynomial expansion of basis random variables  $\boldsymbol{\xi}$ :

$$\mathbf{x} = \mathbf{x}^{(0)} + \sum_{i=1}^m \boldsymbol{\alpha}_i \xi_i + \sum_{i=1}^m \sum_{j \geq i}^m \boldsymbol{\beta}_{ij} \xi_i \xi_j + \dots,$$

where  $\boldsymbol{\alpha}$  and  $\boldsymbol{\beta}$  are column vectors of expansion coefficients,  $m$  is the number of rate coefficients under consideration and  $\mathbf{x}^{(0)}$  is a column vector of normalised rate coefficients which is a zero vector for the nominal reaction model.

- If the  $\mathbf{x}$ 's are independent of each other and normally distributed, then the usual choice for the form of  $\xi$  would be a set of unit-normal random variables.
- If  $\ln u_i$  represents 2 times the standard deviation of  $\ln k_i$  then  $\alpha$  is  $\frac{1}{2} \mathbf{I}_m$ , where  $\mathbf{I}_m$  is the  $m$ -dimensional identity matrix.  $\beta$  and all higher order terms are zero (Sheen et al. 2009).
- In the general case, combining the above two equations and truncating the higher order terms gives:

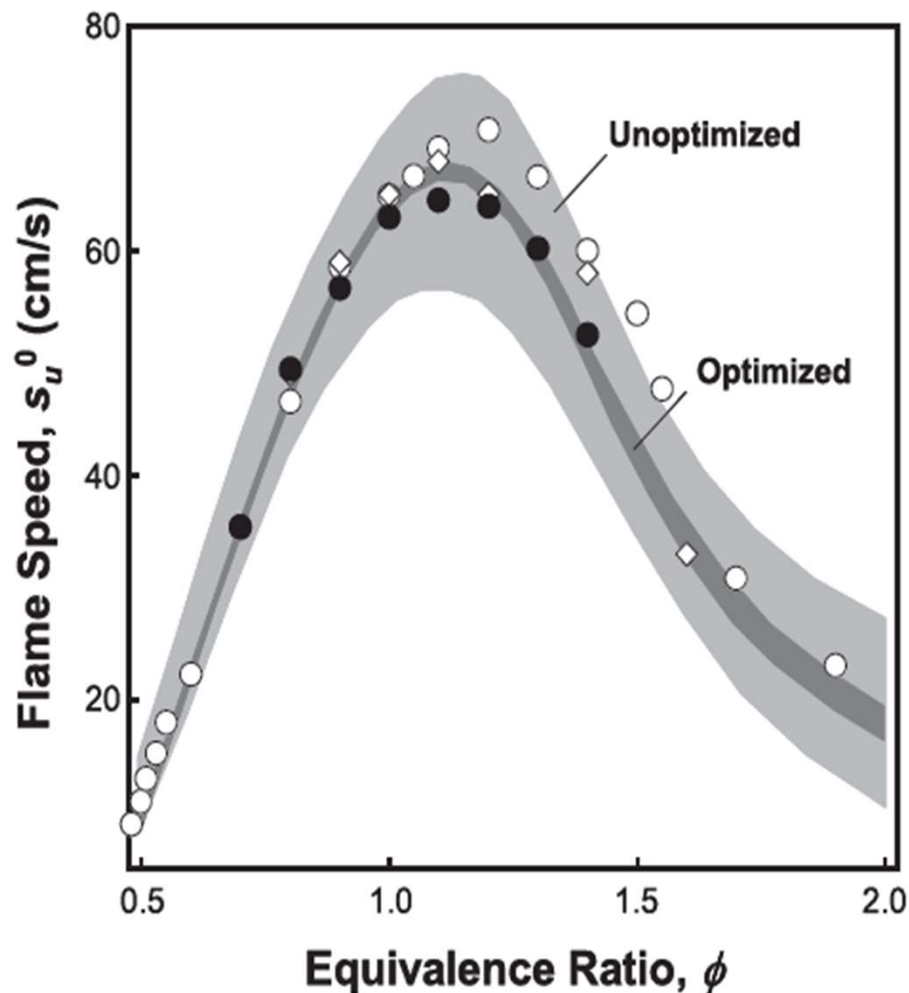
$$\eta_r(\xi) = \eta_r(\mathbf{x}^{(0)}) + \sum_{i=1}^m \hat{\alpha}_{r,i} \xi_i + \sum_{i=1}^m \sum_{j \geq i}^m \hat{\beta}_{r,ij} \xi_i \xi_j + \dots,$$

$$\hat{\alpha}_r = \frac{1}{2} \mathbf{I}_m \mathbf{a}_r \quad \hat{\beta}_r = \frac{1}{4} \mathbf{I}_m^T \mathbf{b}_r \mathbf{I}_m$$

- What this equation shows is that the overall model prediction is given by its **nominal value plus uncertainty contributions from each rate coefficient**.

$$\sigma_r(\xi)^2 = \sum_{i=1}^m \hat{\alpha}_{r,i}^2 + 2 \sum_{i=1}^m \sum_{j>i}^m \hat{\beta}_{r,ij}^2 + \sum_{i=1}^m \sum_{j>i}^m \hat{\beta}_{r,ij}^2$$

# Examples of application



- Experimental data and computed  $2\sigma$  uncertainty bands for the laminar flame speed of ethylene-air mixtures at  $p = 1$  atm. (Sheen et al., 2009).
- Note that following the application of an optimization procedure, the uncertainty bounds are much narrower.
- The polynomial chaos expansion is used within the optimisation procedure.

**ANOVA (ANALYSIS OF  
VARIANCES)  
DECOMPOSITION AND  
HDMR METHODS**

# Variance decomposition

- For independent inputs (i.e. no correlations exist between inputs), a unique decomposition of the unconditional variance  $V(Y)$  can be obtained (Li et al., 2010):

$$V(Y) = \sum_{i=1}^n V_i + \sum_{1 \leq i < j \leq n} V_{ij} + \dots + V_{12\dots n} = \sum_{j=1}^{2^n-1} V_{x_j}$$

$$\sum_{j=1}^{2^n-1} \frac{V_{x_j}}{V(Y)} = \sum_{j=1}^{2^n-1} S_{x_j} = 1$$

- The approach is therefore analogous to the classical approaches described above but instead of directly calculating the conditional variances using e.g. Monte Carlo samples, now a meta-model is developed first and the sensitivity indices are calculated using the meta-model.

# RSM approaches to ANOVA decomposition

- Polynomial chaos expansions were one method to achieve this ANOVA decomposition.
- Other methods are based on **High Dimensional Model Representations** (HDMR).
- HDMR originally developed to provide a straightforward approach to explore input-output mapping of models without requiring large numbers of runs (Sobol', 1990; Rabitz et al., 1999; Li et al., 2001).
- The use of truncated expansions is possible because usually only low-order correlations between inputs have a significant effect on the outputs.
- Because of the **hierarchical form** of HDMR component functions, sensitivity indices can be determined from them in an automatic way in order to rank the importance of input parameters and to explore the influence of parameter interactions.



# Basic mapping

- The mapping between the inputs  $x_1, \dots, x_n$  and the output variable  $Y(\mathbf{x}) = f(x_1, \dots, x_n)$  can be written in the following hierarchical form:

$$Y(\mathbf{x}) = f_0 + \sum_{i=1}^n f_i(x_i) + \sum_{1 \leq i < j \leq n} f_{ij}(x_i, x_j) + \dots + f_{12\dots n}(x_1, x_2, \dots, x_n)$$

- Here the zeroth-order component  $f_0$  denotes the mean effect, which is the expected value of the model output  $f_0 = E(Y)$ .
- The first-order component functions  $f_i(x_i)$  give the effect of variable  $x_i$  acting independently (although generally nonlinearly) upon the output  $Y(\mathbf{x})$ :

$$f_i(x_i) = E(Y|x_i) - f_0$$

- The function  $f_{ij}(x_i, x_j)$  is a second-order term describing the cooperative effects of the variables  $x_i$  and  $x_j$  upon the output  $Y(\mathbf{x})$ :

$$f_{ij}(x_i, x_j) = E(Y|x_i, x_j) - f_i - f_j - f_0$$

If we find an **accurate meta-model** with which to represent the HDMR expansion, we can provide an accurate estimate of the partial variances and therefore the global sensitivity indices.

# Calculation of sensitivity indices

- The HDMR expansion is analogous to the variance decomposition:

$$V(Y) = \sum_{i=1}^n V_i + \sum_{1 \leq i < j \leq n} V_{ij} + \dots + V_{12\dots n} = \sum_{j=1}^{2^n-1} V_{p_j}$$

And the coefficients of the expansion can be used to calculate the sensitivity indices.

$$V_{p_j} = V(f_{p_j}(\mathbf{x}_{p_j}))$$

$$\sum_{j=1}^{2^n-1} \frac{V_{p_j}}{V(Y)} = \sum_{j=1}^{2^n-1} S_{p_j} = 1$$

# QRS-HDMR

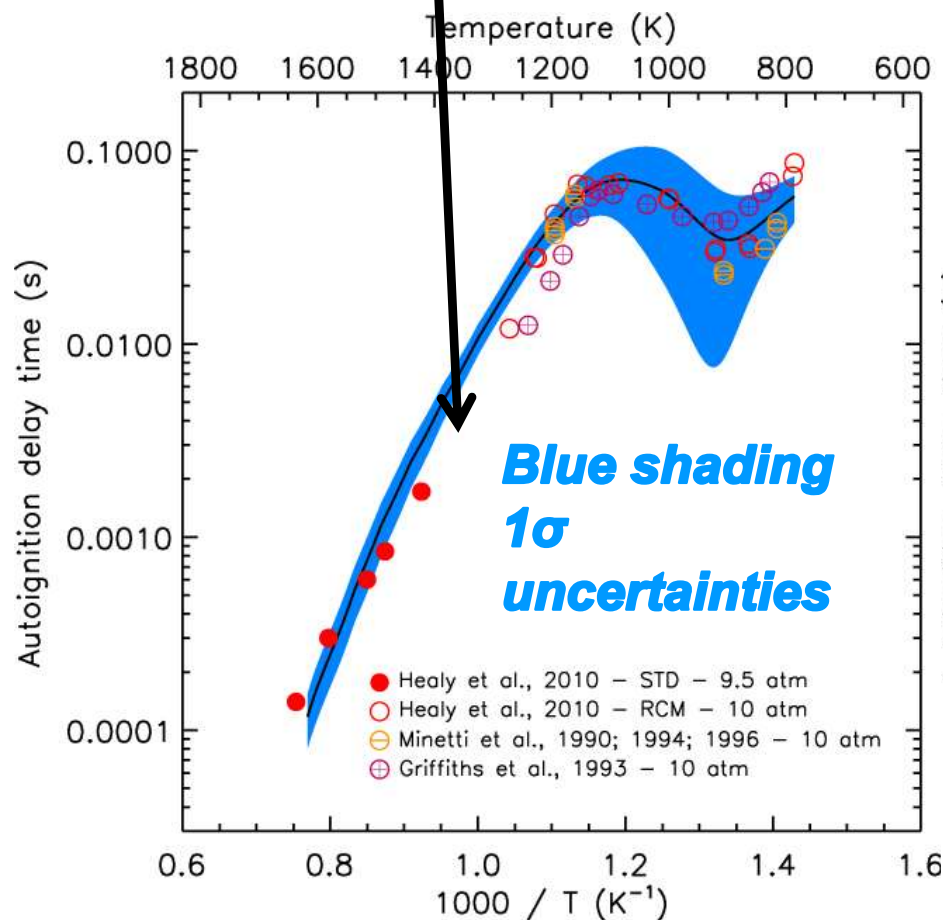
- **Quasi-random sequences** such as a **Sobol sequence** have **better convergence** properties than other sampling approaches.
- Therefore we expect the Sobol' sequence to be a good choice of sampling strategy for fitting an HDMR meta model.
  1. A quasi-random sample is developed for chosen input parameter space.
  2. The full model would be run for each sample (e.g. 1024, 2048, etc) and target outputs stored.
  3. A meta-model would be fitted to the input-output relationships for each target output. **Orthonormal polynomials** are generally used.
  4. The fitted HDMR meta-model would be used to derive global sensitivity indices.
- The accuracy of the meta-model determines the accuracy of the calculated indices and needs to be checked carefully.

# Sample size

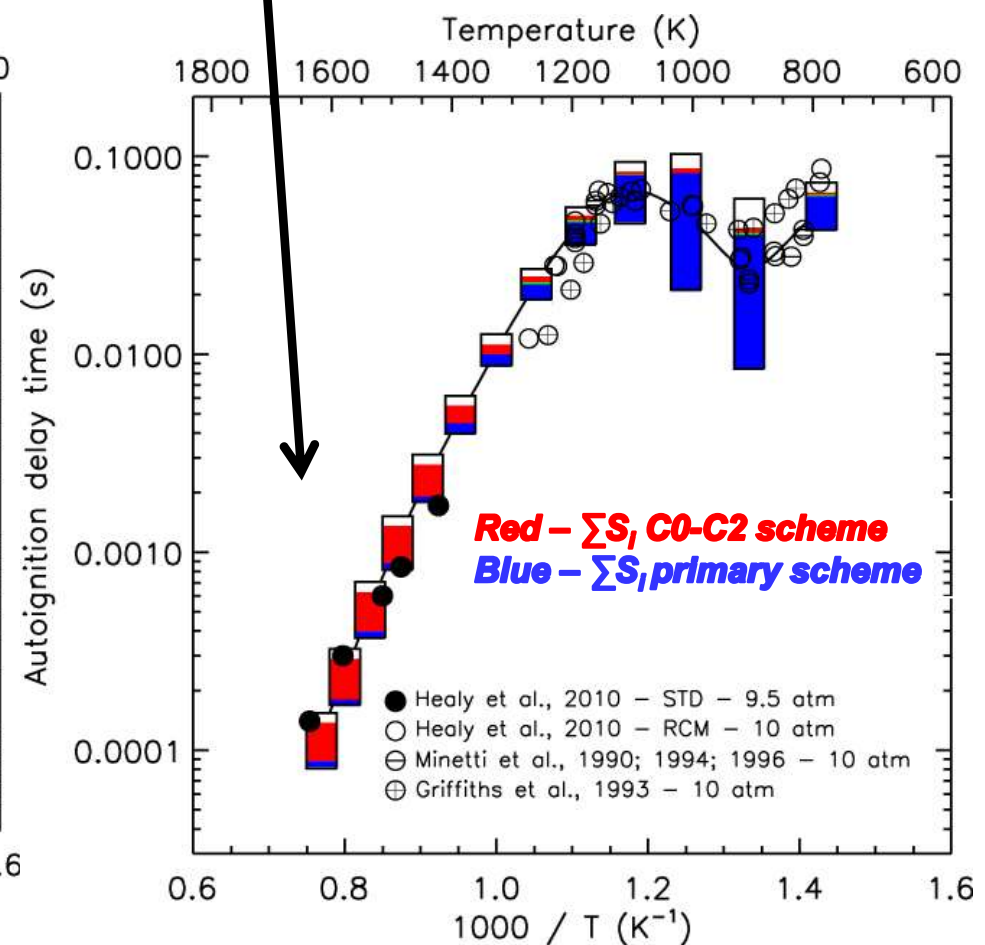
- The coefficients are determined using Monte Carlo integration over the chosen input sample (Li et al., 2002).
- The approximation of the component functions reduces the sampling effort dramatically so that only one set of quasi-random samples  $N$  is necessary in order to determine all RS-HDMR component functions and subsequently the sensitivity indices.
- For first-order indices this sample can usually be quite small (e.g. 1024).
- If significant **second-order effects** are present then the **sample size** will need to be **bigger**.
- **Remember – base 2 system so sample size increases as  $2^{Ns}$** 
  - **2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048 etc!**

# Example: Ignition delays of butane: sources of uncertainties

*Uncertainties lowest at higher temperatures.*



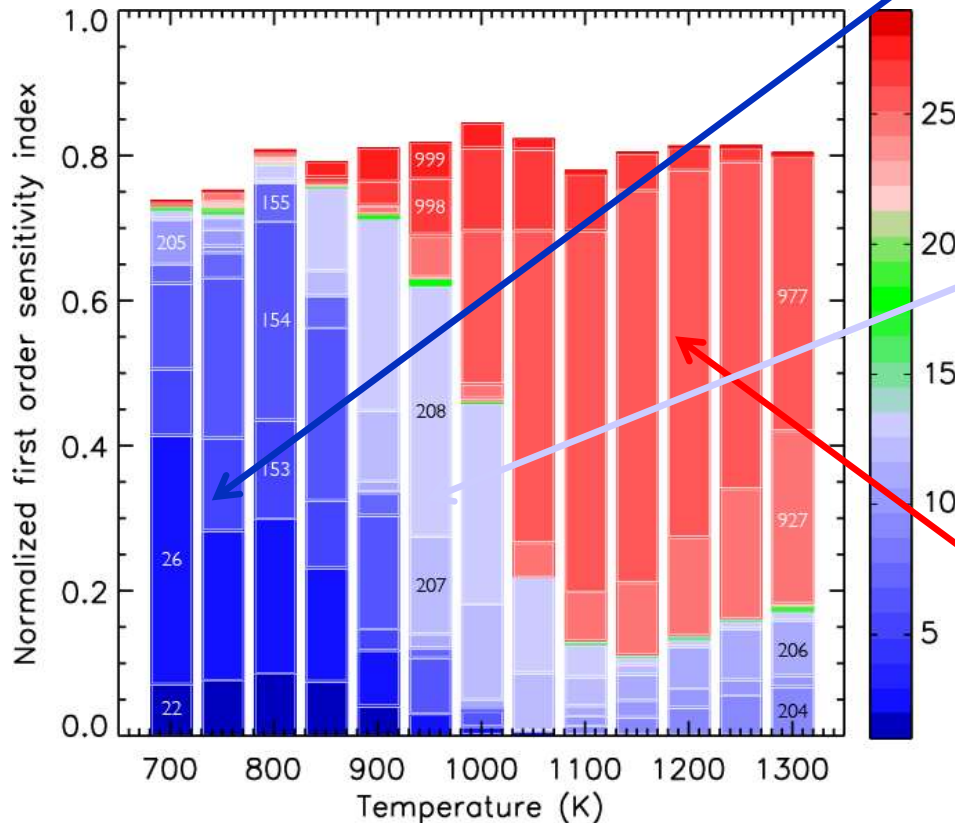
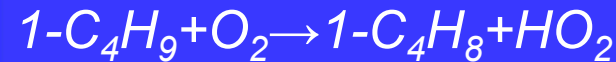
*Dominated by sensitivity to reactions in C0-C2 base scheme*



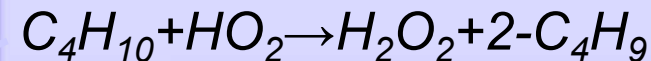
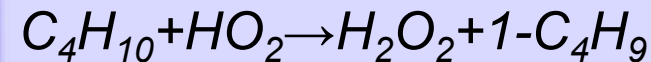
(Hébrard et al., 2015)

# Ignition delays: 1<sup>st</sup>-order global sensitivities

## Key reactions at low temperatures



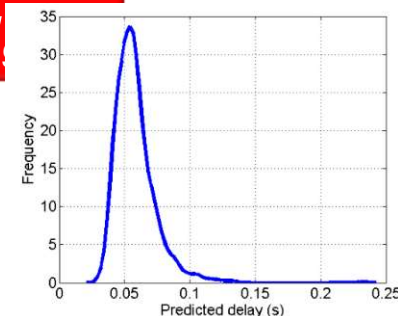
## Key reactions at intermediate temperatures



## Key reactions at high temperatures



higher order effects lead to  
the tail

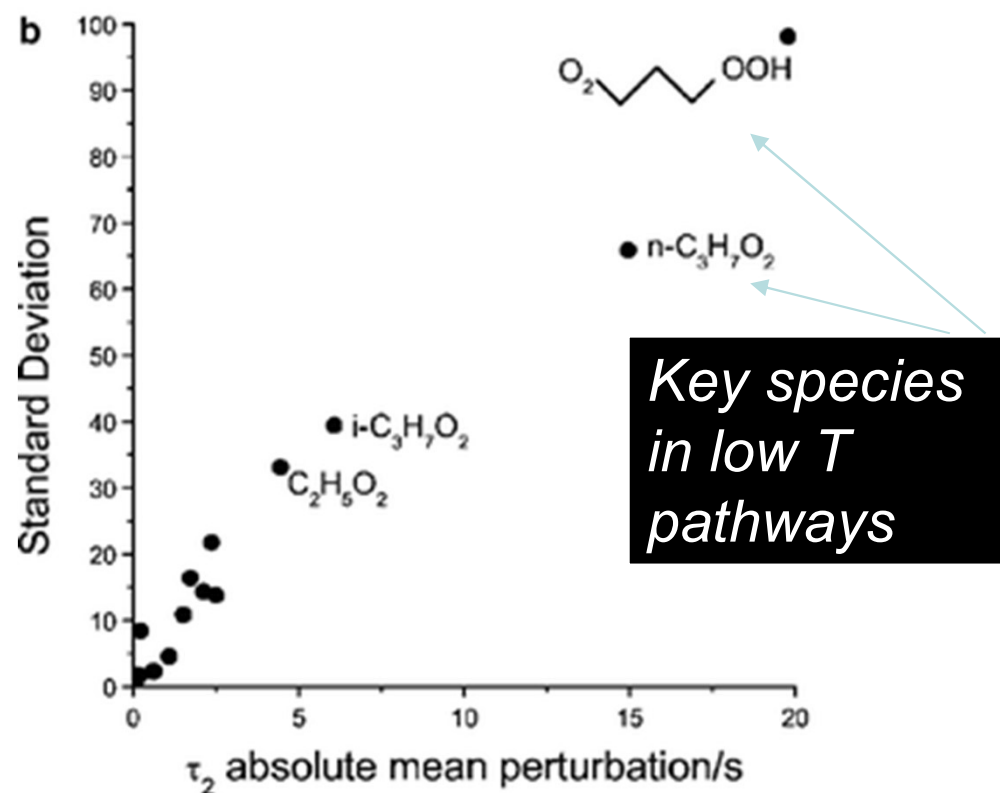
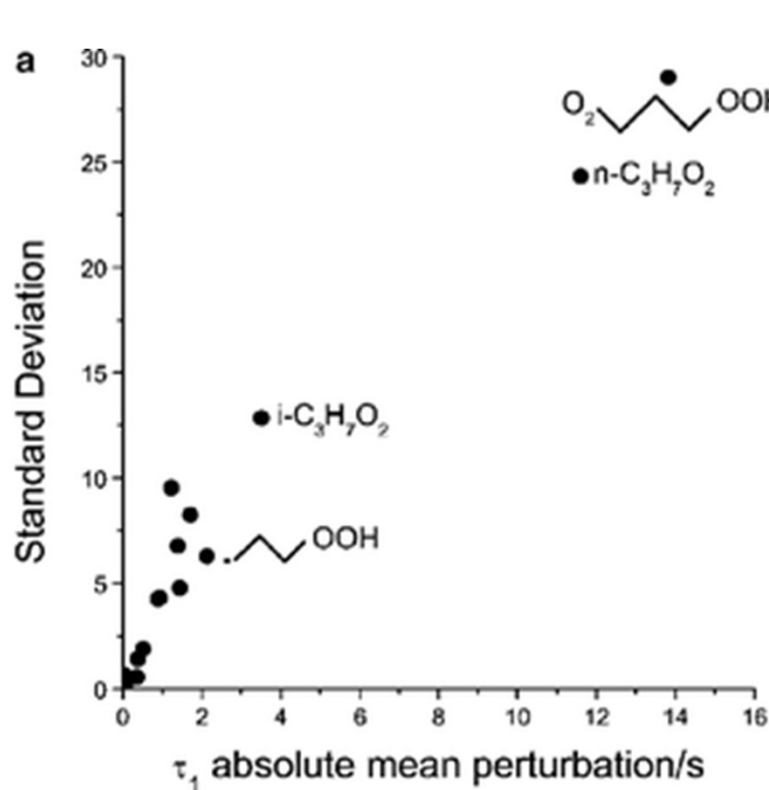


# Importance of uncertainties in thermodynamic data

- Many sensitivity studies focus only on a local analysis of A factors for reactions.
  - Tells us about importance of different reaction steps but not really sufficient for full uncertainty propagation.
- Effects of thermo data often ignored but can be critical for predicting e.g.
  - Heat release
  - Equilibrium between important species in low T reactions such as  $\text{RO}_2$ ,  $\text{QOOH}$
- At simplest level **should involve variability in heats of formation.**
- In reality where data from ATChT – data is highly correlated.

# Some examples

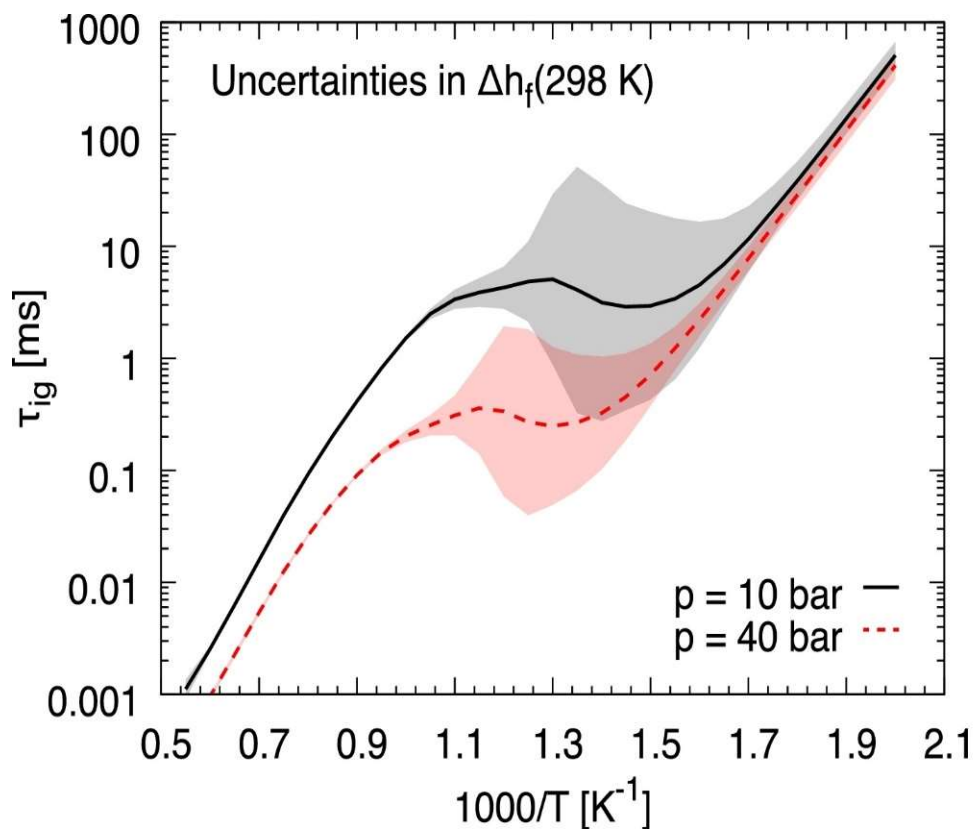
- 2 stage ignition delay times for **stoichiometric propane oxidation**.
- Morris method (Hughes, 2006).
- Variations in  $\Delta H_{f^\circ}$  from NIST Chemistry WebBook where possible.
- Where no quoted error,  $\pm 5$ ,  $+10$  or  $+15$  kJ mol<sup>-1</sup>, used depending on complexity of species.



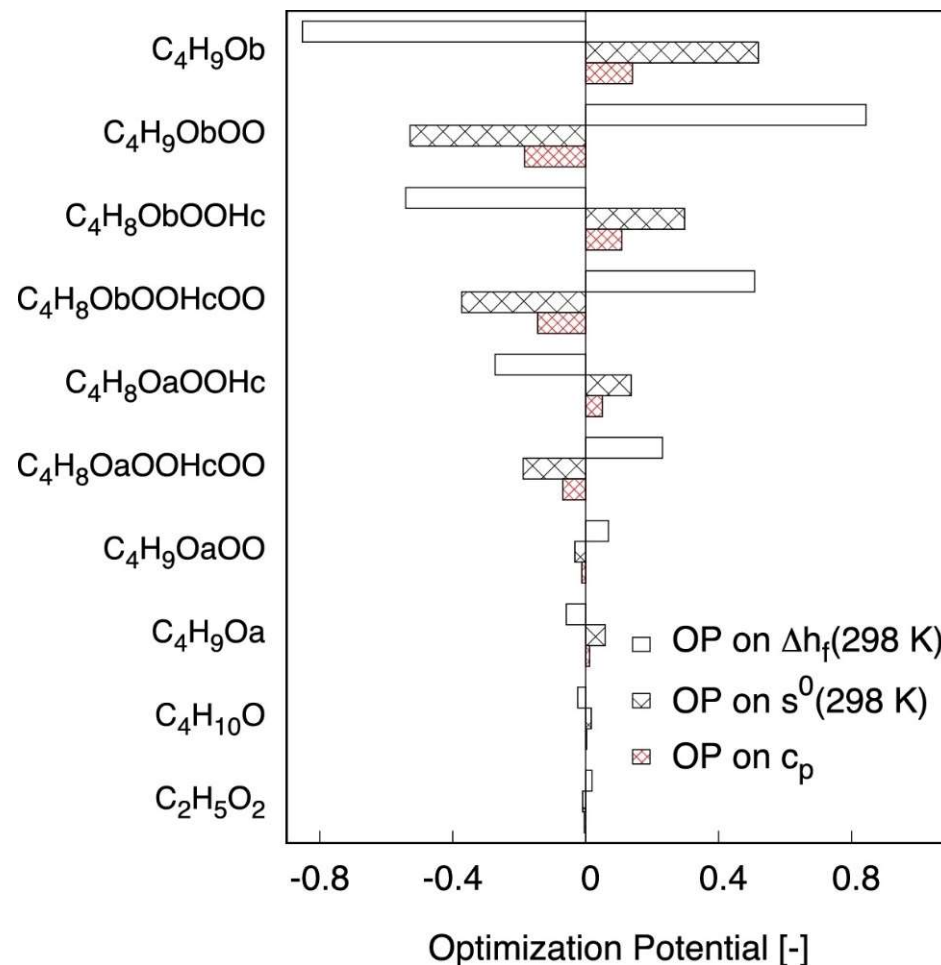


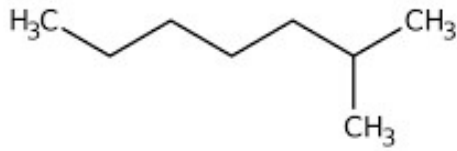
# Impact of thermo uncertainties on diethyl ether oxidation (Vom Lehn, 2019)

- Highest prediction uncertainty observed in the NTC regime.



Optimization potential of  $\tau_{ign}$  for uncertainties in  $\Delta h_f(298 \text{ K})$ ,  $s^0(298 \text{ K})$ , and  $c_p$  for a stoichiometric DEE/air mixture at  $T=769 \text{ K}$  and 10 bar.

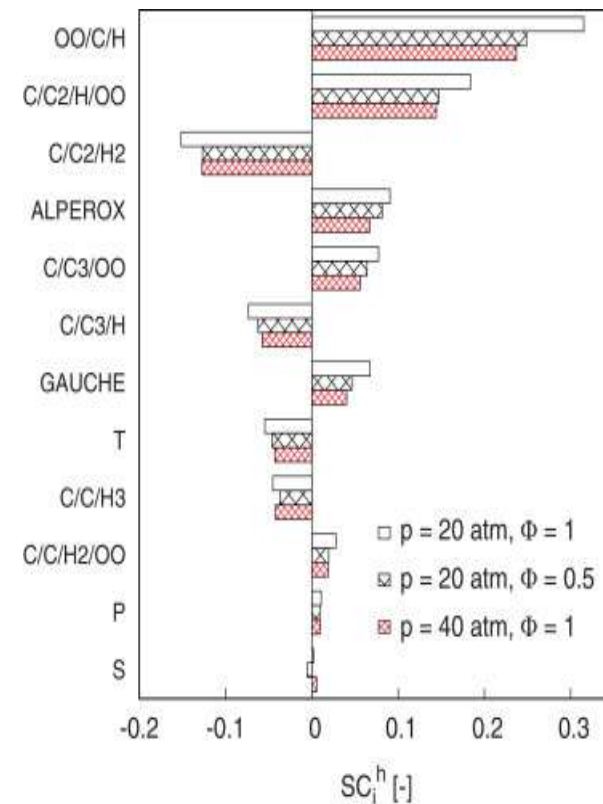
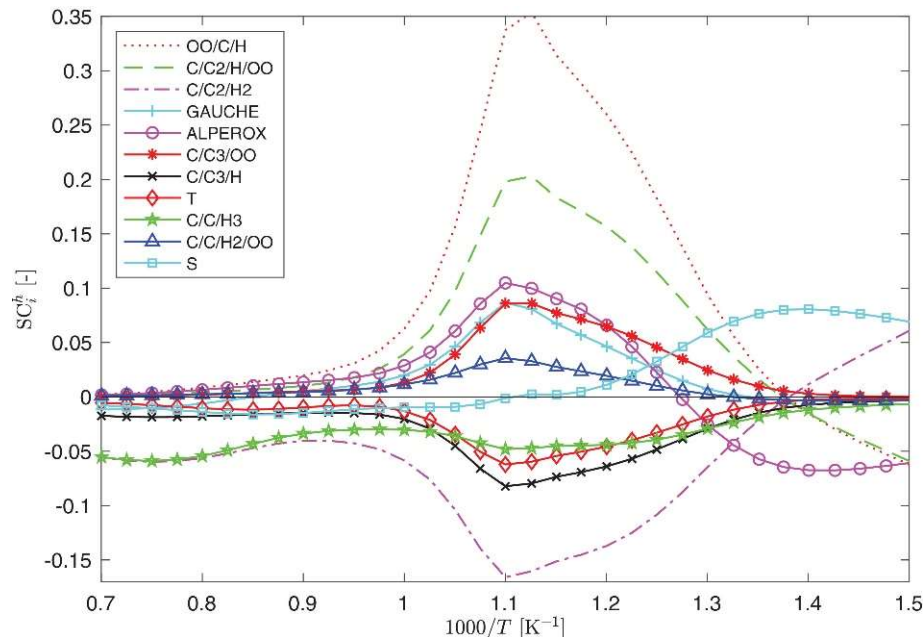
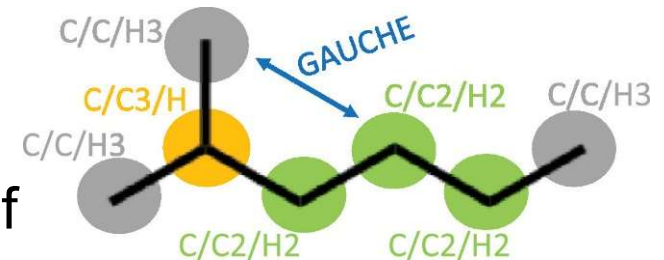




# Impacts of group additivity values on kinetic model predictions (vom Lehn, 2020)

- 2 methyl heptane, shock tube ignition delays.
- Non-dimensionalized sensitivity coefficients of ignition delay time with respect to group parameters  $\psi_j$  defined as:

$$SC_j^\psi = \frac{\partial \tau_{ign}}{\partial \psi_j} \frac{\psi_{norm}}{\tau_0}$$



Sensitivity coefficients of ignition delay time on the group values of  $\Delta h_f(298 \text{ K})$

# Information content and design of experiments (DoE)

- High correlations perhaps suggests a different approach for determining new experiments of value.
- Instead of designing experiments to isolate and improve accuracy of individual reactions, need to think about **minimising uncertainty of system as a whole through optimisation.**
- Methods from information theory useful.
  - Particularly for alternative fuels where little data exists and species are large (limiting application of high level theory).
- Optimal experiments are chosen iteratively one by one.
  - Giving high priority experiments and their order.

**We are not used to simulating experiments before we perform them but we should do it!**

# Example for DME combustion (Vom Lehn, 2021)

- Assuming a multivariate normal distribution for optimized parameters in model, posterior covariance matrix  $\Sigma^*$  of  $\mathbf{x}$  is estimated based on linearization of response surface in neighbourhood of posterior values  $\mathbf{x}^*$  where  $\mathbf{J}_r^*$  is local gradient of model response  $r$  with respect to  $\mathbf{x}$ , evaluated at  $\mathbf{x}^*$ .

$$\Sigma^* = \left[ \sum_{r=1}^n \frac{\mathbf{J}_r^* \mathbf{J}_r^{*T}}{(\sigma_r^{\text{obs}})^2} + 4\mathbf{I} \right]^{-1}.$$

contains the important information about the joint uncertainty space of the optimized parameters

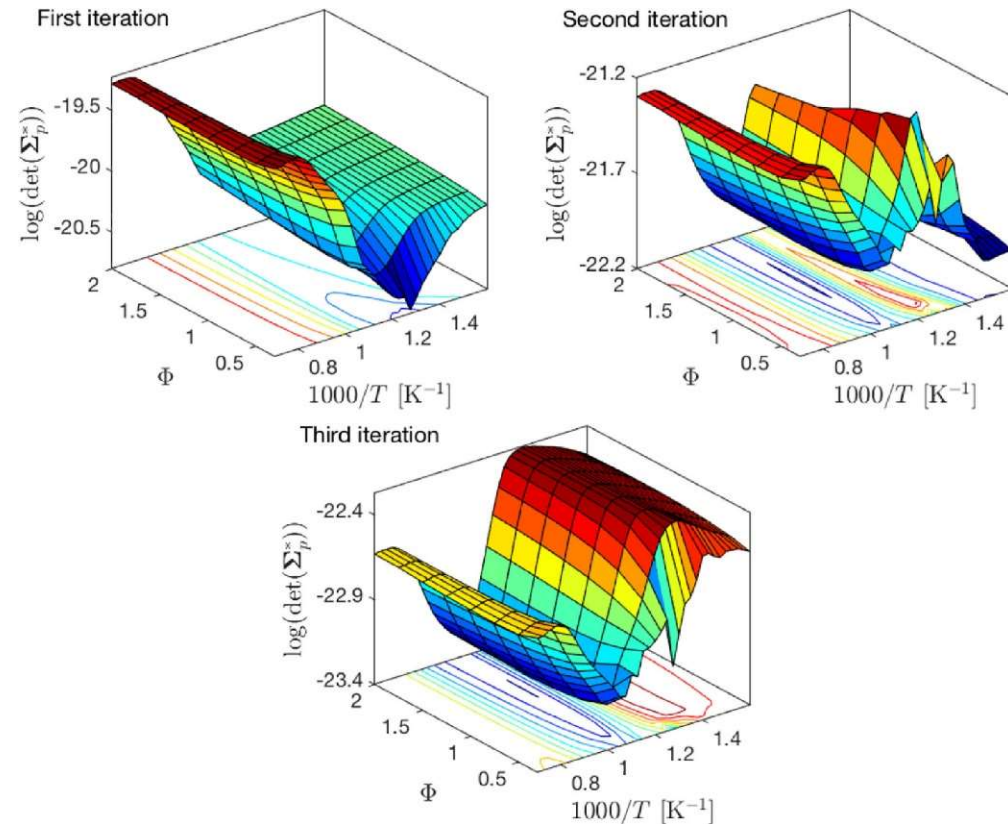
***Aim: efficient minimization of joint parameter uncertainties.***

- Each iteration starts with evaluation of all not yet selected conditions  $p$  in terms of covariance matrices that would result if the experiment  $p$  were selected.
- Nominal model prediction from previous iteration assumed as hypothetical experimental value of experiment  $p$  to determine  $\Sigma_p^*$ . Sum includes all previous exp.

# Example for DME combustion (Vom Lehn, 2021)

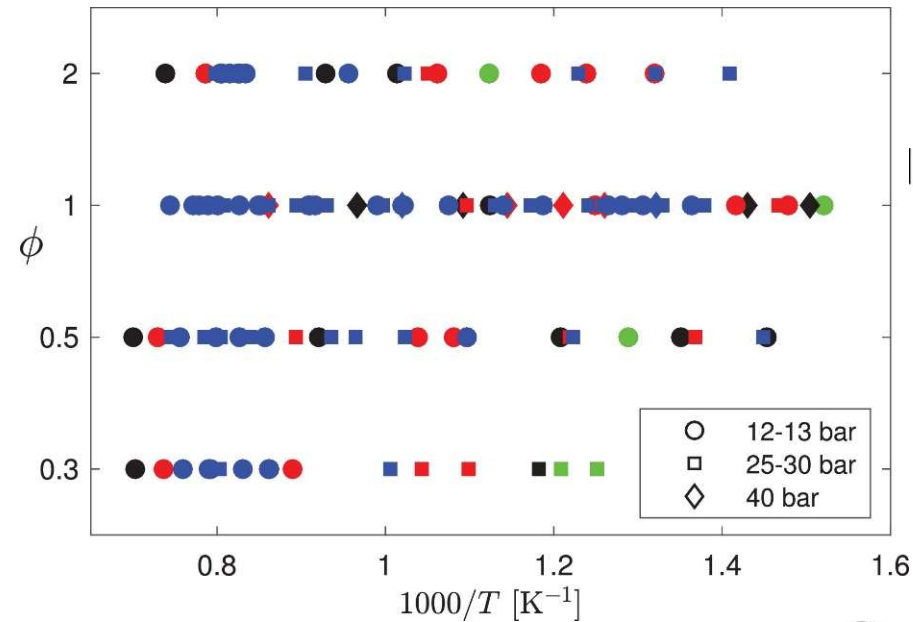
- Equivalent to D-optimal design.
  - select the experiment which minimizes the Shannon entropy of the multivariate normal distribution of model parameters in each step.
- Shannon entropy in this context measures the variability of the multivariate normal distribution of the model parameters.
- Minimization is equivalent to minimizing the determinant of the expected covariance matrix after inclusion of the new experiment:

$$\Psi(p) = \min_p (\det(\Sigma_p^*))$$

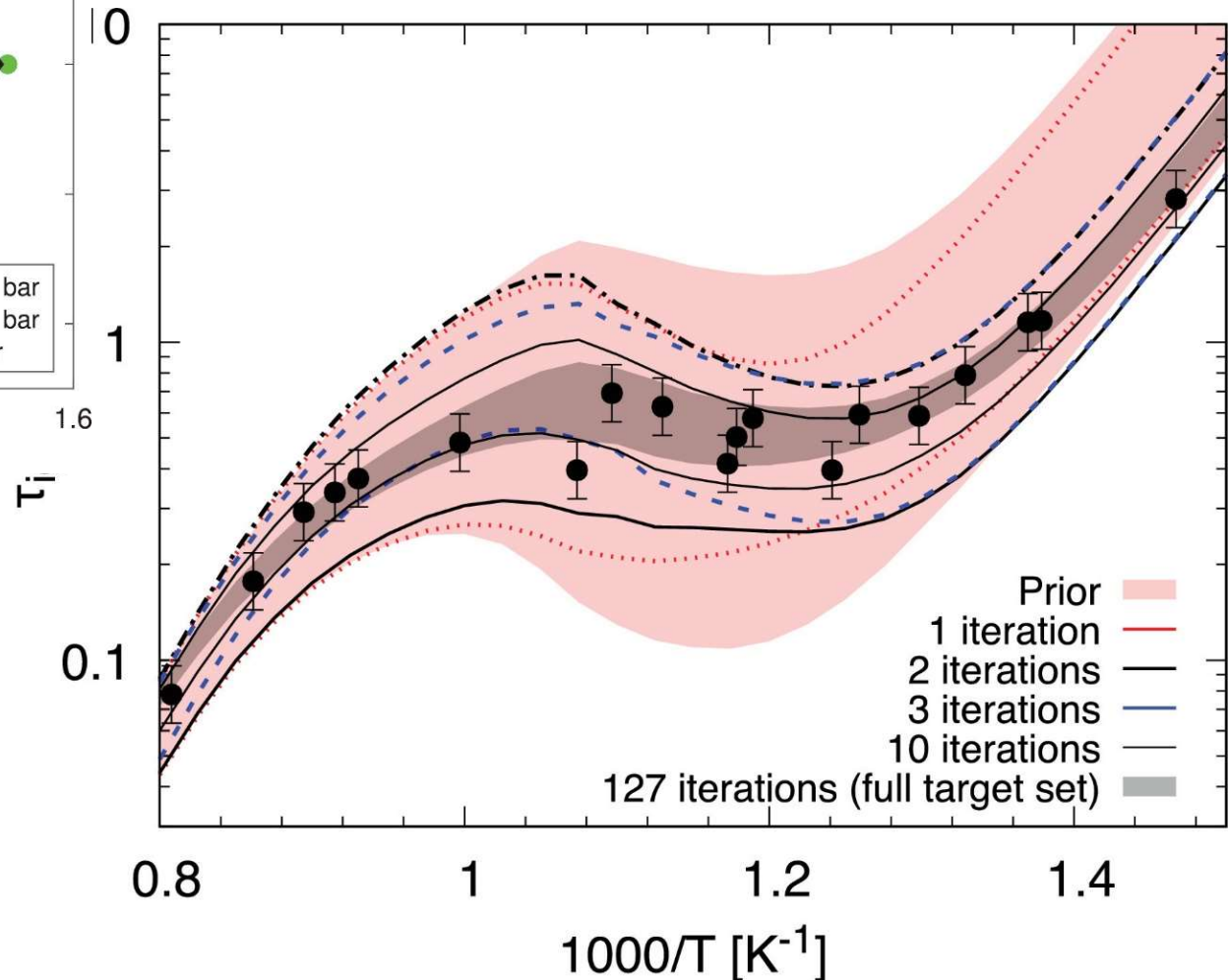
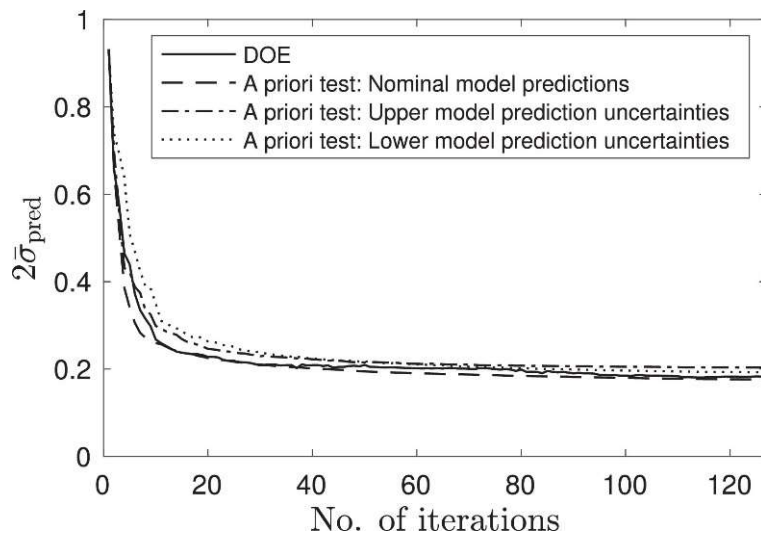




# Example for DME combustion (Vom Lehn, 2021)



● 1-5, ● 6-20  
● 21-50, ● 50-127



$2\sigma$  prediction uncertainties after different DOE iteration steps

# Doing things optimally

- Requires collaboration as a community!
- To combine modelling, experimental design and expertise, statistical methods.
- To think about what experiments/theory calcs are required to reduce system uncertainty and not just what we fancy/might be easiest...

