

# Two or three things I am sure about uncertainty

Pascal PERNOT

Institut de Chimie Physique,  
CNRS & Univ. Paris-Saclay



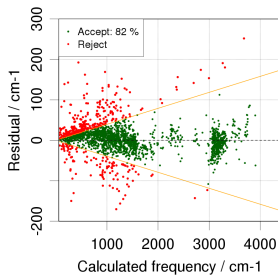
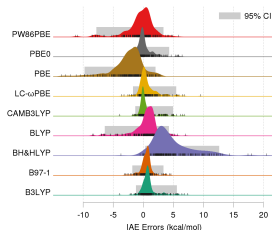
TUMLA, Rennes (2022-07-06)

- 1 Uncertainty representation
- 2 Uncertainty validation
- 3 Statistical treatment of Model Inadequacy
- 4 Conclusions

- **Permanent CNRS researcher**  
([orcid:0000-0001-8586-6222](https://orcid.org/0000-0001-8586-6222))  
(<https://tinyurl.com/52zbchaz>)

- **Main research topics**

- Statistics for benchmarking of comput. chemistry methods
- *Uncertainty management in complex chemical models*
- *Uncertainty quantification in computational chemistry*



# Uncertainty vs. probability

- **Representation of uncertainty**

- one can attribute a "degree of confidence"  $p(X = x)$  to any possible value  $x$  of an uncertain quantity  $X^1$
- consistency and normalization  $\Rightarrow p(X = x)$  follows the laws of probability<sup>2</sup>

- **This approach at the core of the GUM (3.3.1)<sup>3</sup>**

- *"measurement uncertainty reflects the lack of knowledge about the true value of a measurand"*
- *"the corresponding state-of-knowledge is best described by a distribution on the possible values of the measurand"*

- The **Maximum Entropy Principle (Maxent)** states that, among all distributions compatible with a state-of-knowledge, one should select the one which minimizes Shannon entropy  $H(p)$ .

---

<sup>1</sup>P-S. Laplace (1825) *Essai philosophique sur les probabilit s*, 5th ed.

<sup>2</sup>Dupr , Maurice J., Tipler, Frank J. (2009) *Bayesian Analysis* 3:599–606

<sup>3</sup>**JCGM 100:2008** Evaluation of measurement data – Guide to the expression of uncertainty in measurement. (*ISO/IEC Guide 98-3*)

# Uncertainty vs error<sup>4</sup>

*In order to estimate a measurement uncertainty it is assumed that the result of a measurement has been corrected for all recognized significant systematic effects and that every effort has been made to identify such effects.*

## systematic error

component of measurement error that remains constant or varies in a **predictable** manner

## random error

component of measurement error that varies in an **unpredictable** manner

## uncertainty

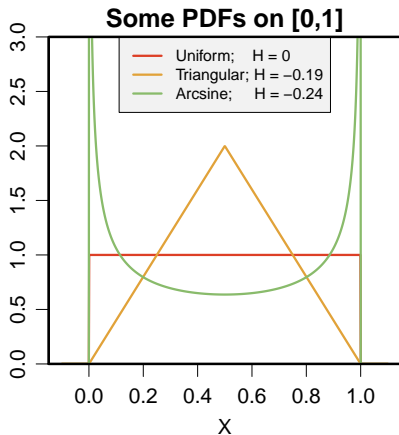
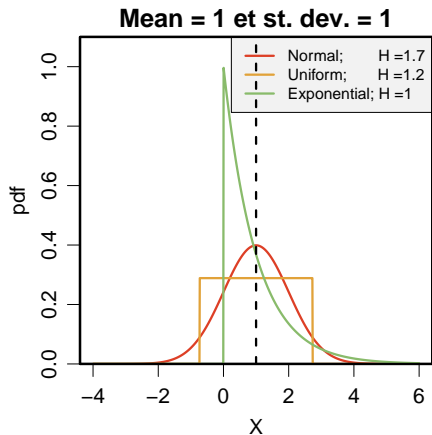
**non-negative parameter** characterizing the **dispersion** of the quantity values being attributed to a measurand

- often estimated by a standard deviation  $u$ , or the half-width of a probability interval  $U_p$

---

<sup>4</sup> *Guide to the expression of uncertainty in measurement (GUM), JCGM 100:2008, International Vocabulary of Metrology (VIM), JCGM 200:2012*

# Examples: entropy vs. pdf



# Multivariate pdfs

To describe several **statistically dependent** variables, one often uses their mean values and a variance-covariance matrix. The pdf is often a multivariate normal.

$$p(x, y | \mu_{X_1}, \mu_{X_2}, \Sigma) \propto (\det \Sigma)^{-1/2} \exp\left(-\frac{1}{2} \mathbf{E}^T \Sigma^{-1} \mathbf{E}\right)$$

$$\mathbf{E} = \begin{pmatrix} x_1 - \mu_{X_1} \\ x_2 - \mu_{X_2} \end{pmatrix}; \Sigma = \begin{pmatrix} u_{X_1}^2 & u_{X_1 X_2}^2 \\ u_{X_2 X_1}^2 & u_{X_2}^2 \end{pmatrix} = \mathbf{S}^T \mathbf{R} \mathbf{S}$$

with

$$\mathbf{S} = \begin{pmatrix} u_{X_1} \\ u_{X_2} \end{pmatrix}; \mathbf{R} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

- $\Sigma$  is the variance-covariance matrix
- $\mathbf{R}$  is the correlation matrix ( $-1 \leq \rho \leq 1$ )

# Sum-constraints

## The case of conservation laws

- $N$  variables such as  $\sum_{i=1}^N X_i = 1$  et  $\{X_i \geq 0; i = 1, N\}$ , one uses *Dirichlet's law* (Maxent)

$$\{X_1, \dots, X_N\} \sim \text{Diri}(\gamma * (\mu_1, \dots, \mu_N))$$

where  $\gamma > 0$  is a precision factor  
(variance increases as  $\gamma$  decreases)

- **examples:** chemical compositions, branching ratios<sup>5</sup>...

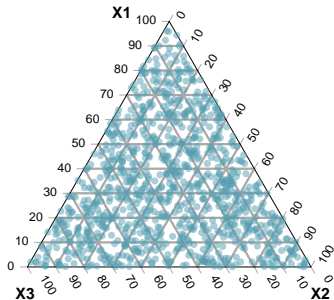
---

<sup>5</sup>Plessis *et al.* (2010) *J. Chem. Phys.* **133**:134110

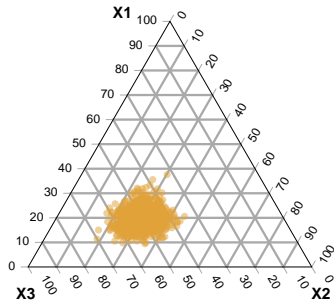


# Dirichlet pdf

Diri( 3 \* (1/3,1/3,1/3) )



Diri( 100 \* (0.2,0.3,0.5) )



# Quantifying uncertainty on rate constants

- chemical equations based on empirical parameters



$$\frac{da(t)}{dt} = -k_{AB}(T, P, M) a(t) b(t)$$

- reaction rates from experiments
  - with [ [very] large ] uncertainty
  - numerous estimated parameters
  - in Titan's chemistry models, less than 10% of rate constants have been measured in representative (T,P,M) conditions. . .

→ **modelers need extrapolation (with uncertainty !)**

# Arrhenius rate analysis (for extrapolation)

- Data

$$\{T_i, k_i, u_{k_i}\}_{i=1}^N$$

- Linear regression

$$\ln k_i = \ln k_0 - E/T_i$$

$$y = a + bx$$

with  $x = -1/T$  and uncertainties  $u_{y_i} = u_{k_i}/k_i$  corresponding to the log-transformation of the rate constants

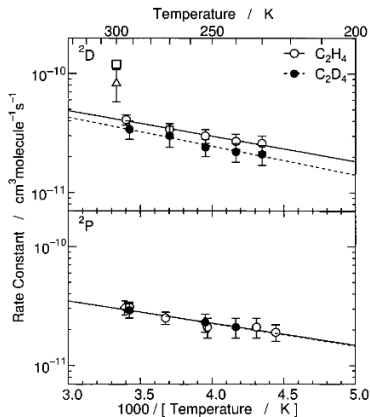
- Uncertainty propagation

$$u_{\ln k}(T) = \left( u_a^2 - \frac{2}{T} u_a u_b \rho_{ab} + \frac{1}{T^2} u_b^2 \right)^{1/2}$$

$$F_k(T) = \exp(u_{\ln k}(T))$$

Example:  $N(^2D) + C_2H_4$ 

Sato et al. (1999) *Phys. Chem. A* **103**:8650–8656

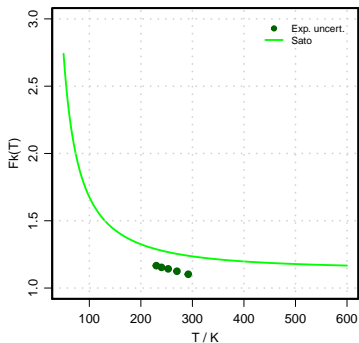


reaction	$A/\text{cm}^3 \text{ molecule}^{-1} \text{ s}^{-1}$	$E_a/\text{kcal mol}^{-1}$	ref
reactions of $N(^2D)$			
$N(^2D) + C_2H_4$	$(2.3 \pm 0.3) \times 10^{-10}$	$1.0 \pm 0.1$	this work
$N(^2D) + C_2D_4$	$(2.4 \pm 0.5) \times 10^{-10}$	$1.1 \pm 0.1$	this work
$N(^2D) + CH_4$	$(7.1 \pm 4.4) \times 10^{-11}$	$1.5 \pm 0.3$	32
$N(^2D) + CD_4$	$(3.3 \pm 1.4) \times 10^{-11}$	$1.4 \pm 0.1$	32
$N(^2D) + C_2H_2$	$(1.6 \pm 0.2) \times 10^{-10}$	$0.53 \pm 0.06$	33
$N(^2D) + C_2D_2$	$(1.4 \pm 0.1) \times 10^{-10}$	$0.48 \pm 0.05$	33
reactions of $N(^2P)$			
$N(^2P) + C_2H_4$	$(1.4 \pm 0.5) \times 10^{-10}$	$0.91 \pm 0.18$	this work
$N(^2P) + C_2D_4$	$(1.3 \pm 0.2) \times 10^{-10}$	$0.87 \pm 0.10$	this work
$N(^2P) + CH_4$	$(5.0 \pm 2.2) \times 10^{-13}$	$0.98 \pm 0.22$	32
$N(^2P) + CD_4$	$(3.1 \pm 1.4) \times 10^{-13}$	$0.96 \pm 0.24$	32
$N(^2P) + C_2H_2$	$(1.0 \pm 0.1) \times 10^{-10}$	$0.88 \pm 0.06$	33
$N(^2P) + C_2D_2$	$(0.71 \pm 0.12) \times 10^{-10}$	$0.76 \pm 0.09$	33

<sup>a</sup> Error represents one standard deviation.

# Extrapolation uncertainty vs. correlations<sup>6</sup>

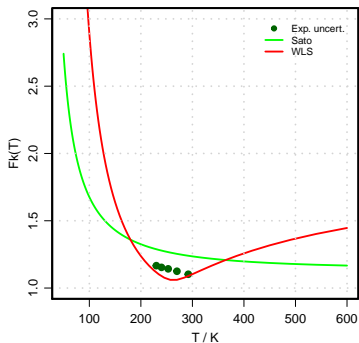
	$\ln A$	$E_a/R$ (K)	$\rho$
Sato <i>et al.</i> (1999)	$-22.193 \pm 0.13$	$503 \pm 50$	N/A



<sup>6</sup>E. H brard *et al.*, *J. Phys. Chem. A* 113:11227-11237 (2009)

# Extrapolation uncertainty vs. correlations<sup>7</sup>

	$\ln A$	$E_a/R$ (K)	$\rho$
Hebrard <i>et al.</i> (2009)	$-22.193 \pm 0.66$	$503 \pm 170$	0.996



<sup>7</sup>E. H brard *et al.*, *J. Phys. Chem. A* 113:11227-11237 (2009)

# Extrapolation uncertainty vs. correlations

- I realized later that the  $\chi^2 \propto \langle R^2/u^2 \rangle$  value for our best fit was **too small**, meaning that the  $u_i$ s were too large to explain the fit residuals
  - it is plausible that *systematic errors* were aggregated into Sato's uncertainties
- Considering the presence of an unknown constant offset  $\xi$ , the statistical model becomes

$$y_i = a + bx_i + \epsilon_i + \xi; i = 1, N$$

with  $E(\epsilon_i) = E(\xi) = 0$  and, noting  $Var(\xi) = \sigma_o^2$ .

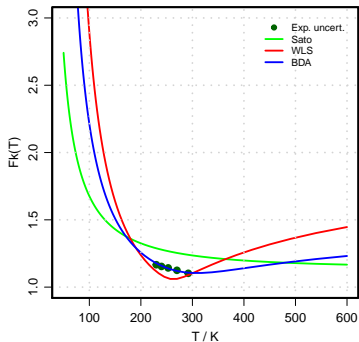
The data are now correlated, with

$$cov(y_i, y_j) = cov(\epsilon_i + \xi, \epsilon_j + \xi) = (u_{y_i}^2 - \sigma_o^2)\delta_{ij} + \sigma_o^2$$

+ new fit, with parameters  $a$ ,  $b$  and  $\sigma_o$  (use Bayesian data analysis)

# Extrapolation uncertainty vs. correlations

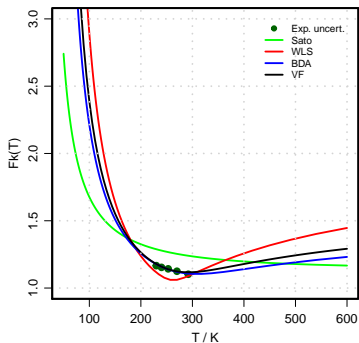
	$\ln A$	$E_a/R$ (K)	$\rho$	$\rho_{sys}$
Pernot (later)	$-22.16 \pm 0.39$	$530 \pm 117$	0.967	0.7





# Extrapolation uncertainty vs. correlations

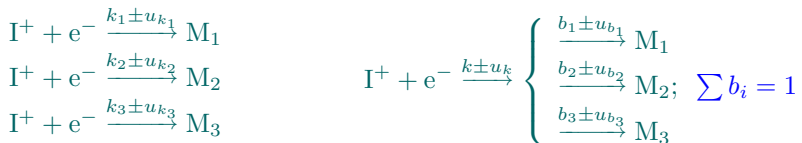
	$\ln A$	$E_a/R$ (K)	$\rho$
Variance Fitting <sup>8</sup>	$n/a \pm 0.46$	$n/a \pm 131$	0.971



<sup>8</sup>Nagy *et al.* (2012) *Rel. Eng. Sys. Saf.* **107**:29–34 (2012)

## Structural correlations - Multi-pathway react

- rate constants ( $k$ ) and branching ratios ( $b_i$ ) (BR) measured in different experiments



- Golden Rule of UP: **Do not mix separate uncertainty sources!**
  - essential for sum-to-one  $b_i$
  - large impact on uncertainty and sensitivity analysis<sup>9</sup>

<sup>9</sup>Carrasco et al. (2007) J. Phys. Chem. A 111:3507

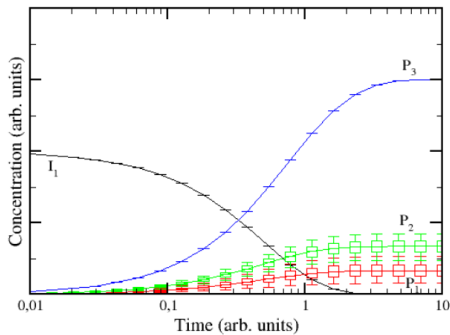
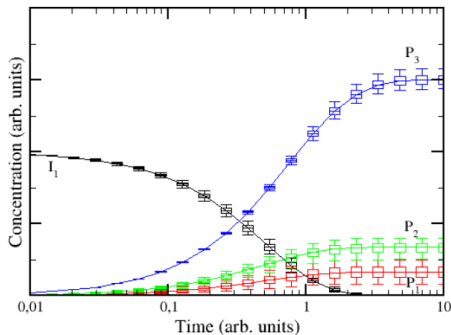
# Effect of ignoring sum-to-one correlation



with  $u(k_1) = u(k_2) = 0$  and  $b_1 + b_2 = 1$

$$k_{11} = k_1 * b_1, k_{12} = k_1 * b_2$$

$$\{k_{11}, k_{12}\} = k_1 * \{b_1, b_2\}_{\Sigma=1}$$



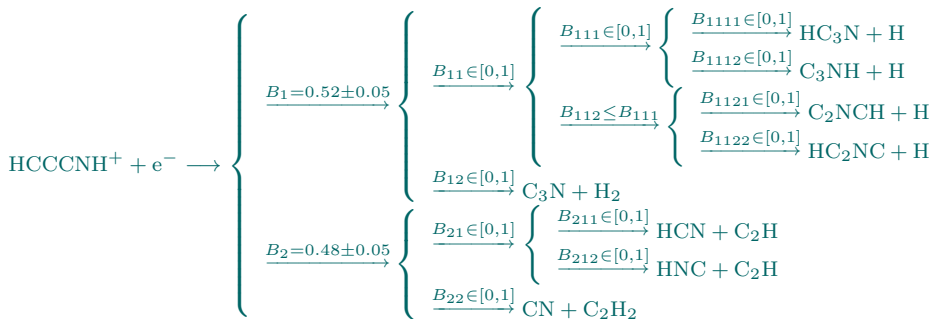
# Application to dissociative recombination

## The case of $\text{HCCCNH}^+$

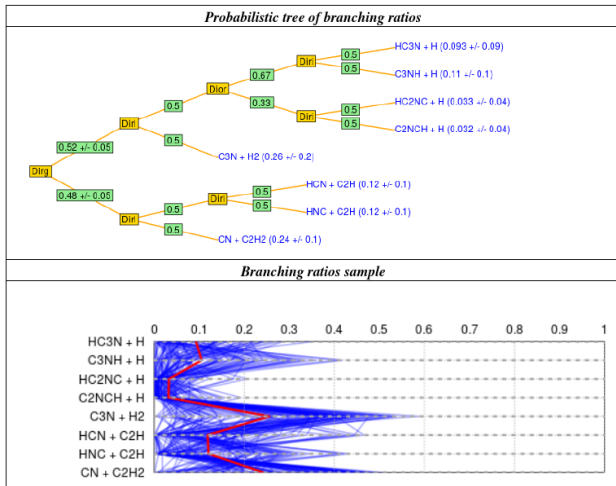
- 1 Geppert et al. [(2004) ApJ 613:1302-1309] measurements for  $\text{DCCCNND}^+$ 
  - $B_1 = 0.52 \pm 0.05$  for  $\{\text{DC}_3\text{N} + \text{D}; \text{C}_3\text{N} + \text{D}_2\}$
  - $B_2 = 0.48 \pm 0.05$  for  $\{\text{DCN} + \text{C}_2\text{D}; \text{CN} + \text{C}_2\text{D}_2\}$
- 2 isotope effects expected to be small: we can transpose information to  $\text{HCCCNH}^+$
- 3 Osamura et al. [(1999) ApJ 519:697-704]
  - interconversion barriers low enough for isomerisation of  $\text{HC}_3\text{N}$
  - formation of  $\text{HC}_3\text{N}$  more likely than  $\text{HC}_2\text{NC}$
- 4  $\text{HCN} / \text{HNC}$

## Application to dissociative recombination

## Resulting probabilistic tree



## Application to dissociative recombination

Implementation in MC-ChemDB<sup>10</sup>

<sup>10</sup>(<https://github.com/ppernot/MC-ChemDB>); Plessis et al. (2010) *J. Chem. Phys.* **133**:134110

# Computational chemistry UQ

- Several methods in 'traditional' CC try to provide prediction uncertainty (Composite methods, Bayesian Ensembles DFT, Statistical corrections. . . )
  - inexistent to disparate validation methods
- Surge of uncertainty validation needs from Machine Learning (for prediction, or for Active Learning)
  - several validation methods with unclear application limits (interval-based, variance-based, ranking-based)<sup>11</sup>
- The Calibration/Tightness framework provides a principled validation approach<sup>12</sup>

<sup>11</sup>Scalia *et al.* (2020) *J. Chem. Inf. Model.* (<https://tinyurl.com/yc7rn7dp>)

<sup>12</sup>Gneiting *et al.* (2007) *Stat. Meth. B* (<https://tinyurl.com/2p8nr3ab>); Pernot (2022) *J. Chem. Phys.* (<https://tinyurl.com/bdfy8snx>); Pernot (2022) (<https://arxiv.org/abs/2204.13477>)

# Calibration / Tightness

Test the statistical compatibility between uncertainties and errors, i.e.

$$E_i \sim D(\mu = 0, \sigma = u_{E_i})$$

where  $D$  is (often) an unknown distribution.

- (Average) calibration (assuming  $\langle E \rangle = 0$ )

$$\text{Var}(E) = \langle u_E^2 \rangle, \text{ or better, } \text{Var}(E/u_E) = 1$$

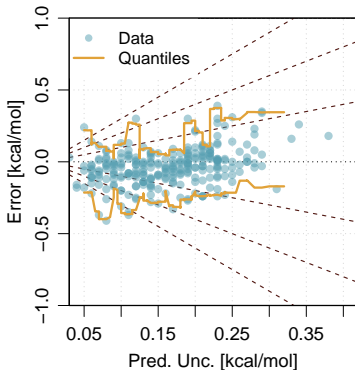
- Tightness (small-scale or local calibration):

$$\text{Var}(E | u_E^2 = \sigma^2) = \sigma^2, \forall \sigma^2$$



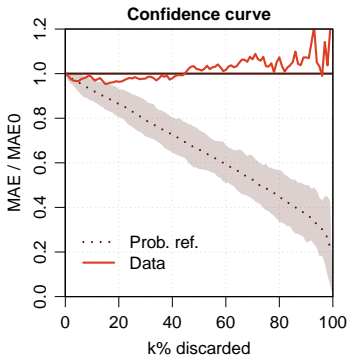
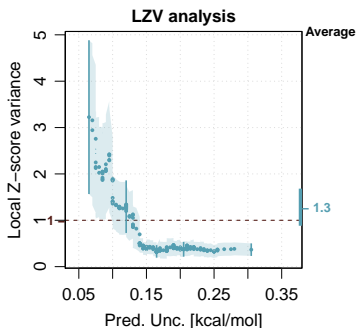
Formation heats by the mBEEF method<sup>13</sup>

- Bayesian Ensembles method inflates *parametric uncertainty* of exchange-correlation model to cover errors amplitude
  - strong functional constraints: tightness ???
  - does not disambiguate model uncertainty from reference data uncertainty
- Set of  $M = 257 \{V_i, R_i, u_{V_i}\}$



<sup>13</sup>Pandey and Jacobsen (2015) *Phys. Rev. B* (<https://tinyurl.com/5dv9spnn>), Pernot (2017) *J. Chem. Phys.* (<https://tinyurl.com/yb6uzwzr>), Pernot and Cailliez (2017) *AIChE J.* (<https://tinyurl.com/2xxcfs2f>)

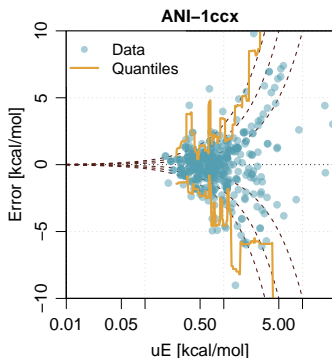
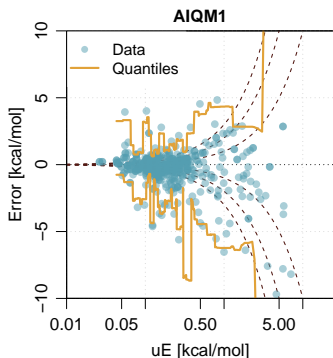
## Formation heats by the mBEEF method



- $\text{Var}(Z) = 1.3(2)$ , average calibration OK
- *Tightness is bad*: the LZV analysis shows that small PUs are underestimated, while large ones are overestimated
- The confidence curve is catastrophic (cannot use for Active Learning)

## Query by Committee UQ

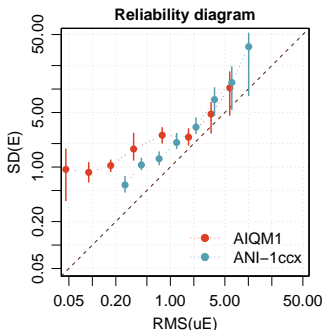
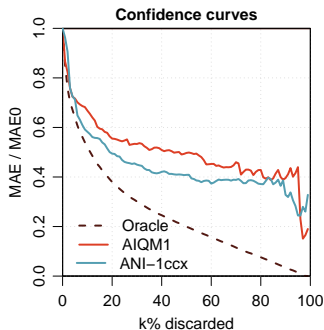
QbC uncertainties ( $n = 8$ ) on formation enthalpies for AIQM1 and ANI-1ccx<sup>14</sup>. Set of  $M = 472 \{E_i, u_{V_i}\}$



<sup>14</sup>Zheng *et al.* (2022) *J. Phys. Chem. Lett.* (<https://tinyurl.com/ccefk79z>)

# Query by Committee UQ

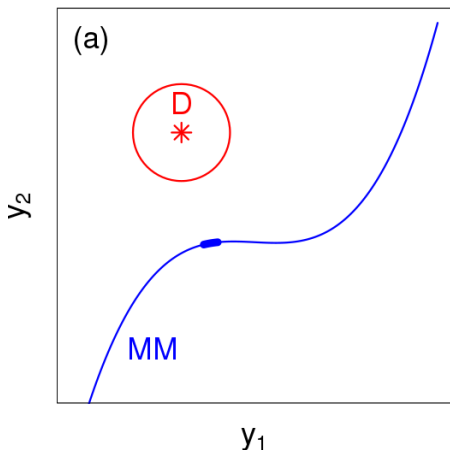
Stat	AIQM1	ANI-1ccx	Target
$Var(Z)$	59	4.3	1.4



- QbC does not provide a prediction uncertainty
- but both methods point reliably to largest errors (good for active learning !)

# Data space representation of model errors<sup>15</sup>

## Model inadequacy



- **Measurement model**

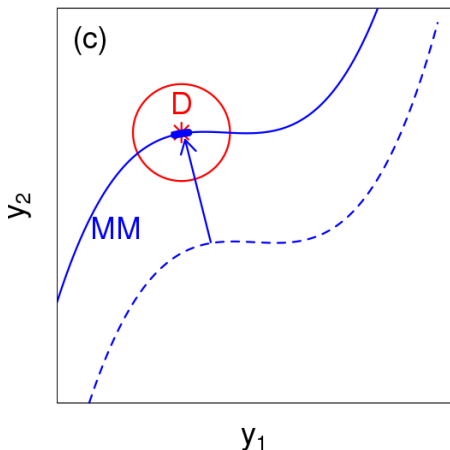
$$y_i = M(x_i; \vartheta) + \epsilon_i$$

- $N_\vartheta < N$  ( $N$  data points)
- **inadequacy**: no intersection between the model manifold (MM) and a high probability hypersphere around the data (D)

<sup>15</sup>Transtrum *et al.* (2011) *Phys. Rev. E* **83**:036701

Correction by model discrepancy function<sup>16</sup>

## Model Correction



- **Measurement model**

$$y_i = \alpha * M(x_i; \vartheta) + \delta(x_i; \tau) + \epsilon_i$$

- $\delta(x_i; \tau)$  deterministic (shift) or stochastic (GP, ML...)

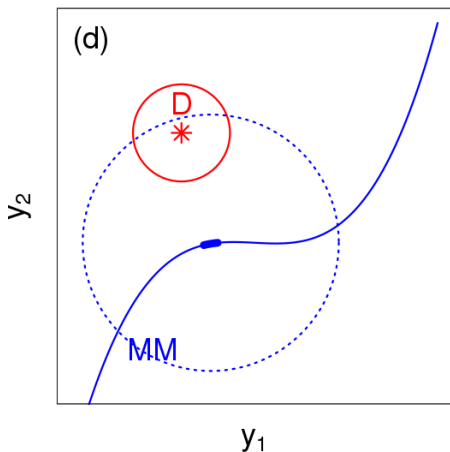
- **Problems**

- parameters identifiability
- $\alpha$  et  $\delta(x_i; \tau)$  not transferable

<sup>16</sup>Kennedy, M.C. & O'Hagan, A. (2001) *J. R. Stat. Soc. B* **63**:425.

External stochastic correction<sup>17</sup>

## Prediction correction



- Measurement model

$$y_i = M(x_i; \vartheta) + \delta + \epsilon_i$$

$$\delta \sim N(0, \sigma)$$

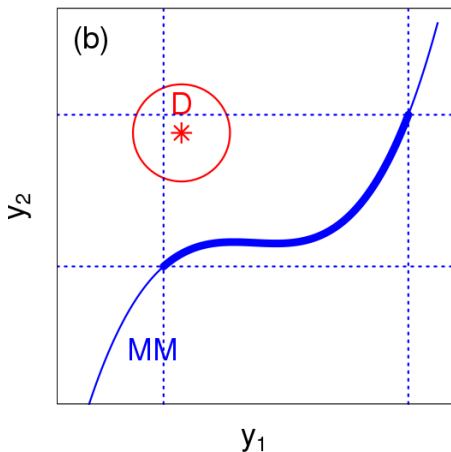
- Problem

- non-transferability of correction model

<sup>17</sup>P. Pernot *et al.* (2015) *J. Phys. Chem. A* **119**:5288-5304; Lejaeghere *et al.* (2014) *Crit. Rev. Sol. St. Mat. Sci.* **39**:1-24

Stochastic embedding<sup>18</sup>

## Uncertainty Inflation



- Measurement model

$$y_i = M[x_i; \vartheta + \delta(\vartheta)] + \epsilon_i$$

$$\delta \sim N(0, \Sigma_\theta)$$

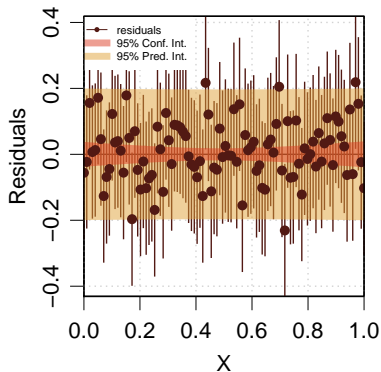
- the model parameters become stochastic variables, with a covariance matrix  $\Sigma_\theta$  adapted for model predictions to **marginally** overlap with the data
- transferability of  $\Sigma_\theta$  possible, but is it pertinent ???

<sup>18</sup>Sargsyan, K. *et al.* (2015) *Int. J. Chem. Kinet.* **47**:246; Pernot and Cailliez (2017) *AIChE J.* **63**:4642-4665; Pernot (2017) *J. Chem. Phys.* **147**:104102; Sargsyan, K. *et al.* (2018) *arXiv*:1801.06768



# Linear Calibration

## The valid case



- $N$  triplets  $\{x_i, y_i, u_{y_i}\}$
- Measurement model (WLS)

$$y_i = a + bx_i + \epsilon_i$$

$$\epsilon_i \sim N(0, u_{y_i})$$

- Confidence and prediction uncertainties at new point  $x_*$

$$y_* = \hat{a} + \hat{b}x_*$$

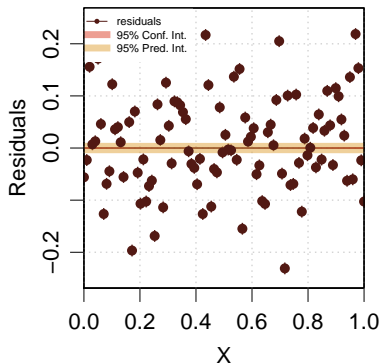
$$u_c^2(y_*) = u_a^2 + (u_b x_*)^2 + 2u_a u_b \rho_{ab} x_*$$

$$u_p^2(y_*) = u_c^2(y_*) + \langle u_y^2 \rangle$$

- **Note.**  $u_c^2(y_*) \rightarrow 0$  as  $N \rightarrow \infty$   
and/or  $u_y \rightarrow 0$

# Failure of standard calibration

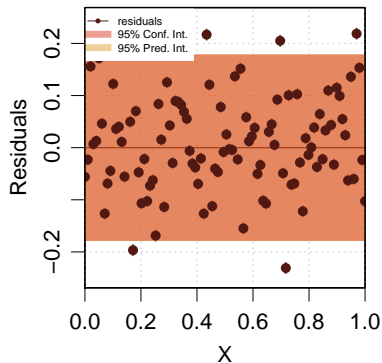
## Model inadequacy



- Same data, with smaller  $u_{y_i}$
- Same measurement model
- The measurement model is inadequate and statically invalid  
 $RMSD \gg \langle u_{y_i} \rangle$

**For inadequate models, one cannot make predictions using the standard calibration setup**

# Alternative models: Statistical Correction



- New measurement model

$$y_i = a + bx_i + \delta + \epsilon_i$$

$$\epsilon_i \sim N(0, u_{y_i})$$

$$\delta \sim N(0, \sigma)$$

Parameters:  $a$ ,  $b$  and  $\sigma$

- Confidence and prediction intervals at new point  $x_*$

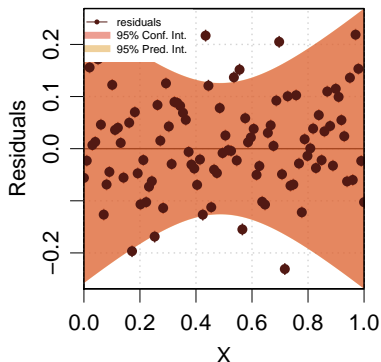
$$y_* = \hat{a} + \hat{b}x_*$$

$$u_c^2(y_*) = u_a^2 + (u_b x_*)^2 + 2u_a u_b \rho_{ab} x_* + \sigma^2$$

$$u_p^2(y_*) = u_c^2(y_*) + \langle u_y^2 \rangle$$

- **Problem:** non-transferability of  $\sigma$

# Alternative models: Stochastic embedding<sup>19</sup>



- Standard measurement model with scaled parameters covariance matrix

$$u'_c(y_*) = T * [u_a^2 + (u_b x_*)^2 + 2u_a u_b \rho_{ab} x_*]$$

i.e.,

$$\Sigma'_\vartheta = T * \Sigma_\vartheta$$

with  $T$  such as  $\langle u'_c{}^2(y_*) \rangle \simeq \text{Var}(R)$

- **Advantage:** transferability of  $\Sigma'_\vartheta$  (?)
- **Problems**
  - confidence intervals geometry is determined by model's jacobian
  - uncontrolled risk of local under/overestimation

<sup>19</sup>Pernot and Cailliez (2017) *AIChE J.* **63**:4642-4665; Pernot (2017) *J. Chem. Phys.* **147**:104102

# Conclusions

- Probabilistic uncertainty representation
  - use adapted pdfs that implement all physical constraints
  - **Correlation is your friend !**
- Uncertainty validation
  - **Calibration is easy, tightness is hard !**
  - Check calibration and tightness with appropriate methods
- Uncertainty estimation with inadequate models
  - Validate the statistical model
  - **Simplest ideas are not necessarily the best ones !**