Validation of Prediction Uncertainty in Computational Chemistry

Pascal PERNOT

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

CECAM (2022-06-22)

CC-UQ validation practices

- **Qualitative appreciation** of conformity between PU and errors amplitude
  - visual check of normality of z-scores histogram[1]
  - visual estimation of local coverage of 95% prediction interval (PI)[2]

- **Statistical estimation**
  - coverage of 95% prediction intervals[3]
  - correlation of uncertainty and absolute errors[4]
  - *calibration/sharpness* (for CC-applied ML methods)[5]

We need a consistent and shared validation framework !

[1] Mortensen *et al.* (2005) *Phys. Rev. Lett.* (https://tinyurl.com/mvwk3fff)

[2] Bakowies and von Lilienfeld (2021) *JCTC* (https://tinyurl.com/ms3dy7yv)

[3] Pernot *et al.* (2015) *J. Chem. Phys.* (https://tinyurl.com/3c9aw28r), Proppe & Kircher (2022) *ChemPhysChem* (https://tinyurl.com/yckxvjkk)

[4] Zheng *et al.* (2022) *J. Phys. Chem. Lett.* (https://tinyurl.com/ccefk79z)

[5] Tran *et al.* (2020) *Mach. Learn.: Sci. Technol* (https://tinyurl.com/2p849fs6), Scalia *et al.* (2020) *J . Chem. Inf. Model.* (https://tinyurl.com/yc7rn7dp)

Uncertainty vs error[6]

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

[6] *Guide to the expression of uncertainty in measurement* (GUM), *JCGM* **100**:2008,*International Vocabulary of Metrology* (VIM), *JCGM* **200**:2012

Error sources in Computational Chemistry[9]

- **Numerical errors**
  - finite arithmetics, stochastic algorithms...
  - mostly *random errors*; assumed to be well controlled[7], except for numerical chaos[8]
- **Parametric uncertainty**
  - semi-empirical methods, statistical corrections...
  - *random errors*; decrease with size of calibration set
- **Model errors**
  - level-of-theory errors (density functional approximation, correlation level, force-field expression...), representation errors (basis-sets, grids)...
  - mostly *systematic errors*; often the dominant error source; **no reason to be normally distributed**

[7] Irikura *et al.* (2004) *Metrologia* **41**:369
[8] Feher and Williams (2012) *J. Chem. Inf. Model.* **52**:3200-3212
[9] Lejaeghere (2020) *Uncertainty Quantification in Multiscale Materials Modeling*, pp. 41–46

CC-UQ outputs

- **Prediction distributions** or **representative samples**
  - available for some methods (Stochastic methods, Statistical models, Bayesian Ensembles. . .)
- Most UQ studies in the CC literature provide statistical summaries:
  - **expanded uncertainties** (U_p, typically for $p = 0.95$)[10]
  - **standard uncertainties** (u)

    Note: no prediction interval without distribution hypothesis

[10] Ruscic (2014) *Int. J. Quantum Chem.* **114**:1097

Validation data sets

Let us consider a typical validation set

- V_i : predicted value at point $i \in 1:M$
- u_{V_i} : uncertainty on V_i (*model* uncertainty)
- R_i : reference value
- u_{R_i} : uncertainty on R_i (*data* uncertainty)

Validation is based on

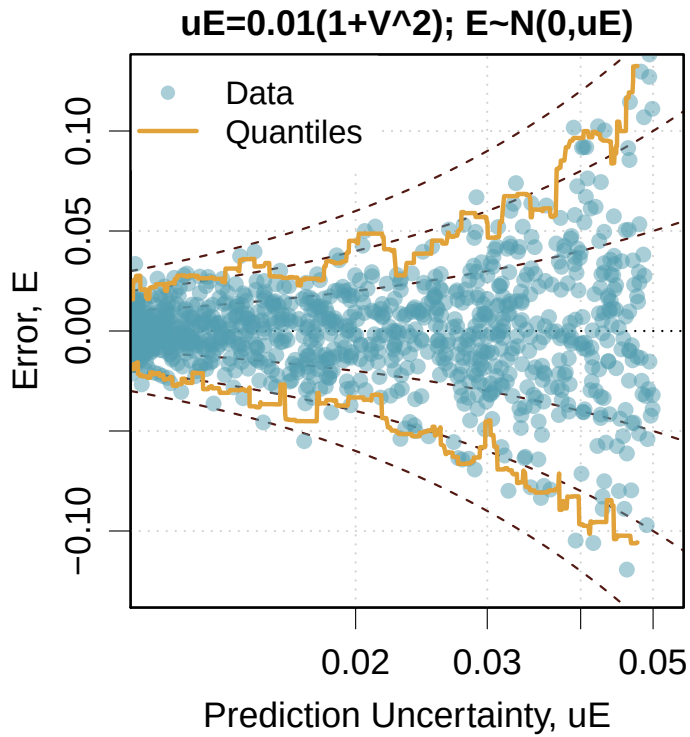
- $E_i = R_i - V_i$: error / prediction error
- for standard uncertainties $u_{E_i} = \sqrt{u_{V_i}^2 + u_{R_i}^2}$ (*prediction* uncertainty)

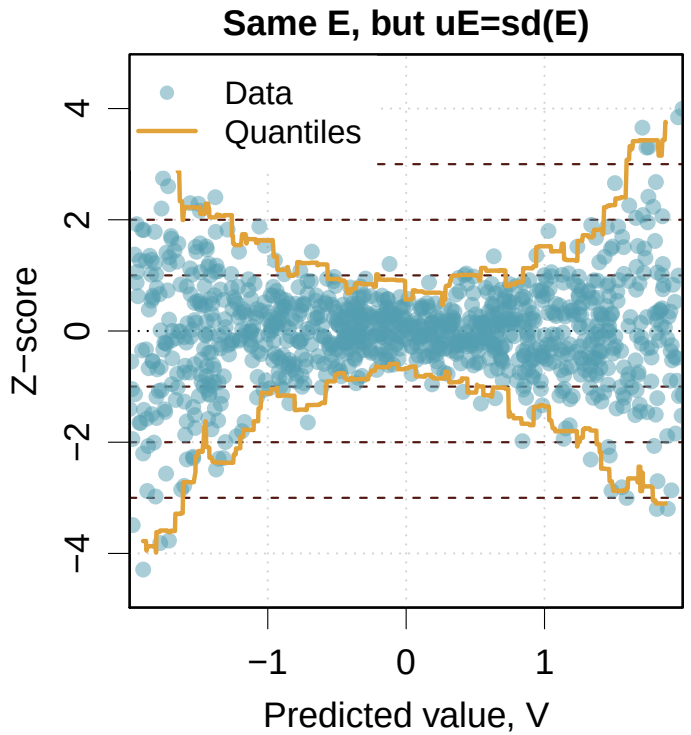
Prediction uncertainty quantifies the dispersion of pred. errors

Visual appreciation

Do the errors scale with uncertainty ?

1. if $u_E \neq c^{te}$, plot E vs u_E and guiding lines $y = k * x$
2. if $u_E = c^{te}$, plot E/u_E (z-score) vs V and guiding lines $y = k$
3. as helper, plot *running quantiles* (CI_{95})





The Calibration/Sharpness framework[12]

Calibration a method is considered to be *calibrated* if the confidence of predictions matches the probability of being correct for all confidence levels

Sharpness the concentration of a predictive distribution in absolute terms. Conditional to calibration

Pb: sharpness is a property of the forecast alone and does not involve the observations.
$\longrightarrow$ useful in benchmarking, not in validation...

Tightness[11] a method is considered to be *tight* if it is calibrated for any relevant subgroup of the validation data (small-scale calibration)

[11] Pernot (2022) (https://arxiv.org/abs/2204.13477)
[12] Gneiting et al. (2007) *Stat. Meth. B* (https://tinyurl.com/2p8nr3ab)

Limits of average calibration

- *Average* calibration does not guarantee tightness
  - in benchmarking, *sharpness* is used to select tighter forecasts
- Stronger calibration modes have been introduced:
  - *group* calibration[13] where calibration is assessed on relevant subgroups of the validation dataset
  - *adversarial* group calibration[14] where calibration is assessed on any random group of useful size
  - *perfect* calibration[15]
  - I propose to use **local** calibration, a variant of group calibration, where the validation set is split into contiguous areas of a chosen coordinate (predicted value, prediction uncertainty. . .)

[13] Chung *et al.* (2021) arXiv:2109.10254; Hébert-Johnson (2017) arXiv:1711.08513
[14] Zhao (2020) arXiv:2006.10288
[15] Levi *et al.* (2020) (http://arxiv.org/abs/1905.11659)

Different validation approaches[18]

- **Interval-based** [16]

$$\lim_{M\to\infty} \frac{1}{M} \sum_{i=1}^{M} \mathbf{1}\left(E_i \in I_{E_i,p}\right) = p,\ \forall\, p \in [0,1]$$

  where $I_{E_i,p}$ is a prediction interval at probability level p

- **Variance-based** [17]

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

  which *does not* operate at the same level as interval-based validation

- **Note**: *ranking-based* methods (correlation between u_E and $|E|$; confidence curves) cannot validate calibration/tightness, but can invalidate tightness...

[16] Kuleshov *et al.* (2018) (http://arxiv.org/abs/1807.00263)
[17] Levi *et al.* (2020) (http://arxiv.org/abs/1905.11659)
[18] Scalia *et al.* (2020) *J . Chem. Inf. Model.* (https://tinyurl.com/yc7rn7dp), Pernot (2022) _J. Chem.

Interval-based validation

An *expanded uncertainty* is the half-width of a prediction interval

$$I_{E_i,p} = [-U_{E_i,p}, U_{E_i,p}]$$

To validate $U_{E,p}$, one should therefore test

$$p \overset{?}{\in} I_{95}(\nu_p, M), \text{ where } \nu_p = \frac{1}{M}\sum_{i=1}^{M} \mathbf{1}\left(|E_i| \leq U_{p,i}\right)$$

and ν_p is a PICP (Prediction Interval Coverage Probability)

- $I_{95}(\nu_p, M)$ (Binomial Proportion CI) is estimated by a method avoiding normality hypothesis (Clopper-Pearson, cc-Wilson, Agresti-Coull...)[19]

[19] Vollset (1993) *Stat. Med.* (https://tinyurl.com/5dps8u3h)

Variance-based validation

Let us consider unbiased errors of *unknown* distribution

$$\mathrm{E}(E_i) = 0;\ \mathrm{Var}(E_i) = u_{E_i}^2$$

Then for *z*-scores $z_i = E_i/u_{E_i}$ one has

$$\mathrm{E}(z_i) = 0;\ \mathrm{Var}(z_i) = 1$$

and for a set of *z*-scores $Z = \{z_i\}_{i=1}^{M}$

$$\mathrm{E}(Z) = 0;\ \mathrm{Var}(Z) = 1$$

To validate $\mathrm{Var}(Z)$, one should therefore test

$$1 \overset{?}{\in} I_{95}\left(\mathrm{Var}(Z), M\right)$$

- $I_{95}\left(\mathrm{Var}(Z), M\right)$ is estimated by an adapted bootstrap method (BC_a, ABC. . .)[20] to avoid the normality-based textbook method

[20] Diciccio and Effron (1996) *Stat. Sci.* (https://tinyurl.com/ssztxy6k)

How to test tightness ?

- **Proposed approach**: interpret tightness as **local calibration** and use calibration tests on subsets of the validation set wrt V or uE (or any other relevant property)
  - LCP analysis: local PICP estimation and test
  - LZV analysis: local *z*-scores variance testing

- **Pb**: partitioning reduces sample size (bad for test power)
  $\longrightarrow$ use overlapping/sliding areas for small datasets.
  Trends in LCP-LZV curves help diagnostic.

- Link of LZV/uE with *perfect* calibration (reliability diagram, RD)[21]

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

  but RD does not deal with homoscedastic datasets.

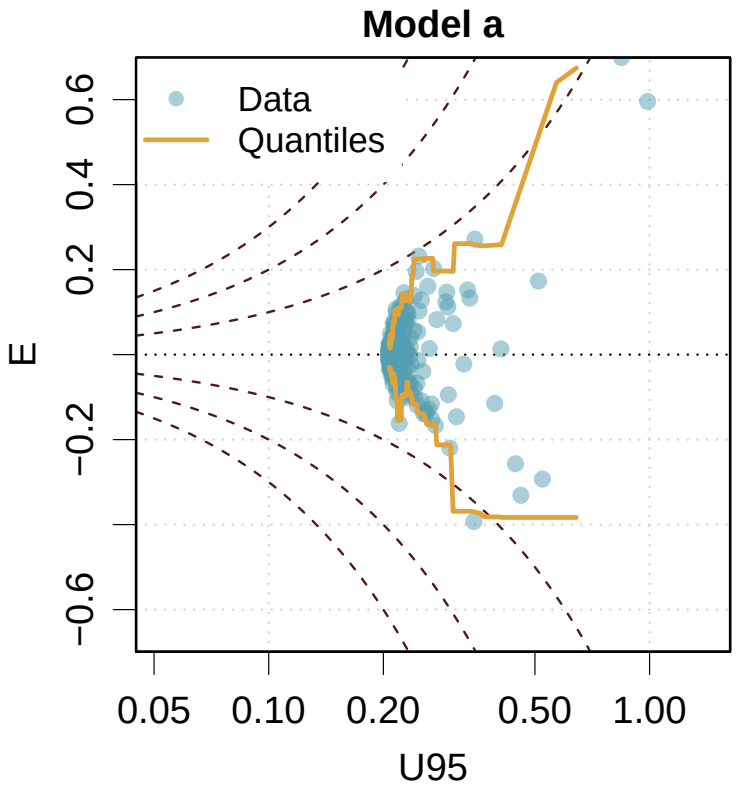
[21] Levi *et al.* (2020) (http://arxiv.org/abs/1905.11659)

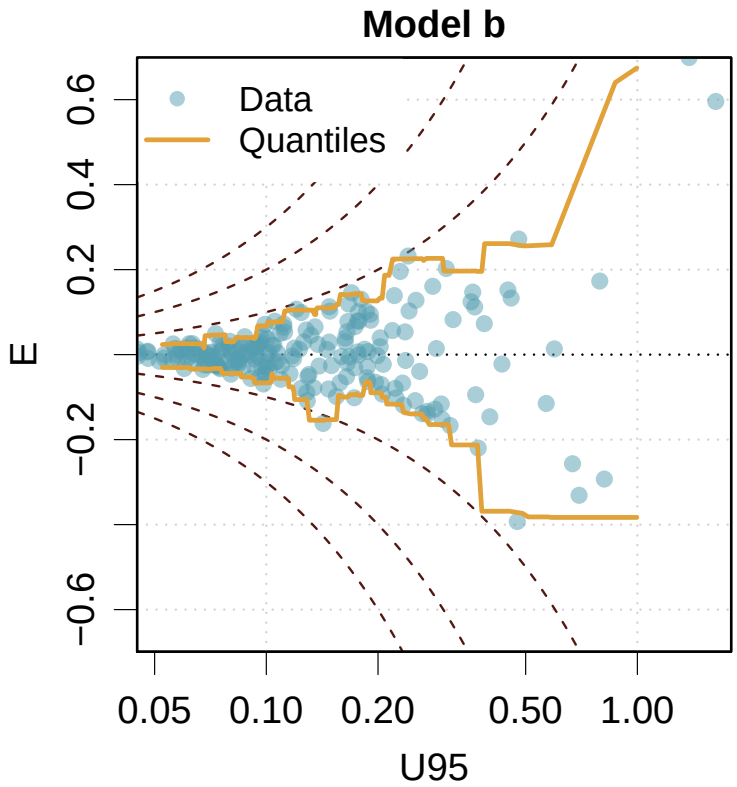
Overview

Diagnostic	Applicability					Validation	
	q_E	u_E	$U_{E,p}$	Homosc.	Heterosc.	Calibrat.	Tightness
Average							
PIT hist.	✓	X	X	✓	✓	✓	X
Calib. curve	✓	X	X	✓	✓	✓	X
PICP	✓	X	✓	✓	✓	✓	X
Var(Z)	✓	✓	X	✓	✓	✓	X
Cor(u_E,$\lvert E\rvert$)	✓	✓	✓	X	✓	X	X*
Local							
LCP/LRR	✓	X	✓	✓	✓	✓†	✓
LZV	✓	✓	X	✓	✓	✓†	✓
Reliab. diag.	✓	✓	X	X	✓	✓†	✓
Confid. curve	✓	✓	✓	X	✓	X	X*

Reactivity Scales[22]

Set of 212 errors for reaction rates by an extended Mayr's reactivity scale and U_{95} values obtained by two UQ methods (*a* and *b*).

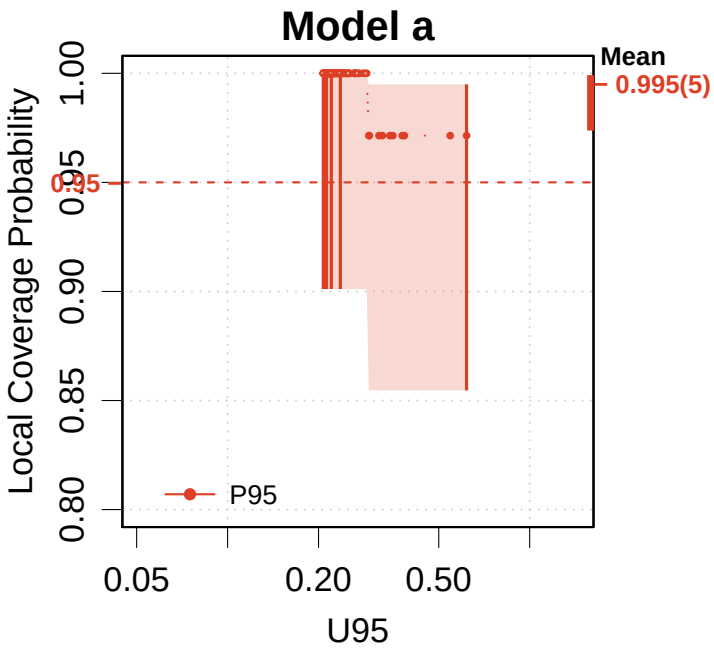


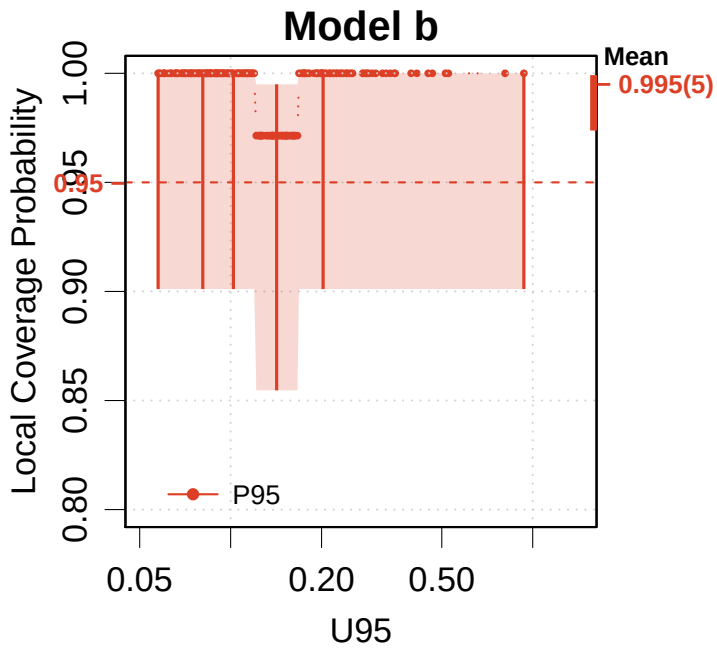


[22] Proppe & Kircher (2022) *ChemPhysChem* (https://tinyurl.com/yckxvjkk)

Reactivity Scales[23]

$U_{95} \Longrightarrow$ interval-based validation

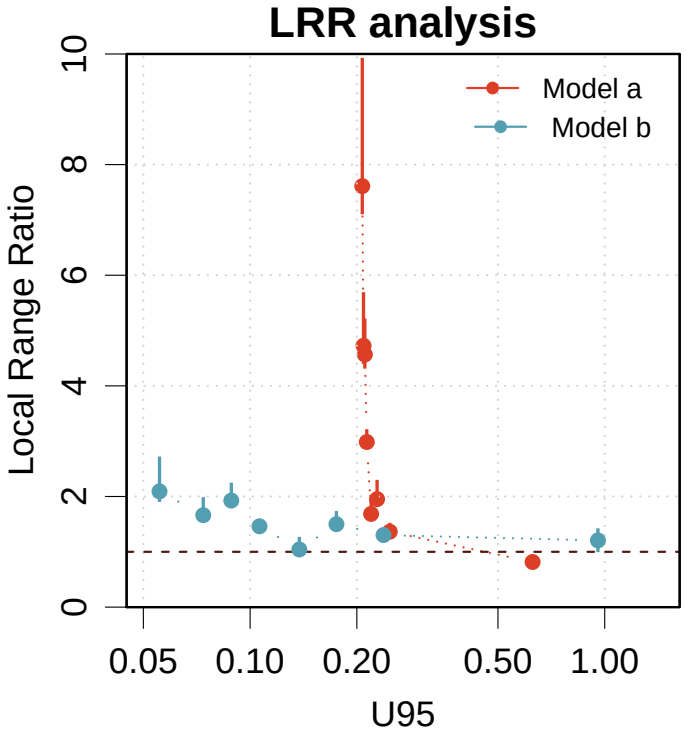


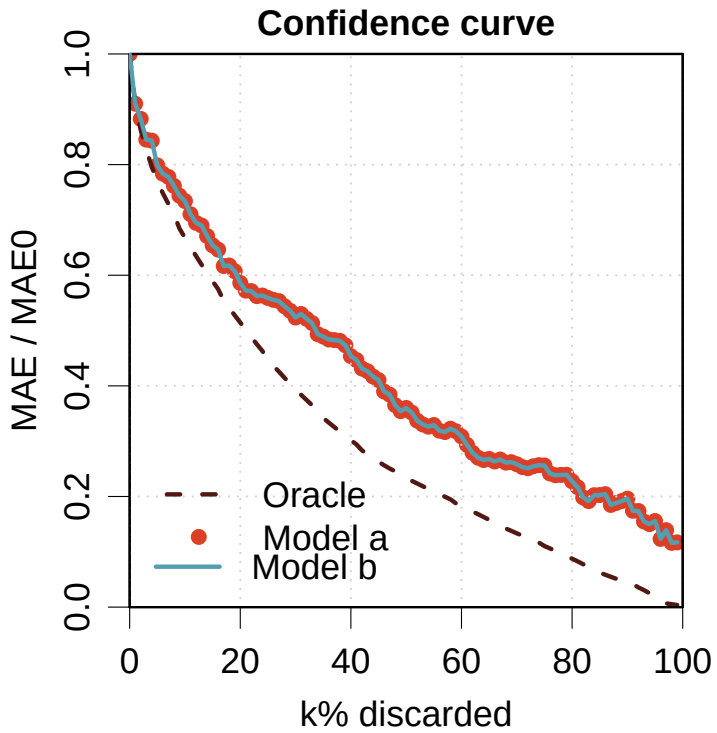


- large local uncertainties because of small groups
- for overestimated uncertainties, the PICP test saturates at 1

[23] Proppe & Kircher (2022) *ChemPhysChem* (https://tinyurl.com/yckxvjkk)

Reactivity Scales[24]





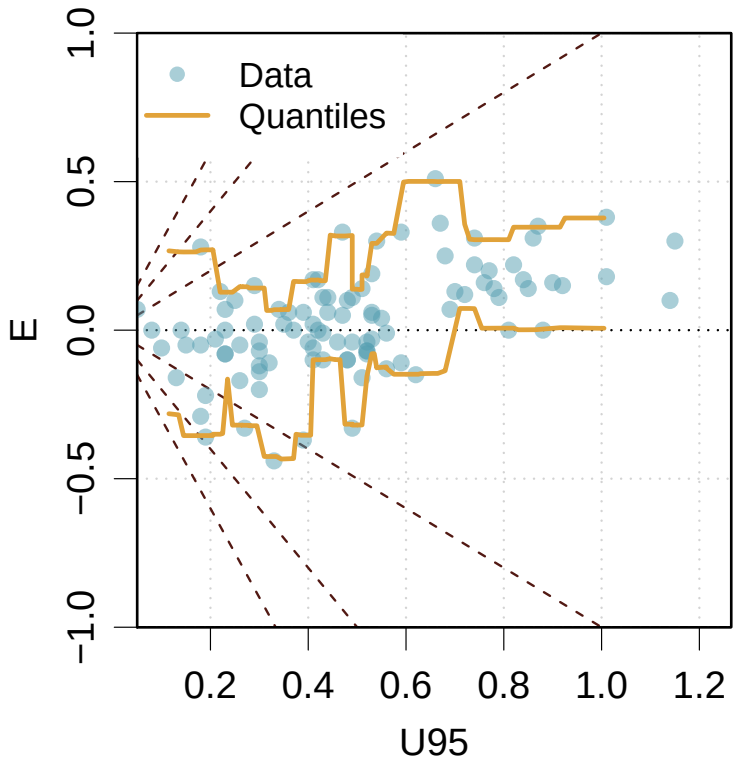
- **Range Ratio** : Mean width of pred. int. / Width of error proba. int.
- even if calibration is rejected, on might reliably use these uncertainties for active learning (Conf. curves)

[24] Proppe & Kircher (2022) *ChemPhysChem* (https://tinyurl.com/yckxvjkk)

ZPE by the ATOMIC-2 composite method[25]

A-posteriori estimation of U_{95} based on a correlation of errors with the fraction of heteroatoms in a molecule.

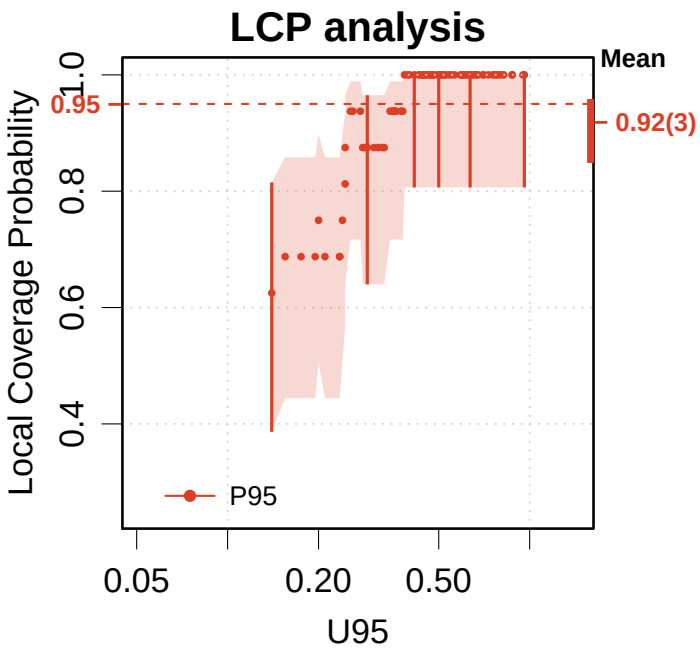
- Small test dataset: $M = 99$ $\{V_i, R_i, U_{95,V_i}\}$
- Reference data: CCSD(T) (no uncertainty)
- Authors validate by visual appreciation of error bars

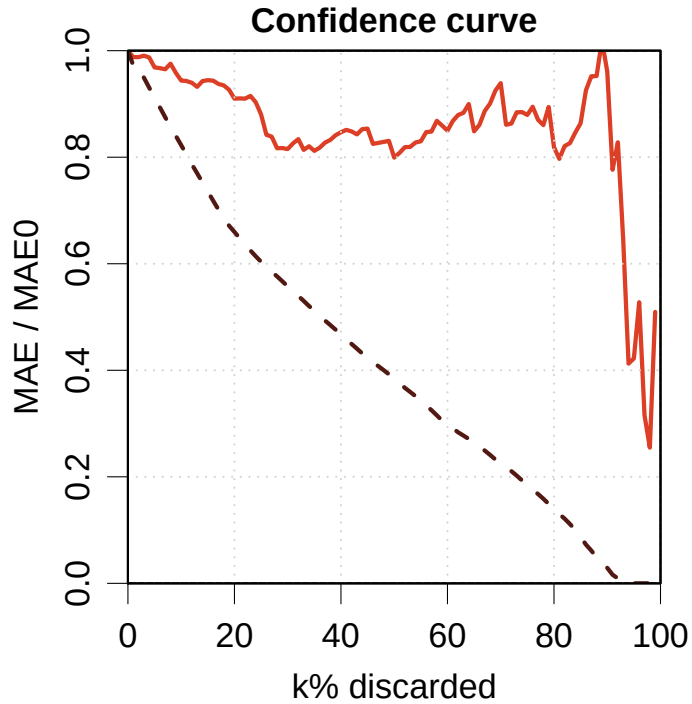


[25] Bakowies and von Lilienfeld (2021) *JCTC* (https://tinyurl.com/ms3dy7yv), Pernot (2022) *J Chem Phys* (https://doi.org/10.1063/5.0084302)

ZPE by the ATOMIC composite method

PICP testing at $p = 0.95$

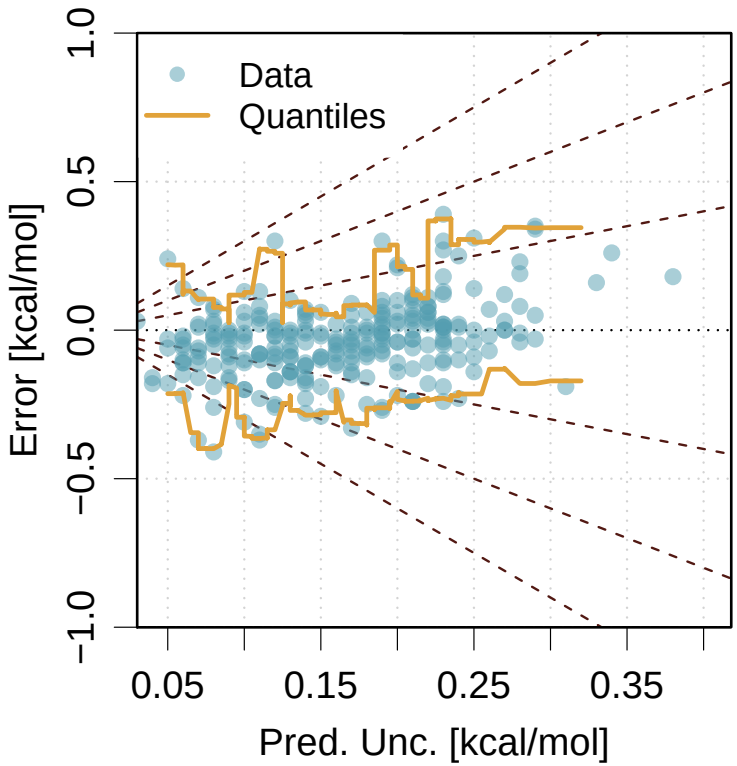




- $\nu_{0.95} = 0.92(3)$: average calibration is OK
- large uncertainty on PICPs, but the **trends are informative**
- from the LCP analysis, one sees a systematic bias:
  small PUs are underestimated, large ones are overestimated

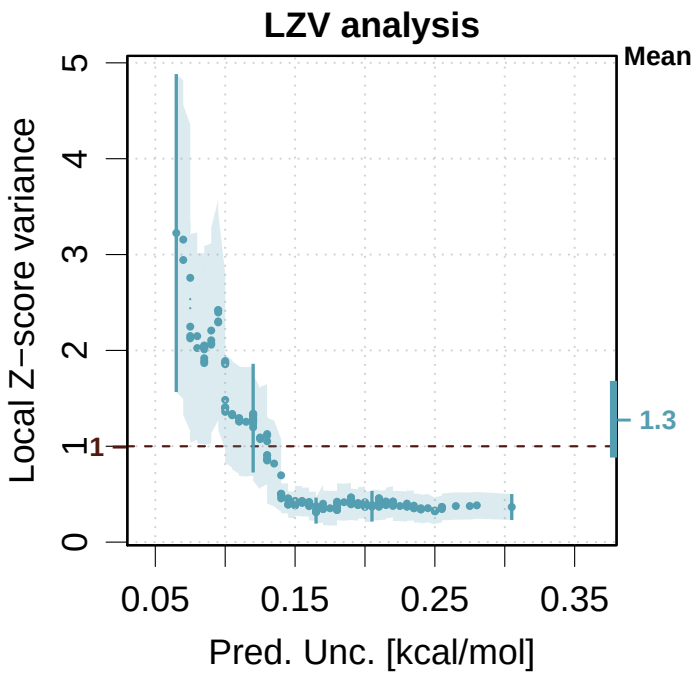
Formation heats by the mBEEF method[26]

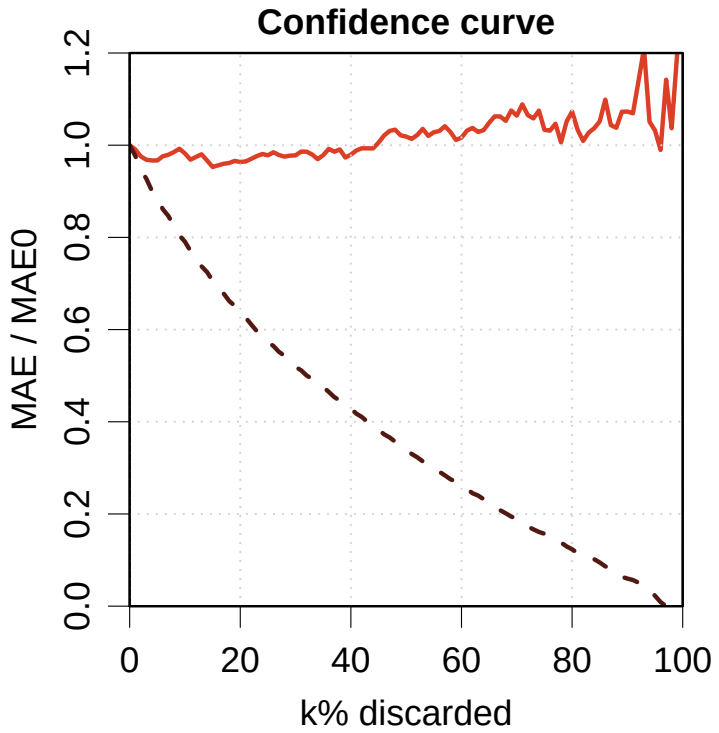
- Bayesian Ensembles method inflates *parametric uncertainty* of exchange-correlation model to cover errors amplitude
  - strong functional constraints: tightness ???
  - does not disambiguate model uncertainty from reference data uncertainty
- Set of $M = 257$ $\{V_i, R_i, u_{V_i}\}$
- R_i experimental, no uncertainty provided



[26] 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

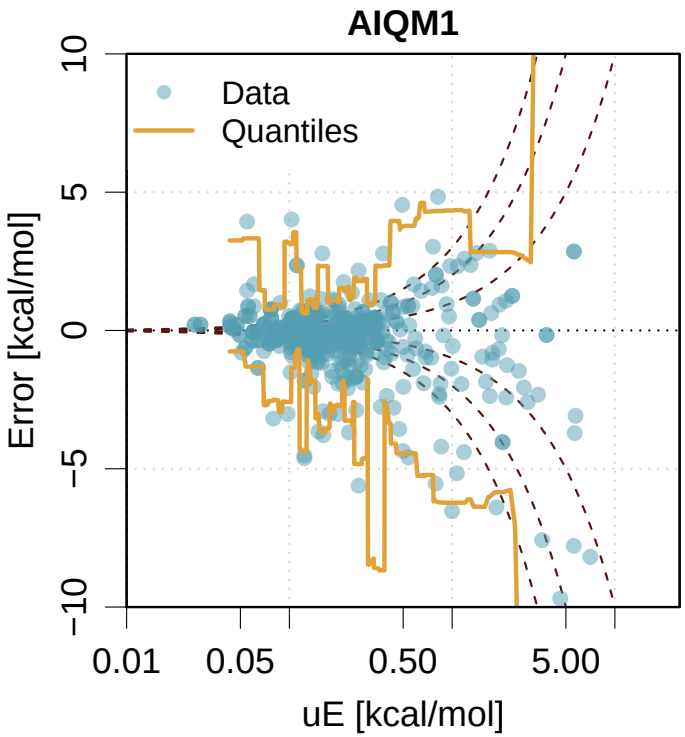


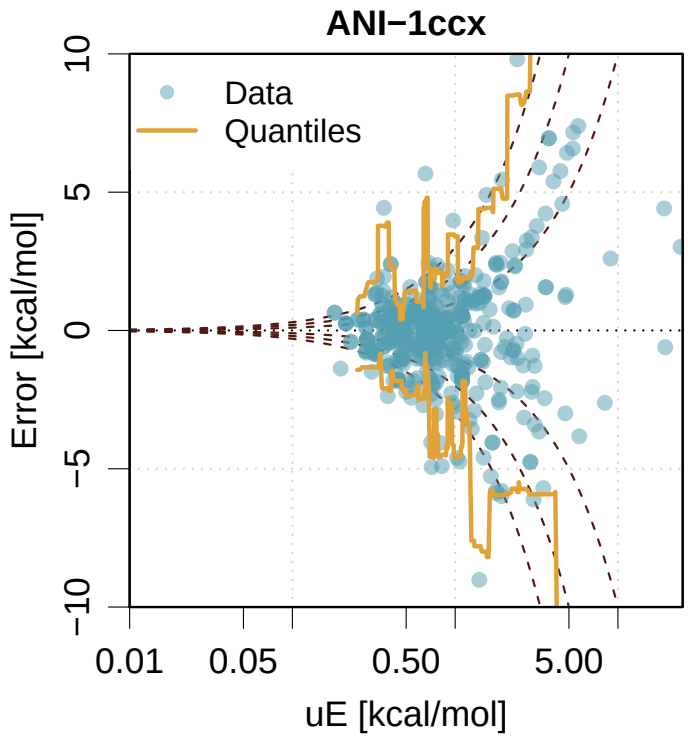


- $\mathrm{Var}(\mathrm{Z}) = 1.3(2)$, *average calibration* OK
- the LZV analysis shows that small PUs are underestimated, while large ones are overestimated
- the confidence curve is catastrophic
- these uncertainties should not be trusted

Query by Commitee UQ

QbC uncertainties ($n = 8$) on formation enthalpies for AIQM1 and ANI-1ccx[27]. Set of $M = 472$ $\{E_i, u_{V_i}\}$

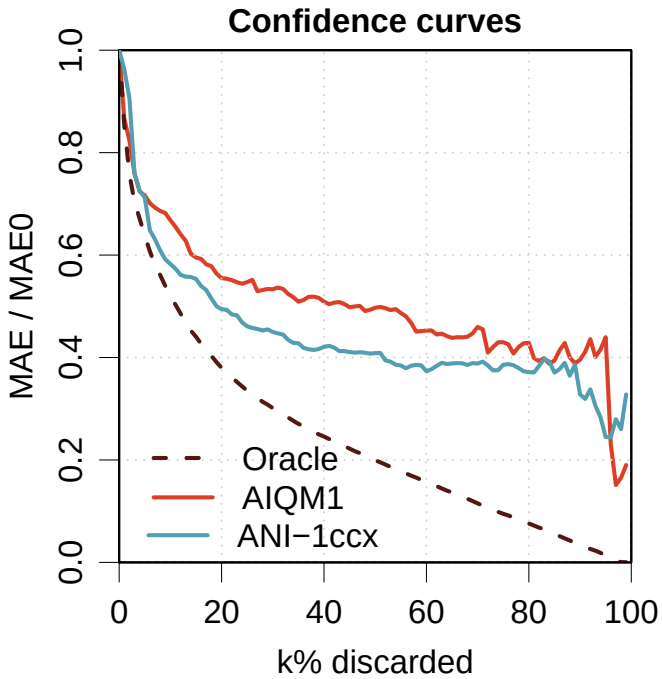


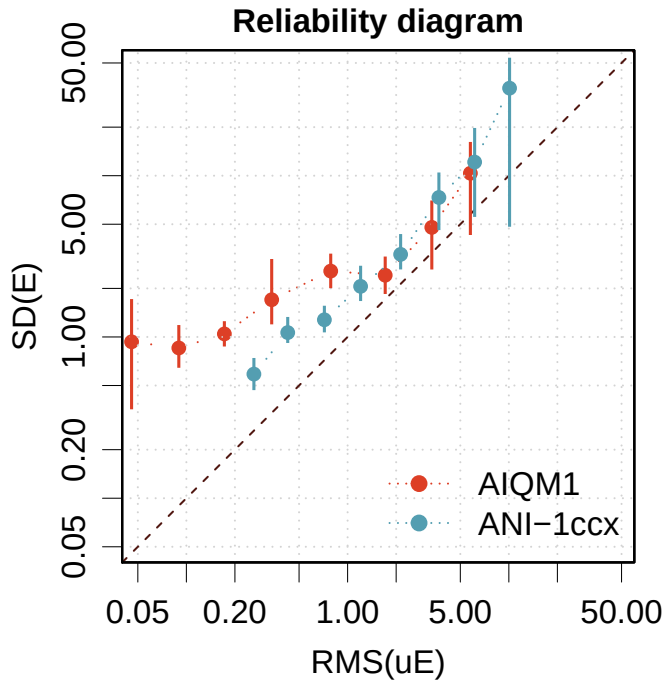


[27] Zheng *et al.* (2022) *J. Phys. Chem. Lett.* (https://tinyurl.com/ccefk79z)

Query by Commitee UQ

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

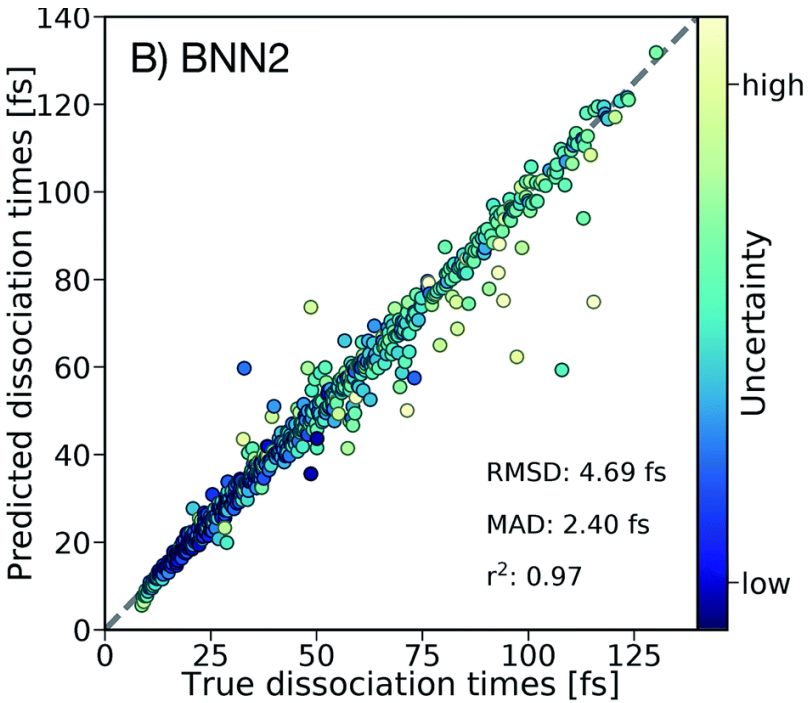


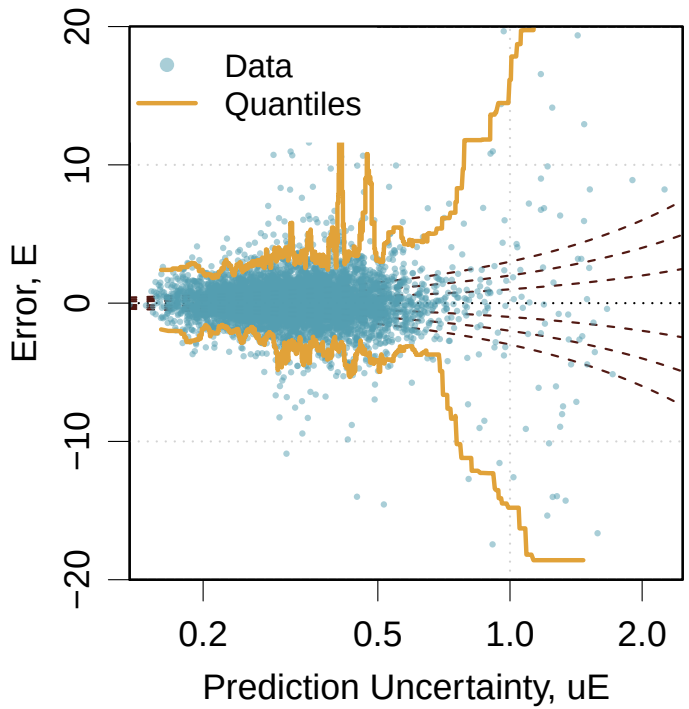


- QbC does not provide a prediction uncertainty
- but both methods point reliably to largest errors (good for active learning !)

Bayesian Neural Network

Data issued from a BNN trained to predict a MD potential[28] ($M = 5923$)

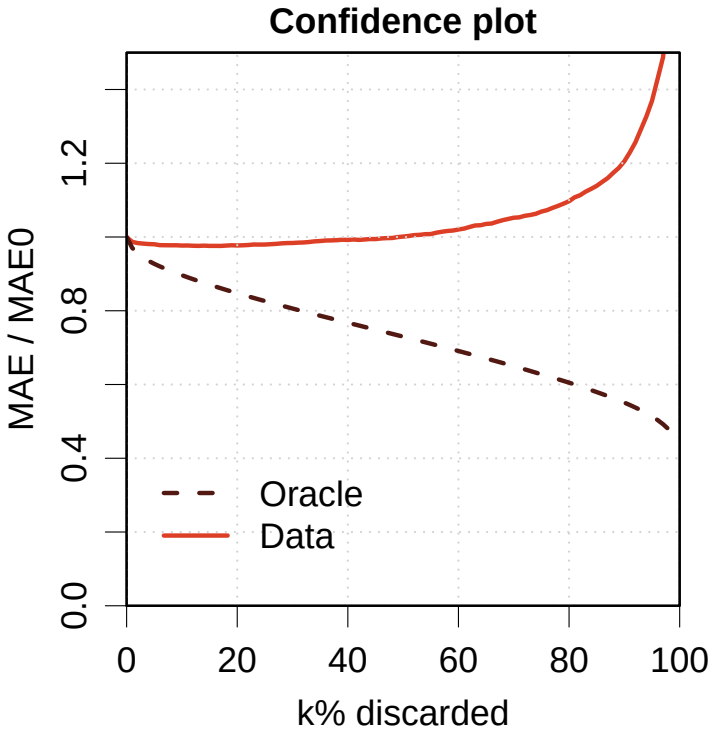


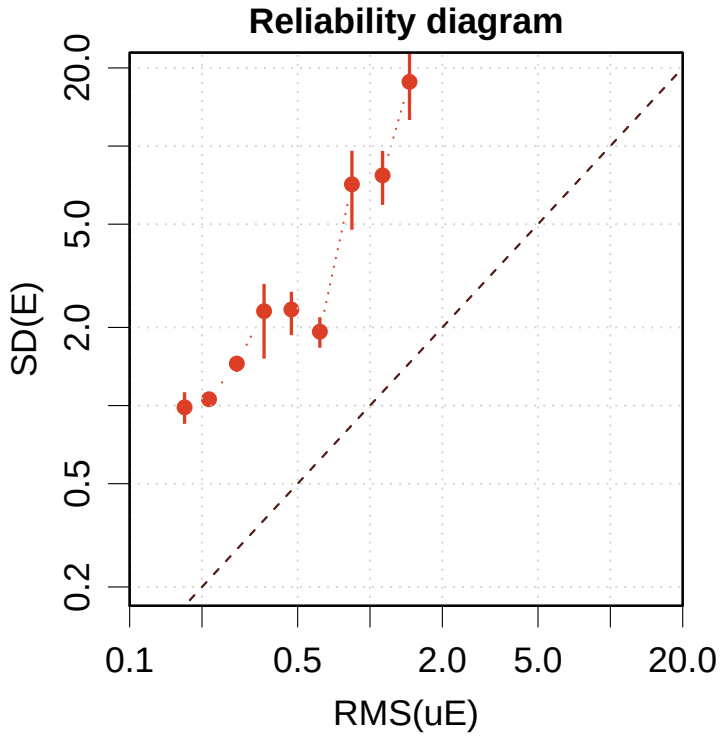


- The color scale for uncertainty is not a proper tool for validation

[28] Häse *et al.* (2019) *Chem. Sci.* **10**:2298

Bayesian Neural Network





- This BNN uncertainty is NOT a prediction uncertainty ($Var(Z) = 30$) and should not be used for active learning...

Conclusions

Calibration is easy, tightness is tough !

- **Calibration/Tightness : a principled framework for UQ validation**
- **CC-adapted C/T validation methods**
  - *standard PU* : test z-scores variance (LZV analysis), or build RD
  - *expanded PU* : test PICP values (LCP/LRR analysis)
- **CC-UQ methods rarely reach calibration and/or tightness**
  - datasets often too small for solid conclusions
  - should we loosen the validation criteria ?
  - how-much mis-calibration/mis-tightness is acceptable for a given application ? (e.g. calibration is not useful for active learning. . .)

Warmful thanks to. . .

- **Andreas SAVIN** (LCT, Jussieu)
  for enlightening discussions
- **Jonny Proppe** (Univ. Göttingen)
  for the PRO2022 dataset
- **Morgane Vacher** (Univ. Nantes)
  for the HAS2019 dataset
- and **YOU**
  for your attention !