# Validation of prediction uncertainty in ML

Pascal PERNOT (pascal.pernot@cnrs.fr)

**Institut de Chimie Physique,**
**UMR8000, CNRS/Univ. Paris-Saclay**


IA Symposium (2023-05-26)

# Why UQ ?

- ML models have many parameters and a high risk to be overly sensitive to small variations of inputs, notably for out-of-the-box predictions.
- Assignment of an uncertainty to each prediction is expected to flag predictions with outstanding values.
- UQ is also a necessity when ML predictions replace physical experiments (Virtual Measurements).

*UQ validation tests the reliability of prediction uncertainty*

# UQ metrics and validation model

## UQ validation methods depend on UQ information

- full distribution
- prediction intervals or expanded uncertainty (half-range of a probability interval)
- **uncertainty** (variance-based UQ metric)

***Prediction uncertainty quantifies the dispersion of errors***

*Validation is based on a probabilistic model*

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

*where* $D(\mu, \sigma)$ *is a distribution of errors (a priori unknown) with mean* $\mu$ *and standard deviation* $\sigma$ *and errors should be unbiased* ($< E >= 0$)

# Notations: UQ validation dataset

Let us consider a typical validation set

- $X_i$ : input feature(s) at point $i \in 1:M$
- $V_i$ : 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)

From which one gets

- $E_i = R_i - V_i$ : [prediction] error
- $u_{E_i} = \sqrt{u_{V_i}^2 + u_{R_i}^2}$ (*prediction* uncertainty)

# Validation goals

**Validation goals depend on the intended use of uncertainty**[1]

- **Internal use** (e.g. *active learning*)
  - small uncertainties should imply small errors
  - calibration is not necessary (need some form of correlation)

- **External use**: prediction uncertainty has to match real world requirements (e.g. high-throughput screening of materials that have to be tested experimentally)
  - uncertainty should be **calibrated**
  - *Consistency*: $E$ and $u_E$ should be *statistically consistent*
  - *Adaptivity*: $u_E$ should be reliable for all input features $X$

[1] Pernot (2022) *J. Chem. Phys.* **157**:144103; Pernot (2023) arXiv:2303.07170

# Variance-based tests of average calibration

Assuming *unbiased errors*, one should have

$$Var(E) \simeq < u_E^2 >$$

or, better (as it accounts for the $(E_i, u_{E_i})$ pairing)[2]

$$Var(Z = E/u_E) \simeq 1$$

*Average calibration is a necessary condition, but it does not guarantee consistency nor adaptivity, as it might result from the compensation of under- and over-estimation of $u_E$.*

[2] Pernot (2022) *J. Chem. Phys.* **157**:144103; Pernot (2023) arXiv:2303.07170

# Variance-based tests of conditional calibration

- **Reliability of uncertainty at all levels**:
  *Consistency* can be expressed as *conditional calibration*[3] wrt $u_E$

$$Var(E|u_E = \sigma) = \sigma^2, \ \forall \sigma > 0$$

or

$$Var(Z|u_E = \sigma) = 1, \ \forall \sigma > 0$$

- **Reliability of uncertainty throughout features space**:
  *Adaptivity* is *conditional calibration*[4] wrt $X$

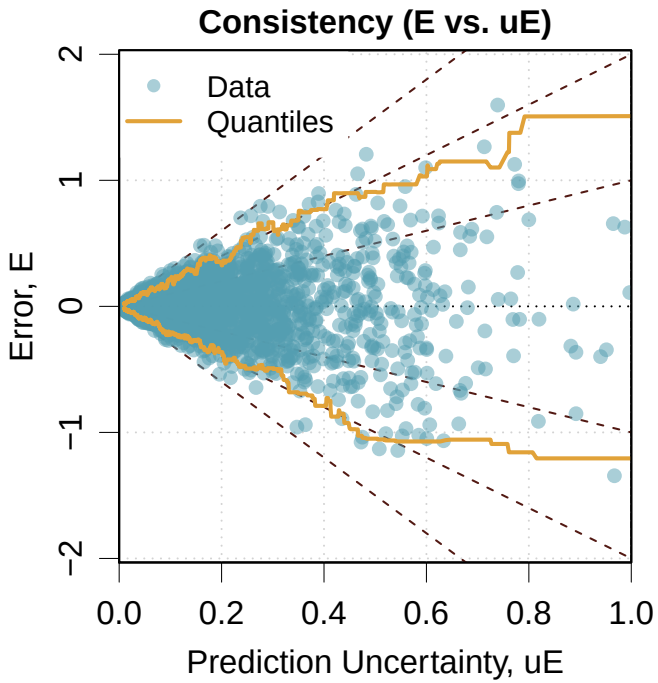
$$Var(Z|X = x) = 1, \ \forall x \in \mathcal{X}$$

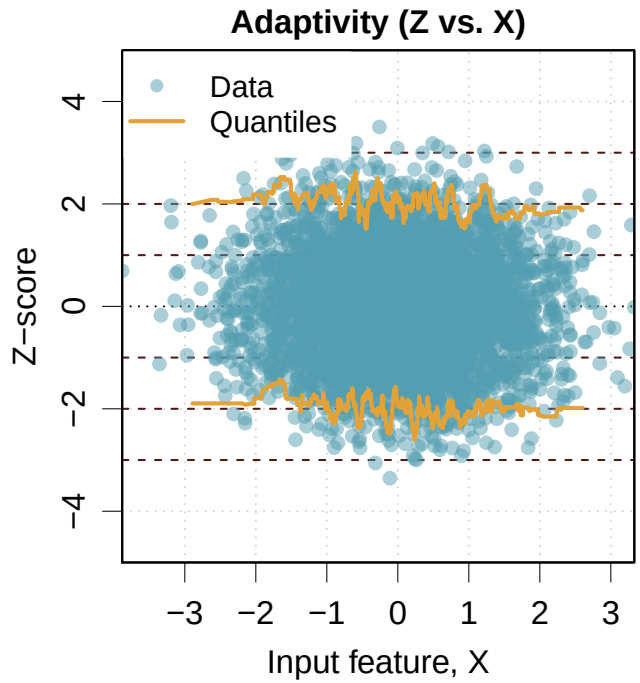
[3] Levi *et al.* (2020) arXiv:1905.11659

[4] Angelopoulos & Bates (2021) arXiv:2107.07511; Pernot (2023) arXiv:2303.07170

# Exploratory plots

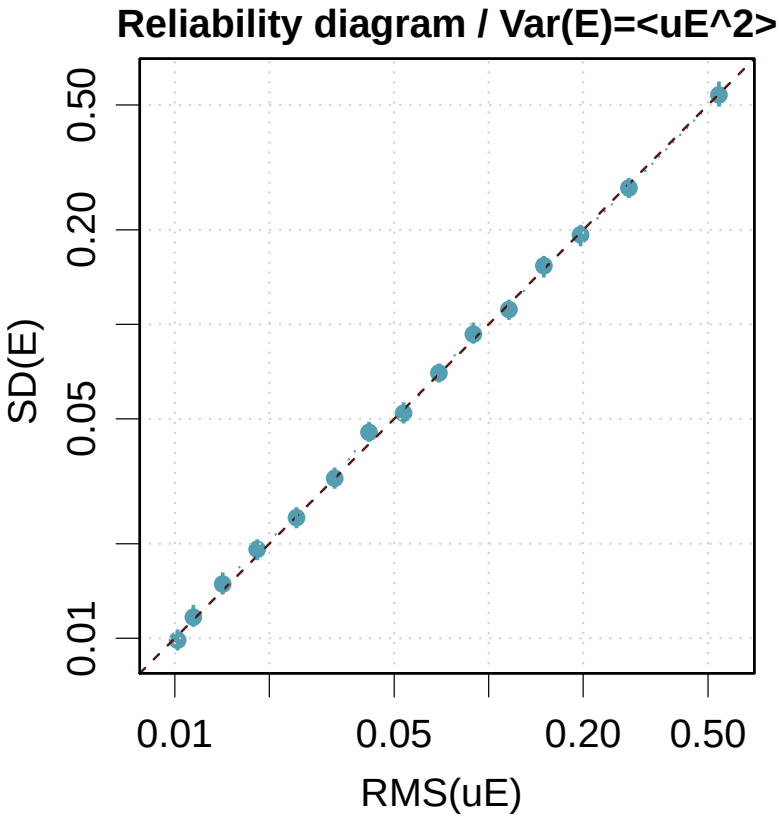
- plot $E$ (or $Z$) vs $u_E$ and guiding lines $y = k * x$ (if $u_E \neq c^{te}$)
- plot $Z = E/u_E$ ($z$-score) vs $X$ and guiding lines $y = k$
- plot *running quantiles* ($CI_{95}$)

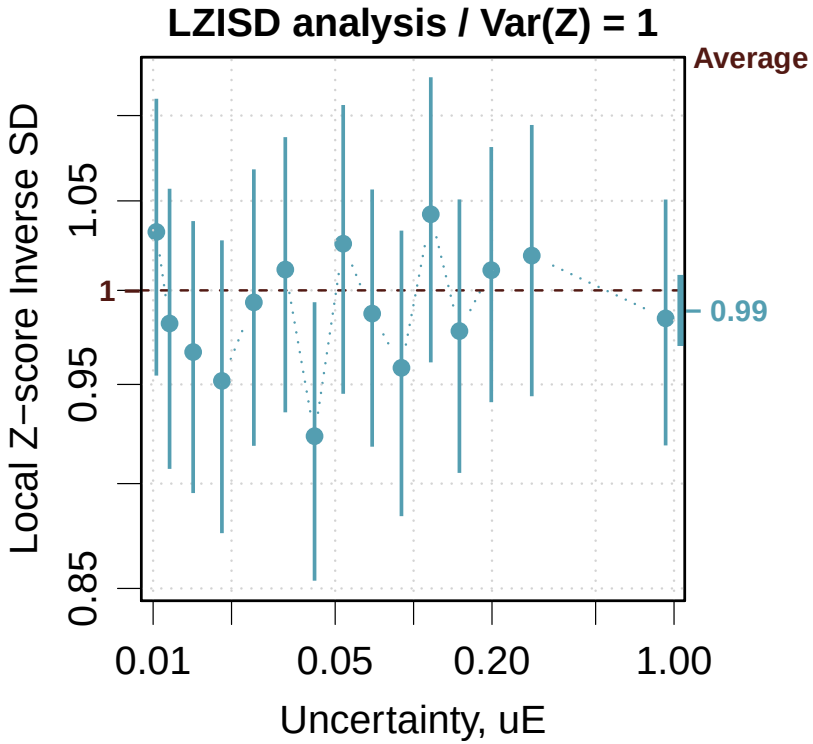



- An incorrect shape is sufficient to reject calibration, consistency or adaptivity, but if the plot seems OK, one needs a more quantitative approach.

# Binning-based consistency tests[5]



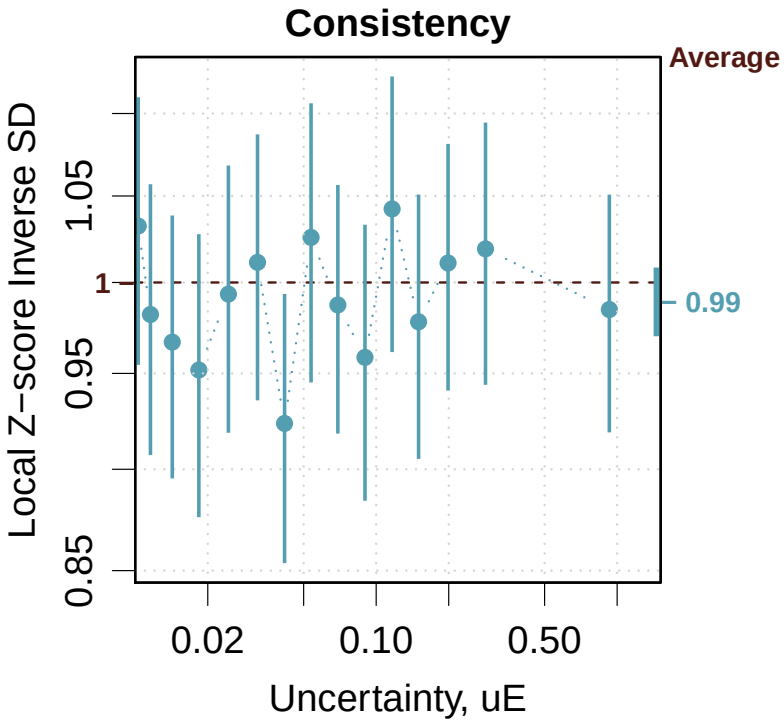



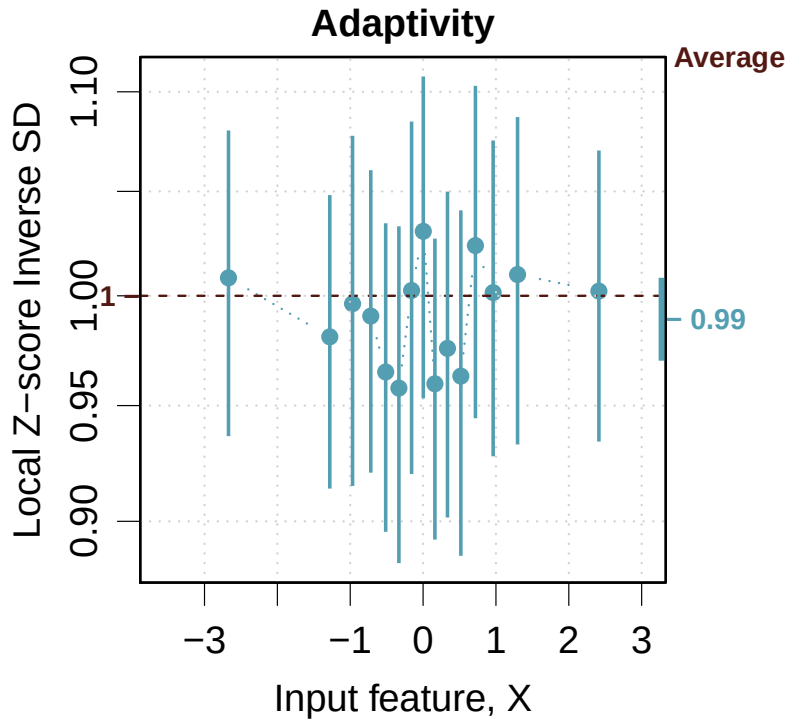

- deviation of a few points, without notable trend, is statistically expected (error bars are 95% probability intervals)

[5] Levi *et al.* (2020) arXiv:1905.11659; Pernot (2022) *J. Chem. Phys.* **157**:144103

# Binning-based adaptivity tests

Conditional calibration implemented through LZV / LZISD analysis wrt $X$



# Ranking-based tests

**The correlation coefficient between $|E|$ and $u_E$ is often reported**

- it is independent on the scales of $E$ and $u_E$ and does not inform us on calibration
- because of the probabilistic link between $|E|$ and $u_E$, one should not expect a strong correlation ($> 0.5$).
  - what is a good value ???
- large errors should derive from large uncertainties, but small errors might come from small uncertainties as well as from large uncertainties

*Correlation/rank tests are mostly useless for variance-based UQ metrics*

# Confidence curves

**How does an error statistic (MAE, RMSE...) change when one removes the errors associated with the largest uncertainties ?**

One estimates

$$c_S(k; E, u_E) = S\left(E \mid u_E < u_k\right)$$

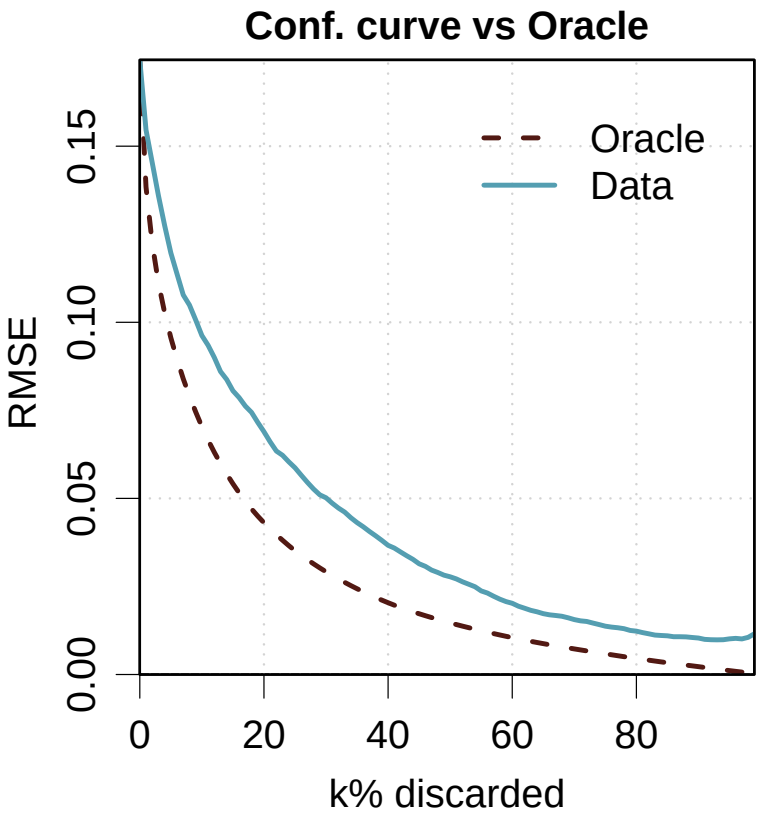
where

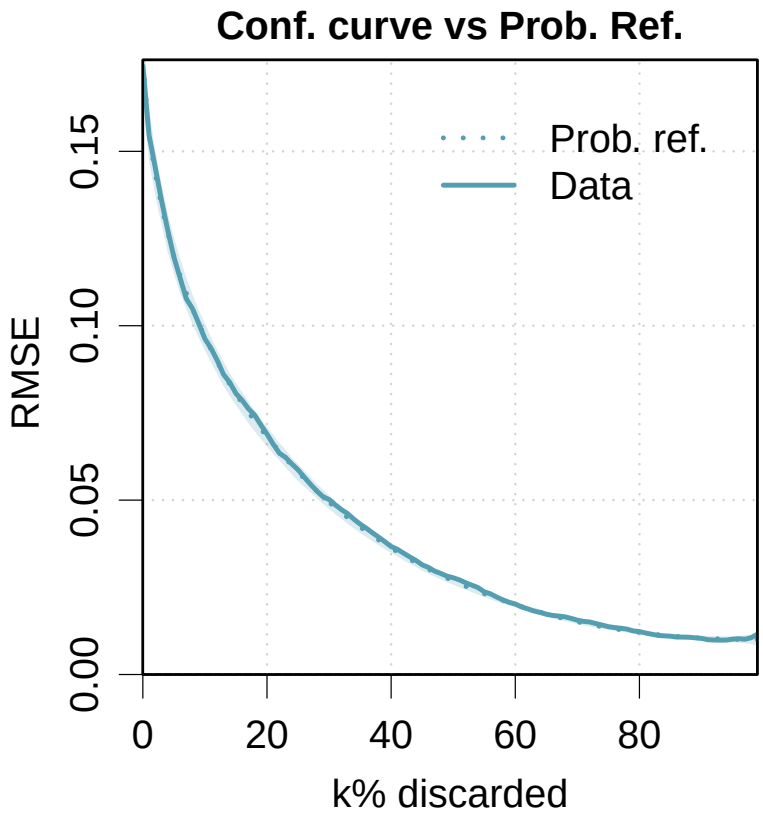
- $S$ is an error statistic (RMSE, MAE...)
- $u_k$ is the largest uncertainty after removal the $k$ % largest uncertainties from $u_E$ ($k \in \{0, 1, \ldots, 99\}$)

A confidence curve is obtained by plotting $c_S(k)$ vs $k$

- a monotonically decreasing confidence curve indicates a good association between large errors and large uncertainties.
  **It is a good validation test for active learning**

# Confidence curve references




- for a consistent dataset as the one treated here, one sees that the *oracle* is of no help for validation[6]

[6] Pernot (2022) arXiv:2206.15272

# Main ML-UQ approaches

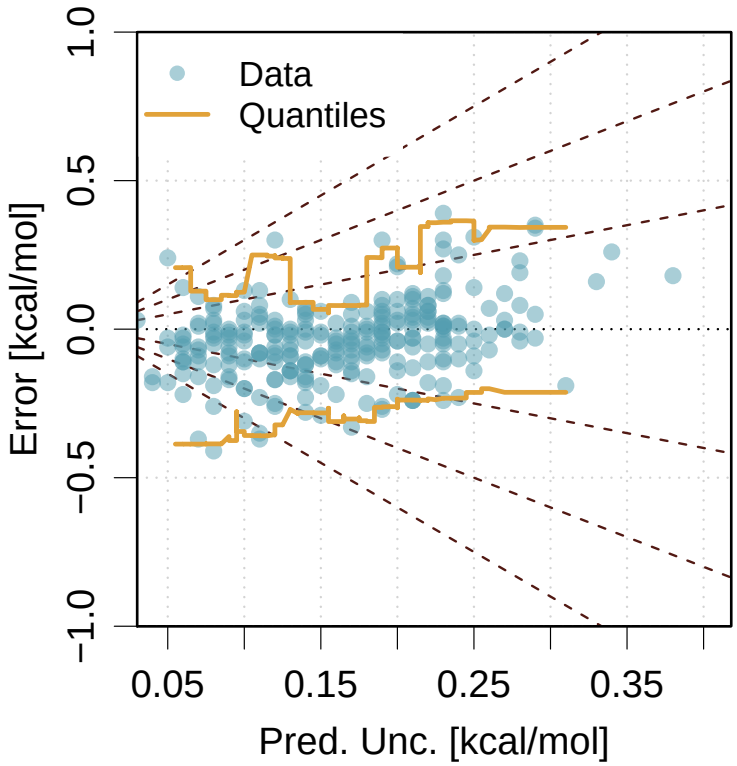
- **Direct methods**[7]
  - Intrinsic methods
    - Gaussian processes, Random Forests, Ridge regression
    - Bayesian neural networks, Evidential deep learning[8]
  - Ensemble methods
    - Dropout, Query by Commitee, Bootstrap...
- **A posteriori / post-hoc methods**
  - *Temperature scaling*[9], *Isotonic regression*[10], *Conformal prediction*[11]...

[7] Tran *et al.* (2020) *Mach. Learn.: Sci. Technol.* **1**:025006
[8] Soleimany *et al.* (2021) *ACS Cent. Sci.* **7**:1356-1367
[9] Mortensen *et al.* (2005) *Phys. Rev. Lett.* **95**:216401
[10] Busk *et al.* (2022) *Mach. Learn.: Sci. Technol.* **3**:015012
[11] Hu *et al.* (2022) *Mach. Learn.: Sci. Technol.* **3**:045028

# Formation heats by the mBEEF method[12]

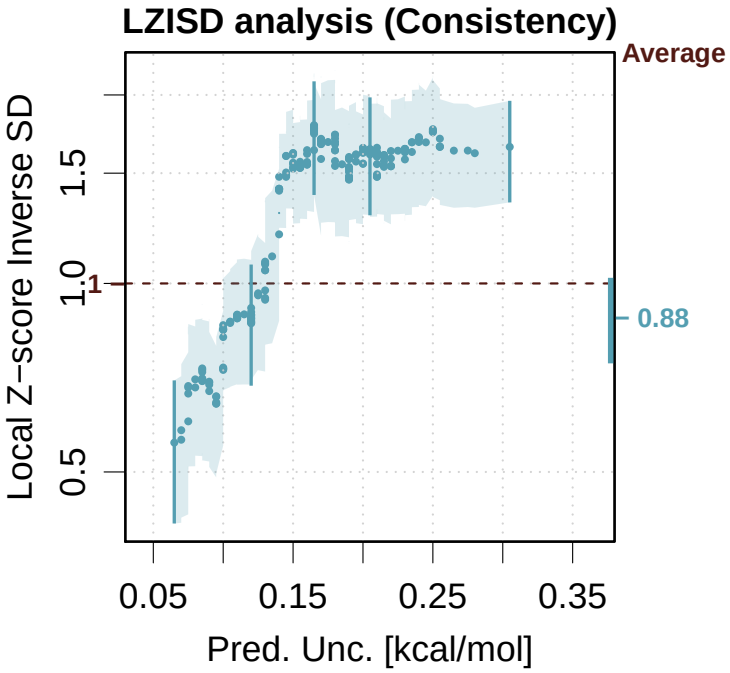
**Bayesian Ensembles** method inflates *parametric uncertainty* of exchange-correlation model to ensure *average calibration*

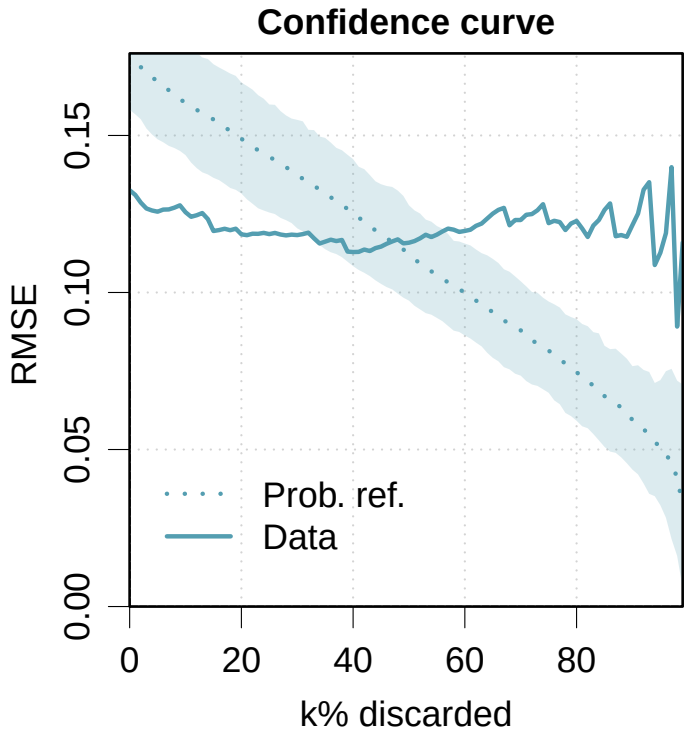
- strong functional constraints: consistency & adaptivity ???
- does not disambiguate model uncertainty from reference data uncertainty
- Set of $M = 257$ $\{V_i, R_i, u_{V_i}\}$




[12] 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

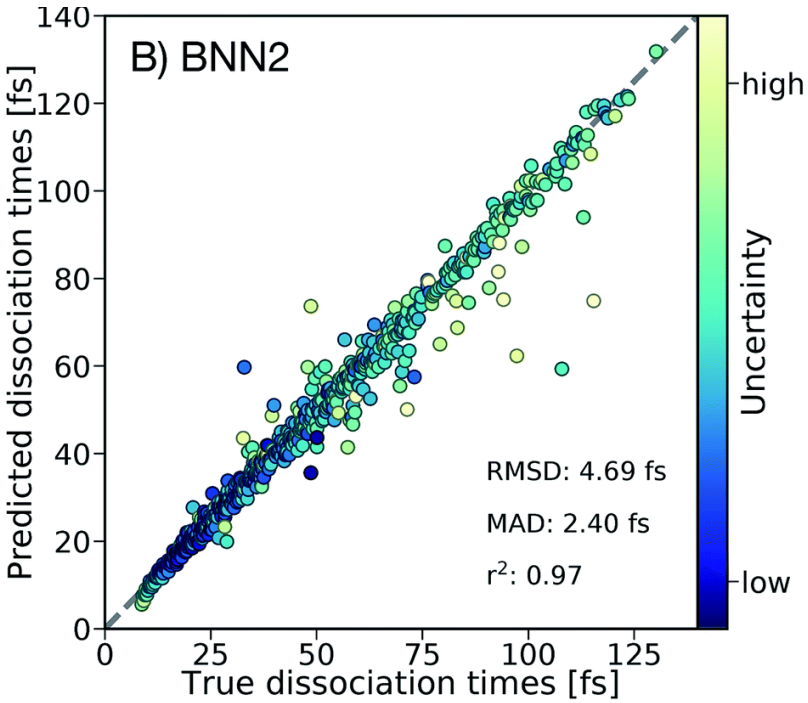


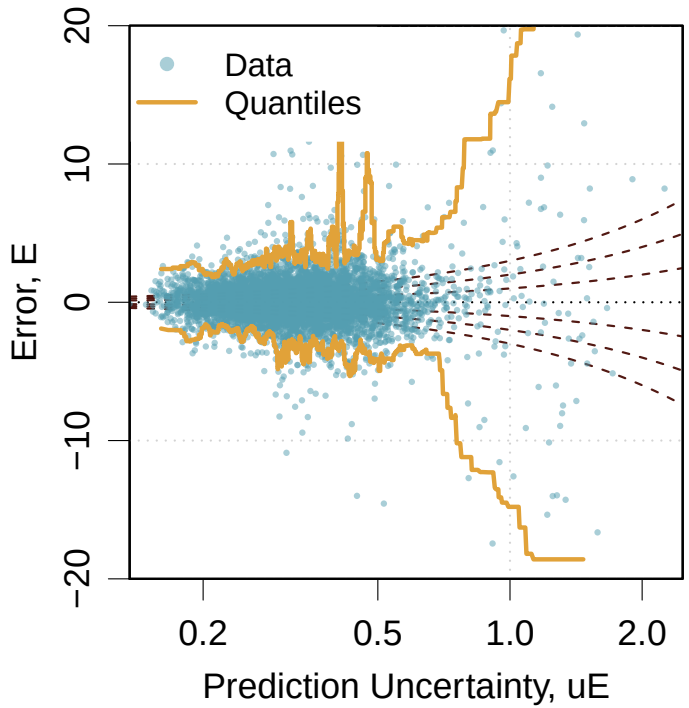




- $\mathrm{Var}(Z) = 1.3(2)$, *average calibration* OK
- the LZISD analysis shows that small PUs are underestimated by a factor up to 2, while large ones are overestimated by up to 60 %
- the confidence curve is not monotonously decreasing

# Bayesian Neural Network

## Data issued from a BNN trained to predict a MD potential[13] ($M = 5923$)

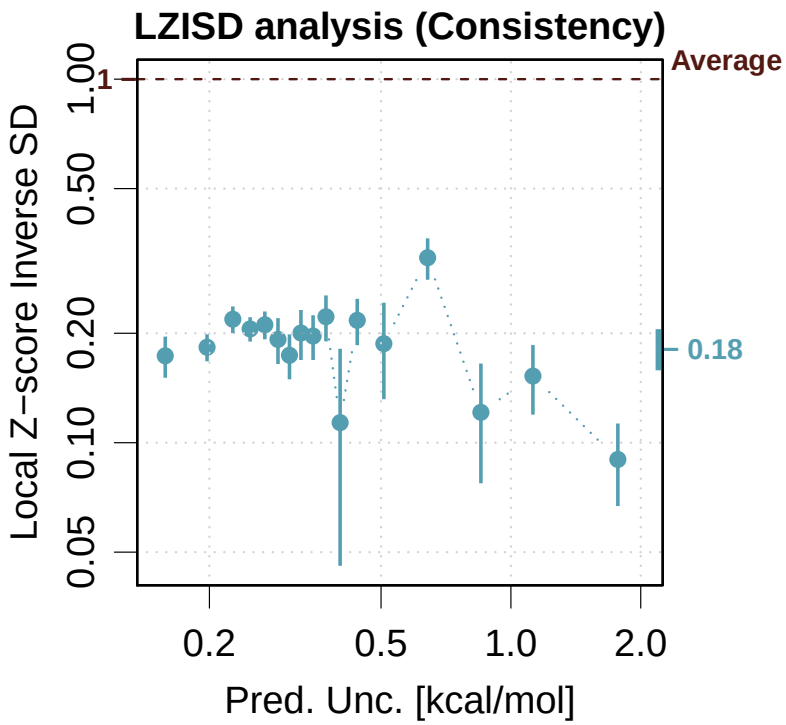


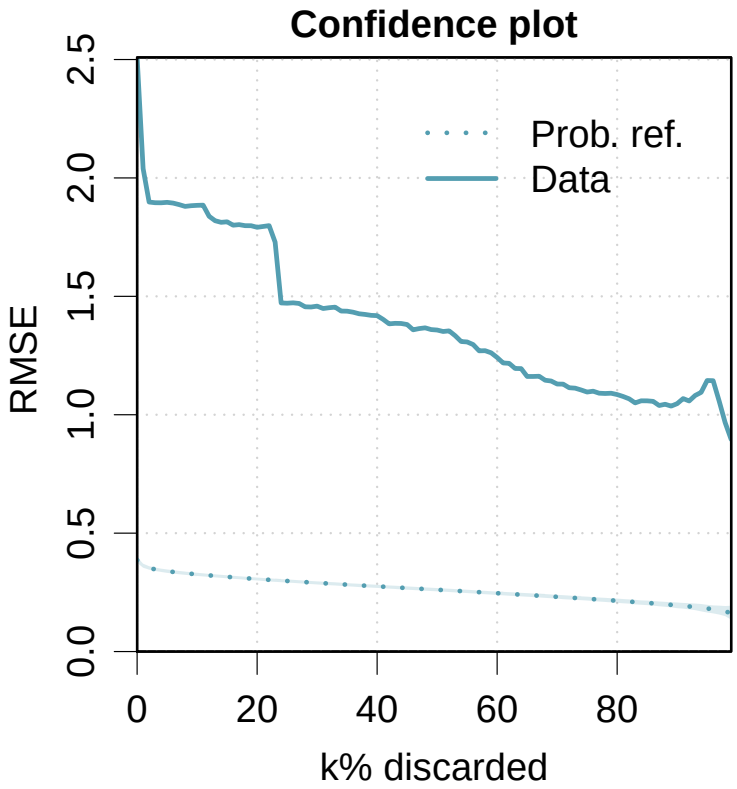




- The color scale for uncertainty is not a proper tool for validation

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

# Bayesian Neural Network

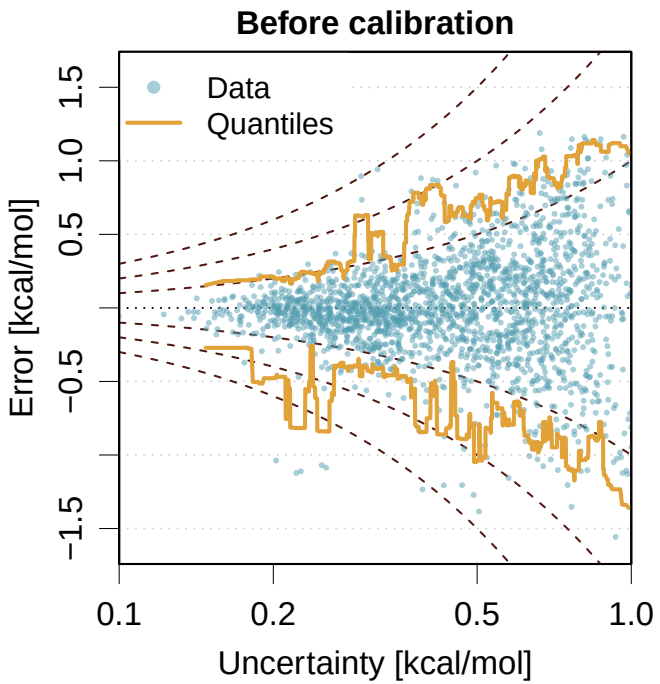


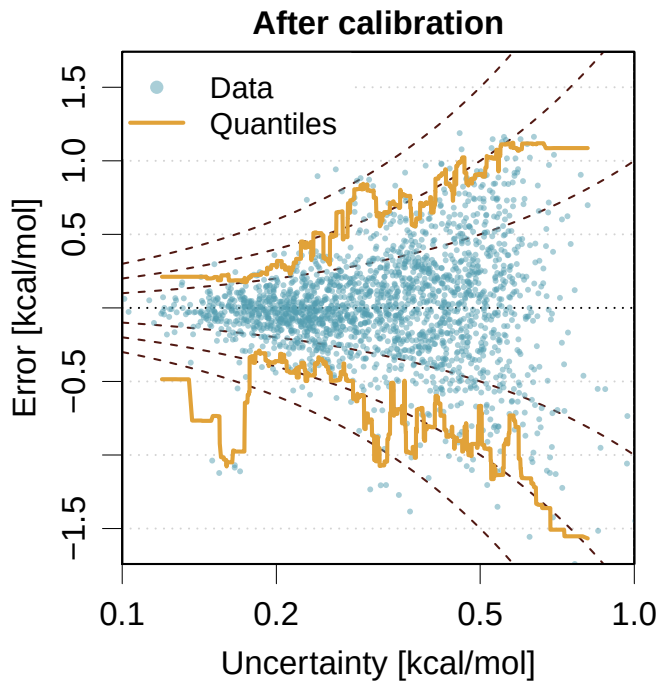




- This BNN uncertainty is NOT calibrated ($Var(Z) = 30$) but might still be used for active learning...

# Calibrated bootstrap for impurities diffusion

**Data issued from a study on a method to obtain calibrated ML uncertainties**[14] ($M = 2040$)

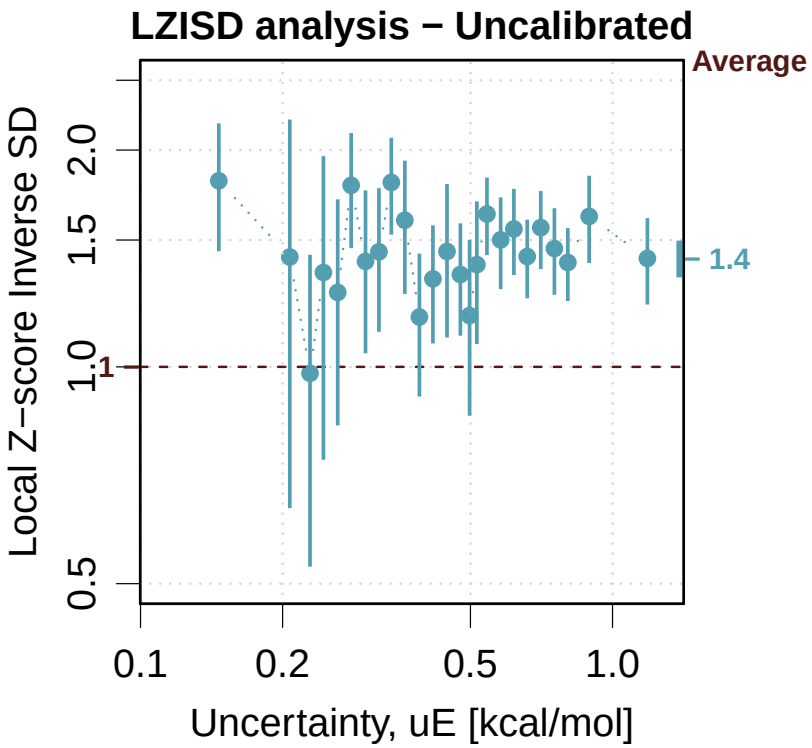


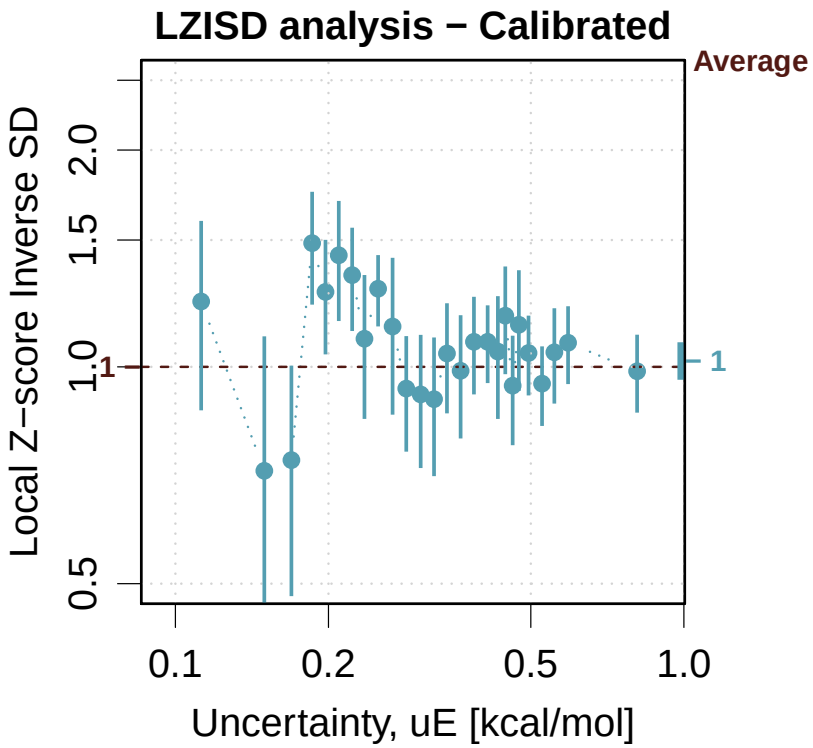




- Calibration seems efficient...

[14] Palmer *et al.* (2022) *npj Comput. Mater.* **8**:1-9; post-hoc calibration by linear transformation of uncal. uncertainties

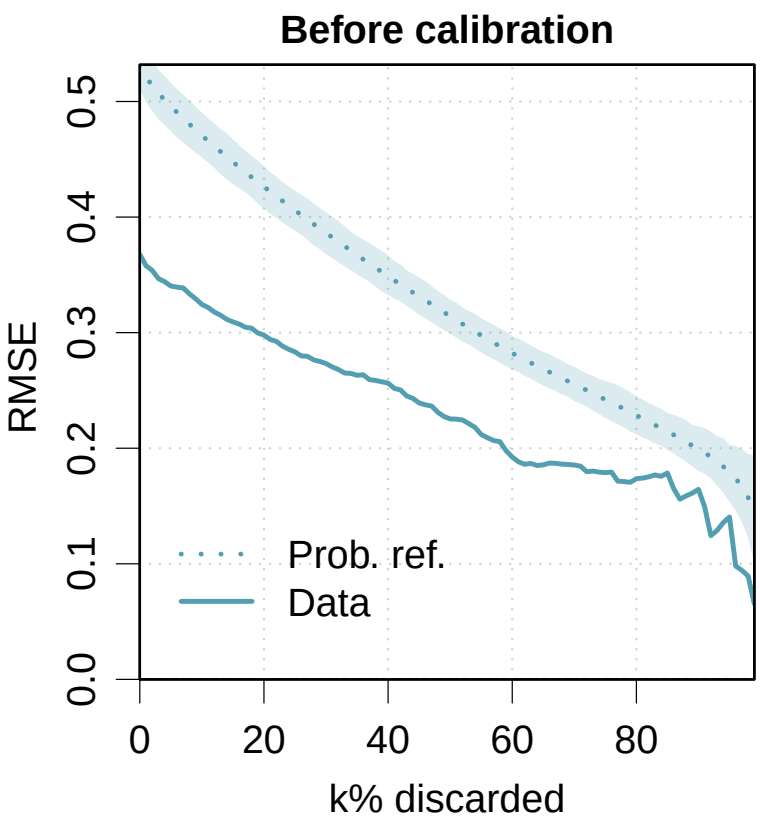
# Calibrated bootstrap for impurities diffusion

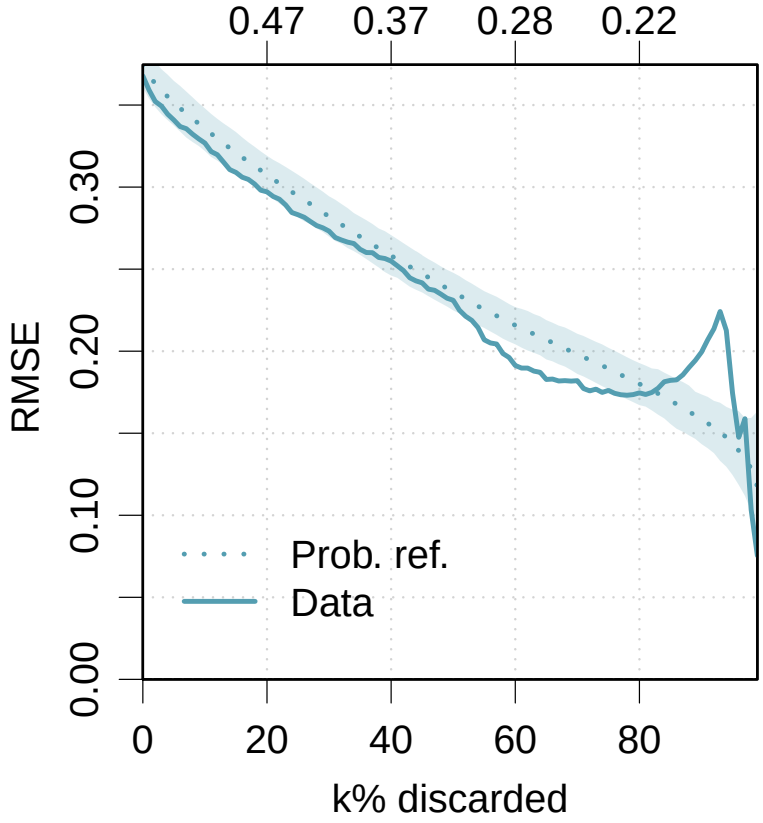



- Average calibration is excellent, but consistency of small uncertainties is not perfect (up to 50% over-estimation around $u_E = 0.2$ kcal/mol)

# Calibrated bootstrap for impurities diffusion

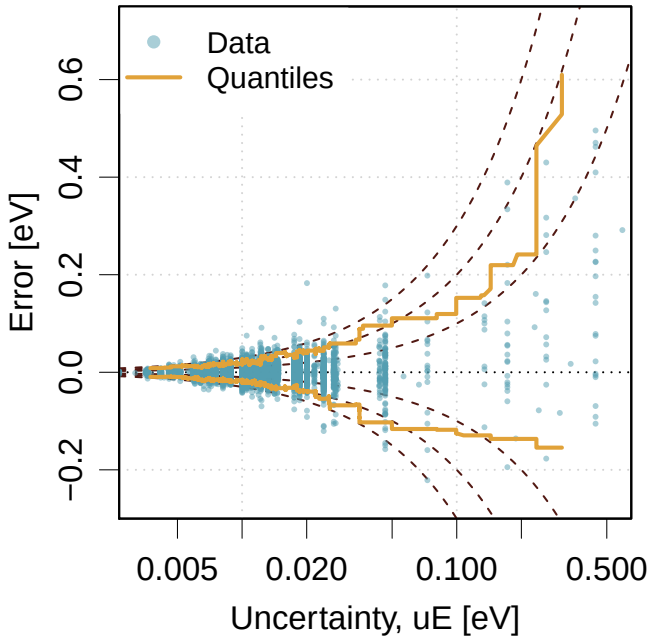


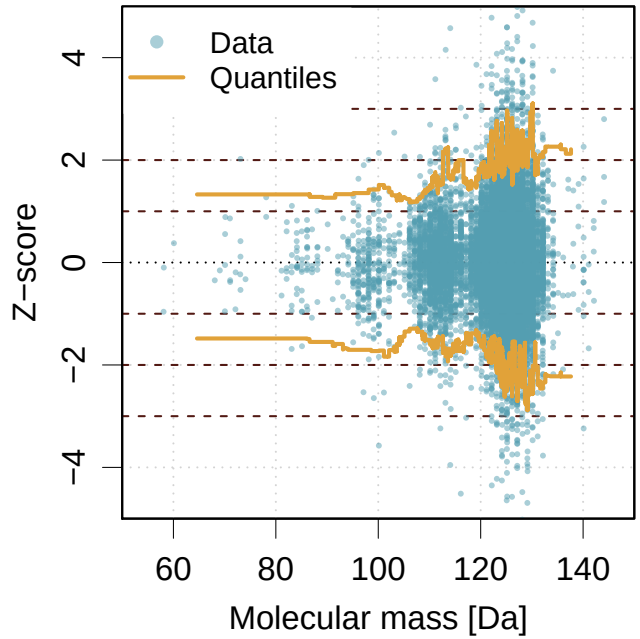




- Efficient calibration but no consistency for the smaller 50% of uncertainties; OK for Active Learning

# Post-hoc calibration of ensemble predictions

## Atomization energy on QM9 dataset[15] ($M = 13885$)

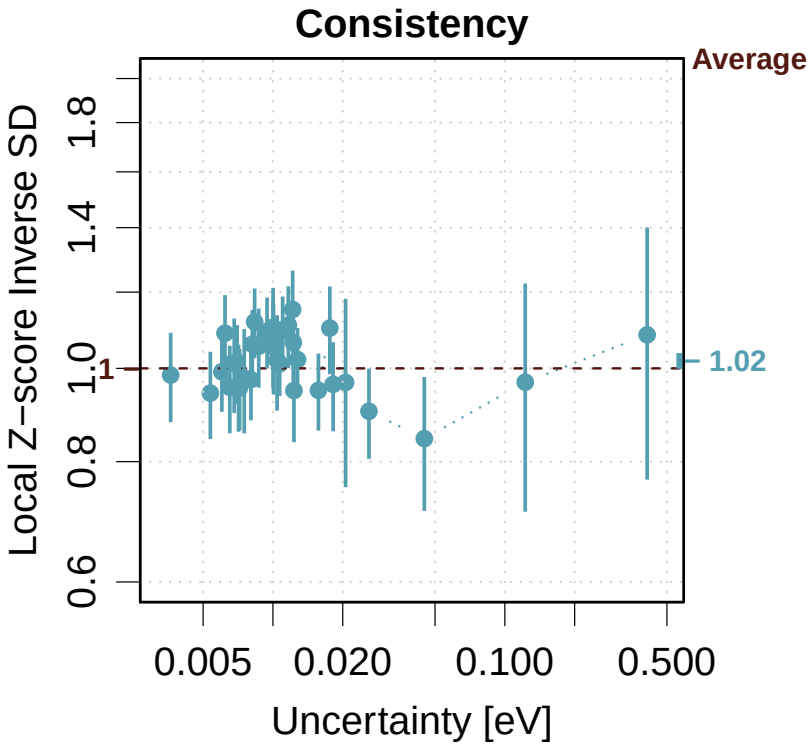


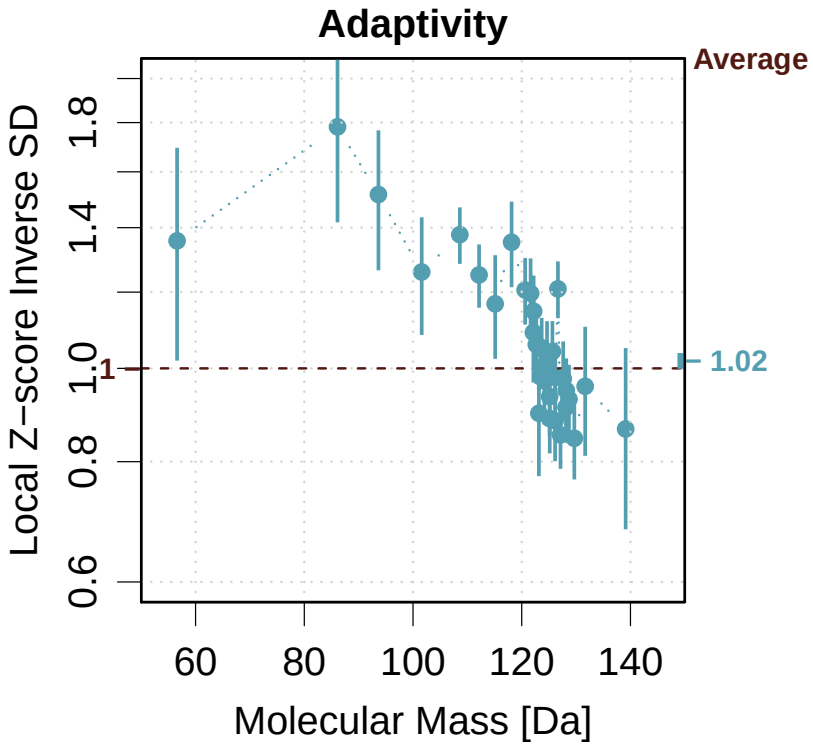




- Consistency seems OK
- Adaptivity seems problematic on "Z vs X" plot

[15] Busk *et al.* (2022) *Mach. Learn.: Sci. Technol.* **3**:015012; post-hoc calibration by non-linear transformation of uncal. uncertainties (isotonic regression)

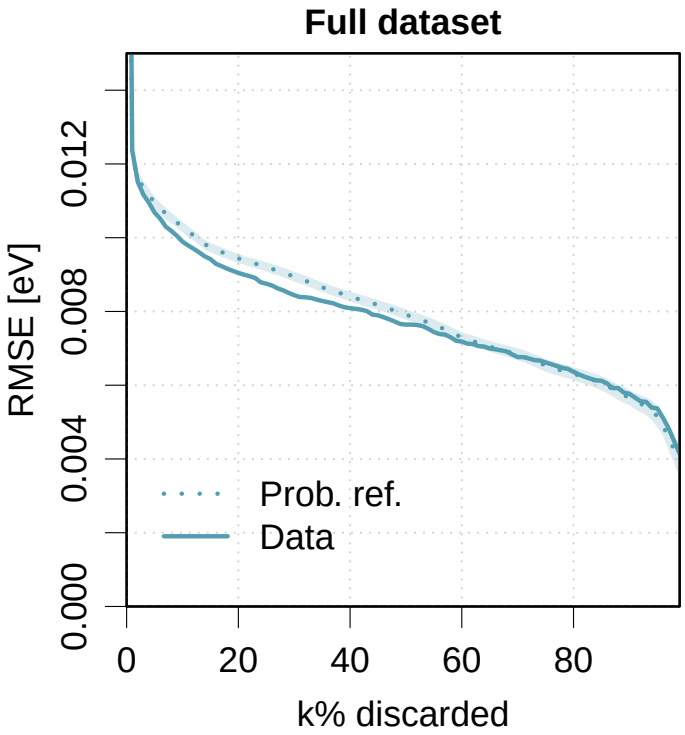
# Consistency and adaptivity

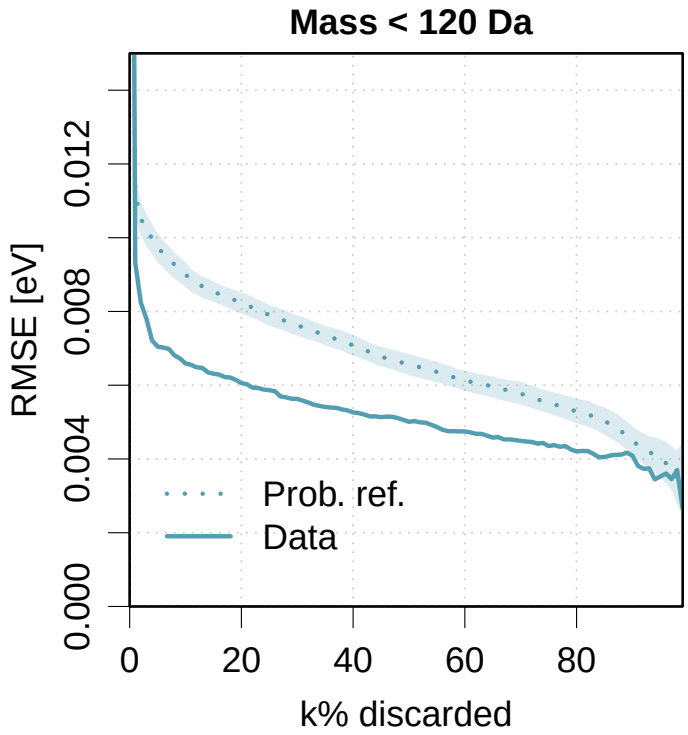



- No major consistency default
- But notable problem of adaptivity, with systematic deviations from the $y = 1$ line. Confirms the diagnostic of the "Z vs X" plot.

# Confidence curves




- The confidence curve tests the *consistency* of $E$ and $uE$
- Consistency might not be fulfilled locally in $X$ space

# Conclusions

- **Calibration/Consistency/Adaptivity**
  A principled framework for UQ validation
  - Calibration is easy, consistency and adaptivity are tough !
  - Adaptivity is presently a (dangerous) blind spot in ML-UQ validation studies.
- **Direct ML-UQ methods do not provide calibrated uncertainty**
  - might still be good for internal use (active learning)
  - strong need for post-hoc calibration methods going beyond average calibration
- **UQ methods used today in computational chemistry rarely reach consistency or adaptivity**
  - how-much mis-calibration is acceptable for a given application ?

# Warmful thanks to...

- **Andreas "Gauss Slayer" SAVIN** (LCT, Jussieu)
  for so many enlightening discussions

- **Morgane VACHER** (Nantes Université)
  **Jonas BUSK** (Technical University of Denmark)
  and many others for providing me with invaluable datasets

- and **YOU**, for your attention !

# Confidence curve references

- The **Oracle** is the confidence curve obtained by assuming a perfect correlation between $|E|$ and $u_E$

$$O(k) = c_S(k; , E, |E|)$$

  - it is unsuitable for variance-based UQ metrics and corresponds to an unrealistic generative model: $E \sim \pm u_E$

- A **Probabilistic** reference can be built instead

$$P(k; u_E) = \left\langle c_S(k; \tilde{E}, u_E) \right\rangle_{\tilde{E}}$$

  where a Monte Carlo average is taken over samples of

$$\widetilde{E}_i \sim D(0, u_{E_i})$$

  - one can thus test the consistency of $E$ and $u_E$

# Consistency tests

Conditional calibration is implementred through binning wrt $u_E$ (*local* calibration)

- **Reliability diagrams** or **RMSE vs RMV plot**[16]
  1. split $u_E$ into bins
  2. estimate $Var(E)$ and $< u_E^2 >$ for each bin
  3. plot $\sqrt{Var(E)}$ vs $\sqrt{< u_E^2 >}$
  4. check for *deviations from the identity line*

- **Local Z-Variance** (LZV) or **Local Z-Inverse SD** (LZISD) plots[17]
  1. split $u_E$ into bins
  2. estimate $Var(Z = E/u_E)$ for each bin
  3. plot $Var(Z)$ or $1/\sqrt{Var(Z)}$ at the center of each bin
  4. check for *deviations from the* $y = 1$ *line*

- **Note**: the diagnostic might depend on the binning strategy.
  Bins should be as small as possible without compromising testing power...

[16] Levi *et al.* (2020) arXiv:1905.11659

[17] Pernot (2022) *J. Chem. Phys.* **157**:144103