Negative impact of heavy-tailed uncertainty and error distributions…

Pascal PERNOT (Groupe ThéoSim, ICP)

2024-06-04

... on the reliability of calibration statistics for machine learning
regression tasks

??? What am I doing here ???

- **Prediction** (ML or other models)
  - **UQ** (*quantify* prediction uncertainty)
    - **UQ validation** (test uncertainty *calibration*)
      - **Test the *reliability* of calibration statistics**

> Our confidence in the UQ approach depends on it !

How to validate prediction uncertainty ?

UQ metrics and validation data

UQ validation methods depend on UQ information[1]

- full distribution (probability-based UQ metrics)
- prediction intervals (interval-based UQ metrics)
- **uncertainty (variance-based UQ metrics)**
  - **Validation dataset**

$$\{X_i, (V_i, u_{V_i}), (C_i, u_{C_i})\}_{i=1}^{M}$$

$$\longrightarrow \left\{X_i, E_i = C_i - V_i, u_{E_i} = \sqrt{u_{C_i}^2 + u_{V_i}^2}\right\}_{i=1}^{M}$$

1. Pernot (2022) *J. Chem. Phys.* **156**:114109; *J. Chem. Phys.* **157**:144103.

Calibration: the generative model

- **Prediction uncertainty quantifies the dispersion of errors**[2]
  $\longrightarrow$ calibration is based on a *probabilistic* model

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

- **Errors have a *compound distribution***

$$p_H(E) = \int_0^\infty p_D(E|u_E)\, p_G(u_E)\, du_E$$

- **Example**: the Normal-Inverse-Gamma (*NIG*) model

$$\text{if } u_E^2 \sim \Gamma^{-1}(\nu/2, \nu/2) \text{ and } D = N(0,1),$$
$$\text{then } E \sim t(\nu) \text{ and } E^2 \sim F(1, \nu)$$

2. BIPM *et al.* (2008) *Evaluation of measurement data - Guide to the expression of uncertainty in measurement (GUM)*, JCGM.

Derived calibration equations

- Law of total variance

$$\begin{aligned}\mathrm{Var}_H(E) &= \langle \mathrm{Var}_D(E|u_E)\rangle_G + \mathrm{Var}_G\left(\langle E|u_E\rangle_D\right)\\ &=< u_E^2 > + \cancel{\mathrm{Var}_G\left(\langle E|u_E\rangle_D\right)}\end{aligned}$$

$$\text{and}$$

$$\begin{aligned}\mathrm{Var}(E) &=< E^2 > - \cancel{< E >^2}\\ \implies &< E^2 >=< u_E^2 >\end{aligned}$$

- Scaled errors (z-scores)

$$\begin{aligned}Z_i &= \frac{E_i}{u_{E_i}} \sim D(0,1)\\ \implies &< Z^2 >= 1\end{aligned}$$

Average Calibration Statistics

- **Relative Calibration Error** (target = 0)

$$RCE = (RMV - RMSE)/RMV$$

where $RMV = \sqrt{< u_E^2 >}$ and $RMSE = \sqrt{< E^2 >}$

- **Z-scores Mean Squares** (target = 1)

$$ZMS = < (E/uE)^2 >$$

> Note that RCE ignores the (E_i, u_{E_i}) pairing, so that **it should be more forgiving** than ZMS

Conditional/Local Calibration

- Average calibration does not guarantee the calibration of individual predictions.
- *Conditional calibration* is estimated by splitting data into N bins and testing calibration for each bin
  - RCE : **Reliability diagram**[3] $< E^2 >_{B_i} = < u_E^2 >_{B_i}$
  - ZMS: **Local ZMS analysis**[4] $< Z^2 >_{B_i} = 1$
- Choice of binning variable:
  - Binning vs u_E : *consistency*
  - Binning vs X : *adaptivity*

3. Levi *et al.* (2022) *Sensors* **22**:5540
4. Pernot (2023) *APL Machine Learning* **1**:046121

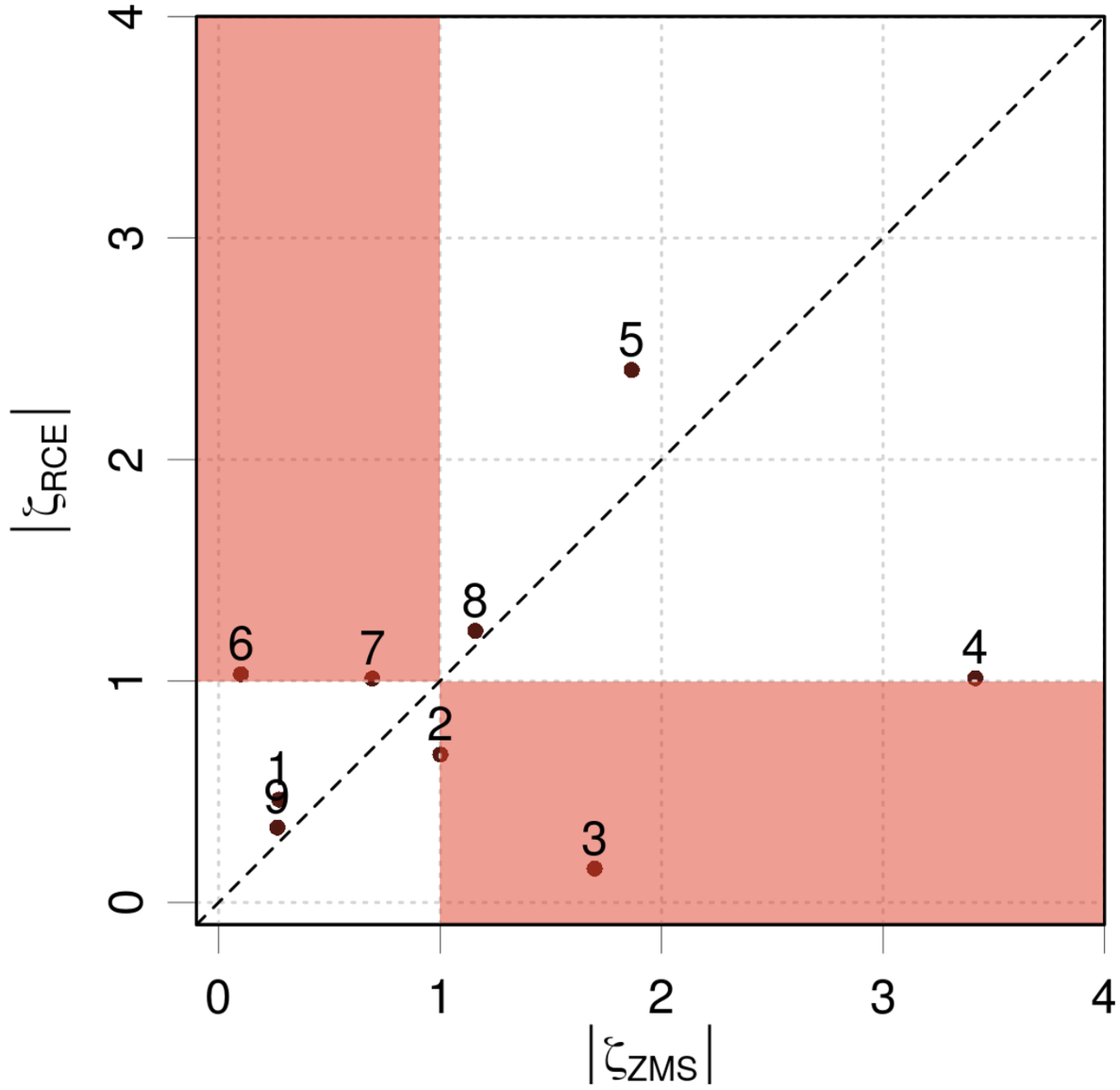
Application to “calibrated” datasets

The datasets[5]

Set #	Name	Size (M)	Reference
1	Diffusion_RF	2040	Palmer et al. (2022)
2	Perovskite_RF	3834	Palmer et al. (2022)
3	Diffusion_LR	2040	Palmer et al. (2022)
4	Perovskite_LR	3836	Palmer et al. (2022)
5	Diffusion_GPR_Bayesian	2040	Palmer et al. (2022)
6	Perovskite_GPR_Bayesian	3818	Palmer et al. (2022)
7	QM9_E (IR)	13885	Busk et al. (2022)
8	logP_10k_a_LS-GCN	5000	Rasmussen et al. (2023)
9	logP_150k_LS-GCN	5000	Rasmussen et al. (2023)

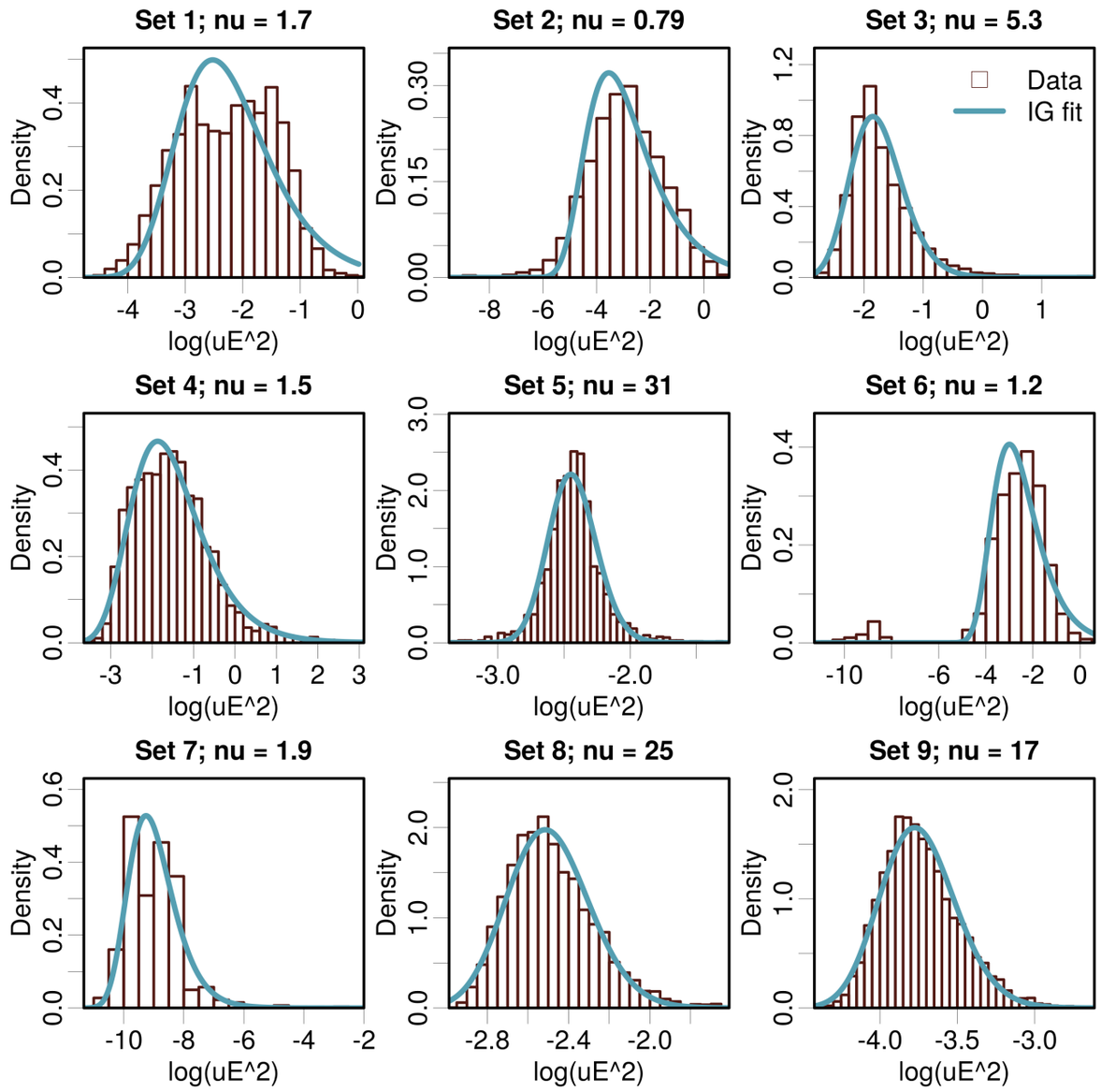
5. Palmer *et al.* (2022) *npj Comput. Mater.* **8**:115; Busk *et al.* (2022) *Mach. Learn.: Sci. Technol.* **3**:015012; Rasmussen *et al.* (2023) *J. Cheminf.* **15**:1-17.

ZMS vs RCE - the problem ![6]

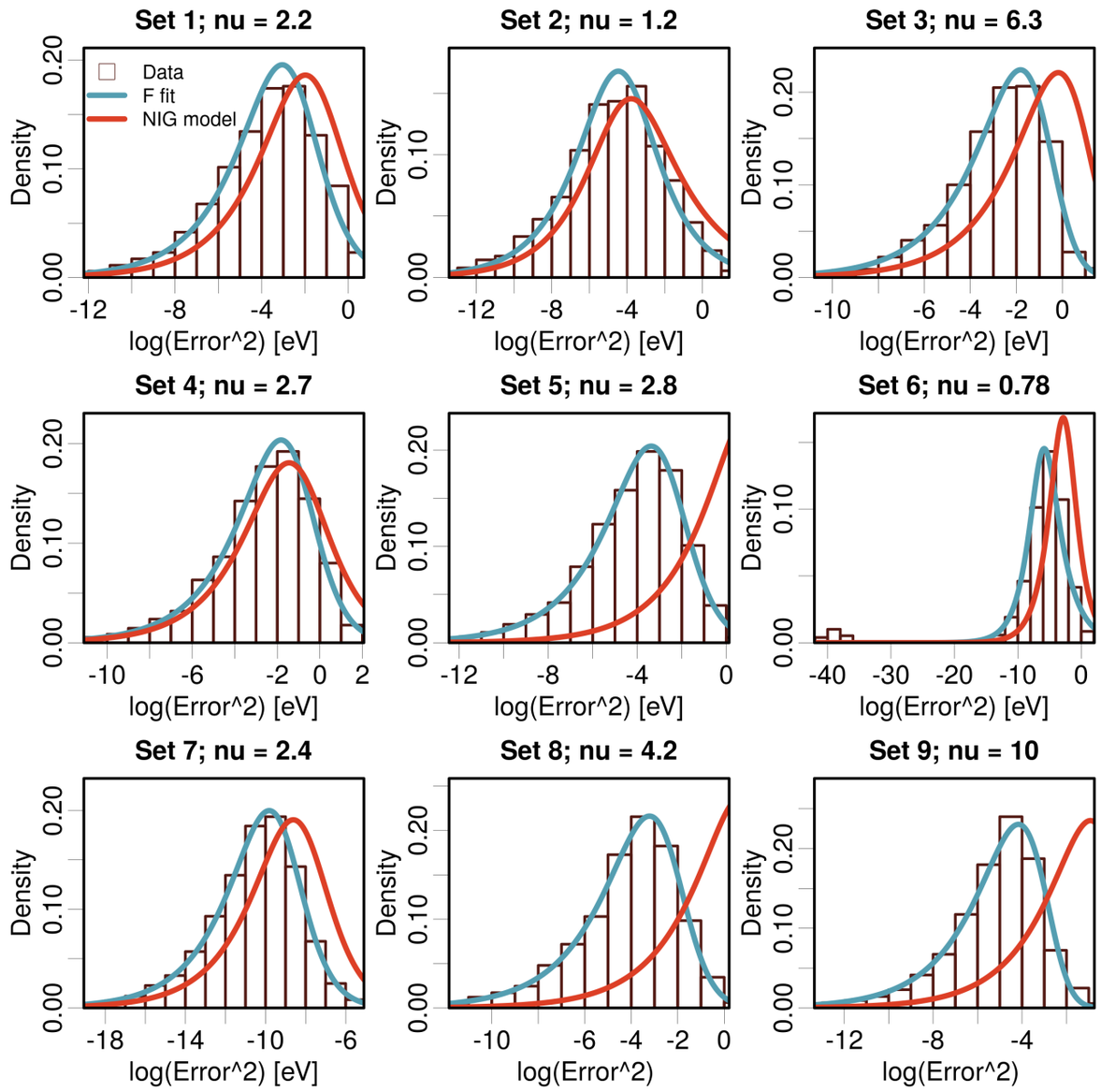


6. $\zeta_\vartheta = (\vartheta - \vartheta_{ref})/U_{\vartheta,95}$; validation by $abs(\zeta_\vartheta) \leq 1$

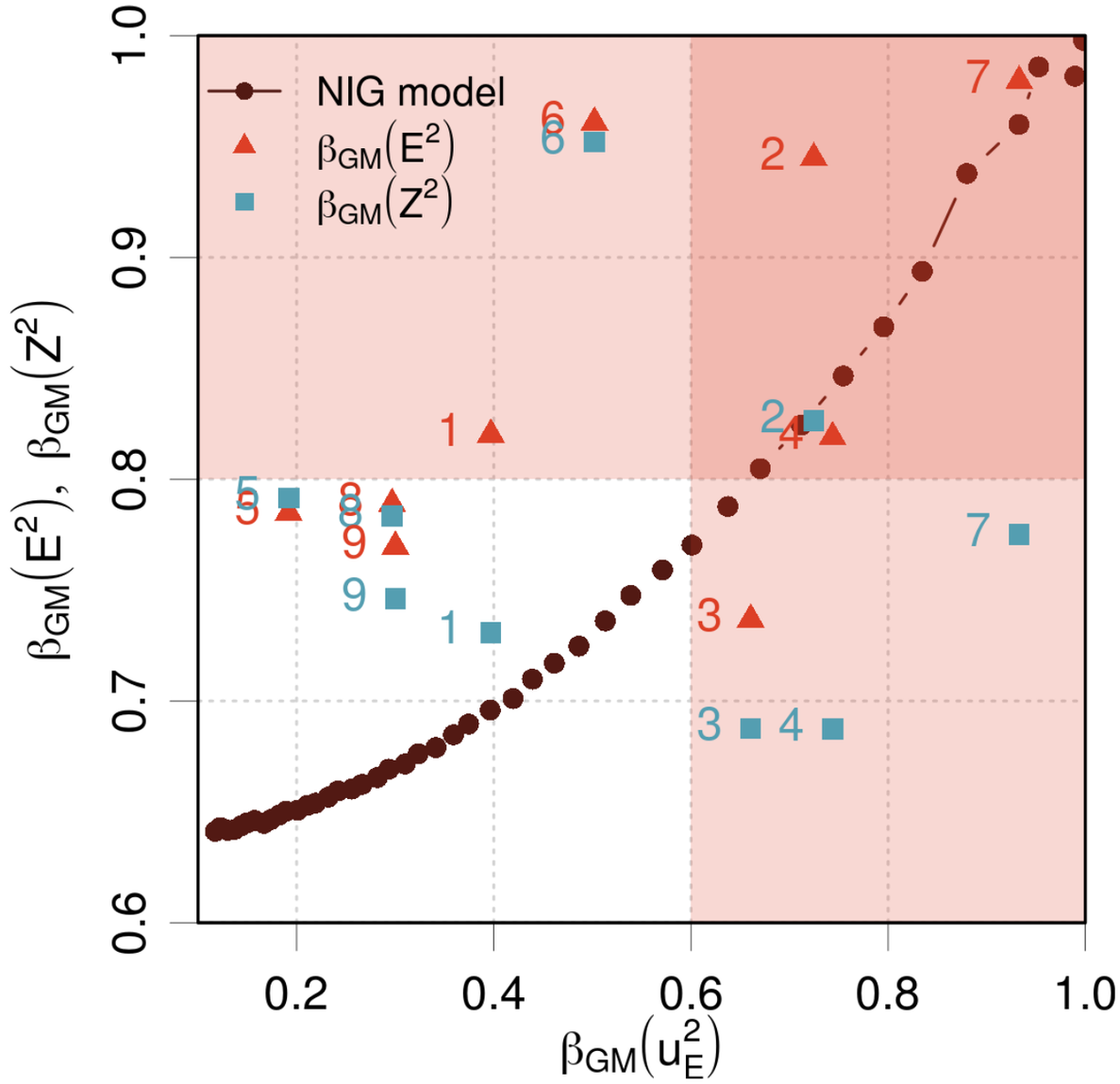
u_E^2 distributions: $\Gamma^{-1}(\nu, \nu)$ fit



E^2 distributions: $F(a, b)$ fit vs. $NIG(\nu)$

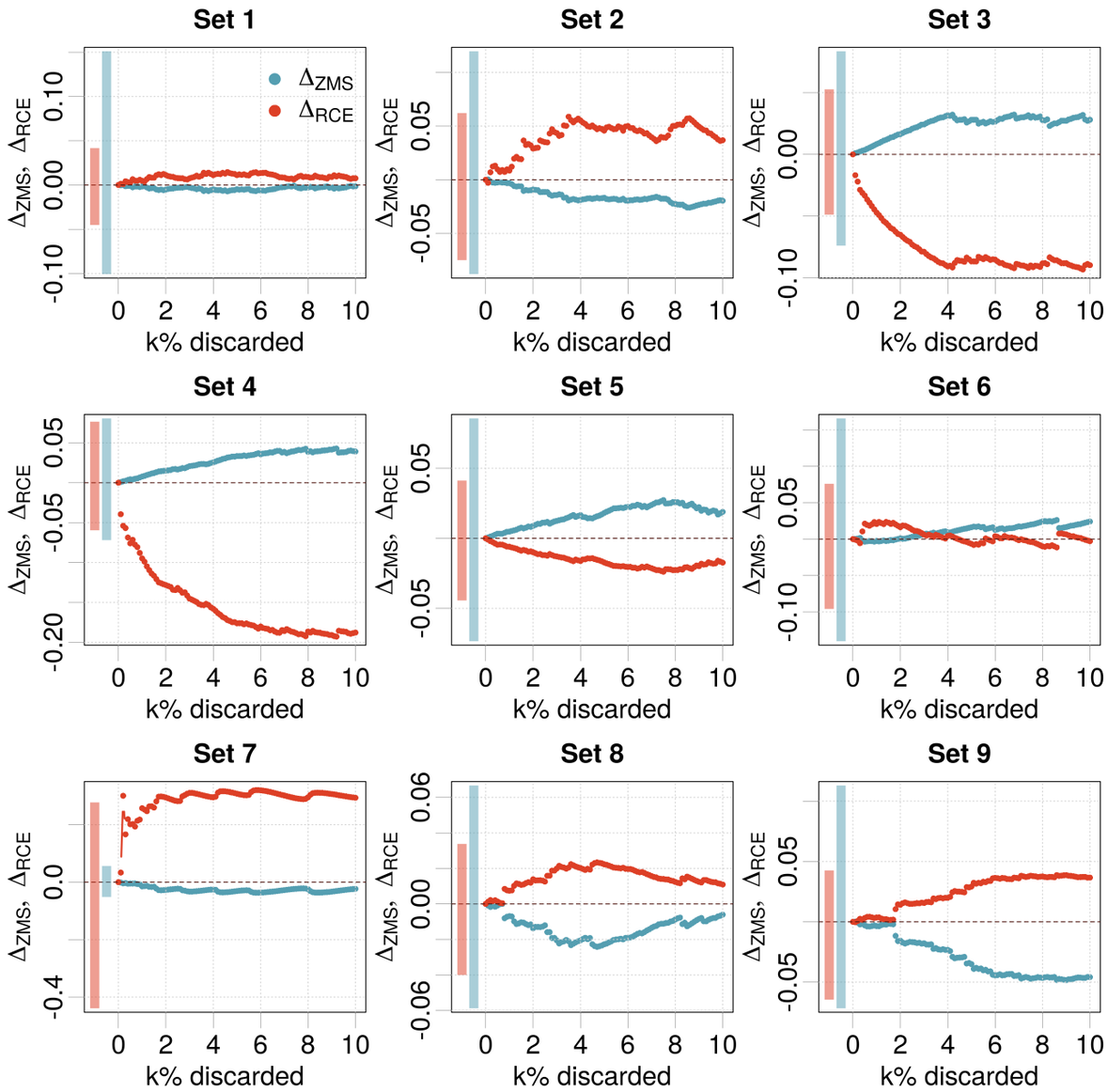


Skewness of u_E^2 and E^2 distributions[7]



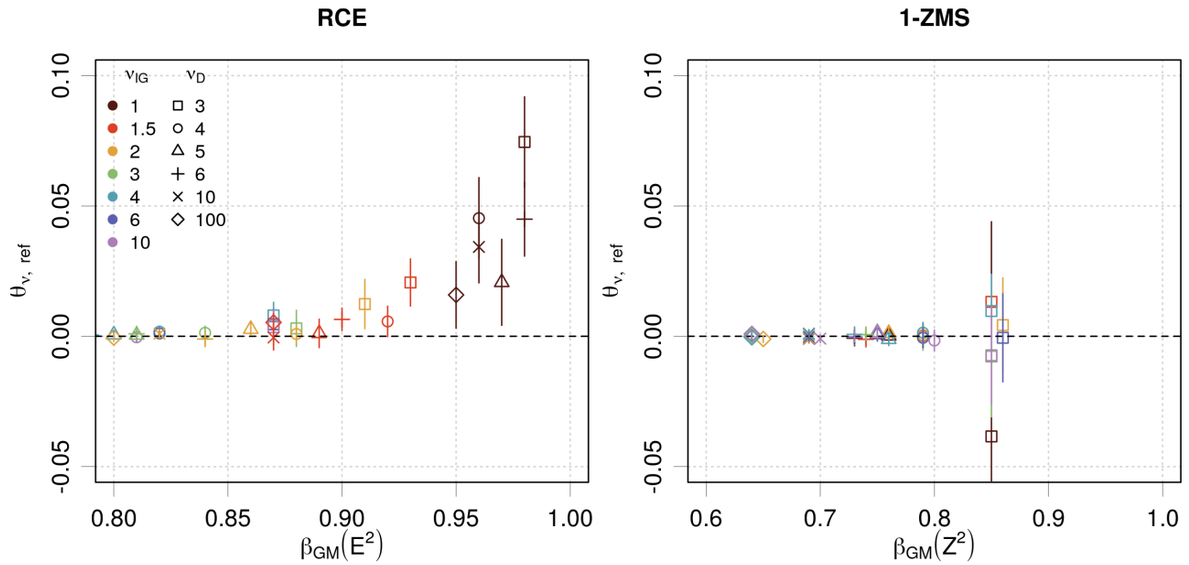
7. β_{GM} is a robust skewness metric (Groeneveld and Meeden (1984), *The Statistician* **33**:391; Pernot and Savin (2021) *Theor. Chem. Acc.* **140**:24).

ZMS vs RCE - Sensitivity to u_E by decimation



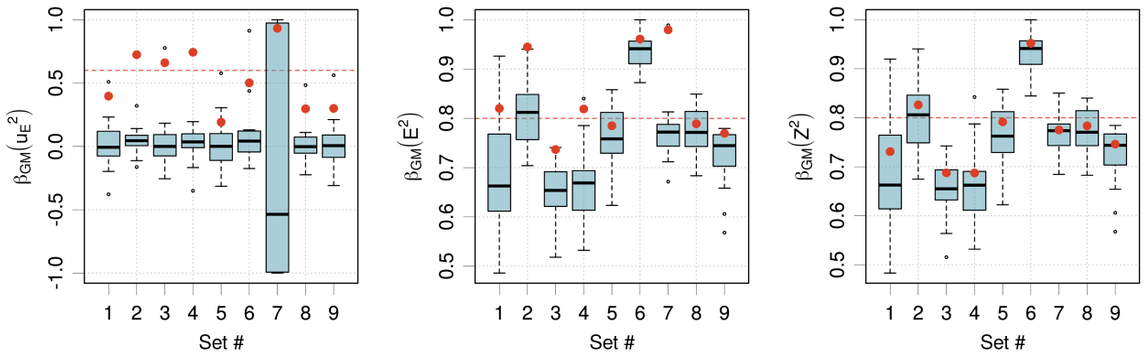
ZMS vs RCE - Sensitivity to E

TIG model: $u_E^2 \sim \Gamma^{-1}(\nu_{IG}, \nu_{IG});\ D = t_u(\nu_D)$

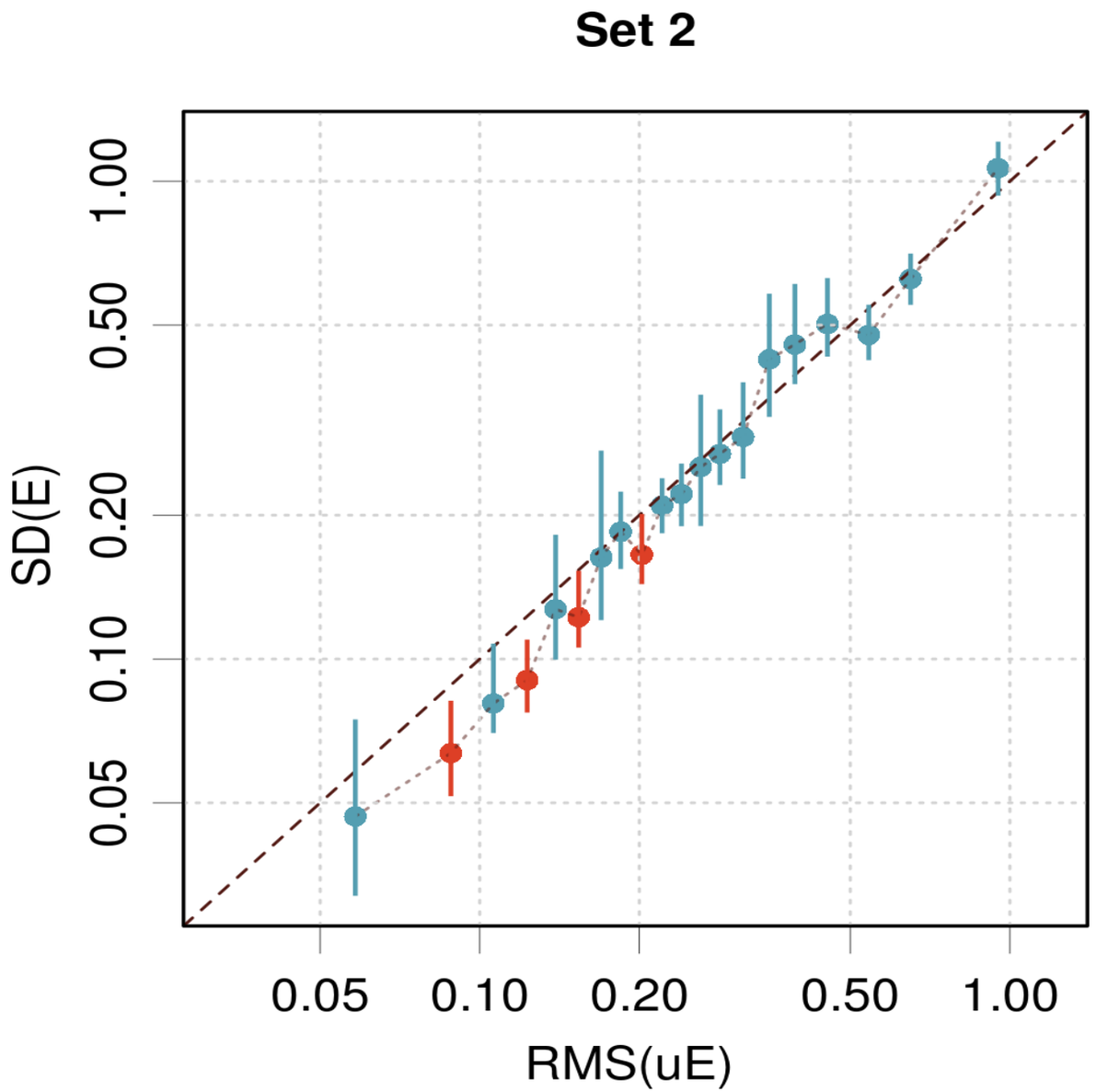


Does binning improve the situation ?

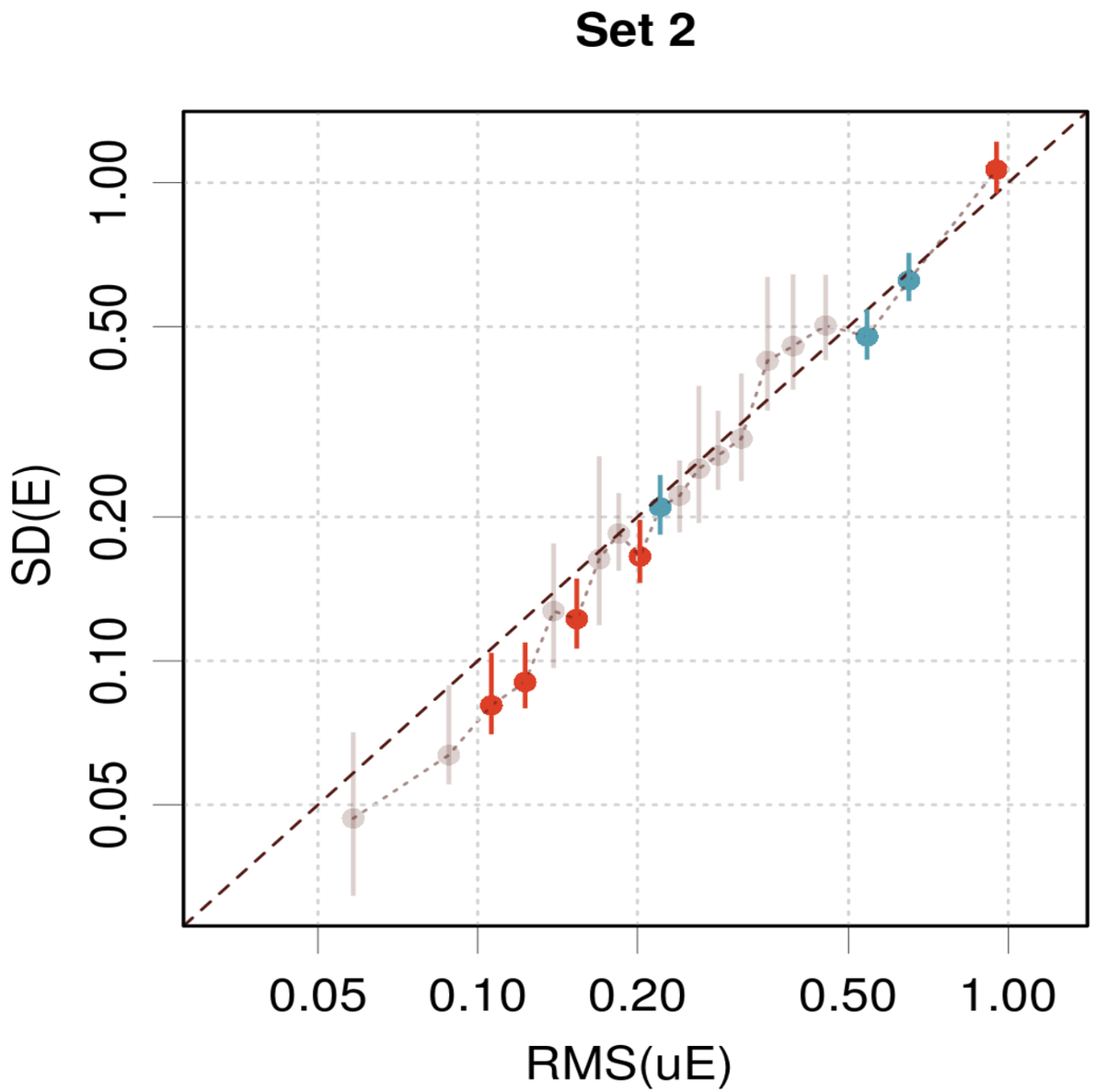
- 20 equal-size bins along u_E
- estimate β_{GM} for each binned variable



Reliability diagram without β_{GM} filtering



Reliability diagram with β_{GM} filtering



Conclusions

- Uncertainty and error distributions often have heavy tails
  - $< X^2 >$ is *not robust to outliers* (cf. RMSD vs. MAE)
    - the RCE statistic is sensitive to $< u_E^2 >$ and/or $< E^2 >$; $< Z^2 >$ is more reliable
  - estimation of *validation CIs* is also problematic
- Binning for conditional calibration statistics does not help
- Robust skewness statistic β_{GM} can be used to detect problematic cases
- Iterative training based on outlying errors and/or uncertainties might help to tame rebel distributions

Perspectives

- Validation of calibration statistics based on variance-based UQ metrics appears simple, but can be excessively tricky

> **It is probably safer to use intervals-based UQ metrics (e.g. using *conformal inference*)**

However,

- standard conformal inference ensures *average* calibration
- general methods for *adaptive* conformal inference are still in development.[8]

8. Angelopoulos & Bates (2023) *Found. Trends Mach. Learn.* **16**:494-591; Susmann *et al.* (2023) *arXiv*:2312.00448

Thanks

- This work benefited from earlier studies in collaboration with **Andreas Savin** on the reliability and improvement of performance metrics for the comparison of computational chemistry methods.[9]

> Thank you for your attention !

9. Pernot and Savin (2018) *J. Chem. Phys.* **148**:241707; (2020) *J. Chem. Phys.* **152**:164108; (2020) *J. Chem. Phys.* **152**:164109; Pernot *et al.* (2020) *Mach. Learn.: Sci. Technol.* **1**:035011; Pernot and Savin (2021) *Theor. Chem. Acc.* **140**:24.

Supplementary Material

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 and is central to **Active Learning**

- UQ is also a necessity when ML predictions replace physical experiments (**Virtual Measurements**)
    - comparisons, conformity testing

> UQ validation tests the reliability of prediction uncertainty

Many ML-UQ approaches…

- **Direct methods**[10]
  - Intrinsic methods
    - *Gaussian processes*, *Random Forests*, *Ridge regression*, *Bayesian neural networks*, *Evidential deep learning*[11]
  - Ensemble methods
    - *Dropout*, *Query by Commitee*, *Bootstrap*

- **A posteriori / post-hoc methods**
  - *Temperature scaling*[12], *Isotonic regression*[13], *Conformal prediction*[14]

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

Reliability of validation ζ-scores

- CIs are estimated by bootstrapping (BCa)