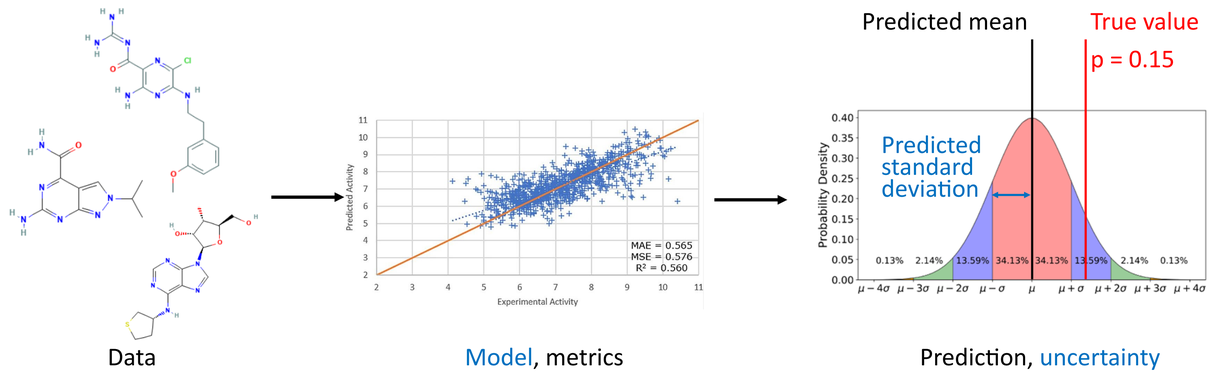
Evaluating and interpreting uncertainty in QSAR models

Charles Gong[1], Katarzyna R. Przybylak[2], Jonathan M. Goodman[1]

1. Centre for Molecular Informatics, Yusuf Hamied Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, UK CB2 1EW
2. Safety & Environmental Assurance Centre, Unilever, Colworth Science Park, Sharnbrook, Bedfordshire, UK MK44 1LQ

1. Introduction

Consumer and environmental safety decisions can be aided by QSAR models built on exposure and hazard data. We can validate models using performance metrics but this is inadequate for high-stakes decisions. Uncertainty quantification can help by allowing models to feedback to us when it is unsure of the given prediction, alerting users that more evidence is required [1]. However, estimates of uncertainty should not be blindly trusted — they need to be validated robustly like any other model.



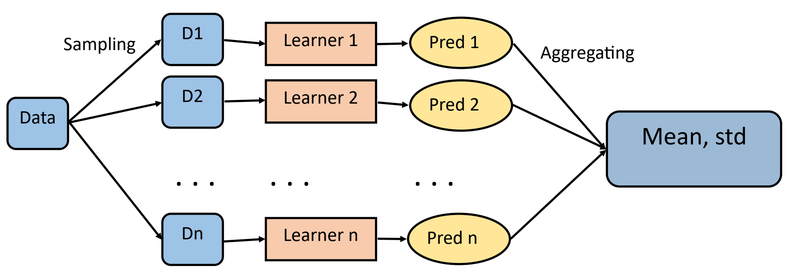
Data — Model, metrics — Prediction, uncertainty

2. Data and methods

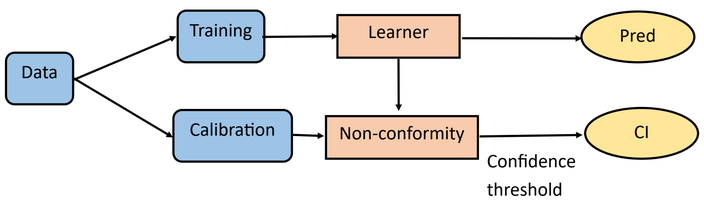
Data for 21 toxicologically relevant targets were obtained from **ChEMBL v23**. External validation data was obtained from **ChEMBL v25**, with training data removed. Further details may be found in Allen et al. [2]

Three uncertainty quantification methods were applied:

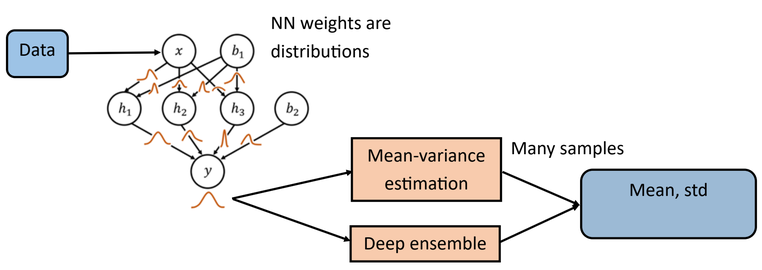
1. **Bayesian bootstrapping (BB)** as described by Rubin [3], and implemented in Python by Cialdella et al. [4]



2. **Conformal prediction (CP)** as described by Vovk et al. [5], and implemented in Python by Linusson et al. [6]



3. A **Bayesian neural network (BNN)** implemented by Allen et al. [2]



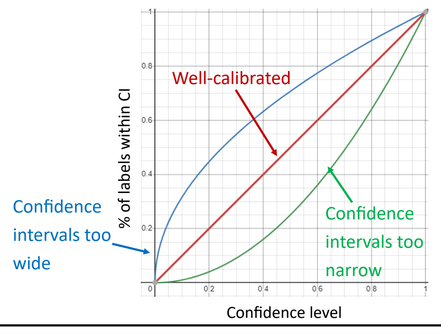
3. Uncertainty metrics

To evaluate the quality of uncertainty quantification, we propose the following uncertainty metrics for each model, designed to be easily interpretable:

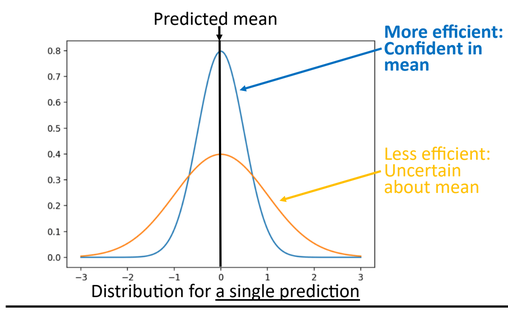
1. **Calibration R2 score** (↑, 1.00 is best)

   For a given confidence level of X, the true value is within the predicted range X of the time.

   This should be true for all 0 ≤ X ≤ 1.



2. **Efficiency score** (↓)



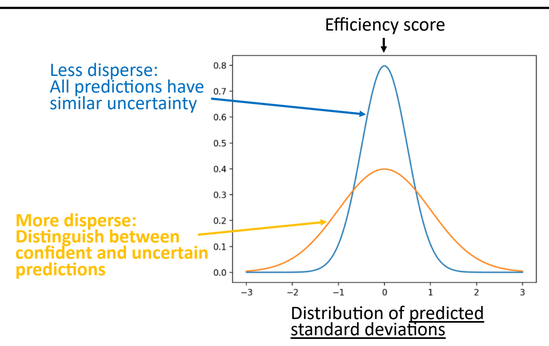
Mean of all predicted standard deviations.

Predictions should have as low uncertainty as possible on average.

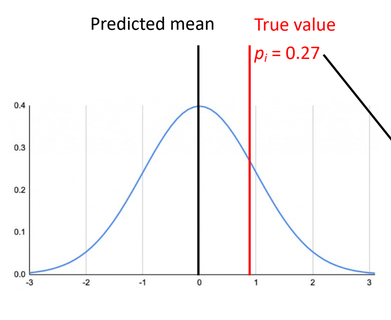
3. **Dispersion score** (↑)

Standard deviation of all predicted standard deviations.

Some predictions must be more uncertain than others.



4. **Geometric mean of probabilities (GMP)** (↑)



Geometric mean of p values for all predictions. This is analogous to the proper scoring rule negative-log-likelihood [7], but has units of probability.

Rewards confidence when the predicted mean is close to the true value and uncertainty when the predicted mean is not.

$$GMP = \left(\prod_{i=1}^{n} p_i\right)^{\frac{1}{n}} = \sqrt[n]{p_1 p_2 \cdots p_n}$$

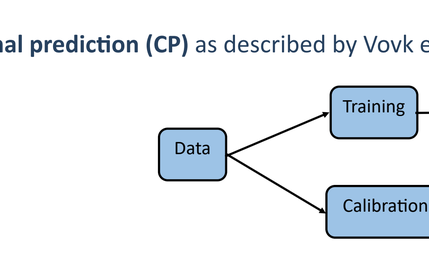
4. Model performance

Models were built for 21 targets, mean metrics across all targets for each uncertainty method are reported below:

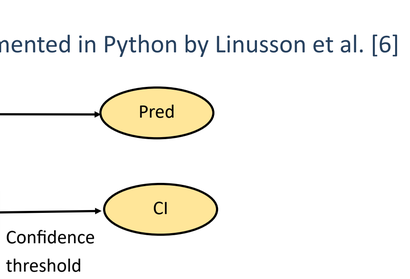
Metric	BB	CP	BNN
Test Calibration	0.087	0.922	0.765
Test Efficiency	0.154	0.856	1.110
Test Dispersion	0.023	0.296	0.096
Test GMP	0.633	0.645	0.281
Val Calibration	-1.003	0.794	0.905
Val Efficiency	0.292	0.985	1.159
Val Dispersion	0.120	0.437	0.102
Val GMP	0.001	0.164	0.219

- BB produces the smallest confidence intervals, but it is often **overconfident** and has poor performance elsewhere.
- CP has good calibration and excellent dispersion scores. It is very reliant on the **calibration set being a good sample of the target chemical space**.
- BNN is the **most robust model**, with the best calibration and GMP in external validation. It is however **too conservative** and predicts very wide confidence intervals even in the test set where it should be more confident.

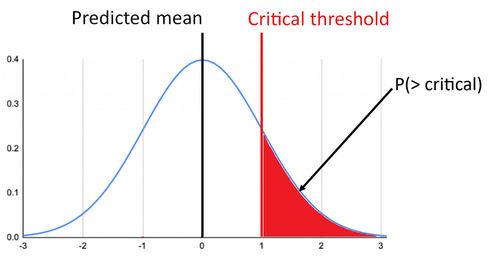
5. Interpretating uncertainty



If model is **well calibrated**, can define **confidence interval**: X% probability that true value is within range (X = 95 shown)



Epistemic uncertainty is correlated with **distance to nearest neighbours**, therefore is intrinsically a notion of **applicability domain**

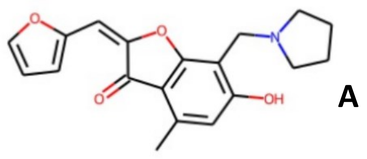


If interested in a **critical threshold**, can find **P(> critical)** which is more meaningful than comparing predicted mean to the threshold.

6. Case studies

Two example molecules from the validation set for the ADRB2 model are shown below.

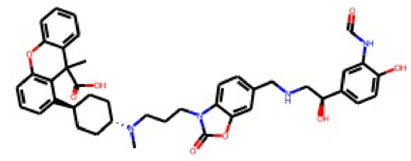
Molecule A has a lower epistemic uncertainty — it is closer to the applicability domain of the model and the model is more confident in the predicted mean. However because the predicted mean is close to the critical threshold of 5, there is still a significant P(> critical) of 0.255. Molecule B shows the reverse situation.



A

SMILES: Cc1cc(O)c(CN2CCCC2)c2c1C(=O)/C(=C\c1ccco1)O2

Predicted mean:	**[4.54]**
95% CI:	**[3.12] — [5.96]**
Epistemic uncertainty:	**[0.244]**
P(> 5):	**[0.255]**



B

SMILES: CN(CCCn1c(=O)oc2cc(CNC[C@H](O)c3ccc(O)c(NC=O)c3)ccc21)

Predicted mean:	**[6.28]**
95% CI:	**[4.53] — [8.04]**
Epistemic uncertainty:	**[0.574]**
P(> 5):	**[0.929]**

7. Conclusion

1. **Interpretable metrics** for evaluating uncertainty quantification
2. Uncertainty includes information about **applicability domain**
3. Uncertainty allows us to make **more useful and nuanced predictions**

References

1. Begoli, E. et al., 2019. *Nature Machine Intelligence*, 1(1), pp.20-23.

2. Allen, T.E.H. et al., 2022. *Computational Toxicology*, 23, p.100228.

3. Rubin, D., 1981. *The Annals of Statistics*, 9(1).

4. *GitHub - lmc2179/bayesian_bootstrap*

5. Vovk, V. et al., 2005. *Algorithmic learning in a random world*.

6. *GitHub - donlnz/nonconformist*

7. Dawid, A. and Musio, M., 2014. *METRON*, 72(2), pp.169-

Declaration

No competing interests to declare.

Contact details

Email: cg588@cam.ac.uk

Acknowledgements

Unilever and Clare College for funding.

Tim Allen for helpful discussions regarding implementation and interpretation of Bayesian neural networks