Gaussian Process Molecular Property Prediction with FlowMO

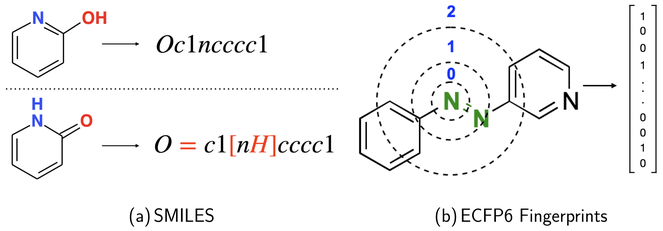
Lancaster University

Henry Moss*, Ryan-Rhys Griffiths*

UNIVERSITY OF CAMBRIDGE

A Gaussian Process Library for Molecules

- FlowMO is a Gaussian Process library for molecules
- Representations include SMILES and ECFP6 fingerprints
- Bespoke kernels for these representations include string kernels for SMILES and the Tanimoto kernel for fingerprints.



(a) SMILES (b) ECFP6 Fingerprints

Figure 1:Molecular Representations in FlowMO.

Regression Benchmark

- We benchmark regression performance on three small molecular datasets: The Photoswitch dataset [1], ESOL [2] and FreeSolv [3].
- Compare against the best reported model from the MoleculeNet benchmark in addition to recently reported SOTA models.
- Achieve best performance on the photoswitch dataset

Table 1:RMSE of the models across the three datasets, with the scores of the best GP model and best overall model highlighted .

	Photoswitch	ESOL	FreeSolv
SSK GP (SMILES)	26.0 ± 3.6	**0.65** ± 0.04	**1.29** ± 0.22
TK GP (Fingerprints)	**22.6** ± 4.0	0.98 ± 0.08	1.85 ± 0.10
ANP	27.2 ± 3.7	1.32 ± 0.13	2.65 ± 0.47
BNN	25.5 ± 5.0	1.01 ±0.11	1.92 ± 0.20
MoleculeNet	**22.0** ± 3.5	**0.58** ± 0.03	1.15 ± 0.02
SMILES-X	-	0.70 ± 0.05	1.14 ± 0.17
SMILES-X (Augm)	-	**0.57** ± 0.07	**0.81** ± 0.22

Uncertainty Calibration

To analyse the calibration achieved by the predictive distributions provided by the probabilistic models (only the GPs, BNN and ANP), we define a calibration score function

$$\mathrm{C(q)} = \frac{1}{|\mathcal{T}|}\sum_{m\in\mathcal{T}}\left[1\left(\left|\frac{\hat{y}(m)-y(m)}{\hat{\sigma}(m)}\right| < \Phi^{-1}\left(\frac{1+q}{2}\right)\right)\right]$$

based on cross-validatory predictive p-values. $y(m)$, $\hat{y}(m)$ and $\hat{\sigma}(m)$ represent true values, predictive means and predictive standard deviations for each test molecule $m \in \mathcal{T}$, and Φ^{-1} is the inverse of the standard Gaussian cumulative distribution function. The indicator 1 is activated only when the true value is contained in the model's $q*100\%$ confidence interval. Therefore, perfect calibration at the q^{th} quantile corresponds to $C(q) = q$. $C(q) > q$ indicates under-confidence through overly large uncertainty estimates (limiting the strength of conclusions that can be drawn from the model) whereas $C(q) < q$ denotes over-confidence (leading to reckless decisions downstream). We plot $C(q)$ for our probabilistic models in Figure 2.

Uncertainty Calibration Benchmark

String kernel GP demonstrates superior calibration across all tasks.

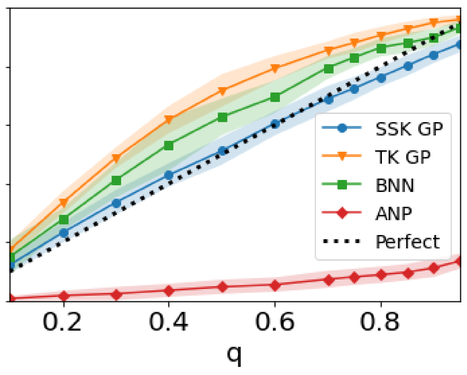


Figure 2:Uncertainty Calibration on the FreeSolv Dataset.

Uncertainty Calibration Benchmark

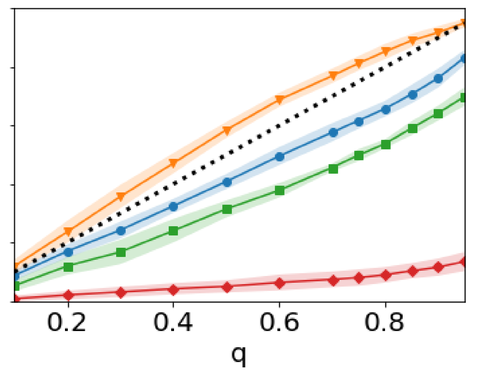


Figure 3:Uncertainty Calibration on the ESOL Dataset.

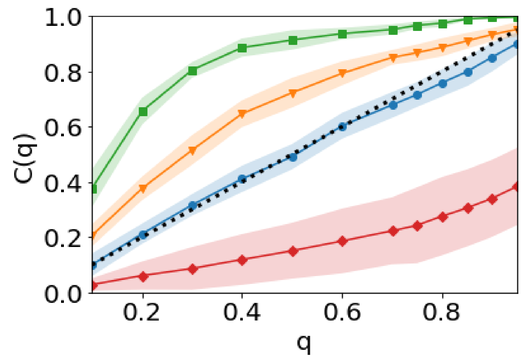


Figure 4:Uncertainty Calibration on the Photoswitch Dataset.

Future Work

We plan to extend the library to cater for graph representations of molecules by making use of graph kernel GPs.

References

[1] A. R. Thawani, R.-R. Griffiths, A. Jamasb, A. Bourached, P. Jones, W. McCorkindale, A. A. Aldrick, and A. A.Lee. *The photoswitch dataset: A molecular machine learning benchmark for the advancement of synthetic chemistry.* arXiv preprint arXiv:2008.03226, 2020

[2] J. S. Delaney. *ESOL: estimating aqueous solubility directly from molecular structure.* Journal of Chemical Information and Computer Sciences, 2004.

[3] D. L. Mobley and J. P. Guthrie. *FreeSolv: a database of experimental and calculated hydration free energies, with input files.* Journal of Computer-Aided Molecular Design, 2014.