Gaussian Process Molecular Property Prediction with FlowMO

Henry B. Moss[*]
STOR-i Centre for Doctoral Training
Lancaster University, UK
`h.moss@lancaster.ac.uk`

Ryan-Rhys Griffiths[*]
Department of Physics
University of Cambridge, UK
`rrg27@cam.ac.uk`

Abstract

We present FlowMO: an open-source Python library for molecular property prediction with Gaussian Processes. Built upon GPflow and RDKit, FlowMO enables the user to make predictions with well-calibrated uncertainty estimates, an output central to active learning and molecular design applications. Gaussian Processes are particularly attractive for modelling small molecular datasets, a characteristic of many real-world virtual screening campaigns where high-quality experimental data is scarce. Computational experiments across three small datasets demonstrate comparable predictive performance to deep learning methods but with superior uncertainty calibration.

1 Introduction

In the early stages of exploring a new class of drug molecules or molecular material, there is often only a small quantity of high-quality experimental data available [Thawani et al., 2020, Griffiths et al., 2018]. In contrast to the big data regime of established molecule classes, where much of the information regarding performance has already been acquired, it is the small data regime which holds most promise for techniques such as Bayesian optimization [Gómez-Bombarelli et al., 2018, Korovina et al., 2020, Griffiths and Hernández-Lobato, 2020, Tripp et al., 2020, Moss et al., 2020a] and active learning [Zhang and Lee, 2019] to expedite the rate at which novel and performant molecules are discovered.

The current trend in molecular machine learning research has been to leverage Bayesian deep learning to provide the uncertainty estimates which underpin active learning and Bayesian optimization methods [Ryu et al., 2019, Zhang and Lee, 2019, Hwang et al., 2020, Scalia et al., 2020]. Deep learning methods, however, are often not the model of choice for small datasets. In fact, leading experts on deep learning have expressed a preference for Gaussian Processes (GPs) [Rasmussen, 2004] for small datasets [Bengio, 2011]. Furthermore, Bayesian deep learning methods rely on approximate inference in order to produce uncertainty estimates. In contrast, GPs admit exact inference at the expense of computationally prohibitive cost for very large datasets. In this paper, we compare calibration of the uncertainty estimates of GPs against a popular Bayesian deep learning method as well as against the recently-introduced attentive neural process [Kim et al., 2019].

Applying GPs to molecules in non-trivial, as the vast majority of existing applications of GPs assume inputs of low and fixed dimension. This assumption is not satisfied for the popular molecular representations of fingerprints [Rogers and Hahn, 2010] and SMILES strings [Weininger, 1988], sparse high-dimensional bit vectors and strings of variable length, respectively. To build GPs over molecules, FlowMO provides GPU-supported implementations of the Tanimoto and string kernels, providing a user-friendly way to make probabilistic predictions, whilst avoiding the expensive architecture tuning and model optimization often required to find effective deep learning models.

Our primary contributions can be summarized as follows. (1) We propose a GP framework for molecular property prediction using Tanimoto and string kernels. (2) We provide an open-source
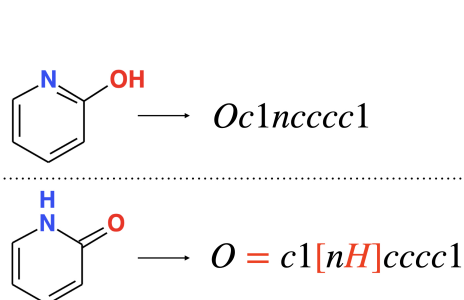
---

[*]Equal contribution

Figure 1: SMILES strings for structurally similar molecules have local differences (red) but common non-contiguous subsequences (black).
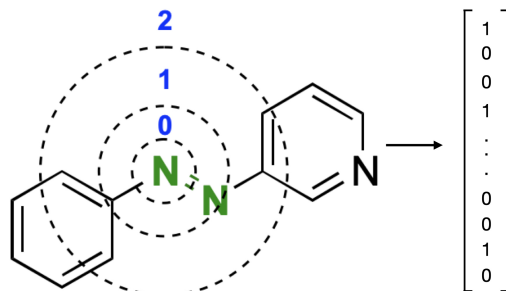

Figure 2: Fingerprints provide high-dimensional binary representations of molecules by encoding fragments whose size is dictated by a pre-specified bond radius.

Python implementation built upon GPflow[De G. Matthews et al., 2017] and RDKit [Landrum, 2013], addressing the absence of GP support in popular libraries such as DeepChem [Ramsundar et al., 2019], DGL-LifeSci [Wang et al., 2020] and ASAP Cheng et al. [2020]. (3) We compare FlowMO with established baselines across three benchmark datasets with plans for a more extensive comparison of uncertainty calibration.

2 Molecular Representations

To apply GPs to molecules, we need a meaningful way to represent molecules. FlowMO currently supports two popular representations: SMILES and ECFP fingerprints.

SMILES strings. The Simplified Molecular-Input Line-Entry System (SMILES) [Weininger, 1988] is a procedure for expressing molecules as character strings. Each SMILES string is assembled by traversing the molecule's chemical graph. Two SMILES strings for structurally similar pyridine derivatives are presented in Figure 1. The alphabet of SMILES strings contains letters representing aliphatic and aromatic atoms $(B, C, N, O, S, P, F, Cl, Br, I, b, c, n, o, s, p)$, parentheses and integers, as well as additional ASCII symbols representing bonds and the presence of rings $(-, =, \#, \&, /, \backslash, ., \%)$.

ECFP fingerprints. ECFP fingerprints are high-dimensional bit vector representations of molecules [Christie et al., 1993] designed to search chemical databases and in the context of kernel-based classification have been used in conjunction with support vector machines [Ralaivola et al., 2005]. In this work, we consider extended-connectivity (ECFP) fingerprints [Rogers and Hahn, 2010] as generated by the RDKit library [Landrum, 2013]. ECFP fingerprints are calculated with an iterative procedure. First, each atom is represented in terms of its local properties, for example its atomic mass and valency. These local representations are then updated based on the representations of its neighbors, iteratively building representations of all fragments containing as many atoms as the chosen bond radius of the fingerprint. Finally, duplicates structures are removed and the remaining representations are hashed into a unique bit vector (see Figure 2). For our experiments, we use 2048-bit ECFP fingerprints with a bond radius of 3.

3 Molecular Property Prediction with Gaussian Processes

Given a training set of (possibly noisy) experimentally-determined molecular properties and a kernel function k measuring intermolecular similarity, Gaussian processes (See Rasmussen [2004] for a comprehensive introduction) provide tractable Gaussian predictive distributions for the properties of any out-of-sample target molecule. In FlowMO, we provide a kernel for each of our supported molecular representations (as introduced above): a Sub-sequence String Kernel (SSK) and a Tanimto Kernel (TK) suitable for measuring the similarity between SMILES strings of varying length and high-dimensional binary fingerprint vectors, respectively.

The computational complexity of training and inference for the GP are $O(n^3 + n^2C)$ and $O(nC)$, where n is the number of training molecules and C is the cost of a single kernel evaluation. Although there exist many methods that reduce the computational complexity of GPs (for example Hensman et al. [2013]), they are designed with smooth input spaces in mind and do not yet support molecular data. In practice, we found standard GPs to be adequate for the datasets considered in this paper, and furthermore they can be comfortably applied to larger data sets using FlowMO's GPU support.

String Kernels. Sub-sequence String Kernels [Lodhi et al., 2002, Cancedda et al., 2003] measure the similarity of strings through the number of shared sub-strings, naturally supporting variable length inputs. The considered sub-strings can be non-contiguous, yielding a rich contextual model of string data that encapsulates the non-contiguity known to be important when characterising inter-SMILES similarity (Figure 1). To avoid enumeration of an exponentially-scaled feature space, SSKs exploit the kernel trick through an efficient dynamic programming algorithm, allowing individual kernel calculations between two strings $\mathbf{s}$ and $\mathbf{s}'$ to be calculated in $C = O(\max(|\mathbf{s}|, |\mathbf{s}'|)^2)$ operations. FlowMO's SSK implements the vectorized formulations of Beck and Cohn [2017] and Moss et al. [2020a], enabling efficient computation on GPUs. FlowMO's SSKs have two kernel parameters λ_m and λ_g controlling the relative weighting of long and/or highly non-contiguous sub-strings. In all our experiments, we set $n = 5$ and learn appropriate values for λ_g and λ_m through the standard practice of maximising the marginal likelihood of the GP.

Tanimoto Kernels. To build a GP model over fingerprint representations of molecules, we employ a kernel based on the Tanimoto similarity measure from the chemoinformatics literature [Gower, 1971, Fligner et al., 2002, Ralaivola et al., 2005]. Note that ECFP fingerprints contain $\gg 1000$ binary entries, each flagging the occurrence of particular features, and so have length and sparsity unsuitable for standard GP kernels designed for low-dimensional continuous spaces. The Tanimoto Kernel (TK) measures the similarity between two molecules via their fingerprints by the number of features present in both molecules normalized by the number of features occurring separately, i.e $k(\mathbf{f}, \mathbf{f}') = \sigma^2 \mathbf{f} \cdot \mathbf{f}' / (\mathbf{f} \cdot \mathbf{f} + \mathbf{f}' \cdot \mathbf{f}' - \mathbf{f} \cdot \mathbf{f}')$, where $\mathbf{f}$ denotes a fingerprint representation and σ^2 is a single kernel hyper-parameter. The complexity of each kernel evaluation is simply $C = O(|\mathbf{f}|)$.

4 Experimental Validation

We now evaluate FlowMO (available at anonymized link) across three small molecular property prediction datasets.

1. The Photoswitch Dataset [Thawani et al., 2020]: a collection of 392 photoswitch molecules and their experimental E isomer $\pi = \pi^*$ transition wavelengths.
2. ESOL [Delaney, 2004]: the logarithmic aqueous solubility for 1128 organic small molecules.
3. FreeSolv [Mobley and Guthrie, 2014]: experimental hydration free energies for 642 molecules.

Both ESOL and FreeSolv are part of the MoleculeNet dataset collection widely used for the benchmarking of molecular property prediction models [Wu et al., 2018], whereas the Photoswitch Dataset is a recent benchmark designed to specifically investigate the properties of light-activated molecules.

We compare the performance of our SSK and TK GPs with a range of existing molecular property prediction models. We report the scores of the best found model during the extensive testing of molecular graph-based models detailed in Wu et al. [2018] namely a graph convolutional network (GCN) [Kipf and Welling, 2016] for the Photoswitch Dataset and a message passing neural network (MPNN) [Gilmer et al., 2017] for ESOL and FreeSolv. These scores are regarded as very strong benchmarks, only recently improved upon for ESOL and FreeSolv by SMILES-X [Lambard and Gracheva, 2020]. SMILES-X is a sophisticated model that relies on extensive hyperparameter optimization and requires data augmentation via augmented random SMILES. We include the performance of SMILES-X pre and post augmentation. Across all three data sets, we also consider a black-box alpha divergence minimization Bayesian Neural Network (BNN) [Hernandez-Lobato et al., 2016] and an Attentive Neural Process (ANP) [Kim et al., 2019] applied to fingerprint representations. Our ANP and BNN follow the exact implementation described by [Thawani et al., 2020], with hyperparameters tuned by grid search. Further details are available in the supplementary information. Reported scores for all methods are based on 20 random train/validation/test splits in an 80:10:10 ratio, except for the GPs which need no validation set and so can be applied to 90:10 train/test splits.

4.1 Predictive Performance

Table 1 demonstrates the predictive performance of the models, showing the mean and a single standard deviation of the root-mean-squared error (RMSE). The best GP models outperform the ANP and BNN, and achieve comparable performance to both the MoleculeNet collection and the pre-augmentation SMILES-X models. Only the substantially more computationally demanding augmented SMILES-X model is able to significantly outperform the GPs and only on the FreeSolv data. The SSK GP outperforms the TK GP on all but the Photoswitch Dataset, suggesting that

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

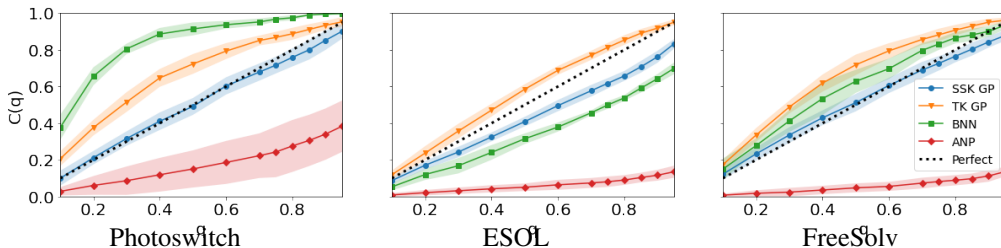|	Photoswitch	ESOL	FreeSolv
SSK GP (SMILES)	26.0 ± 3.6	$\mathbf{0.65} \pm 0.04$	$\mathbf{1.29} \pm 0.22$
TK GP (Fingerprints)	$\mathbf{22.6} \pm 4.0$	0.98 ± 0.08	1.85 ± 0.10
ANP	27.2 ± 3.7	1.32 ± 0.13	2.65 ± 0.47
BNN	25.5 ± 5.0	1.01 ± 0.11	1.92 ± 0.20
MoleculeNet	$\mathbf{22.0} \pm 3.5$	$\mathbf{0.58}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

Figure 3: Calibration scores for the probabilistic models. We see reliable calibration among our GP models, with the SSK GP achieving almost perfect calibration (as represented by the diagonal black dotted line) for Photoswitch and FreeSolv.

SMILES strings are useful representations for predicting aqueous solubility and hydration free energies, but ECFP fingerprints are especially informative for modeling transition wavelengths highlighting the importance of feature engineering for molecular property prediction tasks.

4.2 Calibration

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

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

based on cross-validatory predictive p-values [Marshall and Spiegelhalter, 2003, Leslie et al., 2007]. $y(m)$, $\hat{y}(m)$ 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 $\mathbb{1}$ is activated only when the true value is contained in the model's $q * 100\%$ confidence interval. Therefore, perfect calibration at the qth 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 as Figure 3.

5 Future Work

We have presented an analysis of GPs for molecular property prediction and uncertainty quantification. Future work will extend this analysis in three ways. Firstly, as string kernels operate directly on SMILES strings, we will explore if the data augmentation strategy exploited by SMILES-X can yield a similar performance boost for our SSK GPs. Secondly, we wish to consider GPs built upon additional kernels belonging to the broad class of convolutions kernels that includes our SSK. For example, graph kernels [Vishwanathan et al., 2010] could be used to define GPs directly over molecular graphs, enabling a comparison with recent work on graph-based deep learning methods [Yang et al., 2019, Hwang et al., 2020]. Finally, FlowMO will allow us to apply the extended Bayesian optimization methods designed for GP models to further improve efficiency in automatic wet-lab workflows [MacLeod et al., 2020]. Potential extensions include batch [González et al., 2016], multi-task [Swersky et al., 2013], multi-fidelity [Moss et al., 2020c] and multi-objective [Hernández-Lobato et al., 2016] optimization, as well as optimization with controllable experimental noise [Moss et al., 2020b].

References

D. Beck and T. Cohn. Learning kernels over strings using gaussian processes. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, 2017.

Y. Bengio. *What are some Advantages of Using Gaussian Process Models vs Neural Networks?*, 2011. https://www.quora.com/What-are-some-advantages-of-using-Gaussian-Process-Models-vs-Neural-Networks.

N. Cancedda, E. Gaussier, C. Goutte, and J.-M. Renders. Word-sequence kernels. *Journal of Machine Learning Research*, 2003.

B. Cheng, R.-R. Griffiths, S. Wengert, C. Kunkel, T. Stenczel, B. Zhu, V. L. Deringer, N. Bernstein, J. T. Margraf, K. Reuter, and G. Csanyi. Mapping materials and molecules. *Accounts of Chemical Research*, 53(9): 1981–1991, 2020. doi: 10.1021/acs.accounts.0c00403. URL https://doi.org/10.1021/acs.accounts.0c00403. PMID: 32794697.

B. D. Christie, B. A. Leland, and J. G. Nourse. Structure searching in chemical databases by direct lookup methods. *Journal of chemical information and computer sciences*, 1993.

A. G. De G. Matthews, M. Van Der Wilk, T. Nickson, K. Fujii, A. Boukouvalas, P. León-Villagrá, Z. Ghahramani, and J. Hensman. Gpflow: A gaussian process library using tensorflow. *The Journal of Machine Learning Research*, 18(1):1299–1304, 2017.

J. S. Delaney. Esol: estimating aqueous solubility directly from molecular structure. *Journal of chemical information and computer sciences*, 2004.

M. A. Fligner, J. S. Verducci, and P. E. Blower. A modification of the jaccard–tanimoto similarity index for diverse selection of chemical compounds using binary strings. *Technometrics*, 2002.

J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 2018.

J. González, Z. Dai, P. Hennig, and N. Lawrence. Batch bayesian optimization via local penalization. In *Artificial intelligence and statistics*, 2016.

J. C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 1971.

R. Griffiths, P. Schwaller, and A. Lee. Dataset bias in the natural sciences: A case study in chemical reaction prediction and synthesis design. 2018.

R.-R. Griffiths and J. M. Hernández-Lobato. Constrained bayesian optimization for automatic chemical design using variational autoencoders. *Chemical science*, 11(2):577–586, 2020.

J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian processes for big data. In *Uncertainty in Artificial Intelligence*, 2013.

D. Hernández-Lobato, J. Hernandez-Lobato, A. Shah, and R. Adams. Predictive entropy search for multi-objective Bayesian optimization. In *International Conference on Machine Learning*, 2016.

J. Hernandez-Lobato, Y. Li, M. Rowland, T. Bui, D. Hernández-Lobato, and R. Turner. Black-box alpha divergence minimization. In *International Conference on Machine Learning*, pages 1511–1520, 2016.

D. Hwang, G. Lee, H. Jo, S. Yoon, and S. Ryu. A benchmark study on reliable molecular supervised learning via bayesian learning. *arXiv preprint arXiv:2006.07021*, 2020.

H. Kim, A. Mnih, J. Schwarz, M. Garnelo, A. Eslami, D. Rosenbaum, O. Vinyals, and Y. W. Teh. Attentive neural processes. *arXiv preprint arXiv:1901.05761*, 2019.

T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

K. Korovina, S. Xu, K. Kandasamy, W. Neiswanger, B. Poczos, J. Schneider, and E. Xing. Chembo: Bayesian optimization of small organic molecules with synthesizable recommendations. In *International Conference on Artificial Intelligence and Statistics*, pages 3393–3403. PMLR, 2020.

G. Lambard and E. Gracheva. Smiles-x: autonomous molecular compounds characterization for small datasets without descriptors. *Machine Learning: Science and Technology*, 2020.

G. Landrum. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling, 2013.

D. S. Leslie, R. Kohn, and D. J. Nott. A general approach to heteroscedastic linear regression. *Statistics and Computing*, 2007.

H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text classification using string kernels. *Journal of Machine Learning Research*, 2002.

B. P. MacLeod, F. G. Parlane, T. D. Morrissey, F. Häse, L. M. Roch, K. E. Dettelbach, R. Moreira, L. P. Yunker, M. B. Rooney, J. R. Deeth, et al. Self-driving laboratory for accelerated discovery of thin-film materials. *Science Advances*, 2020.

E. Marshall and D. Spiegelhalter. Approximate cross-validatory predictive checks in disease mapping models. *Statistics in medicine*, 2003.

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.

H. B. Moss, D. Beck, J. Gonzalez, D. L. Leslie, and P. Rayson. Bayesian optimisation over string spaces. In *Advances in neural information processing systems*, 2020a.

H. B. Moss, D. S. Leslie, and P. Rayson. Bosh: Bayesian optimization by sampling hierarchically. *arXiv preprint arXiv:2007.00939*, 2020b.

H. B. Moss, D. S. Leslie, and P. Rayson. Mumbo: Multi-task max-value bayesian optimization. *arXiv preprint arXiv:2006.12093*, 2020c.

L. Ralaivola, S. J. Swamidass, H. Saigo, and P. Baldi. Graph kernels for chemical informatics. *Neural networks*, 2005.

B. Ramsundar, P. Eastman, P. Walters, V. Pande, K. Leswing, and Z. Wu. *Deep Learning for the Life Sciences*. O'Reilly Media, 2019. https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837.

C. E. Rasmussen. Gaussian processes in machine learning. In *Advanced Lectures on Machine Learning*. 2004.

D. Rogers and M. Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 2010.

S. Ryu, Y. Kwon, and W. Y. Kim. A bayesian graph convolutional network for reliable prediction of molecular properties with uncertainty quantification. *Chemical Science*, 10(36):8438–8446, 2019.

G. Scalia, C. A. Grambow, B. Pernici, Y.-P. Li, and W. H. Green. Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. *Journal of Chemical Information and Modeling*, 2020.

K. Swersky, J. Snoek, and R. P. Adams. Multi-task Bayesian optimization. In *Advances in Neural Information Processing Systems*, 2013.

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.

A. Tripp, E. Daxberger, and J. M. Hernández-Lobato. Sample-efficient optimization in the latent space of deep generative models via weighted retraining. *arXiv preprint arXiv:2006.09191*, 2020.

S. V. N. Vishwanathan, N. N. Schraudolph, R. Kondor, and K. M. Borgwardt. Graph kernels. *Journal of Machine Learning Research*, 2010.

M. Wang, D. Zheng, Z. Ye, Q. Gan, M. Li, X. Song, J. Zhou, C. Ma, L. Yu, Y. Gai, T. Xiao, T. He, G. Karypis, J. Li, and Z. Zhang. Deep graph library: A graph-centric, highly-performant package for graph neural networks, 2020.

D. Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 1988.

Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 2018.

K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388, 2019.

Y. Zhang and A. A. Lee. Bayesian semi-supervised learning for uncertainty-calibrated prediction of molecular properties and active learning. *Chemical Science*, 10(35):8154, 2019.

A   Broader Impact

As mentioned in the main body of the paper, calibrated uncertainty estimates are highly important for the successful implementation of techniques which depend on them such as Bayesian Optimization and active learning. To this end, Gaussian Processes are an important, yet somewhat neglected method (in the molecular machine learning domain) for producing well-calibrated uncertainty estimates for small datasets. It is our hope that through the provision of a bespoke library for Gaussian Processes for molecular property prediction, we can enhance the efficacy of downstream Bayesian Optimization and active learning campaigns in discovering promising drug molecules and molecular materials in the laboratory.

B   Hyper-parameter Settings

For the ESOL and FreeSolv data, hyper-parameters were selected by validation on an 80/10/10 train/validation/test split. For the Photoswitch data a split of 80/20 train/test was required in order to maintain consistency with the results reported in Thawani et al. [2020]. In this latter case, validation was performed by taking the original train set (80) and splitting this in the ratio of 90/10. Following hyper-parameter selection, the models were then re-trained using these hyper-parameters on the original train set.

For the BNN, a grid search was performed over the number of nodes per layer, the learning rate, the batch size as well as the number of iterations for the ADAM optimizer. The α hyper-parameter was fixed at 0.5 and the number of samples was fixed at 100. ReLU activations were the only activation functions considered and the number of layers was fixed to 2.

For the ANP, a grid search was performed over the number of nodes in the deterministic encoder, the number of nodes in the latent encoder and the number of nodes in the decoder in addition to the learning rate, the batch size and the number of training iterations. The dimensionality of the context encoding was fixed to 8. The number of layers for the deterministic encoder, the latent encoder and the decoder was fixed to 2 in each case. The number of samples to take of the context set was fixed to 10.