Drug Discovery under Covariate Shift with Domain-Informed Prior Distributions over Functions

Leo Klarner[1], Tim G. J. Rudner[1], Michael Reutlinger[2], Torsten Schindler[2], Garrett. M. Morris[1], Charlotte M. Deane[1], Yee Whye Teh[1]

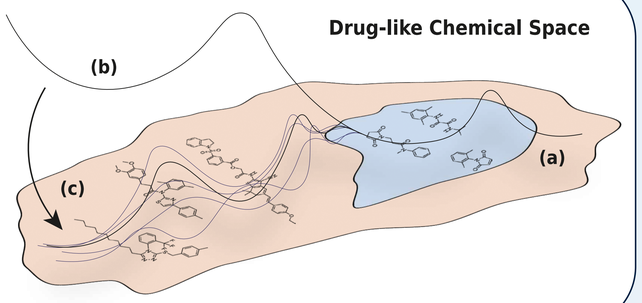
[1] *Department of Statistics, University of Oxford,* [2] *Pharma Research and Early Development, Roche*

Drug Discovery under Covariate Shift

- Machine learning models that can reliably predict **clinically relevant molecular properties** have the potential to accelerate key steps in early-stage drug discovery
- In practical settings, predictions are often most useful for **novel compounds** that are **structurally or functionally dissimilar** to known molecules (a)
- Standard deep learning algorithms perform poorly in this **out-of-distribution** regime, yielding both **incorrect and highly overconfident** predictions (b)
- We propose **Q-SAVI**: a framework to specify **explicit prior knowledge** of drug-like chemical space beyond (a) as a **regularizing prior distribution** over the **induced function space** of a neural network (c)

@leoklarner

/in/leo-klarner



Q-SAVI: A Framework to Specify Explicit Prior Knowledge

- Encoding domain-knowledge as a prior distribution over the parameters of a neural network is difficult

$$p_\theta(\boldsymbol{\theta}) \rightarrow p_f(f(\mathbf{x};\boldsymbol{\theta})) = \int_{\mathbb{R}^P} p_\theta(\boldsymbol{\theta}')\,\delta\left(f(\mathbf{x};\boldsymbol{\theta}) - f(\mathbf{x};\boldsymbol{\theta}')\right)\mathrm{d}\boldsymbol{\theta}'$$

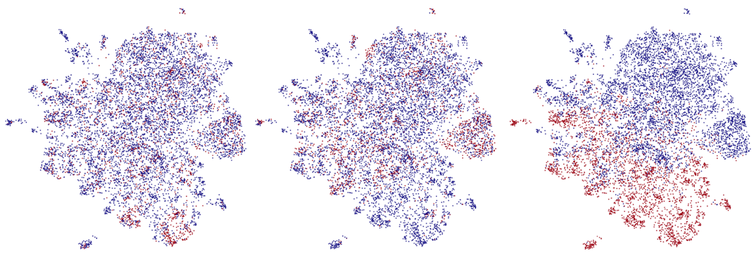
- Instead, we consider the **function space** induced by a given neural network architecture (evaluated at a set of **context points**)

$$p_{f|\mathcal{D}}(f(\mathbf{x};\boldsymbol{\theta})|\mathcal{D}) \propto p_{\mathcal{D}|f}(\mathcal{D}|f(\mathbf{x};\boldsymbol{\theta})) \cdot p_f(f(\mathbf{x};\boldsymbol{\theta}))$$

- We then rephrase the **inference problem** of learning a distribution over parameters as **learning a distribution over the functions** these parameters encode
- This enables us to formulate a prior distribution over the space of **Q**uantitative **S**tructure-**A**ctivity mappings and perform **V**ariational **I**nference in the resulting probabilistic model, a framework we refer to as **Q-SAVI**
- Q-SAVI allows us to restrict a neural network's hypothesis space by enabling the specification of **explicit, domain-informed prior knowledge** by encoding:
  (1) problem-specific modeling preferences in the **function-space prior** itself
  (2) and providing set of (potentially unlabelled) **context points** it is enforced at

Experimental Setup

- We use Q-SAVI to specify an uninformative prior that encourages **high predictive uncertainty** away from the training data and enforce it on unlabeled molecules sampled uniformly from ZINC
- To construct a challenging and meaningful evaluation setup, we:
  - Curate and pre-process a practically relevant dataset with **high-quality bioactivity labels**
  - Define appropriate **statistics to quantify covariate and label shift** in chemical space
  - Split molecules by **molecular weight** and using **spectral clustering**, inducing more **meaningful covariate shift** than the standard approach of splitting by scaffold



Scaffold split. Molecular weight split. Spectral split.

Results & Conclusion

	Model & Featurization	Spectral Split ECFP	Spectral Split rdkitFP	Weight Split ECFP	Weight Split rdkitFP
AUC-ROC (↑)	Logistic Regression	.583±.000	.551±.000	.626±.000	.632±.000
	Random Forest	.576±.009	.552±.006	.592±.006	.567±.004
	MLP	.574±.006	.571±.003	.614±.004	.577±.005
	Deep Ensemble	.589±.006	.571±.002	.644±.001	.594±.002
	GIN	.549±.009	.551±.007	.582±.007	
	GIN (attr masking)	.588±.004	.559±.010	.625±.004	
	GIN (context pred)	.541±.005	.566±.009	.621±.003	
	Grover	.574±.002	.544±.006	.623±.003	
	Q-SAVI	**.606**±.003	**.603**±.006	**.650**±.002	**.643**±.003

- We compare Q-SAVI to a range of **self-supervised pre-training** and **domain adapation techniques**, outperforming all of them in terms of predictive accuracy and most of them in terms of calibration
- Imbuing neural networks with **contextualized prior knowledge of the data-generating** process substantially improves their performance in extrapolative, out-of-distribution regimes
- Q-SAVI also presents researchers with a transparent and **probabilistically principled framework** to encode additional, **problem-infoded modeling preferences**, such as synthesizability, patentability, likely adverse side-effects, etc.