A new framework for evaluating model out-of-distribution generalization for the biochemical domain

Raúl Fernández-Díaz*, Thanh Lam Hoang, Vanessa Lopez, Denis Shields

University College Dublin

IBM Research

Introduction

Motivation

Machine learning models in scientific discovery are expected to work in previously unseen data, which are out-of-distribution (OOD).

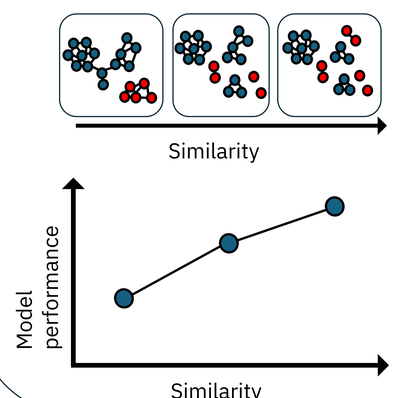
Research gaps

1. How can we choose the similarity metric, without relying on domain expertise?
2. How can we prevent training/testing leakage, without losing data?
3. How can we estimate model performance given an arbitrary number of deployment distributions, without repeating experiments?

Hestia-GOOD framework

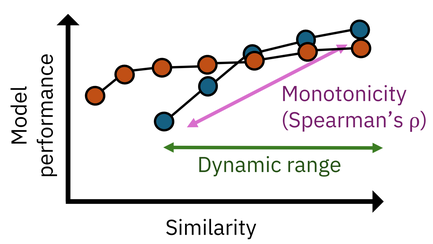
GOOD curve

Model performance as a function of train-test similarity



Similarity metric selection

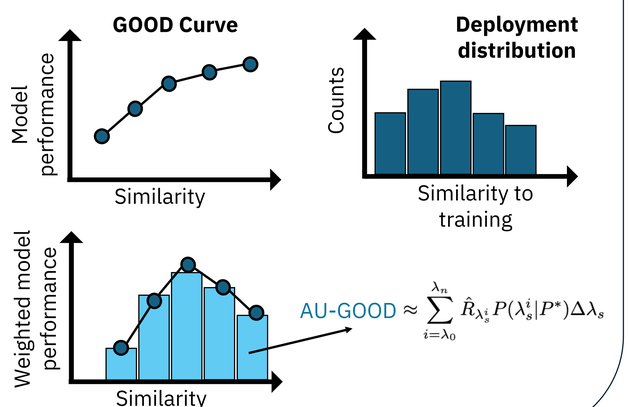
Quantitative analysis of best similarity function for a given task/dataset



- **Monotonicity:** Is model performance a function of train-test similarity?
- **Dynamic range:** What is the resolution of the similarity metric?

AU-GOOD metric

Estimation of model performance conditioned on a deployment distribution



Choice of similarity metric

Features

- ECFP
- MAPc
- Lipinski
- BUTCD
- Molformer-XL
- ChemBERTa-2

Indices

- Jaccard/Tanimoto
- Sokal
- Dice
- Rogot-Goldberg
- Euclidean

Best metrics

Dataset	Similarity metric	Dynamic range	Monotonicity
Ames	MAPc - Jaccard	0.75	0.82 ± 0.02
DILI	MAPc - Jaccard	0.85	0.81 ± 0.02
PAMPA-NCATS	Molformer	0.80	0.13 ± 0.02
Caco2	Molformer	0.85	0.76 ± 0.03
Half-life	Lipinski	0.60	0.46 ± 0.06
LD50	MAPc - Jaccard	0.70	0.94 ± 0.02

Method comparison

AU-GOOD

Comparsion conditioned on the deployment distribution. It simulates a realistic scenario in pharmaceutical industry, where the virtual screening library is known.

Significant rank

Robust statistics using *post-hoc* Wicoxon test

	KNN		SVM		RF		LightGBM	
Dataset	AU-GOOD	S. Rank	AU-GOOD	S. Rank	AU-GOOD	S. Rank	AU-GOOD	S. Rank
Ames	0.89 ± 0.01	2	0.88 ± 0.00	4	0.90 ± 0.01	1	0.88 ± 0.01	4
DILI	0.95 ± 0.00	3	0.96 ± 0.00	1	0.95 ± 0.00	3	0.77 ± 0.04	4
Caco2	0.87 ± 0.00	4	0.92 ± 0.00	1	0.92 ± 0.00	2	0.88 ± 0.00	3
Half life	0.93 ± 0.00	2	0.97 ± 0.00	1	0.69 ± 0.04	3	0.23 ± 0.01	4
LD50	0.86 ± 0.00	2	0.83 ± 0.00	4	0.89 ± 0.00	1	0.83 ± 0.00	4
Average	0.90 ± 0.01	2.6	0.91 ± 0.01	2.2	0.87 ± 0.02	**2.0**	0.72 ± 0.05	4.8

Conclusions

Hestia-GOOD framework allows:

1. Selection of best similarity metrics, based on quantitative analysis
2. Model comparison conditioned on deployment distributions.
3. Analysis of model performance as a function of the similarity between a prediction instance and its training data.

Github Repository:

Contact info:

ICLR International Conference On Learning Representations