A new framework for evaluating machine learning in biochemistry and its application for peptides and small molecules

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

IBM Research - Europe – Dublin | School of Medicine – UCD | Conway Institute – UCD | SFI CRT for Genomics Data Science

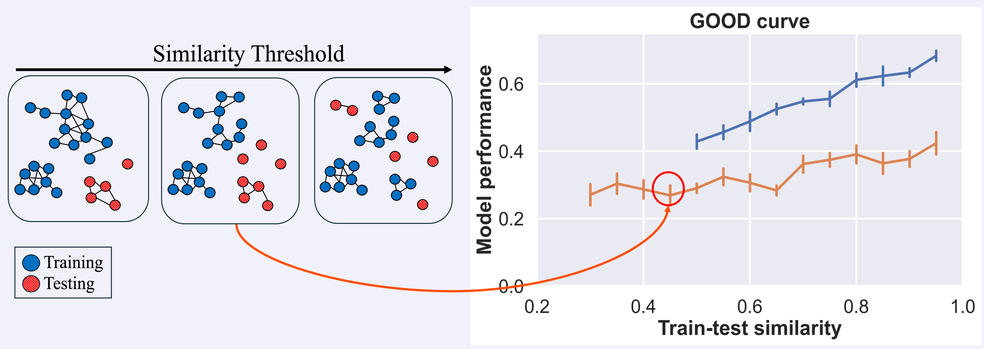
Introduction

- **Out-of-distribution (OOD):** Data that is different from the training data used to fit a model.
- **OOD evaluation is important for biochemistry** because models are expected to predict the properties of new molecules. More accurate estimations lead to **more trust** by the experimental community.
- We build a new **framework** for defining **OOD generalisation as a function of molecular similarity**.
- We provide metrics to **rationally choose the best molecular similarity function**, given a new task/dataset.
- We define a new generalisation metric, the **AU-GOOD**, that estimates model performance against any arbitrary target distribution(s).
- We present **Hestia-GOOD,** a suite of **Python** tools for leveraging and implementing this new framework across a **variety of biomolecules** (e.g. biosequences, protein structures, small drug-like organic compounds, etc.).

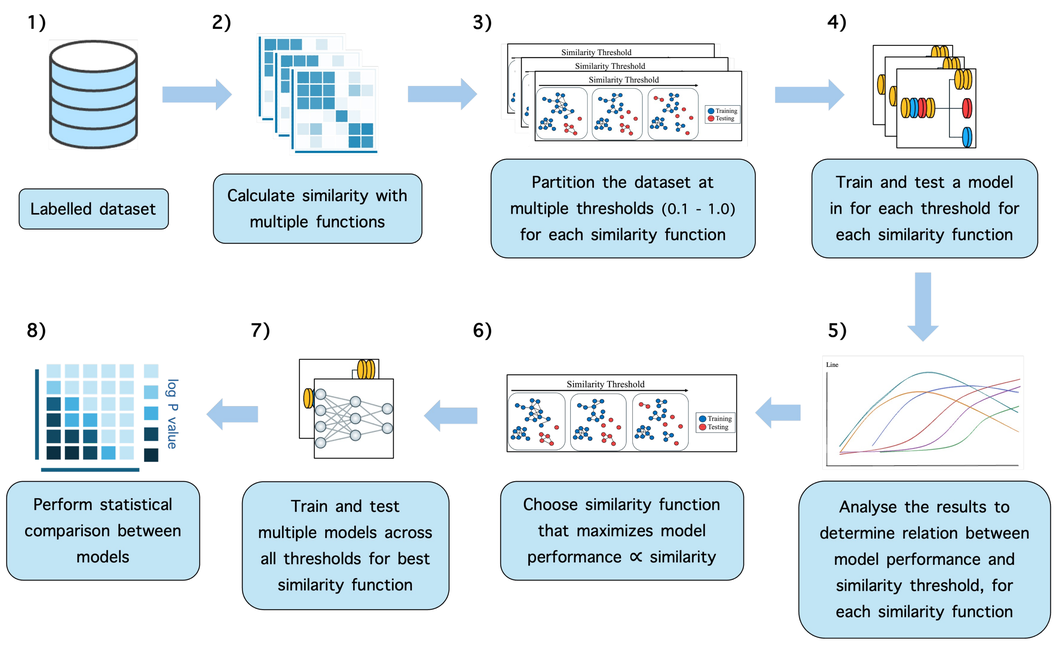
GOOD Curve

- We study model performance as a function of train-test similarity under multiple similarity functions
- Best similarity function will:

1. Have a reasonable dynamic range (resolution)
2. Show a good correlation between train-test similarity and model performance (otherwise the assumption would not be true)



Hestia-GOOD framework



Search for similarity function: Application to peptides

Similarity functions considered

Bioinformatics (canonical peptides)

Local sequence alignment:
- MMSeqs
- MMSeqs+Prefilter

Global sequence alignment
- Needleman-Wunsch

PLM embedding similarity
- ESM2-8M

Chemoinformatics (canonical and non-canonical)

Fingerprint similarity:
- MAPc (diameter: 4 to 20)
- ECFP (diameter: 4 to 20)

Chemical Language Model embedding similarity:
- Molformer-XL

Best functions per dataset

Dataset	Task	Similarity function	Dynamic range	Monotonicity (Spearman's rho)
Protein-peptide binding affinity (canonical)	Regression	MAPc-8	70	0.8 ± 0.1
Protein-peptide binding affinity (non-canonical)	Regression	MAPc-20	80	0.95 ± 0.03
Antibacterial (canonical)	Classification	MAPc-8	60	0.97 ± 0.02
Antibacterial (non-canonical)	Classification	MAPc-12	50	0.9 ± 0.1
Antiviral (canonical)	Classification	MMSeqs2 (prefilter)	90	0.95 ± 0.06
Antiviral (non-canonical)	Classification	ECFP-12	70	0.6 ± 0.2
Cell penetration (canonical)	Classification	MAPc-8	60	0.95 ± 0.06
Cell penetration (non-canonical)	Classification	MAPc-12	60	0.5 ± 0.2

Model performance estimation OOD: Application to small molecules