Effect of dataset partitioning strategies for evaluating out-of-distribution generalisation for predictive models in biochemistry

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):** From the perspective of a machine learning (ML) model, data that is different from the training data.
- **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 present **CCPart,** a new **dataset partitioning algorithm** that creates the **most OOD training-testing splits** possible given a dataset.
- We build a new **mathematical framework** for defining **OOD generalisation** as a **function of molecular similarity**. We define a new generalisation metric, the **AU-GOOD.**
- We present **Hestia,** 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.).

Mathematical framework

- Model: $f_\theta(x) \simeq y$ where $(x, y) \sim \mathcal{Z}$
- Partitioning strategy: $\Phi: \mathcal{Z} \rightarrow \mathcal{T},\ \mathcal{E}$
  - Training subset: $\mathcal{T}$
  - Evaluation subset: $\mathcal{E}$
- Population risk: $\mathcal{R}(\theta) = \mathbb{E}_{(x,y) \sim \mathcal{Z}}[\mathcal{L}(f_\theta(x), y)]$
- Empirical risk: $\hat{R}_E = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(f_{\theta,\mathcal{T}}(x_i), y_i)$
- Similarity-based partitioning: Φ_{λ_S}
- Empirical risk as function of similarity:

$$\hat{R}_{\mathcal{E}\sim\Phi(\mathcal{Z},\lambda_S)} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(f_{\theta,\mathcal{T}\sim\Phi(\mathcal{Z},\lambda_S)}(x_i), y_i)$$

- Generalisation to new data ($\mathcal{W}$) definition:

$$\mathcal{G}(\Phi(\mathcal{Z}|\mathcal{W})) = \mathbb{E}_{\Phi(\mathcal{Z})}\ \hat{R}_{\mathcal{E}\sim\Phi(\mathcal{Z}|\mathcal{W})} =$$

$$= \int_0^1 \hat{R}_{\mathcal{E}\sim\Phi(\mathcal{Z}|\mathcal{W})}\, p(\lambda_S|\mathcal{W})\, d\lambda_S$$

- Geometrical interpretation: area under the generalisation to OOD data (**AU-GOOD**) curve

Availability

Links to source code (Github Repository) and the biorxiv pre-print can be found through the QR-code:

Biomolecular similarity

Function $\mathcal{S}(x_i, x_j)$ such that:

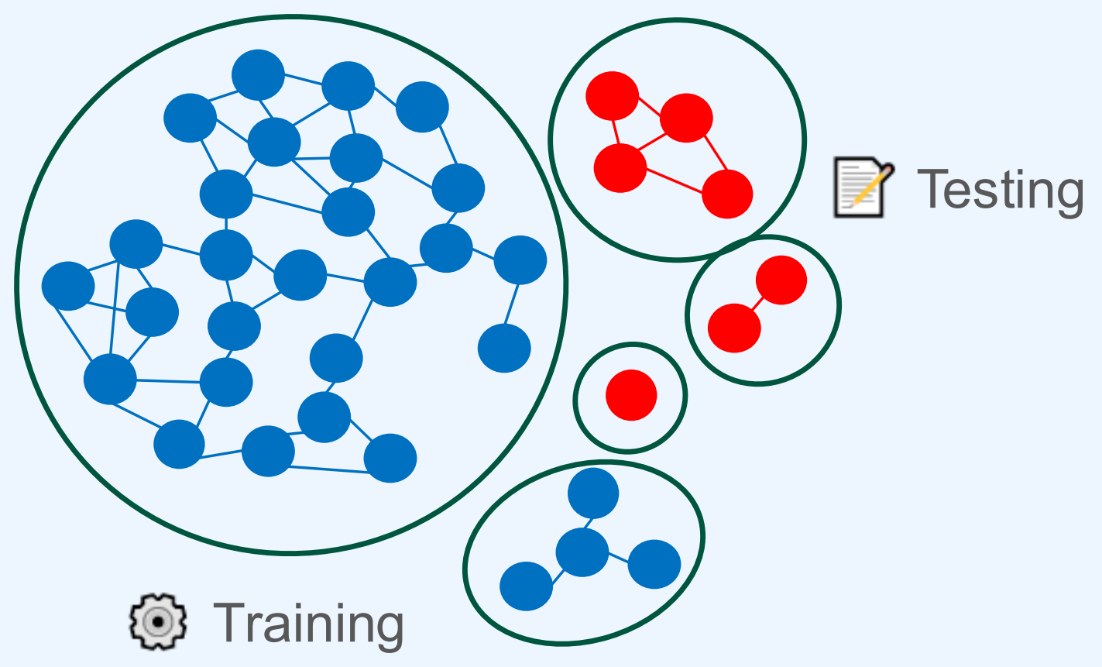
- It is normalized: $\mathcal{S} \in [0, 1]$
- It is symmetric: $\mathcal{S}(x_i, x_j) = \mathcal{S}(x_j, x_i)$
- The similarity between a molecule and itself is maximal: $\mathcal{S}(x_i, x_i) = 1$

Examples of molecular similarity metrics:

- Sequence identity (or e-value) in sequence alignment
- Tanimoto similarity between molecular fingerprints
- TM-score between protein structural alignments
- Manhattan (or Hamming) distance between multi-point mutants

CCPart algorithm

1. Calculate all pairwise similarities between the biomolecules in the dataset.
2. Given a similarity threshold, define a graph where the nodes are the biomolecules and the edges the similarities above the threshold.
3. Identify all unconnected subgraphs within that graph.
4. Iteratively assign the smallest unconnected subgraphs to testing subset until it reaches the desired size.



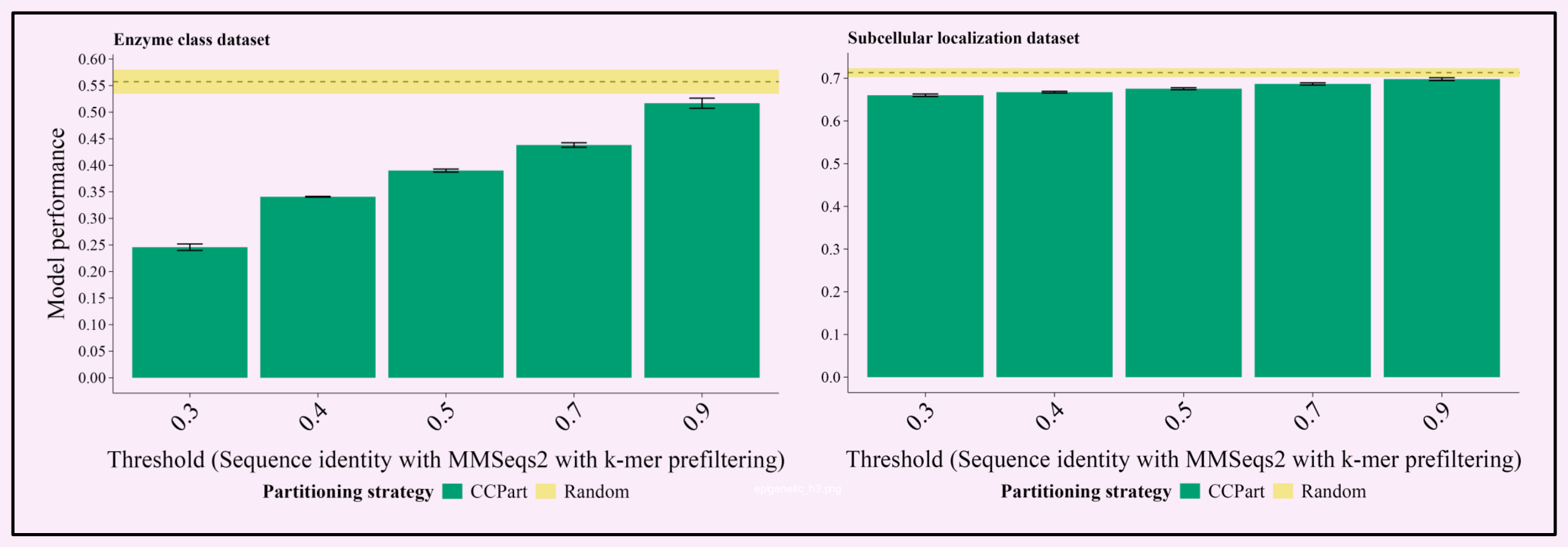
Understanding Molecular Language Models: A case study

- **Molecular language models** are machine learning models that model the conditional probability of a token (minimal component) in a molecule given the rest of the molecule ($p(t_i|m_{-t_i})$).
- **Experiments**:
  - **Settings:** All models are finetuned for 20 epochs with a MLP layer for classification/regression
  - **Protein Language Model (ESM2 8M):** similarity metric is sequence identity in MMSeqs2 pairwise alignments (with k-mer prefiltering).
  - **SMILES Language Model (MolFormer-XL):** similarity metric is Tanimoto similarity with extended-connectivity fingerprints (ECFP) with radius 2 and 1,204 bits.
  - **DNA Language Model (multi-species NucleotideTransformer 250M):** similarity metric is sequence identity in MMSeqs2 pairwise alignments (with k-mer prefiltering).

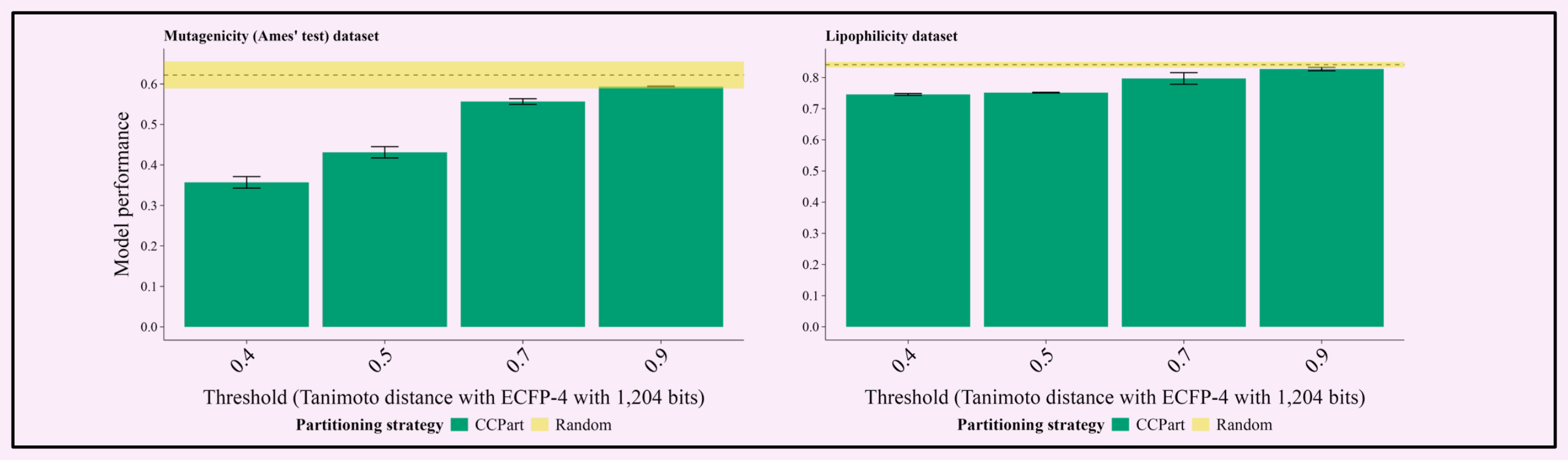
Understanding Molecular Language Models: A case study

- **Molecular Language Models tend generalize better to tasks mediated by short-range patterns**
- **Current tasks for DNA language models are not effective for testing model generalisation**

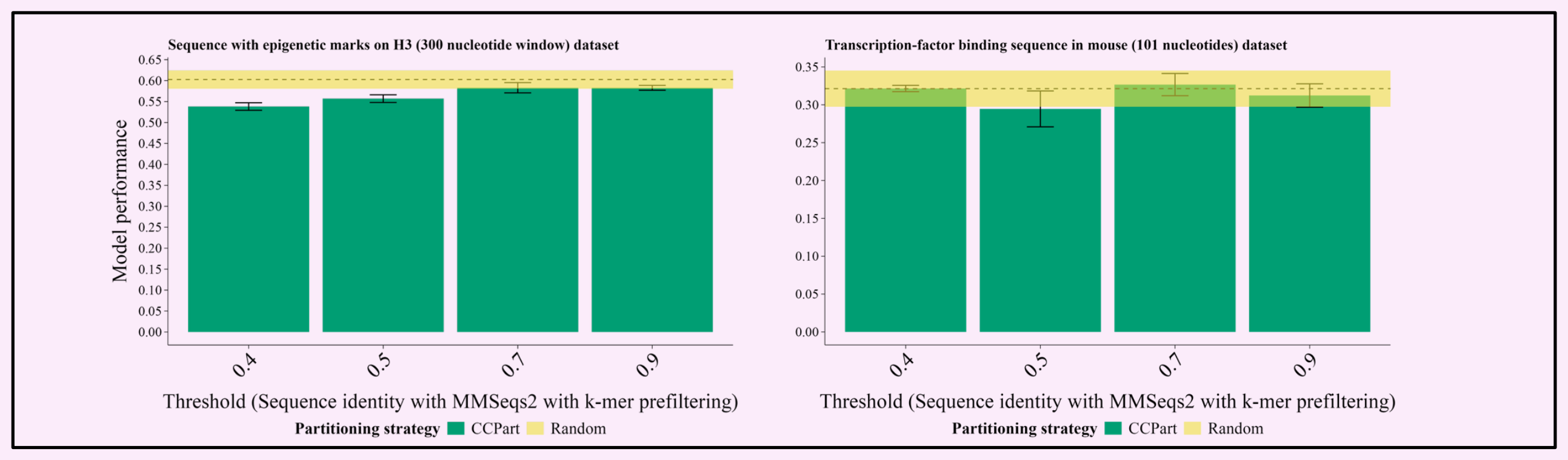
Protein tasks



Small molecule tasks

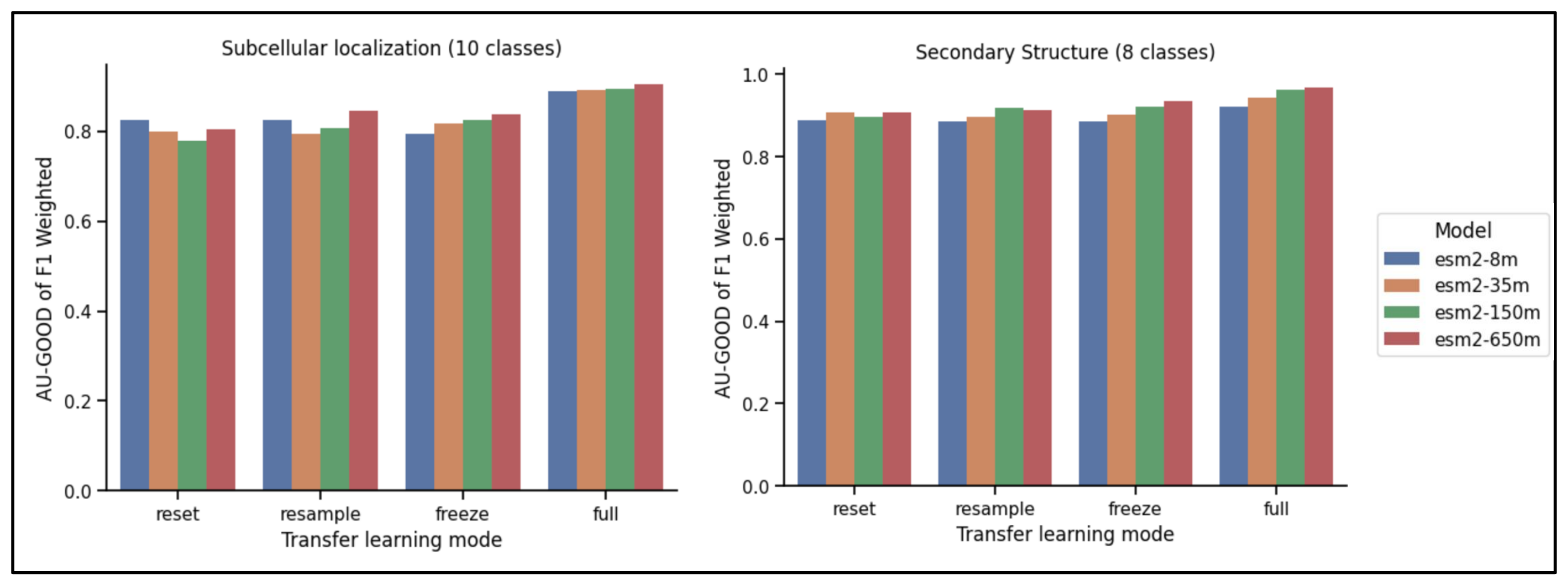


DNA tasks



Protein Language Model pretraining examination

- **AU-GOOD** as generalization metric to compare between models.
- **Experiments**:
  - **Reset:** No pre-training. Model weights randomly initialized.
  - **Resample:** Gross statistics. Model weights randomly permuted.
  - **Freeze:** Model weights frozen. Only finetuning the MLP head.
  - **Full:** Full model finetuning.
- **Results:** Model size scaling improves generalization both for local range tasks like secondary structure prediction and subcellular localization



- **Future work:** examining global range tasks like enzyme classification or thermostability