Rethinking the generalization of drug target affinity prediction algorithms via similarity aware evaluation

Chenbin Zhang*, Zhiqiang Hu*, Chuchu Jiang*, Wen Chen, Jie Xu, Shaoting Zhang

Oral in 2025 at ICLR

29.04.2025

Clemence Reda

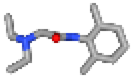
Background

1. The drug-target binding affinity prediction problem
2. Of the importance of proper data splitting
3. Issues with random splitting
4. SotA on fair predictive evaluation

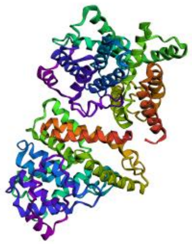
The drug-target binding affinity prediction problem screens drugs for interactions on a specific target.

e.g., **Protein docking** connect molecule and target and compute how strong the connection is

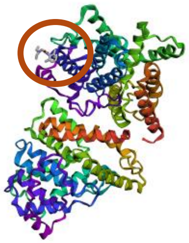
Drug
Lidocaine
cid:3676



Target (fixed)
PDE10A
pdb:4ajd



Predictive model
(e.g., MolModa **[1]**)



Docking score
-3.28 kcal/mol << 0

Other scores can be considered: KI or IC_{50}

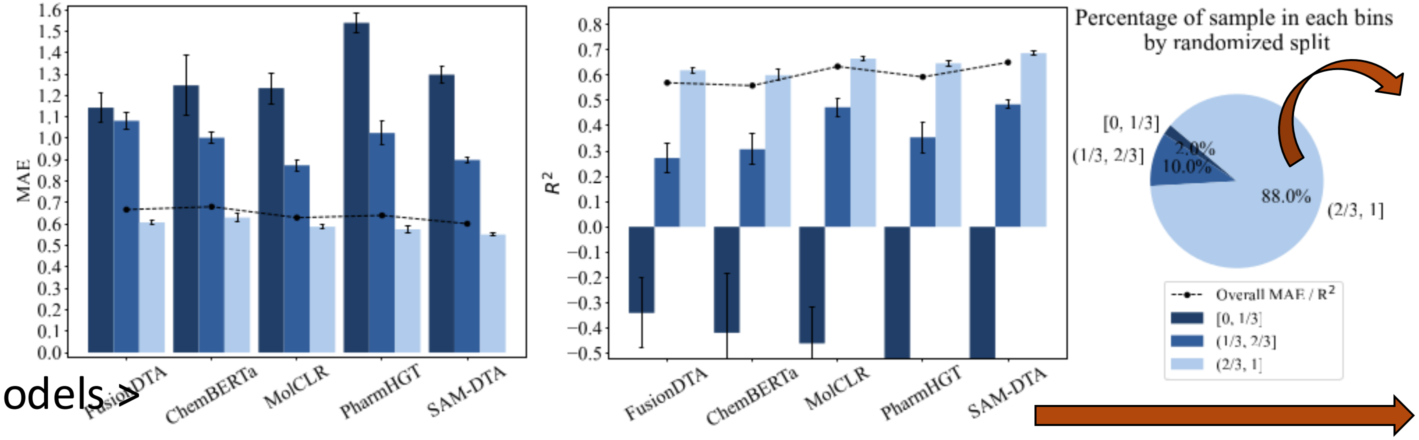
Binding is (often) an inhibitory action

- A prefiltering task in drug discovery / repurposing
- Lots of literature (structure-based, sequence-based, similarity-based, …)

A classifier can be trained with text or tabular data, and evaluated on R^2 or MSE metrics

[1] Kochnev, … & Durrant. (2024). MolModa: accessible and secure molecular docking in a web browser. *Nucleic Acids Research*, *52*(W1), W498-W506. *https://durrantlab.pitt.edu/molmoda/#*

Of the importance of proper data splitting related to fair evaluation of model generalizability.

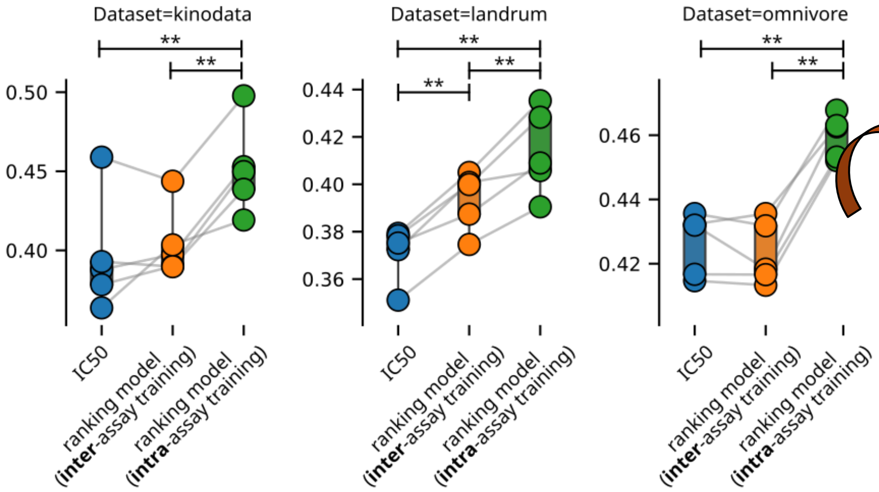


Models ->

From Fig. 1 in the paper.

Most drugs in the test set are **similar** to drugs in the train set

Driving the *seemingly* good performance on the test set (≈ data leakage)



Drugs often cluster by assay origin
Might lead to less efficient training due to batch effect and unfair comparisons

Aggregated assay-wise evaluation and training

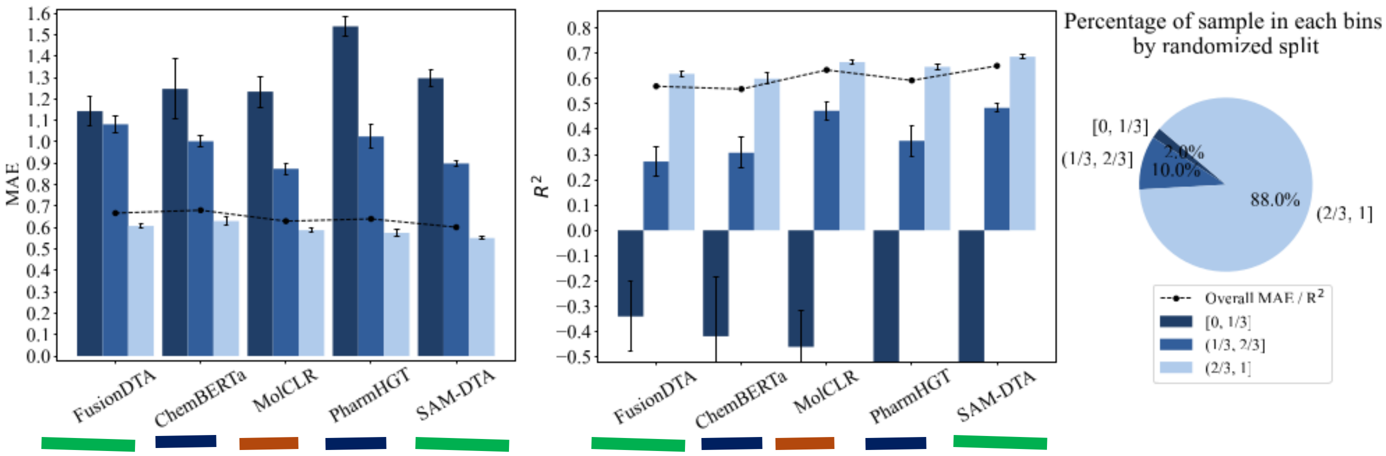
From Fig. 4 in Backenköhler et al. (2025). Assay-Based Machine Learning: Rethinking Evaluation in Drug Discovery. *ChemRxiv*.

Issues with random splitting and why and when random splitting can be applied.

Random splitting "balances out" the sample similarity between the train/test sets
if the samples are **drawn iid = strong assumption which does not hold for drugs (too optimistic)**

Tested drugs:

- Different exposure times and doses but same molecule
- Some drugs are more common (with potentially similar mechanisms of action) e.g., cancer
- Assay-related batch effect
- Relatively small data sets (< 5k drugs per target in open data sets)



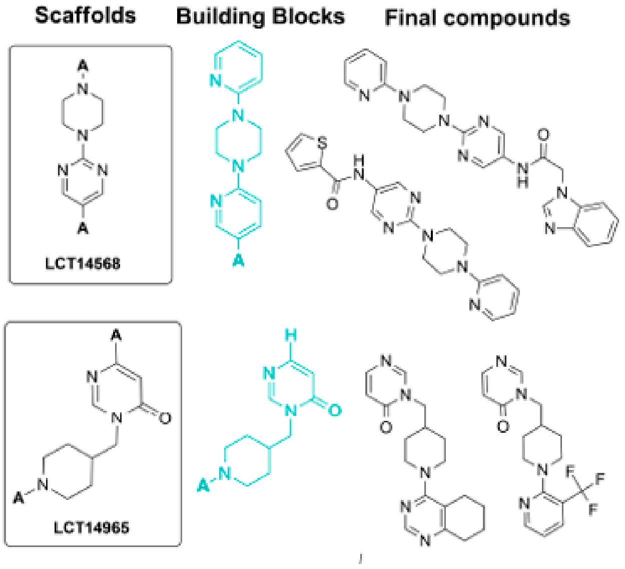
From Fig. 1 in the paper.

Observations still stand for other data sets, drug features, predictive models, evaluation metrics and similarity measures

- 🟧 GCN on atom graphs
- 🟩 NNs on fingerprints
- 🟦 Transformers

Fair predictive evaluation in SotA alternatives to random splitting in related works.

Scaffold splitting = identify (e.g., Markush, Murcko **[2]**) scaffolds and segregate molecules with the same scaffolds (≈ in the same structural class) in the train or test sets



Adapted from **[3]**

Pros

Scaffolds are a quick way to assess the structural similarity between drugs

Cons **[4]**

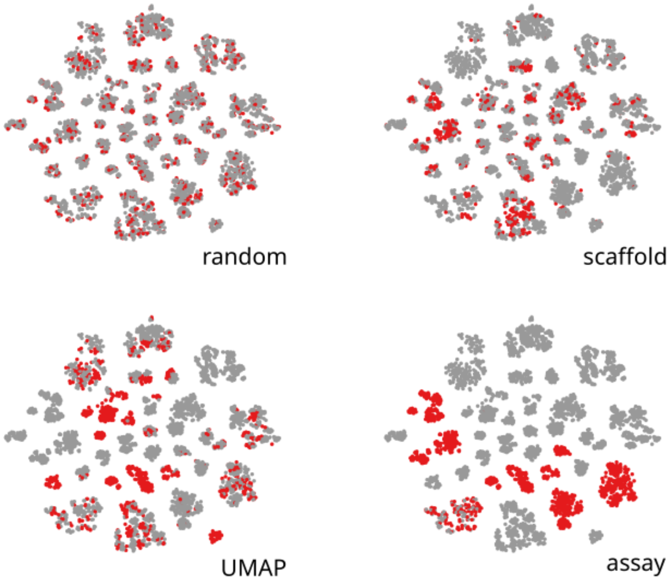
No standard scaffold-finding algorithm

Scaffolds are not necessarily relevant to the mechanism of action

[2] Bemis, & Murcko. (1996). The properties of known drugs. 1. Molecular frameworks. *Journal of medicinal chemistry*, *39*(15), 2887-2893.
[3] *https://lifechemicals.com/screening-libraries/scaffolds-and-scaffold-based-compounds*
[4] *https://greglandrum.github.io/rdkit-blog/posts/2024-05-31-scaffold-splits-and-murcko-scaffolds1.html*

Fair predictive evaluation in SotA alternatives to random splitting in related works.

Similarity splitting = identify similarity groups of drugs (e.g., Taylor-Butina clustering, UMAP **[5]** or stratification **[6]** on features) and segregate same-group molecules in the train or test sets



T-SNE map of drugs **[7]**. 🔴 test ● train

Pros

Seems to be fairer than scaffold splits **[5]**

Cons

No standard dimension-reduction / clustering

No control on the distribution of similarities in the train and test set

[5] Guo, … & Ballester . (2024, September). Scaffold Splits Overestimate Virtual Screening Performance. In *International Conference on Artificial Neural Networks* (pp. 58-72). Cham: Springer Nature Switzerland.
[6] Farias, … & Bastos-Filho. (2020). Similarity Based Stratified Splitting: an approach to train better classifiers. *ArXiv*.
[7] Backenköhler et al. (2025). Assay-Based Machine Learning: Rethinking Evaluation in Drug Discovery. *ChemRxiv*.

Fair predictive evaluation in SotA alternatives to random splitting in related works.

SIMPD **[8]** = use a multi-objective genetic algorithm to mimic time-based-splitting, where assays produced within the same time frame are grouped together

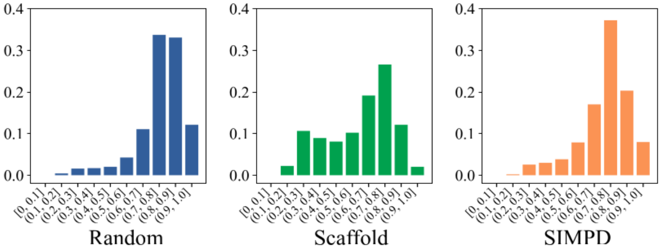
Candidate: Cluster drugs and assign clusters to train/test at random until |test| is 20% of all data
Fitness: Chemical requirements (e.g., median #heavy atoms)
Criteria: For train and test
Entropy < 0.9 x log2(#clusters)

- At each iteration, recombine the fittest candidates to produce new solutions
- Stop when a solution fits the criteria

Pros **[8]**

Less pessimistic than similarity-based splits

Cons



Adapted from **Fig. 4** in the paper. Sample similarity histograms b/w train and test sets.

Sometimes produces the same similarity distribution as random

[8] Landrum, ... & Riniker. (2023). SIMPD: an algorithm for generating simulated time splits for validating machine learning approaches. *Journal of cheminformatics*, *15*(1), 119.

Content of the paper

"SAE [...] a framework of **similarity aware evaluation** in which a novel split methodology is proposed to adapt to any desired distribution"

Objectives:

- Split drugs into train / test subsets according to their similarity
- "Controllable" and tractable approach even for larger data sets

Competitive advantages wrt SotA:

- Target similarity distribution is often uniform but SAE can reproduce the distribution in an external test set

Content of the paper

"SAE [...] a framework of **similarity aware evaluation** in which a novel split methodology is proposed to adapt to any desired distribution"

1. Target optimization problem
2. Tractable optimization algorithm for data splitting
3. Experimental results

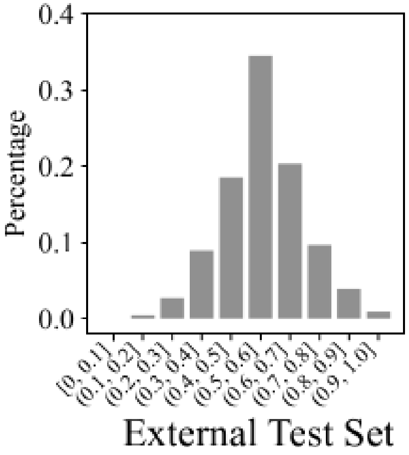
Target optimization problem find a test subset of size αN matching a target sample similarity histogram.

K = #similarity bins
N = #drugs in total

o^c_k = #drugs i with c_i=1 and in bin k

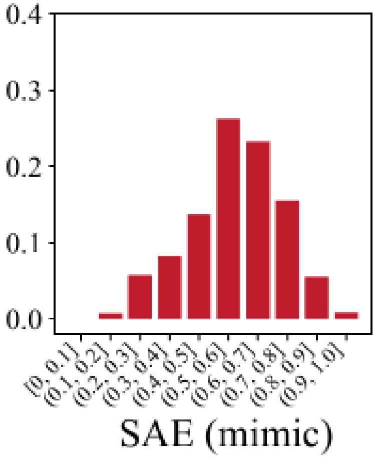
e_k = target %drugs in bin k
e.g., balanced would yield $e_k = 1/K$

Fig. 4 in paper.



$$\min_{\substack{\mathbf{c}\in\{0,1\}^N \\ |\mathbf{c}|^{1}=[\alpha N]}} \quad f(\mathbf{c}) = \sum_{k<K} (o^{\mathbf{c}}_k - \alpha N e_k)^2 \,/\, (\alpha N e_k)$$

< max similarity with training samples for each test sample



1. Compute $(e_k)_k$ from target histogram

2. Solve the minimization problem in **c**

3. Use the elements associated with 1's in **c** as testing subset

f(c) looks like the Pearson χ^2 statistic with K-1 degrees of freedom

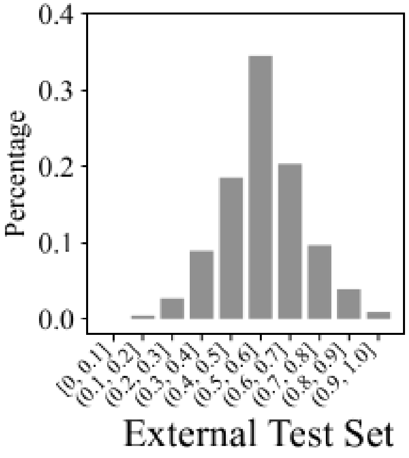
Target optimization problem find a test subset of size αN matching a target sample similarity histogram.

K = #similarity bins
N = #drugs in total

o^c_k = #drugs i with c_i=1 and in bin k

e_k = target %drugs in bin k
e.g., balanced would yield $e_k = 1/K$

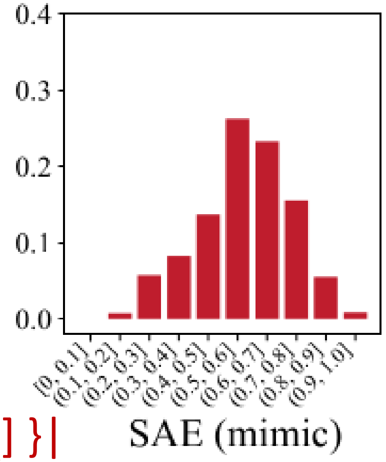
Fig. 4 in paper.



$$\min_{\substack{\mathbf{c}\in\{0,1\}^N \\ |\mathbf{c}|^1=[\alpha N]}} f(\mathbf{c}) = \sum_{k<K} (o^{\mathbf{c}}_k - \alpha N e_k)^2 / (\alpha N e_k)$$

$|\{\ i \mid c_i=1$ and $\max_{cj=0} \mathrm{sim}(i,j) \in [\mathrm{bin}_{k-1}, \mathrm{bin}_k]\ \}|$

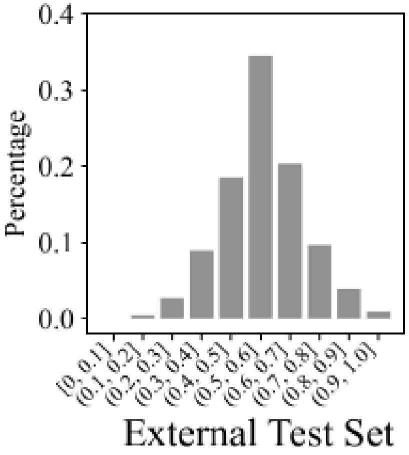
< max similarity with training samples for each test sample



1. Compute $(e_k)_k$ from target histogram
2. ~~Solve the minimization problem in **c**~~
3. Use the elements associated with 1's in **c** as testing subset

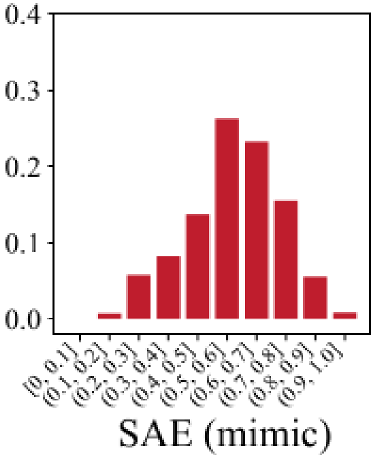
Tractable optimization algorithm for data splitting relax the problem to avoid integer prog and non-diff functions.

K = #similarity bins
N = #drugs in total



$$\min_{\substack{\mathbf{w}\in[0,1]^N \\ |\mathbf{w}|^1=[\alpha N]}} \quad f(\mathbf{w}) = \sum_{k<K} (o^{\mathbf{w}}_k - \alpha N e_k)^2 / (\alpha N e_k) + \lambda \ell_{\text{entropy}}(\mathbf{w})$$

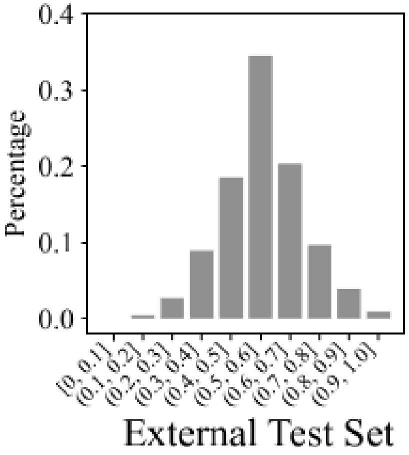
Fig. 4 in paper.



Tractable optimization algorithm for data splitting relax the problem to avoid integer prog and non-diff functions.

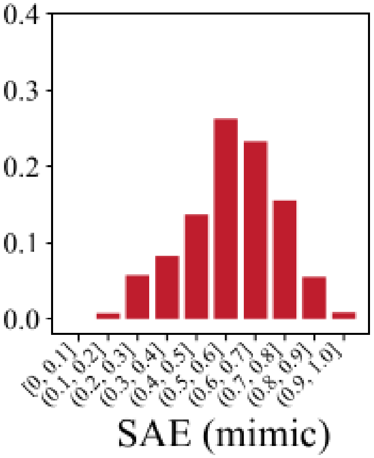
K = #similarity bins
N = #drugs in total

Fig. 4 in paper.



$$\min_{\substack{\mathbf{w}\in[0,1]^N \\ |\mathbf{w}|^1=[\alpha N]}} \quad f(\mathbf{w}) = \sum_{k<K} (o^{\mathbf{w}}_k - \alpha N e_k)^2 / (\alpha N e_k) + \lambda \ell_{\text{entropy}}(\mathbf{w})$$

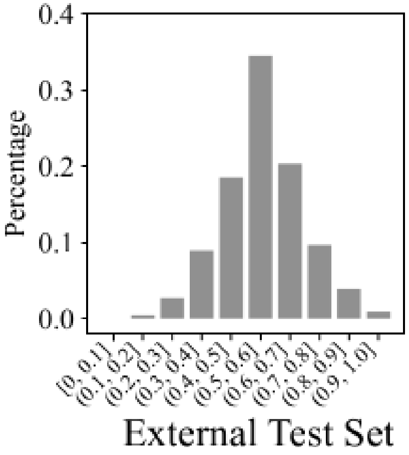
$$\approx |\{\, i \mid c_i=1 \text{ and } \max_{c_j=0} \text{sim}(i,j) \in [\text{bin}_{k-1}, \text{bin}_k] \,\}|$$



Tractable optimization algorithm for data splitting relax the problem to avoid integer prog and non-diff functions.

K = #similarity bins
N = #drugs in total

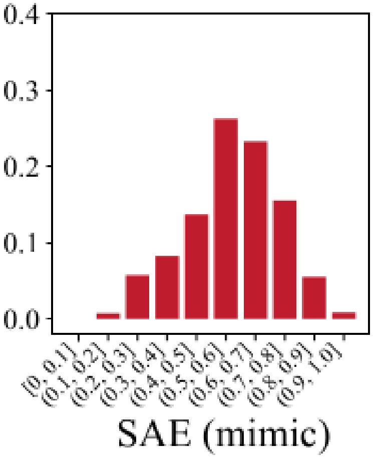
Fig. 4 in paper.



$$\min_{\substack{\mathbf{w}\in[0,1]^N \\ |\mathbf{w}|^1=[\alpha N]}} \quad f(\mathbf{w}) = \sum_{k<K} (o^{\mathbf{w}}_k - \alpha N e_k)^2 / (\alpha N e_k) + \lambda \ell_{\text{entropy}}(\mathbf{w})$$

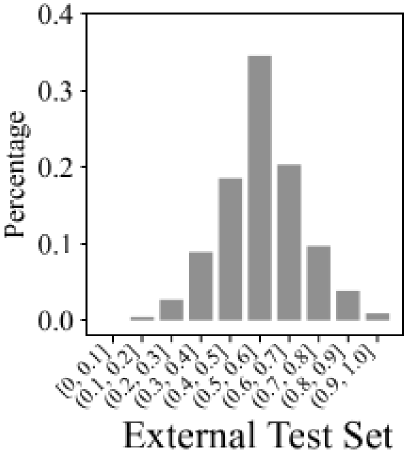
$$\approx |\{ i \mid c_i=1 \text{ and } \max_{cj=0} \text{sim}(i,j) \in [\text{bin}_{k-1}, \text{bin}_k] \}|$$

$$\approx \sum_{i<N} w_i \, I(\max_{cj=0} \text{sim}(i,j) \in [\text{bin}_{k-1}, \text{bin}_k])$$



Tractable optimization algorithm for data splitting relax the problem to avoid integer prog and non-diff functions.

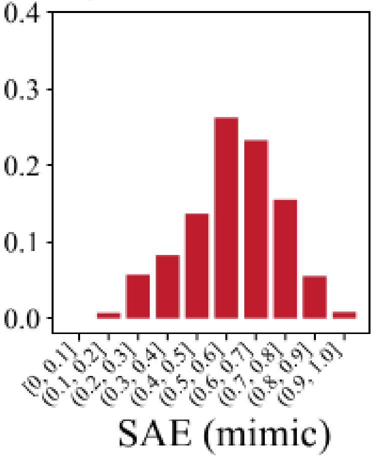
K = #similarity bins
N = #drugs in total



$$\min_{\substack{\mathbf{w}\in[0,1]^N \\ |\mathbf{w}|^1=[\alpha N]}} \quad f(\mathbf{w}) = \sum_{k<K} (o^{\mathbf{w}}_k - \alpha N e_k)^2 / (\alpha N e_k)$$

$$+ \lambda \ell_{entropy}(\mathbf{w})$$

Fig. 4 in paper.



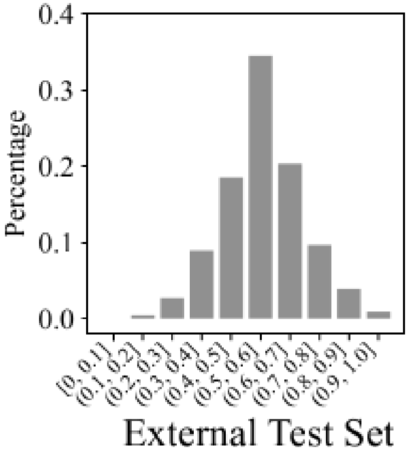
$\approx |\{ i \mid c_i=1 \text{ and } \max_{cj=0} \text{sim}(i,j) \in [bin_{k-1}, bin_k] \}|$

$\approx \sum_{i<N} w_i \, I(\max_{cj=0} \text{sim}(i,j) \in [bin_{k-1}, bin_k])$

$\approx \sum_{i<N} w_i \, I(\max_j (1-w_j)\text{sim}(i,j) \in [bin_{k-1}, bin_k])$

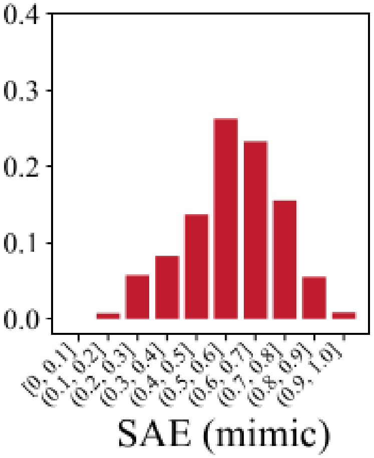
Tractable optimization algorithm for data splitting relax the problem to avoid integer prog and non-diff functions.

K = #similarity bins
N = #drugs in total



$$\min_{\substack{\mathbf{w}\in[0,1]^N \\ |\mathbf{w}|^1=[\alpha N]}} \quad f(\mathbf{w}) = \sum_{k<K} (o^{\mathbf{w}}_k - \alpha N e_k)^2 / (\alpha N e_k) + \lambda \ell_{\text{entropy}}(\mathbf{w})$$

Fig. 4 in paper.



$$\approx |\{ i \mid c_i=1 \text{ and } \max_{cj=0} \operatorname{sim}(i,j) \in [\text{bin}_{k-1}, \text{bin}_k] \}|$$

$$\approx \sum_{i<N} w_i \, I(\max_{cj=0} \operatorname{sim}(i,j) \in [\text{bin}_{k-1}, \text{bin}_k])$$

$$\approx \sum_{i<N} w_i \, I(\max_j (1-w_j)\operatorname{sim}(i,j) \in [\text{bin}_{k-1}, \text{bin}_k])$$

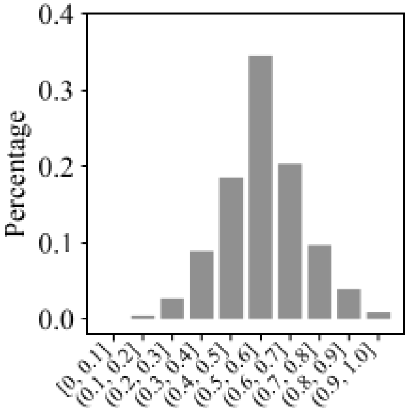
$$\approx \sum_{i<N} w_i \, I(\operatorname{LogSumExp}((1-\mathbf{w}) \text{ x } \mathbf{sim(i,.)}) \in [\text{bin}_{k-1}, \text{bin}_k])$$

$$\operatorname{LogSumExp}(\boldsymbol{x}) = \beta^{-1} \log\Big(\sum_{i<N} \exp(\beta x_i)\Big)$$
= multivariate SoftPlus

$$\max_i x_i \leq \operatorname{LogSumExp}(\boldsymbol{x}) \leq \max_i x_i + \underbrace{\log(N)/\beta}_{\text{Controls the accuracy}}$$

Tractable optimization algorithm for data splitting relax the problem to avoid integer prog and non-diff functions.

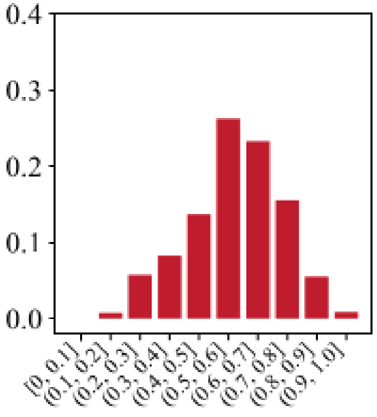
K = #similarity bins
N = #drugs in total



External Test Set

$$\min_{\substack{\mathbf{w}\in[0,1]^N \\ |\mathbf{w}|^1=[\alpha N]}} \quad f(\mathbf{w}) = \sum_{k<K} (o^{\mathbf{w}}_k - \alpha N e_k)^2 / (\alpha N e_k) + \lambda \ell_{\text{entropy}}(\mathbf{w})$$

Fig. 4 in paper.



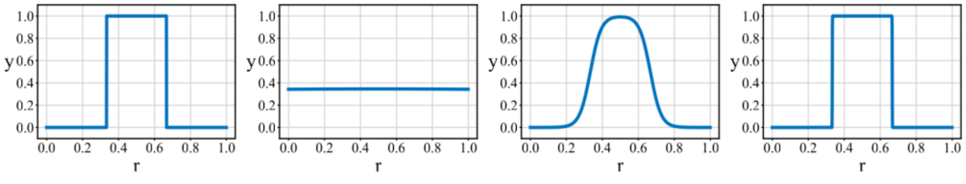
SAE (mimic)

$$\approx |\{ i \mid c_i=1 \text{ and } \max_{c_j=0} \text{sim}(i,j) \in [\text{bin}_{k-1}, \text{bin}_k] \}|$$

$$\approx \sum_{i<N} w_i \, \mathbb{I}(\max_{c_j=0} \text{sim}(i,j) \in [\text{bin}_{k-1}, \text{bin}_k])$$

$$\approx \sum_{i<N} w_i \, \mathbb{I}(\max_j (1-w_j)\text{sim}(i,j) \in [\text{bin}_{k-1}, \text{bin}_k])$$

$$\approx \sum_{i<N} w_i \, \mathbb{I}(\underbrace{\text{LogSumExp}\big((1-\mathbf{w}) \times \textbf{sim(i,.)}\big)}_{= r_i} \in [\text{bin}_{k-1}, \text{bin}_k])$$



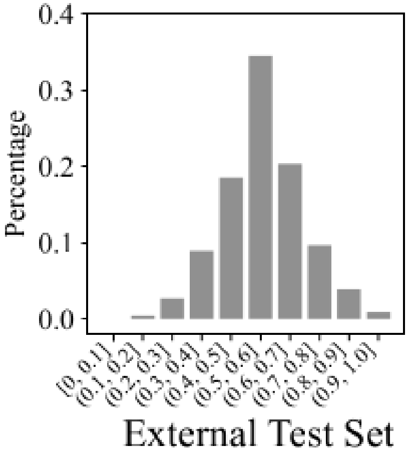
(a) $\mathbb{I}(\frac{1}{3} < r \leq \frac{2}{3})$ (b) $\sigma = 1$ (c) $\sigma = 0.1$ (d) $\sigma = 0.01$

Assume Prob($r_i \in [\text{bin}_{k-1}, \text{bin}_k]$) = Norm($r_i$; c_k, σ_k) where $c_k = (\text{bin}_k + \text{bin}_{k-1})/2$ is the center of the bin and σ_k controls the accuracy

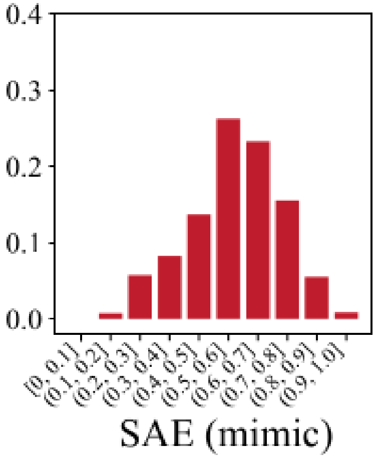
Tractable optimization algorithm for data splitting relax the problem to avoid integer prog and non-diff functions.

K = #similarity bins
N = #drugs in total

Fig. 4 in paper.



$$\min_{\substack{\mathbf{w}\in[0,1]^N \\ |\mathbf{w}|^1=[\alpha N]}} \quad f(\mathbf{w}) = \sum_{k<K} (o^{\mathbf{w}}_k - \alpha N e_k)^2 / (\alpha N e_k) + \lambda \ell_{entropy}(\mathbf{w})$$

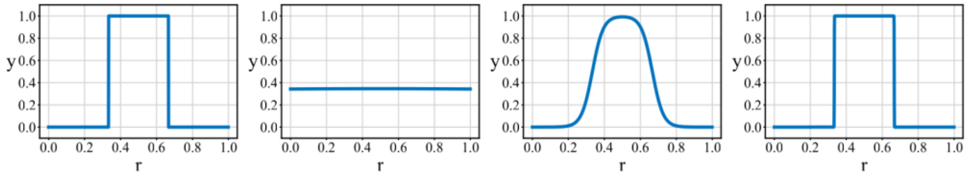


$$\approx |\{ i \mid c_i=1 \text{ and } \max_{c_j=0} sim(i,j) \in [bin_{k-1}, bin_k] \}|$$

$$\approx \sum_{i<N} w_i \, I(\max_{c_j=0} sim(i,j) \in [bin_{k-1}, bin_k])$$

$$\approx \sum_{i<N} w_i \, I(\max_j (1-w_j) sim(i,j) \in [bin_{k-1}, bin_k])$$

$$\approx \sum_{i<N} w_i \, I(\underbrace{\text{LogSumExp}\big((1-\mathbf{w}) \times \mathbf{sim(i,.)}\big)}_{= r_i} \in [bin_{k-1}, bin_k])$$



$(a)\ \mathbb{I}(\frac{1}{3} < r \leq \frac{2}{3})$ $(b)\ \sigma = 1$ $(c)\ \sigma = 0.1$ $(d)\ \sigma = 0.01$

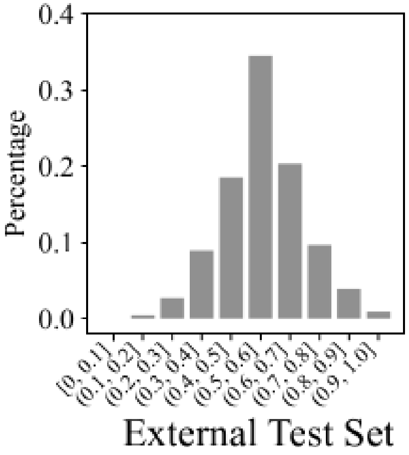
Assume $\text{Prob}(r_i \in [bin_{k-1}, bin_k]) = \text{Norm}(r_i ; c_k, \sigma_k)$
where $c_k = (bin_k + bin_{k-1})/2$ is the center of the bin
and σ_k controls the accuracy

$$\propto \sum_{i<N} w_i \, \text{Norm}(r_i ; c_k, \sigma_k)$$

Tractable optimization algorithm for data splitting relax the problem to avoid integer prog and non-diff functions.

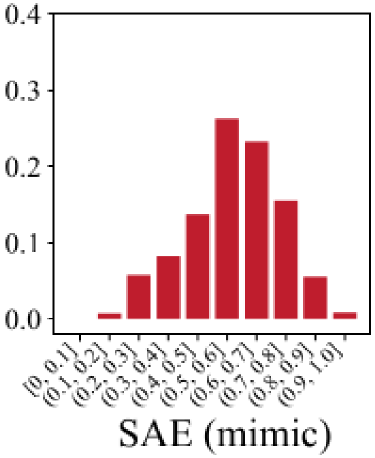
K = #similarity bins
N = #drugs in total

e_k = target %drugs in bin k
e.g., balanced would yield $e_k = 1/K$ Fig. 4 in paper.



$$\min_{\substack{\mathbf{w}\in[0,1]^N \\ |w|^1=[\alpha N]}} f(\mathbf{w}) = \sum_{k<K} (o^{\mathbf{w}}_k - \alpha N e_k)^2 / (\alpha N e_k) + \lambda \ell_{\text{entropy}}(\mathbf{w})$$

$$\propto \sum_{i<N} w_i \,\text{Norm}(r_i\,;\, c_k, \sigma_k)$$

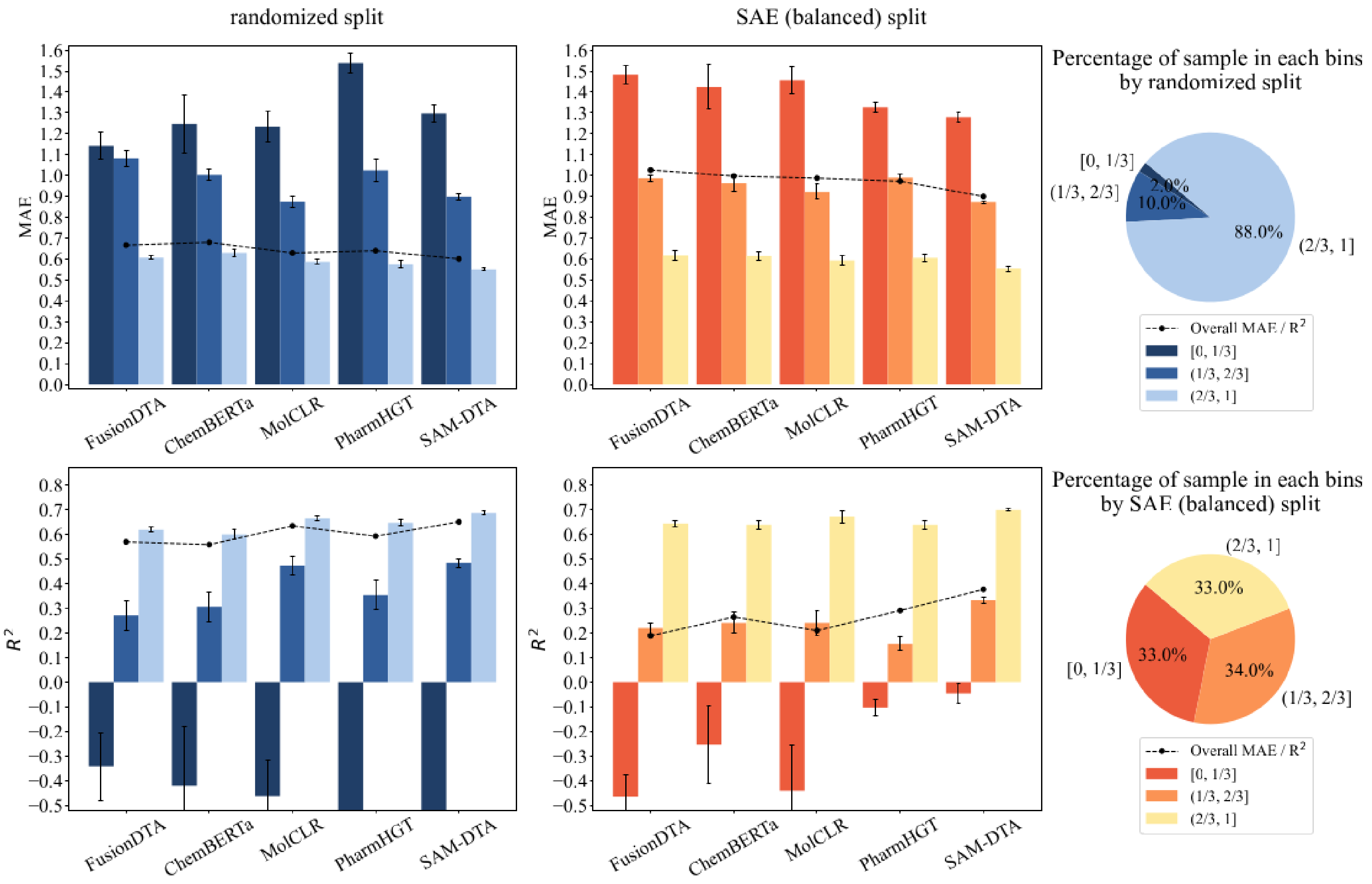


1. Compute $(e_k)_k$ from target histogram

2. Solve the minimization problem in **w** with standard numerical approaches for convex optimization

3. Use the elements associated with 1's in **w** as testing subset

Experimental results Fairer evaluation and balanced similarity between the train and test subsets.

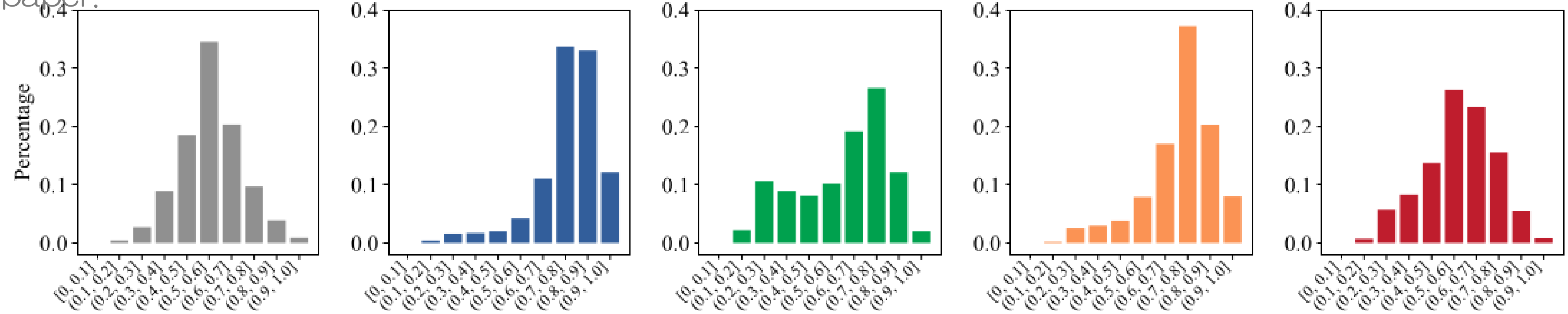


for all bin k, e_k=1/K

Fig. 1 in the paper.

Experimental results Reproduce similar conditions as in a target testing subset or condition ("any distribution").

Fig. 4 in the paper.



(a) External Test Set

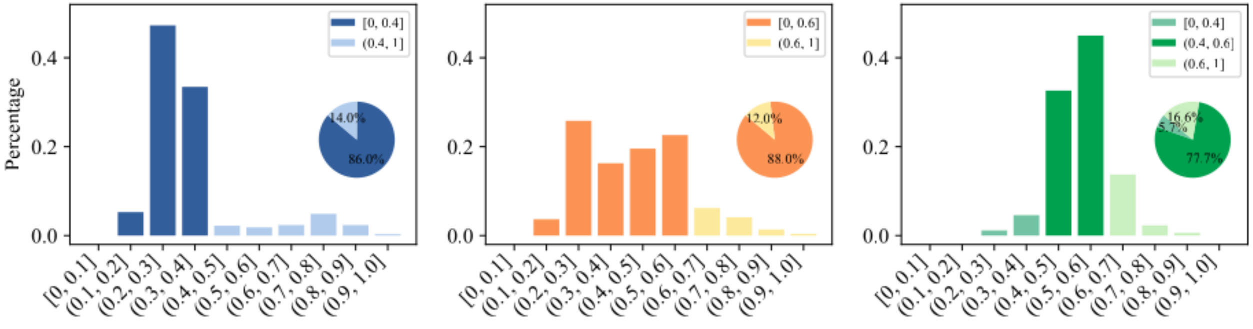
(b) Random

(c) Scaffold

(d) SIMPD

(e) SAE (mimic)

Custom $(e_k)_k$ from target histogram



(a) Test data distribution with 0~0.4 split

(b) Test data distribution with 0~0.6 split

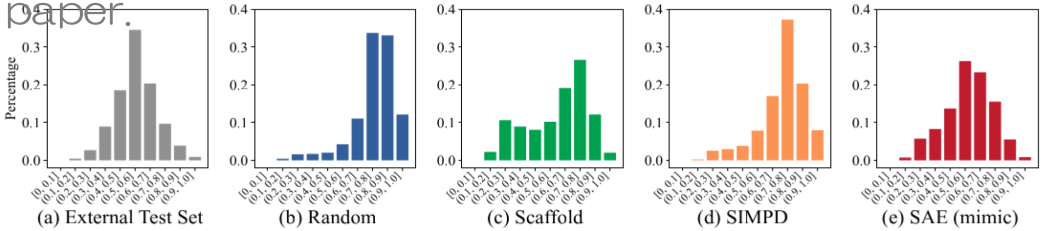
(c) Test data distribution with 0.4~0.6 split

Constraint the range of train-test sample similarities

Fig. 6 in the paper.

Experimental results Reproduce similar conditions as in a target testing subset or condition ("any distribution").

Fig. 4 in the paper.



(a) External Test Set (b) Random (c) Scaffold (d) SIMPD (e) SAE (mimic)

Custom $(e_k)_k$ from target histogram

More reproducible performance on external test sets

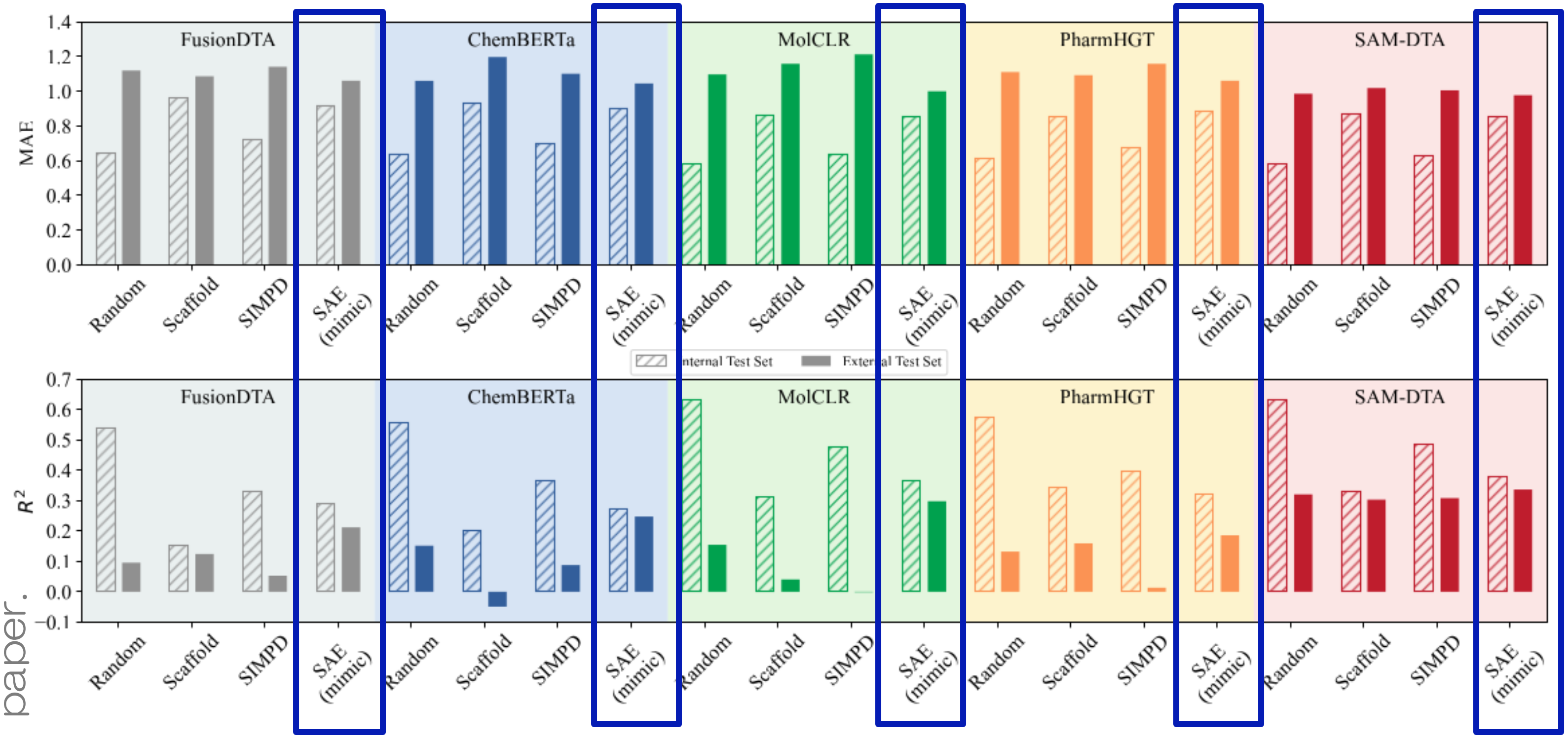


Fig. 5 in the paper.

Perspectives

1. Comments on the paper
2. Why is it interesting for BioComp?

My comments on the paper

Strengths:

- Paper is well-written
- Topic is interesting and their experiments on random splits are useful
- Algorithm is flexible and computationally efficient (even for large data sets)

Weaknesses:

- Experimental results on the "mimic" (unbalanced target distribution) are not impressive (Fig. 4 and 6)
- Does not address three-way splitting (training + testing + validation) but it is discussed in the OpenReview page **[9]**

Your comments?

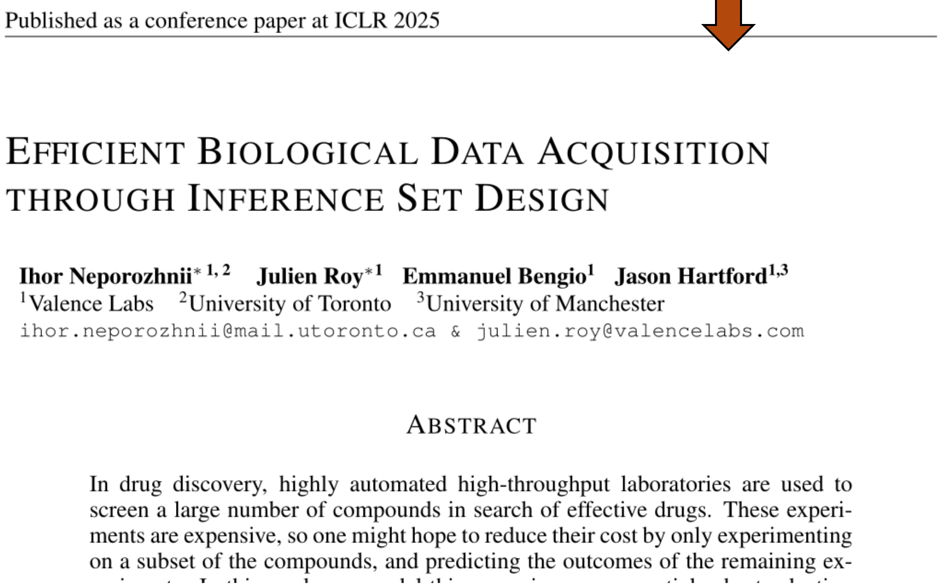
[9] *https://openreview.net/forum?id=j7cyANIAxV*

Why is it interesting for BioComp? Fairer evaluation, model generalizability = better understanding.

Especially for biological data: random splits might be tricky

→ Need to remove / balance out confounders for the target outcome (like in clinical trials!)

Might be connected to active learning in biology: a careful selection of the training set is (iteratively) done (because the training phase is expensive or because data is scarce)



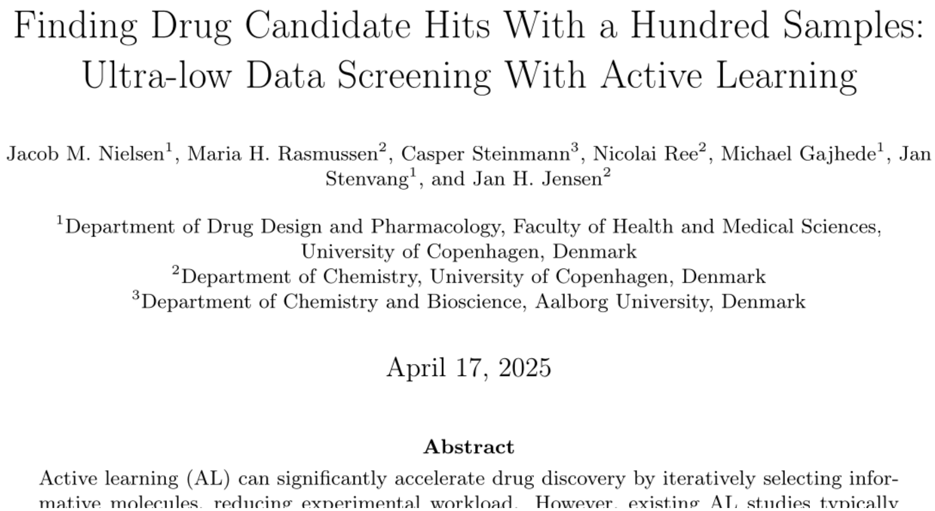
Published as a conference paper at ICLR 2025

EFFICIENT BIOLOGICAL DATA ACQUISITION THROUGH INFERENCE SET DESIGN

Ihor Neporozhnii*[1,2] Julien Roy*[1] Emmanuel Bengio[1] Jason Hartford[1,3]
[1]Valence Labs [2]University of Toronto [3]University of Manchester
ihor.neporozhnii@mail.utoronto.ca & julien.roy@valencelabs.com

ABSTRACT

In drug discovery, highly automated high-throughput laboratories are used to screen a large number of compounds in search of effective drugs. These experiments are expensive, so one might hope to reduce their cost by only experimenting on a subset of the compounds, and predicting the outcomes of the remaining ex-



Finding Drug Candidate Hits With a Hundred Samples: Ultra-low Data Screening With Active Learning

Jacob M. Nielsen[1], Maria H. Rasmussen[2], Casper Steinmann[3], Nicolai Ree[2], Michael Gajhede[1], Jan Stenvang[1], and Jan H. Jensen[2]

[1]Department of Drug Design and Pharmacology, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark
[2]Department of Chemistry, University of Copenhagen, Denmark
[3]Department of Chemistry and Bioscience, Aalborg University, Denmark

April 17, 2025

Abstract

Active learning (AL) can significantly accelerate drug discovery by iteratively selecting informative molecules, reducing experimental workload. However, existing AL studies typically