Towards a large-scale benchmark of collaborative filtering in drug repurposing (#507716)

[1] Funded by the European Union — Systems Biology Bioinformatics Rostock

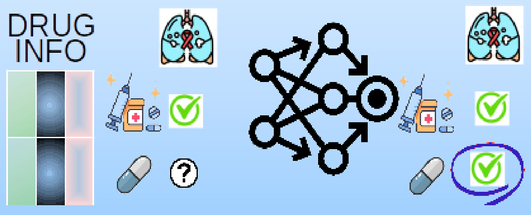
[2] Soda, Inria

[3] Universität Rostock, Traditio et Innovatio; Leibniz Institute for Food Systems Biology at the Technical University of Munich; stias Stellenbosch Institute for Advanced Study

Clémence Réda[1], Jill-Jênn Vie[2], Olaf Wolkenhauer[3]

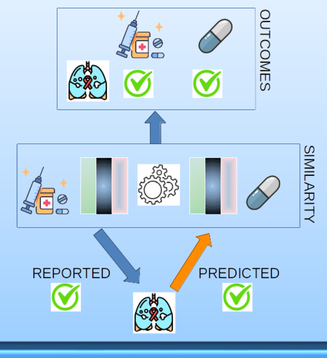
Drug development is expensive, prone to high failure rate in commercialization. Incentives tend to focus on profitable diseases, which penalizes rare / tropical neglected disease research. [1]

➡ **Drug repurposing** screens documented molecules in a systematic way to uncover new therapeutic ("positive") drug-disease associations



Yet • there is a large imbalance of outcomes between known drug-disease associations
• there is *implicit* information to exploit

➡ **Collaborative filtering (CF)** filters for patterns in associations by implementing collaboration across entities (ex. drugs, diseases)
Returns a matrix Â of drug-disease pairs



I. Standardized, reproducible datasets and pipelines to evaluate drug repurposing models

Two new reproducible datasets from biological data

Dataset	Data type	#drug	#drug features	#disease	#disease features	#positive (negative)
TRANSCRIPT [2]	Gene expression	204	12,096	116	12,096	401 (11)
PREDICT [3]	Chemical, Transcript., ...	1,351	6,265	1,066	2,914	5,624 (152)

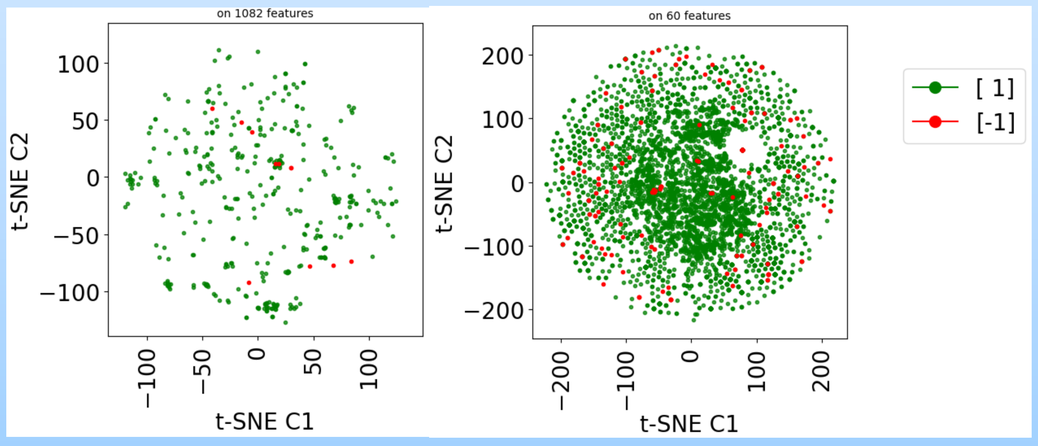


Fig. 1. *t*-SNE plots of TRANSCRIPT (left) and PREDICT.

Two Python packages to enable benchmarking [4]

- **stanscofi** automates data processing, model training and evaluation
- **benchscofi** implements ~20 state-of-the-art **CF** algorithms

Benchmark on 6 datasets and 11 algorithms

- **Datasets** 1 synthetic (**S**), 2 text-mining (**T**), 4 biological data-based (**B**)
- **Algorithms** 5 matrix factorization (**M**), 3 neural networks (**N**), 3 graph-based (**G**)

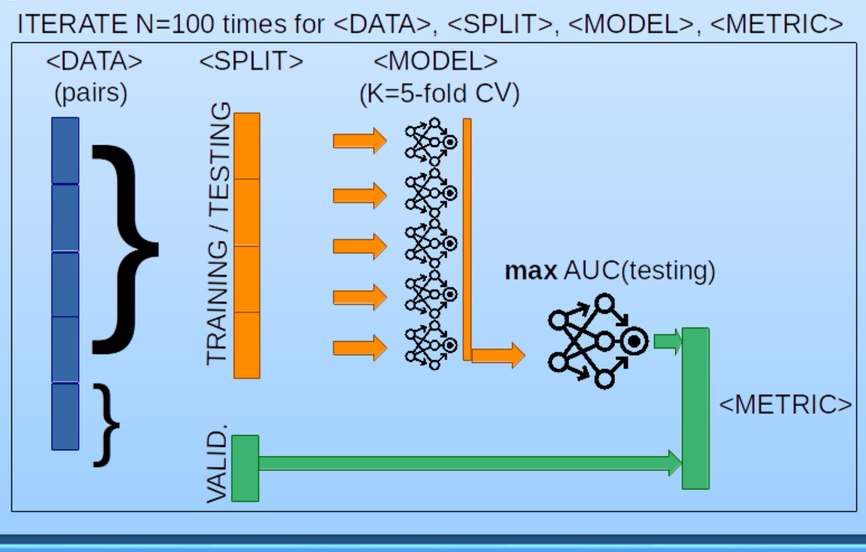


Fig. 2. Pipeline for fixed dataset, data splitting, algorithm, validation metric, iterated 100 times.

II. Guidelines: Q1. Which metric? Q2. Which dataset? Q3. How to measure the generalization error?

Q1. Choice of validation metric

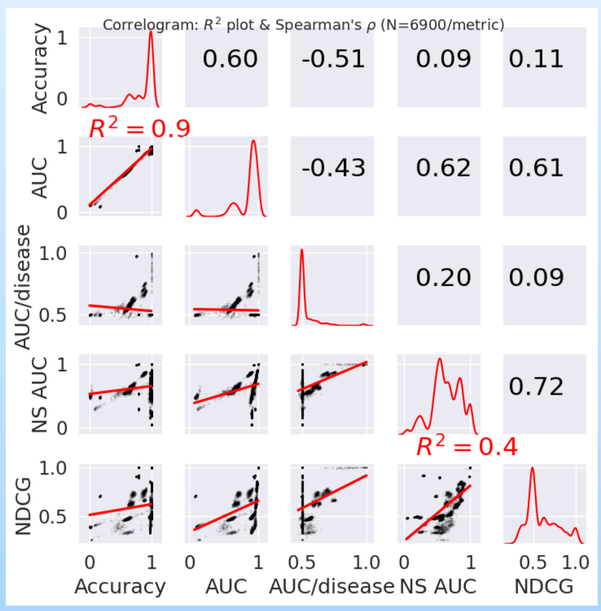
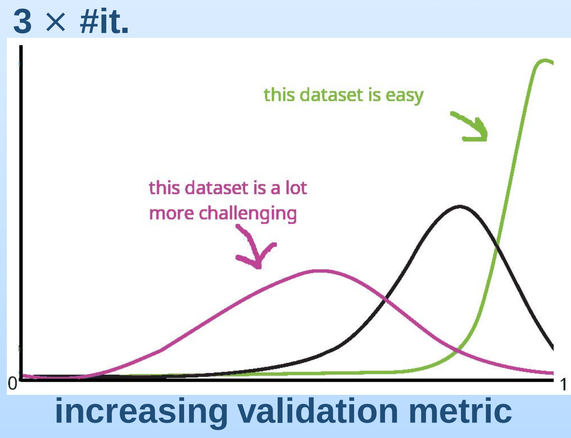


Fig. 4. Correlation plot on metrics
NS AUC [5] / disease *d*
$\propto \sum_{(m,d)>0} \sum_{(m',d)<0} \mathbb{1}(\hat{A}[m,d] \geq \hat{A}[m',d])$

Q2. Choice of a dataset (with μ=NS-AUC)

- is it a challenging one?

3 × #it.



Median on Top-3 algs, 100 iterations		μ
Synthetic	(S)	1.00
Gottlieb	(B)	0.87
LRSSL	(B)	0.87
Cdataset	(T)	0.86
Fdataset	(T)	0.83
PREDICT	(B)	0.78
TRANSCRIPT	(B)	0.70

- are features useful for classification?

for each dataset, test H_0 : $\mu_{\text{alg w/ feat.}} = \mu_{\text{alg w/o feat}}$
with a Kruskal-Wallis H-test, α=1%, N_{feat}=600, $N_{w/o}$=500

Yes for all datasets but the synthetic one (which makes sense).

Q3. Approximation of the generalization error

➡ a non-random "cheap" data splitting method for maximizing dissimilarity b/w training & validation [6]

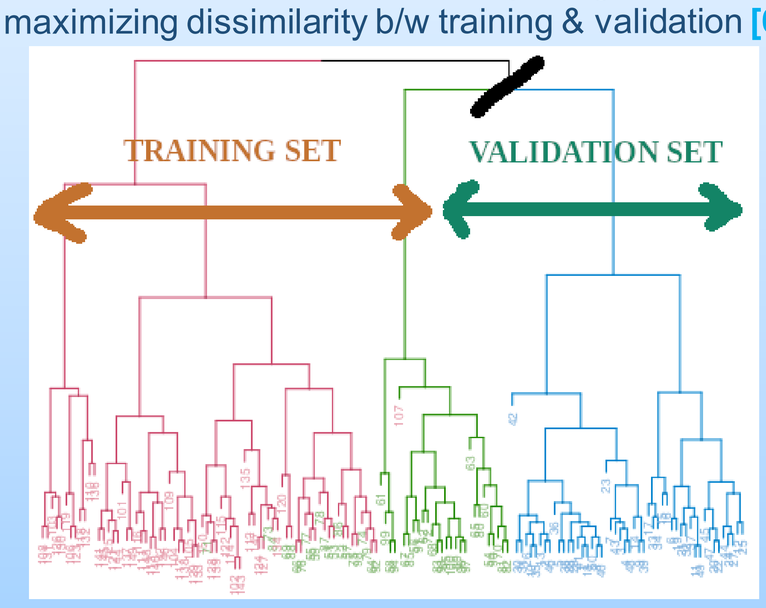
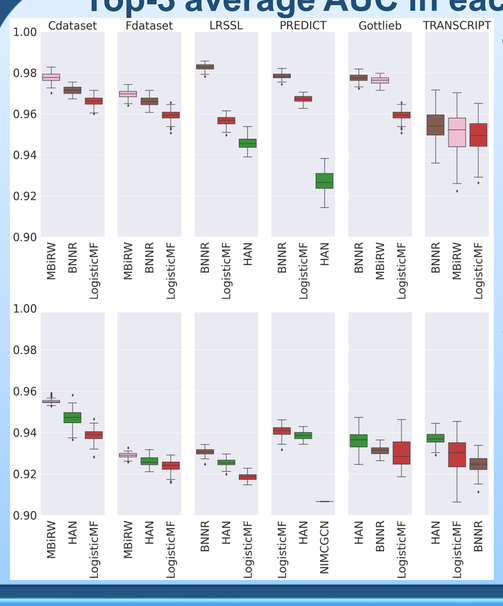


Fig. 5. *Weakly correlated* training / validation sets from the dendrogram computed on drugs

III. Benchmark results: approximation and generalization errors

Top-3 average AUC in each dataset

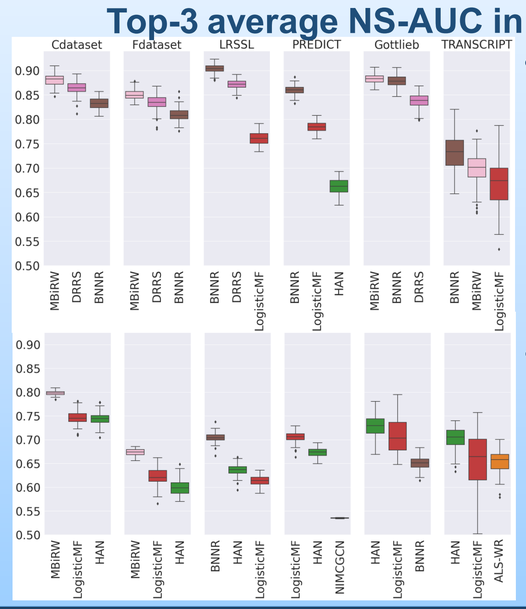


- randomly split sets ≈ approximation error
- weakly correlated sets ~ generalization error

Algorithms (types) among Top-3:

ALS-WR	(M)	LogisticMF	(M)
BNNR	(G)	MBiRW	(G)
DRRS	(M)	NIMCGCN	(N)
HAN	(N)		

Top-3 average NS-AUC in each dataset



- randomly split sets ≈ approximation error
- weakly correlated sets ~ generalization error

- is there a clear winner?

BNNR [7] is almost in all Top-3
➡ future papers should try to beat it!

- is a type of method: (M), (G) or (N) consistently better?

for each dataset, test
$H_0 : \mu_{(M).} = \mu_{(G)} = \mu_{(N)}$
Kruskal-Wallis H-test, α=1%
$N_{(M)}$=500, $N_{(G)}$=300, $N_{(N)}$=300

➡ graph-based (G) are better!

Discussion

Three novel contributions: • richer, larger datasets, • standardized evaluation
• medium-scale reproducible benchmark

➡ ensure a fair assessment of the technological improvement by a method
➡ yield a healthier ecosystem and easier development of drug repurposing

[1] Philippidis. (2023). DOI: 10.1089/genedge.5.1.39
[2]TRANSCRIPT. DOI: 10.5281/zenodo.7982976
[3] PREDICT. DOI: 10.5281/zenodo.7983090
[4] C.R., J.-J. V., O.W. (2024). DOI: 10.21105/joss.05973
[5] Yu, Bilenko, Lin. DOI: 10.1137/1.9781611974973
[6] Chekroud et al. DOI: 10.1126/science.adg8538
[7] Yang et al. DOI: 10.1093/bioinformatics/btz331

GitHub benchmark code repository

clemence.reda@uni-rostock.de
https://recess-eu-project.github.io