JELI: an interpretable embedding-learning recommender system for drug repurposing #129

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

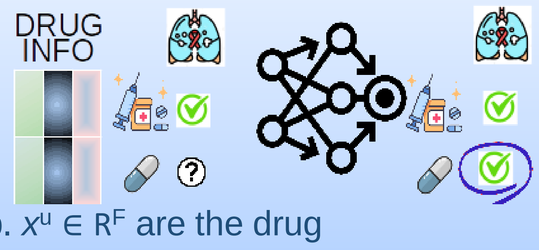
1 Funded by the European Union — SYSTEMS BIOLOGY BIOINFORMATICS ROSTOCK

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

Drug development is expensive, prone to high failure rate in commercialization.

➡ **Drug repurposing** screens documented molecules to uncover therapeutic ("positive") drug-disease associations from unknown pairs



GOAL Learn classifier $C(x^i,x^u) \in \{0, 1\}$, x^i, resp. $x^u \in R^F$ are the drug (resp. disease) feature vectors of size F

➡ **Interpretability** can be obtained by feature-attribution methods [1-2]
$\varphi^C(x^i,x^u) \in R^F$ where $\varphi^C(x^i,x^u)_j$ is the importance of feature j, $j \leq F$

But *post hoc* approaches might lead to unreliable interpretations [3-4]

QUESTION Build a generic (drug repurposing) recommender system with embedded importance scores

Methods – Joint Embedding Learning-classifier for improved Interpretability (JELI)

CLASSIFIER A linear **Redundant Higher-Order Factorization Machine*** of dimension d and order m=2 has parameters $\omega^0 \in R$, $\omega^1 \in R^d$, $\omega^2 \in R$, $W \in R^{Fxd}$

$$C(x^i,x^u) \triangleq \mathbb{1}[\,RHOFM(x^i, x^u) > 0.5\,]$$

and $RHOFM(x^i, x^u) \triangleq \underbrace{\omega^0 + \omega^{1T}(x^i + x^u)W}_{\text{linear regression}} + \omega^2 \underbrace{\Sigma_{f<f' \leq 2F} \langle W_{f\%F,:}, W_{f'\%F,:} \rangle x^i_{f\%F}\, x^u_{f'\%F}}_{\text{pairwise interaction term for (f,f')}}$

* can be defined for any structure and order m>1

EMBEDDING $e^j = W_{j,:}$ is *also* the d-dimensional embedding of feature j whereas $e^h = x^hW$ (linear structure) for h a drug or a disease

KNOWLEDGE GRAPH PRIOR G(V, T) where drugs, diseases and features are included in V, and T contains edges (h, r, t) where h, t ∈ V and r ∈ {+, – , …}
+ (resp. –) for positive (resp. negative) drug-disease pairs

JOINT LEARNING It should minimize the soft margin ranking loss

$$L(\omega^0, \omega^1, \omega^2, W) \triangleq \underbrace{\Sigma_{(h, r, t) \in T}}_{\text{edges in prior}} \underbrace{\Sigma_{(\underline{h}, \underline{r}, \underline{t}) \notin T}}_{\text{edges } \textit{not} \text{ in prior}} \log(1+\exp(1+score(h, r, t)-score(\underline{h}, \underline{r}, \underline{t})))$$

where $score(h, r, t) \triangleq RHOFM(x^h, x^t)$ if r = + else $-RHOFM(x^h, x^t)$ if r = –
else $MuRE(e^h, e^r, e^t)$ [5]

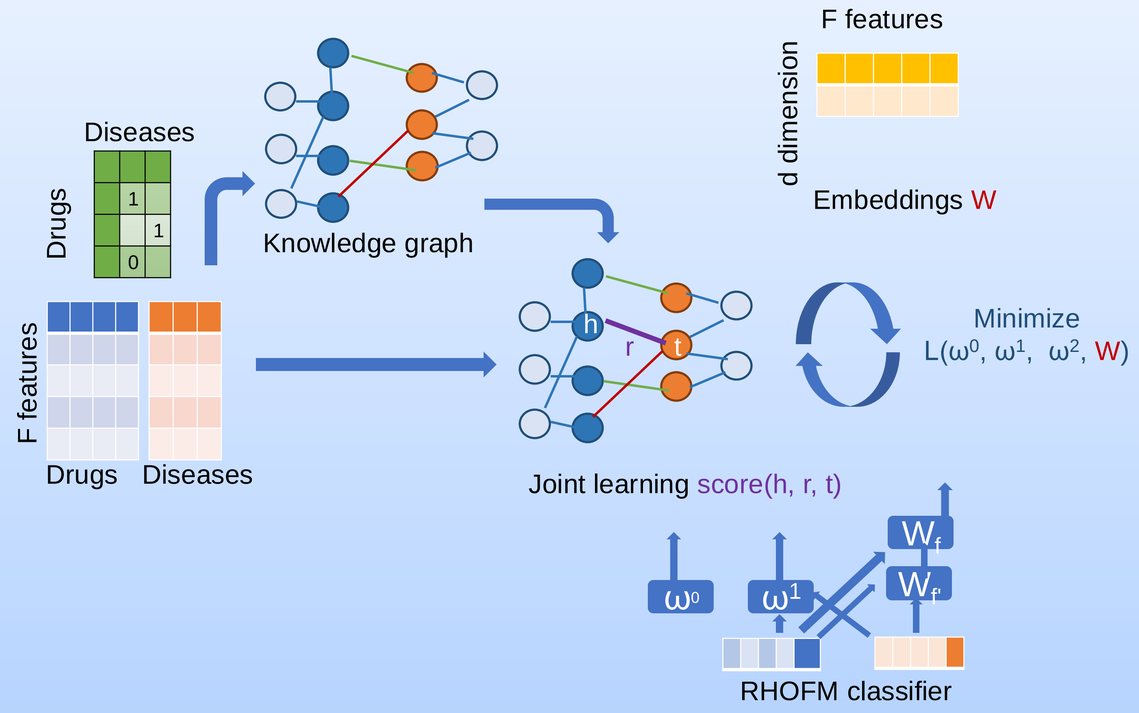


Fig. 1. Architecture of JELI for drug repurposing.

Results – Interpretability and drug repurposing performance

Predict a score for a drug-disease pair (h, t) – $score(h, +, t) = RHOFM(x^h, x^t)$

Compute an embedding for a new drug / disease h – for a linear RHOFM $e^h = x^hW$

Feature-wise importance scores – for a linear RHOFM
$\varphi^C_f = \Sigma_{k \leq d} W_{f,k}$ for feature $f \leq F$

➡ **JELI reliably retrieves ground truth importance scores and is robust in synthetic data sets**

Interpretable "deviated" models with sparsity number $s \in (0,1)$ = % of unknown drug-disease pairs

$x^i , x^u , W^* \sim \mathcal{N}(0, 0.1)$
$C^*(x^i,x^u) \leftarrow MASK_s \circ \sigma((x^i+x^u)W^*)$

Average (avg) Spearman's ρ across data sets
$\rho(\Sigma_{k \leq d} W^*_{f,k}, \Sigma_{k \leq d} W_{f,k}) = 0.92$

on 10 data sets with F=10, d=2, $n^u=n^i=173$

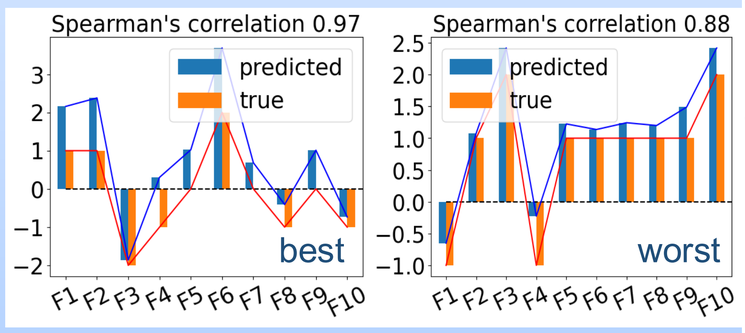


Fig. 2. Barplots of true and predicted feature importance scores in synthetic data sets.

Average AUC JELI 0.79 HAN 0.75 [6]
NS-AUC [7] $\propto \Sigma_{(h,+,t) \in T} \Sigma_{(\underline{h},-,t) \in T \text{ or } (\underline{h},\{-,+\},t) \notin T} \mathbb{1}(C(x^h,x^t) \geq C(x^{\underline{h}},x^t))$

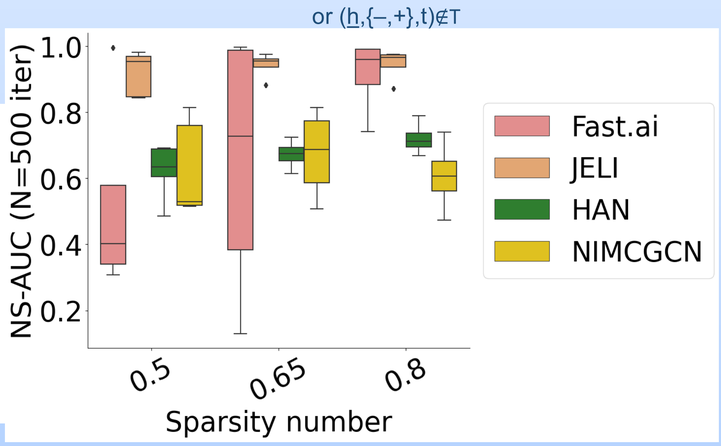


Fig. 3. Boxplots of NS-AUCs across synthetic data sets of variable sparsity numbers.

➡ **JELI is predictive on drug repurposing data sets**

Name	Year	#drugs	#diseases	Sparsity
Gottlieb	2016	593	313	98.96%
LRSSL	2017	763	681	99.41%
PREDICT-Gottlieb	2022	593	313	98.96%

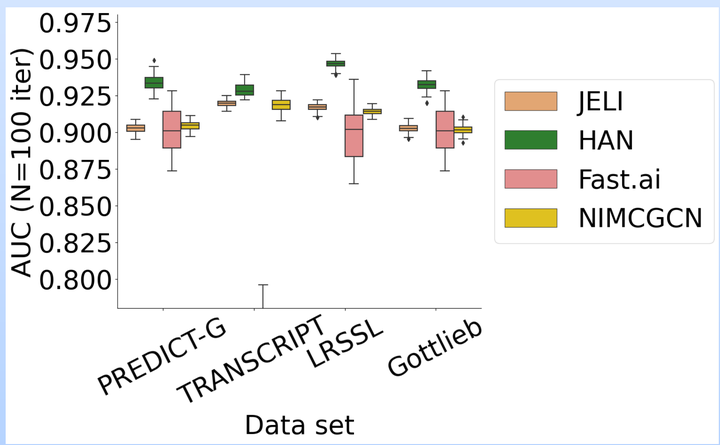


Fig. 4. Boxplots of AUCs of JELI and state-of-the-art across drug repurposing data sets.

➡ **JELI can integrate any graph prior**

TRANSCRIPT transcriptomic data set [8]

Name	F	#drugs	#diseases	Sparsity
TRANSCRIPT	12,096	204	116	98.26%

Sim (default)– drug-drug / disease-disease similarities, drug/disease-gene connections

Sim+PPI– Sim connections + gene-gene connections using STRING [9]-extracted protein-protein interaction networks in humans

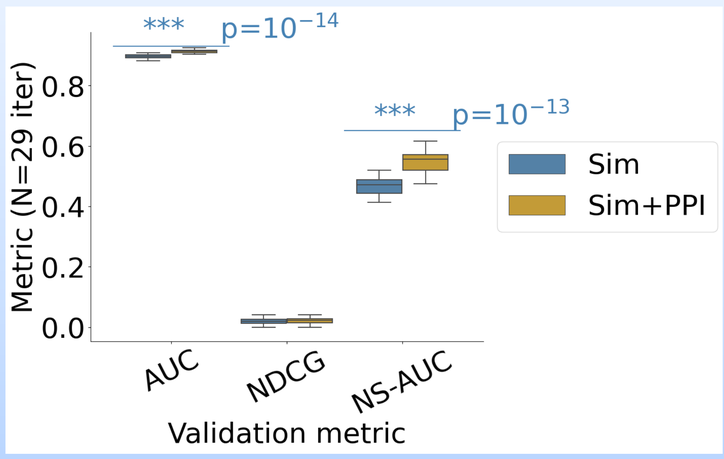


Fig. 5. Boxplots of AUC, NS-AUC and NDCG scores in JELI depending on the graph prior.

➡ **Ablation study: structure & joint learning are crucial**

FMs without structure
FM 2nd order factorization machine
CrossFM on only drug-disease terms
Separate RHOFM/Embedding (SELT)
SELT (PCA*) PCA embeddings
SELT (KGE) KG embeddings

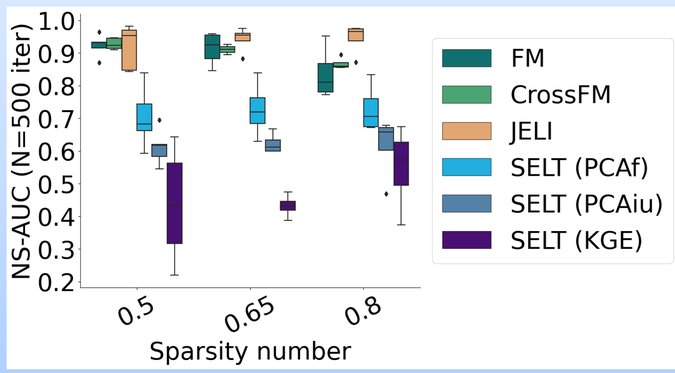


Fig. 6. Boxplots of NS-AUCs across the ablation study on synthetic data sets in JELI.

Discussion

JELI is a novel importance score-based approach for drug repurposing, that flexibly encodes a graph-based regularization constraint on drugs and diseases

➡ JELI explicitly includes the feature-wise importance scores
➡ JELI can be readily applied beyond the task of drug repurposing

[1] Lundberg & Lee. NeurIPS, 2017.
[2] Ribeiro, Singh & Guestrin. SIGKDD, 2016.
[3] Fokkema, de Heide & van Erven. JMLR, 2023.
[4] Swamy, Radmehr et al. arXiv:2207.00551, 2022.
[5] Balazevic, Allen & Hospedales. NeurIPS, 2019.
[6] Wang, … & Yu. WWW, 2019.
[7] Yu, Bilenko, Lin. DOI: 10.1137/1.9781611974973
[8] TRANSCRIPT. DOI: 10.5281/zenodo.7982976
[9] Szklarczyk … & von Mering. Nucl.Ac.Res., 2023.

← **GitHub**
JELI Python package

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