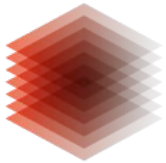
Interaction Network Analysis Using Semantic Similarity Based on Translation Embeddings



TIB
LEIBNIZ-INFORMATIONSZENTRUM
TECHNIK UND NATURWISSENSCHAFTEN
UNIVERSITÄTSBIBLIOTHEK

Fraunhofer
IAIS

Presenter: Damien Graux

Authors: Awais Manzoor Bajwa,
Diego Collarana, and Maria-Esther Vidal

SMART
DATA
ANALYTICS
FROM DATA TO KNOWLEDGE

SEMANTiCS
Karlsruhe 2019

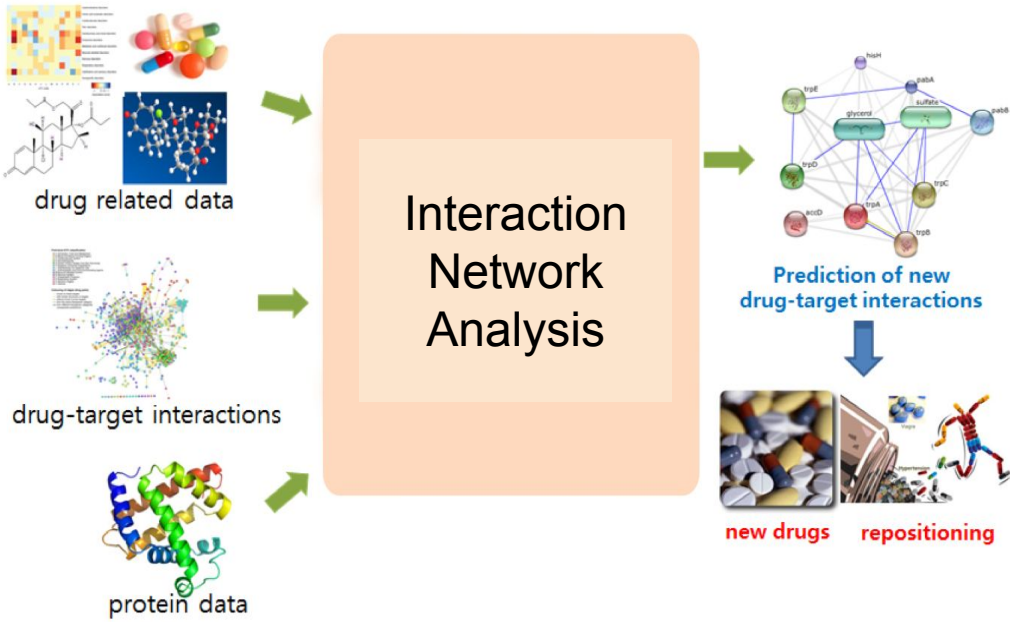
UNIVERSITÄT BONN

WHAT

HOW

Motivation

- In vitro & vivo identification of drug target interactions is expensive, time consuming, and very laborious;
- Bringing a new drug to the market, **costs≈$1.8 billion** and takes more than 10 years;
- Computational approaches predict such interactions to be then verified, helping in reducing the cost.



Question

Network based drug-target interaction prediction with probabilistic soft logic

Different solutions already exist.

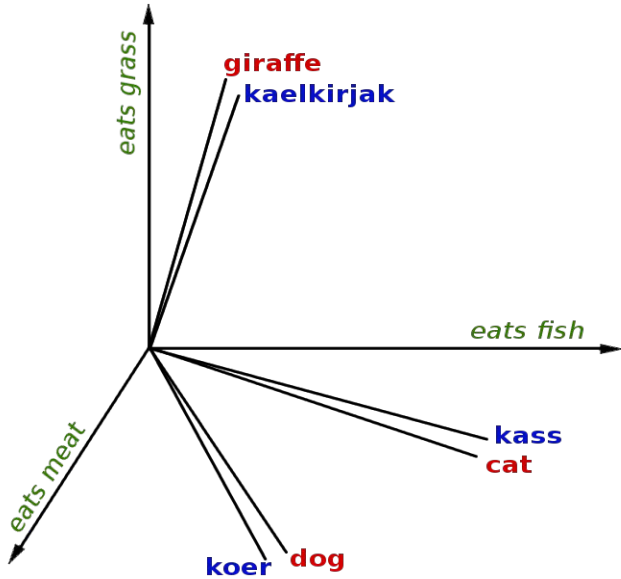
Why work on this problem?

Kernelized bayesian matrix factorization with twin kernels

Drug-Target interaction prediction using semantic similarity and edge partitioning

Vector Based Approaches

- Simple, multidimensional and computationally efficient vector based representation;
- Can be used as underlying input to other machine learning algorithms;
- More information can be embedded;
- Entities are comparable to each other in different dimensions;
- Simpler to visualize in vector space.



http://www.marekrei.com/blog/multilingual-semantic-models/

WHY

WHAT

HOW

Research question and goals

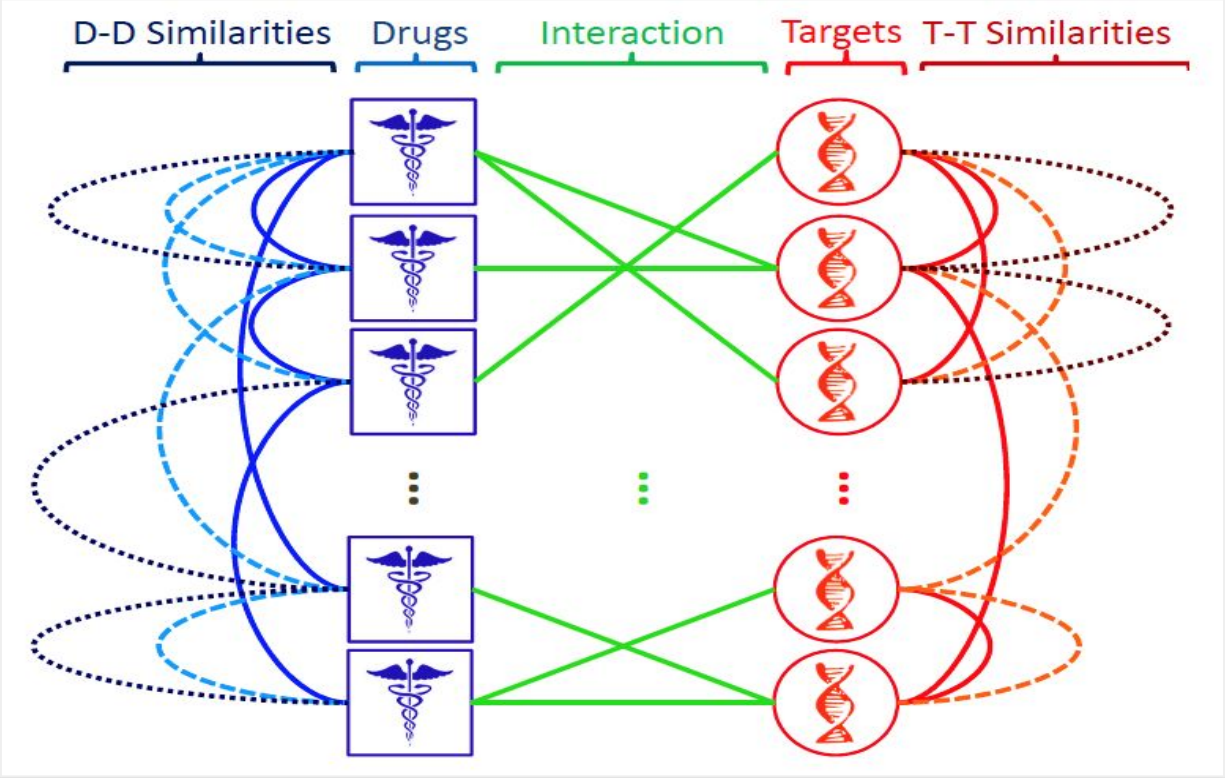
- We want to answer the following research question:

  "Can semantically vector based representation of drug-target interaction network identify novel drug-target interactions?"

- We want to understand entity to vector (e2v) translation approaches;
- We want to develop a novel e2v approach that encodes semantic similarity value between entities too;
- We evaluate link prediction between drugs and targets.

Why embed semantic similarity values too?

- The **homophily principle** is the tendency of individuals to **interact** and bond with similar others;
- The presence of homophily has been discovered in a vast array of network studies;
- Individuals in homophilic relationships share common characteristics (e.g.,beliefs, values, education) that make communication and relationship formation easier;



Drug-Target interaction network with similarity based interactions represented with dotted lines.

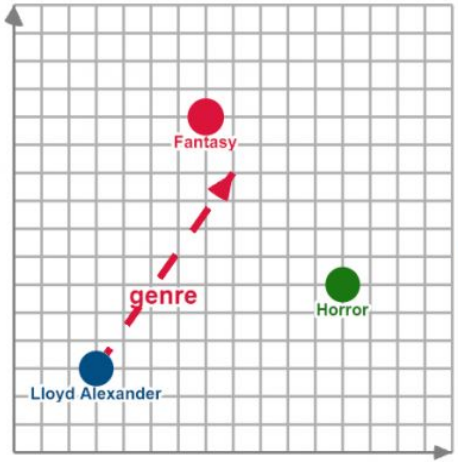
https://www.semanticscholar.org/paper/Network-Based-Drug-Target-Interaction-Prediction-Fakhraei-Huang/074f3e9f973e6127dfe2eb74a51c15594ce15e25

We extend TransE

- Published in NIPS 2013 by Antoine Bordes et al;
- True interactions are closer, corrupted interactions are moved away;
- Tree representation are considered;
- Stochastic gradient descent is used;
- It is an energy based model.

$$subject + predicate \approx object$$
$$subject + predicate \not\approx object$$

$$\mathcal{L} = \sum_{(h,\ell,t)\in S} \sum_{(h',\ell,t')\in S'_{(h,\ell,t)}} \left[\gamma + d(\boldsymbol{h}+\boldsymbol{\ell},\boldsymbol{t}) - d(\boldsymbol{h}'+\boldsymbol{\ell},\boldsymbol{t}')\right]_+$$



http://pyvandenbussche.info/2017/translating-embeddings-transe/

WHY

WHAT

HOW

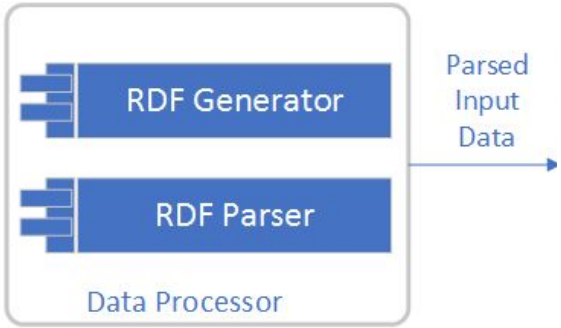
SimTransE

- End-to-end approach for drug-target interaction prediction;
- Two interaction types are considered:
  - Actual drug-target interactions
  - Similarity based reinforced interactions
- Vector embeddings representing each drug & target
- Predictions made based on similarities between embeddings

SimTransE

Interaction Functions	Objective Functions
$\begin{cases} h + l \approx t, & \text{if h interacts (l) t} \\ h + l \not\approx t, & \text{otherwise} \end{cases}$	$L_i = \sum_{(h,\ell,t)\in S} \sum_{(h',\ell,t')\in S'_{(h,\ell,t)}} \left[\gamma + d(h+\ell,t) - d(h'+\ell,t')\right]_+$
$\begin{cases} h1 + l \approx h2, & \text{if h1 similar h2} \\ h1 + l \not\approx h2, & \text{otherwise} \end{cases}$	$L_s = \sum_{(h,\ell,t)\in S} \sum_{(h',\ell,t')\in SI_{(h,\ell,t)}} \left[d(h+\ell,t) - d(h'+\ell,t')\right]_+$

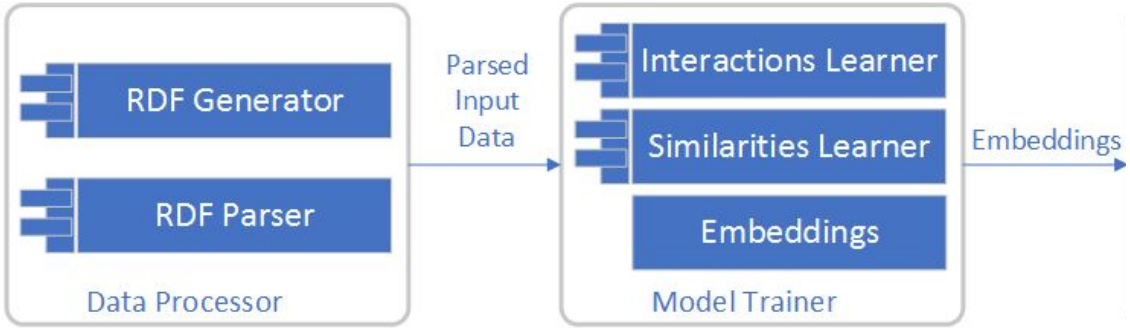
The Architecture



The **Data Processor** parses data to RDF and creates:

- Three sets of, i.e., the subjects (s), relational entities (p), and the objects (o).
- Two matrices one representing the positive and negative interactions of entities and second with the similarity values among entities.

The Architecture



The **Model Trainer** receives as input dictionaries and matrices and:

- It resorts to the **stochastic gradient descent** method to optimize the position and direction of the embeddings in a vector space;
- It uses interactions and similarities between entities to solve the optimization problem, and generates embeddings as output.

The Architecture

The **Predictor** component takes the generated embedding vectors, and thresholds and:

- It iterates over all the entities and predicts new interactions of each entity with every other entity;
- It calculates the Precision, Recall and additionally, the Area Under Receiver (AUC) and the Area Under the Precision-Recall Curve (AUPRC).

Experiments and Results

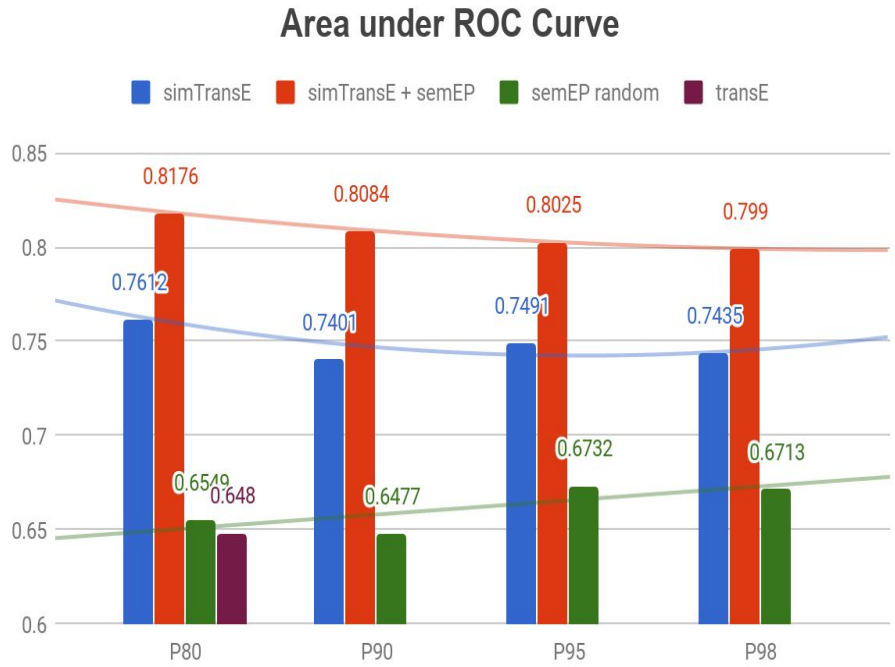
Drug-Target Interaction Prediction

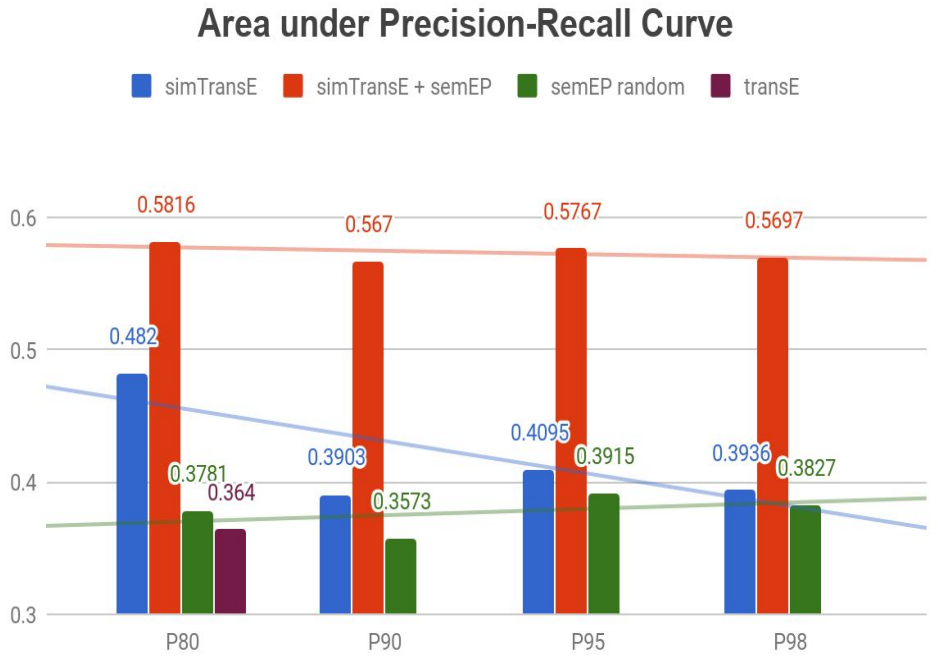
- Drug-target data obtained from KEGG BRITE, BRENDA, SuperTarget, and DrugBank. (Nov 2007)
- Similarity scores are computed by the SIMCOMP score (Hattori et al, J.Ame.Chem.Soc, 2003)
- 10 fold cross validation, 4 Percentiles P80, P90, P95, P98
- 90% interactions used for training, 10% for test
- Precision, Recall, AUC, AUPRC

Statistics	Nuclear Receptor	Ion channel
# of drugs	54	210
# of targets	26	204
# of drug-target interactions	90	1476

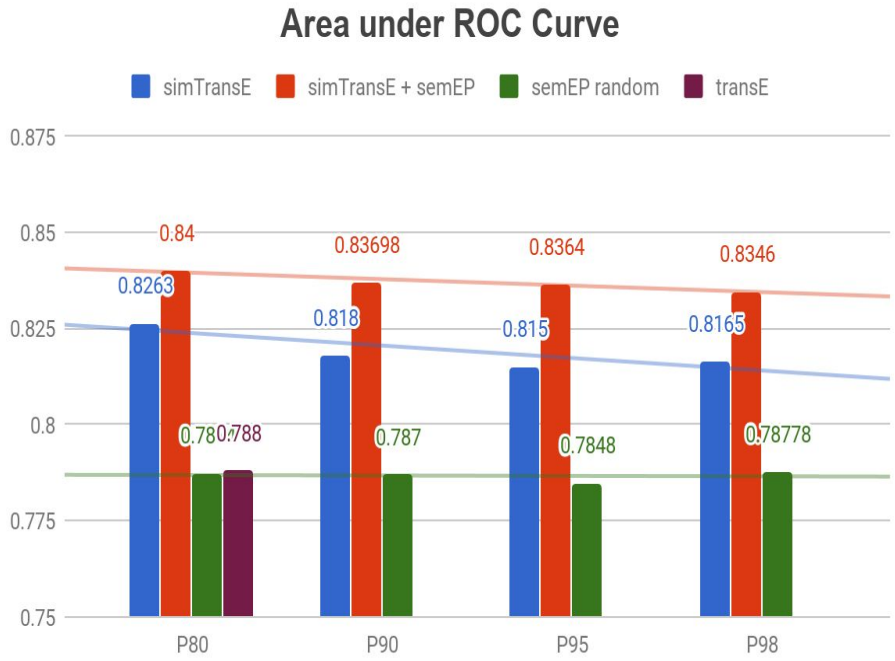
Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2718640/

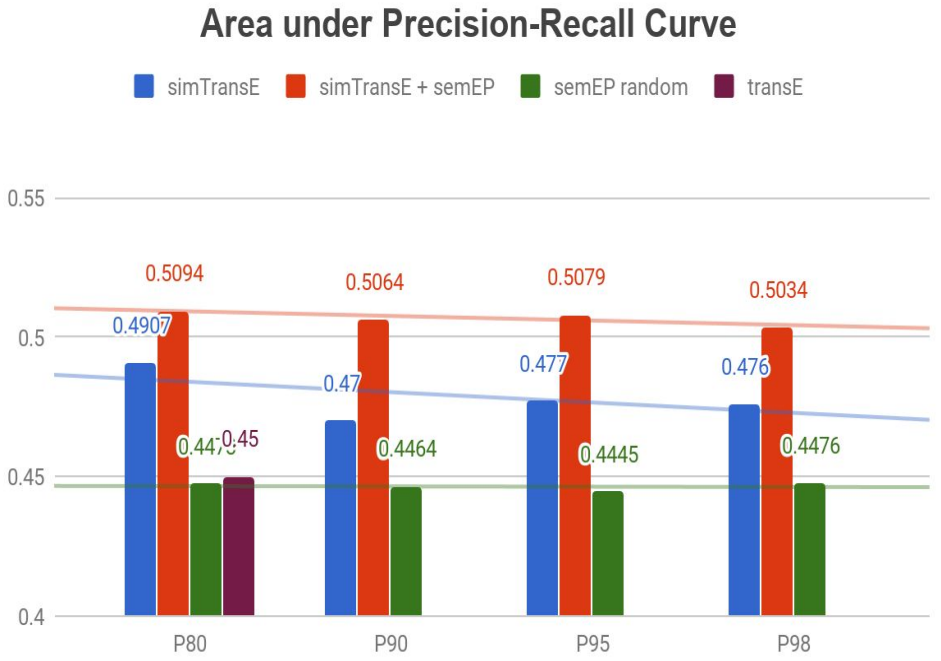
Nuclear Receptor





Ion channel


Area under ROC Curve
simTransE
simTransE + semEP
semEP random
transE
0.875
0.85
0.825
0.8
0.775
0.75
0.84
0.83698
0.8364
0.8346
0.8263
0.818
0.815
0.8165
0.788
0.787
0.7848
0.78778
P80
P90
P95
P98


Area under Precision-Recall Curve
simTransE
simTransE + semEP
semEP random
transE
0.55
0.5
0.45
0.4
0.5094
0.5064
0.5079
0.5034
0.4907
0.47
0.477
0.476
0.45
0.4464
0.4445
0.4476
P80
P90
P95
P98

Conclusion

1. Presented results suggest that vector-based drug-target embeddings are a suitable solution for predicting interactions.

2. Including values of similarities in the training process improves the predictive performance.

3. SimTransE is able to produce drug-target predictions with significant results.

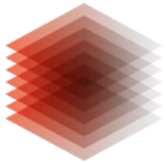
Future Work

1. Work on the problem of imbalanced classes. Decrease the mixed up negative examples in positive space

2. See if the results are improved while using TrasH as underlying approach

3. Include other input data sources in the training process e.g. side effect similarities.

Thanks for your Attention
Questions?



TIB
LEIBNIZ-INFORMATIONSZENTRUM
TECHNIK UND NATURWISSENSCHAFTEN
UNIVERSITÄTSBIBLIOTHEK

Fraunhofer
IAIS

Questions:

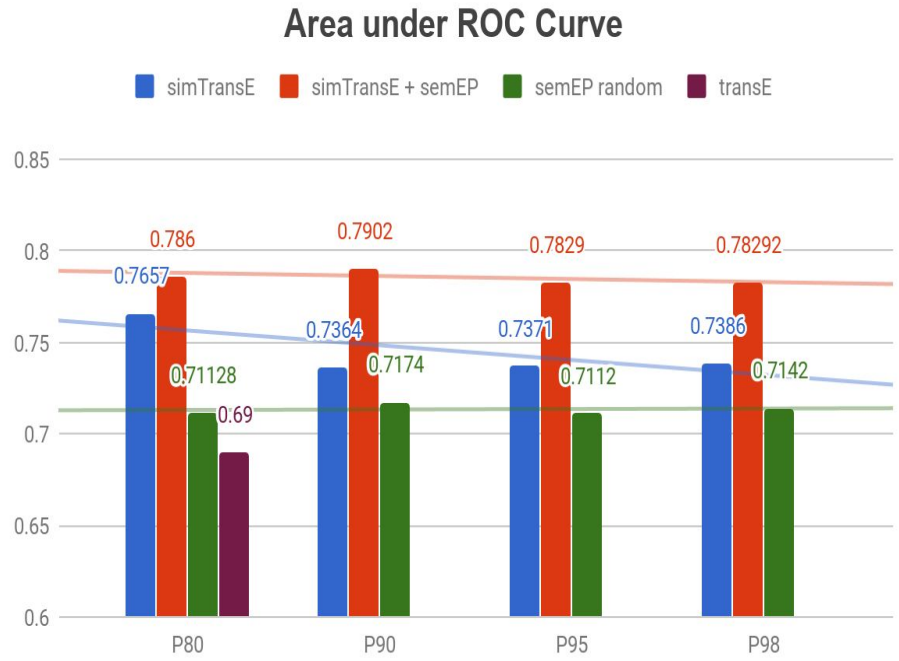
<Diego.Collarana.Vargas@iais.fraunhofer.de>

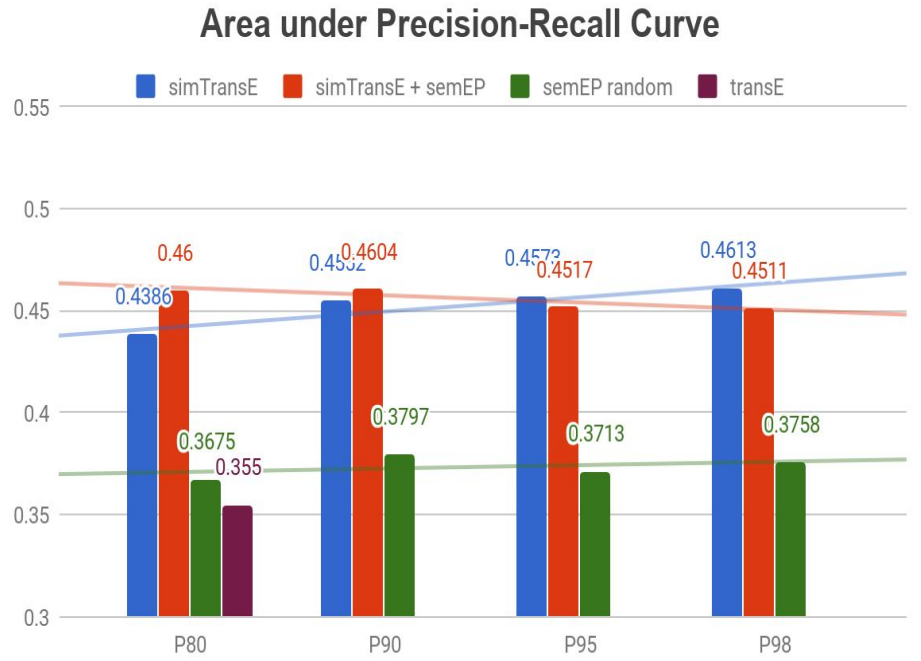
SMART
DATA
ANALYTICS
FROM DATA TO KNOWLEDGE

SEMANTiCS
Karlsruhe 2019

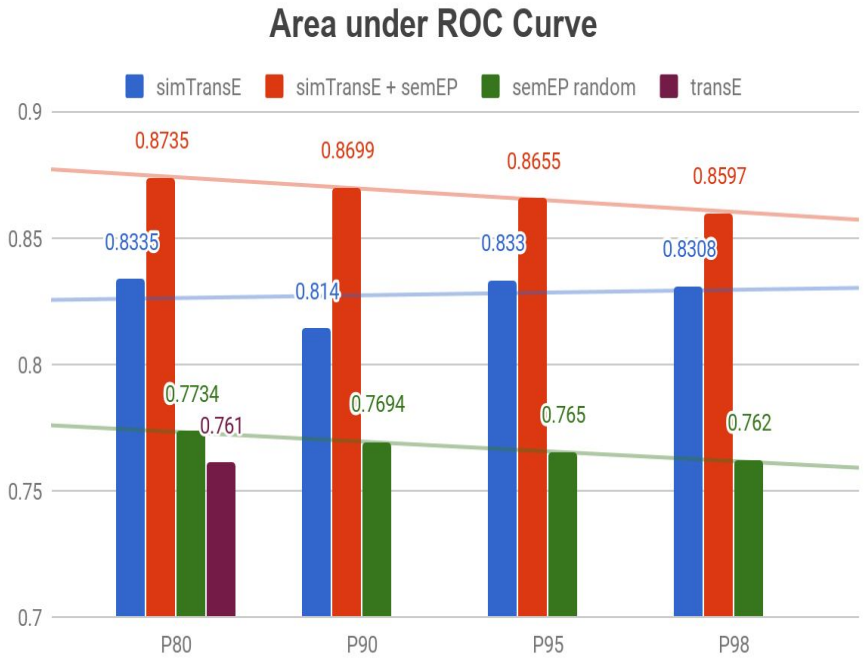
UNIVERSITÄT BONN

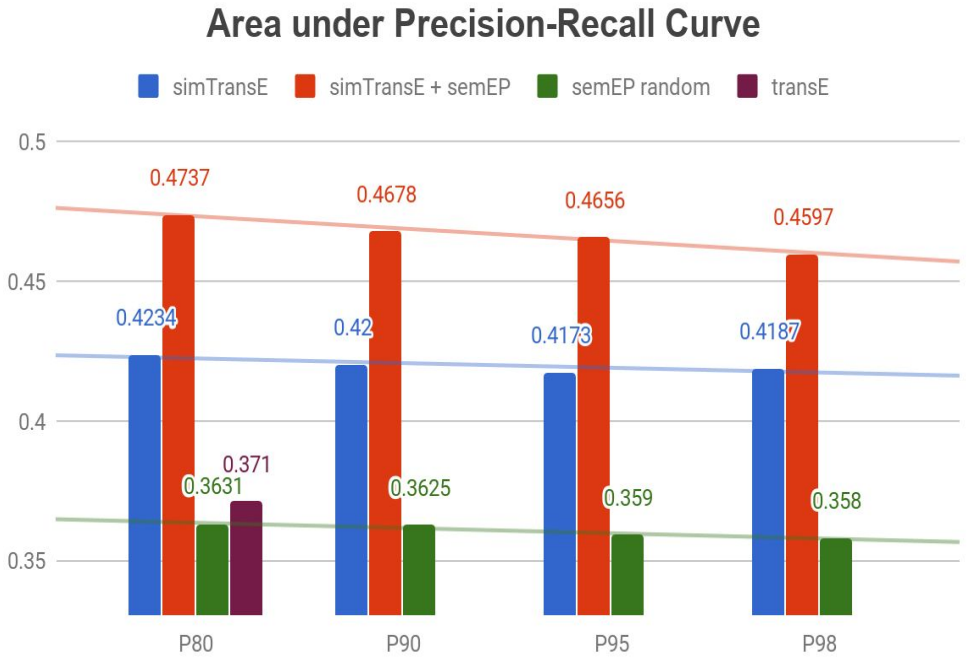
GPCR


Area under ROC Curve
simTransE
simTransE + semEP
semEP random
transE
0.85
0.8
0.75
0.7
0.65
0.6
0.786
0.7902
0.7829
0.78292
0.7657
0.7364
0.7371
0.7386
0.71128
0.7174
0.7112
0.7142
0.69
P80
P90
P95
P98


Area under Precision-Recall Curve
simTransE
simTransE + semEP
semEP random
transE
0.55
0.5
0.45
0.4
0.35
0.3
0.46
0.4604
0.4517
0.4613
0.4511
0.4386
0.4532
0.4573
0.3675
0.355
0.3797
0.3713
0.3758
P80
P90
P95
P98

Enzyme


Area under ROC Curve
simTransE
simTransE + semEP
semEP random
transE
0.9
0.85
0.8
0.75
0.7
0.8735
0.8699
0.8655
0.8597
0.8335
0.814
0.833
0.8308
0.7734
0.761
0.7694
0.765
0.762
P80
P90
P95
P98


Area under Precision-Recall Curve
simTransE
simTransE + semEP
semEP random
transE
0.5
0.45
0.4
0.35
0.4737
0.4678
0.4656
0.4597
0.4234
0.42
0.4173
0.4187
0.371
0.3631
0.3625
0.359
0.358
P80
P90
P95
P98

Algorithm to learn simTransE vector embeddings

Inputs:

- set $S = \{(h, \ell, t)\}$ of training data.
- set E *and* L representing the entities and relations.
- square matrix for each entity type representing the similarity among each entity i.e. the value at index (i, j) represents the similarity value between entity i and entity j.
- hyper parameters learning rate **lr**, margin γ, number of embedding dimensions k, number of similar interactions t, similarity threshold **st**, number of epochs **ep** .

Initialize:

similarInteractions $\leftarrow$ generateSimilarInteractions(simMatrix,S,t,**st**)

$\boldsymbol{\ell} \leftarrow$ uniform$(-\frac{6}{\sqrt{k}}, \frac{6}{\sqrt{k}})$ for each relation $\ell \in$ L

$\mathbf{e} \leftarrow$ uniform$(-\frac{6}{\sqrt{k}}, \frac{6}{\sqrt{k}})$ for each entity e $\in$ E

$iterationCounter \leftarrow 0$

loop

$iterationCounter+ = 1$

if $(iterationCounter == ep)$ **then**

break

end if

$\mathbf{e} \leftarrow \boldsymbol{e}/||\boldsymbol{e}||$ for each entity e $\in$ E

$S_{intBatch} \leftarrow sample(S, b)$ //sample a minbatch of size b $\in S$

$S_{simBatch} \leftarrow sample(similarInteractions, b)$ //sample a minibatch of size b $\in similarInteractions$

$S_{batch} \leftarrow S_{intBatch}$

$T_{batch} \leftarrow \phi$

for $(h, \ell, t) \in S_{batch}$ **do**

$(h', \ell, t') \leftarrow sample(S'_{(h,\ell,t)})$ // sample a corrupted triplet

$T_{batch} \leftarrow T_{batch} \cup \{((h, \ell, t), (h', \ell, t'))\}$ // original triple should not go as a corrupted triple

end for

$S_{batch} \leftarrow S_{simBatch}$

$Sim_{batch} \leftarrow \phi$

for $(h, \ell, t) \in S_{batch}$ **do**

$(h', \ell, t') \leftarrow sample(S_{simBatch(h,\ell,t)})$ // sample a soft similar triplet

$Sim_{batch} \leftarrow Sim_{batch} \cup \{((h, \ell, t), (h', \ell, t'))\}$ // original triple should not go as a similar triple

end for

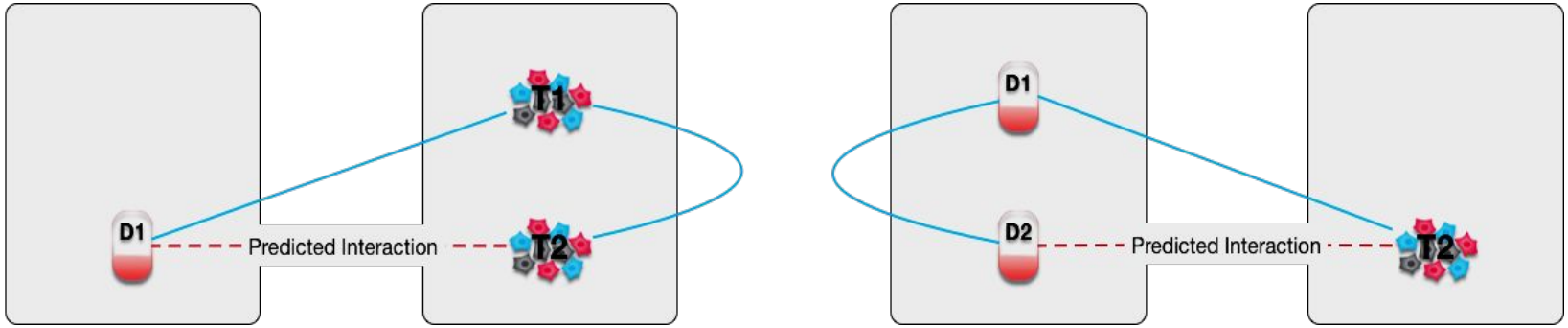
Update embeddings w.r.t

$$\sum_{((h,\ell,t),(h',\ell,t'))\in T_{batch}} \nabla\left[\gamma + d(\boldsymbol{h} + \boldsymbol{\ell}, \boldsymbol{t}) - d(\boldsymbol{h'} + \boldsymbol{\ell}, \boldsymbol{t'})\right]_+$$

$$\sum_{((h,\ell,t),(h',\ell,t'))\in Sim_{batch}} \nabla\left[d(\boldsymbol{h} + \boldsymbol{\ell}, \boldsymbol{t}) - d(\boldsymbol{h'} + \boldsymbol{\ell}, \boldsymbol{t'})\right]_+$$

end loop

Predicted interactions based on triad rule



Predicted interactions based on tetrad rule