Global and local evaluation of link prediction tasks with neural embeddings

Asan Agibetov, Matthias Samwald
Institute for Artificial Intelligence and Decision Support, Medical University of Vienna
SemDeep-4 @ ISWC 2018, Oct 8, 2018, Monterey CA, USA

Link prediction task

- Given two entities *e1*, *e2* (nodes in the KG) and their vector representations (*V(e1)*, *V(e2)*), predict whether a link with label *r1* exists between them
- Example:
  - e1 = TRIM28 gene
  - e2 = negative regulation of transcirption by RNA polymerase II
  - r1 = has function
  - Does „e1“ have function „e2“?
- Research question
  - How to obtain „good“ vector representations (i.e., embeddings) V(e1), V(e2)?

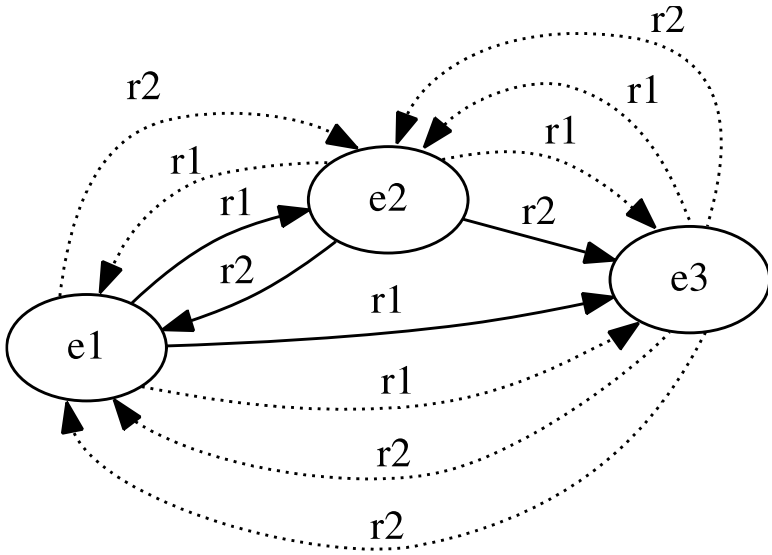
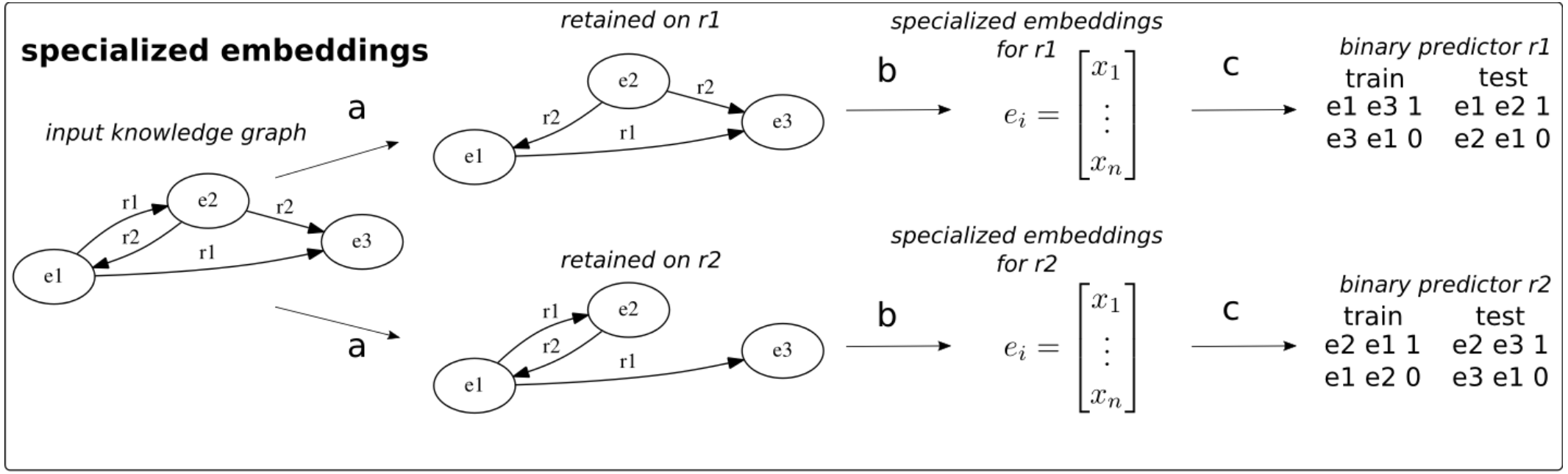


Fig. Caption:
In bold existing (positive) links, dashed non-existing (negative) links

Global vs. Local evaluation



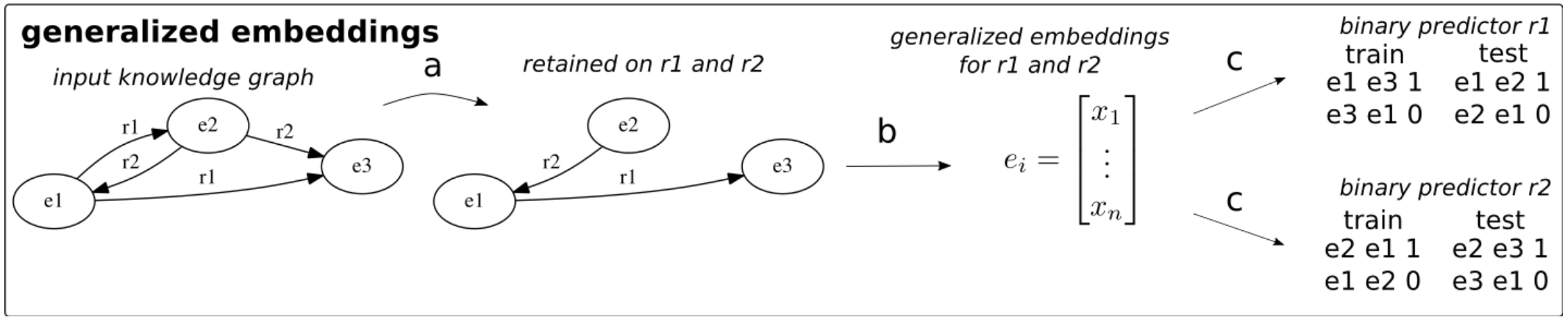


Fig. Caption: a) split positive and negative links into train and test, b) train embeddings on train positives, c) train binary classifiers for each relation type on learned embeddings

Datasets used

Dataset	# relation types	# entities	max # links per relation type	min # links per relation type	mean # links per relation type
WN11	11	40943	37221	86	8459
FB15k-237	237	14541	16391	45	1308
UMLS	46	137	1021	1	142
BIO-KG	9	346225	554366	6159	179915

Fig. Caption: WN11 – WordNet, FB – Freebase, UMLS – subset of Unified Medical Language System semantic network, BIO-KG (Alshahrani et al., Bioinformatics 2017) comprehensive and curated knowledge graph from several biological ontologies (e.g., GO) and databases (e.g., SIDER). In WN11 and FB15K-237 all inverse relations removed (see Toutanova et al., ACL 2016)

Flattening knowledge graphs

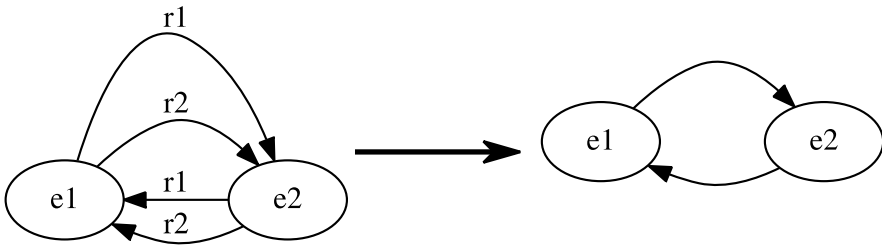


Fig. Caption: Turn KG into unlabelled directed graph, s.t., no pair of nodes is connected with more than one arc (directed edge)

Dataset	# pairs connected with > 1 relation types
WN11	124/93003 (0.133%)
FB15-237	23700/310116 (7.642%)
UMLS	1343/6527 (20.576%)
BIO-KG	0/1619239 (0%)

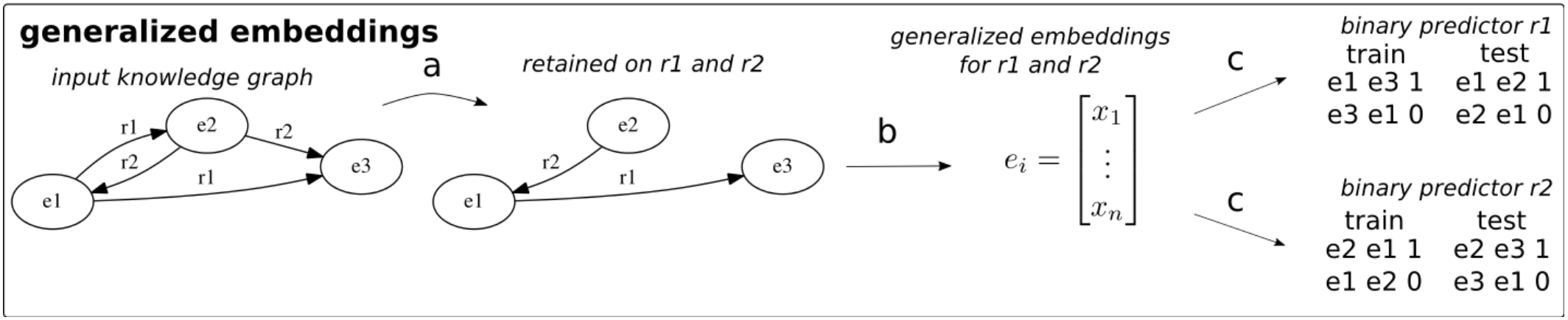
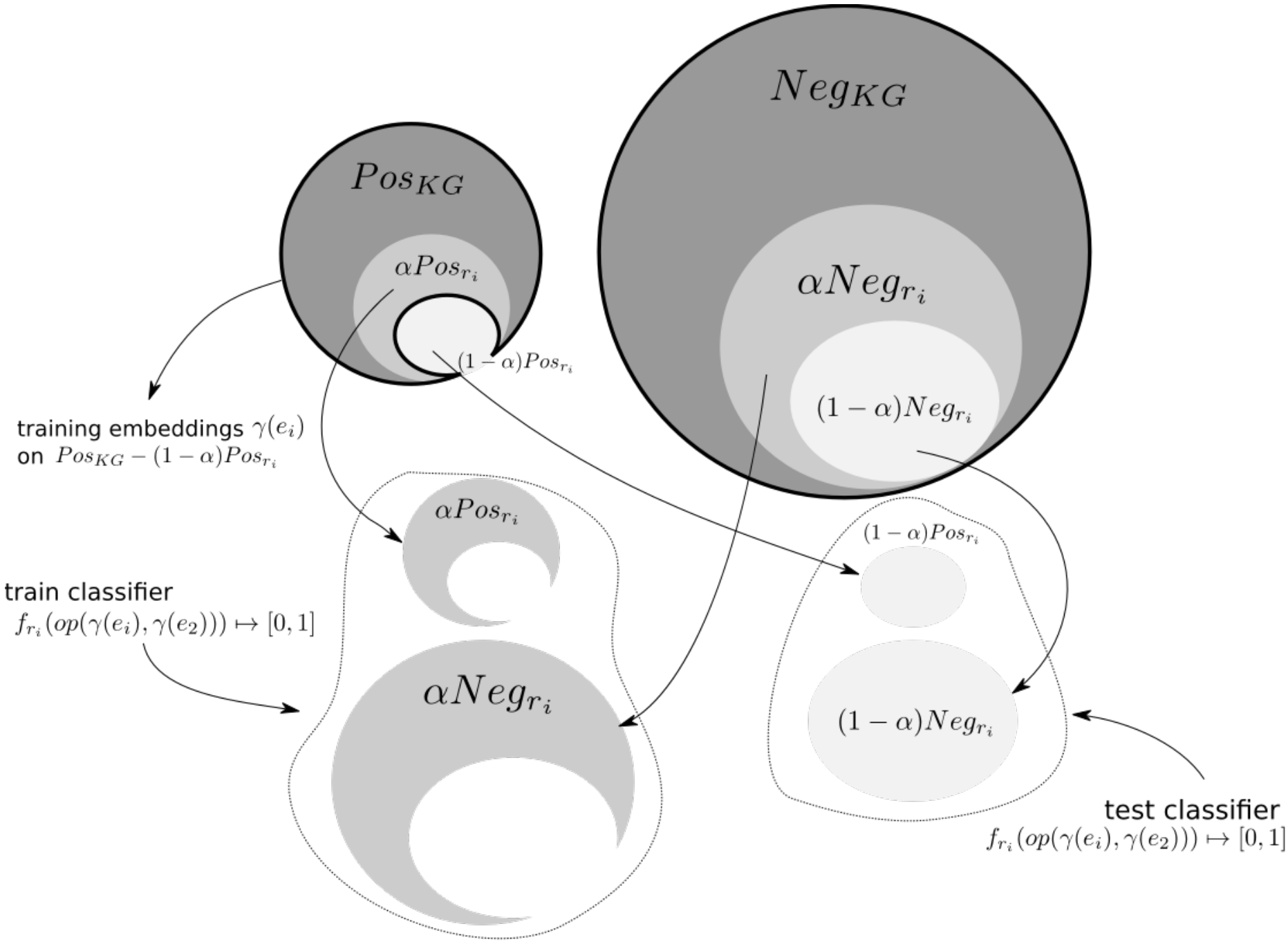


Fig. Caption: Compensate reduced information with binary classifiers fine tuned for each relation type

Evaluation pipeline



Generation of negatives

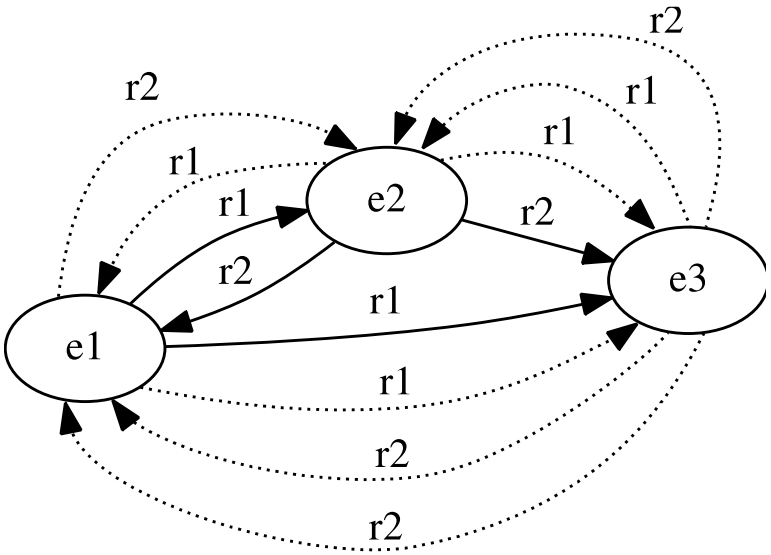
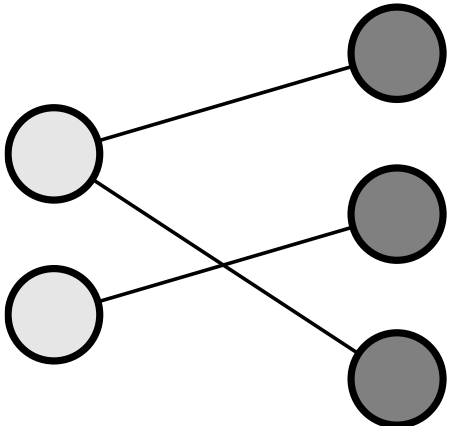


Fig. Caption:
In bold existing (positive) links, dashed non-existing (negative) links

$$\mu_{r_i} = \frac{3}{2 \times 3} = 0.5$$



$$\mu_{r_i} = \frac{6}{2 \times 3} = 1.0$$

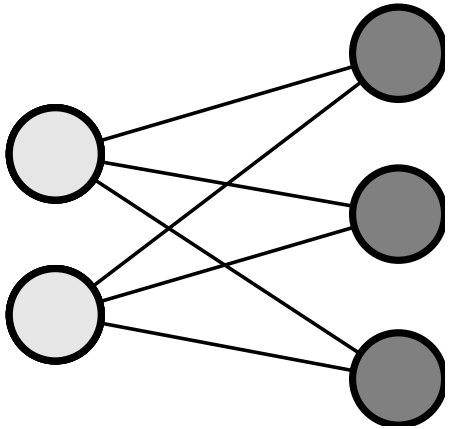
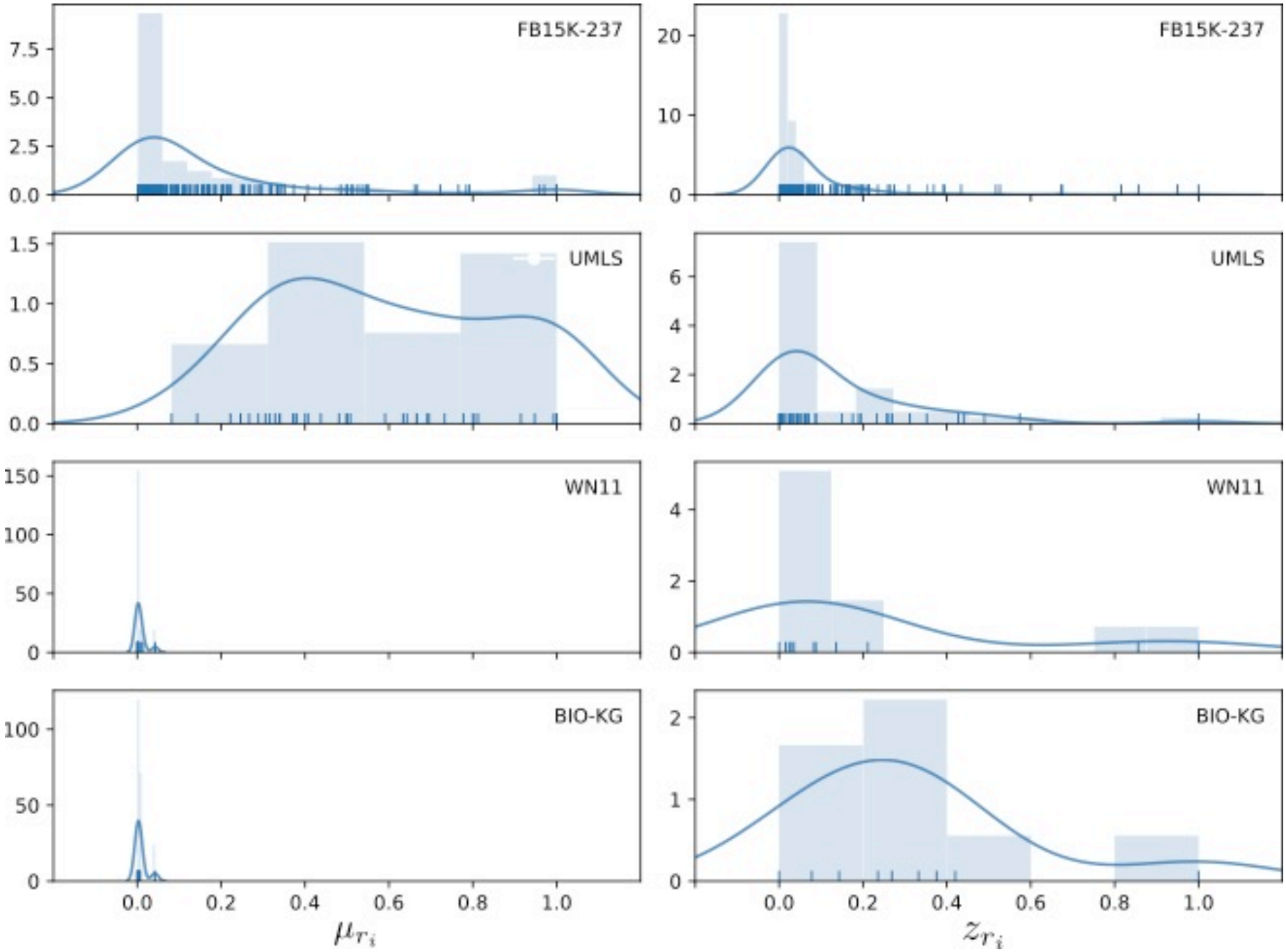
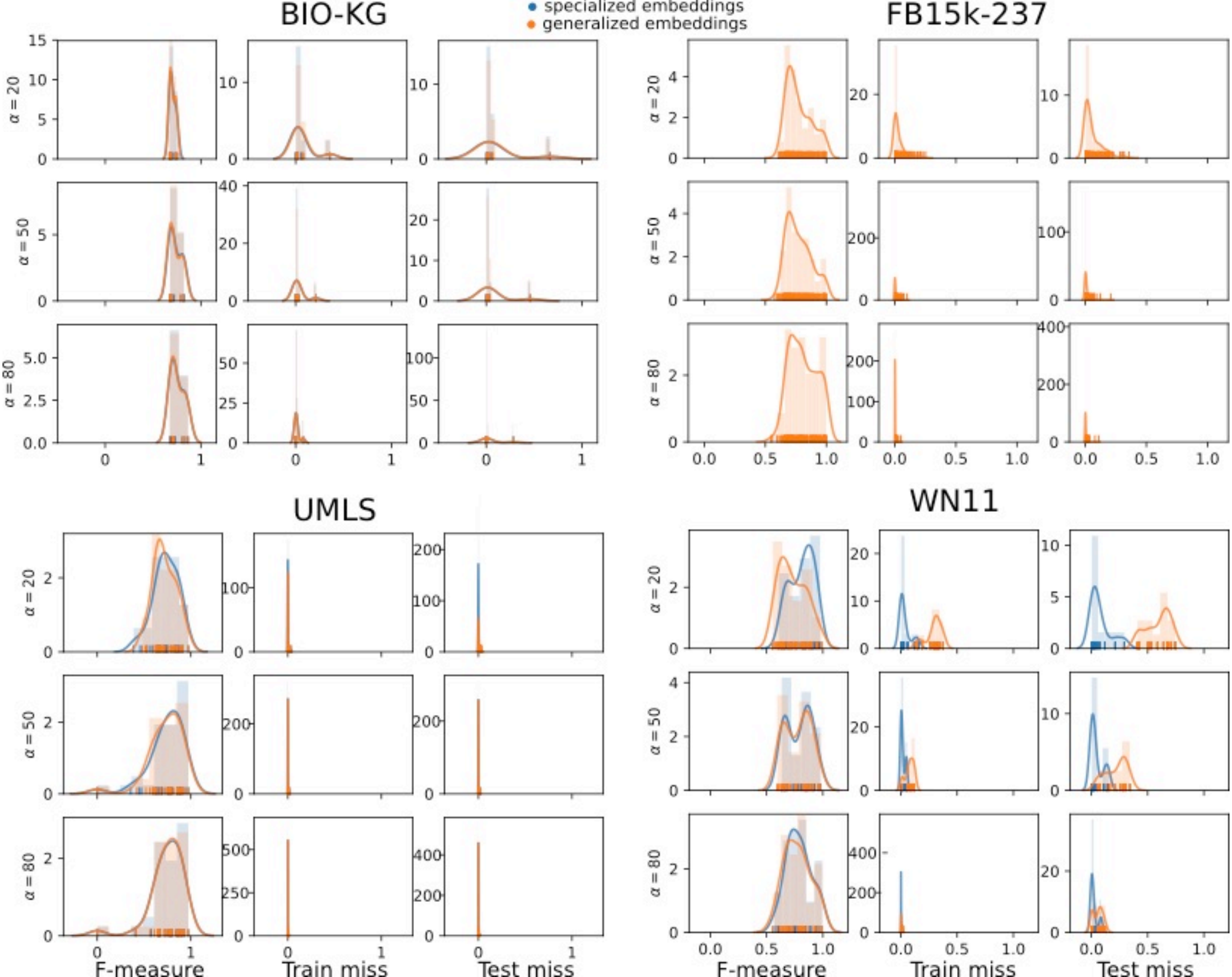


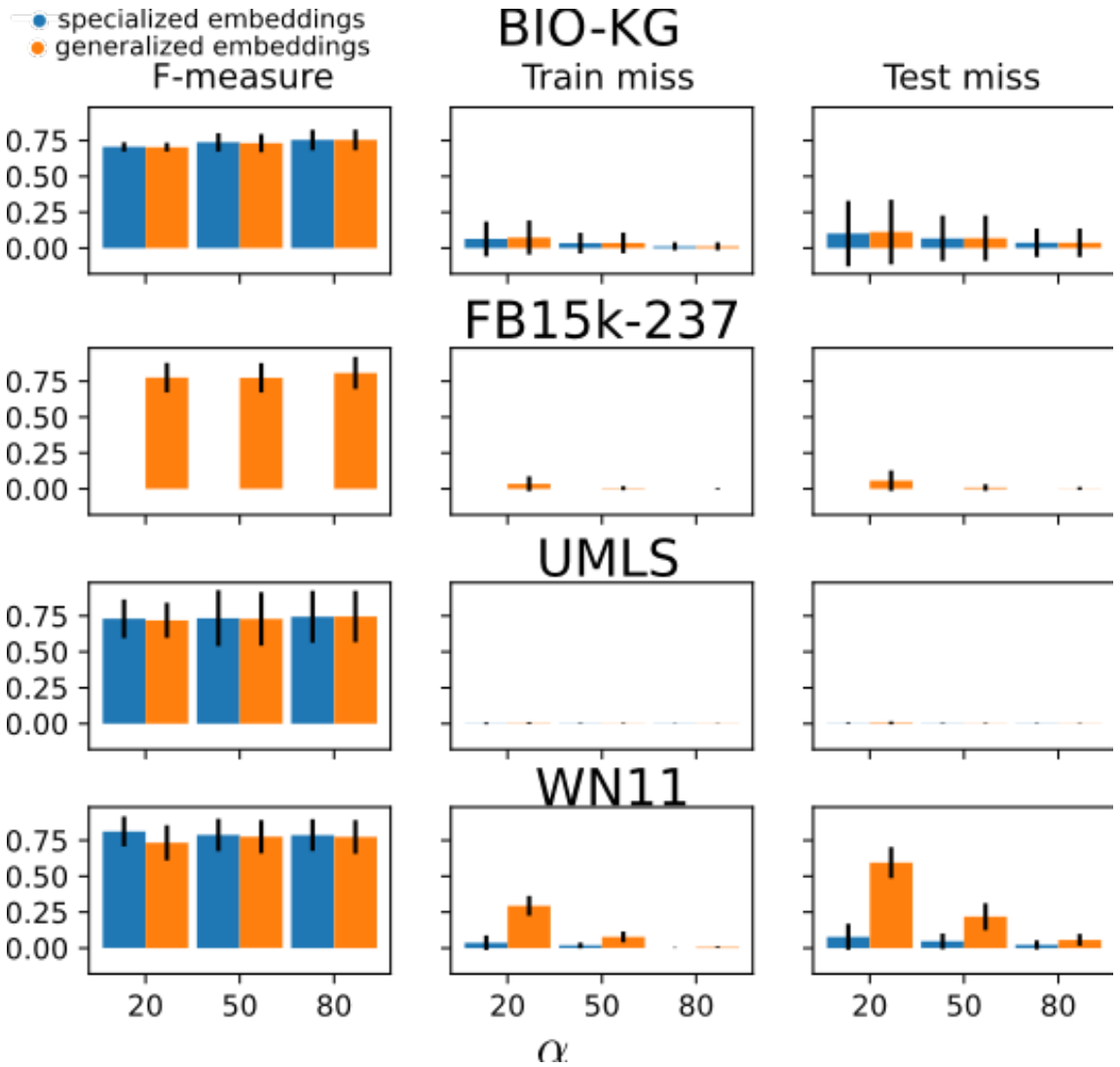
Fig. Caption: Capacity to produce negatives. μ measures how much positive information we have for a relation type

Connectivity characterization

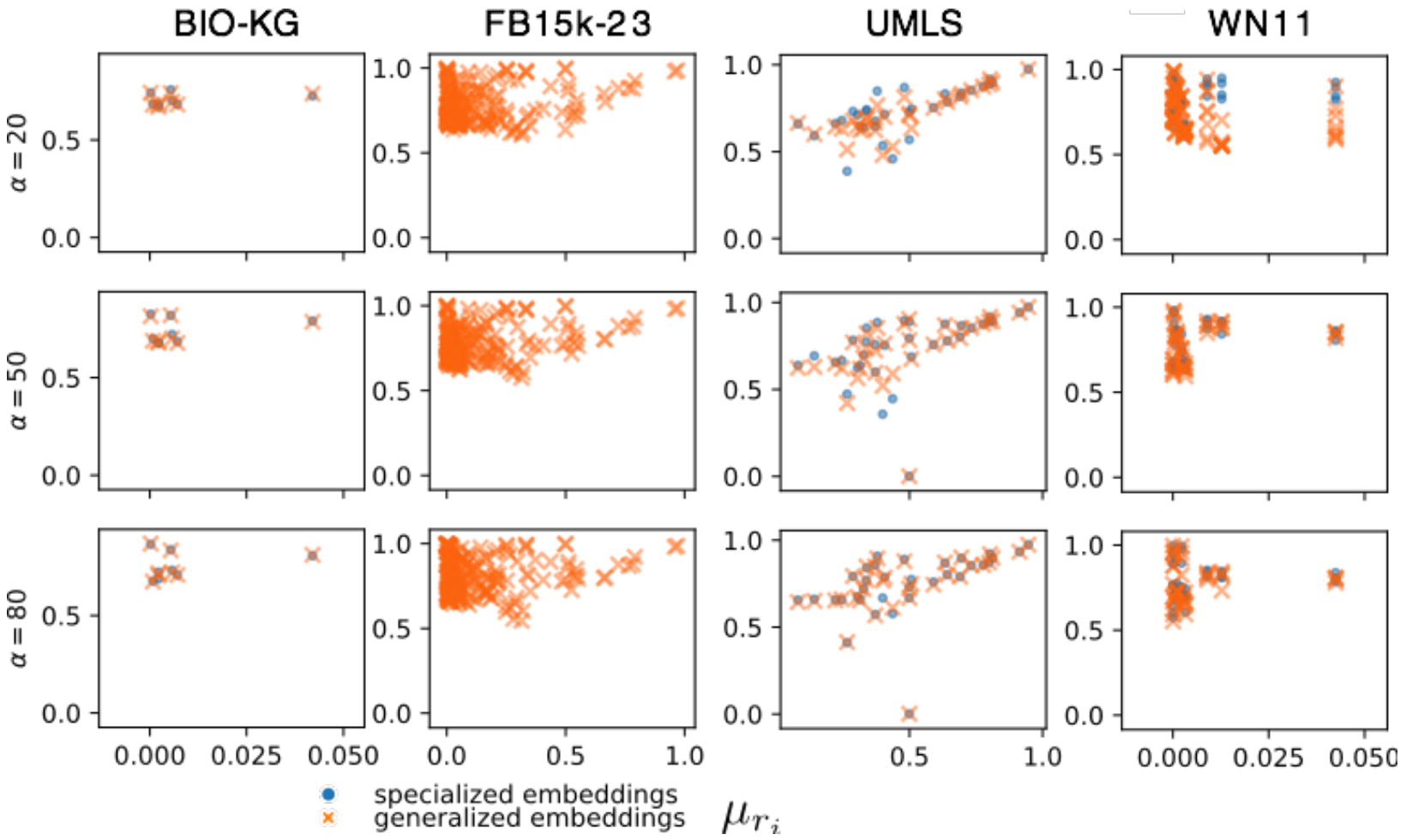



specialized embeddings
generalized embeddings
BIO-KG
FB15k-237
UMLS
WN11
α = 20
α = 50
α = 80
F-measure
Train miss
Test miss

Averaged classification scores



Correlation with μ



Conclusions

- Generalized embeddings can perform as good as specialized embeddings
  - Main advantage is scalability: we train embeddings only once for the whole KG
- Best results are achieved for dense and very well connected graphs OR sparse graphs with many positive links per relation type
- Evaluation pipeline is open source (collaborations more than welcome!)
  - https://github.com/plumdeq/neuro-kglink.git