Universität Stuttgart
IPVS – Institute for Parallel and Distributed Systems
Analytic Computing

Faithful Embeddings for EL++ Knowledge Bases

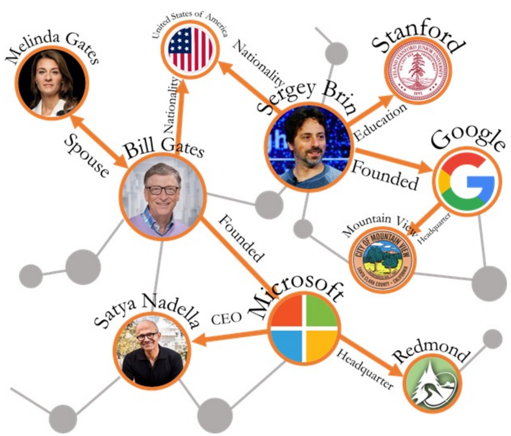
Bo Xiong[1], Nico Potyka[2], Trung-Kien Tran[3], Mojtaba Nayyeri[1], and Steffen Staab[1,4]

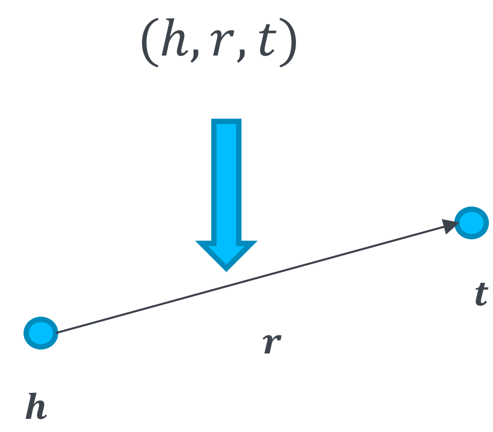
[1] University of Stuttgart, Germany
[2] Imperial College London, United Kingdom
[3] Bosch Center for Artificial Intelligence, Germany
[4] University of Southampton, United Kingdom

ISWC 2022 Virtual
02.09.2022

Knowledge Graphs (KGs)

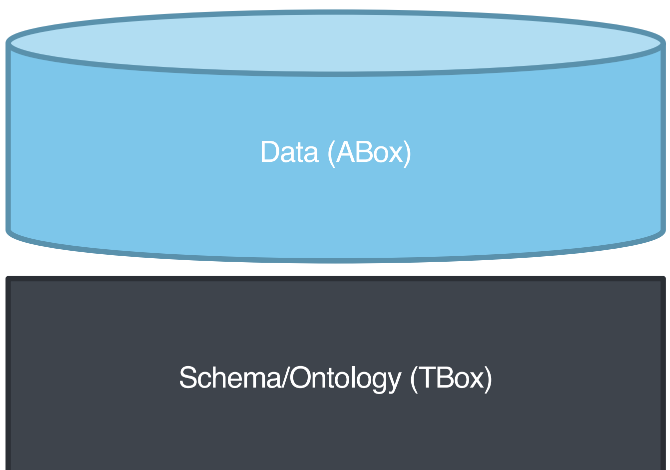
- KG is often considered as a set of tripes (h,r,t) defined over
  - a set of entities $\mathcal{E}$ (nodes) and
  - a set of relations $\mathcal{R}$ (edges)
- KG embeddings embed entities and relations into vector spaces
  - such that the relational structure is preserved
  - and new plausible links can be inferred





Data vs Conceptual knowledge

- A knowledge base (KB) can be divided into
  - ABox: instance information (data level) and
  - TBox: class information (concept level)
- Knowledge is often expressed via logical statements/assertions in Description Logic



- $(alice, type, DataScientist)$
- $(bob, type, SoftwareEngineer)$
- $(alice, \text{has\_}employer, ibm)$
- $(bob, has\_employer, ibm)$
- $(ibm, type, TechCompany)$

- $DataScientist \sqsubseteq Employee$
- $SoftwareEngineer \sqsubseteq Employee$
- $Employee \equiv \exists has\_employer.Employer$
- $TechCompany \sqsubseteq Company$
- $Company \sqsubseteq Employer$

- Can we embed KBs with conceptual knowledge?

EL++ Knowledge Bases

- EL++ is a lightweight description logic that
  - balances well between expressive power and reasoning complexity (polynomial)
  - has many applications in large-scare ontologies (e.g., Gene Ontology)

EL++ ABox contains:
1. concept assertion: C(a)
2. role assertion: r(a,b)

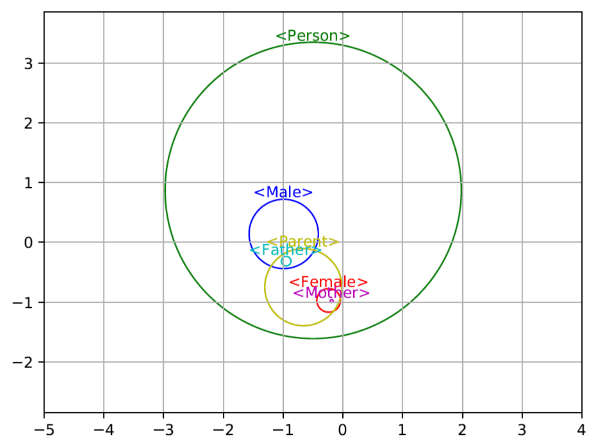
EL++ TBox statements can be normalized into the following forms
1. concept subsumption: $C \sqsubseteq D$,
2. concept intersection: $C_1 \sqcap C_2 \sqsubseteq D$,
3. right existential: $\exists r. C_1 \sqsubseteq D$,
4. left existential: $C_1 \sqsubseteq \exists r. C_2$,

Baader, F., Brandt, S., & Lutz, C. Pushing the EL envelope. In IJCAI 2005.

Ball EL++ Embedding

- Ball EL++ embedding embeds
  - concepts/entities as (convex) regions (balls)
  - relations as translation (like TransE)

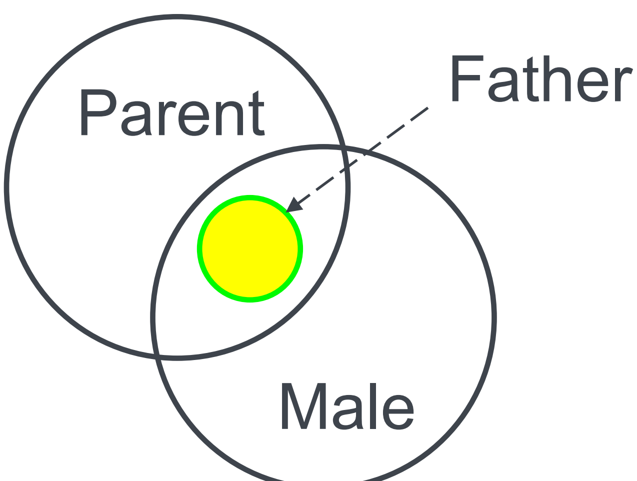
$$
\begin{aligned}
Male &\sqsubseteq Person \\
Female &\sqsubseteq Person \\
Father &\sqsubseteq Male \\
Mother &\sqsubseteq Female \\
Father &\sqsubseteq Parent \\
Mother &\sqsubseteq Parent \\
Female \sqcap Male &\sqsubseteq \bot \\
Female \sqcap Parent &\sqsubseteq Mother \\
Male \sqcap Parent &\sqsubseteq Father \\
\exists hasChild.Person &\sqsubseteq Parent \\
Parent &\sqsubseteq Person \\
Parent &\sqsubseteq \exists hasChild.\top
\end{aligned}
$$

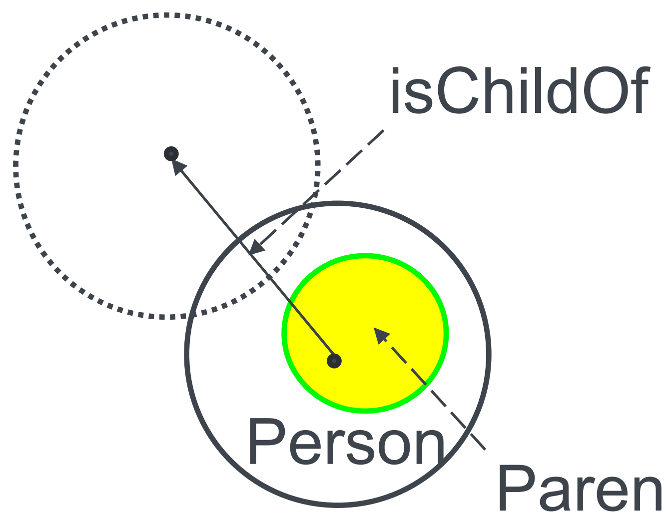


Kulmanov, M., Liu-Wei, W., Yan, Y., & Hoehndorf, R. EL Embeddings: Geometric Construction of Models for the Description Logic EL++. In IJCAI 2019: 6103-6109.

Ball EL++ Embedding

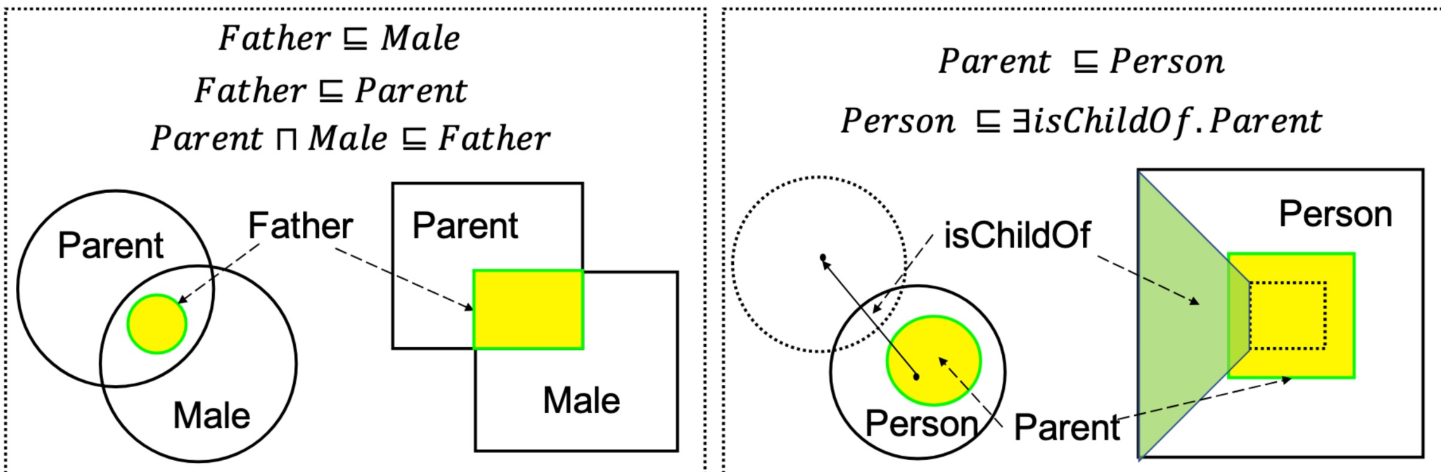
- Advantages: parameterization is simple
- Disadvantages:
  - Balls cannot faithfully represent concept intersection
  - Translation causes issues for concepts with varying size
  - No distinction between entities/concepts





Our idea: Box EL++ Embedding

- Box EL++ embedding represents
  - concepts by hyper-rectangulars (boxes)
  - relations by affine-transformations (translation + scaling)
  - entities by points in R^n

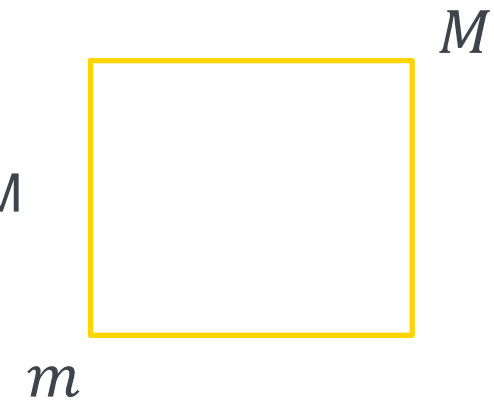


(a) Ball and Box embedding (b) Translation and affine transformation

Parameterization

- Box (concept) is parameterized by a lower-left corner + upper-right corner

$$\mathrm{Box}_w(C) = \{x \in \mathbb{R}^n \mid m_w(C) \leq x \leq M_w(C)\},$$



- Point (entity) is a special case of Box where m=M

- Affine transformation (relation) is parameterized by a diagonal scaling matrix + a translation vector

$T_w^r(x) = D_w^r x + b_w^r$, where D_w^r is an $(n \times n)$ diagonal matrix with non-negative entries

Geometric Interpretation

Idea: mapping **logical constraints/axioms** to **geometric (soft) constraints**

- we encode the axioms by designing one loss term for every axiom
- such that the axiom is satisfied by the geometric interpretation when the loss is 0

ABox embedding

- Concept assertion C(a)
  - Demanding point a to be inside the box of C

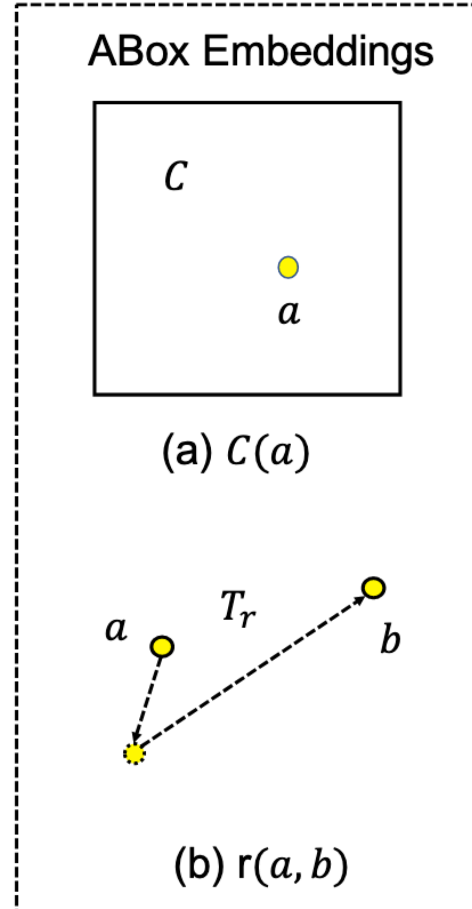
$$\mathcal{L}_{C(a)}(w) = \sum_{i=1}^{n} \|\max(0, m_w(a)_i - M_w(C)_i)\|_2 + \sum_{i=1}^{n} \|\max(0, m_w(C)_i - m_w(a)_i)\|_2 \,.$$

- Role assertion r(a, b)
  - Point a should be mapped b by an affine transformation

$$\mathcal{L}_{r(a,b)}(w) = \|T_w^r(m_w(a)) - m_w(b)\|_2 \,.$$

Proposition 1. *We have*

1. *If* $\mathcal{L}_{C(a)}(w) = 0$, *then* $\mathcal{I}_w \models C(a)$,
2. *If* $\mathcal{L}_{r(a,b)}(w) = 0$, *then* $\mathcal{I}_w \models r(a, b)$.



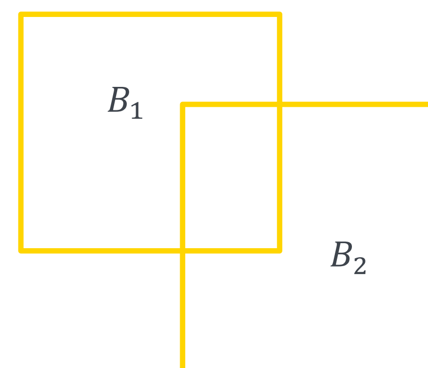
TBox embedding

Definition 2 (Disjoint measurement). *Given two boxes B_1, B_2, the disjoint measurement can be defined by the (modified) volumes of B_1 and the intersection box $B_1 \cap B_2$,*

$$\mathrm{Disjoint}(B_1, B_2) = 1 - \frac{\mathrm{MVol}(B_1 \cap B_2)}{\mathrm{MVol}(B_1)}. \tag{4}$$

We have the following guarantees.

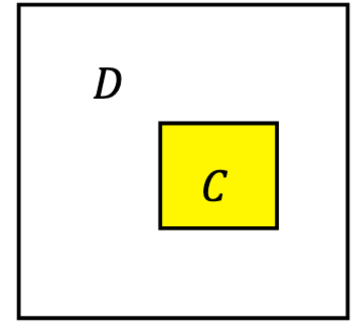
Lemma 1. *1.* $0 \leq \mathrm{Disjoint}(B_1, B_2) \leq 1$,
2. $\mathrm{Disjoint}(B_1, B_2) = 0$ *implies* $B_1 \subseteq B_2$,
3. $\mathrm{Disjoint}(B_1, B_2) = 1$ *implies* $B_1 \cap B_2 = \emptyset$.



TBox embedding

- concept subsumption $C \sqsubseteq D$

$$\mathcal{L}_{C \sqsubseteq D}(w) = \mathrm{Disjoint}(\mathrm{Box}_w(C), \mathrm{Box}_w(D)).$$

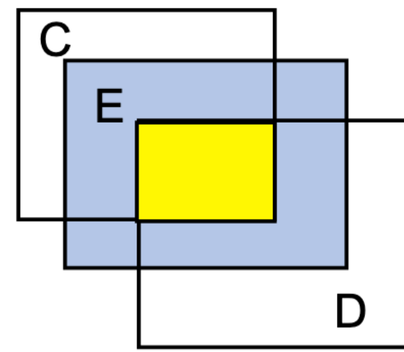


(c) $C \sqsubseteq D, D \neq \perp$

Proposition 2. *If $\mathcal{L}_{C \sqsubseteq D}(w) = 0$, then $\mathcal{I}_w \models C \sqsubseteq D$, where we exclude the inconsistent case $C = \{a\}, D = \perp$.*

- concept intersection

$$\mathcal{L}_{C \sqcap D \sqsubseteq E}(w) = \mathrm{Disjoint}(\mathrm{Box}_w(C) \cap \mathrm{Box}_w(D), \mathrm{Box}_w(E)).$$



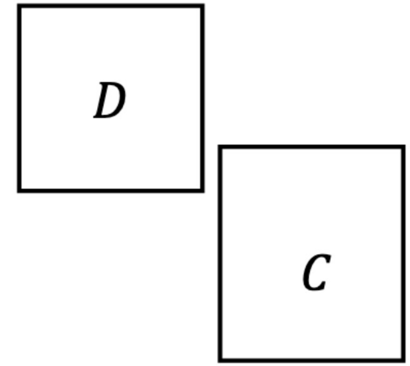
(d) $C \sqcap D \sqsubseteq E, E \neq \perp$

Proposition 3. *If $\mathcal{L}_{C \sqcap D \sqsubseteq E}(w) = 0$, then $\mathcal{I}_w \models C \sqcap D \sqsubseteq E$, where we exclude the inconsistent case $a \sqcap a \sqsubseteq \perp$ (that is, $C = D = \{a\}, E = \perp$).*

TBox embedding

- Concept disjointedness

$$\mathcal{L}_{C \sqcap D \sqsubseteq \perp}(w) = \frac{\mathrm{MVol}(\mathrm{Box}_w(C) \cap \mathrm{Box}_w(D))}{\mathrm{MVol}(\mathrm{Box}_w(C)) + \mathrm{MVol}(\mathrm{Box}_w(D))}.$$

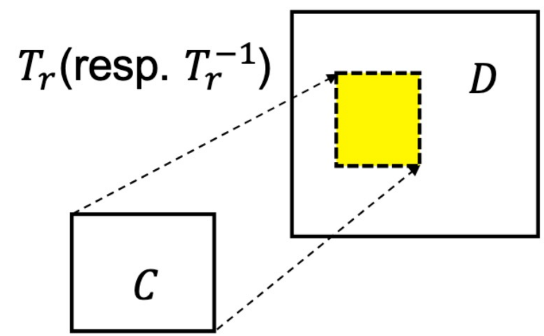


(e) $C \sqcap D \sqsubseteq \perp$

- Right/Left existential

$$\mathcal{L}_{C \sqsubseteq \exists r.D}(w) = \mathrm{Disjoint}(T_w^r(\mathrm{Box}_w(C)), \mathrm{Box}_w(D)).$$

Proposition 4. *If* $\mathcal{L}_{C \sqsubseteq \exists r.D}(w) = 0$, *then* $\mathcal{I}_w \models C \sqsubseteq \exists r.D$.

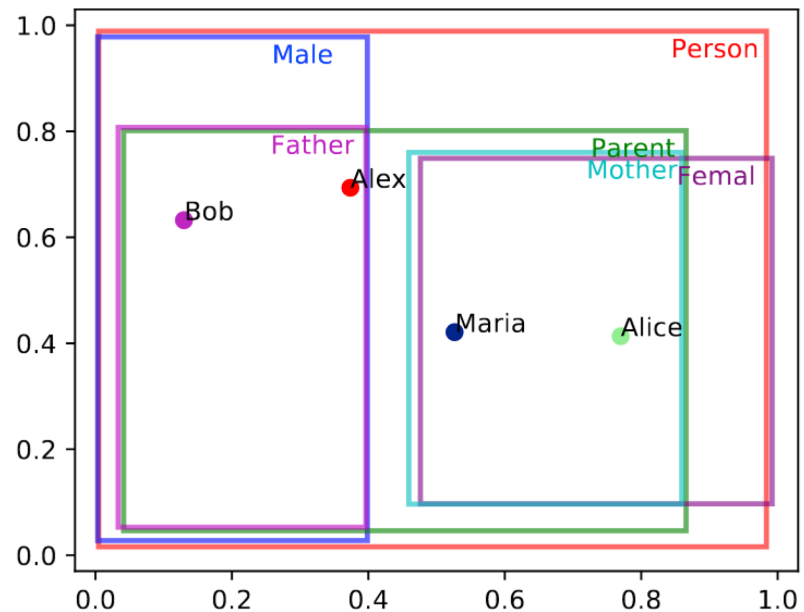


(f) $C \sqsubseteq \exists r.D$ (resp. $\exists\, r.C \sqsubseteq D$)

Toy family example

- We sum up all loss terms and use Adam as optimizer

Male $\sqsubseteq$ Person	Female $\sqsubseteq$ Person
Father $\sqsubseteq$ Male	Mother $\sqsubseteq$ Female
Father $\sqsubseteq$ Parent	Mother $\sqsubseteq$ Parent
Female $\sqcap$ Male $\sqsubseteq \perp$	Female $\sqcap$ Parent $\sqsubseteq$ Mother
Male $\sqcap$ Parent $\sqsubseteq$ Father	$\exists$hasChild.Person $\sqsubseteq$ Parent
Parent $\sqsubseteq$ Person	Parent $\sqsubseteq \exists$ hasChild.Person
Father(Alex)	Father(Bob)
Mother(Marie)	Mother(Alice)



- **Run the toy example from scratch?**

https://colab.research.google.com/drive/17U5olNtQotVXFT9kfr2p9K8RM_x2qH40?usp=sharing

Subsumption reasoning

Table 4: The ranking based measures of embedding models for sumbsumtion reasoning on the testing set. ∗ denotes the results from [20].

Dataset	Metric	TransE∗	TransH∗	DistMult∗	ELEm	EmEL^{++}	**BoxEL**
GO	Hits@10	0.00	0.00	0.00	0.09	0.10	0.03
	Hits@100	0.00	0.00	0.00	0.16	0.22	0.08
	AUC	0.53	0.44	0.50	0.70	0.76	0.81
	Mean Rank	-	-	-	13719	11050	8980
GALEN	Hits@10	0.00	0.00	0.00	0.07	0.10	0.02
	Hits@100	0.00	0.00	0.00	0.14	0.17	0.03
	AUC	0.54	0.48	0.51	0.64	0.65	0.85
	Mean Rank	-	-	-	8321	8407	3584
ANATOMY	Hits@10	0.00	0.00	0.00	0.18	0.18	0.03
	Hits@100	0.01	0.00	0.00	0.38	0.40	0.04
	AUC	0.53	0.44	0.49	0.73	0.76	0.91
	Mean Rank	-	-	-	28564	24421	10266

Protein-Protein Interactions

- Knowledge base constructed from
  - STRING database (ABox)
  - Gene Ontology (TBox)

STRING

https://string-db.org/

GENEONTOLOGY
Unifying Biology

http://geneontology.org/

- Link prediction evaluation: for $interact(P_1, P_2)$, identify rank of P_2 among all proteins

$$P(\text{interacts}(P_1, P_2)) = \left\| T_w^{\text{interacts}}(m_w(P_1)) - m_w(P_2) \right\|_2 .$$

Kulmanov, M., Liu-Wei, W., Yan, Y., & Hoehndorf, R. EL Embeddings: Geometric Construction of Models for the Description Logic EL++. In IJCAI 2019: 6103-6109.

Protein-protein interaction

Table 6: Prediction performance on protein-protein interaction (human).

Method	Raw Hits@10	Filtered Hits@10	Raw Hits@100	Filtered Hits@100	Raw Mean Rank	Filtered Mean Rank	Raw AUC	Filtered AUC
TransE	0.05	**0.11**	0.24	0.29	3960	3891	0.78	0.79
BoxE	0.05	0.10	0.26	0.32	2121	2091	0.87	0.87
SimResnik	0.05	0.09	0.25	0.30	1934	1864	0.88	0.89
SimLin	0.04	0.08	0.20	0.23	2288	2219	0.86	0.87
ELEm	0.01	0.02	0.22	0.26	1680	1638	0.90	0.90
EmEL^{++}	0.01	0.03	0.23	0.26	1671	1638	0.90	0.91
Onto2Vec	0.05	0.08	0.24	0.31	2435	2391	0.77	0.77
OPA2Vec	0.03	0.07	0.23	0.26	1810	1768	0.86	0.88
BoxEL (Ours)	**0.07**	0.10	**0.42**	**0.63**	**1574**	**1530**	**0.93**	**0.93**

Takeaway

- We propose Box EL++ KB embedding that
  - faithfully encodes concepts and relations
  - provides soundness guarantee for the underlying logical structure
  - shows significant improvements in biomedical KBs
  - Code is open available: https://github.com/Box-EL/BoxEL

- Future work
  - Incorporating such background knowledge into ML tasks
  - Embedding probabilistic description logic

Thank You