



EpiMap: Predicting Epigenetic Targets with KG Embedding Models*

HCLS @ KGC 2022



Payal Mitra



Thom Pijnenburg



Ted Slater

May 2, 2022



Krishna
Bulusu



Vishwa
Nellore



Sophie
Kirschner



Tom
Plasterer



Today



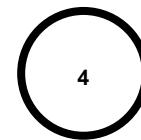
**Background: EpiMap –
Why and what**



**Building KGE models for
link Prediction**



**Results & Challenges
In KG ML**



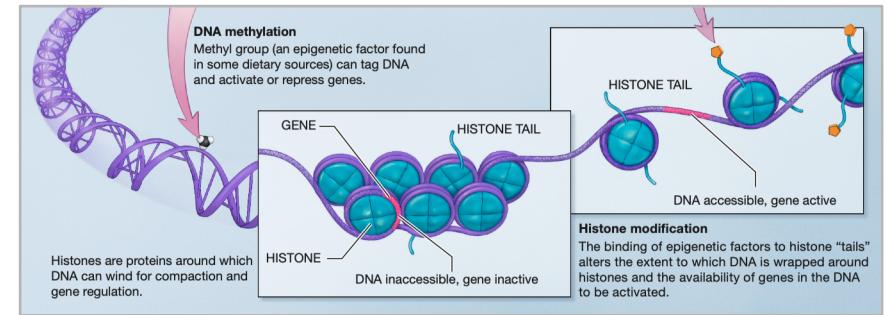
Applications and Impact

Epigenetics – factors that impact cellular function

- Genes encode instructions that are “read” to produce proteins; proteins do most of the physiological work in cells.

Abnormalities in protein synthesis can lead to disease

- Genetic causes: mutations, deletions, etc
- Epigenetic causes: genes become physically (in)accessible within to transcription machinery.

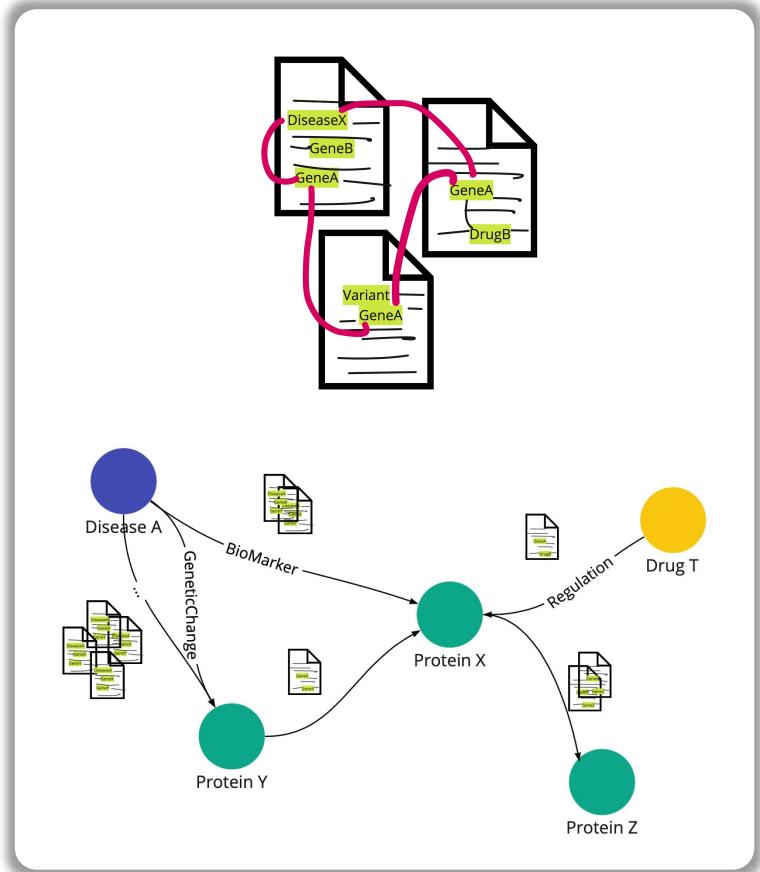


Opportunities

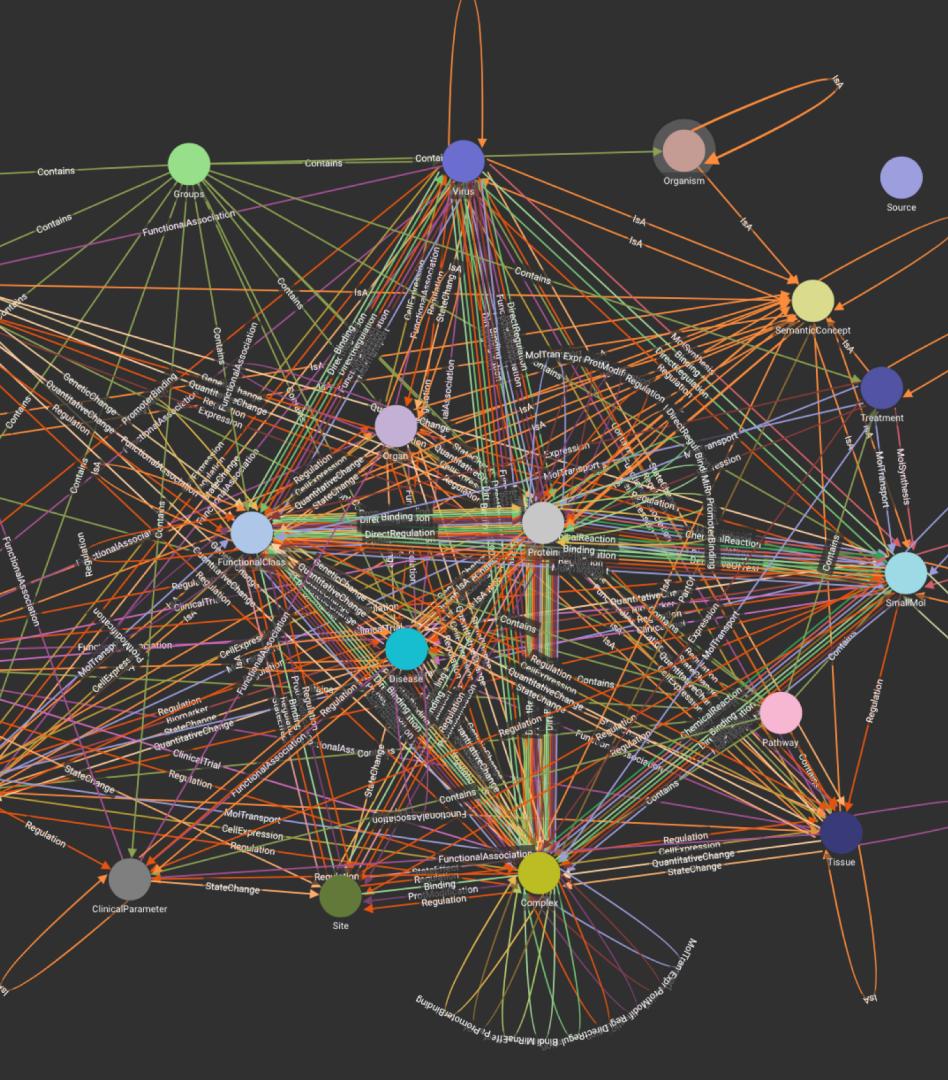
- >12,000 Epi publications each year.
- Unmet need to
 - Consolidate all epigenetic information.
 - Use the existing knowledge to identify novel epigenetic targets.
- Key strategic pillar for Oncology R&D community.

EpiMap KG unlocks Text

- The scientific literature contains vast amounts of information relevant to pharmaceutical R&D.
- Getting that information out of natural language and into a usable form is a huge challenge, because the scientific community uses different terminologies and formats.
- Literature-mined KGs address biological and technological complexity
 - Inject semantics and represent relations in context, e.g. directionality.
 - Support multi-hop path traversals, and explainable hypotheses
 - Capture evolution of knowledge in text-sources.
- EpiMap** is Elsevier's Biology Knowledge Graph ("ResNet") + assertions mined from literature describing epigenetic effects.



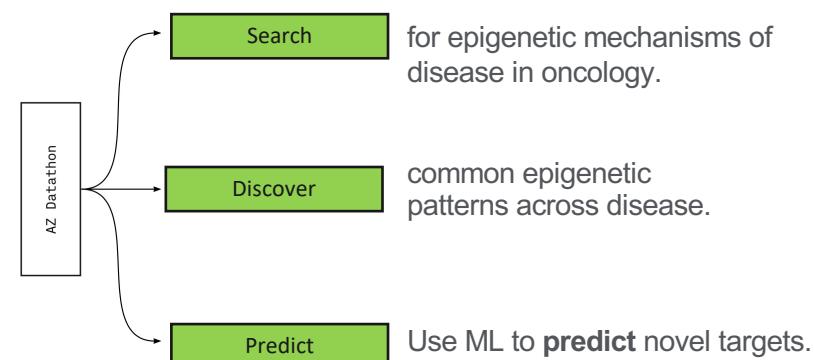
About EpiMap KG



EpiMap KG has ..

- *1.5M Vertices*
- *13.2M Edges*
- *Extracted from 7M documents*

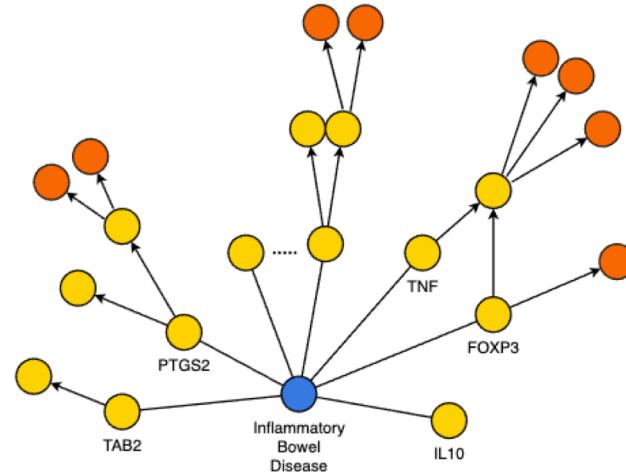
Collaboration with AstraZeneca Oncology with the goal to



EpiMap KG for Search and Discover

- Researcher Q:
Could Drug Target Interactions (DTI) with genes that are x hops away from Inflammatory Bowel Syndrome (IBS) potentially regulate IBS? I want to view these DTI paths.

- Step 1: # 1hop genes associated to IBS – 256
- Step 2: Friends of friends, i.e., 2nd hop genes – 10988
- Step 3: Filter only those genes which have known DTIs
Genes - 2477
DTI Paths – 5891

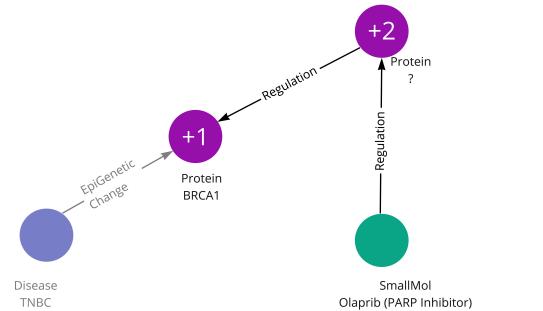


EpiMap provides deep insights into drug activity space

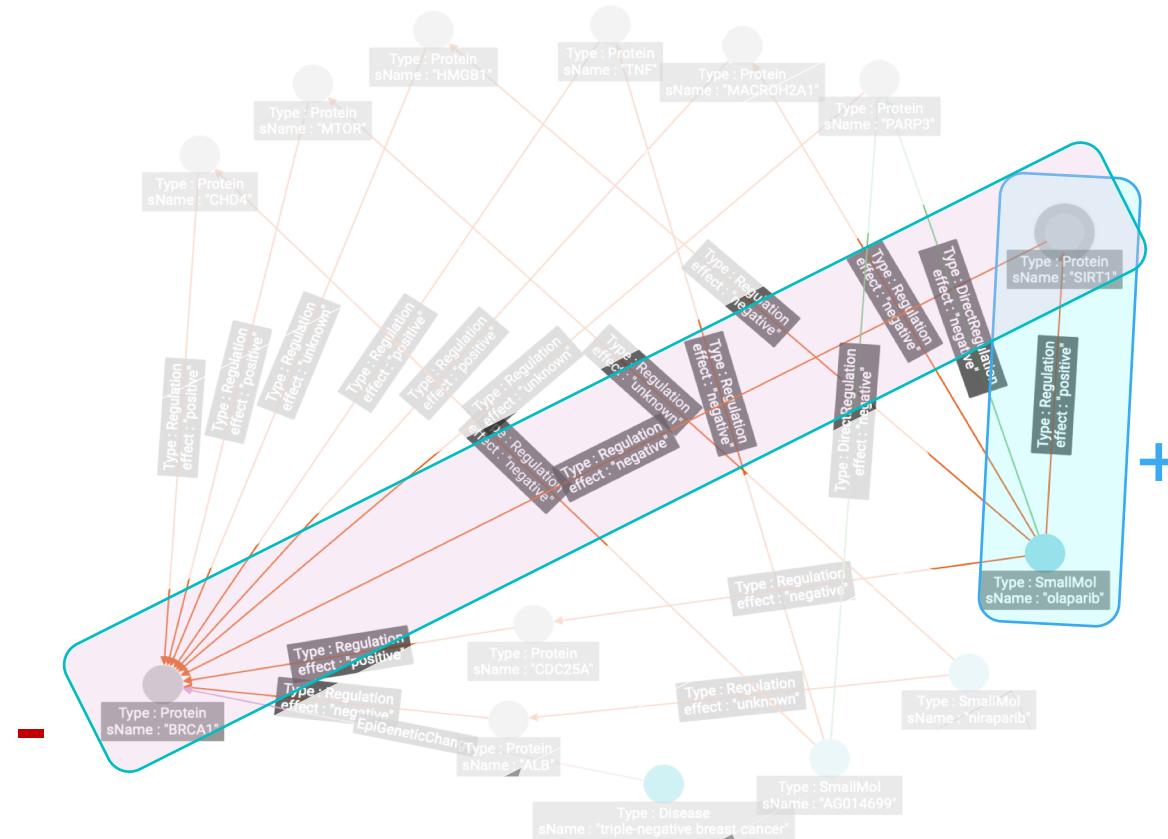


Which Genes co-sensitize to PARP inhibitors in context of disease?

- Leverage multi-hop patterns.
 - Identify alternate ‘directed’ paths driving drug response/resistance.
 - Generate novel hypotheses informing prospective validation studies.



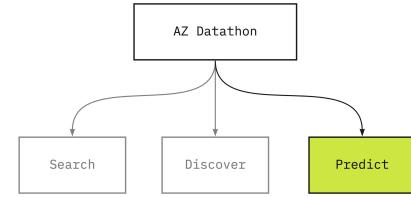
TNBC – Triple-Negative Breast Cancer



Today



Predicting novel targets using ML



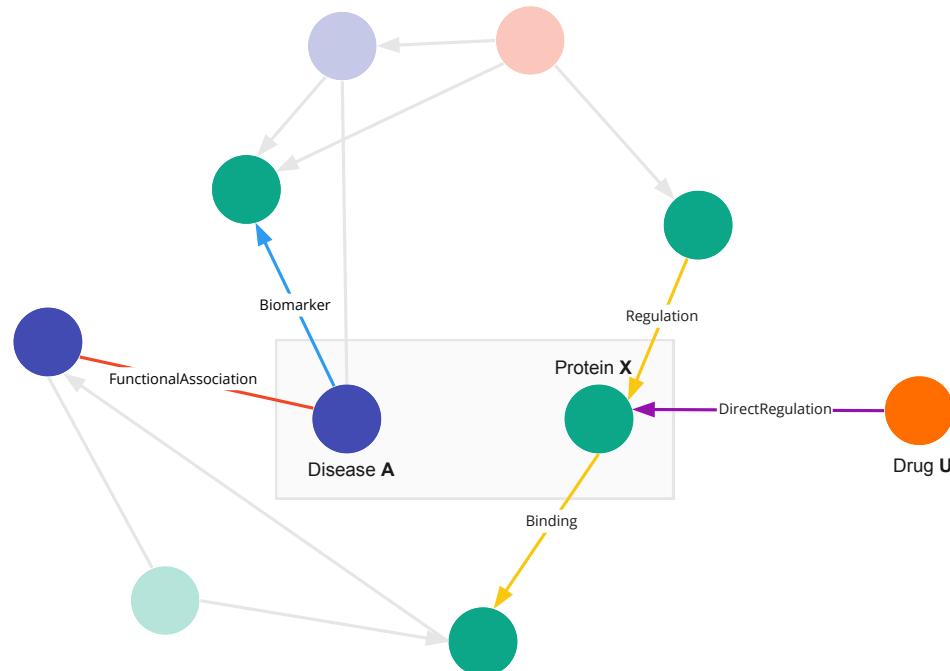
Can we predict potential new targets for specific cancer segments?

This could help researchers in:

1. Reprioritisation of existing leads.
2. Identification of new leads.

Intuitively what would we want to put into a model?

- Similarities between disease functions
- Known biomarkers of disease
- Protein interactions
- Known drug interactions
- ...



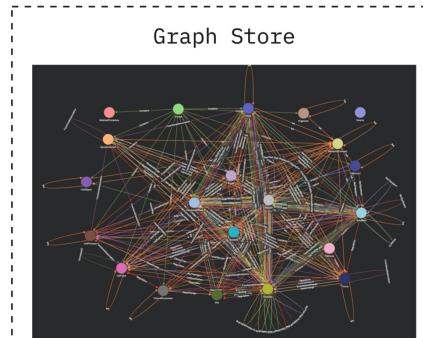
Apply ML to infer from graph what is important for new link prediction.

- How? - Train Knowledge Graph Embedding (KGE) Models to derive vector representations of entities and relations

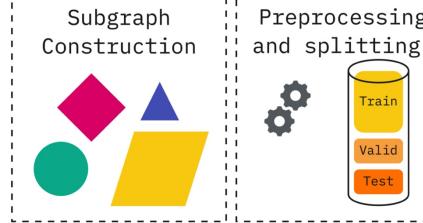
Building KGE models for link prediction

- An end to end pipeline for building, selecting and applying KGE models for link prediction*

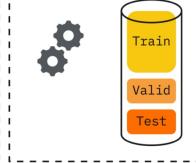
Data Selection



Subgraph Construction



Preprocessing and splitting



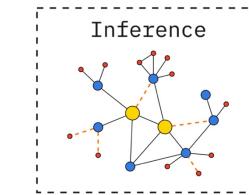
Model Training



Evaluation



Model Selection



*Utilised the PyKEEN library to support development

Ali, M., Berrendorf, M., Hoyt, C. T., Vermue, L., Sharifzadeh, S., Tresp, V., & Lehmann, J. (2021). PyKEEN 1.0: A Python Library for Training and Evaluating Knowledge Graph Embeddings. *Journal of Machine Learning Research*, 22(82), 1–6. Retrieved from <http://jmlr.org/papers/v22/20-825.html>



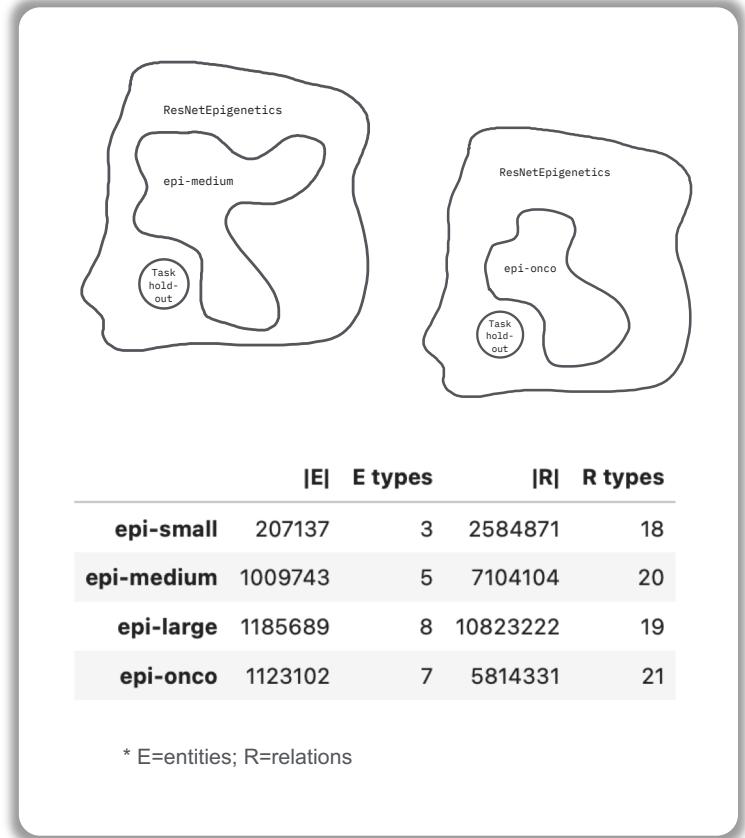
Data selection – Subgraph variants

Full graph contains

- 1.5M vertices
 - Include entities like protein, disease, but also processes, locations and groups (e.g. cell death, liver, complex)
- 13.2M edges (66.8M supporting references)
 - Physical interactions, disease/cell processes, gene expression

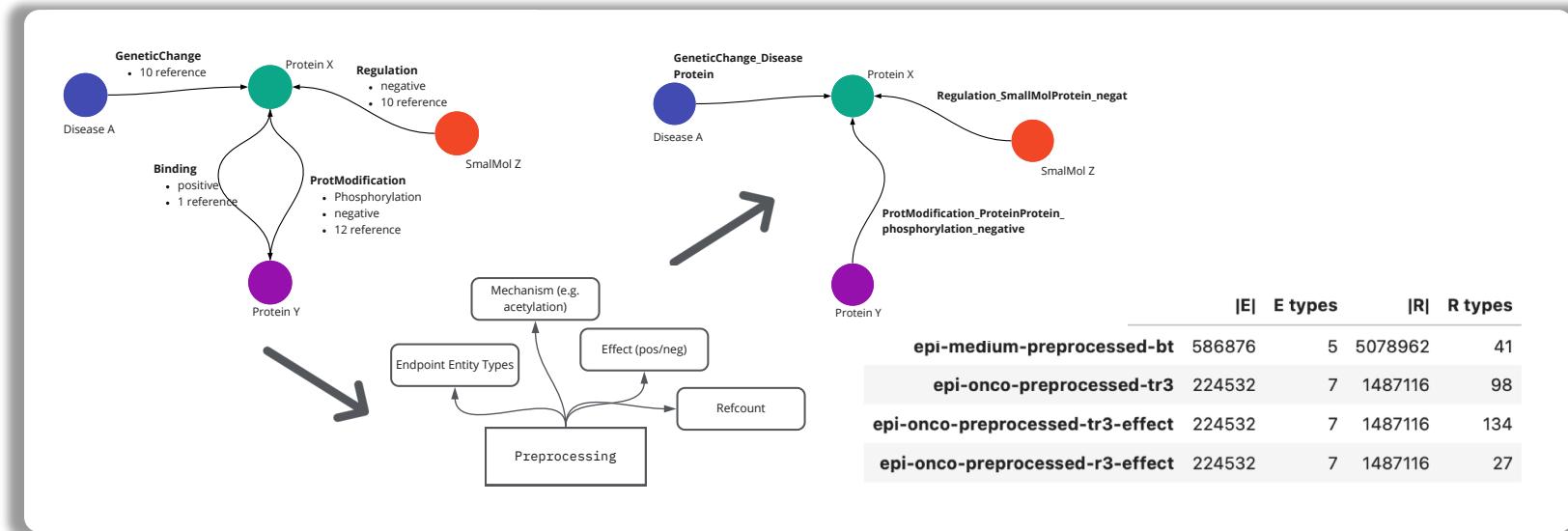
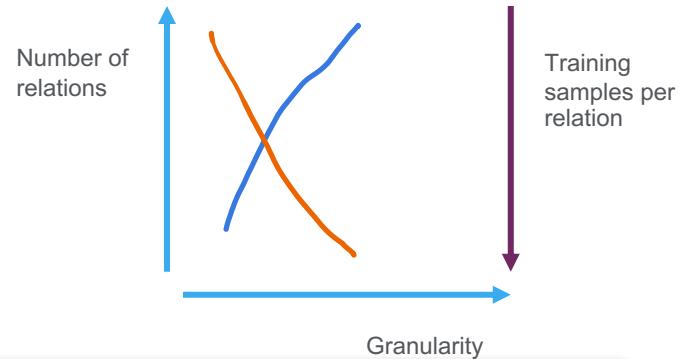
KGE models learn from triples, we have the freedom to choose what entities and relations to include

- Concise vs all-encompassing graphs
- Focus models on specific processes, e.g. PPI
- Isn't more data always better?



Data selection - Preprocessing

- Control granularity with various encodings
 - Can we make vague patterns more explicit?
- Trade-off granularity vs n_rels vs n_train
- Control confidence or FP rate with refcount filter



Model Selection - Data splitting and evaluation

Evaluation procedure

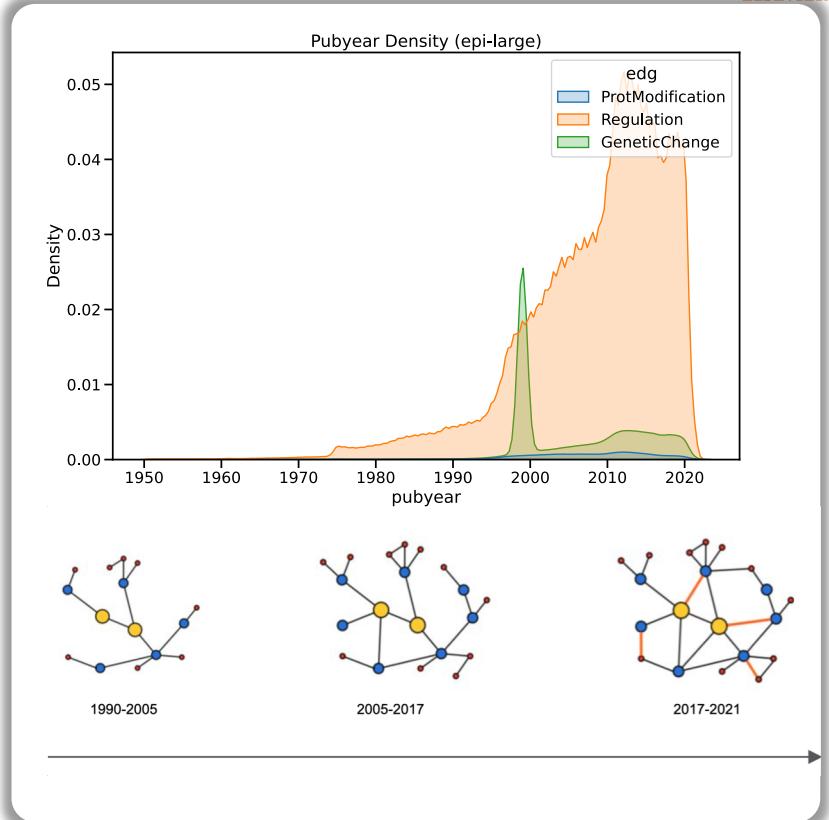
- Task holdout (disease target prediction) across all models and subgraphs
- Per subgraph 90/5/5 train/valid/test split
- Corroborate predictions with OpenTargetsPlatform

Option to split

- Randomly
- Time-based

Metrics from Information Retrieval

- Hits@k
- Variations of mean rank, e.g. Inverse Arithmetic Mean Rank (IAMR) - (0, 1]
- Evaluate models on hard task: all entities in context of all relations.
- New metrics pop up (Berrendorf et al., 2020)



Berrendorf, M., Faerman, E., Vermue, L., & Tresp, V. (2020). On the Ambiguity of Rank-Based Evaluation of Entity Alignment or Link Prediction Methods. arXiv. <https://doi.org/10.48550/ARXIV.2002.06914>

Model Selection - KG Embedding Models

- KG Embeddings (KGE) derive vector representations of entities and relations in the graph.
- Scores and ranks all possible entities by their likelihood of completing the link $\{ \text{head}, \text{rel}, \text{???} \}$.
- Lookup embeddings with enforced structure through scoring function, e.g. TransE vs ComplEx vs RotatE.

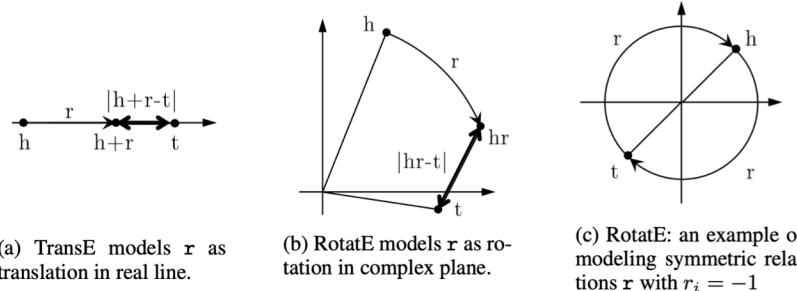


Figure 1: Illustrations of TransE and RotatE with only 1 dimension of embedding.

- TransE: Bordes, A., Usunier, N., Garcia-Durán, A., Weston, J., & Yakhnenko, O. (2013). Translating Embeddings for Modeling Multi-Relational Data. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2* (pp. 2787–2795). Red Hook, NY, USA: Curran Associates Inc.
- ComplEx: Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., & Bouchard, G. (2016). Complex Embeddings for Simple Link Prediction. arXiv.
<https://doi.org/10.48550/ARXIV.1606.06357>

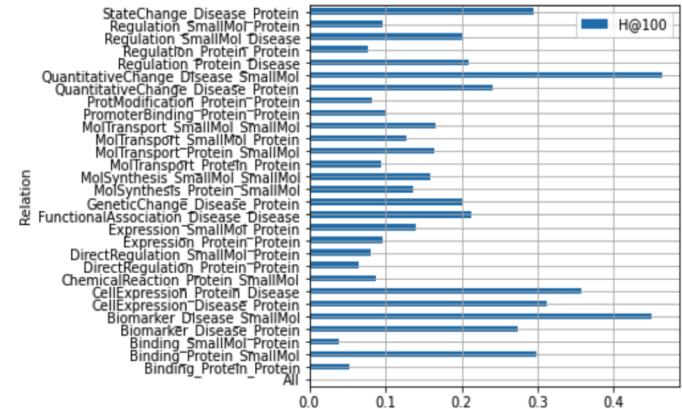
Model	Score Function	Symmetry	Antisymmetry	Inversion	Composition
SE	$-\ \mathbf{W}_{r,1}\mathbf{h} - \mathbf{W}_{r,2}\mathbf{t}\ $	✗	✗	✗	✗
TransE	$-\ \mathbf{h} + \mathbf{r} - \mathbf{t}\ $	✗	✓	✓	✓
TransX	$-\ g_{r,1}(\mathbf{h}) + \mathbf{r} - g_{r,2}(\mathbf{t})\ $	✓	✓	✗	✗
DistMult	$\langle \mathbf{h}, \mathbf{r}, \mathbf{t} \rangle$	✓	✗	✗	✗
ComplEx	$\text{Re}(\langle \mathbf{h}, \mathbf{r}, \mathbf{t} \rangle)$	✓	✓	✓	✗
RotatE	$-\ \mathbf{h} \circ \mathbf{r} - \mathbf{t}\ $	✓	✓	✓	✓

Table 2: The pattern modeling and inference abilities of several models.

Resources from: Sun, Z., Deng, Z.-H., Nie, J.-Y., & Tang, J. (2019). RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. arXiv.
<https://doi.org/10.48550/ARXIV.1902.10197>

Model Selection - Results and nuances

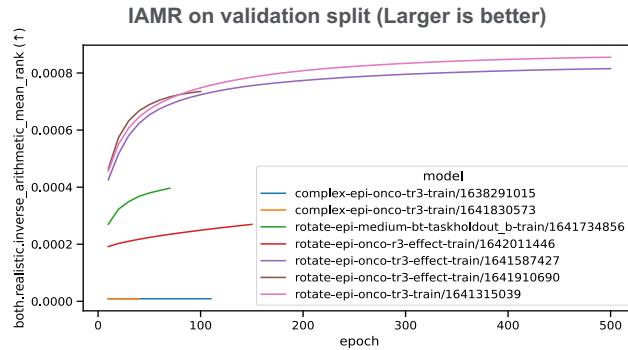
- Choice of meaningful metric on single model
 - Tradeoff between Link Prediction training objective and end-task specificity
 - Link Prediction KGE models are trained to learn embedding representations of KG entities and relations such that they can predict ‘all’ relations in context of each other, and not singular relation task
 - However, end task favours selecting model based on performance of predicting disease-gene association relations only
- Metric type
 - Ranking and information retrieval metrics used, and not classification metrics of PRF
 - Graph density influences prediction performance -> Model is better at predicting ranks of high-degree nodes in test set, rather than in sparser regions?



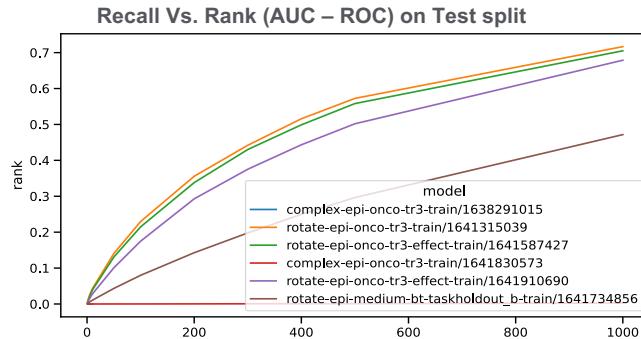
Model Selection - Results and nuances

- Choice across models trained with different subgraph and algorithm variants

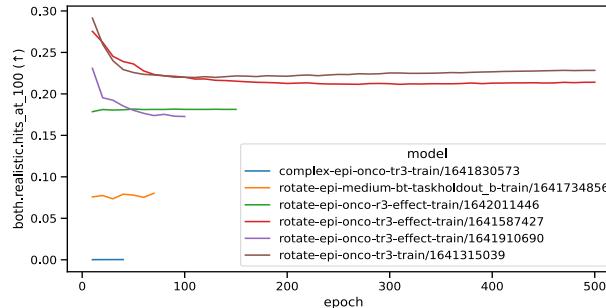
Training curves



Recall Vs Rank curves



Hits@100 on validation split (Larger is better)



Observations

- Oncology specific smaller subgraph > larger subgraphs
- RotatE > ComplEx for our data
- Is this a consequence of structure in underlying data model?
- Proxy measures indicate underlying data model to have xx extent of composition relations

Model Selection - Hyperparameter optimisation

- KGE models very sensitive to training setup, hyperparameters, parameter initialisation seeds and different splits in the dataset. (Bonner et al., 2021)
- Need to experiment A LOT
- Can we be smart about how we expend computational budget?

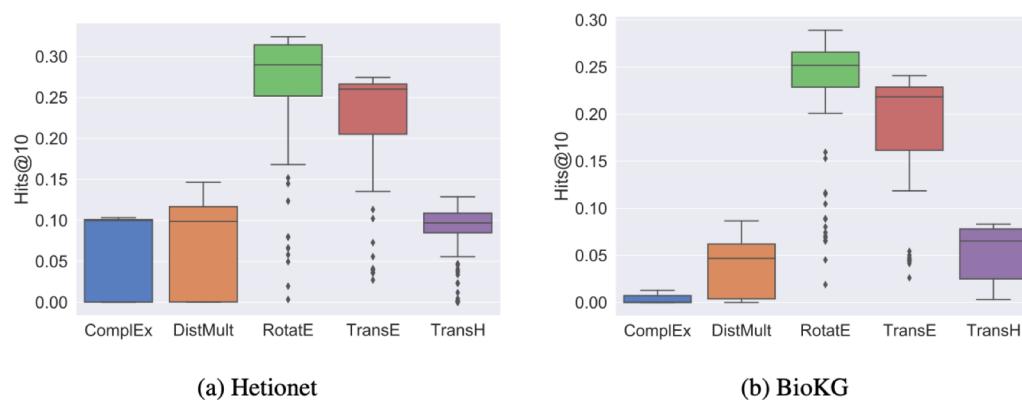


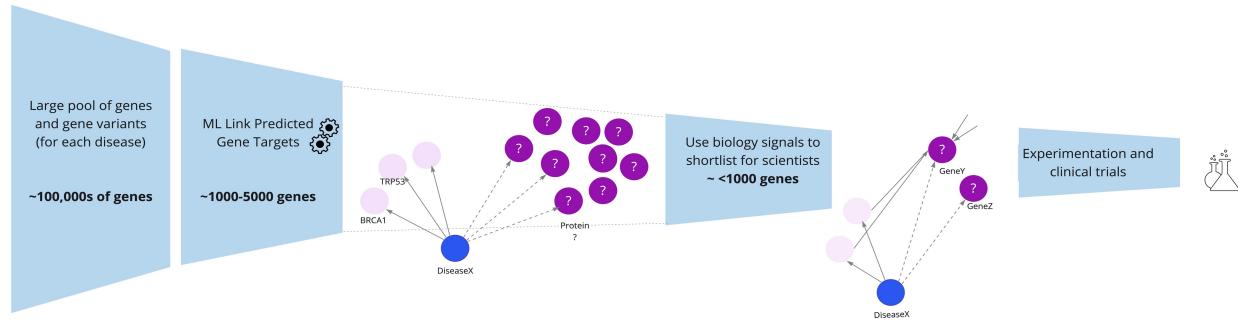
Figure 4: Distribution of the Hits@10 scores across all 100 runs of different parameters.

Interpreting ML scores - Validating target predictions with experts



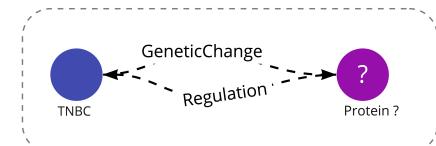
How to validate a “novel” / unseen gene target prediction?

- Experts assess plausibility of predicted targets
- Model outputs large number of predictions. To shortlist before expert review, infuse biology priors, and reduce ‘desk to lab-bench time’



How did we shortlist?

- Reinforce prediction signals -> An ideal gene target should have multiple desired characteristics (Ex: Expression, druggability, mode of action, etc).
Obtain intersection of genes predicted to have multiple association routes



Today



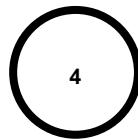
**Background: EpiMap –
Why and what**



**Building KGE models for
link Prediction**



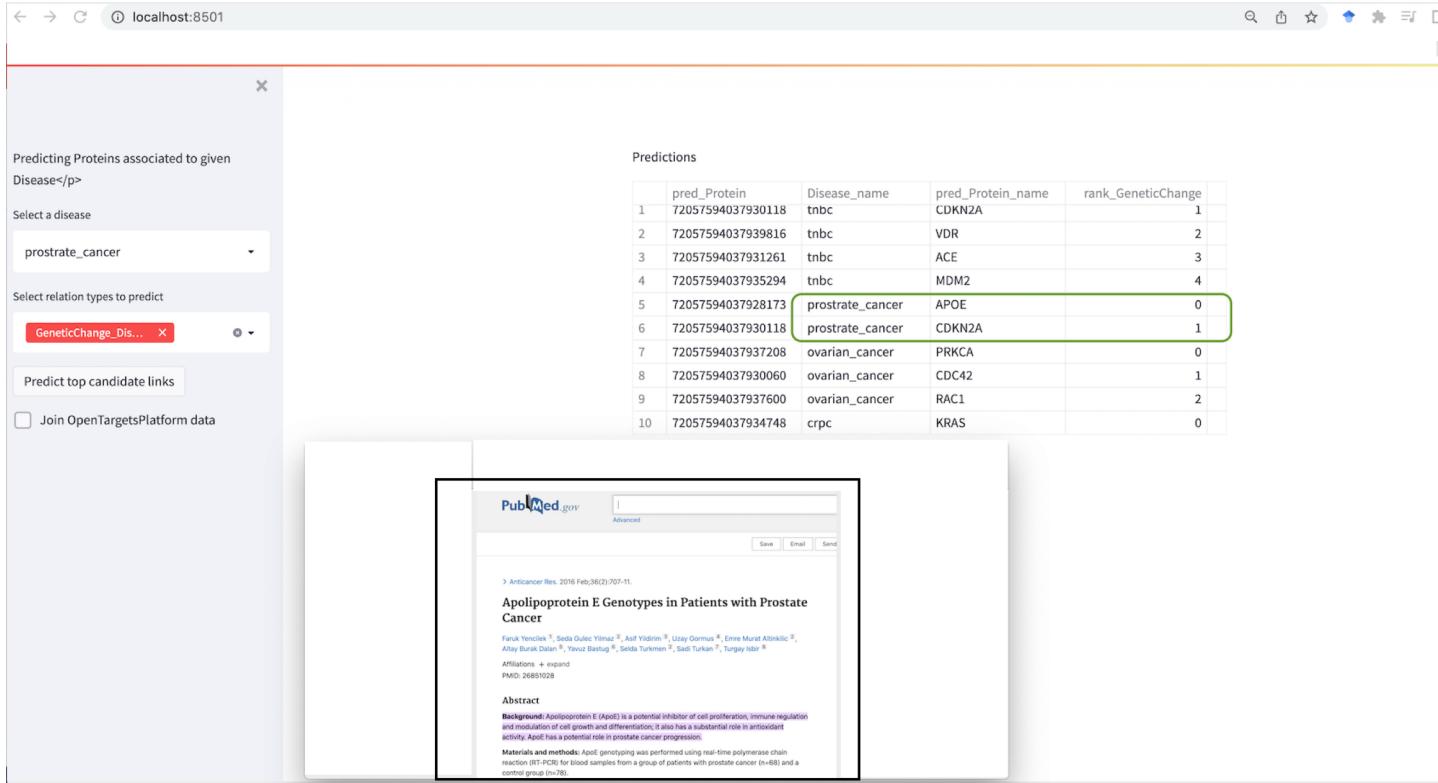
**Results & Challenges
in KG ML**



Applications and Impact

Demo

Quick look at KGE based Link Predictions for Disease Target associations



Predicting Proteins associated to given Disease</p>

Select a disease

prostate_cancer

Select relation types to predict

GeneticChange_Dis... X

Predict top candidate links

Join OpenTargetsPlatform data

	pred_Protein	Disease_name	pred_Protein_name	rank_GeneticChange
1	72057594037930118	tnbc	CDKN2A	1
2	72057594037939816	tnbc	VDR	2
3	72057594037931261	tnbc	ACE	3
4	72057594037935294	tnbc	MDM2	4
5	72057594037928173	prostate_cancer	APOE	0
6	72057594037930118	prostate_cancer	CDKN2A	1
7	72057594037937208	ovarian_cancer	PRKCA	0
8	72057594037930060	ovarian_cancer	CDC42	1
9	72057594037937600	ovarian_cancer	RAC1	2
10	72057594037934748	crpc	KRAS	0

PubMed.gov

> Anticancer Res. 2016 Feb;36(2):707-11.

Apolipoprotein E Genotypes in Patients with Prostate Cancer

Furuk Yenicek ¹, Seda Gulec Yilmaz ², Asif Yildirim ³, Uzay Gormus ⁴, Emre Murat Attilikci ³, Altay Burak Duman ⁵, Yavuz Bastug ⁶, Selda Turkmen ³, Sadi Turkmen ⁷, Turgay Isbir ⁸

Affiliations + expand

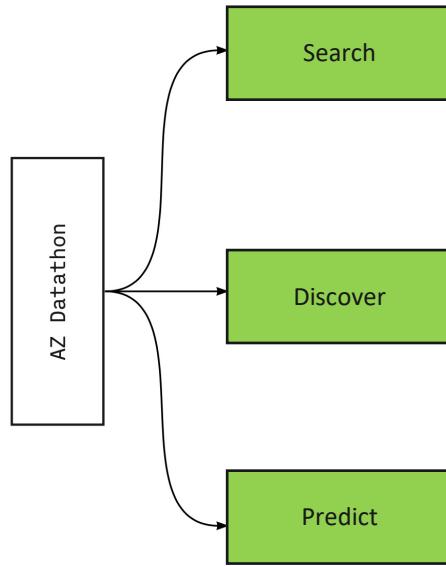
PMID: 26861028

Abstract

Background: Apolipoprotein E (Apol) is a potential inhibitor of cell proliferation, immune regulation and modulation of cell growth and differentiation. It also has a substantial role in antioxidant activity. Apol has a potential role in prostate cancer progression.

Materials and methods: Apol genotyping was performed using real-time polymerase chain reaction (RT-PCR) for blood samples from a group of patients with prostate cancer (n=68) and a control group (n=78).

Validation of EpiMap results with AZ knowledge/expertise shows value in Knowledge Graph approach



EpiMap helps reprioritize internal Epigenetic targets

Pan-indication epigenetic signature



Co-location/modulation with drugs of interest



EpiMap identifies resistance paths in patients treated with PARPi*

Identified established & validated resistance mechanisms
Identified >100 testable pathways of drug resistance.



Expanded our search space by ~10x



Parting note

- First text-mined epigenetics Knowledge Graph spanning disease segments, including 13.2M context-specific relationships mined from 7M documents.
- Demonstrated that traversable FAIR KGs derived from scientific literature are valuable resources in complex domains, and complementary to scientist's expertise through the scale and usability they offer.
- ML applied to KGs, such as link prediction, can help **discover and prioritise potential therapeutic interventions** and improve **understanding of disease biology, mechanisms of drug resistance**, and more.
- Nuances in KGE for link prediction
 - Performance of KGE models affected by many factors: No free lunch!
 - Predictions are context dependent and infusing biological signals assists scientists in novelty validation.
 - Validation of 'novel' target candidates is hard, but ongoing validation by AstraZeneca experts seems promising
- Identified need for ranked hypothesis rather than ranked genes

References



- TransE: Bordes, A., Usunier, N., Garcia-Durán, A., Weston, J., & Yakhnenko, O. (2013). Translating Embeddings for Modeling Multi-Relational Data. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2* (pp. 2787–2795). Red Hook, NY, USA: Curran Associates Inc.
- ComplEx: Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., & Bouchard, G. (2016). Complex Embeddings for Simple Link Prediction. arXiv. <https://doi.org/10.48550/ARXIV.1606.06357>
- Ali, M., Berrendorf, M., Hoyt, C. T., Vermue, L., Sharifzadeh, S., Tresp, V., & Lehmann, J. (2021). PyKEEN 1.0: A Python Library for Training and Evaluating Knowledge Graph Embeddings. *Journal of Machine Learning Research*, 22(82), 1–6. Retrieved from <http://jmlr.org/papers/v22/20-825.html>
- Bonner, S., Barrett, I.P., Ye, C., Swiers, R., Engkvist, O., Hoyt, C.T., Hamilton, W.L.: Understanding the performance of knowledge graph embeddings in drug discovery (2021). <https://doi.org/10.48550/ARXIV.2105.10488>, <https://arxiv.org/abs/2105.10488>
- Rossi, A., & Matinata, A. (2020). Knowledge Graph Embeddings: Are Relation-Learning Models Learning Relations? In *EDBT/ICDT Workshops*.
- Berrendorf, M., Faerman, E., Vermue, L., Tresp, V.: Interpretable and fair comparison of link prediction or entity alignment methods with adjusted mean rank. arXiv preprint arXiv:2002.06914 (2020)
- Berrendorf, M., Faerman, E., Vermue, L., & Tresp, V. (2020). On the Ambiguity of Rank-Based Evaluation of Entity Alignment or Link Prediction Methods. arXiv. <https://doi.org/10.48550/ARXIV.2002.06914>



- <Placeholder for sample tutorial notebook using public data>
- Related Work:  **DiscoveryLab**

Also find out more about our AI on KG related research at
discoverylab.ai or icai.ai/discovery-lab



Want to know more? Drop us a line:

- p.mitra@elsevier.com
- t.pijnenburg@elsevier.com
- t.slater@elsevier.com

