Stanford University

Bridging the Usability-Expressivity Gap in Biomedical Data Discovery

STANFORD SCHOOL OF MEDICINE
Stanford University Medical Center

Maulik R. Kamdar (maulikrk@stanford.edu)

Motivation

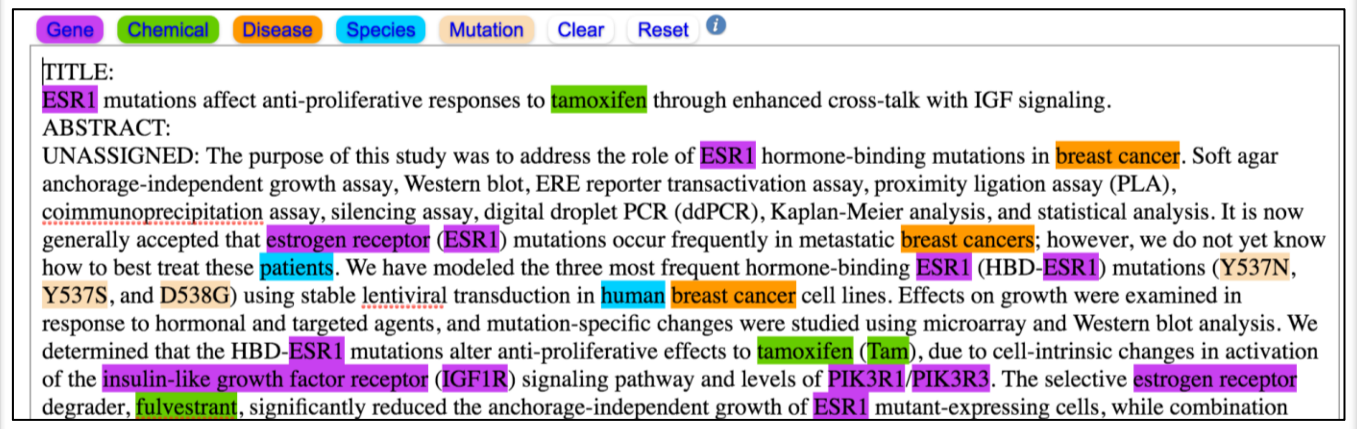
- Questions asked during the Hypothesize-test-evaluate cycle: *"Which antineoplastic agents target IDH1 gene in glioma patients?"*
- Requires precise answers (in above example, the corresponding drugs) or relevant *-omics* datasets for further analysis.
- Semantic Web and Linked Data technologies (RDF and SPARQL) towards tackling the integrative bioinformatics challenges.
- Querying Life Sciences Linked Open Data (LSLOD) Cloud has a steep learning curve.
- Natural language querying method over the LSLOD network to enable scalable, autonomous discovery of relevant answers and datasets for evaluating hypotheses.

Word Embeddings

PubMed
MEDLINE U.S. National Library of Medicine

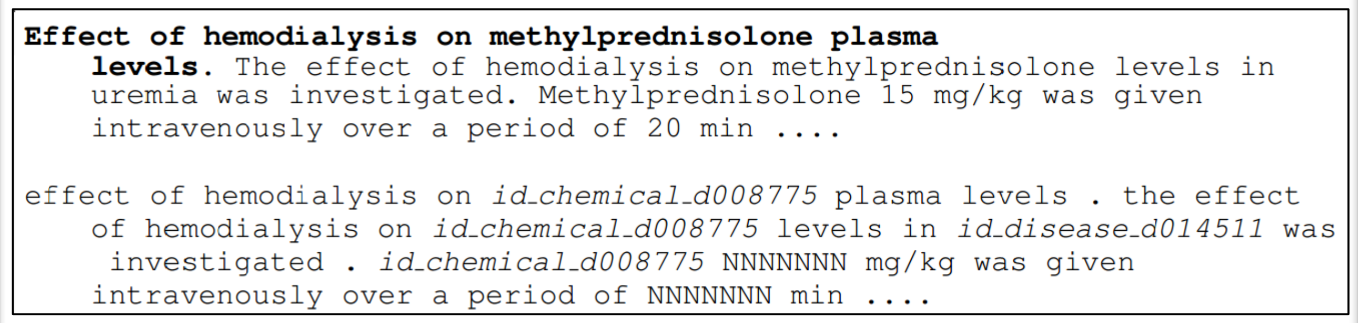
1.5B+ abstracts, 3B+ Tokens
2.2M+ unique words
0.14M biomedical entities

PubTator ⇓ Annotation



TITLE:
ESR1 mutations affect anti-proliferative responses to tamoxifen through enhanced cross-talk with IGF signaling.
ABSTRACT:
UNASSIGNED: The purpose of this study was to address the role of ESR1 hormone-binding mutations in breast cancer. Soft agar anchorage-independent growth assay, Western blot, ERE reporter transactivation assay, proximity ligation assay (PLA), coimmunoprecipitation assay, silencing assay, digital droplet PCR (ddPCR), Kaplan-Meier analysis, and statistical analysis. It is now generally accepted that estrogen receptor (ESR1) mutations occur frequently in metastatic breast cancers; however, we do not yet know how to best treat these patients. We have modeled the three most frequent hormone-binding ESR1 (HBD-ESR1) mutations (Y537N, Y537S, and D538G) using stable lentiviral transduction in human breast cancer cell lines. Effects on growth were examined in response to hormonal and targeted agents, and mutation-specific changes were studied using microarray and Western blot analysis. We determined that the HBD-ESR1 mutations alter anti-proliferative effects to tamoxifen (Tam), due to cell-intrinsic changes in activation of the insulin-like growth factor receptor (IGF1R) signaling pathway and levels of PIK3R1/PIK3R3. The selective estrogen receptor degrader, fulvestrant, significantly reduced the anchorage-independent growth of ESR1 mutant-expressing cells, while combination

⇓ Tokenization



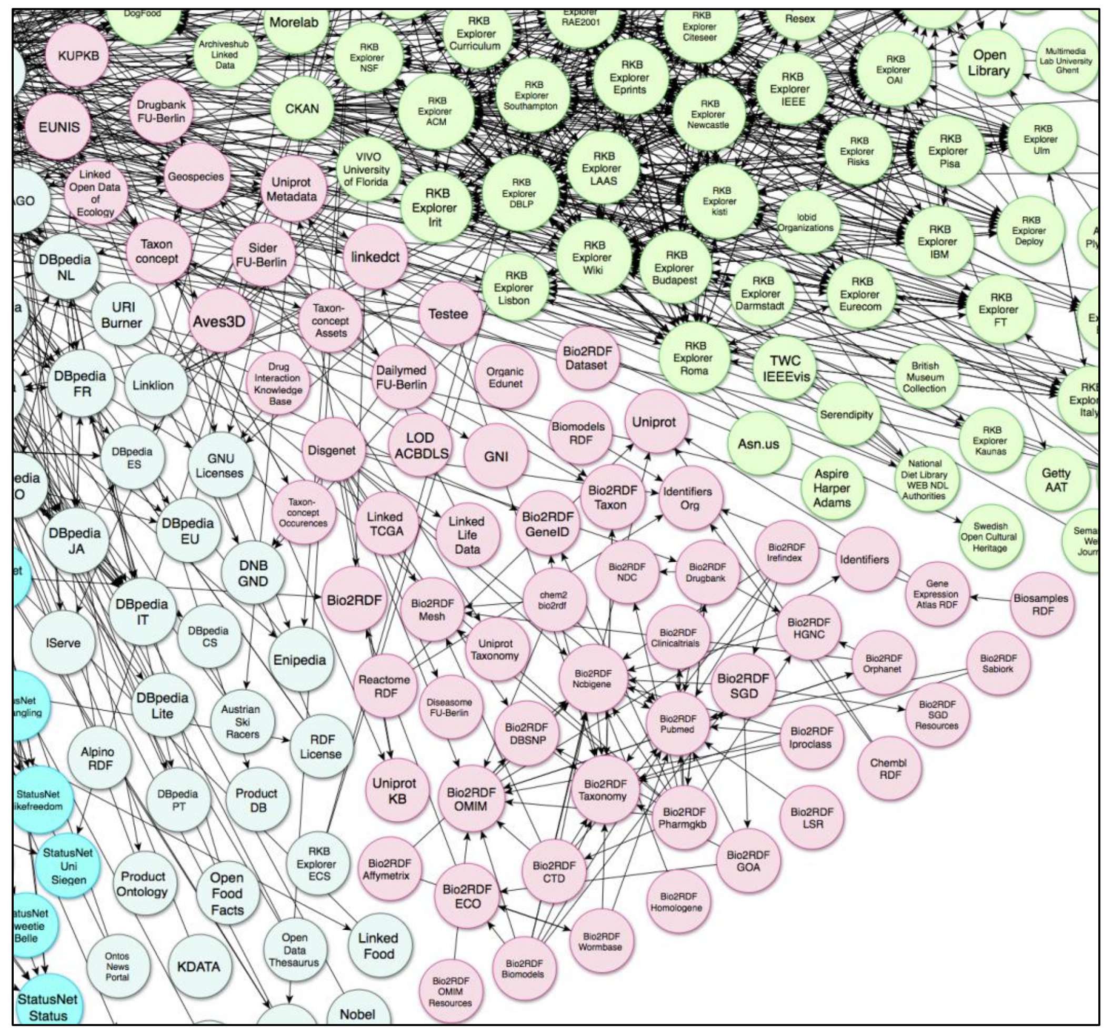
```
Effect of hemodialysis on methylprednisolone plasma
    levels. The effect of hemodialysis on methylprednisolone levels in
    uremia was investigated. Methylprednisolone 15 mg/kg was given
    intravenously over a period of 20 min ....

effect of hemodialysis on id_chemical_d008775 plasma levels . the effect
    of hemodialysis on id_chemical_d008775 levels in id_disease_d014511 was
     investigated . id_chemical_d008775 NNNNNNN mg/kg was given
    intravenously over a period of NNNNNNN min ....
```

⇓

GloVe

Iterations: 20, α: 0.75
Dimensions: 100

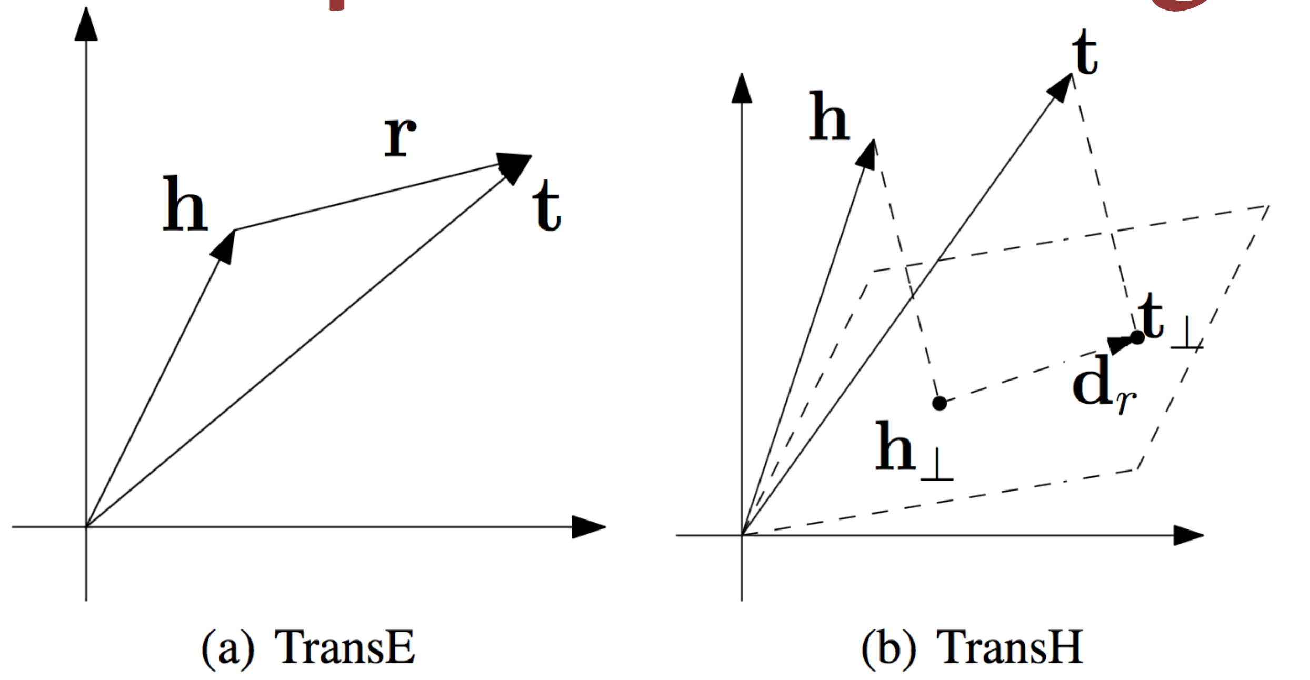
Linked Data



Reference: http://lod-cloud.net/

- Life Sciences Linked Open Data Cloud: 1T+ triples from 80+ biomedical sources.
- Develop Bio-mashups and Linked Biomedical Dataspaces facilitating *in silico* data discovery

Graph Embeddings



(a) TransE (b) TransH

$$f_r(\mathbf{h},\mathbf{t}) = \|(\mathbf{h} - \mathbf{w}_r^\top \mathbf{h}\mathbf{w}_r) + \mathbf{d}_r - (\mathbf{t} - \mathbf{w}_r^\top \mathbf{t}\mathbf{w}_r)\|_2^2$$

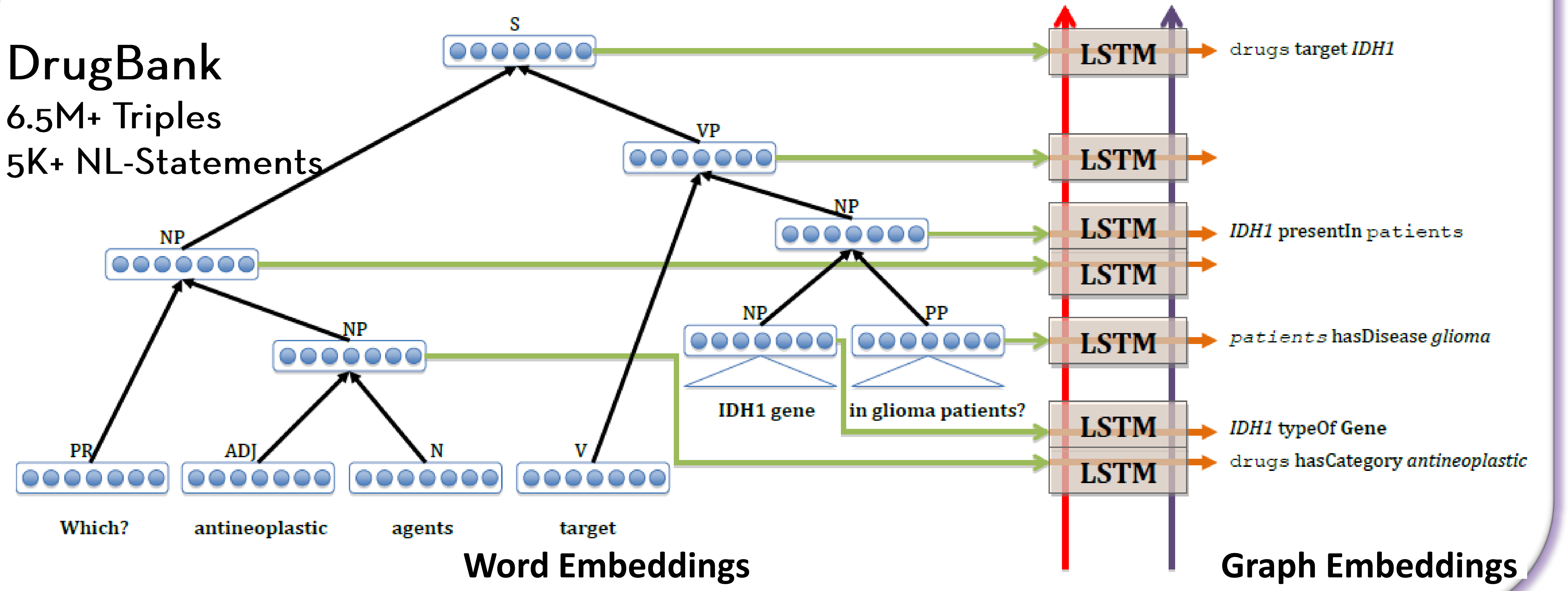
$$\mathcal{L} = \sum_{(h,r,t)\in\Delta} \sum_{(h',r',t')\in\Delta'_{(h,r,t)}} \left[f_r(\mathbf{h},\mathbf{t}) + \gamma - f_{r'}(\mathbf{h}',\mathbf{t}')\right]_+ + C\left\{\sum_{e\in E}\left[\|\mathbf{e}\|_2^2 - 1\right]_+ + \sum_{r\in R}\left[\frac{(\mathbf{w}_r^\top \mathbf{d}_r)^2}{\|\mathbf{d}_r\|_2^2} - \epsilon^2\right]_+\right\}$$

TransH: 6.5M+ DrugBank Triples, α = 0.05, Dimensions = 100, λ = 1, C = 0.25,

Reference: Wang, Zhen, et al. "Knowledge Graph Embedding by Translating on Hyperplanes." *AAAI*. 2014.

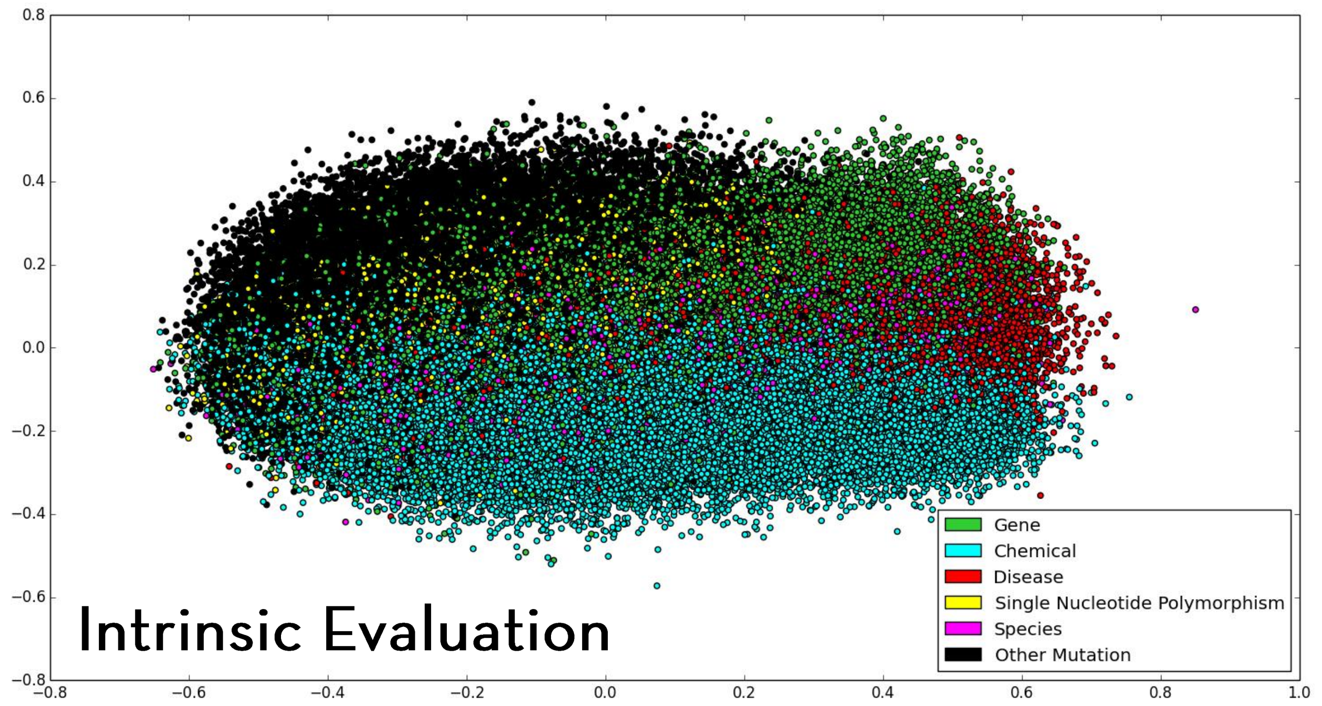
Neural Network Architecture

DrugBank
6.5M+ Triples
5K+ NL-Statements



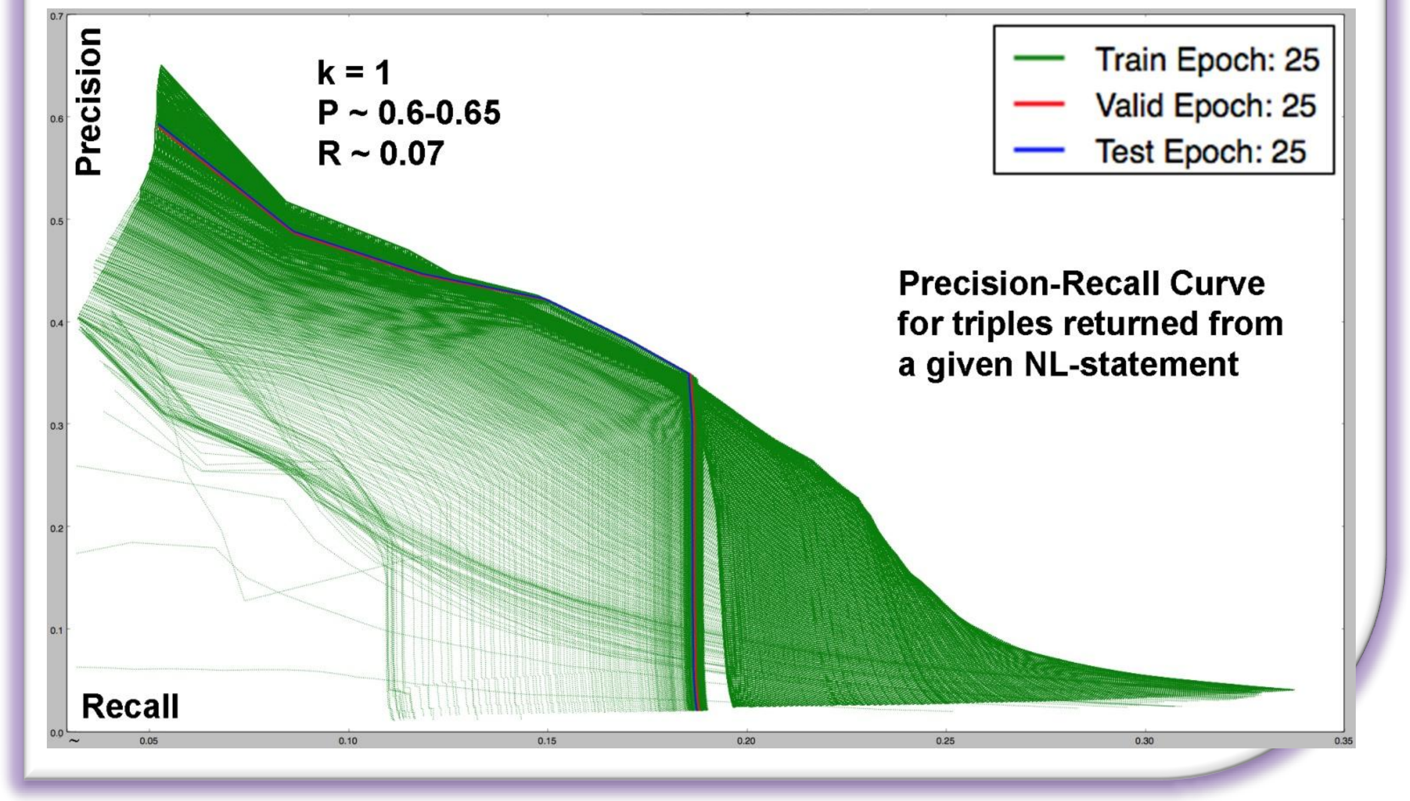
Acknowledgements I would like to acknowledge CS 224D staff and Dr. Mark Musen

Evaluation



Extrinsic Evaluation: MESH Recommendations

- •Each biomedical abstract is provided with Medical Subject Headings (MESH) manually.
- •**Train:** 8.3M+, **Valid:** 2.7M+, **Test:** 2.7M+ abstracts
- •Baseline is KNN algorithm and simple TF-IDF vectors (K=6):- Precision: 0.31, Recall: 0.45
- •2-layer Neural Network, with 20 and 300 hidden nodes respectively, 25 Epochs, 27K MESH Terms (α= 0.001, λ= 0.001) :- Precision: 0.86, Recall: 0.22.



Conclusion

- Word & Graph embeddings and Neural Network architecture that can translate NL-queries to structured triples, and provide a usable interface to tackle integrative bioinformatics challenges.
- More rigorous evaluations of graph embeddings and method, and scale it to LSLOD network.