COMPUTATIONAL DRUG REPOSITIONING USING KNOWLEDGE GRAPH AND AI

The LNM Institute of Information Technology

Team Members

Om Gupta	21ucs140
Pitani Nitin Srichakri	21ucs149
Tarun Agarwal	21ucs219

Guided by
Dr. Abhijit Adhikari

Co-guided by
Dr. Poulami Dalapati

TABLE OF CONTENTS

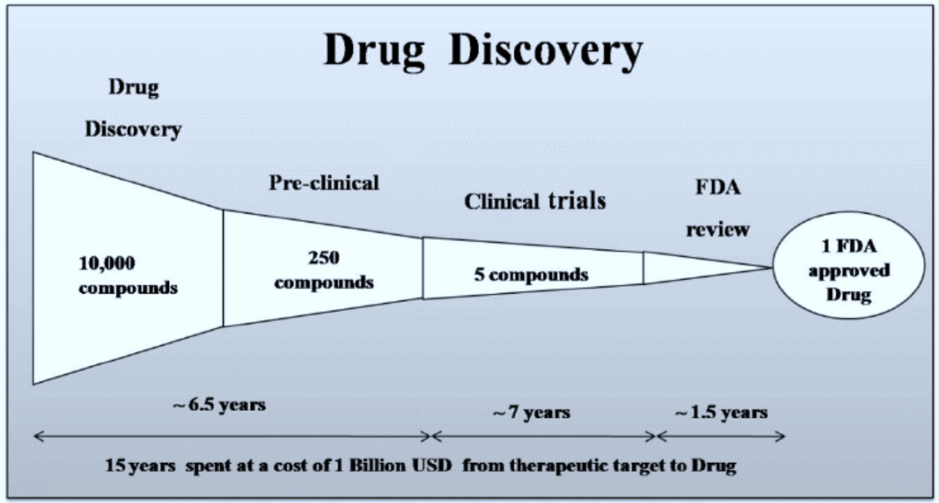
INTRODUCTION

- What is Drug Discovery ?
  - Drug discovery is the process by which new medications are identified and developed.
- What are stages of drug discovery ?
  - Identification of a target
  - Lead compound development
  - Preclinical and clinical testing

Target Discovery → Target Validation → Lead Optimization → Development → Trials

INTRODUCTION

- Limitations of Drug discovery
  - Costly and lengthy process.
  - Four to five main stages.



This is where "Drug Repurposing" is helping

INTRODUCTION

- Areas where Computer Science can contribute?
  - Suggesting candidate drugs.
  - Risk Assessment done by FDA.
  - Dose-Response Analysis can be helped through data science.

- The research paper aims to contribute by identifying existing drugs that could potentially be repurposed for the treatment of other diseases.

INTRODUCTION

- Which tools can be used in perform the above task?
  - Clustering and Classification Algorithms
    - Meticulous analysis and pattern recognition by grouping similar compounds and predicting their efficacy efficiently identify potential drug candidates
    - Undergone extensive research and has reached a saturation point.
  - Regression Analysis
    - Modeling relationships between variables aid in predicting drug effectiveness but not efficient in suggesting candidate drugs.
    - Cornerstone in the realm of drug discovery
  - Deep Learning
    - Forefront of drug discovery research offering immense promise.
    - Our base paper also focuses here.

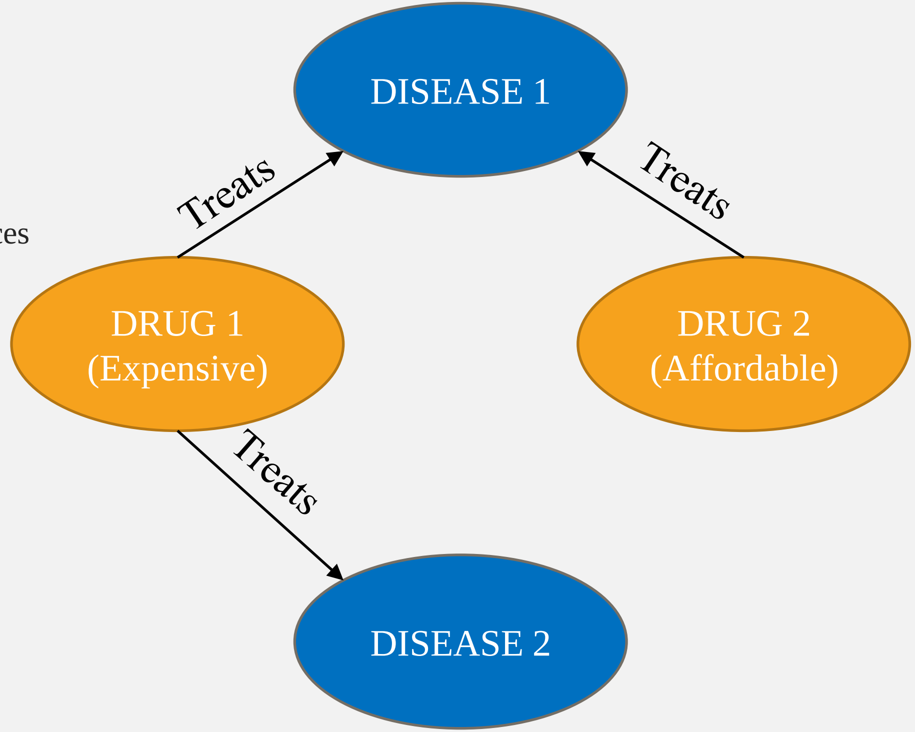
MOTIVATION


Drug repurposing
De novo drug design
Takes 3-12 years
Takes 10-15 years
Has 30-75% success rate
Has < 10% success rate
Costs up to $300 million
Costs up to $2.5 billion

SOME DEFINITIONS

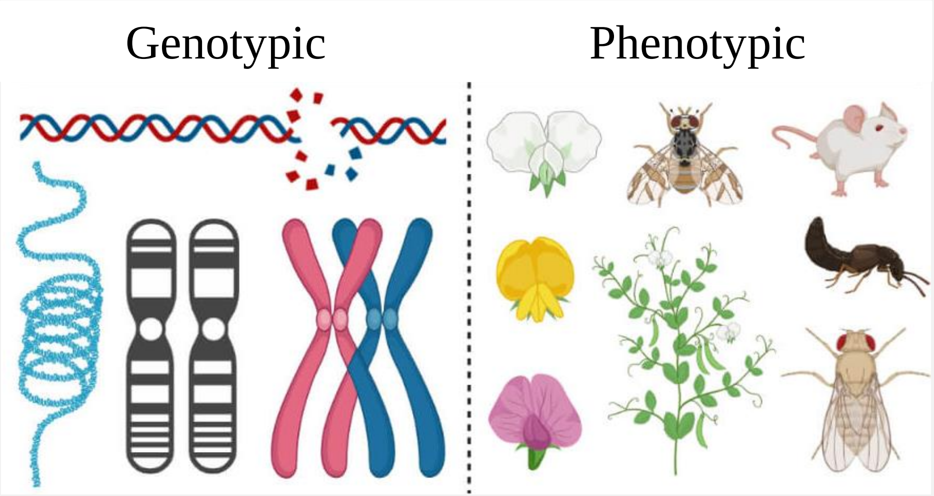
- Knowledge Graph
    - Represents a network of real-world entities
    - Also known as a semantic network
    - Typically made up of datasets from various sources

Our base paper leverages Knowledge Graph technology to facilitate the process of drug repositioning.



SOME DEFINITIONS

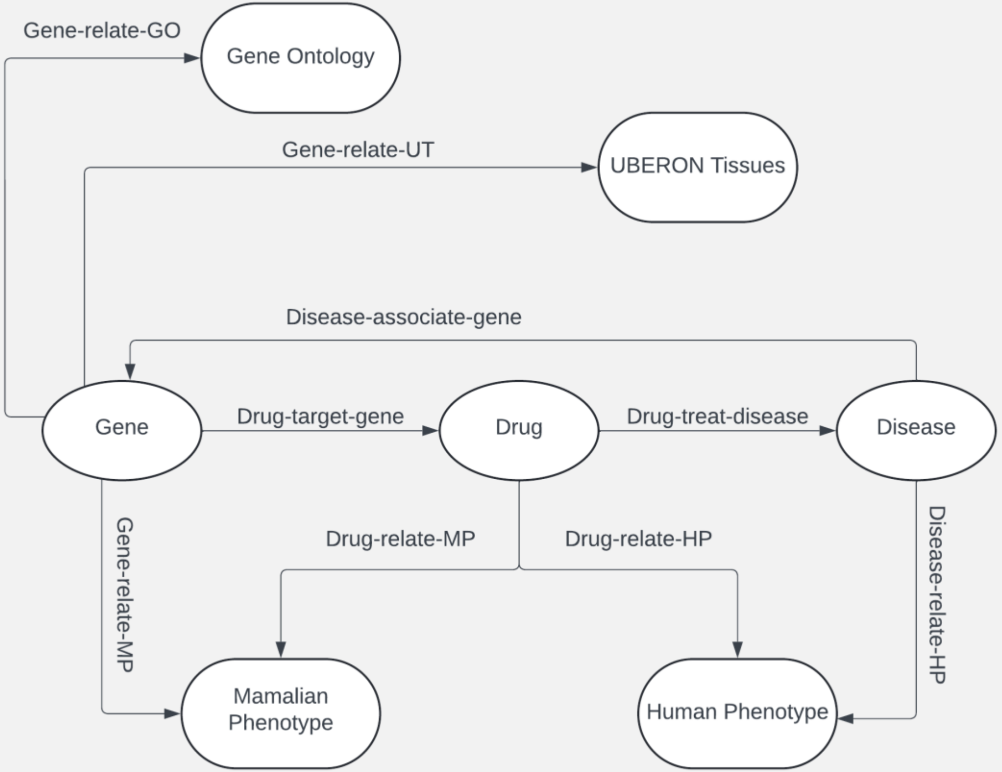
- What is Genotypic and Phenotypic?
  - Genotypic genes are responsible for unique traits or characteristics.
  - Phenotypic genes are responsible for physical appearance or characteristic of an organism.



SOME DEFINITIONS

- Knowledge Graph Construction
  - 61,146 nodes with 7 node types
  - 1,246,726 edges
  - Categorized into 9 semantic relationships
    - 2 from genome-level knowledge databases
    - 1 from text-mined knowledge bases
    - 6 from phenome-level knowledge databases

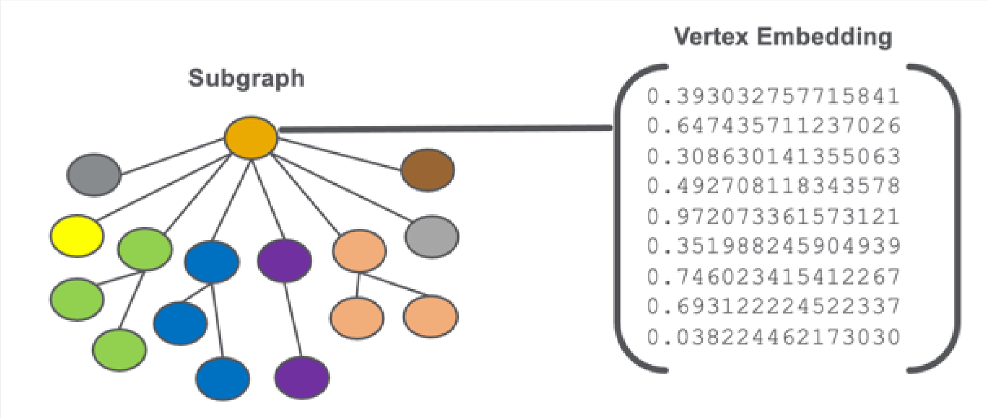
LITERATURE SURVEY



SOME DEFINITIONS

- Traditional machine learning or deep learning algorithms typically operate on fixed-size vectors or matrices.

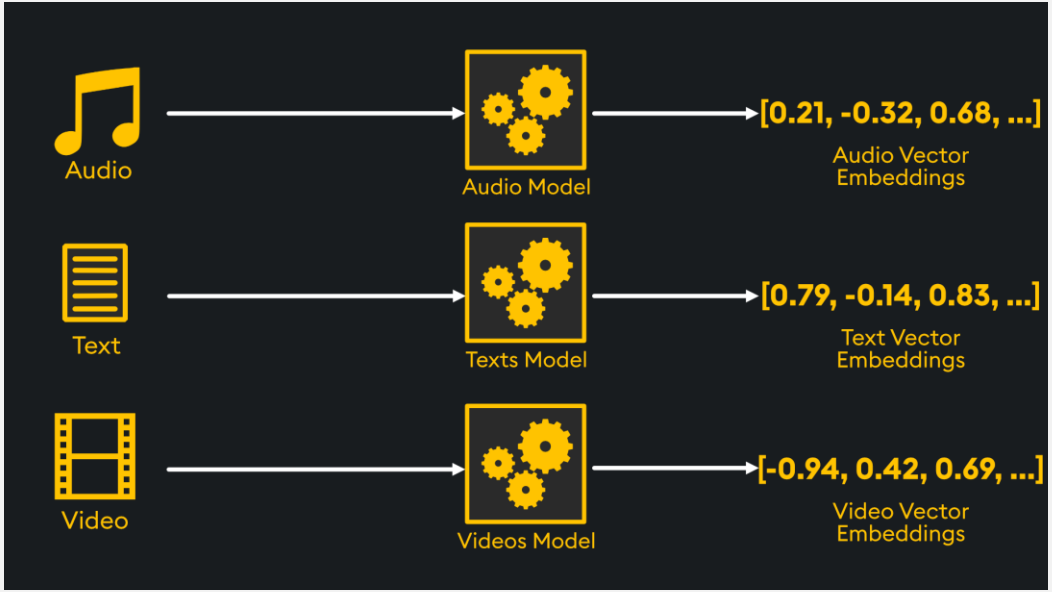
- How to convert this complex graph into a simpler representation while preserving important information?
    - Graph Embedding - Embeddings encode objects or concepts in a continuous vector space.



SOME DEFINITIONS

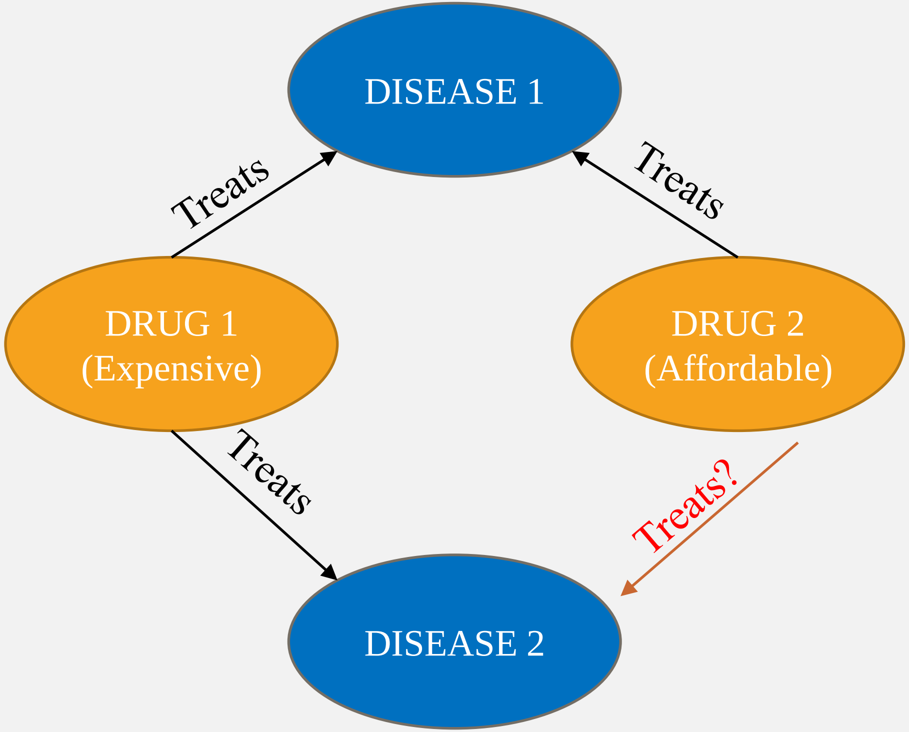
LNMIIT
The LNM Institute of Information Technology

- What do these Numbers Signify?



KNOWLEDGE GRAPH COMPLETION

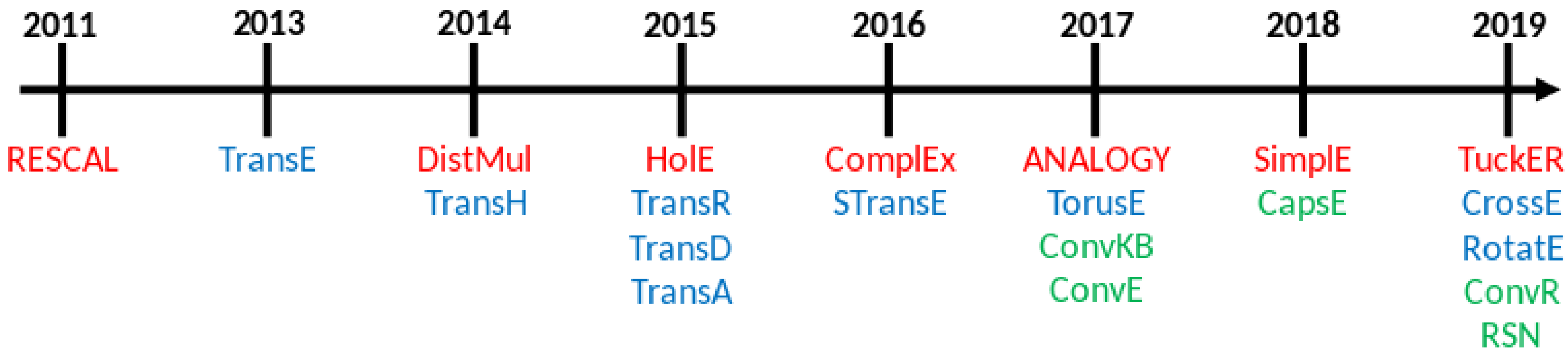
- Given an enormous KG, can we complete the KG?
  - For a given (head, relation), we predict missing tails.
- The graph should contain multiple types of relationships otherwise, you can use the standard link prediction techniques.
- This graph could also include the origin of drugs and the relationship between them.



KG REPRESENTATION

- Edges in KG are represented as triples (h, r, t)
    - head (h) has relation (r) with tail (t)

- Associate entities and relations with shallow embeddings
- Given a triple (h, r, t) the goal is that the embedding of (h, r) should be close to the embedding of t.
    - How to embed (h, r) ?
    - How to define score $f_r\,(h, t)$?
    - Score f_r is high if (h, r, t) exists, else f_r is low



KG EMBEDDING MODELS

LITERATURE SURVEY

- We started looking into pre-existing state of arts model involving different embeddings.
  - TransE
    - Given by- *A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston and O. Yakhnenko.*
  - RotatE
    - Given by- *Z. Sun, Z.H. Deng, J.Y. Nie and J. Tang.*
  - DistMult
    - Given by- *B. Yang, W.T. Yih, X. He, J. Gao and L. Deng*

FOUR RELATION PATTERNS

- Symmetric (Antisymmetric) Relations:
  - $r(h,t) \Rightarrow r(t,h)$ and $r(h,t) \Longrightarrow \neg r(t,h) \quad \forall h,t$
  - Example:
    - Symmetric – Family, Roommate
    - Antisymmetric – Hypernym (colour and red)

- Inverse Relations:
  - $r_1(h,t) \Longrightarrow r_2(t,h)$
  - Example: Advisor and Advisee

FOUR RELATION PATTERNS

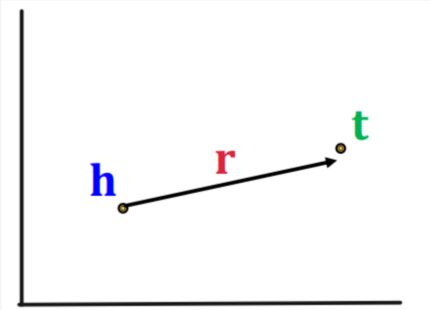
- Composition (Transitive) Relations:
  - $r_1(x, y)$ & $r_2(y, z) \Rightarrow r_3(x, z) \quad \forall x, y, z$
  - Example: My mother's husband is my father.

- 1-to-N Relations:
  - $r(h, t_1),\ r(h, t_2),\ \ldots, r(h, t_n)$ are all true.
  - Example: ***r*** is the '*student_of*'

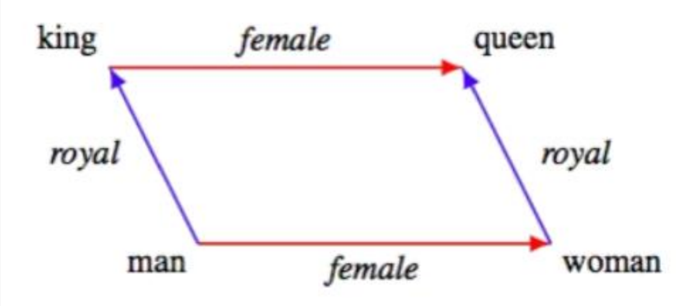
TransE

LNMIIT
The LNM Institute of Information Technology

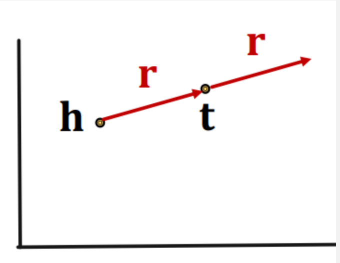
- Translation:
  - For a triple (h, r, t), let $\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t} \in \mathbf{R}^{\mathbf{d}}$ be embedding vectors.
    - if the given link exists - $\boldsymbol{h} + \boldsymbol{r} \approx \boldsymbol{t}$
    - else $\boldsymbol{h} + \boldsymbol{r} \neq \boldsymbol{t}$



- Entity scoring function: $f_r(h, t) = -|| \, h + r \, - t \, ||$

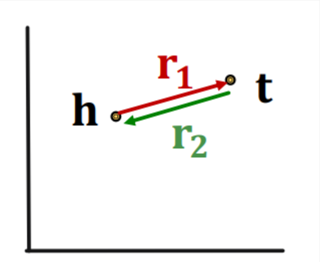


TransE



Antisymmetric:
$r(h,t) \Longrightarrow \neg r(t,h)$

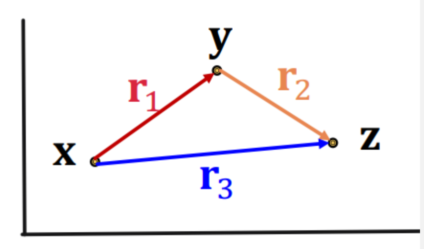
$h + r = t \; but \; t + r \neq h$



Inverse:
$r_1(h,t) \Longrightarrow r_2(t,h)$

$h + r_2 = t$
$and \; we \; can \; set$

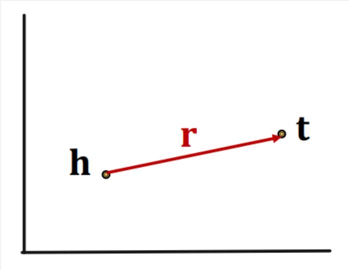
$r_1 = -r_2$



Composition:
$r_1(x,y) \;\&\; r_2(y,z)$
$\Rightarrow r_3(x,z) \quad \forall x,y,z$

$r_1 + r_2 = r_3$

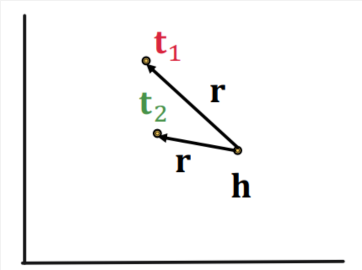
TransE - Limitation



Symmetric:

$r(h,t) \Longrightarrow r(t,h)$

$Only\ if\ r = 0\ and\ h = t$



1-to-N:

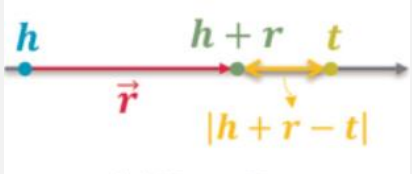
$r(h,t_1), r(h,t_2),\ \ldots, r(h,t_n)$ are all true.

$t_1 = h + r = t_2$

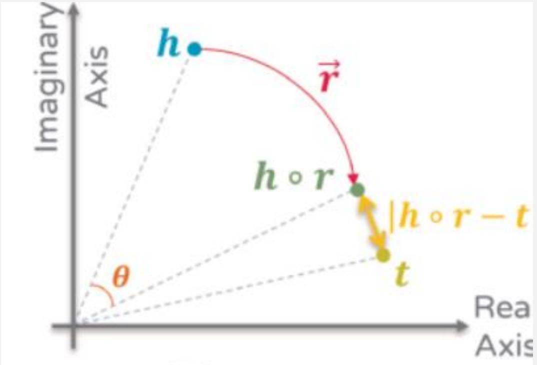
$t_1\ and\ t_2$ will map to the same vector although they are different entities.

RotatE

- Translation:
  - For a triple (h, r, t), let $\boldsymbol{h, r, t} \in \mathbf{R^d}$ be embedding vectors.
    - if the given link exists - $\boldsymbol{h} \circ \boldsymbol{r} \approx \boldsymbol{t}$
    - else $\boldsymbol{h} \circ \boldsymbol{r} \neq \boldsymbol{t}$
    - where '∘' is the Hadamard product on the embeddings.



(a) TransE

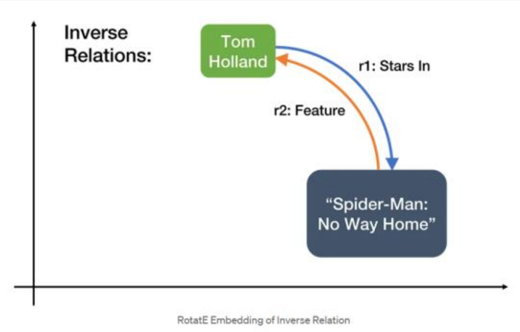


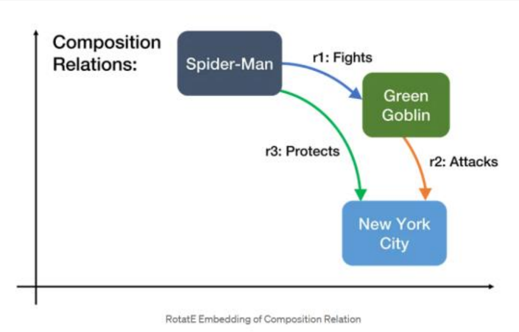
(b) RotatE

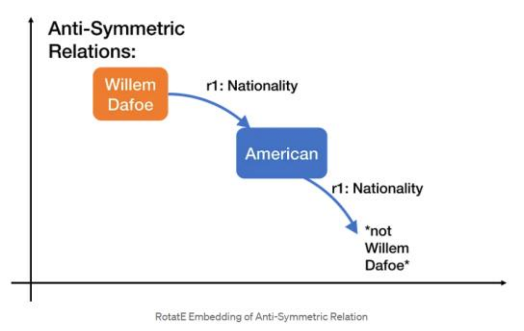
- Entity scoring function: $f_r(h, t) = -||h \circ r \ - t\ ||^2$

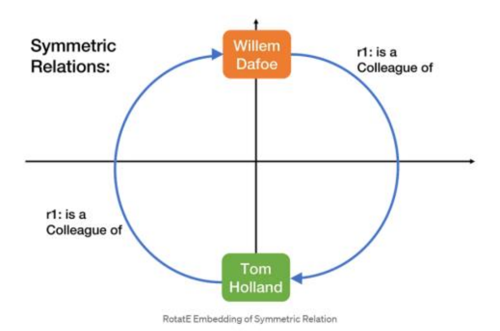
The Hadamard product is a binary operation that takes in two matrices of the same dimensions and returns a matrix of the multiplied corresponding elements.

RotatE



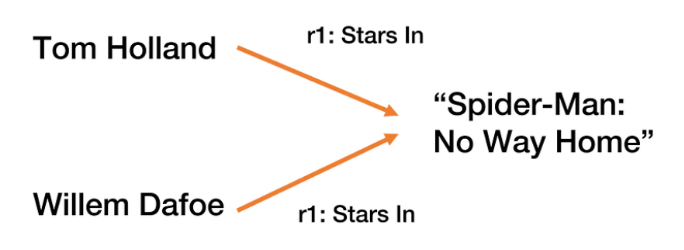




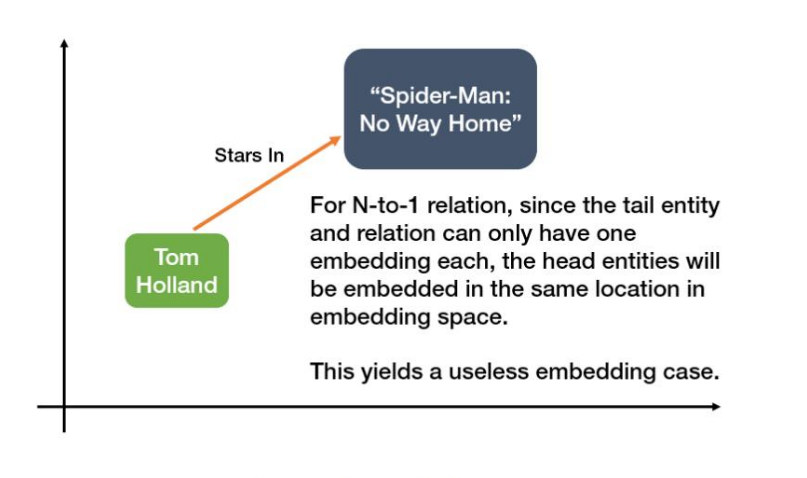


RotatE can model symmetric relations that TransE cannot express! It does so as 180-degree rotations.

RotatE - Limitation



Knowledge Graph with N-to1 Relation



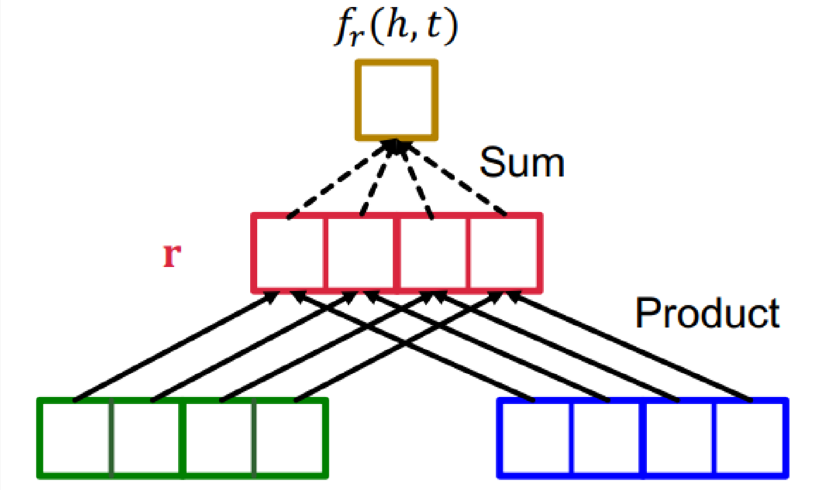
TransE Embedding of N-to-1 Relation

DistMult

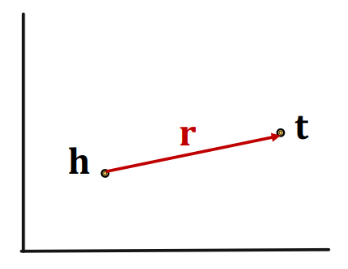
- For a triple (h, r, t), let $\mathbf{h}, \mathbf{r}, \mathbf{t} \in \mathbf{R}^{\mathbf{d}}$ be embedding vectors.

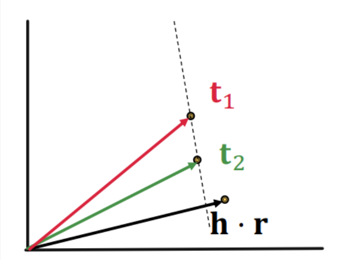
- Score function: $f_r(h,t) = < h, r, t >$
  $= \sum_i \; h_i \,.\, r_i \,.\, t_i$

- Intuition of score function:
  - Cosine similarity between $h \,.\, r \; and \; t$



DistMult





Symmetric:

$r(h,t) \implies r(t,h)$

$$f_r(h,t) = \langle h, r, t \rangle = \sum_i h_i \,.\, r_i \,.\, t_i$$
$$= \langle t, r, h \rangle = f_r(t,h)$$

1-to-N:

$r(h,t_1), r(h,t_2), \ \ldots, r(h,t_n)$ are all true.

$$\langle h, r, t_1 \rangle = \langle h, r, t_2 \rangle$$

DistMult - Limitations

Antisymmetric:
$r(h,t) \Longrightarrow \neg r(t,h)$

$f_r(h,t) = < h, r, t >$
$= \sum_i h_i . r_i . t_i$
$= < t, r, h > = f_r(t,h)$

Inverse:
$r_1(h,t) \Longrightarrow r_2(t,h)$

$< h, r_1, t > = < t, r_2, h >$
$r_1 = r_2$ solves this

But semantically it does not make sense
(advisor is not equal to advisee)

Composition:
$r_1(x,y) \,\&\, r_2(y,z)$
$\Rightarrow r_3(x,z) \quad \forall x,y,z$

Since dot product is commutative, it can distinguish between head and tail entities

Summary – Embedding Models

Model	**Score**	**Symmet.**	**Antisym.**	**Inverse**	**Compos.**	**1-to-N**
TransE	$-\|\| h+r-t \|\|$	✗	✓	✓	✓	✗
RotatE	$-\|\|h \circ r-t \|\|^2$	✓	✓	✓	✓	✗
DistMult	$< h,r,t >$	✓	✗	✗	✗	✓

LITERATURE SURVEY

LNMIIT
The LNM Institute of Information Technology

- This Led us to our Base Paper “KG-Predict: A knowledge graph computational framework for drug repurposing”
- How well did KG-Predict outperformed the other state of art embedding methods ?

Data	Method	Hits@1	Hits@3	Hits@10	MRR
GP-KG	TransE	0.116	0.226	0.399	0.209
	DistMult	0.103	0.207	0.379	0.191
	ConvE	0.126	0.232	0.399	0.216
	RotatE	0.119	0.231	0.403	0.212
	KG-Predict	**0.174**	**0.266**	**0.447**	**0.261**

LITERATURE SURVEY

The LNM Institute of Information Technology

- This Led us to our Base Paper “KG-Predict: A knowledge graph computational framework for drug repurposing”.
- The authors extended the Composition based multi regional graph Convolutional network (CompGCN).

LITERATURE SURVEY

- Let's take a look at the algorithm what the authors used
  - Input : Feature Vector x_v of node v (e.g. drug, disease) and feature vector S_r of edge r (e.g. Drug-treat-disease).
  - CompGCN Layer : The model Stacked several convolutional layers.
    - Used 2 different functions.
      - $AGGREGATE_n(.)$
      - $CONCAT(.)$
    - KG-Predict uses learnable transformational matrix W_{rel}^{n} which project all relations to same embedding space as entities and relations which are to update
  - Output : z_v represent the set of entity embedding and z_r represent the set of relation embedding

LITERATURE SURVEY

LNMIIT
The LNM Institute of Information Technology

Pseudo Code for the explained algorithm

Algorithm 1: GP-KG Embedding Algorithm

Input: $G = (V, E, X, R, S)$, $\forall n \in 1, 2, ..., N$, aggregator function $AGGREGATE_n(.)$, concat function $CONCAT(.)$, relation-specific coefficient matrix W^n_{rel}, self-specific coefficient matrix W^n_o, learnable transformation matrix W^n_{rel}, the set of entity v's neighbors $N(v)$.

Output: entity embedding Z_V, relation embedding Z_R.

$h^0_v \leftarrow X_v$, $h^0_r \leftarrow S_r$;

for $n = 1, 2, ..., N$ **do**

  for $v \in V$ **do**

    $h^n_{N(v)} \leftarrow AGGREGATE_n(W^n_\lambda \psi\left(h^{n-1}_u, h^{n-1}_r\right)), u, r \in N(v)$;

    $h^n_v \leftarrow f(CONCAT(h^n_{N(v)}, W^n_o h^{n-1}_v))$;

  for $r \in R$ **do**

    $h^n_r \leftarrow W^n_{rel} h^{n-1}_r$;

$Z_V \leftarrow \{h^N_v\}, \forall v \in V$;

$Z_R \leftarrow \{h^N_r\}, \forall r \in R$;

LITERATURE SURVEY

- How well did KG-Predict outperformed the other state of art embedding methods ?

Data	Method	Hits@1	Hits@3	Hits@10	MRR
	TransE	0.116	0.226	0.399	0.209
	DistMult	0.103	0.207	0.379	0.191
GP-KG	ConvE	0.126	0.232	0.399	0.216
	RotatE	0.119	0.231	0.403	0.212
	KG-Predict	**0.174**	**0.266**	**0.447**	**0.261**

OBJECTIVE

The LNM Institute of Information Technology

- Modify the GP-KG embedding process to construct a unique embedding methodology
- Investigate state-of-the-art embedding and make necessary alteration to them.
- Use Altered embedding to develop new technique and use them to find results on Prediction model.

THANKYOU