A Study on the Advanced Development of Environmental Disease Prediction Models Using Knowledge Graphs and State-of-the-Art Algorithms

Jong-June Jeon[1]

[1]Department of Statistics, University of Seoul

Introduction

Graph

A **graph** G is an ordered pair:

$$G = (V, E)$$

where:

- V is a set of **vertices** (nodes),
- $E \subseteq \{\{u, v\} \mid u, v \in V, u \neq v\}$ is a set of **edges** for undirected graphs,
- or $E \subseteq \{(u, v) \mid u, v \in V\}$ for directed graphs (digraphs).

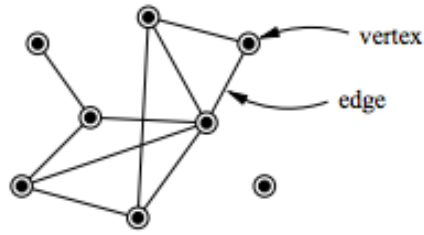


Figure: Illustration of Graph

Homogeneous vs. Heterogeneous Graphs

Homogeneous Graph:

- All nodes and edges are of a single type.
- $G = (V, E)$
- Example: Social network with only "person" nodes and "friendship" edges.

Heterogeneous Graph:

- Contains multiple types of nodes and/or edges.
- $G = (V, E, \phi, \psi)$, with:
  - $\phi : V \to \mathcal{T}_V$ (node types)
  - $\psi : E \to \mathcal{T}_E$ (edge types)
- Example: Paper-author-institution graph with citation, authorship, and affiliation edges.

Knowledge Graph as a representative heterogeneous graph

- A graph representing information
- KG encodes factual knowledge as triples in $(head,\ relation,\ tail)$, for example (Da Vinci, painted, Mona Lisa) or (James, likes, Mona Lisa).
- Call the head and tail as entities.

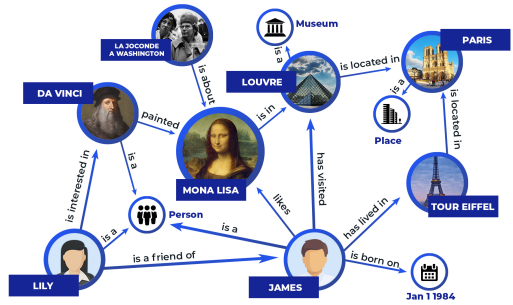


Figure: Example of the knowledge graph.

Why is Building a Knowledge Graph Important in AI?

1. **Semantic Integration of Heterogeneous Data**
   - Enables combining information from different sources with consistent meaning.
2. **Facilitating Semantic Search and Question Answering**
   - Improves retrieval accuracy by understanding intent and relationships beyond keywords.
3. **Enhancing Inference and Reasoning in AI Systems**
   - Supports logical inference through ontology and relation paths.
4. **Enabling Explainable and Trustworthy AI**
   - Transparent structure helps trace decisions back to interpretable facts.

Task 1: Link Prediction for relation existence or type

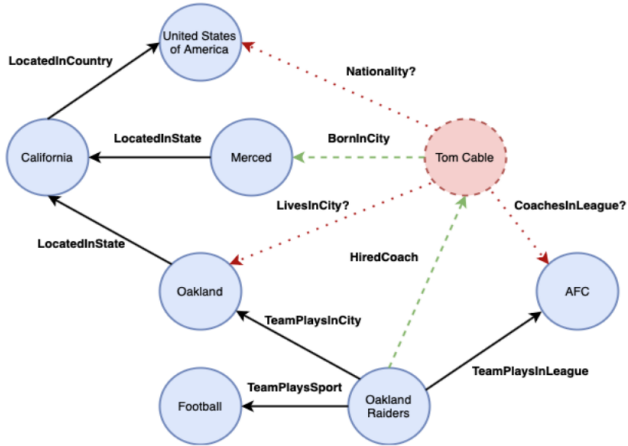


Figure: An example of link prediction.

Task 2: **Multi-hop Reasoning** (Path Prediction)

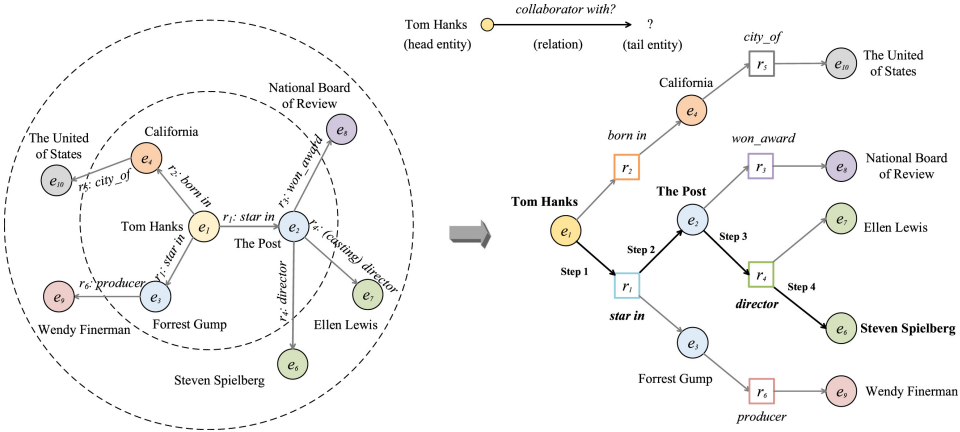


Figure: An example of multi-hop reasoning.

Research Motivation: Using the knowledge graph to reduce animal experimentation

- **Ethics**: reduces or eliminates animal suffering.
- **Human relevance**: some animal data don't translate well to people.
- **Regulation & public opinion**: growing support for humane science.
- **Cost & speed**: non-animal methods can be quicker and cheaper.

We focus on estimating toxicity on the process of

$$\text{Chemical} \rightarrow \text{Gene} \rightarrow \text{Disease}$$

Figure: Illustration of animal experimentation

In-Silico Method in Toxicology

Running experiments inside a computer rather than in a petri dish (*in vitro*) or in an animal (*in vivo*).

- **How it works**
  - Mathematical simulations of how a substance interacts with cells or organs.
  - Uses large data sets and AI to predict toxicity, absorption, or effectiveness.
- **Essential Machine Learning**
  - Dataset produced by High-throughput screening (HTS), a method for scientific discovery
  - Rendering predictors for classification task (Graph Neural Network embedding chemicals, Description augmentation by LLM)
  - Constrastive Learning

Dataset: Comparative Toxicogenomics Database

Dataset: Comparative Toxicogenomics Database (CTD)

CTD

- A biological database that collects information on the effects of environmental exposures on human health.
- Composed of entities such as chemicals, genes, and disease.
- Constructed through literature-based curation by domain experts.
  - Therapeutic
  - Marker/Mechanism
  - NA (cooccurrence in a paper)

ChemicalID	DiseaseID	DirectEvidence
C046983	MESH:D054198	therapeutic
C112297	MESH:D006948	marker/mechanism
C112297	MESH:D006948	NA
C534883	MESH:D000230	NA
C534883	MESH:D000505	NA
D015054	MESH:D012769	marker/mechanism
D015054	MESH:D014605	marker/mechanism

Table: Example of chemical-disease association.

*https://ctdbase.org/about/

Dataset: Comparative Toxicogenomics Database (CTD)

Entities and Relations

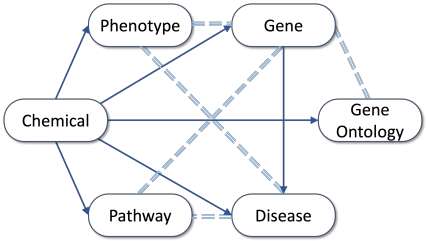


Figure: Overview of the CTDKG: The solid and dashed lines denote antisymmetric and symmetric relations.

Name	#Entity Types	#Entities	#Relation Types	#Triplets	Average Node Deg.
CD	2	23,143	2	3,346,161	289.2
CGD	3	79,785	141	39,058,546	979.1
CGPD	4	88,112	143	39,212,822	890.1
CTDKG	6	308,928	155	58,099,654	376.1

Table: Summary of CTDKG.

Dataset: Comparative Toxicogenomics Database (CTD)

CTD Examples: Chemical–Gene–Pathway

- **Chemical**: Bisphenol A (BPA)
- **Target Gene**: *ESR1* (Estrogen Receptor 1)
- **Pathway**: Estrogen signaling pathway
- **Mechanism**: BPA mimics estrogen and activates *ESR1*, leading to hormone-related gene expression.

Comparative Toxicogenomics Database Knowledge Graph (CTDKG)

- Association between two entities can be represented as triplets.
- Entity information (e.g. gene information) is generally complex and some is completely unknown.
- Consideration of adding a new entity (a new chemical)

ChemicalID	DiseaseID	DirectEvidence
C046983	MESH:D054198	therapeutic
C112297	**MESH:D006948**	**marker/mechanism**
C112297	MESH:D006948	NA
C534883	MESH:D000230	NA
C534883	MESH:D000505	NA
D015054	MESH:D012769	marker/mechanism
D015054	MESH:D014605	marker/mechanism

$\Longrightarrow$

(C112297, chem_curated_dis, MESH:D006948)
(C534883, chem_inferred_dis, MESH:D000230)
(C534883, chem_inferred_dis, MESH:D000505)
(D015054, chem_curated_dis, MESH:D012769)
(D015054, chem_curated_dis, MESH:D014605)

Figure: Example of chemical-disease triplet construction.

Chemical Molecular Fingerprints

- A digital "barcode" that encodes a molecule's structure into 0s and 1s.
- Predictor is given by a binary vector.

Popular Fingerprint Types

- **Structural keys** (e.g. MACCS): checks for *specific* sub-structures from a fixed dictionary.
- **Circular fingerprints** (e.g. ECFP/Morgan): captures each atom's neighbourhood within a chosen radius.
- **Physicochemical descriptors**: summarises counts and properties (molecular weight)

Rendoring Predictors

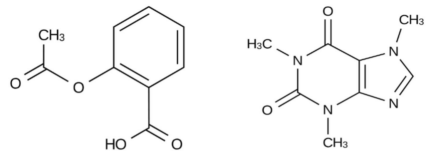
SMILES (Simplified Molecular Input Line Entry System) is a textual representation of chemical structures using ASCII strings.

Molecule	**SMILES**
Water	`O`
Methane	`C`
Ethanol	`CCO`
Acetic acid	`CC(=O)O`
Benzene	`c1ccccc1`
Toluene	`Cc1ccccc1`
Aspirin	`CC(=O)Oc1ccccc1C(=O)O`
Caffeine	`Cn1cnc2c1c(=O)n(c(=O)n2C)C`

Table: Exmaples of SMILES

Key Characteristics of SMILES

- Atoms are represented by their atomic symbols (e.g., `C`, `O`, `N`).
- Bonds:
  - Single: omitted or `-` (e.g., `CC`)
  - Double: `=`, Triple: `#`
- Branches: represented with parentheses, e.g., `CC(C)C`
- Rings: indicated by numbers, e.g., `c1ccccc1` for benzene
- Aromatic atoms: lowercase letters like `c`, `n`



Acetylsalicylic Acid (Aspirin)
CC(=O)Oc1ccccc1C(=O)O

Epinephrine
CNC[C@H](O)c1ccc(O)c(O)c1

Figure: Example of SMILES

The molecule has a graph strucure, which can be easily process by python package (RDKit).

Rendoring Predictors: Chemical Embedding

Before CNNs

- Vision relied on handcrafted features (e.g., SIFT, HOG).
- Performance plateaued due to human-designed limitations.

After CNNs

- Learned representations from raw pixels.
- Enabled end-to-end learning with superior performance.

Graph Neural Network

GNN renders node (atom) representations updated via messages from their neighbors.

$$h_v^{(k)} = \text{UPDATE}^{(k)}\left(h_v^{(k-1)}, \text{AGGREGATE}^{(k)}\left(\{h_u^{(k-1)} : u \in \mathcal{N}(v)\}\right)\right)$$

- $\mathcal{N}(v)$: neighbors of node v
- AGGREGATE: sum, mean, max, attention
- UPDATE: neural network (e.g., MLP, GRU)

Suppose that a graph consists of m nodes and each node is represented by d-dimensional vector. Then the graph is usually represented by $m \times n$ matrix.

Graph Neural Network

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l)}W^{(l)}\right)$$

- $\tilde{A} = A + I$: adjacency matrix with self-loops
- $\tilde{D}$: degree matrix of $\tilde{A}$
- $H^{(l)}$: node embeddings at layer l; $W^{(l)}$: learnable weights; $H^{(l)}W^{(l)}$ is a feature transformation
- $\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}$: aggregation map
- σ: activation function

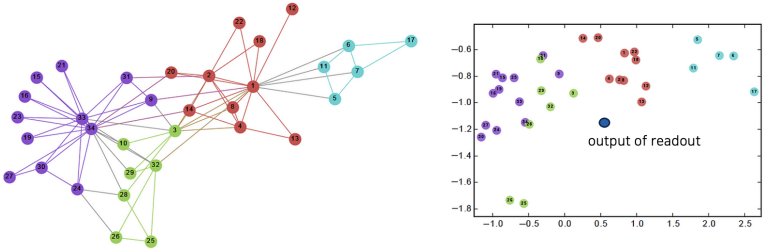
The graph remains an $m \times n$ matrix after each layer, with features updated using 1-hop neighbor information.

Graph Readout Layer: Aggregate all node embeddings into a single graph representation.

$$h_G = \text{READOUT}(\{h_v^{(K)} \mid v \in V\})$$

- $h_v^{(K)}$: final node representation from GNN layers
- h_G: graph-level embedding
- Common functions (Permutation Invariant Function):
  - `sum`, `mean`, `max`
- Used for graph classification or regression tasks.

Rendoring Predictors: Chemical Embedding



Graph data -> Graph Feature (node feature) on $\mathbb{R}^2$-> Readout (Graph feature)

Rendoring Predictors: Chemical Embedding

- Loss function: Mean Squared Error (MSE)

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 ,$$

  where $\hat{y}_i$ is the output of GNN regression model.
- Evaluation metrics:
  - RMSE, MAE
  - Pearson/Spearman correlation
- Data split: scaffold or random split

Link prediction by contrastive Learning

Contrastive Learning for Link Prediction

Goal: Predict whether an edge exists between two nodes using contrastive learning.

- Encode graph to obtain node embeddings z_u, z_v.
- Construct **positive pairs** from existing edges.
- Sample **negative pairs** from unconnected nodes.
- Apply contrastive loss to bring positive pairs closer and push negative pairs apart:

$$\mathcal{L} = -\log \frac{\exp(\mathsf{sim}(z_u, z_v)/\tau)}{\sum_k \exp(\mathsf{sim}(z_u, z_k)/\tau)}$$

- Use learned embeddings to score link existence:

$$\mathsf{score}(u, v) = \sigma(z_u^\top z_v)$$

Theoretical Framework of Contrastive Learning

- Joint distribution of a positive pair: $(z_u, z_v) \sim p(z_u, z_v)$
- Joint distribution of a negative pair: $(z_u, z_v{}') \sim p(z_u)p(z_v)$
- Modeled by parameterized function (Mutual information or Lift measure)

$$r(z_u, z_v; \theta, \eta) = \log \frac{p(z_u, z_v)}{p_{z_u}(z_u)p_{z_v}(z_v)}$$

Contrastive Loss Function with the logistic loss:

$$J_{CL}(\theta, \eta) = -E_{p(z_u, z_v)} \left[\log \frac{\exp(r(z_u, z_v))}{1 + \exp(r(z_u, z_v))} \right] - E_{p_{z_u}(z_u)p_{z_v}(z_v{}')} \left[\log \frac{1}{1 + \exp(r(z_u, z_v{}'))} \right]$$

The log density ratio is estimated by the constrastive learning and the embedding function is trained by maximizing the mutual information of the joint distribution [7, 6].

Limitations of Node Embeddings in Chemical–Gene Link Prediction

- Fixed Predictor (Unlearnable Predictor)
  - for chemicals: fingerprints and molecular descriptors
  - for genes and diseases: questionable!
- Learnable Predictor
  - for chemicals: molecular embedding via pre-trained model based on GNN ($e.g.$, MGSSL [12], GraphMVP [3], Uni-Mol [14])
  - for genes and diseases: entity ID embedding via neural network
- Embedding problem: semantic information is not included if an ID is employed.
- How to use features of new emerging chemical? (a case that a new entity is added: expansion of knowledge!)

Textual enrichiment for entity representation

- Collection of the entity descriptions from 9 databases (Wikipedia, Wikigenes, and MeSH).
- Descriptions of entities were embedded using BioT5+ [5].
- If one of a text description and a molecular structural information is avaiable, a feature of the entity can be included in KG.

Entity type	#Entity	#Raw Desc.	#Processed Desc.	#No Desc.
Chemical	18,708	17,155	15,876	2,832
Gene	237,018	82,574	69,080	167,938
Disease	7,263	4,366	4,366	2,897
Phenotype	20,223	20,223	20,136	87
Pathway	2,363	2,044	2,017	346
GO	23,353	23,353	23,163	190

Table: Summary of the number of entity descriptions.

Comparative Toxicogenomics Database Knowledge Graph (CTDKG)

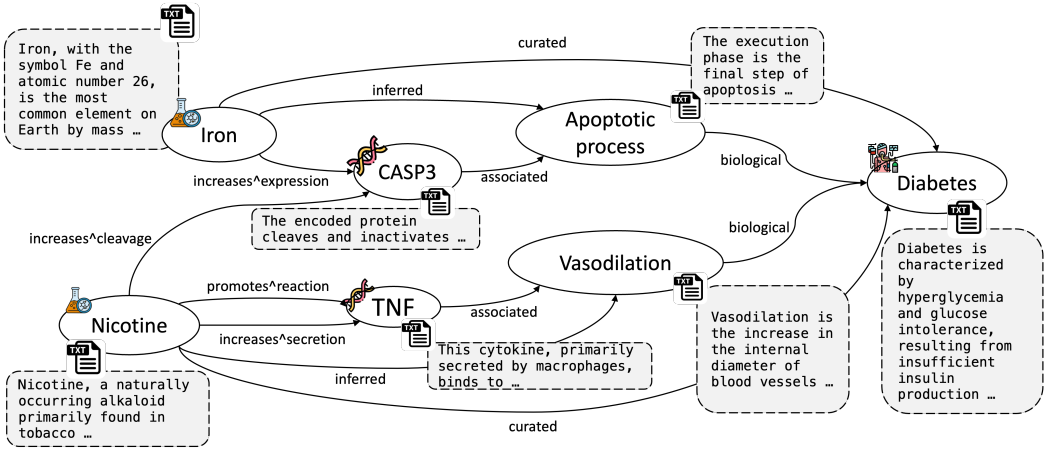


Figure: Illustration of CTDKG.

Analysis Framework

1. Constructing a text-augmented knowledge graph focusing on chemical-induced disease
2. Benchmarking experiments

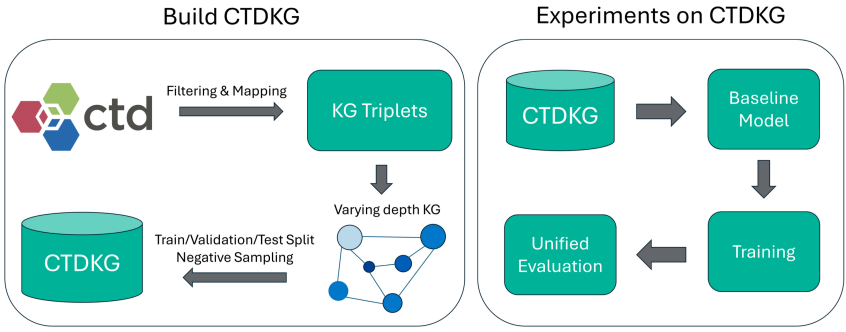


Figure: Overview of our research.

Task 1: Link Prediction

- Translation-based model (distance based model): head (chemical) + relation (curated) = tail (disease)
- Information-based models (angle based model): Relational angle of head and tail $\simeq 0$ degree and non-relational angle $\simeq 90$ degree.

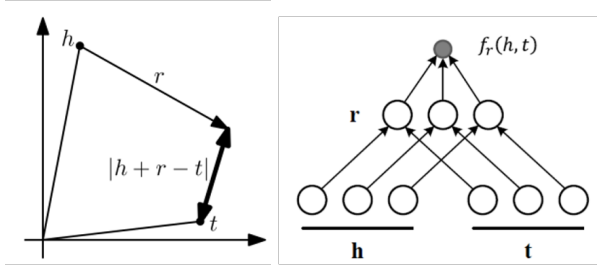


Figure: (Left) TransE; (Right) DistMult.

Task 2: Path Prediction

- The model identifies relational paths connecting head to tail through one or more intermediate nodes. (the shortest path $\simeq$ the the most likely path)

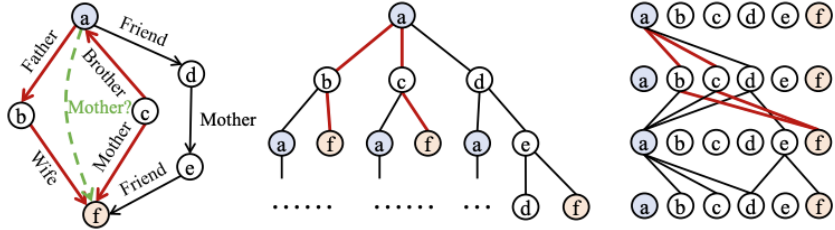


Figure: (a) Given a query (a, Mother, ?), only a few important paths (shown in red) are necessary for reasoning. (b) An exhaustive search algorithm enumerates all paths in exponential time. (c) Bellman-Ford algorithm computes all paths in polynomial time.

The loss function is defined by

$$L = -\sum_{x=(h,r,t)\in\mathcal{T}} \left(\log \sigma(s_r(h,t) + \gamma) + \sum_{(h',r',t')\in\mathcal{N}_x} p(h',r',t') \log \sigma(-s_{r'}(h',t') - \gamma) \right), \tag{1}$$

where $\gamma > 0$ and $\alpha > 0$ are constants, σ is the sigmoid function and

$$p(h',r',t') = \exp(\alpha s_{r'}(h',t'))/\sum_{(\tilde{h},\tilde{r},\tilde{t})\in\mathcal{N}_x} \exp(\alpha s_{\tilde{r}}(\tilde{h},\tilde{t}))$$

is a weight of negative samples within $\mathcal{N}_x$.

Evaluation settings

- Tasks:
  - Head prediction: predict h given (r, t)
  - Tail prediction: predict t given (h, r)

- Negative Sampling:
  - For a positive triplet $x = (h, r, t)$ (head, relation, tail):

$$\mathcal{N}_{1,x} = \{(h', r, t) \mid h' \in V, (h', r, t) \notin G\}$$

$$\mathcal{N}_{2,x} = \{(h, r, t') \mid t' \in V, (h, r, t') \notin G\}$$

  - 1,000 negatives per triplet: 500 for head, 500 for tail

- Ranking Computation:
  - $rank_{1,x}$: among $\mathcal{N}_{1,x} \cup \{x\}$
  - $rank_{2,x}$: among $\mathcal{N}_{2,x} \cup \{x\}$

- Metrics for a test set $\mathcal{T}_e$ (evaluation triplets):

$$MR = \frac{1}{2|\mathcal{T}_e|} \sum_{x \in \mathcal{T}_e} (rank_{1,x} + rank_{2,x})$$

$$MRR = \frac{1}{2|\mathcal{T}_e|} \sum_{x \in \mathcal{T}_e} \left(\frac{1}{rank_{1,x}} + \frac{1}{rank_{2,x}} \right)$$

$$Hits@k = \frac{1}{2|\mathcal{T}_e|} \sum_{x \in \mathcal{T}_e} [I(rank_{1,x} \leq k) + I(rank_{2,x} \leq k)]$$

Benchmark Models for Link Prediction

Model	Embedding	Score Function $s_r(h,t)$	Complexity
TransE [1]	$h,r,t \in \mathbb{R}^d$	$-\|h+r-t\|$	$O(\|V\|d+\|R\|d)$
RotatE [8]	$h,r,t \in \mathbb{C}^d$	$-\|h \circ r - t\|$	$O(2\|V\|d+2Rd)$
HAKE [13]	$h_m,t_m \in \mathbb{R}^d, r_m \in \mathbb{R}^d_+,$ $h_p,r_p,t_p \in [0,2\pi)^d$	$-\|h_m \circ r_m - t_m\|$ $-\lambda\|\sin(h_p+r_p-t_p)/2\|$	$O(2\|V\|d+2\|R\|d)$
Triplere [11]	$h,r_h,r_m,r_t,t \in \mathbb{R}^d$	$-\|h \circ r_h - t \circ r_t + r_m\|$	$O(\|V\|d+3\|R\|d)$
Rotate4D [2]	$h,r,t \in \mathbb{H}^d$	$\|W_r \times (h \circ r) - t\|$	$O(4\|V\|d+4\|R\|d)$
DistMult [10]	$h,r,t \in \mathbb{R}^d$	$h^\top \mathrm{diag}(r) t$	$O(\|V\|d+\|R\|d)$
ComplEx [9]	$h,r,t \in \mathbb{C}^d$	$\mathrm{Re}(h^\top \mathrm{diag}(r) \bar{t})$	$O(2\|V\|d+2\|R\|d)$
QuatRE [4]	$h,r,r_h,r_t,t \in \mathbb{H}^d$	$((h \otimes r_h^\triangleleft) \otimes r^\triangleleft)) \bullet (t \otimes r_t^\triangleleft)$	$O(4\|V\|d+12\|R\|d)$

Table: Comparisons of scoring functions in various knowledge graph embedding models. The number of entities and relation types is denoted as $|V|$ and $|R|$. Details of score functions are skipped.

Task 1-1: Link prediction in naive cases (positive samples:negative samples = 1:500)

- Goal: Explore the potential for discovering new relationships under transductive simulation setting.
- Predictor: ID embedding + description embedding

Model	#Params	Validation					Test				
		MR	MRR	Hits@1	Hits@3	Hits@10	MR	MRR	Hits@1	Hits@3	Hits@10
Baseline	-	251	0.014	0.002	0.006	0.020	251	0.014	0.002	0.006	0.020
DistMult	31.1M	$6.0_{\pm 0.4}$	$0.665_{\pm 0.013}$	$0.545_{\pm 0.015}$	$0.742_{\pm 0.013}$	$0.892_{\pm 0.007}$	$6.0_{\pm 0.4}$	$0.664_{\pm 0.013}$	$0.545_{\pm 0.015}$	$0.742_{\pm 0.013}$	$0.892_{\pm 0.007}$
ComplEx	62.2M	$4.6_{\pm 0.3}$	$0.726_{\pm 0.015}$	$0.619_{\pm 0.018}$	$0.802_{\pm 0.014}$	$0.923_{\pm 0.007}$	$4.6_{\pm 0.3}$	$0.726_{\pm 0.014}$	$0.619_{\pm 0.018}$	$0.801_{\pm 0.014}$	$0.923_{\pm 0.007}$
QuatRE	124.7M	$\mathbf{3.9_{\pm 0.3}}$	$\mathbf{0.757_{\pm 0.016}}$	$\mathbf{0.657_{\pm 0.019}}$	$\mathbf{0.830_{\pm 0.014}}$	$\mathbf{0.938_{\pm 0.007}}$	$\mathbf{3.9_{\pm 0.3}}$	$\mathbf{0.757_{\pm 0.016}}$	$\mathbf{0.657_{\pm 0.019}}$	$\mathbf{0.830_{\pm 0.014}}$	$\mathbf{0.938_{\pm 0.007}}$
TransE	31.1M	$12.7_{\pm 0.0}$	$0.585_{\pm 0.001}$	$0.466_{\pm 0.001}$	$0.652_{\pm 0.001}$	$0.814_{\pm 0.000}$	$12.7_{\pm 0.0}$	$0.585_{\pm 0.001}$	$0.467_{\pm 0.001}$	$0.652_{\pm 0.001}$	$0.814_{\pm 0.001}$
RotatE	62.1M	$10.2_{\pm 0.0}$	$0.656_{\pm 0.001}$	$0.551_{\pm 0.001}$	$0.721_{\pm 0.001}$	$0.855_{\pm 0.001}$	$10.2_{\pm 0.0}$	$0.656_{\pm 0.001}$	$0.551_{\pm 0.001}$	$0.721_{\pm 0.001}$	$0.855_{\pm 0.000}$
HAKE	62.3M	$10.0_{\pm 0.0}$	$0.637_{\pm 0.001}$	$0.529_{\pm 0.001}$	$0.700_{\pm 0.001}$	$0.842_{\pm 0.000}$	$10.0_{\pm 0.0}$	$0.637_{\pm 0.001}$	$0.530_{\pm 0.001}$	$0.700_{\pm 0.001}$	$0.842_{\pm 0.000}$
Triplere	31.3M	$15.1_{\pm 1.7}$	$0.474_{\pm 0.023}$	$0.347_{\pm 0.023}$	$0.531_{\pm 0.026}$	$0.730_{\pm 0.023}$	$15.1_{\pm 1.7}$	$0.475_{\pm 0.023}$	$0.348_{\pm 0.023}$	$0.532_{\pm 0.026}$	$0.731_{\pm 0.023}$
Rotate4D	124.5M	$26.8_{\pm 0.9}$	$0.355_{\pm 0.001}$	$0.240_{\pm 0.001}$	$0.390_{\pm 0.001}$	$0.596_{\pm 0.003}$	$26.8_{\pm 0.9}$	$0.355_{\pm 0.001}$	$0.240_{\pm 0.001}$	$0.391_{\pm 0.001}$	$0.596_{\pm 0.003}$

Table: Link prediction results on CTDKG obtained by embedding descriptions with BioT5+ and concatenating them with the simple embedding. Baseline is computed by random permutation-based ranking.

Task 1-2: Link prediction for evaluting a new chemical (inductive simulation setting)

- Goal: Evaluate the ability of the model to predict associations between a novel chemical and a disease.
- Predictor: description and molecular embedding for chemicals, and description and ID embedding for diseases.

Model	#Params	Validation		Test	
		MRR	Hits@10	MRR	Hits@10
Baseline	-	0.014	0.020	0.014	0.020
DistMult	0.9M	$0.196_{\pm 0.001}$	$0.372_{\pm 0.001}$	$0.153_{\pm 0.000}$	$0.294_{\pm 0.000}$
ComplEx	1.8M	$0.198_{\pm 0.000}$	$0.375_{\pm 0.001}$	$0.155_{\pm 0.000}$	$0.294_{\pm 0.001}$
QuatRE	3.7M	$0.206_{\pm 0.001}$	$0.385_{\pm 0.002}$	$0.156_{\pm 0.003}$	$0.297_{\pm 0.003}$
TransE	0.9M	$0.228_{\pm 0.007}$	$0.447_{\pm 0.012}$	$0.137_{\pm 0.001}$	$0.277_{\pm 0.001}$
RotatE	1.7M	$0.247_{\pm 0.006}$	$0.471_{\pm 0.010}$	$0.148_{\pm 0.001}$	$0.290_{\pm 0.002}$
HAKE	1.9M	$0.269_{\pm 0.006}$	$0.503_{\pm 0.011}$	$0.153_{\pm 0.001}$	$0.295_{\pm 0.002}$
Triplere	1.1M	$\mathbf{0.282_{\pm 0.004}}$	$\mathbf{0.526_{\pm 0.005}}$	$0.159_{\pm 0.002}$	$0.299_{\pm 0.003}$
Rotate4D	3.7M	$0.177_{\pm 0.002}$	$0.309_{\pm 0.002}$	$\mathbf{0.169_{\pm 0.001}}$	$\mathbf{0.303_{\pm 0.001}}$

Table: Results for CD under the inductive setting.

Task 1-3: Link prediction for prioritization of Gene-Disease associations (inductive simulation settings)

- Goal: Evaluating the contribution of link prediction to prioritizing existing inferred relationships.
- A CGD subgraph was constructed around ten consumer-product chemicals used as preservatives or surfactants.
- We treated the curated and inferred gene-disease associations as positive and negative samples.

Model	#Params	Validation		Test	
		MRR	Hits@10	MRR	Hits@10
Baseline	-	0.005	0.007	0.005	0.006
DistMult	3.2M	$0.114_{\pm 0.005}$	$0.209_{\pm 0.010}$	$0.105_{\pm 0.005}$	$0.198_{\pm 0.010}$
ComplEx	6.3M	$0.125_{\pm 0.005}$	$0.233_{\pm 0.008}$	$0.115_{\pm 0.005}$	$0.223_{\pm 0.011}$
QuatRE	12.8M	$0.133_{\pm 0.004}$	$0.242_{\pm 0.008}$	$0.120_{\pm 0.005}$	$0.233_{\pm 0.009}$
TransE	3.2M	$0.153_{\pm 0.003}$	$0.277_{\pm 0.005}$	$0.137_{\pm 0.004}$	$\mathbf{0.266_{\pm 0.005}}$
RotatE	6.2M	$\mathbf{0.155_{\pm 0.002}}$	$0.281_{\pm 0.006}$	$0.138_{\pm 0.002}$	$0.258_{\pm 0.007}$
HAKE	6.5M	$0.142_{\pm 0.002}$	$0.252_{\pm 0.004}$	$0.127_{\pm 0.002}$	$0.225_{\pm 0.006}$
Triplere	3.4M	$0.144_{\pm 0.005}$	$0.260_{\pm 0.009}$	$0.128_{\pm 0.003}$	$0.241_{\pm 0.010}$
Rotate4D	12.7M	$\mathbf{0.155_{\pm 0.002}}$	$\mathbf{0.284_{\pm 0.005}}$	$\mathbf{0.139_{\pm 0.002}}$	$\mathbf{0.266_{\pm 0.005}}$

Table: Results for CGD under the curated-vs-inferred setting.

Benchmark Models and Results II

- Negative samples within the top 10 were analyzed for test cases where a positive triplet was correctly ranked first.
- The association between gene MIR150 and atherosclerosis has been reported as plausible in the literature.
- These findings suggest the model can effectively prioritize potential relationships for subsequent in vivo or in vitro experimental validation.

	GeneSymbol	GeneID	DiseaseName	DiseaseID	Rank
positive	MR150	406942	Heart Failure	MESH:D006333	1
	MIR150	406942	Atherosclerosis	MESH:D050197	2
	MIR150	406942	Brain Injuries	MESH:D001930	3
	MIR150	406942	Diabetes Mellitus, Experimental	MESH:D003921	4
	MIR150	406942	Diabetes Mellitus, Type 2	MESH:D003924	5
negative	MIR150	406942	Inflammation	MESH:D007249	6
	MIR150	406942	Myocardial Reperfusion Injury	MESH:D015428	7
	MIR150	406942	Neoplasm Metastasis	MESH:D009362	8
	MIR150	406942	Non-alcoholic Fatty Liver Disease	MESH:D065626	9
	MIR150	406942	Reperfusion Injury	MESH:D015427	10

Table: Top 10 gene-disease pairs for MR150 screened by Rotate4D.

Task 2: Path Prediction (Multi-hop Reasoning)

- Goal: Infer the paths underlying the indirect connection between chemicals and diseases provided by CTD to fill knowledge gaps.
- For path-based representation learning, we employed NBFNet [15], which incorporates GNN and a generalized Bellman-Ford algorithm (the shortest path algorithm replaces the distance with dissimilarity defined by the conditional probability).

Weight	Query: <C005451, chem_inferred_dis, MESH:D006816>
6.102	<C005451, chem_decreases^expression_gene, 3725> → <3725, gene_inferred_dis, MESH:D006816>
Weight	Query: <MESH:D006816, chem_inferred_dis^{-1}, C005451>
4.295	<MESH:D006816, gene_curated_dis^{-1}, 4968> → <4968, chem_increases^expression_gene^{-1}, C005451>

Table: Example of path prediction. Inverse relations are denoted with a superscript^{-1}.

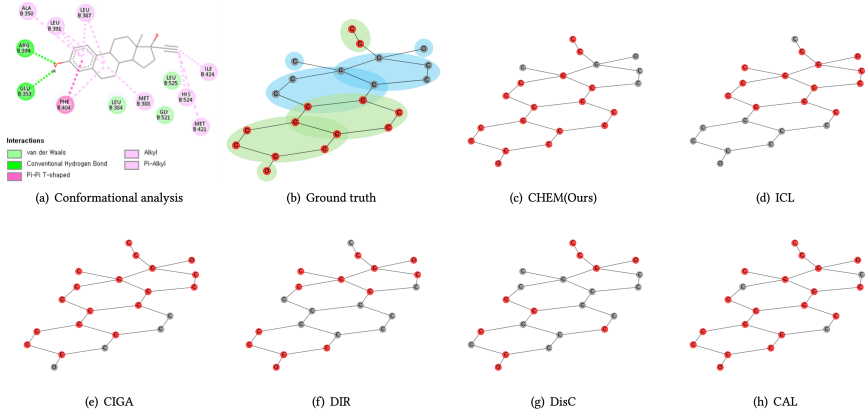
Conclusion & Future work

Conclusion and Limitations

- We constructed CTDKG, a text-augmented knowledge graph that integrates diverse biological knowledge.
- The model demonstrated competitive performance on some benchmarks.
- The absence or imbalance of entity descriptions may impact model effectiveness, highlighting the need for more extensive collection and refinement of textual data.

Conclusion & Future work: WP1

- Causal inference by identifying a subgraph (node selection method)



(a) Conformational analysis (b) Ground truth (c) CHEM(Ours) (d) ICL

(e) CIGA (f) DIR (g) DisC (h) CAL

Figure: Example of causal subgraph analysis on NCGC00091533-04. (a) Result of conformational analysis. (b) True causal subgraph. Functional groups or carbon rings that bind to the receptor are marked green; those that do not are marked blue. (c)-(f) Estimated causal subgraph. Red nodes are causal subgraph nodes.

Conclusion & Future work: WP2

- Improving generalized performance
  - Inductive simulation result indicates a significant drop in predictive performance of existing models.
  - It is difficult to evaluate an overfitting problem even by using the scaffold split method (a representative data split method to avoid overfitting).
  - Data collection process implies an inherent bias problem in CTD (to an experiment design called the assay)
- Integration of prior knowledge (more databases and qualitative analysis results)
  - The discovery by an AI model should always be investigated by a domain expert.
  - In addition, a final discovery aims to be connected with a regulation. Thus, the most important discovery is based on the confirmatory analysis.

Conclusion & Future work: Working Group

(Domestic Research Group)

- **Toxicology, Assessment of hazardous materials, Database**: ChemBAI, Prof Jinhee Cho (School of Environmental Engineering, University of Seoul)
- **AI models**: Prof. Chanwon Lim (Department of Statistics, Chung-Ang University)
- **Epidemiology**: Prof Yoon-Hyeong Choi (College of Health Science, Korea University),
- **In vitro experiment**: Prof. Seung Min Oh (Department of Animal Health and Welfare, Hoseo University)

(International Research Group): EU

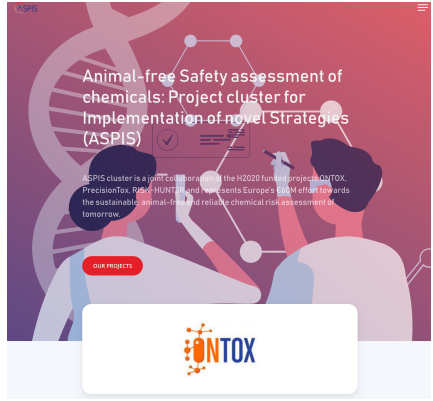


Figure: EU Research Group

Conclusion & Future work: Working Group

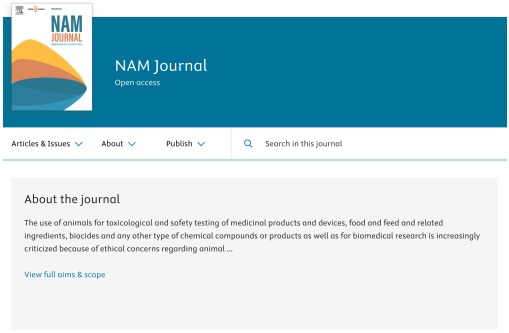


Figure: Journal: New Approach Methodologies

Figure: Horizon Project

Thank you!

References I

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko.
Translating embeddings for modeling multi-relational data.
Advances in neural information processing systems, 26, 2013.

Thanh Le, Huy Tran, and Bac Le.
Knowledge graph embedding with the special orthogonal group in quaternion space for link prediction.
Knowledge-Based Systems, 266:110400, 2023.

Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang.
Pre-training molecular graph representation with 3d geometry.
arXiv preprint arXiv:2110.07728, 2021.

Dai Quoc Nguyen, Thanh Vu, Tu Dinh Nguyen, and Dinh Phung.
Quatre: Relation-aware quaternions for knowledge graph embeddings.
In *Companion Proceedings of the Web Conference 2022*, pages 189–192, 2022.

References II

Qizhi Pei, Lijun Wu, Kaiyuan Gao, Xiaozhuan Liang, Yin Fang, Jinhua Zhu, Shufang Xie, Tao Qin, and Rui Yan.
Biot5+: Towards generalized biological understanding with iupac integration and multi-task tuning.
arXiv preprint arXiv:2402.17810, 2024.

Hiroaki Sasaki and Takashi Takenouchi.
Representation learning for maximization of mi, nonlinear ica and nonlinear subspaces with robust density ratio estimation.
Journal of Machine Learning Research, 23(231):1–55, 2022.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori.
Density Ratio Estimation in Machine Learning.
Cambridge University Press, USA, 1st edition, 2012.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang.
Rotate: Knowledge graph embedding by relational rotation in complex space.
arXiv preprint arXiv:1902.10197, 2019.

References III

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard.
Complex embeddings for simple link prediction.
In *International conference on machine learning*, pages 2071–2080. PMLR, 2016.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng.
Embedding entities and relations for learning and inference in knowledge bases.
arXiv preprint arXiv:1412.6575, 2014.

Long Yu, Zhicong Luo, Huanyong Liu, Deng Lin, Hongzhu Li, and Yafeng Deng.
Triplere: Knowledge graph embeddings via tripled relation vectors.
arXiv preprint arXiv:2209.08271, 2022.

Zaixi Zhang, Qi Liu, Hao Wang, Chengqiang Lu, and Chee-Kong Lee.
Motif-based graph self-supervised learning for molecular property prediction.
Advances in Neural Information Processing Systems, 34:15870–15882, 2021.

Zhanqiu Zhang, Jianyu Cai, Yongdong Zhang, and Jie Wang.
Learning hierarchy-aware knowledge graph embeddings for link prediction.
In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 3065–3072, 2020.

References IV

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke.
Uni-mol: A universal 3d molecular representation learning framework.
2023.

Zhaocheng Zhu, Zuobai Zhang, Louis-Pascal Xhonneux, and Jian Tang.
Neural bellman-ford networks: A general graph neural network framework for link prediction.
Advances in Neural Information Processing Systems, 34:29476–29490, 2021.