

# Evidence Chain Mining based on Domain Knowledge Graph: a Case in Alzheimer's Disease

---

Shirui yu<sup>1</sup>, **Zhengyin Hu**<sup>1</sup>, Haiyun Xu<sup>2</sup>,

<sup>1</sup> National Science Libray(Chengdu), Chinese Academy of Sciences

<sup>2</sup> Shandong University of Technology

Contact email: [huzy@clas.ac.cn](mailto:huzy@clas.ac.cn)



Nice, France, May 18-22, 2025



# Outline

**1. Introduction**

**2. Methodology**

**3. Case Study**

**4. Discussion**






# Introduction

---

## Objective

- This study proposed an **evidence chains** mining based **Knowledge Graph (KG)** , which aims to discover the meaningful relationship among concepts in literatures , help infer feasible new hypotheses for researchers and provide interpretable evidence to support the generated hypothesis.
  - **Evidence chains** consist of more than three knowledge paths from different fields, which can express specific semantic information, effectively integrate dispersed knowledge units, and enable reasonable semantic interpretation for hypothesis generation. Evidence chains mining fosters the potential for tacit knowledge discovery and enables a deeper exploration into the underlying mechanisms of the research.
- 

# Introduction

**Knowledge Graph (KG)** is a kind of semantic network, which aims to describe the concepts, entities, events, and their relationships based on the graph model. In KG, the nodes are knowledge entities, the edges are relationships between them, and multiple connected nodes and edges form knowledge paths.

KG Uses a knowledge representation formalism (e.g., RDF, Subject-Predicate-Object **SPO triples**) to implement semantic descriptions of entities and their relationships.

- Entities: real world objects (e.g., ACTB gene, **CD34 positive cell**, Zebrafish) and concepts (e.g., gene, **cell**, lab animal)
- Relationships: describe relations among entities with a well defined meaning (e.g., **TREATS**)

**SPO triple** also called semantic predication, consists of a subject argument, an object argument, and the relation that binds them and helps to translate complicated free text into structured formats with rich semantic information. It is a kind of mature knowledge representation framework, which is the basis of the construction of a KG.

NAD-dependent protein deacetylases-**AFFECTS**-apoptosis  
down-regulation-**AUGMENTS**-autophagy

# Introduction

**Knowledge Path (KP)** is a relationship chain between nodes in a Knowledge Graph (KG) = (A, R) with the representation as:

$$A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \dots \xrightarrow{R_l} A_{l+1}, \text{ which defines a relationship chain between } A_1 \text{ and } A_{l+1}.$$

**Knowledge Path:** “traf6”- [*stimulates*] - “nf-kappa b”- [*associated\_with*] - “alzheimer's disease”

The knowledge path above indicates that the nf-kappa b signaling pathway can be activated by the transcription factor traf6, which subsequently impacts AD.

The knowledge paths reveal the probable potential links between TF and AD, and also provide an interpretable semantic information pertaining to these links.

# Introduction

**Evidence chains** are selected based on knowledge paths. They connect important entities and relationships, which can reveal key information of a specific field. They are also deemed of high importance amidst all paths.

Evidence chains can connect dispersed knowledge from various research directions, providing interpretable evidence support for the association between the source node and the target node.

**Evidence chain:** “lrwd1”- [*predisposes*] - “tumorigenesis”- [*affects*] -“apoptosis”- [*affects*] -“transcription genetic ”- [*associated\_with*] -“mutation” [*causes*] -“alzheimer's disease”



**KP1:** “lrwd1”- [*predisposes*] - “tumorigenesis”<sup>[1]</sup>

**KP2:** “tumorigenesis”- [*affects*] -“apoptosis”<sup>[2]</sup>

**KP3:** “apoptosis”- [*affects*] -“transcription genetic ”- [*associated\_with*] -“mutation” [*causes*] -“alzheimer's disease”<sup>[3]</sup>

*The interpretable evidence chain can accurately reveal the molecular cascade evidence between the antibody lrwd1 and Alzheimer's disease(AD), which provides a new perspective for the study of the etiology of AD.*

[1] Hung J H, Cheng H Y, Tsai Y C, et al. LRWD1 expression is regulated through DNA methylation in human testicular embryonal carcinoma cells[J]. Basic and Clinical Andrology, 2021, 31: 1-10.

[2] Tong X, Tang R, Xiao M, et al. Targeting cell death pathways for cancer therapy: recent developments in necroptosis, pyroptosis, ferroptosis, and cuproptosis research[J]. Journal of hematology & oncology, 2022, 15(1): 174.

[3] Gazestani V, Kamath T, Nadaf N M, et al. Early Alzheimer's disease pathology in human cortex involves transient cell states[J]. Cell, 2023, 186(20): 4438-4453. e23.

# Outline

**1. Introduction**

**2. Methodology**

**3. Case Study**

**4. Discussion**

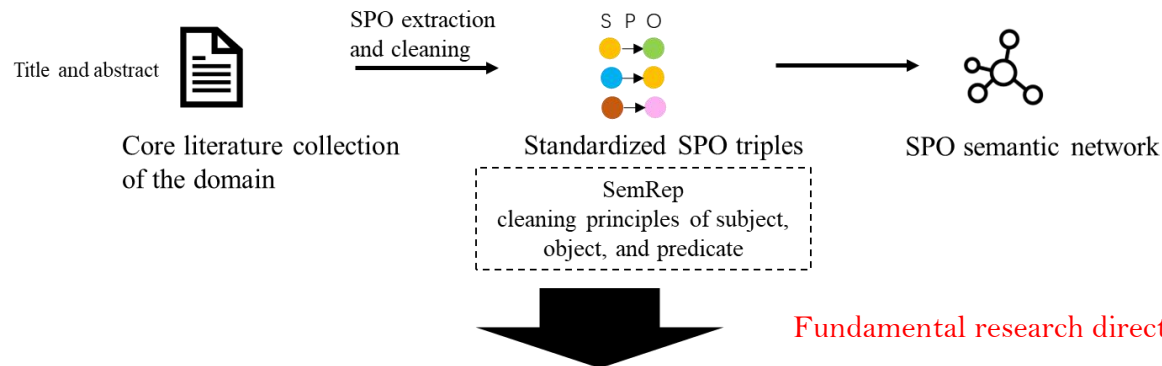




# Framework

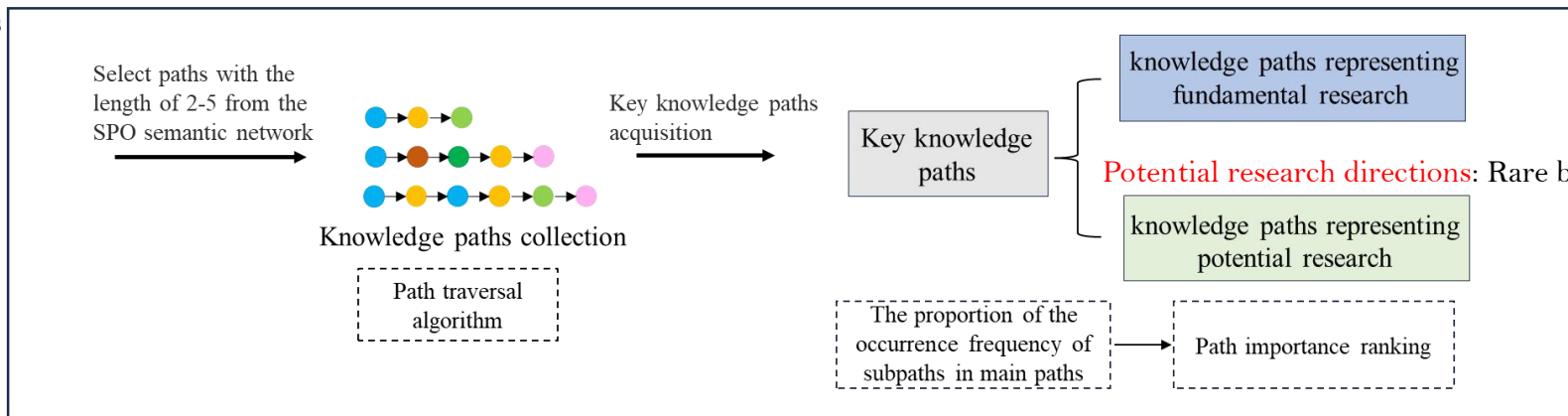
## Step1:

### Knowledge graph construction



## Step2:

### Key knowledge paths mining



**Fundamental research directions:** frequent and important research.

This type of paths need to contain a portion of the knowledge units in order to demonstrate that this kind of knowledge paths can reflect the vital knowledge of the field.

sub-paths appear more than 5 times in the main paths

**Potential research directions:** Rare but important.

The knowledge units in this type of paths should be included in the main paths of the field as much as possible.

the length of the paths affects the occurrence of sub-paths in the main paths.

set thresholds for different path lengths for separate selection.

25% for path length=2, 50% for path length=3, 60% for path length=4, and 60% for path length=5.

The **main paths** are obtained by combining k-shortest path, PageRank and Betweenness Centrality algorithms.

The **main paths** represent important research directions. These kinds of paths typically have significant bridging roles, are crucial to the diffusion and flow of knowledge, and reflect core directions of the field.

$$\text{Strength of SPO } (A, r, B) = \frac{p(A,r,B) - P(A,r)P(r,B)}{\sqrt{p(A,r)(1-P(A,r))} \sqrt{p(r,B)(1-P(r,B))}}$$

$$= \frac{zm - xy}{\sqrt{xy(m-x)(m-y)}}$$

$$\text{KPS (path1)} = \sum_n \text{KPS}(A, r, B) / n$$



# Framework

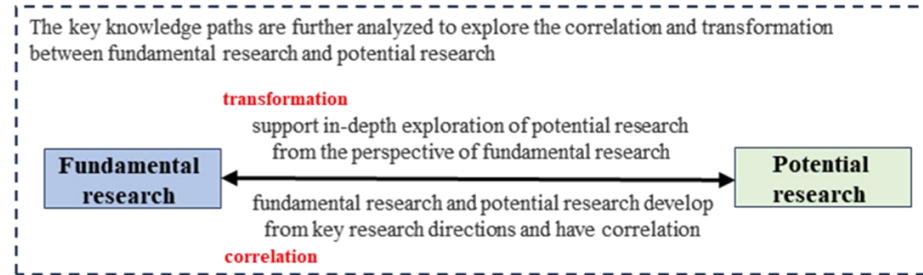
## Step3:

Domain knowledge  
discovery

Fundamental research and potential  
research knowledge discovery



Theoretical basis



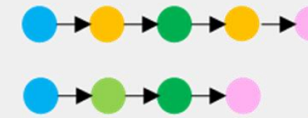
Fundamental  
research

representing the core  
basic research

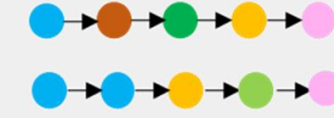
Potential  
research

representing potential  
innovation research

Dimension 1: Important knowledge paths analysis based on semantic expressed

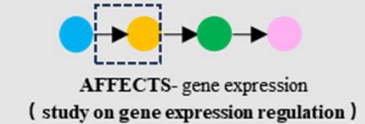


Potential knowledge correlation/weakly connected paths

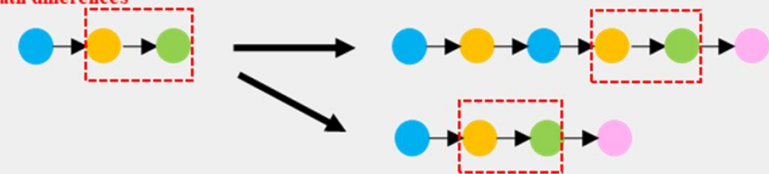


Knowledge paths ranking  
based on its importance

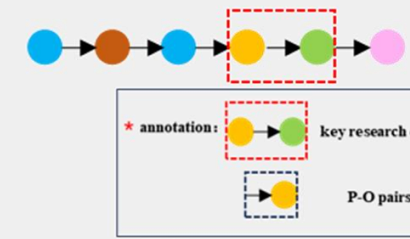
Dimension 2: Major research direction analysis based on P-O pairs



Dimension 3: Correlation analysis based on knowledge path differences



two different research branches are formed based on key research directions : fundamental research and potential research



# Outline

**1. Introduction**

**2. Methodology**

**3. Case Study**

**4. Discussion**

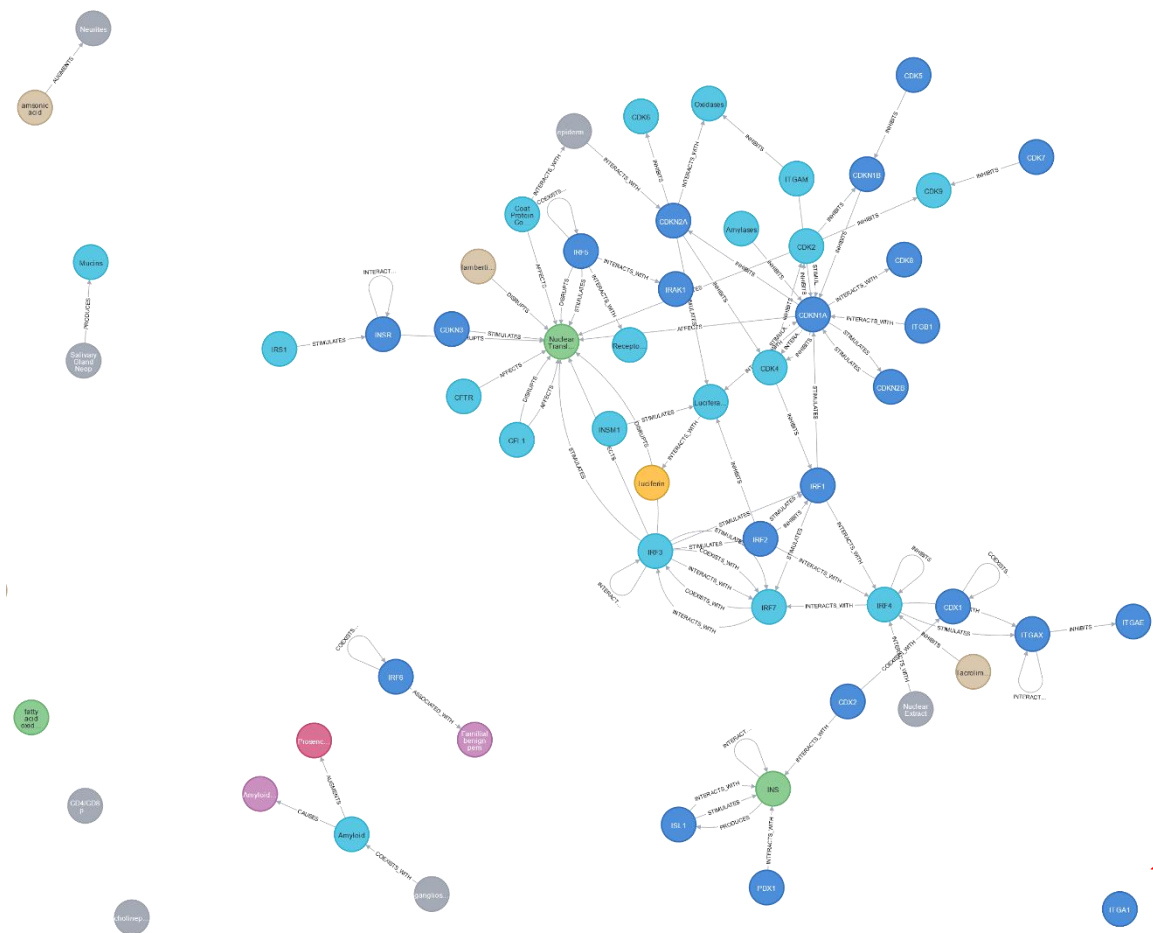


# Alzheimer's Disease (AD)

## Explore the associations between transcription factor (TF) and AD

### ➤ Knowledge graph construction

#### AD KG



#### Retrieval strategies

Data source	Dataset	Search strategy	Literature number
PubMed	TF ∪ AD	(transcription factor[MeSH Major Topic]) OR (Alzheimer's Disease [MeSH Major Topic]) Filters: Journal Article, from 2015 - 2024	190,694

#### SPOs cleaning

follow the principles of relevance, connectivity, uniqueness and salience.

includes three steps: entity cleaning, predicate cleaning and SPO screening.

**16 semantic relationships,**  
**totaling 51,380 entities and 554,834 relationships**

## ➤ Evidence chains (key knowledge paths) mining

- Knowledge paths extraction

traverse the paths between nodes TF and AD with lengths of 2-5.

- Evidence chains acquisition

In this study, evidence chains represent the fundamental and potential research directions of this field.

**Evidence chains** are obtained based on the proportion of the occurrence frequency of sub-paths in main paths.

Combine k-shortest path, PageRank and Betweenness Centrality algorithms to mine **main paths**.

**Fundamental research paths**: paths with sub-paths appearing more than 5 times in the main paths (totaling 260 paths).

**Potential research paths**: select paths whose sub-paths occur more frequently in the main paths, and set thresholds for different path lengths for selection (totaling 240 paths).

- Evidence chains ranking

A simple KP strength

(measure the strength of the Path using the Pearson correlation):

$$\begin{aligned}\text{Strength of SPO (A, r, B)} &= \frac{p(A,r,B) - P(A,r)P(r,B)}{\sqrt{p(A,r)(1-P(A,r))}\sqrt{p(r,B)(1-P(r,B))}} \\ &= \frac{zm - xy}{\sqrt{xy(m-x)(m-y)}}\end{aligned}$$

$$\text{KPS (path1)} = \sum_n \text{KPS}(A, r, B) / n$$

- Evidence chains

Examples of top basic research paths

Path num	Knowledge paths	Strength
Path 1	xanthine oxidase-COEXISTS_WITH-reactive oxygen species-AFFECTS- aging-ASSOCIATED_WITH-malignant neoplasms-COEXISTS_WITH-alzheimer's disease	0.0347
Path 2	JUNB-PREVENTS-malignant neoplasms-COEXISTS_WITH-alzheimer's disease,	0.0205
Path 3	catechol oxidase-INHIBITS-multicatalytic endopeptidase complex-CAUSES-autophagy-AUGMENTS-transcription, genetic -CAUSES-alzheimer's disease	0.0191

.....  
Examples of top potential research paths

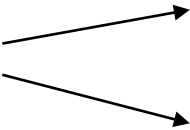
Path num	Knowledge paths	Strength
Path 1	co-repressor proteins-DISRUPTS-transcription, genetic-CAUSES-alzheimer's disease	0.0282
Path 2	SMARCC2-DISRUPTS-gene expression-AFFECTS-autophagy-AUGMENTS-transcription, genetic-CAUSES-alzheimer's disease	0.0272
Path 3	SLC11A2-DISRUPTS-growth-COEXISTS_WITH-autophagy-AUGMENTS-transcription, genetic-CAUSES-alzheimer's disease	0.0246

- Correlation between fundamental research and potential research

Based on a key research direction, some develop into potential research directions, others develop into the fundamental research directions.

Key research directions

autophagy-AUGMENTS -transcription, genetic-CAUSES-alzheimer's disease



Research branches

Paths of fundamental research: NES gene-AFFECTS -metabolism-AFFECTS -autophagy-AUGMENTS -transcription, genetic - CAUSES-alzheimer's disease

Paths of potential research: MST1- DISRUPTS-apoptosis-CAUSES- down-regulation - COEXISTS\_WITH- autophagy- AUGMENTS-transcription, genetic-CAUSES -alzheimer's disease

- Major research directions

Use P-O pairs to reveal the specific research directions in the field.

Main research directions of fundamental research:

Neuronal protection, Signaling pathway, Oxidative stress, Pathological mechanism, The co-morbidities, associations, and mechanisms of AD and cancer.

Main research directions of potential research:

Gene expression regulation, Pathogenic indicators, Apoptosis, Autophagy regulation, Neuroinflammation.

## Example of an evidence chain

**SLC11A2-DISRUPTS-growth-COEXISTS\_WITH-autophagy-AUGMENTS-transcription, genetic-CAUSES-alzheimer's disease**

### Explanation

This evidence chain suggests that certain genetic variants in the SLC11A2 gene may affect an individual's susceptibility to AD and play a role in the onset of AD.

#### **SLC11A2-DISRUPTS-growth**

The transporter proteins that the SLC11A2 gene encodes may have an indirect impact on cell growth and function

#### **COEXISTS\_WITH-autophagy**

as well as autophagy processes by altering the intracellular environment and ion homeostasis [1].

#### **AUGMENTS-transcription, genetic-CAUSES-alzheimer's disease**

There is strong evidence of an allelic association between AD and SNP rs407135 of SLC11A2, where the variant allele is protective [2].

This evidence chain can provide a thorough understanding of SLC11A2's role in AD, which will help provide crucial clues for the exploration of the pathogenesis of AD in the future, and may also yield potential targets for new treatment strategies.

### Evaluation of evidence chains

The evidence chains obtained in this study have been reviewed by experts, indicating that the results are somewhat reasonable.

[1] Robertson, K.V., Rodriguez, A.S., Cartailier, J.-P., Shrestha, S., Schleh, M.W., Schroeder, K.R., Valenti, A.M., Kramer, A.T., Harrison, F.E., and Hasty, A.H.: 'Knockdown of microglial iron import gene, Slc11a2, worsens cognitive function and alters microglial transcriptional landscape in a sex-specific manner in the APP/PS1 model of Alzheimer's disease', Journal of Neuroinflammation, 2024, 21, (1), pp. 238

[2] Roca-Agujetas, V., de Dios, C., Abadin, X., and Colell, A.: 'Upregulation of brain cholesterol levels inhibits mitophagy in Alzheimer disease', Autophagy, 2021, 17, (6), pp. 1555-1557

# Outline

1. Introduction
2. Methodology
3. Case Study
4. Discussion





# Discussion

---

## Advantages

- Employ evidence chains for mining, which is more **in-depth** in mining knowledge, more **comprehensive** in revealing knowledge associations, provides **finer semantic granularity** of knowledge, and **more interpretable** in knowledge discovery results.
- The mining of evidence chains can reveal key research directions in the field. This will contribute to the generation of hypotheses and the inspiration of research ideas, thereby promoting knowledge discovery and driving major breakthroughs in the field.

## Challenges

- The mining of the evidence chains is not deep enough and the correlation of knowledge units in the chains is not that strong. The mining was done on a local network, so the evidence chains may be incomplete. When mining was done on larger datasets, the evidence chains are likely to be susceptible to noise, and chains with cycles and even contradictions may be seen.

# Discussion

---

## Future Directions

- Leverage the prompt engineering of LLM(Large language model) to facilitate the automatic mining and filtering of **evidence chains**. This approach enhances the synergistic inference between LLM and evidence chains, thereby improving the performance of downstream tasks. Consequently, it increases the reliability of reasoning results while providing interpretable evidence to support these results.
- The process of evidence chains mining should consider the characteristics of the field in order to **generate domain-specific evidence chains**. Moreover, these evidence chains **require evaluation by domain experts** to verify their efficacy and ensure they represent meaningful research directions of the field. **Weak signal and potential knowledge associations are especially important**. This will facilitate an in-depth exploration of evidence chains.



**Thank You**

---

