Knowledge Graphs for Early Target Discovery

Vlad Kim

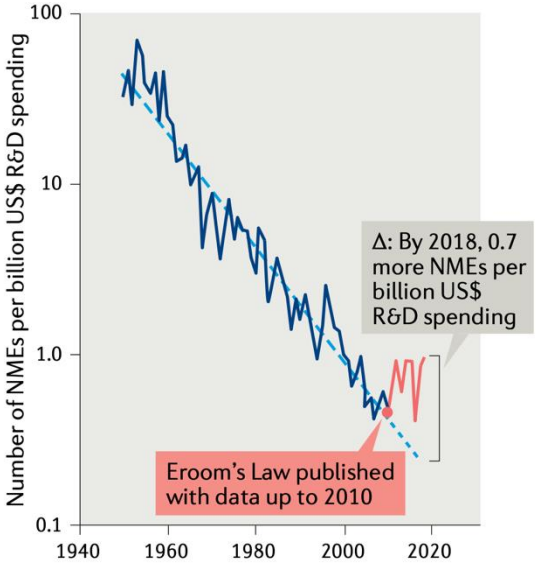
Bayer Pharma R&D

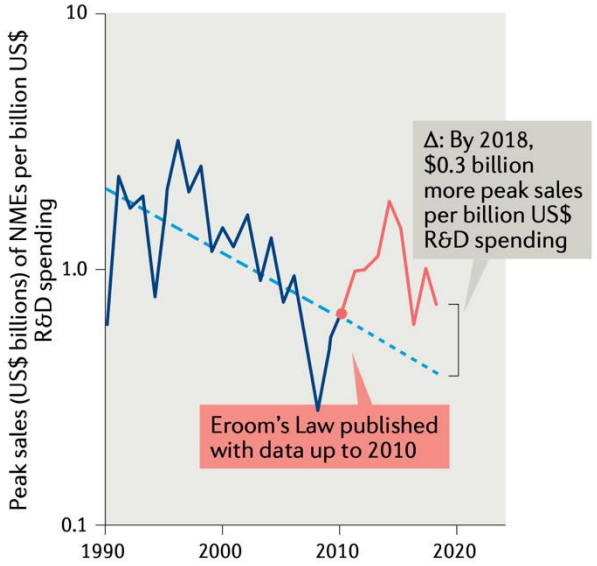
July 2, 2026

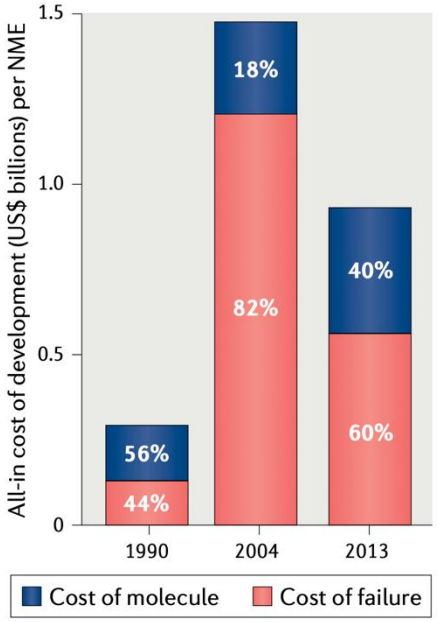
neo4j GraphTalk Pharma & Life Sciences

Introducing target discovery

Drug discovery and develpoment are slow and costly







Ringel et al. "Breaking Eroom's Law" *Nat. Rev. Drug. Discov.* **2020**

Cost estimates: Paul *et al.* (**2010**) *Nat. Rev. Drug. Discov.* **9**, 203−214

Time estimates: Ashburn *et al.* (**2004**) *Nat. Rev. Drug. Discov.* **3**, 673−683

Recent overall cost estimate is $1.0 − 1.6 Billion (95% CI) on average:
Wouters et al. (**2020**) JAMA 323, 9, 844−853

Disease prioritization and target discovery are earliest discovery stages

What are areas of unmet medical need?

↓

What disease to focus on?

↓

What subtypes of this disease are in scope?

↓

What genes are relevant for this subtype?

↓

...

Drug discovery is a multistage process, with **disease prioritization** and **target discovery** as the earliest stages.

Discovering novel biological targets is challenging

1 COMPLEXITY OF BIOLOGY

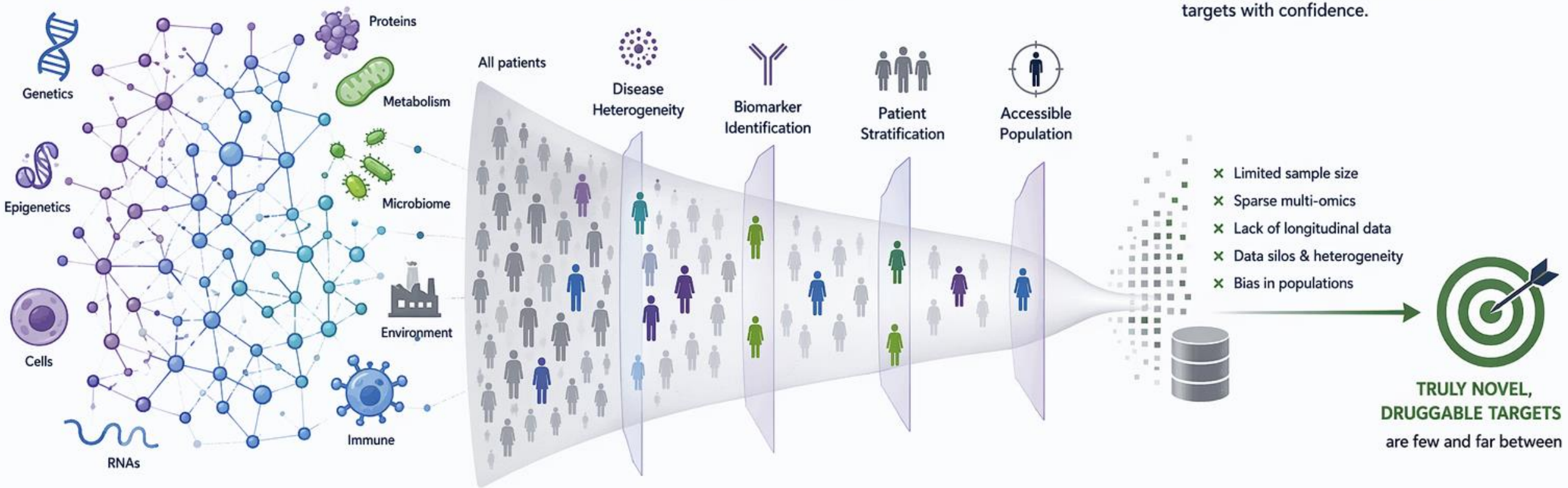
Diseases arise from highly interconnected networks, not single pathways.

2 NEED FOR A WELL DEFINED PATIENT POPULATION

Only a subset of patients have the right biology — identifiable via biomarkers.

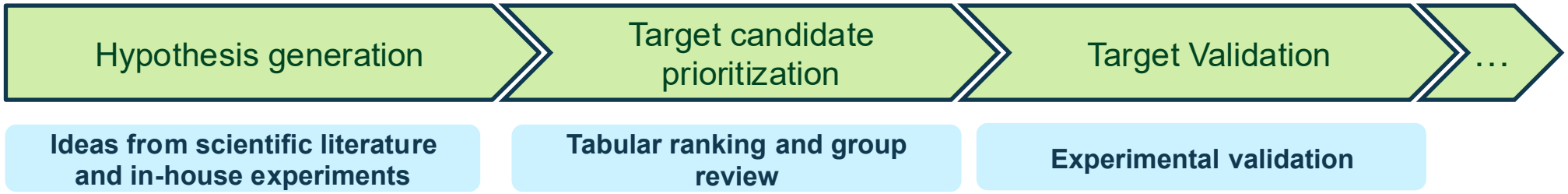
3 LACK OF BROAD & DEEP PATIENT BIOLOGICAL DATASETS

Limited, biased, or shallow data makes it hard to see true biology and find novel targets with confidence.



Target discovery in focus

Target discovery stages

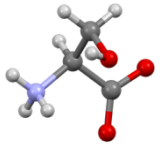


- As a whole, target discovery is not a standardized process.
- However, certain stages include “routine” tasks that could be automated.

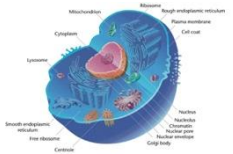
What makes a good target?

Target selection considerations:

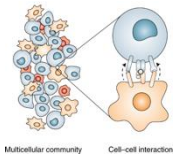
- Established mechanistic gene-disease link
- Supporting experimental evidence in cellular and animal models
- Evidence in human data (biobanks, clinical trials)
- Druggable/actionable target
- Safety assessment
- Competitive landscape (unmet need)
- Identified disease subpopulation affected by the target
- Associated biomarker availability



Molecules

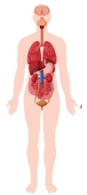


Intracellular processes



Individual cells

Tissues and organs



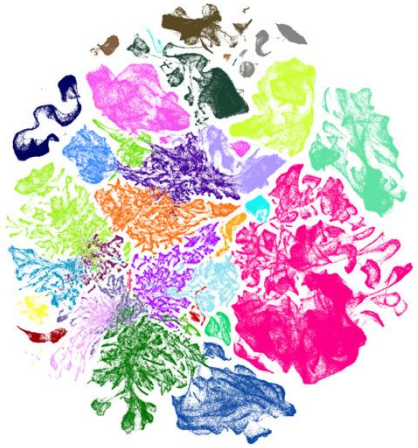
Organism

Population

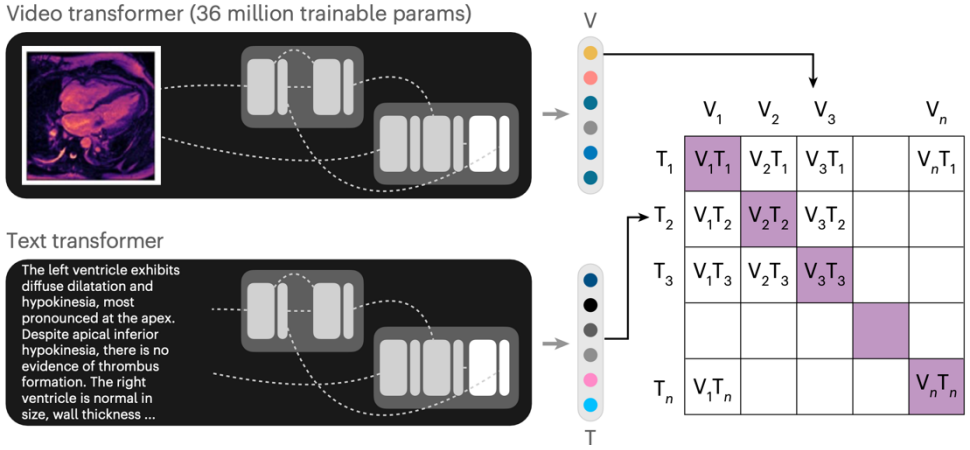
AI/ML for target discovery

What are areas in which AI/ML excels at?

Pattern recognition in high-dimensional data



Multimodal data integration

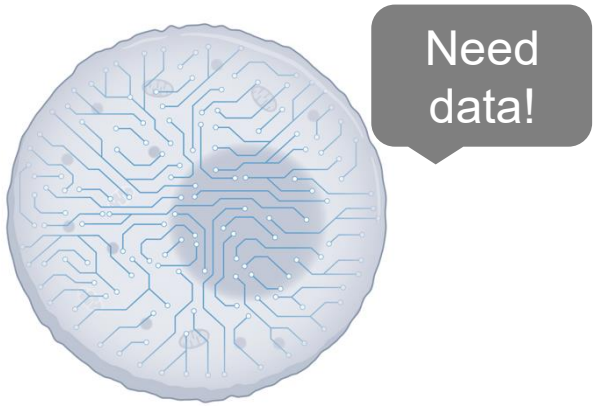


Classification and regression

Generative modelling

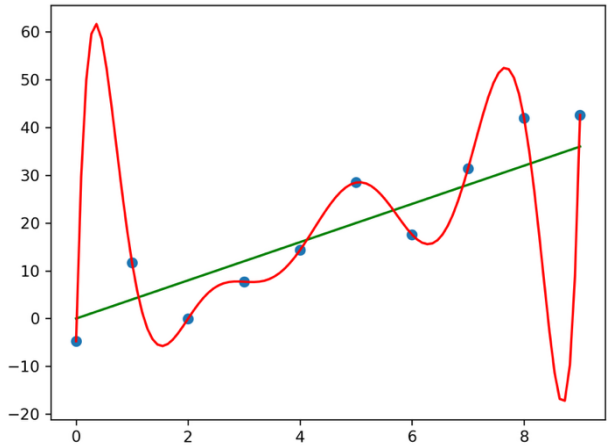
Areas in which AI/ML struggles

Knowledge generation *de novo*



Experimental data generation without any training data.

Predictive ML with sparse data



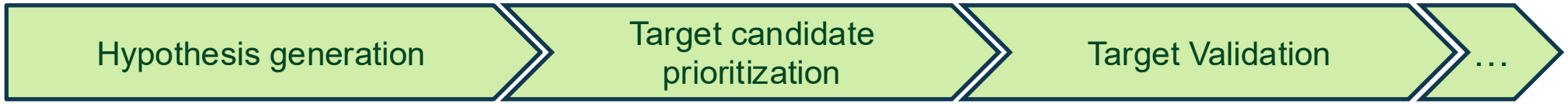
Not enough data to train a model => either a simple model or overfitting

Nested questions with vast knowledge gaps

Questions that require a chain of real-world experiments

How can AI/ML accelerate target discovery?

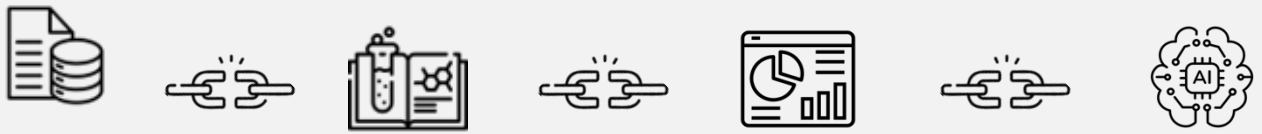
Target discovery stages



Ideas from scientific literature and in-house experiments

Tabular ranking and group review

Experimental validation

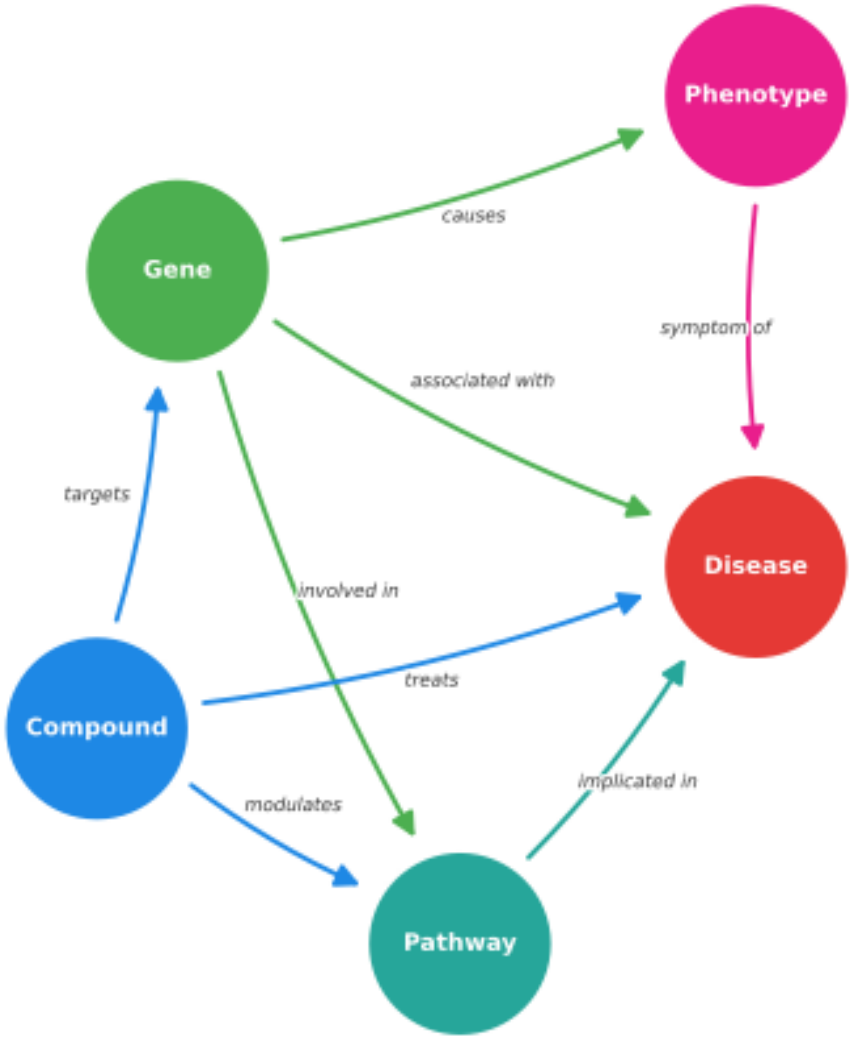


Knowledge aggregation | **Literature mining** | **Analytics and Insights** | **Project-specific ML-models**

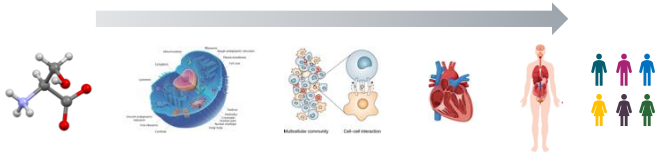
AI/ML could accelerate

- **Data integration** from literature, omics, internal and external databases.
- **Hypothesis generation** and candidate prioritization through knowledge retrieval.
- Target validation through bespoke ML models

Knowledge graphs: integrating any data modality in a unified data resource



From molecules to disease subpopulations



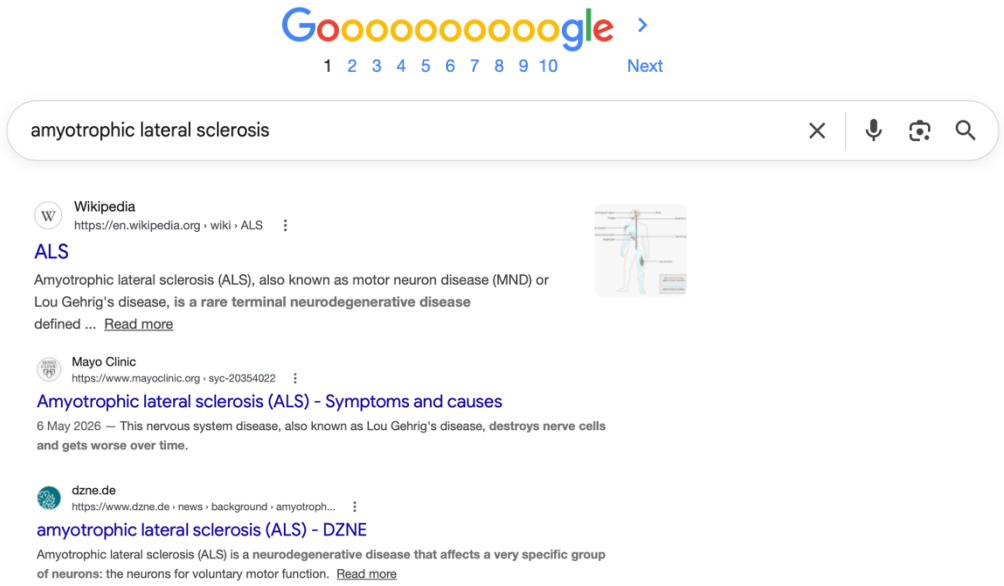
Impact: knowledge aggregation across various data sources, modalities, experimental assays and biological subdomains.

By itself, a KG is a convenient data structure**. How can we best harness this wealth of data?**

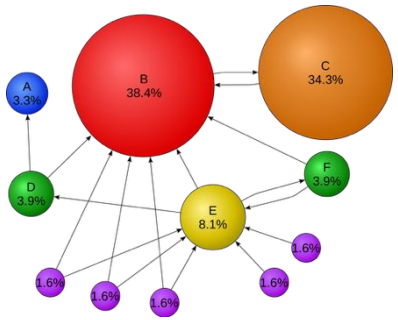
Knowledge graph reasoning

Invisible hand of evidence ranking in information retrieval

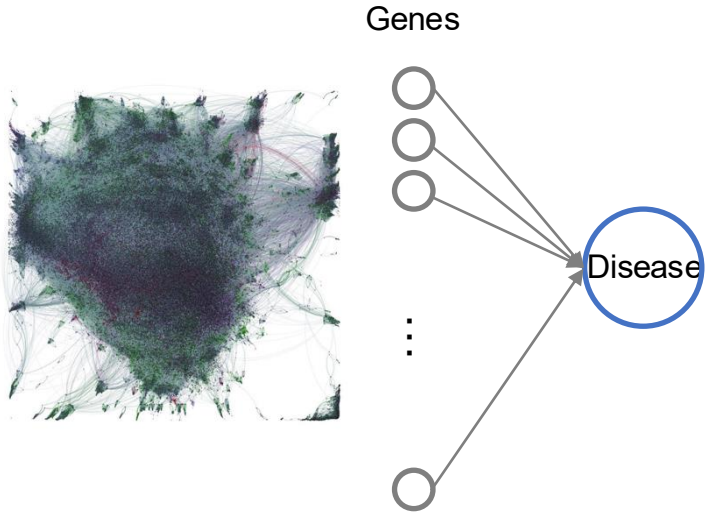
Search result ranking by relevance



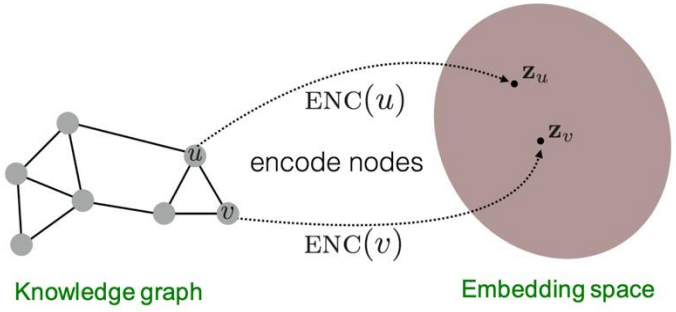
Graph algorithm (PageRank)



Ranking evidence in knowledge graphs

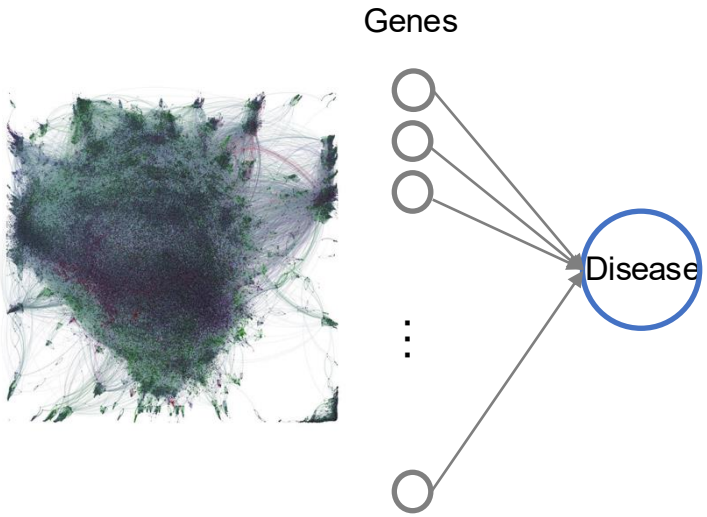


Knowledge graph reasoning



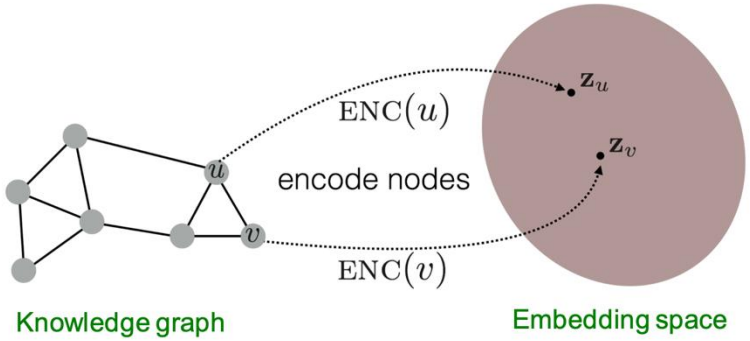
Evidence ranking and link prediction are two sides of KG reasoning

Evidence ranking



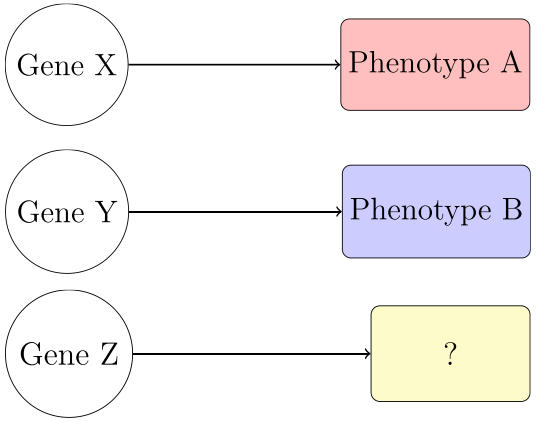
Given a knowledge graph with millions of relations, how can we rank evidence?

KG reasoning



Use graph structure to learn embeddings of entities (e.g., genes and diseases) and *reason in the embedding space*.

Link prediction



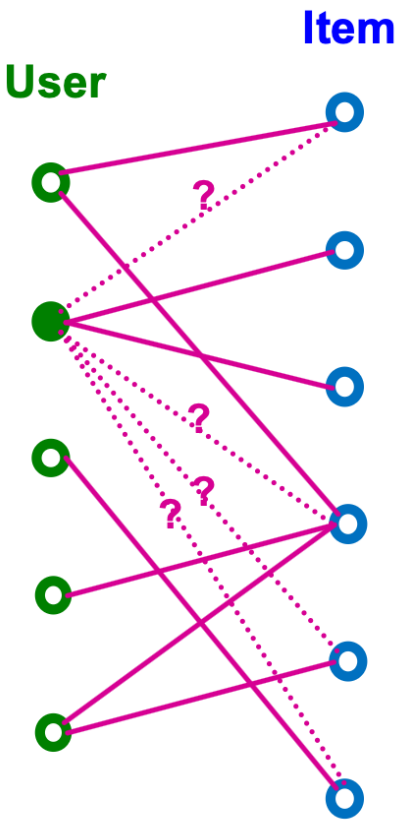
Can we use existing knowledge to predict unobserved relations?

Graphic from Jure Leskovec's slide deck

Graph data needs graph ML!

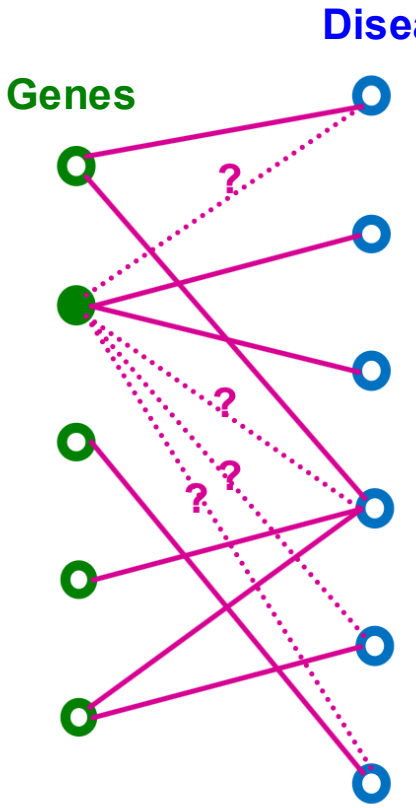
Recommender systems for biomedical knowledge graphs

Recommender system



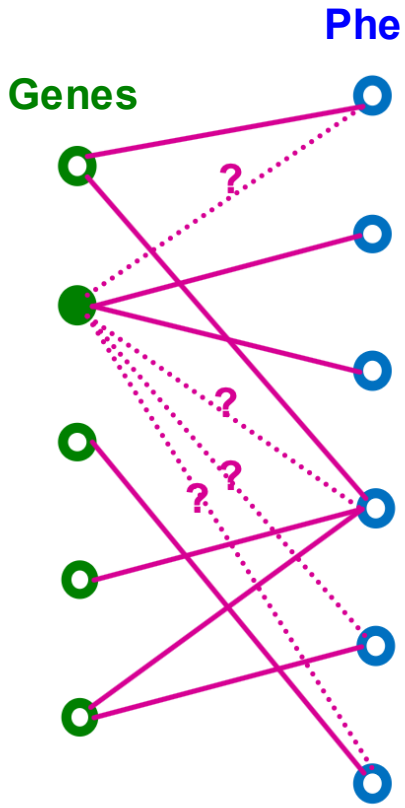
Given a set of users, items and the past transactions, predict the next item that a user is likely to buy.

Gene-disease “recommendations”



Given a set of genes, diseases, and reported associations among them: **predict most likely disease-linked genes.**

Gene-phenotype associations



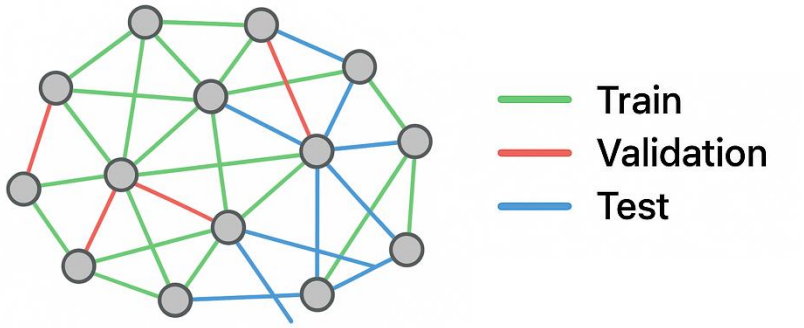
Given a set of genes, phenotypes, and reported associations among them: **predict phenotypes associated with genes.**

...

Can be extended to any KG relation!

Graphic adapted from Jure Leskovec’s slide deck

How do we evaluate KG reasoning models?



Evaluation metric: hit@k
On a held-out test set, check the frequency at which an existing link occurs in the top *k* predictions.

Internal benchmarking results

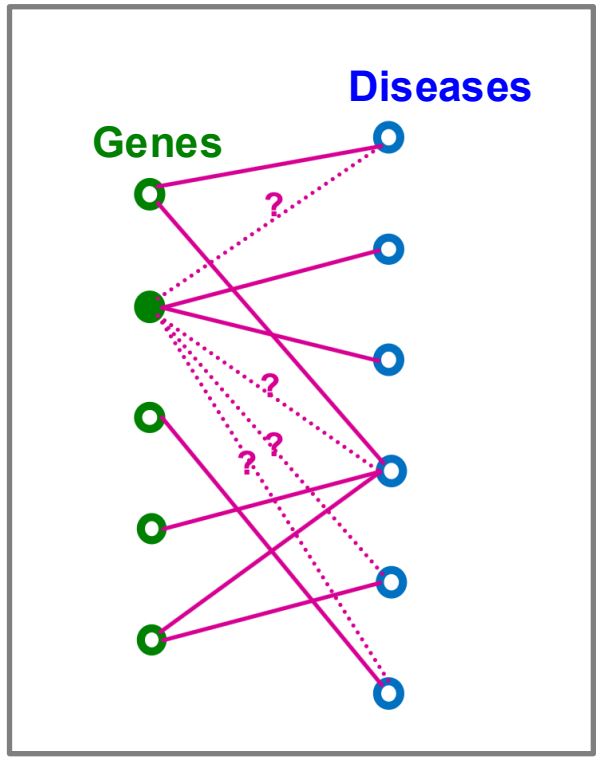
Model	Hit@10	Hit@20	Hit@50	Hit@100
Embedding-based	**0.44**	0.52	0.59	0.64
Graph convolutional network	0.40	0.47	0.55	0.61
Transformer	0.40	**0.61**	**0.86**	**0.91**

Embedding-based approaches are good enough in most cases. A transformer model performs the best, but only predicts well the relations of high-degree nodes.

Top-ranked genes are significantly more likely to make it past the pre-clinicial stage than differentially expressed genes or gene-disease pairs with high citation count.

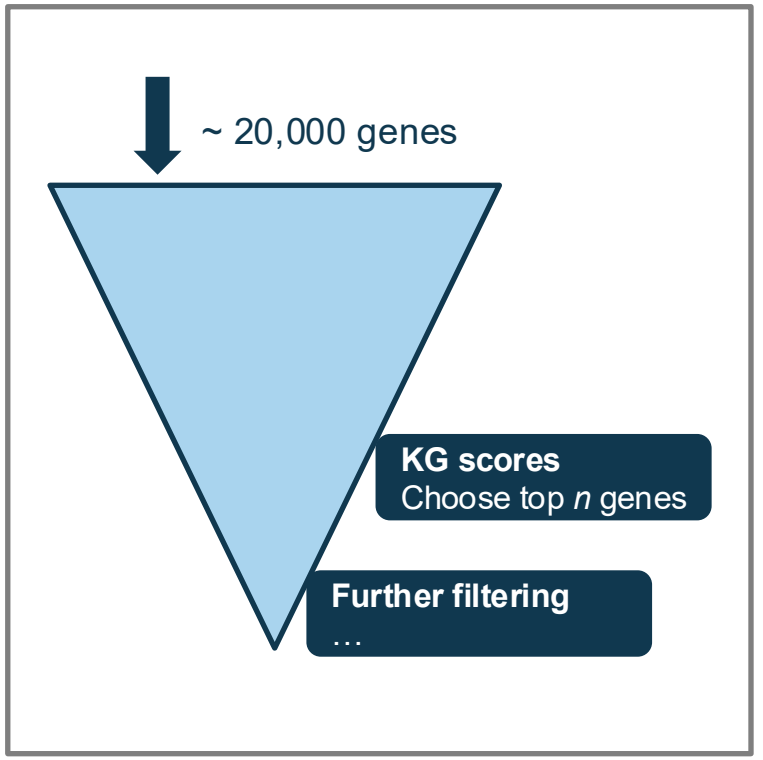
KG reasoning for hypothesis generation and candidate prioritization

Ranked gene-disease associations



Predict disease-associated genes using knowledge graph reasoning models.

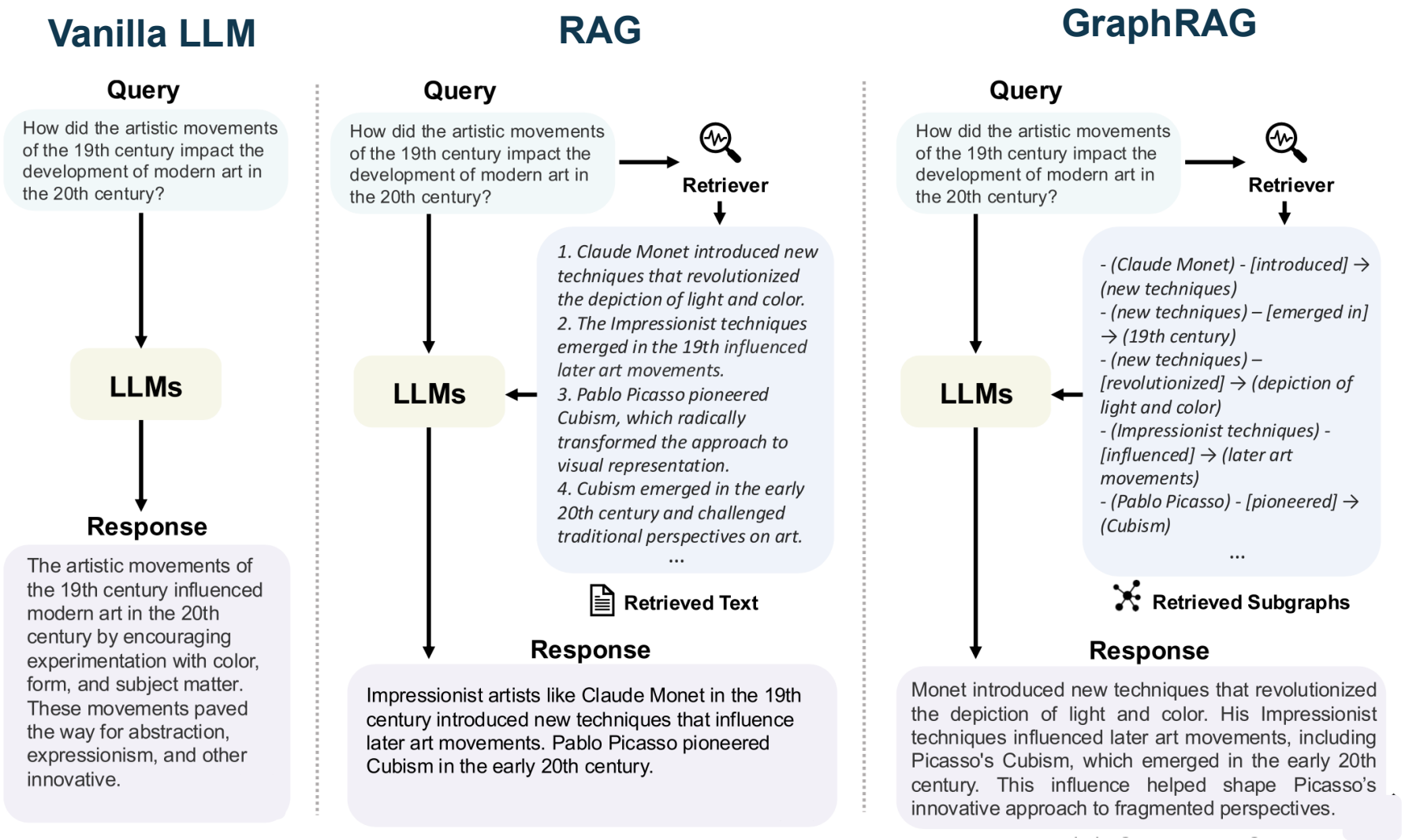
Target candidate nomination



Select genes with the highest KG evidence scores and apply further filtering criteria.

AI-based knowledge retrieval

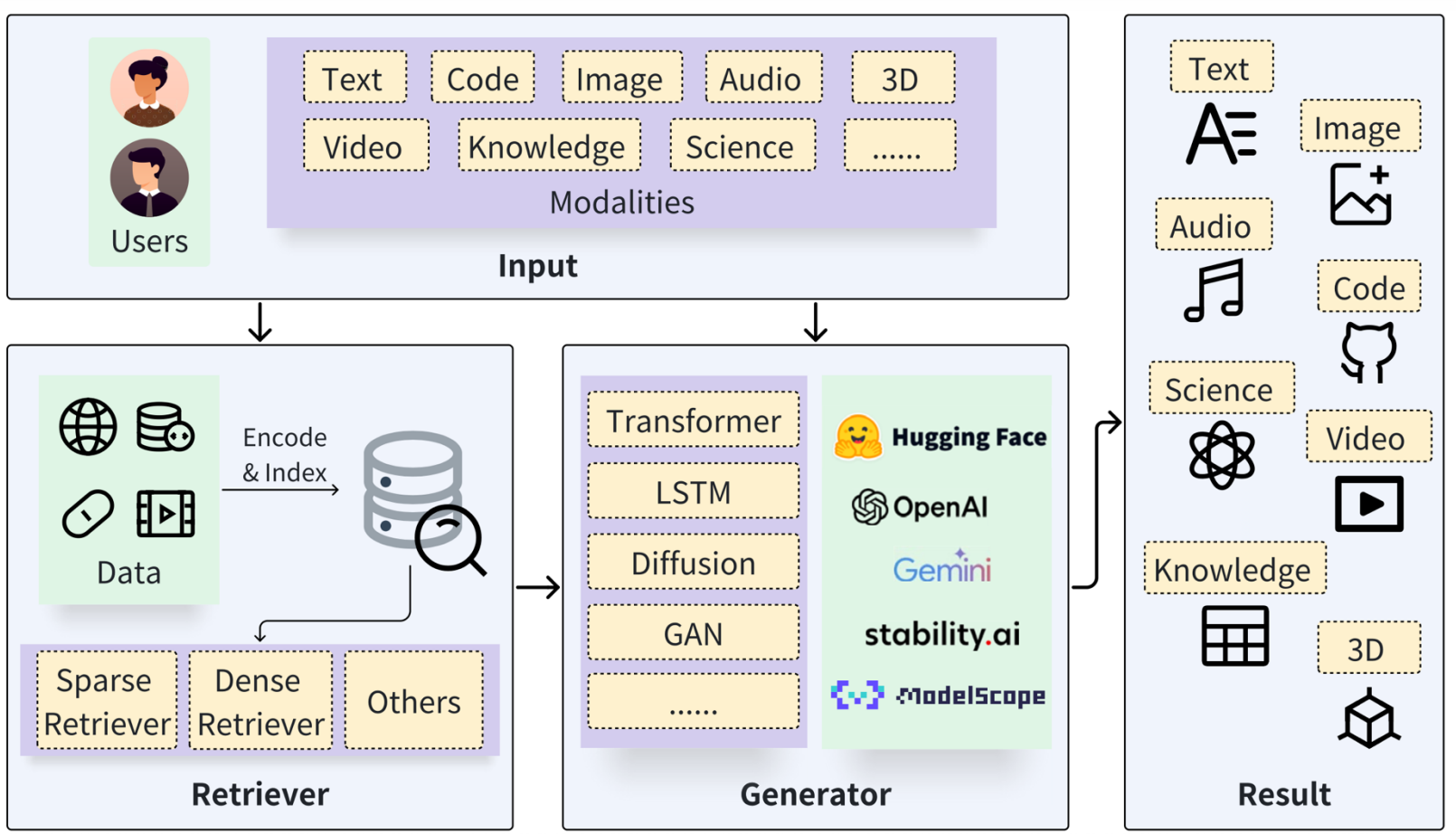
Knowledge retrieval with LLMs



Retrieval augmented generation (RAG) helps us ground LLM responses with knowledge graph links or literature citations.

Peng et al., *ACM Transactions on Information Systems, Vol. 44, Issue 2,* **2026**

Multimodal RAG



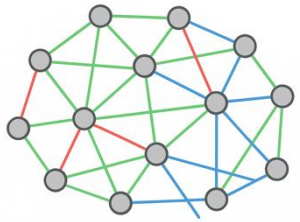
RAG can be implemented for any data modality: text, graphs, images, molecules, etc.

Zhao et al., *Data Science and Engineering - Springer,* **2026**

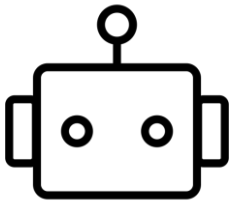
Hybrid Literature-GraphRAG for target discovery

Scientific literature

Knowledge graph



LLM

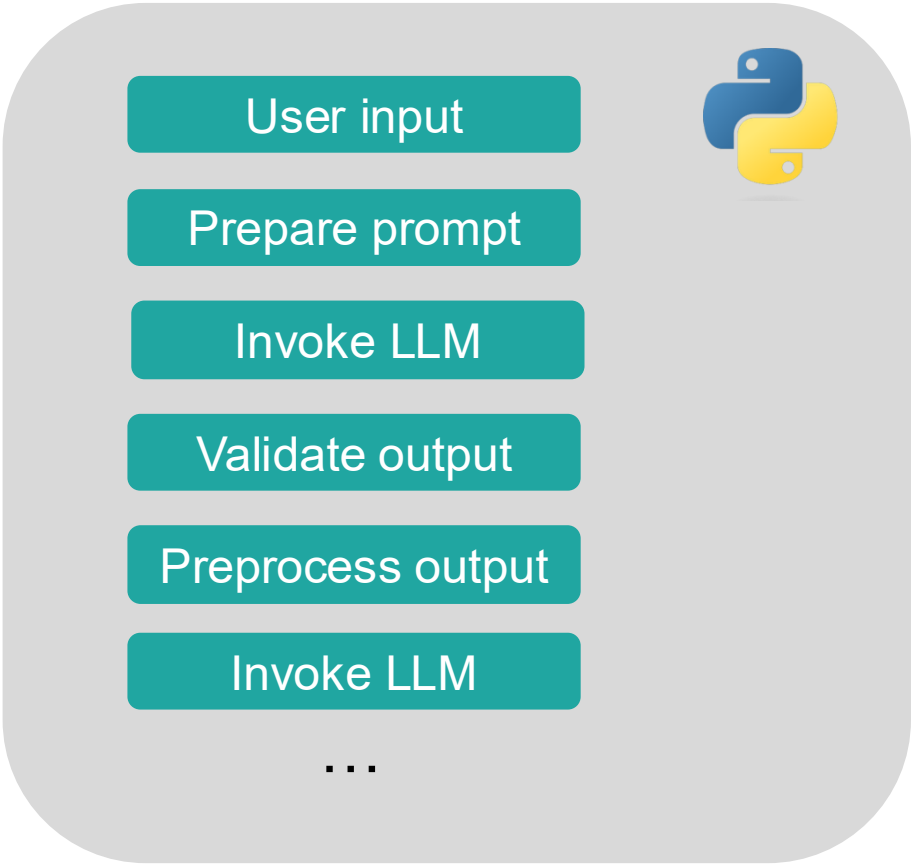


- AI-based **evidence search, data aggregation and report generation**.
- The workflow provides references to scientific publications and KG data, such as gene-gene interactions and pathways.
- Internally, two GraphRAG flavors: MCP server or direct Cypher queries.

Evaluation of RAG workflows is not straightforward. We use retrieval metrics, output variability and token efficiency.

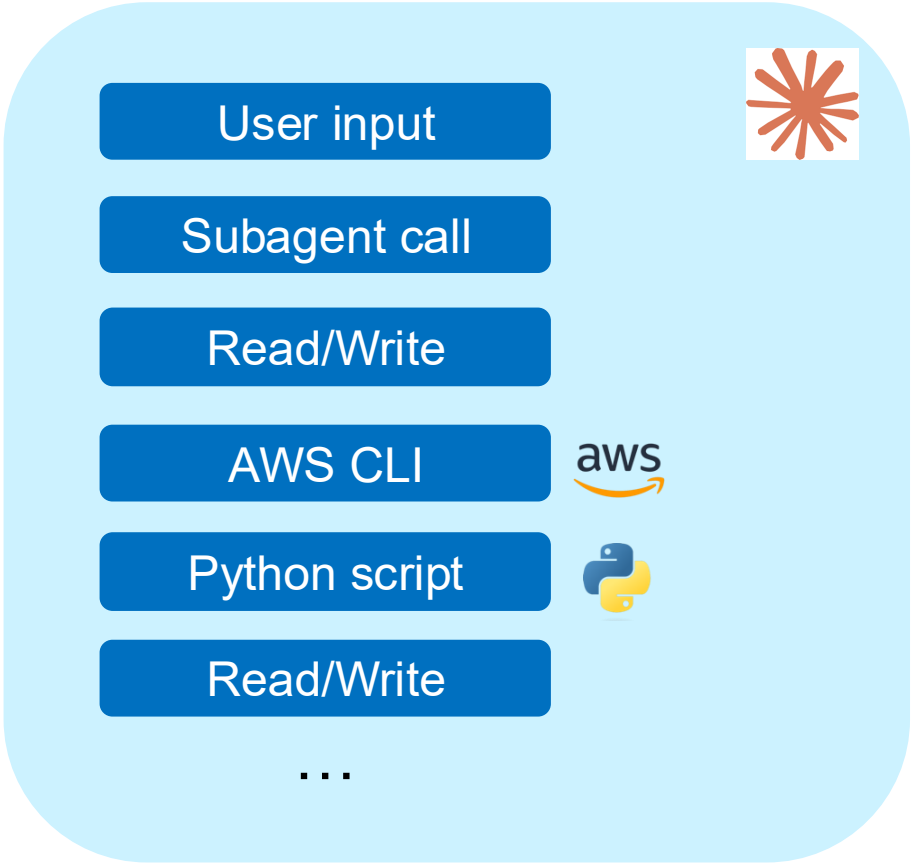
Determinism vs flexibility tradeoff for RAG implementations

Semideterministic Python workflow



More robust, but less flexible

LLM-based workflow

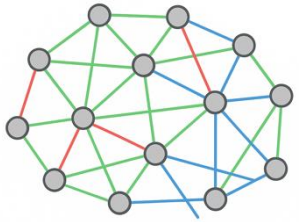


More flexible, but less reproducible

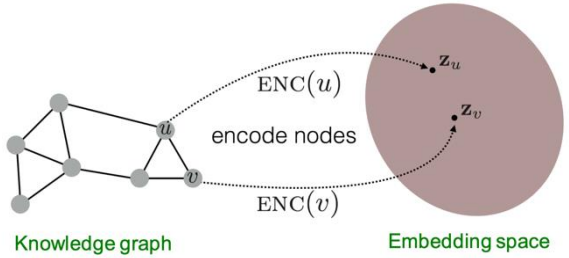
Combining GraphRAG and KG reasoning

Scientific literature

Knowledge graph

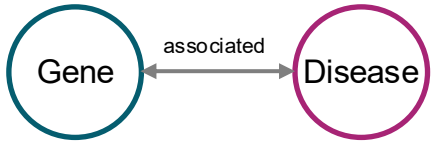


KG reasoning



Extract relations from the literature and ingest into a KG

Use KG reasoning to rank literature-mined relations



LLM evidence score	KG evidence score
Strong	0.7
Strong	0.5
Strong	0.3

- LLMs struggle with assigning quantitative evidence scores. We can ingest the RAG data into the KG and use graph ML to obtain **evidence scores**.
- **To save tokens in GraphRAG applications**, consider only top *n* relations in each graph-based query.

Do users trust our ML outputs?

Common roadblocks to AI adoption

Lack of transparency

Sharing CSV files without context

Domain gap

“How exactly does this ML model work?”

Cognitive overload

Too many ML models, AI solutions, apps, etc.

Scientific skepticism

Rejecting the ”null hypothesis”

Mitigation strategies

Show results and all necessary context, including training data sources.

Let the data speak for itself: data and objectives first, ML second

Produce less, but higher quality

ML model outputs will always be questioned by scientists!

Thank you for listening!

Huge thanks to my Data Science & Artificial Intelligence (DS&AI) colleagues at Bayer for all the amazing work and their infallible support
🙏