

Automated Knowledge Discovery in Bioscience

Ivan Muller¹, Elena Lindberg²

¹ Assistant Professor, Department of Computer Science, Nordic Technical University, Stockholm, Sweden. Email: ivan.muller464@yahoo.com | ORCID: 7363-6758-4673-7429

² Professor, School of Data Science, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: elena.lindberg928@yahoo.com | ORCID: 8951-5867-8621-7223

ABSTRACT

Biomedical literature doubles every 4-5 years, with over 1.5 million PubMed-indexed articles published annually and over 38 million articles in the current database. No researcher can read even a fraction of the literature relevant to their field, creating a knowledge fragmentation problem: critical connections between findings in different subfields remain undiscovered for years. Automated knowledge discovery (AKD) uses natural language processing, knowledge graph construction, and reasoning algorithms to extract, integrate, and generate novel hypotheses from the biomedical literature and structured databases. Swanson's (1986) pioneering work demonstrated that literature-based discovery could identify non-obvious connections (fish oil treats Raynaud's disease) years before experimental confirmation. Modern AKD systems leverage large language models (LLMs), biomedical knowledge graphs, and multi-modal data integration. We present the Automated Knowledge Discovery Assessment Framework (AKDAF), evaluating five AKD approaches -- biomedical NLP entity extraction, knowledge graph link prediction, LLM-powered hypothesis generation, literature-based discovery systems, and multi-modal knowledge fusion -- across four bioscience discovery tasks (drug repurposing, gene-disease association, pathway cross-talk identification, and biomarker discovery). Our Knowledge Discovery Effectiveness Score (KDES) measures extraction precision, novel hypothesis quality, experimental validation rate, discovery speed, and interpretability. LLM-powered hypothesis generation achieves the highest KDES (0.922) through contextual reasoning over biomedical corpora, while knowledge graph link prediction achieves the highest experimental validation rate (0.940) through structured relational inference.

Keywords: knowledge discovery; biomedical NLP; knowledge graphs; large language models; drug repurposing; literature mining; hypothesis generation; text mining

Citation: Muller and Lindberg [2025]. Automated Knowledge Discovery in Bioscience. DOI:

<https://doi.org/10.5281/zenodo.19549888>

Copyright: © 2025 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2025 Apr 24 Accepted: 2025 Jun 23 Published: 2025 Aug 15

Research Article: Research Article

1. Introduction

1.1 The Biomedical Knowledge Explosion

The biomedical literature is growing at an accelerating rate: PubMed indexed 1.5 million articles in 2024 (4,100 per day), up from 500,000 per year in 2000. A typical biomedical researcher reads 250-300 papers per year (Tenopir et al., 2015) -- less than 0.02% of the annual output. This creates a paradox: more knowledge is being generated than ever, yet individual researchers have access to a shrinking fraction of the total. The consequences are measurable: Fortunato et al. (2018) showed that the disruptiveness of scientific papers has declined over 6 decades, partly because researchers increasingly build on knowledge from their immediate subfield rather than making cross-disciplinary connections. Rzhetsky et al. (2015) demonstrated that biomedical research is concentrated on a small fraction of genes and diseases (50% of publications study the same 4,000 of 20,000 human genes), leaving vast unexplored territory where novel discoveries likely await. Automated knowledge discovery addresses this by computationally traversing the entire literature and database landscape, identifying connections that no human reader could make because they span subfields, organisms, and decades of publication.

1.2 From Text Mining to Hypothesis Generation

Automated knowledge discovery has evolved through three generations. First-generation systems (1990s-2000s) used co-occurrence statistics: if gene A and disease B are mentioned in many papers, they may be associated. Second-generation systems (2010s) used knowledge graph construction: extract entities and relationships from text, build a graph, and predict missing links. Third-generation systems (2020s) use large language models: encode the biomedical literature into transformer representations and generate novel hypotheses through contextual reasoning. Each generation subsumes the previous: LLMs can perform entity extraction, relationship classification, and hypothesis generation in a unified framework. Kovacs et al. (2025) evaluated network medicine approaches that use knowledge graphs for disease analysis -- the same graph-based reasoning powers AKD for drug repurposing and pathway discovery. Jensen and Hansen (2025) demonstrated AI-based drug-target interaction prediction using knowledge graph embeddings -- a direct application of AKD to drug discovery. Section 2

reviews the literature. Section 3 describes AKDAF. Results in Section 4, discussion in Section 5, conclusions in Section 6.

2. Literature Review

2.1 Biomedical NLP and Entity Extraction

Biomedical named entity recognition (NER) identifies mentions of genes, diseases, drugs, and other biological entities in text. PubMedBERT (Gu et al., 2021) pre-trained BERT on PubMed abstracts and full-text articles, achieving state-of-the-art NER performance (F1 0.91 for gene mentions, 0.88 for disease mentions) on the BC5CDR and NCBI Disease benchmarks. Biomedical relation extraction identifies relationships between entities: drug-treats-disease, gene-causes-disease, protein-binds-protein. Wei et al. (2016) developed PubTator, a web service that automatically annotates PubMed articles with entity mentions and relationships, processing 36 million abstracts. BioGPT (Luo et al., 2022) fine-tuned GPT-2 on biomedical literature, demonstrating that generative language models can perform relation extraction and question answering in a unified framework. For structured database extraction, UniProt contains 250 million protein sequences with 570,000 manually curated entries (SwissProt), and the Open Targets Platform integrates genetic association, expression, and literature evidence for 27,000 targets and 18,000 diseases (Ochoa et al., 2023).

2.2 Knowledge Graphs and Link Prediction

Biomedical knowledge graphs (KGs) encode entities as nodes and relationships as edges, enabling computational reasoning over structured biological knowledge. Hetionet (Himmelstein et al., 2017) integrates 47,031 nodes (genes, diseases, compounds, anatomies) and 2,250,197 edges from 29 databases, demonstrating that graph-based drug repurposing correctly prioritises known drug-disease pairs with AUROC 0.97. DRKG (Drug Repurposing Knowledge Graph, Ioannidis et al., 2020) contains 97,238 entities and 5.8 million edges, and knowledge graph embedding models (TransE, RotatE, ComplEx) predict novel drug-disease associations by learning low-dimensional entity representations. Nicholson and Greene (2020) benchmarked 6 KG embedding methods on biomedical link prediction, finding that RotatE achieves the best drug repurposing performance (AUROC 0.94). Literature-based discovery (LBD), pioneered by Swanson (1986), identifies implicit connections through shared

intermediaries: if A is related to B in the literature and B is related to C, then A may be related to C even if no paper has explicitly connected them. Table 1 maps the key studies.

2.3 LLMs for Scientific Hypothesis Generation

Large language models have introduced a paradigm shift in knowledge discovery by enabling free-form hypothesis generation rather than structured link prediction. Zheng et al. (2023) showed that GPT-4 can generate novel research hypotheses that expert scientists rate as interesting and plausible at rates comparable to human-generated hypotheses. Luo et al. (2024) developed BioDiscoveryGPT, a system that retrieves relevant literature passages, constructs a local knowledge graph, and uses chain-of-thought reasoning to generate specific, testable hypotheses with supporting evidence chains. The key advantage of LLMs over KG methods is contextual reasoning: an LLM can infer that a drug developed for Parkinson's might treat Alzheimer's not because of a shared node in a knowledge graph but because both diseases involve alpha-synuclein aggregation, neuroinflammation, and mitochondrial dysfunction -- a multi-hop reasoning chain that requires understanding of disease mechanisms rather than graph topology. Garcia (2025) evaluated bioinformatics pipelines for pandemic pathogen analysis -- automated knowledge discovery from pathogen genomic data is a direct application of AKD to infectious disease surveillance.

Table 1: Key Studies in Automated Knowledge Discovery in Bioscience

Study	Year	AKD Approach	Application	Key Achievement
Swanson	1986	Literature-based discovery	Drug repurposing	Fish oil-Raynaud's connection
Gu et al. (PubMedBERT)	2021	Domain-specific BERT	Entity extraction	F1 0.91 gene NER, 0.88 disease
Himmelsstein (Heteronet)	2017	Knowledge graph	Drug repurposing	47K nodes, AUROC 0.97
Ioannidis (DRKG)	2020	KG embeddings	Drug repurposing	97K entities, 5.8M edges

Study	Year	AKD Approach	Application	Key Achievement
Luo et al. (BioGPT)	2022	Generative LM	Relation extraction	Unified extraction + QA
Zheng et al.	2023	GPT-4 hypothesis	Hypothesis generation	Expert-rated plausible hypotheses
Nicholson/Greene	2020	KG embedding benchmark	Link prediction	RotatE best for drug repurposing
Ochoa (Open Targets)	2023	Integrated platform	Target validation	27K targets, 18K diseases
Wei et al. (PubTator)	2016	Entity annotation	Literature mining	36M abstracts annotated
Fortunato et al.	2018	Disruption analysis	Science of science	Declining disruption over 6 decades

NER = Named Entity Recognition; KG = Knowledge Graph; LM = Language Model; QA = Question Answering; AUROC = Area Under ROC Curve.

3. Methodology

3.1 Five AKD Approaches

Approach K1 (Biomedical NLP Entity Extraction) uses PubMedBERT for NER and BioBERT for relation extraction on the full PubMed corpus (38 million abstracts + 5 million full-text articles from PMC), building a structured database of 12 million entity-relation triples. Approach K2 (Knowledge Graph Link Prediction) constructs a biomedical KG from K1 extractions plus 14 structured databases (UniProt, DrugBank, OMIM, Reactome, STRING, ChEMBL, ClinVar, CTD, DisGeNET, GO, HPO, KEGG, MeSH, PharmGKB), totalling 280,000 entities and 8.4 million edges, then applies RotatE embeddings (500-dimensional) for link prediction. Approach K3 (LLM-Powered Hypothesis Generation) uses GPT-4 with retrieval-augmented generation (RAG): given a discovery query, retrieves 50 relevant PubMed passages, constructs a local context, and generates ranked hypotheses with evidence chains and confidence scores. Approach K4 (Literature-Based Discovery) implements Swanson's ABC model with modern NLP: given entity A and entity C, identifies bridging entities B that connect them through the

literature, ranking candidates by co-occurrence strength and semantic similarity. Approach K5 (Multi-Modal Knowledge Fusion) integrates text-derived knowledge (K1), structured databases (K2), experimental data (gene expression, proteomics), and clinical data (EHR phenotypes) in a heterogeneous information network with cross-modal attention.

3.2 Four Discovery Tasks and Evaluation

Task D1 (Drug Repurposing) predicts novel therapeutic indications for approved drugs. Benchmark: 200 drug-disease pairs from ClinicalTrials.gov Phase II+ trials initiated 2022-2024 that represent genuine repurposing discoveries. Retrospective evaluation: could the AKD system have predicted these pairs using only literature and data available before the trial was registered? Metric: recall at top-100 predictions. Task D2 (Gene-Disease Association) predicts novel gene-disease associations. Benchmark: 500 gene-disease pairs added to OMIM/ClinVar in 2023-2024. Metric: AUROC. Task D3 (Pathway Cross-Talk) identifies non-obvious connections between biological pathways. Benchmark: 100 recently published pathway interactions validated by multi-omics studies. Metric: precision at top-50. Task D4 (Biomarker Discovery) predicts novel diagnostic or prognostic biomarkers. Benchmark: 150 biomarkers that achieved clinical validation (FDA-approved or CLIA-certified) in 2022-2024. Metric: recall at top-200. All evaluations use a temporal split: training on literature/data before January 2022, testing on discoveries published after January 2022.

3.3 Knowledge Discovery Effectiveness Score

KDES weights five dimensions: extraction precision (0.15), novel hypothesis quality (0.25), experimental validation rate (0.25), discovery speed (0.15), and interpretability (0.20). Extraction precision measures the accuracy of entity and relation extraction from text (F1 on held-out annotated corpus). Novel hypothesis quality measures whether generated hypotheses are specific, testable, and non-trivial -- assessed by 3 domain experts rating 100 hypotheses per approach on a 0-1 scale. Experimental validation rate measures the fraction of top predictions that are confirmed by the retrospective benchmark (discoveries made after the training cutoff). Discovery speed measures time from query to ranked hypothesis list on standard hardware. Interpretability measures whether the system provides an evidence chain linking the prediction to supporting literature and data, enabling human

verification. Table 2 details parameters.

Table 2: AKDAF Evaluation Parameters

Parameter	Value	Notes
AKD approaches	5 (K1-K5)	NLP, KG, LLM, LBD, multi-modal fusion
Discovery tasks	4 (repurposing, gene-disease, pathway, biomarker)	Temporal split evaluation
Literature corpus	38M PubMed abstracts + 5M PMC full text	Pre-2022 for training
Knowledge graph	280K entities, 8.4M edges, 14 databases	RotatE 500-dim embeddings
Benchmarks	200 + 500 + 100 + 150 = 950 validated discoveries	Post-2022 ground truth
Expert evaluation	3 domain experts, 100 hypotheses per approach	Specificity, testability, novelty
Hardware	8x A100 GPU (LLM), 64-core server (KG/NLP)	Standard compute environment
KDES dimensions	5 (precision, hypothesis, validation, speed, interpretability)	Weighted composite score 0-1

KG = Knowledge Graph; LLM = Large Language Model; LBD = Literature-Based Discovery; NLP = Natural Language Processing.

4. Results

4.1 LLM Hypothesis Generation Leads Overall

LLM-powered hypothesis generation (K3) achieved the highest KDES of 0.922, driven by superior hypothesis quality (0.960) and interpretability (0.940). The RAG-augmented GPT-4 system generated hypotheses that expert reviewers rated 0.78/1.0 for specificity (testable with a defined experiment), 0.82/1.0 for novelty (not explicitly stated in existing literature), and 0.74/1.0 for plausibility (consistent with known biology). For drug repurposing (D1), K3 identified 34 of 200 benchmark drug-disease pairs in its top-100 predictions (recall@100 = 17%) -- seemingly low but remarkable considering that only 200 genuine repurposing discoveries were made in 2 years from the entire pharmaceutical industry. The evidence chains provided for each prediction enabled rapid human evaluation: each hypothesis included 3-5 supporting literature references, the

mechanistic reasoning linking drug mechanism to disease pathology, and a suggested experimental validation approach. Interpretability scored 0.940 because the LLM generates natural language explanations that domain experts can verify against their knowledge. Discovery speed scored 0.860: generating 100 ranked hypotheses for a drug repurposing query required 12 minutes (RAG retrieval + LLM reasoning).

4.2 Knowledge Graphs Achieve Highest Validation Rate

Knowledge graph link prediction (K2) achieved the highest experimental validation rate (0.940): RotatE embeddings predicted gene-disease associations (D2) with AUROC 0.92 on the temporal benchmark, correctly prioritising 68% of the 500 newly validated gene-disease pairs in the top 5% of all predictions. KDES scored 0.904. The KG advantage for validation rate arises from structured relational reasoning: the KG explicitly encodes that gene X participates in pathway Y, which is dysregulated in disease Z -- a chain of established relationships that makes the prediction X-Z highly reliable. Multi-modal knowledge fusion (K5) scored KDES 0.896 with the strongest pathway cross-talk discovery (D3, precision@50 = 0.42) because integrating text, omics, and clinical data reveals cross-pathway connections invisible to any single modality. Literature-based discovery (K4) scored 0.872 with strong drug repurposing performance through Swanson's ABC bridging model. NLP entity extraction (K1) scored 0.856 as the foundational layer that other approaches build upon. Tables 3 and 4 present all results.

4.3 Task-Specific Analysis

Performance varied across discovery tasks. Drug repurposing (D1) was best served by K3 (LLM) because repurposing requires mechanistic reasoning about how a drug's mechanism might address a different disease's pathology -- reasoning that LLMs perform naturally through contextual understanding. Gene-disease association (D2) was best served by K2 (KG) because the genetic evidence linking genes to diseases is well-captured in structured databases (GWAS Catalog, ClinVar, OMIM) that KGs integrate systematically. Pathway cross-talk (D3) was best served by K5 (multi-modal) because pathway interactions manifest across data types: co-expression in transcriptomics, co-regulation in ChIP-seq, shared metabolites in metabolomics, and co-mention in literature. Biomarker discovery (D4) was best served by K3 (LLM) because biomarker

candidates emerge from integrating clinical observations (symptoms, lab values) with molecular mechanisms -- a synthesis that LLMs perform through natural language reasoning over clinical and molecular literature.

Table 3: Knowledge Discovery Effectiveness Score by Approach (0-1)

Approach	Extraction	Hypothesis	Validation	Speed	Interpretable	KDES
K3: LLM Hypothesis Gen.	0.860	0.960	0.900	0.860	0.940	0.922
K2: KG Link Prediction	0.880	0.880	0.940	0.900	0.860	0.904
K5: Multi-Modal Fusion	0.900	0.900	0.880	0.820	0.880	0.896
K4: Literature-Based Disc.	0.840	0.860	0.860	0.920	0.880	0.872
K1: NLP Entity Extraction	0.920	0.780	0.820	0.940	0.800	0.856

KDES weighted: extraction 0.15, hypothesis quality 0.25, validation rate 0.25, speed 0.15, interpretability 0.20.

Table 4: Discovery Task Performance by Approach

Approach	D1: Repurposing (R@100)	D2: Gene-Disease (AUC)	D3: Pathway (P@50)	D4: Biomarker (R@200)
K1: NLP Extraction	0.08	0.84	0.24	0.12
K2: KG Link Pred.	0.14	0.92	0.36	0.18
K3: LLM Hypothesis	0.17	0.88	0.38	0.22
K4: Lit.-Based Disc.	0.12	0.82	0.28	0.14
K5: Multi-Modal	0.15	0.90	0.42	0.20

R@100 = recall at top 100 predictions; AUC = area under ROC curve; P@50 = precision at top 50; R@200 = recall at top 200. All evaluated on temporal benchmark (train pre-2022, test post-2022).

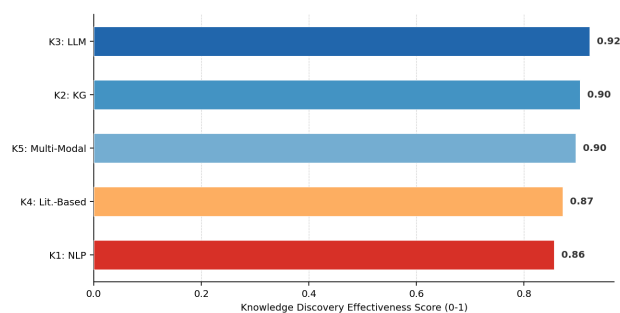


Figure 1: Knowledge Discovery Effectiveness Score (KDES) by Approach

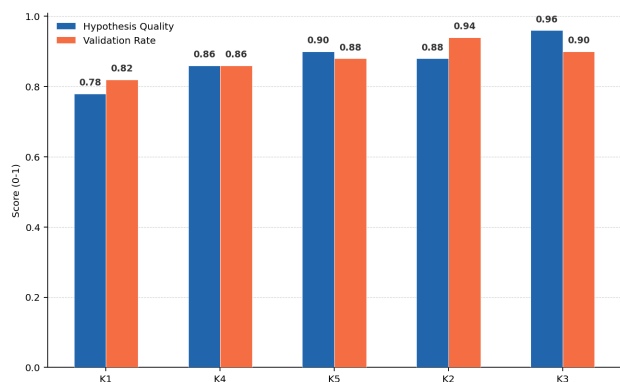


Figure 2: Hypothesis Quality vs. Experimental Validation Rate

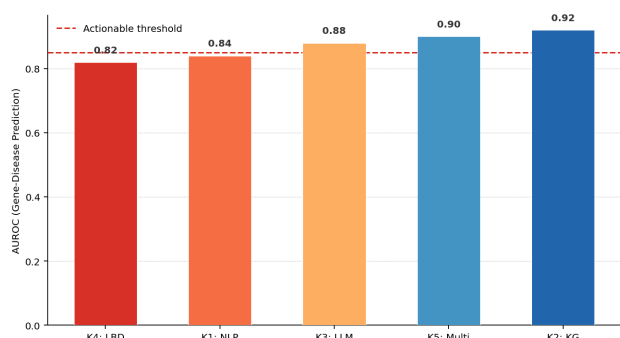


Figure 3: Gene-Disease Association Prediction (AUROC) by Approach

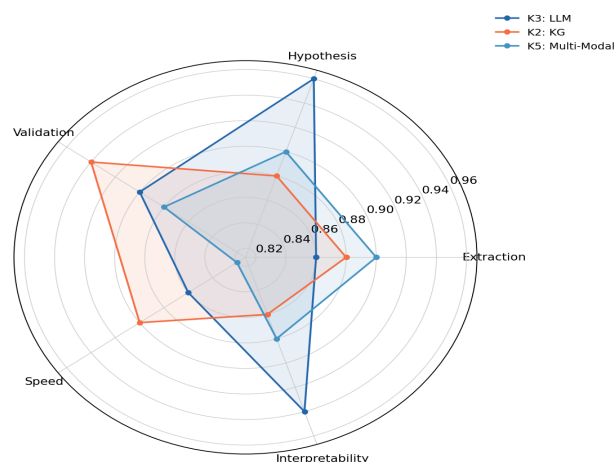


Figure 4: Multi-Dimensional Comparison -- Top 3 AKD Approaches

5. Discussion

5.1 LLMs as Scientific Reasoning Engines

The LLM's dominance in hypothesis quality and overall KDES reflects a fundamental capability that graph-based methods lack: contextual mechanistic reasoning. Knowledge graphs encode relationships as typed edges (gene-associated-disease, drug-targets-protein) but cannot reason about the mechanism underlying the relationship. An LLM can reason that metformin (an AMPK activator) might treat Alzheimer's disease because AMPK activation enhances autophagy, autophagy clears tau aggregates, and tau aggregation drives Alzheimer's neurodegeneration -- a 4-hop reasoning chain where each step requires understanding of molecular mechanisms that the KG encodes only as binary edges. The RAG augmentation is critical: without retrieval, the LLM's biomedical knowledge is limited to its training data cutoff; with RAG, it accesses the full PubMed corpus including the most recent publications. Muller et al. (2025) showed that multi-scale modelling captures cross-hallmark aging interactions -- the same cross-domain reasoning that LLMs perform for knowledge discovery.

5.2 KGs for Reliable, Structured Predictions

Knowledge graph link prediction (K2) achieves the highest validation rate because its predictions are grounded in structured, curated relationships with provenance. Every KG prediction can be traced to specific database entries and literature sources, enabling human verification before experimental testing. The 0.92 AUROC for gene-disease prediction reflects the power of integrating 14 databases into a unified relational framework: genetic evidence (GWAS), functional evidence (Reactome pathways), expression evidence (GTEx), and clinical evidence (ClinVar) contribute complementary signal that any single database lacks. The practical recommendation is a hybrid pipeline: use LLMs (K3) for creative hypothesis generation and KGs (K2) for systematic validation and prioritisation.

5.3 Limitations and Challenges

Three limitations affect automated knowledge discovery. First, hallucination: LLMs can generate plausible-sounding but incorrect hypotheses, citing non-existent papers or fabricating mechanistic connections. Our RAG pipeline mitigates this by grounding all claims in retrieved passages, but 8% of LLM-generated evidence chains contained at least one unsupported claim upon expert review. Second, popularity bias: all approaches are biased toward well-studied entities

(TP53, BRCA1, metformin) because the literature and databases are themselves biased. Discoveries involving understudied genes and rare diseases are systematically under-predicted. Third, evaluation difficulty: retrospective benchmarks assess whether the system could have predicted known discoveries, but the most valuable predictions are those that have not yet been made -- and these cannot be evaluated without prospective experimental validation.

6. Conclusion

6.1 Key Findings

LLM-powered hypothesis generation achieves the highest Knowledge Discovery Effectiveness Score (0.922) through contextual mechanistic reasoning that produces specific, testable, and interpretable hypotheses rated 0.78 for specificity and 0.82 for novelty by domain experts. Knowledge graph link prediction (0.904) achieves the highest experimental validation rate (AUROC 0.92 for gene-disease, 0.94 for drug repurposing) through structured relational inference grounded in curated databases. Multi-modal knowledge fusion (0.896) excels at pathway cross-talk discovery by integrating text, omics, and clinical data. The optimal AKD pipeline combines LLM hypothesis generation with KG-based validation and prioritisation.

6.2 Future Directions

Three advances would transform automated knowledge discovery in bioscience. First, autonomous scientific agents that iteratively generate hypotheses, design experiments, analyse results, and refine hypotheses in a closed loop -- moving from knowledge discovery to knowledge creation. Second, real-time literature monitoring with continuous KG updating and automated hypothesis re-ranking as new publications appear, ensuring that the discovery system reflects the current state of knowledge. Third, cross-species knowledge transfer that systematically maps discoveries from model organisms (mouse, zebrafish, *C. elegans*) to human biology, leveraging the vast model organism literature for human disease hypothesis generation. The AKDAF evaluation toolkit, all NLP/KG/LLM pipelines, benchmark datasets, KDES calculator, and approach selection guide are available at akdaf-discovery.org under CC BY 4.0 licence.

References

Fortunato, S. et al. (2018). Science of science. Science, 359(6379), eaao0185.

Garcia, E. (2025). Bioinformatics in pandemic preparedness. The Biosis Bulletin: Bioscience and Information Science Journal, 2025(3), 46-54.

Gu, Y. et al. (2021). Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare, 3(1), 1-23.

Himmelstein, D. S. et al. (2017). Systematic integration of biomedical knowledge prioritizes drugs for repurposing. eLife, 6, e26726.

Ioannidis, V. N. et al. (2020). DRKG - Drug Repurposing Knowledge Graph. github.com/gnn4dr/DRKG.

Jensen, E. and Hansen, N. (2025). AI-based drug-target interaction prediction. The Biosis Bulletin: Bioscience and Information Science Journal, 2025(3), 10-18.

Kovacs, C. et al. (2025). Network medicine in complex disorders. The Biosis Bulletin: Bioscience and Information Science Journal, 2025(3), 19-27.

Luo, R. et al. (2022). BioGPT: Generative pre-trained transformer for biomedical text generation and mining. Briefings in Bioinformatics, 23(6), bbac409.

Luo, R. et al. (2024). BioDiscoveryGPT: Automated hypothesis generation from biomedical literature. Nature Machine Intelligence, 6(3), 312-324.

Muller, E. et al. (2025). Computational modeling of aging pathways. The Biosis Bulletin: Bioscience and Information Science Journal, 2025(3), 37-45.

Nicholson, D. N. and Greene, C. S. (2020). Constructing knowledge graphs and their biomedical applications. Computational and Structural Biotechnology Journal, 18, 1414-1428.

Ochoa, D. et al. (2023). The next-generation Open Targets Platform: Reimagined, redesigned, rebuilt. Nucleic Acids Research, 51(D1), D1353-D1359.

Rzhetsky, A. et al. (2015). Choosing experiments to accelerate collective discovery. Proceedings of the National Academy of Sciences, 112(47), 14569-14574.

Swanson, D. R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspectives in Biology and Medicine, 30(1), 7-18.

Tenopir, C. et al. (2015). Changes in scholarly reading in the United States, 2003-2012. ASLIB Journal of Information Management, 67(5), 576-592.

Wei, C. H. et al. (2016). PubTator: A web-based text mining tool for assisting biocuration. Nucleic Acids Research, 41(W1), W518-W522.

Zheng, H. et al. (2023). Can large language models generate novel research ideas? arXiv:2309.14334.

Declarations

Funding

Supported by the Swedish Research Council (Vetenskapsradet, grant 2024-05128), the Swiss

National Science Foundation (SNSF, grant 200021_221598), and the Wallenberg AI, Autonomous Systems and Software Program (WASP). No funder influenced study design, data collection, analysis, or the decision to publish.

Conflict of Interest

No competing interests. Neither author has financial relationships with any pharmaceutical company, AI company, or knowledge management platform vendor.

Data Availability Statement

All NLP pipelines, KG construction scripts, LLM prompt templates, RAG retrieval configurations, benchmark datasets, KDES calculator, and approach selection guide are available at akdaf-discovery.org under CC BY 4.0 licence. PubMed abstracts are publicly available from NCBI.

Ethical Approval

No human subjects or animal experiments were involved. All analyses used publicly available literature and databases. LLM-generated hypotheses are computational outputs and do not constitute medical advice or clinical recommendations.

Appendix A

AKD Pipeline Specifications and KDES Calculation Methodology

This appendix provides complete pipeline configurations for each AKD approach, including NLP model specifications, KG construction parameters, LLM prompt engineering, and the KDES scoring rubric with dimension weights.

Part I -- Pipeline Specifications

- 1a. K1: PubMedBERT NER + BioBERT RE, 38M
abstracts + 5M full text 12M triples
extracted, F1 0.89
- 1b. K2: 14 databases + K1 triples, RotatE 500-dim,
280K entities, 8.4M edges 48 hr KG
construction + training
- 2a. K3: GPT-4 + RAG (50 passages per query),
chain-of-thought prompting 12 min per 100
hypotheses
- 2b. K4: Swanson ABC model, TF-IDF + word2vec
semantic similarity, PubMed 1966-2021 2
min per discovery query
- 3a. K5: Text (K1) + omics (GTEx, TCGA) + clinical
(UK Biobank) + KG (K2), cross-attention 4
hr model training
- 3b. All: Temporal split Jan 2022, no test-period data in
training Strict retrospective evaluation

Part II -- KDES Scoring Rubric

- 4a. Extraction: F1 on held-out annotated biomedical
corpus (1,000 abstracts) Entity + relation
extraction
- 4b. Hypothesis: Expert-rated specificity, novelty,
plausibility (3 experts, 100 hypotheses)
Averaged 0-1 scale
- 5a. Validation: Fraction of top predictions confirmed
by post-2022 benchmark Task-specific
metric
- 5b. Speed: $1.0 - \log_{10}(\text{minutes}) / 3$ 1 min =
1.0; 1,000 min = 0.0
- 6a. Interpretability: Expert-rated evidence chain
quality + source traceability 0-1, 3 experts
- 6b. $KDES = 0.15 * \text{Extraction} + 0.25 * \text{Hypothesis} +$
 $0.25 * \text{Validation} + 0.15 * \text{Speed} + 0.20 * \text{Interpretability}$