*Briefings in Bioinformatics*, 2025, **26(1)**, bbaf070
https://doi.org/10.1093/bib/bbaf070
**Problem Solving Protocol**

# A large language model framework for literature-based disease–gene association prediction

Peng-Hsuan Li[1], Yih-Yun Sun[1], Hsueh-Fen Juan[1,2,3,4], Chien-Yu Chen[1,3,4,5], Huai-Kuang Tsai[1,6], Jia-Hsin Huang[1,*]

[1]Taiwan AI Labs, 6F., No. 70, Sec. 1, Chengde Road, Datong Dist., Taipei 10355, Taiwan
[2]Department of Life Science, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei 10617, Taiwan
[3]Center for Computational and Systems Biology, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei 10617, Taiwan
[4]Center for Advanced Computing and Imaging in Biomedicine, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei 10617, Taiwan
[5]Department of Biomechatronics Engineering, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei 10617, Taiwan
[6]Institute of Information Science, Academia Sinica, No. 128, Academia Road, Section 2, Nankang, Taipei 11529, Taiwan

*Corresponding author. Taiwan AI Labs, 6F., No. 70, Sec. 1, Chengde Road, Datong Dist., Taipei 10355, Taiwan. E-mail: jiahsin.huang@gmail.com
Biographical note: This study explores biomedical informatics and artificial intelligence, leveraging large language models and knowledge graphs to advance precision medicine and enhance the discovery of disease–gene relationships.

## Abstract

With the exponential growth of biomedical literature, leveraging Large Language Models (LLMs) for automated medical knowledge understanding has become increasingly critical for advancing precision medicine. However, current approaches face significant challenges in reliability, verifiability, and scalability when extracting complex biological relationships from scientific literature using LLMs. To overcome the obstacles of LLM development in biomedical literature understating, we propose LORE, a novel unsupervised two-stage reading methodology with LLM that models literature as a knowledge graph of verifiable factual statements and, in turn, as semantic embeddings in Euclidean space. LORE captured essential gene pathogenicity information when applied to PubMed abstracts for large-scale understanding of disease–gene relationships. We demonstrated that modeling a latent pathogenic flow in the semantic embedding with supervision from the ClinVar database led to a 90% mean average precision in identifying relevant genes across 2097 diseases. This work provides a scalable and reproducible approach for leveraging LLMs in biomedical literature analysis, offering new opportunities for researchers to identify therapeutic targets efficiently.

**Keywords**: literature mining; biomedical relation extraction; NLP; knowledge graph; large language model

## Introduction

Knowledge curation from scientific literature is fundamental to the advancement of research across disciplines [1–3]. Traditionally, domain-specific knowledge accumulates incrementally and relies heavily on human expert review processes. The biomedical literature, for instance, is a key resource for identifying causal genetic elements associated with diseases and offering insights into clinical practice. Several expert-curated databases, such as ClinVar [4], COSMIC [5], OMIM [6], and PharmGKB [7], provide invaluable assessments of literature evidence. However, such resources are limited in scale because of the broad scope and the rapid expansion of scientific publications [8, 9]. Many computational approaches have been applied to enhance automation in biomedical literature-based discovery for different tasks, such as gene–disease association prediction [10–12], text mining and curation [13–17], and biomedical entity relation extraction [18–21]. However, these methods primarily focus on extracting isolated sentences or paragraphs containing entities of interest rather than synthesizing comprehensive information across multiple sources and contexts. Thus, substantial efforts are required to create task-specific datasets and train models for each literature domain.

On the other hand, Machine Reading Comprehension (MRC) [22], wherein machines answer questions based on textual context, serves as a promising complement to human expertise in reading vast amounts of literature. Recent advancements in natural language processing, particularly the development of Large Language Models (LLMs), have significantly enhanced MRC capabilities to potentially accelerate knowledge synthesis [23–25]. Recent LLMs, such as GPT-4, have demonstrated remarkable capabilities in textual comprehension across diverse domains; however, LLMs face challenges when it comes to reliability and verifiability [23, 26]. Specifically, LLMs are prone to hallucination, a phenomenon whereby plausible but factually incorrect information is generated. Moreover, the opaque parametric memory of LLMs poses a substantial obstacle to traceability—sources of evidence supporting their statements are often unclear. To mitigate these concerns, researchers have applied Retrieval Augmented Generation (RAG) in LLM-based chatbots [27, 28]. RAG restricts the information source of an LLM to an explicit but small set of retrieved texts per user query. However, owing to scalability constraints, fast but shallow sentence similarity-based retrieval is required, leading to incomplete information capture as nuanced content and relevant articles are often missing.

**Received:** November 15, 2024. **Revised:** January 9, 2025. **Accepted:** February 6, 2025

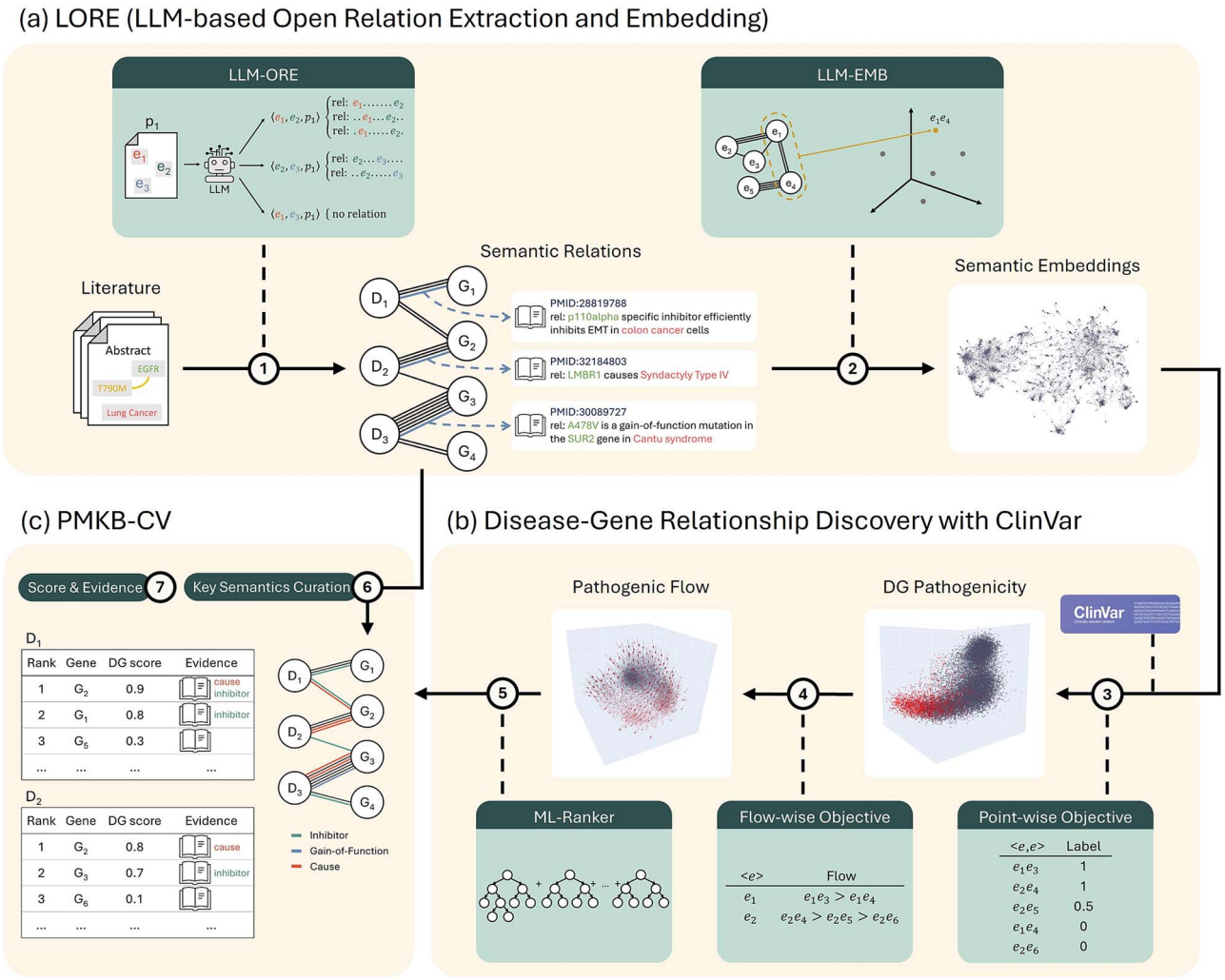



Figure 1. Overview of the literature-semantics framework. Given a large body of literature containing expert domain knowledge, the proposed framework creates a comprehensive unsupervised knowledge graph and numerical embeddings between entities. This enables large-scale supervised modeling of downstream tasks, where model predictions are accompanied by verifiable evidence of relations that can be traced back to the original articles. (a) The framework is applied to PubMed literature, and a knowledge graph containing semantic relations between diseases and genes and their mutations is created. (b) An embedding for disease–gene relationships, where each point in space contains the literature-semantic knowledge of a DG, is created. (c) The embedding is shown to contain a latent structure of DG pathogenicity. (d) We further analyze the disease-wise pathogenic flow and find that it is consistent across diseases and even smoother than the point-wise distribution. (e) An ML-ranker is trained to model the flow and to predict pathogenic genes for each disease, where the prediction scope is 200× larger than expert-curated supervision. (f) We curated 105 key semantics about DG pathogenicity with linguistic lemmas to automatically tag relations. (G) With the proposed framework, we facilitate future research on DG pathogenicity with a literature-scale knowledge base of predicted DG scores and supporting evidence.

In essence, scalable knowledge curation calls for a computational method that is inductive across domains and is capable of capturing nuanced textual context for knowledge synthesis, all while maintaining essential reliability and verifiability. To this end, we introduce LORE (LLM-based Open Relation extraction and Embedding), a novel literature semantics framework that encompasses the best of both worlds and is tested true in capturing disease–gene relationships across the PubMed literature (Fig. 1).

LORE leverages a two-stage reading methodology comprising LLM-based Open Relation Extraction (LLM-ORE) and LLM-based embedding (LLM-EMB) (Fig. 1a). First, LORE employs LLM-ORE to comprehend each article and generate atomic statements about entity relations therein. Curated knowledge is explicitly derived from individual articles, making the generated relations reliable and verifiable. Furthermore, this operation creates a comprehensive unsupervised knowledge graph of the literature, the conciseness of which makes it possible for LORE to then use LLM-EMB to encode the full relation knowledge between each entity pair and create numerical embeddings for downstream task-specific applications. In this work, we applied LORE to curate disease–gene relationship knowledge in the PubMed literature, using disease, gene, and variant annotations from pubmedKB [2].

The ClinVar [4] database is an important annotation repository of the relationships between genes and diseases. Using key disease–gene pairs from ClinVar as references, we evaluated the effectiveness of LORE to explore the latent space of gene-wise pathogenicity and disease-wise pathogenic flow across genes (Fig. 1b). In addition, we constructed machine learning models with the supervision of pathogenic genes from ClinVar to rank the relevance of gene pathogenicity using the semantic embeddings (Fig. 1b).

Finally, we curated a taxonomy of key semantics to use as tags for relations (Fig. 1c). We created PMKB-CV (pubmedKB-ClinVar) dataset, a novel resource that expands the scope of disease–gene

relationship data. Notably, PMKB-CV encompasses more than 2097 diseases and covers disease–gene pairs (DGs) at a scale 200 times larger than that covered by ClinVar. Moreover, PMKB-CV provides rich annotations, including semantic embeddings, predicted DG scores, and verifiable knowledge graph relations with tags and source article IDs (Fig. 1c). In summary, our LORE framework harnesses the power of LLM-based MRC and enables literature-scale knowledge graph construction and downstream modeling. Importantly, PMKB-CV further bridges the gap between large-scale computational analysis and human assessment, helping to advance our understanding of disease genetics and potential therapeutic targets.

## Results

### LLM-ORE curates semantic relations knowledge from literature

We applied LLM-ORE to PubMed abstracts for annotating disease–gene relationships and created a comprehensive unsupervised knowledge graph (Fig. 1a). A total of 11 million relations across 1.7 million abstracts were obtained by prompting GPT-3.5 [29]. Using text-continuation prompts [30], we employed LLMs to analyze individual articles and extract atomic statements describing relations between pairs of entities. Figure 2 illustrates the prompt structure used to guide GPT-3.5 in this task. This prompt is composed of two key sections. The first section demonstrates a domain-agnostic Open Relation Extraction (ORE) [31]. The text serves as a primer, independent of the specific biomedical context, to establish the expected format and level of detail for the extraction (see more details in Methods). Following the same structure as the demonstration, the second section applies the ORE process to the target article and entities under investigation. Notably, this approach allows for a generalized ORE from the literature, which is not constrained to predefined relation types or entity pairs.

A total of 358,888 distinct semantic lemmas are present across the 11 million disease–gene relations. We reviewed 282 high-coverage lemmas in the knowledge graph and curated a taxonomy of 105 key semantics about pathogenicity (Fig. 3). The key semantics, automatically tagged to relations, served as a set of indexes to access the knowledge graph. We grouped the key pathogenicity semantics into four main classes.

Class 1, Relation, includes key semantics that directly describe the relation between a gene and a disease. For example, the key semantic 'mutation cosegregates with disease' in Class 1.3 describes the correlation between occurrence of a certain genetic mutation and whether the genome is from a patient of a certain disease.

Class 2, Mutation, conveys information about genetic mutations. This type of information implicates that the corresponding gene has a role in a certain disease.

Class 3, Disease, conveys information on the genetic aspects of diseases. This semantic implies that a certain gene is at play.

Class 4 contains cohort and miscellaneous information that entails or hints at pathogenicity.

### LLM-EMB embeddings capture underlying gene pathogenicity

In the second stage of LORE, we applied LLM-EMB to the PubMed DG knowledge graph created in the first stage using LLM-ORE. For each pair of entities, all their relations across all articles were encoded to a single vector, which contained the literature knowledge about their relationship. A dense 512-dimensional representation was created for each DG.

To analyze the latent pathogenicity structure within the literature-semantic embedding, we first displayed the embedding space using 2D UMAP [32] (Fig. 4a–d). Each point represented a DG. When points were colored by co-occurrence frequency in PubMed abstracts, high-frequency DGs were observed to be distributed throughout the space (Fig. 4a). When points were colored by ClinVar pathogenicity labels, pathogenic DGs were observed to cluster in a subspace (Fig. 4b). This subspace was well-captured by the pathogenic score prediction of ML-Ranker (See more details about the ML-Ranker in the following sections and in Methods). With an optimal pathogenic score threshold to split the DGs by color into red or gray, the distribution was close to that of the ClinVar labels (Fig. 4d). In contrast to the ClinVar labels, which were human-curated and sparse, the stratified pathogenic score predicted by ML-Ranker delineated a smooth landscape of pathogenicity (Fig. 4c). Thus, high-score DGs not curated yet in ClinVar will be of high interest to the biomedical community.

Next, we displayed the 3D linear subspaces of the LLM-EMB embedding and observed the DG pathogenicity distribution (Fig. 4e–g). The axes were calculated using Principal Component Analysis (PCA) or ridge regression, a regularized version of ordinary least squares regression, against ClinVar pathogenicity labels. For the PCA subspace (Fig. 4e), the top three dimensions explained 12.8% of the variance of the embedding. A latent structure of two manifolds was observed—a dense spherical cluster of gray DG points without ClinVar annotations and a curved pattern of gradual transition from gray to red (i.e., known pathogenic DGs annotated in ClinVar). For the subspaces spanned by a combination of PCA axes and regression axes (Fig. 4f,g), clearer manifolds were observed when more regression axes were included. With two regression axes, a distribution pattern aligned with ClinVar annotations, signifying an association with pathogenicity, was revealed. We noted that the regression axes spanned a smaller subspace of the embedding compared with the PCA axes. Nevertheless, they provided an analytical pathogenicity perspective of the unsupervised embedding.

We further analyzed the literature-semantics structure by visualizing the distribution of important semantic lemmas in the 3D PCA subspace (Fig. 4h–j). The frequency of each semantic lemma, such as 'cause,' was represented in log scale per DG using a cool-blue-to-warm-red color gradient. Distinct patterns emerged for various lemmas, particularly those lemmas with a flatter distribution, such as 'associate,' 'mutation,' and 'cause.' The semantic 'associate' (Fig. 4h) was observed to increase along a linear axis but was more prevalent (indicated by greener colors) in the curved arm than in the sparse parts (bluer). The semantic 'mutation' (Fig. 4i) was primarily distributed along the curved arm. Similarly, the semantic 'cause' (Fig. 4j) was concentrated along the curved arm, with a significant presence at the space correlated with ClinVar annotations. This visualization of semantic lemmas revealed a connection between the intrinsic semantic structure of LLM-EMB and the latent DG point-wise pathogenicity structure.

In short, our analyses revealed the connection between the literature semantics captured by LORE and the known pathogenic DGs in the ClinVar database, as illustrated in Fig. 4. Using both UMAP and PCA visualization techniques, we examined how these known pathogenic DGs are distributed within the semantic LLM-EMB embedding space. Notably, these pathogenic DGs clustered in specific subspaces that aligned with disease-related semantic concepts – particularly terms like 'mutation' and 'cause'. This clustering pattern demonstrates the potential value of LLM-EMB


Figure 2. Annotating articles with LLM-ORE (open relation extraction). Text-continuation prompts are used to make LLM write down its understanding of an article in the form of atomic factual statements. The ORE task is crafted to extract concise relations between entities at the article level. For example, 'Martin dreamed about fishing' requires comprehension and rewriting of several sentences. Also, common unwanted behaviors are avoided by providing examples of bad relations. We applied the task-agnostic prompt to extract an open set of diverse and comprehensive entity relationships for academic literature.

embeddings as robust features for developing our ML-Ranker system for pathogenicity prediction.

## Pathogenic flow in the literature-semantic space

In this section, we explore the embedding space underlying the DG pathogenicity distribution (Fig. 5). First, to model pathogenicity, a straightforward approach would be to model the distribution of pathogenic DGs (i.e., red points) in the embedding space (Fig. 5a,b). Clear subspaces of pathogenicity and nonpathogenicity were evident in the literature-semantic space visualized using 3D linear axes (Fig. 4e–g). These low-rank subspaces can be seen in the gray balls and the ends of the gray-to-red arm. However, the gray and red dots co-occur in some places, such as the center of the arm. To distinguish between gray and red dots, high-dimensional hyperplanes or complex nonlinear subspaces are needed, requiring more data and hampering generalization.

However, the fundamental task is to predict the most relevant pathogenic genes for each disease. If the gray and red dots for each disease are distributed separately to the two ends of a linear axis or, more generally, a smooth curve, pathogenic and nonpathogenic genes can be distinguished by splitting the curve. Furthermore, if the curves are consistent across diseases, the relevant pathogenic genes for every disease can be identified by monotonically raising the predicted relevance along the curves. Under this condition, perfect modeling can be achieved even if nonpathogenic and pathogenic DGs from different diseases are mixed in the embedding space (Fig. 5c).

In this study, we defined pathogenic flow as the direction and magnitude of nonpathogenicity to pathogenicity at each location in space. We started by calculating the disease-wise flow direction at each DG. Then, we quantized the flow direction at each location in space, using flow magnitude to reflect consistency. We observed a smooth, cross-disease consistent field of pathogenic flow in the literature-semantic embedding (Fig. 5d). A consistent flow of nonpathogenicity to pathogenicity was noted along the arm-shaped manifold of the gray-to-red transition, implying that although DG point-wise modeling was difficult in the central mixture of the arm, disease-wise modeling was smoother and much more linear.

| Semantic class | Curated key semantics | LLM-ORE annotation |
|---|---|---|
| Class 1 Relation | 1.1 Descriptions of "a mutation causes a disease" | |
| | cause | PMID:32184803, LMBR1 causes Syndactyly Type IV |
| | driver gene/mutation | PMID:30854234, L858R is an oncogenic driver in non-small cell lung cancer (NSCLC) |
| | 1.2 Descriptions of "a mutation is known for a disease" | |
| | mutation encountered in disease | PMID:35989815, TP53 mutations are frequently encountered in Li-Fraumeni Syndrome patients |
| | mutation previously reported | PMID:15317752, R208X was previously identified in patients with infantile neuronal ceroid lipofuscinosis |
| | 1.3 Descriptions of "a mutation correlates to a disease" | |
| | mutation cosegregates with disease | PMID:9851430, Tyr329Ser mutation in GBE gene cosegregates with Adult polyglucosan body disease |
| | predisposition to disease | PMID:14644139, NOD2 is involved in the predisposition to Blau syndrome |
| | 1.4 Descriptions that implicate relation | |
| | molecular basis of disease | PMID:10909845, N21I is part of the molecular basis of hereditary pancreatitis |
| | precise diagnosis | PMID:9268242, G to T transversion at position 7314 can help in the precise diagnosis of FAP |
| | 1.5 Other genetics-disease descriptions | |
| | haploinsufficiency | PMID:16429401, SALL1 haploinsufficiency results in Townes-Brocks syndrome |
| | autosomal dominant | PMID:23954459, NF1 is inherited in an autosomal dominant pattern in spinal neurofibromatosis |
| Class 2 Mutation | 2.1 Various mutation types | |
| | nonsense mutation | PMID:31162149, UBE3B has a nonsense variant (c.518C > A, p.Ser173Ter) in Kaufman oculocerebrofacial syndrome |
| | gain-of-function mutation | PMID:30089727, A478V is a gain-of-function mutation in the SUR2 gene in Cantu syndrome |
| | 2.2 Other commonsense mutation types | |
| | novel mutation | PMID:36320120, c.528dupT is a novel variant in cardiospondylocarpofacial syndrome |
| | unreported mutation | PMID:18553553, IVS12-1G > A is a previously unreported mutation in Johanson-Blizzard syndrome |
| | 2.3 Paraphrases of "mutation" | |
| | mutant | PMID:19289107, S773P is an AE1 mutant associated with renal tubular acidosis |
| | sequence variant | PMID:32812400, de novo COL4A5 sequence variants are detected in X-linked Alport syndrome |
| | 2.4 Other mentions of mutation | |
| | mutational hot spot | PMID:26056022, R767W is a hot spot in autosomal dominant osteopetrosis |
| | gene harbours mutation | PMID:7920660, L1CAM harbours mutations leading to X-linked hydrocephalus |
| | 2.5 Various defect types | |
| | protein folding defect | PMID:12655546, P244L is a folding defect in Phenylketonuria |
| | missing protein | PMID:22749141, CSB missing or defective in cells of patients with Cockayne syndrome |
| Class 3 Disease | 3.1 Descriptions of genetic disorder | |
| | congenital | PMID:9589634, S281N is associated with congenital nonautoimmune hyperthyroidism |
| | spontaneous | PMID:11380448, prothrombin mutation risk for spontaneous recurrent venous thromboembolism |
| | 3.2 Disease outcome severity | |
| | fatal outcome | PMID:23095120, c.244dup1 causes an unusual and very early fatal outcome in newborns with MCAD deficiency |
| | milder phenotype | PMID:25043250, c.300delC leads to a milder form of PMD |
| | 3.3 Other disease descriptions | |
| | progression-free survival | PMID:33283711, FLT3 mutation affects PFS of AML patients |
| | atypical phenotype | PMID:16729790, p.R544G is a mutation found in atypical XLA |
| Class 4 Other | 4.1 Regarding cohort | |
| | founder effect | PMID:15140538, S143P has a founder effect in Finnish DCM patients |
| | proband | PMID:11536076, SMAD4 described in probands with JPS |
| | 4.2 Miscellaneous | |
| | inhibitor | PMID:28819788, p110alpha specific inhibitor efficiently inhibits EMT in colon cancer cells |
| | mosaicism | PMID:10923037, OCRL1 somatic and germinal mosaicisms common in Lowe syndrome |

Figure 3. Curated key semantics. We created a taxonomy of four main classes, 15 subclasses, and 105 key semantics about disease–gene pathogenicity. With their corresponding linguistic lemmas, tags are added to relations automatically. Here, two key semantics and sample relations are shown for every subclass.

## Literature-scale pathogenicity prediction by ML-ranker

We built an ML-Ranker that can model disease-wise pathogenic flow and score DGs that co-occur in PubMed abstracts. We applied the lambda objective with Gradient-Boosted Decision Trees (GBDT) [33] to directly model disease-wise pathogenic flow.

We compiled the PMKB-CV dataset to validate the proposed approach (Fig. 6a). PMKB-CV contains 2097 diseases that are present in both pubmedKB and ClinVar. For these diseases, 652 701 DGs co-occurred in PubMed abstracts, whereas only 3004 DGs were human-curated by ClinVar. Paper abstract co-occurrence covered 94.8% of the 3004 known pathogenic DGs, whereas LLM-ORE relations covered 71.4% (Fig. 6b). We included pubmedKB annotations, such as the number of paper abstracts in which a DG co-occurs in the dataset, and also used them as model features. Moreover, we used zero embedding for DGs without relations. Consequently, all DGs in the dataset can be uniformly used for training and prediction.

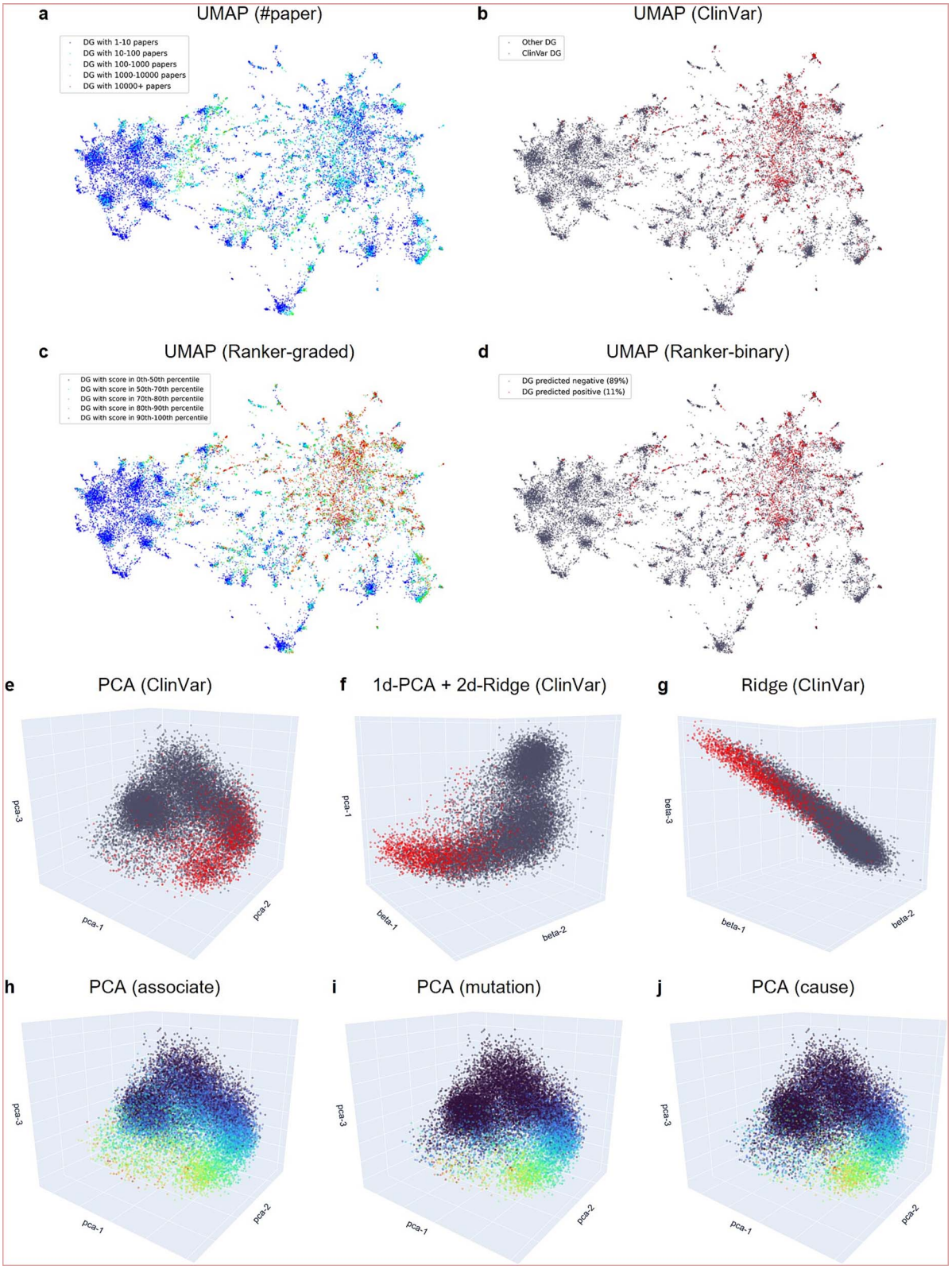



Figure 4. LLM-EMB literature-semantic embedding visualization. (a–d) Visualization with UMAP to analyze the latent pathogenicity structure within the literature-semantic embedding. Points are colored by the number of papers (#paper) (a), ClinVar pathogenicity labels (b), graded ML-ranker prediction (c), and binary ML-ranker prediction (d). The sparse ClinVar-curated red pathogenic DGs are seen clustering toward a subspace, captured well by ML-ranker, which also provides a smooth landscape of graded predictions for uncurated DGs. (e–g) Visualization with linear axes calculated using PCA (e), ridge (g), and their combination (f). A point is colored red if it is a known pathogenic DG in ClinVar, and DGs with unknown pathogenicity are colored gray. A latent structure of two manifolds—a gray ball of nonpathogenicity and a curved arm of transition from nonpathogenicity to pathogenicity—resides in the semantic space. (h–j) PCA visualization colored by distributions of literature semantics 'associate' (h), 'mutation' (i), and 'cause' (j). The connection between the smooth semantics distribution and the sparsely-curated ClinVar pathogenicity distribution can be seen.

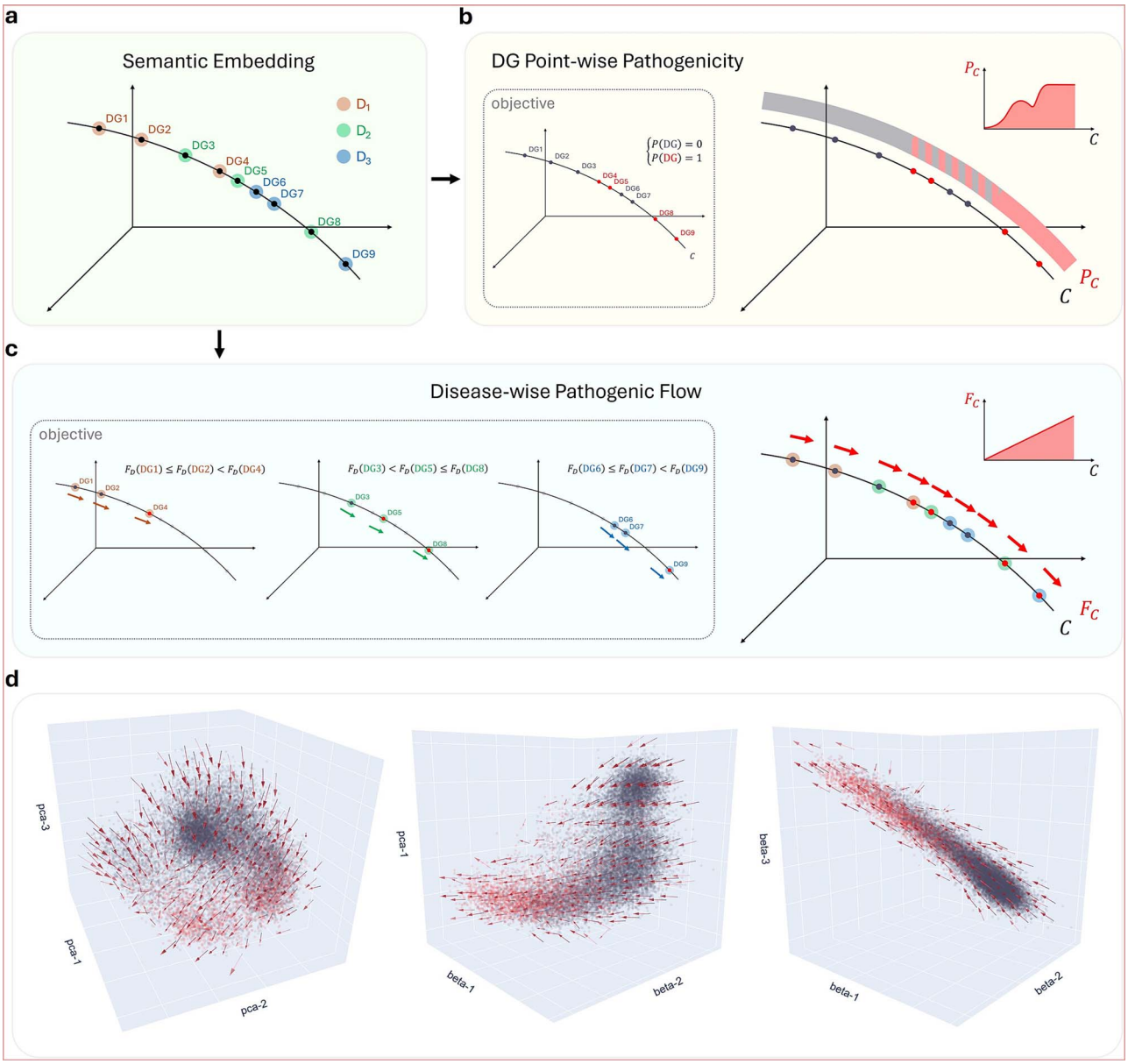



Figure 5. Pathogenic flow in the literature-semantic space. (a) Suppose in the semantic embedding, nine DGs from three different diseases reside on a curve. (b) For the DG point-wise objective, the disease group information is not used, and absolute zero–one labels are the prediction target. As a result, a non-linear function along the curve must be learned. (c) For the disease-wise flow objective, DGs are grouped by disease, and relative flow directions are the prediction target. Because of the cross-disease consistency of the flow, a linear function along the curve will be learned to rank DGs for every disease perfectly. (d) Visualization of the actual pathogenic flow. A smooth, cross-disease consistent field of pathogenic flow is seen residing in the literature-semantic space.

We used leave-one-disease-out cross-validation to evaluate the performance of ML-Ranker for predicting pathogenic genes for each disease iteratively. An average precision (AP) score was calculated for each disease and Mean Average Precision (MAP) was used to evaluate the overall performance of the ranker. As a baseline, we directly asked GPT-3.5 (ver. 2023-06-13) and GPT-4o (ver. 2024-05-13) about the pathogenicity of each DG. In addition, to put into perspective the effectiveness of the curated key semantics in identifying crucial literature evidence, we tested a DG pathogenicity prediction method of simply counting the number of tagged relations per DG. Of note, the curated key semantics were only used in the experiment 'LLM-ORE (key semantics)' (Fig. 6b).

For all DGs in PMKB-CV, ML-Ranker achieved an MAP of 79.9%, which was a significant enhancement over the 69.4% MAP of co-occurrence paper counting and the 31.7% MAP of GPT-4o (Fig. 6b). Similar performance (MAP = 79.2%) was observed when applying 5-fold cross-validation of disjoint genes subsets (Supplementary Fig. 1).

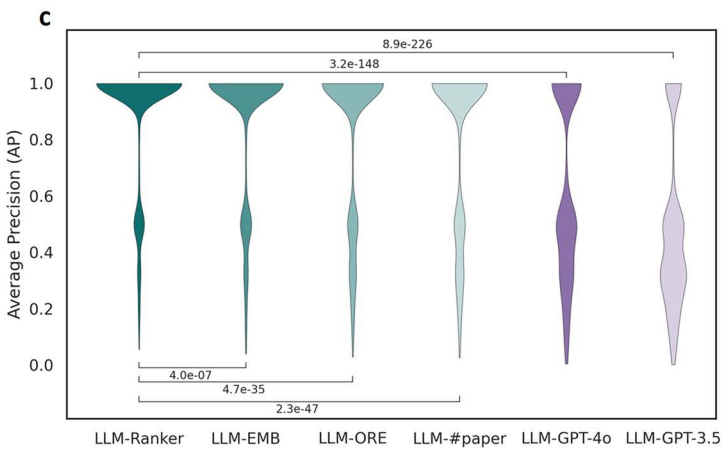
When focusing specifically on those DGs with LLM-ORE annotations, ML-Ranker yielded a remarkable MAP of 90.0% (Fig. 6b). The predictive performance (AP) of ML-Ranker was statistically significantly better than other methods including co-occurrence paper counting (LLM-#paper), literature-semantic relation counting (LLM-ORE), and latent pathogenic flow modeling (LLM-EMB) (Fig. 6c). For the highly prevalent diseases that co-occurred with the highest number of genes in pubmedKB, ML-Ranker achieved a robust MAP of 81%, compared with the 67% MAP of co-occurrence paper counting (Fig. 6d).

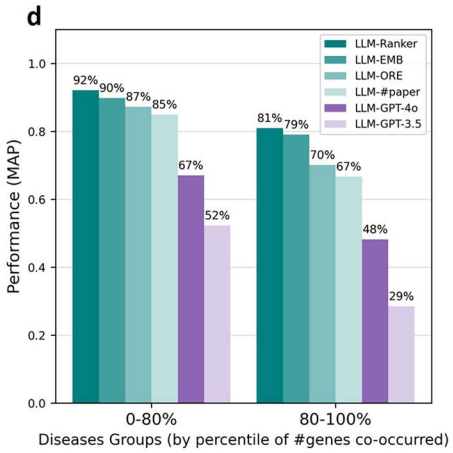
**a**

| PMKB-CV | Distinct diseases | Distinct genes | DGs |
|---|---|---|---|
| PMKB ∩ CV (DG set intersection) | 2,097 | 1,824 | 2,848 |
| PMKB | 2,097 | 17,750 | 652,701 (229x prediction scope) |
| CV | 2,097 | 1,872 | 3,004 (94.8% covered) |

**b**

| Method | Performance (MAP) | ClinVar Coverage |
|---|---|---|
| All-GPT-3.5 | 15.9% | **94.8%** |
| All-GPT-4o | 31.7% | **94.8%** |
| LLM-GPT-3.5 | 47.8% | 71.4% |
| LLM-GPT-4o | 63.5% | 71.4% |
| All-#paper | 69.4% | **94.8%** |
| All-Ranker | 79.9% | **94.8%** |
| LLM-#paper | 81.5% | 71.4% |
| LLM-ORE | 84.0% | 71.4% |
| LLM-ORE (key semantics) | 87.7% | 70.2% |
| LLM-EMB | 87.7% | 71.4% |
| LLM-Ranker | **90.0%** | 71.4% |




**e**

| Method | LLM-ORE Model | LLM-ORE Papers | Performance (MAP) | ClinVar DGs with relations | ClinVar DGs Predicted | #relations |
|---|---|---|---|---|---|---|
| All-Ranker | Llama-8B | 3,997,496 | **81.6%** | **91.3%** | 94.8% | 74,132,940 |
| | Llama-8B | 1,745,538 | 79.5% | 70.4% | 94.8% | 10,485,288 |
| | GPT-3.5 | 1,745,538 | 79.9% | 71.4% | 94.8% | 11,285,095 |

**f**

| Disease | Gene | ClinVar | DG Score | Evidence |
|---|---|---|---|---|
| (MESH:D005879) Tourette Syndrome | (NCBI:114798) SLITRK1 | Y | (1) 100.0% | key: 43, other: 96 |
| (MESH:D005879) Tourette Syndrome | (NCBI:3067) HDC | Y | (2) 80.8% | key: 27, other: 30 |
| (MESH:D005879) Tourette Syndrome | (NCBI:83943) IMMP2L | N | (3) 76.4% | key: 14, other: 18 |
| cause | PMID:17043892, IMMP2L possible rare cause of Gilles de la Tourette syndrome | | | |
| deletion mutation | PMID:24549057, IMMP2L has intragenic deletions in patients with Tourette syndrome | | | |
| phenotype | PMID:27932244, IMMP2L dysfunction contributes to the neurodevelopmental phenotype of Tourette syndrome | | | |
| predisposition to disease | PMID:15103711, IMMP2L shows a tendency for association with Tourette syndrome predisposition | | | |
| mutation previously reported | PMID:21386874, IMMP2L has previously been proposed as a candidate gene for Tourette syndrome | | | |
| ... | ... | | | |
| (MESH:D005879) Tourette Syndrome | (NCBI:114781) BTBD9 | N | (4) 74.1% | key: 7, other: 42 |
| (MESH:D005879) Tourette Syndrome | (NCBI:3356) HTR2A | N | (5) 72.9% | key: 17, other: 56 |
| (MESH:D005879) Tourette Syndrome | (NCBI:13) AADAC | N | (6) 66.4% | key: 4, other: 15 |
| (MESH:D005879) Tourette Syndrome | (NCBI:1268) CNR1 | N | (7) 66.0% | key: 6, other: 43 |
| ... | ... | ... | ... | ... |

Figure 6. PMKB-CV dataset and ranking performance. (a) Statistics of the PMKB-CV dataset. For 2097 diseases, the literature semantics framework has a 200× prediction scope against curated DGs. (b) Mean average precision (MAP) of the ranking performance and the ClinVar coverage of different methods. GPT-4o does not perform well. In comparison, LLM-EMB linear regression alone achieves 87.7% performance, whereas the full-fledged ML-ranker provides higher performance at 90.0% or higher coverage at 94.8%. (c) Dispersion of ranking performance across diseases for each method and the p-values of distribution differences by one-sided Wilcoxon signed-rank test. LLM-ranker significantly outperformed baseline methods. (d) Performance across diseases of different scopes. The naive paper counting method encountered difficulties while ranking DGs for diseases that co-occurred with many genes in PMKB, but our semantic embedding approach remained robust. (e) Extending the scope of LORE using the public llama-8B model. Replacing GPT-3.5 with llama-8B resulted in marginal performance loss. The smaller model was further applied to all 3.9 million papers with DG co-occurrence. The final extracted relations covered 91.3% of ClinVar DGs and achieved 81.6% ranking performance. (f) Top-ranked DGs, seven out of 586 in PMKB, for Tourette syndrome accompanied by their literature relation evidence.

Furthermore, we extended LORE to use the open-source Llama-8B model (ver. Llama-3.1-8B-Instruct). Using a much smaller and accessible model, ML-Ranker achieved a comparable 79.5% MAP to the 79.9% MAP of using GPT-3.5. Leveraging Llama-8B, we processed all 3,997,496 pubmedKB abstracts with DG co-occurrence and curated 74,132,940 relations. The resulting relations covered 91.3% of ClinVar DGs, enabling ML-Ranker to achieve a notable ranking performance of 81.6% MAP (Fig. 6e).

Finally, the DG scores and ranking provided by ML-Ranker are accompanied by literature evidence (Fig. 6f). Our approach facilitates future expert assessment of DG pathogenicity by a quick grasp of literature knowledge with key semantics relations and relevant articles.

## Discussion

Recent advancements in LLMs aim to automate complex sense-making as human endeavors in reading and connecting information across large collections of scientific literature [34]. Our study introduces LORE, a novel literature semantics framework that fundamentally reframes how we leverage LLMs to extract and use knowledge from scientific literature.

The LORE framework offers several key advantages. First, knowledge synthesis using the LORE approach constructs a literature knowledge graph of verifiable factual statements linked to the sources. Second, LORE offers a scalable framework for knowledge synthesis from large amounts of article texts. LORE extracts original article texts and transforms them into a concise knowledge graph. The approach is more efficient than traditional retrieval augmented generation approaches that select only a small set of articles for an LLM to read. The knowledge graph is much more concise compared with the original articles. This reduction in size and complexity allows for a more efficient representation of information; all the relevant knowledge can then be embedded for downstream tasks. In addition, LORE allows new publications to be annotated, thereby continually expanding the knowledge graph. Third, this approach places much less demand on the capability of LLMs, compared with directly asking LLMs expert domain questions. Using LORE, we have captured gene pathogenicity with GPT-3.5 and Llama-8B (Fig. 6e), a feat far from being achieved by directly asking GPT-4o (Fig. 6b). Indeed, small and open-source LLMs have been demonstrated to be competent for article-level comprehension [35, 36], hence the methodology is not constrained to enterprise LLMs. Finally, our framework demonstrates remarkable efficacy in capturing disease–gene relationships through unsupervised relationship extraction and embedding, and users can also employ LORE with prompt engineering and fine-tuning to annotate task-specific knowledge across various domains of scientific inquiry [37, 38].

When applying LORE to the complex landscape of DG relationships, we demonstrated the presence of a latent smooth field of cross-disease consistent pathogenic flow in the unsupervised literature-semantic embeddings. This discovery reveals that although pathogenic and nonpathogenic DGs from different diseases may occupy similar locations in the embedding space, a consistent directional flow of pathogenicity exists in terms of semantics. To illustrate, consider a simplified one-dimensional embedding axis where rare and common diseases coexist (Supplementary Fig. 2). Suppose that D1 is a rare disease reported in a few studies and that D2 is a common disease whose association with many genes is discussed in a multitude of papers; in this scenario, the following literature annotations, DG locations, and pathogenicity labels are possible:

*G1 is not related to D1.* ($x = 0$, *y = non-pathogenic*).

*G2 mutation is found in a D1 patient.* ($x = 1$, *y = pathogenic*).

*G3 mutation is found in a D2 patient; G3 is associated with D2.* ($x = 2$, *y = non-pathogenic*).

*G4 mutation is found in a D2 patient; G4 causes D2.* ($x = 3$, *y = pathogenic*).

Although the pathogenic D1G2 and the non-pathogenic D2G3 are mixed in the center of the axis, literature evidence about pathogenicity consistently increases along the axis. As a result, even when pathogenic and non-pathogenic associations are interspersed due to inter-disease differences such as popularity, the literature-semantic axis provides for a cross-disease consistent linear flow, enabling accurate pathogenicity modeling across diseases.

Our initial analyses of this pathogenic flow revealed clusters of disease-specific pathogenic curves. Notably, we found that these clusters often form a continuum, with the endpoint of one cluster serving as the starting point for another. This observation suggests a broader, interconnected field of pathogenic relationships across diseases, offering new perspectives on the complex landscape of genetic pathogenicity.

LORE curates knowledge for entity pairs that co-occur in literature articles. For the study of disease–gene pathogenicity, we noted that the potential curation scope was larger than the PMKB-CV dataset. PMKB-CV contained 2097 diseases that had both pubmedKB DG co-occurrence and known ClinVar pathogenic DGs (Fig. 6a); the full pubmedKB contained 3 128 402 DGs co-occurred in abstracts, spanning 8894 diseases (Supplementary Fig. 3). In this study, we focused on those 2097 diseases that could be validated by ClinVar, but the potential curation scope was as large as the 3 128 402 DGs. On the other hand, we also noted that the full ClinVar contained 4311 known pathogenic DGs, 1307 of which had no pubmedKB abstract co-occurrence. This was the inherent limitation to article abstract-based MRC.

## Conclusion

In summary, our study makes three significant contributions to the field. First, it presents a novel literature semantics framework that addresses the long-standing challenges of comprehensiveness, reliability, and verifiability in machine reading comprehension. Second, it demonstrates the efficacy of LORE in capturing complex pathogenic relationships across diseases to reveal new insights into pathogenic flow. Finally, it provides a literature-scale dataset that not only complements existing resources such as ClinVar but also offers a knowledge graph of DG relationships with graded pathogenicity scores for genetic prioritization in clinical practice. The methodology of LORE is a general improvement on LLM-based machine reading comprehension, paving the way for bridging vast literature resources and actionable scientific knowledge to realize accelerated discoveries across scientific disciplines.

## Methods

### Two-stage reading comprehension of LORE

In the first stage of LORE, we applied LLM-ORE by prompting LLMs. Figure 2 shows the actual prompt we used. The demonstration consists of an article with a topic as Martin Likes Fish, target entity pair ‘Martin’ and ‘fish’, along with lists of unwanted and desired annotations. This section implicitly specifies the ORE task and the following required properties.

(1) *The annotations should be concrete statements of fact implied by the article.*

(2) *The annotations should follow a structured format, specifically a list of relational triplets.*

(3) *Each relational triplet should be* < *'subject', 'predicate', 'object'* > .

(4) *The subject should contain the first entity, and the first entity should only appear in the subject.*

(5) *The object should contain the second entity, and the second entity should only appear in the object.*

(6) *The predicate should be a concise, self-contained description of the relation between the subject and the object.*

(7) *Abstract understanding of the article is allowed beyond sentence-level syntax.*

The approach allows for abstract understanding beyond sentence-level syntax. The requirements are better understood through demonstration rather than explicit definitions. For instance, instead of explaining the terms 'subject,' 'predicate,' and 'object,' the examples and counterexamples in the demonstration make the concept clear. In addition, the examples illustrate behaviors that are easier to grasp instinctively rather than by complex rules. Examples include the removal of 'also' from 'Martin also loves eating fish', the rewriting of 'large fish scare him' to 'scare of', and the digested understanding of 'Martin dreamed about fishing' from 'he has a dream. The dream is about fishing'. Finally, we note that the ending '-' is important in ensuring that LLM follows the desired list format.

Formally, the desired generation

$$g = f_{\text{LLM-ORE}}(p, e_1, e_2; m)$$

where $p$ is the target article, $e_1$ and $e_2$ are the names of the target entities, and $m$ is the LLM model. In this work, we use the paper title and abstract as $p$. For $m$, gpt-3.5-turbo-0613 is used. The function maps $< p, e_1, e_2 >$ to $g$, the list of parsed relational triplets, using the text-continuation prompt shown in Fig. 2.

In the second stage of LORE, LLM reads all relations between an entity pair and produces a numerical representation of knowledge about their relationships. For example, the following relations between the entity $e_1$ and the entity $e_2$ are read by LLM as the following document.

*'e1', 'causes', 'e2'.*
*'e1 mutations', 'are frequently encountered in', 'e2 patients'.*
*'e1 haploinsufficiency', 'results in', 'e2'.*

*Document: 'e1 causes e2. e1 mutations are frequently encountered in e2 patients. e1 haploinsufficiency results in e2.'.*

If the document is larger than the allowed context of an LLM, it is split into multiple sub-documents, each of which contains as many relations as possible.

Formally, the embedding vector of an entity pair is given by

$$v = f_{\text{LLM-EMB}}(D_{e_1,e_2}; m) = \frac{\sum_{d \in D_{e_1,e_2}} emb(d; m) \times |\, d \,|}{\sum_{d \in D_{e_1,e_2}} |\, d \,|}$$

where $e_1$ and $e_2$ are the target entities, $D_{e_1,e_2}$ is the set of all sub-documents, usually just one full document, containing the relations between $e_1$ and $e_2$, and $m$ is the LLM model. In this work, we employed the text-embedding-3-large from OpenAI for $m$, and the length of each sub-document $|\, d \,|$ is its number of tokens according to $m$.

## Modeling the pathogenic flow with ML-ranker

To visualize the disease-wise pathogenic flow, we define the flow as a unit vector at each DG that points to the pathogenic direction of disease D at the embedding location of DG. Formally, the flow vector is given by

$$u'_{DG} = \begin{cases} v_{DG} - \frac{1}{|S_D^0|} \sum_{DG' \in S_D^0} v_{DG'} \text{ if } P(DG) = 1 \\ \frac{1}{|S_D^1|} \sum_{DG' \in S_D^1} v_{DG'} - v_{DG} \text{ if } P(DG) = 0 \end{cases}$$

$$u_{DG} = \frac{u'_{DG}}{|\, u'_{DG} \,|}$$

where $v$ denotes the embedding vector by LLM-EMB, $S_D^0$ and $S_D^1$ denote the sets of non-pathogenic and pathogenic disease–gene pairs of $D$ respectively, and $P$ maps non-pathogenic and pathogenic pairs to 0 and 1, respectively. In other words, each non-pathogenic DG has a unit flow vector directed at the average embedding of the pathogenic DGs of the same disease, and each pathogenic DG has a unit flow vector directed from the average embedding of the non-pathogenic DGs of the same disease. Then, we quantize the flow vectors for each cube in space by averaging them. Formally, a quantized flow vector is given by

$$u_L = \frac{1}{|\, L \,|} \sum_{DG \in L} u_{DG}$$

where $L$ is the set of DGs in a cube subspace. Finally, we show the field of pathogenic flow in Fig. 5d, where each arrow corresponds to a quantized flow vector. The arrow direction is the aggregated disease-wise flow direction from non-pathogenicity to pathogenicity, and the arrow length, proportional to $|\, u_L \,|$, reflects the degree of cross-disease consistency at that location.

As shown in Fig. 5c, the essence of modeling the pathogenic flow is to model the difference between non-pathogenic and pathogenic genes per disease. Specifically, suppose $DG1$ and $DG2$ correspond to the same disease but $P(DG1) = 1$ and $P(DG2) = 0$. Let $s$ denote the predicted pathogenicity score. Then one would want to maximize the probability $\mathbf{Pr}(P(DG1) > P(DG2))$ by minimizing the following RankNet [39] loss.

$$L_{DG1,DG2} = -\log \mathbf{Pr}(P(DG1) > P(DG2)) = \log\left(1 + e^{-\sigma(s_{DG1} - s_{DG2})}\right)$$

To make sure the most relevant genes are ranked on top for each disease, the final gradient, $\lambda$, is the gradient of the RankNet loss multiplied by the change in Normalized Discounted Cumulative Gain (NDCG) [40]. Formally, this LambdaRank [41] gradient is given by

$$\lambda_{DG1,DG2} = \frac{\partial L}{\partial s_{DG1}} \cdot \Delta \mathbf{NDCG}(DG1, DG2)$$

In this work, we use **λGBDT**, the lambda objective combined with GBDT [33], as the ML-Ranker to explicitly model the pathogenic flow in the literature-semantic space.

To evaluate the ranking performance for each disease, AP is used to see if the known pathogenic genes for a disease are ranked on top. Formally, for each disease,

$$AP = \frac{1}{R} \sum_{k=1}^{N} \mathbf{known}(k) \times \mathbf{precision}(k)$$

where $N$ is the number of ranked genes for that disease, and $R$ is the number of ranked known pathogenic genes for that disease.

$\mathbf{known}(k) = 1$ if the gene ranked at $k$ is a known pathogenic gene for that disease; otherwise $\mathbf{known}(k) = 0$. $\mathbf{precision}(k)$ is the percentage of known pathogenic genes among genes ranked top-$k$ for that disease.

## Key semantics curation

The curation process consists of three main steps: linguistic lemma extraction, important lemma identification, and manual taxonomy construction. In the first step, the sentential relations extracted by LLM-ORE are tokenized to bags of words, and word inflections are lemmatized to dictionary form. For example, *cause*, *causes*, *caused*, and *causing* all correspond to the same linguistic lemma. At this stage, entity names are also filtered out.

In the second step, important lemmas are identified using a coverage filter and a precision filter. The coverage filter demands a lemma to appear in LLM-ORE relations of at least $n$ DGs. The precision filter requires that a certain proportion $r$ of all the relations involving a lemma be from known pathogenic DGs, as indicated by ClinVar. The parameters can be adjusted according to the desired scope of key semantics. In this work, we selected a coverage parameter $n = 100$ and a precision parameter $r = 50\%$ and resulted in 282 important lemmas.

The final step involves manually curating a taxonomy of 105 key semantics by examining the LLM-ORE relations associated with these important lemmas During curation, we sampled 10 relations from known pathogenic DGs and 10 relations from other DGs to inspect for each lemma. The resulting key semantics were then used to tag all relevant relations in the respective lemmas. As a result, the LLM-ORE knowledge graph contains sentential relations linked to the semantic taxonomy of pathogenicity.

Furthermore, we experimented with a DG pathogenicity prediction method where the number of tagged relations for each DG is directly used as its pathogenic score. We note that the curation process has used the ClinVar information, so the generalizability of the ranking performance of this method is not directly comparable to other methods. Nevertheless, we put the effectiveness of the curation method into perspective, showing what performance and scope DG pathogenicity researchers can expect when using the key semantics tags to grasp the literature knowledge and identify relevant relations and articles for their DGs of interest.

## Constructing the PMKB-CV dataset

We constructed the PMKB-CV dataset as a large-scale complement of ClinVar and an evaluation benchmark for our proposed methodology. The scope of the dataset is defined using disease IDs from Medical Subject Headings (MeSH), a vocabulary thesaurus maintained by the Nation Library of Medicine (NLM), and *Homo sapiens* protein-coding gene IDs from the National Center for Biotechnology Information (NCBI).

The PMKB-CV dataset comprises two main components including literature-based and expert database parts. For the literature part, we utilized the annotations from pubmedKB [2]. The gene and variant mentions are both indexed by NCBI gene IDs, and we considered the occurrence of either a gene or its variant as an occurrence. This approach yielded 3,128,402 DG pairs with co-occurrence within abstracts, encompassing 8894 diseases and 18,393 genes. In addition, we extracted a subset of PubMed articles as the most relevant literature for the study of disease–gene relationships via a bootstrapping iteration with pubmedKB annotations as features. Using leave-one-disease-out training, we predicted a bootstrap score for each DG. Then, the literature subset is formed by the articles associated with the top three genes for each disease, the top three diseases for each gene, and the top 15 K DGs. The resulting subset contains 1,745,538 articles, for which we applied LORE and curated 11,285,095 relations.

The expert database part was derived from ClinVar [4], which provides gene-disease relationship data. As ClinVar uses OMIM (Online Mendelian Inheritance in Man) [6] numbers for disease indexing, we applied UMLS (Unified Medical Language System) [42], a vocabulary alignment dataset maintained by NLM, to map OMIM numbers to MeSH IDs. This process resulted in 4311 known pathogenic DGs, spanning 3175 distinct diseases and 2416 distinct genes.

The final PMKB-CV dataset was created by including those diseases that have both pubmedKB co-occurrence DGs and ClinVar known pathogenic DGs. Statistics of the resulting dataset are shown in Fig. 6a.

**Key Points**

- We present a scalable framework that achieves 90% mean AP in identifying pathogenic gene associations across 2097 diseases, demonstrating remarkable accuracy in automated literature interpretation while effectively mitigating LLM hallucination risks.
- Our analysis of literature-based semantic embeddings revealed a consistent directional pattern in how pathogenic genes are represented across different diseases. While both pathogenic and non-pathogenic disease–gene pairs cluster similarly in the embedding space, we discovered a distinct semantic flow that indicates pathogenicity. This pattern could help automate the identification of disease-causing genes from scientific literature.
- The framework provides a reproducible methodology for leveraging LLMs in biomedical literature analysis, offering a valuable tool for researchers and clinicians in understanding disease mechanisms and identifying potential therapeutic targets.

## Acknowledgements

This manuscript was edited by Wallace Academic Editing. The authors also thank Dau-Ming Niu and Yun-Ru Chen at the Taipei Veterans General Hospital in Taiwan for their support (NSTC 113-2634-F-A49-003).

## Author contributions

P.H.L. conceived this study. P.H.L. and Y.Y.S. implemented pathogenic flow modeling and created figures. P.H.L. and J.H.H. wrote the manuscript. H.K.T., C.Y.C., and H.F.J. helped conceptualization and in the preparation of manuscript. J.H.H. contributed to conceptualization, writing, and project supervision. All authors are involved in discussion and finalization of the manuscript.

## Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Conflict of interest: None declared.

## Funding

This work was supported in part by the Center for Advanced Computing and Imaging in Biomedicine (NTU-113 L900701) from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education in Taiwan.

## Code availability

The code supporting the conclusions of this study is available on GitHub at https://github.com/ailabstw/LORE.

## Data availability

The PMKB-CV datasets supporting the findings of this study are available at https://doi.org/10.5281/zenodo.14607639.

## References

1. Fiorini N, Leaman R, Lipman DJ. *et al.* How user intelligence is improving PubMed. *Nat Biotechnol* 2018;**36**:937–45. https://doi.org/10.1038/nbt.4267.
2. Li PH, Chen TF, Yu JY. *et al.* pubmedKB: an interactive web server for exploring biomedical entity relations in the biomedical literature. *Nucleic Acids Res* 2022;**50**:W616–22. https://doi.org/10.1093/nar/gkac310.
3. Jin Q, Leaman R, Lu Z. PubMed and beyond: biomedical literature search in the age of artificial intelligence. *EBioMedicine* 2024;**100**:104988. https://doi.org/10.1016/j.ebiom.2024.104988.
4. Landrum MJ, Lee JM, Benson M. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* 2016;**44**:D862–8. https://doi.org/10.1093/nar/gkv1222.
5. Tate JG, Bamford S, Jubb HC. *et al.* COSMIC: the catalogue of somatic mutations In cancer. *Nucleic Acids Res* 2019;**47**:D941–7. https://doi.org/10.1093/nar/gky1015.
6. Amberger JS, Bocchini CA, Scott AF. *et al.* OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res* 2019;**47**:D1038–43. https://doi.org/10.1093/nar/gky1151.
7. Whirl-Carrillo M, Huddart R, Gong L. *et al.* An evidence-based framework for evaluating pharmacogenomics knowledge for personalized medicine. *Clin Pharmacol Ther* 2021;**110**:563–72. https://doi.org/10.1002/cpt.2350.
8. Hunter L, Cohen KB. Biomedical language processing: what's beyond PubMed? *Mol Cell* 2006;**21**:589–94. https://doi.org/10.1016/j.molcel.2006.02.012.
9. Baumgartner WA, Jr, Cohen KB, Fox LM. *et al.* Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics* 2007;**23**:i41–8. https://doi.org/10.1093/bioinformatics/btm229.
10. Bravo A, Pinero J, Queralt-Rosinach N. *et al.* Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC Bioinform* 2015;**16**:55. https://doi.org/10.1186/s12859-015-0472-9.
11. Pinero J, Bravo À, Queralt-Rosinach N. *et al.* DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res* 2017;**45**:D833–9. https://doi.org/10.1093/nar/gkw943.
12. Wu Y, Luo R, Leung HCM. *et al.* RENET: a deep learning approach for extracting gene-disease associations from literature. *Research in Computational Molecular Biology* 2019;**11467**:272–84. https://doi.org/10.1007/978-3-030-17083-7_17.
13. Legrand J, Gogdemir R, Bousquet C. *et al.* PGxCorpus, a manually annotated corpus for pharmacogenomics. *Sci Data* 2020;**7**:3. https://doi.org/10.1038/s41597-019-0342-9.
14. Pinto BGG, Oliveira AER, Singh Y. *et al.* ACE2 expression is increased in the lungs of patients with comorbidities associated with severe COVID-19. *J Infect Dis* 2020;**222**:556–63. https://doi.org/10.1093/infdis/jiaa332.
15. Buch AM, Vértes PE, Seidlitz J. *et al.* Molecular and network-level mechanisms explaining individual differences in autism spectrum disorder. *Nat Neurosci* 2023;**26**:650–63. https://doi.org/10.1038/s41593-023-01259-x.
16. Pu Y, Beck D, Verspoor K. Graph embedding-based link prediction for literature-based discovery in Alzheimer's disease. *J Biomed Inform* 2023;**145**:104464. https://doi.org/10.1016/j.jbi.2023.104464.
17. Szklarczyk D, Kirsch R, Koutrouli M. *et al.* The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res* 2023;**51**:D638–46. https://doi.org/10.1093/nar/gkac1000.
18. Xu D, Zhang M, Xie Y. *et al.* DTMiner: identification of potential disease targets through biomedical literature mining. *Bioinformatics* 2016;**32**:3619–26. https://doi.org/10.1093/bioinformatics/btw503.
19. Percha B, Altman RB. A global network of biomedical relationships derived from text. *Bioinformatics* 2018;**34**:2614–24. https://doi.org/10.1093/bioinformatics/bty114.
20. Lai PT, Wei CH, Luo L. *et al.* BioREx: improving biomedical relation extraction by leveraging heterogeneous datasets. *J Biomed Inform* 2023;**146**:104487. https://doi.org/10.1016/j.jbi.2023.104487.
21. Wei CH, Allot A, Lai PT. *et al.* PubTator 3.0: an AI-powered literature resource for unlocking biomedical knowledge. *Nucleic Acids Res* 2024;**52**:W540–6. https://doi.org/10.1093/nar/gkae235.
22. Liu S, Zhang X, Zhang S. *et al.* Neural machine reading comprehension: methods and trends. *Appl Sci* 2019;**9**:3698. https://doi.org/10.3390/app9183698.
23. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI Chatbot for medicine. *N Engl J Med* 2023;**388**:1233–9. https://doi.org/10.1056/NEJMsr2214184.
24. Singhal K, Azizi S, Tu T. *et al.* Large language models encode clinical knowledge. *Nature* 2023;**620**:172–80. https://doi.org/10.1038/s41586-023-06291-2.
25. Hou W, Ji Z. Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. *Nat Methods* 2024;**21**:1462–5. https://doi.org/10.1038/s41592-024-02235-4.
26. Jin Q, Leaman R, Lu Z. Retrieve, summarize, and Verify: how will ChatGPT affect information seeking from the medical literature? *J Am Soc Nephrol* 2023;**34**:1302–4. https://doi.org/10.1681/ASN.0000000000000166.
27. Lewis P, Petroni F, Karpukhin V. *et al.* Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv Neural Inf Process Syst* 2020;**33**:9459–74.
28. Gao Y, Xiong Y, Gao X. *et al.* Retrieval-augmented generation for large language models: a survey. arXiv Preprint arXiv: 2312.10997, 2023. https://doi.org/10.48550/arXiv.2312.10997.
29. Ouyang L, Wu J, Jiang X. *et al.* Training language models to follow instructions with human feedback. *Adv Neural Inf Process Syst* 2022;**35**:27730–44.
30. Brown T, Mann B, Ryder N. *et al.* Language models are few-shot learners. *Adv Neural Inf Process Syst* 2020;**33**:1877–901.
31. Mesquita F, Schmidek J, Barbosa D. Effectiveness and efficiency of open relation extraction. In: *Proceedings of the 2013 Conference*

*on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Seattle, Washington, USA, pp. 447–57, 2013.
32. McInnes L, Healy J, Saul N. *et al.* UMAP: uniform manifold approximation and projection. *J Open Source Softw* 2018;**3**:861. https://doi.org/10.21105/joss.00861.
33. Wu Q, Burges CJC, Svore KM. *et al.* Adapting boosting for information retrieval measures. *Inf Retr* 2009;**13**:254–70. https://doi.org/10.1007/s10791-009-9112-1.
34. Research AI4Science, M, Azure Quantum M. The impact of large language models on scientific discovery: a preliminary study using GPT-4. arXiv Preprint arXiv:2311.07361, 2023. https://doi.org/10.48550/arXiv.2311.07361.
35. Zhang T, Ladhak F, Durmus E. *et al.* Benchmarking large language models for news summarization. *Trans Assoc Comput Linguist* 2024;**12**:39–57. https://doi.org/10.1162/tacl_a_00632.
36. Team G, Mesnard T, Hardin C. *et al.* Gemma: open models based on Gemini research and technology. arXiv Preprint arXiv: 2403.08295, 2024. https://doi.org/10.48550/arXiv.2403.08295.
37. Dettmers T, Pagnoni A, Holtzman A. *et al.* QLoRA: efficient Finetuning of quantized LLMs. *Adv Neural Inf Process Syst* 2023;**36**: 10088–115.
38. Chen Z, Cano AH, Romanou A. *et al.* MEDITRON-70B: scaling medical Pretraining for large language models. arXiv Preprint arXiv: 2311.16079, 2023. https://doi.org/10.48550/arXiv.2311.16079.
39. Burges C, Shaked T, Renshaw E. *et al.* Learning to rank using gradient descent. In: *Proceedings of the 22nd international conference on Machine learning - ICML '05*, Association for Computing Machinery, New York, NY, USA, pp. 89–96, 2005.
40. Järvelin K, Kekäläinen J. Cumulated gain-based evaluation of IR techniques. *ACM Trans Inf Syst* 2002;**20**:422–46. https://doi.org/10.1145/582415.582418.
41. Burges C, Ragno R, Le Q. Learning to rank with nonsmooth cost functions. *Adv Neural Inf Process Syst* 2006;**19**:193–200. https://doi.org/10.7551/mitpress/7503.003.0029.
42. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;**32**:267D–0. https://doi.org/10.1093/nar/gkh061.


*Briefings in Bioinformatics*, 2025, **26(1)**, bbaf070
https://doi.org/10.1093/bib/bbaf070
Problem Solving Protocol