Experiment Plan: Discovery of Novel Conceptual Relations Based on Biomedical Literature

Rurui Guo*

November 9, 2026

Abstract

With the large-scale knowledge bases provided by artificial intelligence, I have been thinking about whether AI can be used to explore new ideas and viewpoints. In this paper, I attempt to discover new possible connections between concepts. I screen concept groups that appear infrequently in existing papers but are semantically plausible, and use these concept groups together with the relevant paper sentences as evidence supporting the relations between them to learn new relations. Finally, I use the learned new relations to generate new concept groups and examine the plausibility of the newly generated concept groups.

1 Construction of Common Conceptual Relations

1.1 Abstract-Based Concept and Claim Extraction

We retrieve up to 1,000 non-review Europe PMC records with abstracts published in 2022–2023 [4]. The title/abstract query uses the union of spatial-biology terms (e.g., spatial biology, spatial transcriptomics, and spatial omics) and gene-regulation terms (e.g., gene regulation, gene regulatory network, and transcriptional regulation). Because the pipeline retrieves abstracts, it does not extract text from the full Introduction or Discussion sections. A local Qwen3 model [1] extracts concepts, explicit claims, and claim evidence sentences from each abstract. MPNet [2] produces fixed concept embeddings for graph construction and semantic-distance filtering. The same Qwen3 model is also used for sentence-level evidence review; therefore, its review labels are not independent confirmation of its own extraction or candidate-generation outputs.

1.2 Construction of the Concept Relation Graph

After extracting and normalizing scientific concepts from biological literature, we construct a weighted concept relation graph to characterize the statistical structure of scientific knowledge. Let the corpus consist of N documents:

$$D = \{d_1, d_2, \ldots, d_N\}$$

and let the extracted concept set be:

$$C = \{c_1, c_2, \ldots, c_M\}$$

where each c_i represents a distinct scientific concept.

For each concept c_i, we define its document frequency as:

$$f(c_i) = \sum_{k=1}^{N} \mathbb{I}(c_i \in d_k)$$

*Email: 2400010876@stu.pku.edu.cn. Code: `https://github.com/GR-wyd/NewRelationLearning.git`.

where $\mathbb{I}(\cdot)$ is an indicator function that returns 1 if concept c_i occurs in document d_k, and 0 otherwise. The document frequency $f(c_i)$ is used as the node weight to quantify the prominence of each concept within the corpus.

To model relationships between concepts, we calculate the co-occurrence frequency of each concept pair (c_i, c_j):

$$f(c_i, c_j) = \sum_{k=1}^{N} \mathbb{I}(c_i \in d_k \wedge c_j \in d_k)$$

where $f(c_i, c_j)$ denotes the number of documents in which concepts c_i and c_j appear simultaneously. However, raw co-occurrence frequency tends to favor highly frequent concepts and may introduce spurious associations. Therefore, we normalize the co-occurrence strength using Pointwise Mutual Information (PMI) [3]:

$$PMI(c_i, c_j) = \log \frac{P(c_i, c_j)}{P(c_i)P(c_j)}$$

where the marginal and joint probabilities are estimated as:

$$P(c_i) = \frac{f(c_i)}{N},$$

$$P(c_i, c_j) = \frac{f(c_i, c_j)}{N}.$$

A positive PMI value indicates that two concepts co-occur more frequently than expected under statistical independence, suggesting a meaningful semantic association.

Based on these statistics, we construct a weighted undirected concept graph:

$$G = (V, E, W)$$

where each node $v_i \in V$ corresponds to a scientific concept c_i, and the node weight is defined as:

$$W_V(v_i) = f(c_i).$$

An edge $e_{ij} \in E$ is established between two concepts when their PMI value is positive, with the corresponding edge weight defined as:

$$W_E(e_{ij}) = PMI(c_i, c_j).$$

2 Temporal Discovery and Prospective Validation of Novel Conceptual Relations

The experiment uses three temporally separated Europe PMC [4] abstract corpora. The historical corpus $\mathcal{D}_{\text{hist}}$ contains 1,000 non-review records published in 2022–2023 and retrieved using the spatial-biology OR gene-regulation title/abstract query. The validation corpus $\mathcal{D}_{\text{evidence}}$ contains up to 3,000 records from 2024. The held-out corpus $\mathcal{D}_{\text{future}}$ contains up to 3,000 records from 2025 and is not used to select seed relations, learn relation patterns, generate candidates, or tune thresholds. These are Europe PMC records, not a Nature-only collection.

2.1 Historical Concept Relation Graph

For each concept c, we obtain an MPNet embedding e_c using MPNet [2]. Let $f(c)$ denote the number of historical documents containing c and $f(c_i, c_j)$ denote the number of documents containing both concepts. We estimate their probabilities as

$$P(c_i) = \frac{f(c_i)}{|\mathcal{D}_{\text{hist}}|}, \qquad P(c_i, c_j) = \frac{f(c_i, c_j)}{|\mathcal{D}_{\text{hist}}|}. \tag{1}$$

The statistical association between two concepts is measured using pointwise mutual information [3]:

$$PMI(c_i, c_j) = \log \frac{P(c_i, c_j)}{P(c_i)P(c_j)}. \tag{2}$$

We define a candidate distant concept pair (A, B) when

$$PMI(A, B) \leq \delta_{\text{PMI}}, \tag{3}$$

$$f(A, B) \leq \delta_{\text{freq}}, \tag{4}$$

and

$$\cos(e_A, e_B) \leq \delta_{\text{sem}}. \tag{5}$$

These conditions identify concept pairs that have weak historical association and are not trivially close in semantic embedding space.

2.2 Independent Literature Evidence

For each candidate pair (A, B), the primary evaluation searches the 2024 Europe PMC abstract corpus for sentences in which both concepts are explicitly mentioned. An optional targeted mode searches Europe PMC by title/abstract and falls back to PubMed through NCBI E-utilities when Europe PMC is unavailable. Historical DOI overlaps are excluded. Records sharing a normalized title across preprint and journal versions are deduplicated. A local Qwen3 reviewer [1] labels each matched sentence using only that sentence. Let

$$I(D_i, A, B) = \begin{cases} 1, & \text{if Qwen3 labels a sentence in } D_i \text{ as explicitly supporting} \\ & \text{a biological relation between } A \text{ and } B, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

The evidence count is

$$N_{\text{support}}(A, B) = \sum_{D_i \in \mathcal{D}_{\text{evidence}}} I(D_i, A, B). \tag{7}$$

A candidate is retained as a model-labeled, literature-supported under-observed relation when

$$N_{\text{support}}(A, B) \geq k, \tag{8}$$

where the primary threshold is $k = 3$ distinct articles. Articles are deduplicated using DOI, then PMID or database record identifier. No threshold sensitivity analysis is reported in the current experiment.

These are automated Qwen3 labels, not expert annotations or independent model verification; support metrics are interpreted as model-labeled literature evidence rather than ground truth.

Table 1 lists the 18 low-frequency `ours` pairs among the 500 tested that received model-labeled support from at least three distinct 2024 articles. Article counts are deduplicated by DOI, PMID, or database record identifier.

Table 1: Eighteen historically low-frequency pairs supported by 2024 literature.

Candidate rank	Concept pair	Supporting articles
97	differentiation – single-cell RNA sequencing	4
100	aging – single-cell resolution	3
101	spatial omics – tissues	8
107	gene expression patterns – mouse	3
121	development – tumor microenvironment	7
162	gene expression patterns – human	3
174	development – single-cell resolution	4
177	human – single-cell RNA sequencing	12
181	human – single-cell RNA-sequencing	3
192	aging – cell types	3
203	chromatin accessibility – development	7
224	resolution – tissue section	4
237	throughput – tissues	4
247	spatial location – tissues	3
320	single-cell RNA sequencing – spatial distribution	3
419	development – molecular mechanisms	19
436	spatial locations – transcriptomics	9
487	differentiation – transcriptomics	12

2.3 Latent Relation Learning

Let $A = (a_s, a_t)$ be a low-frequency concept pair selected from the historical relation graph. We define $\mathcal{H}_A$ as the set of historical claim triples and their evidence sentences in the graph that involve the same unordered concept pair. Let $\mathcal{V}_A$ be the sentences retrieved from Europe PMC that explicitly support a relation between a_s and a_t. Support is assigned by the sentence- level evidence reviewer. We count distinct articles, using DOI when available and otherwise PMID or the database record identifier:

$$N_{\text{articles}}(A) = \left|\{\text{id}(D) : D \text{ contains a supporting sentence in } \mathcal{V}_A\}\right|. \tag{9}$$

The pair is retained as a supported seed relation only when $N_{\text{articles}}(A) \geq 3$. Multiple sentences from the same article count once. Historical and validation evidence are then combined to infer a reusable relation pattern:

$$\rho_A = \text{Qwen3}\left(A, \mathcal{H}_A, \mathcal{V}_A\right). \tag{10}$$

The prompt requires the pattern to be grounded in the supplied sentences and prohibits adding mechanisms not stated in the evidence. To represent this evidence-conditioned pattern, let $\mathcal{X}_A$ contain the pattern description and the historical and supporting evidence sentences. MPNet encodes each text x as $u(x)$; the stored relation prototype is the normalized mean:

$$r_A = \frac{\bar{u}_A}{\|\bar{u}_A\|_2}, \qquad \bar{u}_A = \frac{1}{|\mathcal{X}_A|} \sum_{x \in \mathcal{X}_A} u(x). \tag{11}$$

This procedure constructs a prototype for each supported seed from its evidence. It does not train a relation encoder or cluster the prototypes; Qwen3 performs relation-pattern abstraction, while MPNet provides a fixed embedding representation.

2.4 Relation-Based Candidate Generation

For each retained seed relation A, Qwen3 applies the inferred pattern ρ_A to the fixed historical concept vocabulary $\mathcal{C}_{\text{hist}}$ and proposes new pairs $B = (b_s, b_t)$. The seed pair itself is excluded. These proposals are hypotheses, not relations already established by the validation articles. A proposal is retained only if both concepts belong to $\mathcal{C}_{\text{hist}}$ and the pair satisfies the historical novelty filters:

$$PMI(b_s, b_t) \leq \delta_{\text{PMI}}, \qquad f(b_s, b_t) \leq \delta_{\text{freq}}, \qquad \cos(e_{b_s}, e_{b_t}) \leq \delta_{\text{sem}}. \tag{12}$$

The retained proposals are ranked across supported seed relations, and at most 500 are written to the learned candidate set. The MPNet prototype r_A is stored with its seed and evidence for analysis; in this implementation, candidate generation is conditioned on the Qwen3 text pattern ρ_A, rather than nearest-neighbor retrieval with r_A. Generated candidates require subsequent literature validation and are not treated as discoveries solely because their seed relation has supporting evidence.

Table 2 reports the 50 unique concept pairs generated from the supported seed patterns. The seed column identifies the evidence-supported relation pattern used for each proposal.

Table 2: Fifty new concept pairs generated from evidence-supported relation patterns.

Rank	Generated concept pair	Supporting seed pair
1	gene regulatory network – spatial transcriptomics	development – molecular mechanisms
2	spatial transcriptomics – transcription factor	chromatin accessibility – development
3	gene regulatory network – spatially resolved transcriptomics	single-cell RNA sequencing – spatial distribution
4	spatially resolved transcriptomics – transcription factor	chromatin accessibility – development
5	spatially resolved transcriptomics – tumor microenvironment	throughput – tissues
6	cell-cell interactions – spatially resolved transcriptomics	development – single-cell resolution
7	cell-cell interactions – tumor microenvironment	development – molecular mechanisms
8	gene regulatory network – spatial gene expression	chromatin accessibility – development
9	cell-cell interactions – spatial omics	single-cell RNA sequencing – spatial distribution
10	cell-cell interactions – tissue architecture	development – tumor microenvironment
11	gene expression profiles – transcriptional regulation	gene expression patterns – mouse
12	gene regulatory networks – spatial domains	spatial location – tissues
13	spatial domains – spatially variable genes	spatial omics – tissues
14	spatial domains – transcriptional regulation	chromatin accessibility – development
15	spatial omics – tissue architecture	throughput – tissues
16	cell-cell interactions – spatial resolution	spatial location – tissues
17	gene expression regulation – transcription factor	differentiation – transcriptomics
18	gene regulation – tissue architecture	chromatin accessibility – development

Table 2: Generated concept pairs (continued).

Rank	Generated concept pair	Supporting seed pair
19	spatial clustering – tissue architecture	single-cell RNA sequencing – spatial distribution
20	cell-cell interactions – ligand-receptor interaction	spatial locations – transcriptomics
21	cell-cell interactions – spatial gene expression	throughput – tissues
22	chromatin accessibility – gene expression patterns	development – tumor microenvironment
23	chromatin accessibility – gene regulation	spatial omics – tissues
24	differentially expressed genes – gene expression profiles	development – molecular mechanisms
25	enhancer – spatially variable genes	chromatin accessibility – development
26	gene expression patterns – spatial domains	single-cell RNA sequencing – spatial distribution
27	gene expression profiles – tumor microenvironment	spatial location – tissues
28	spatial domains – Visium spatial gene expression	resolution – tissue section
29	spatial resolution – tissue regions	development – single-cell resolution
30	cell states – single-cell transcriptomics	differentiation – transcriptomics
31	cell-type-specific gene regulation – spatial clustering	resolution – tissue section
32	chromatin accessibility – cis-regulatory elements	gene expression patterns – mouse
33	gene expression patterns – transcription factor	development – single-cell resolution
34	spatial transcriptome – transcription factor binding sites	chromatin accessibility – development
35	spatially variable genes – tissue regions	single-cell RNA sequencing – spatial distribution
36	cell types – spatial gene expression	single-cell RNA sequencing – spatial distribution
37	data integration – single-cell multi-omics data	gene expression patterns – mouse
38	gene expression regulation – transcription factors	development – tumor microenvironment
39	spatial transcriptomics – transcriptional regulation	human – single-cell RNA sequencing
40	cell heterogeneity – spatially resolved transcriptomics	single-cell RNA sequencing – spatial distribution
41	cell states – spatial gene expression	spatial location – tissues
42	cell type deconvolution – single-cell transcriptomics	throughput – tissues
43	gene expression patterns – tissue regions	chromatin accessibility – development
44	gene expression patterns – transcriptional activity	differentiation – transcriptomics
45	spatial domains – spatial gene expression	development – single-cell resolution
46	biological insights – spatial transcriptome	single-cell RNA sequencing – spatial distribution
47	cell states – spatial clustering	throughput – tissues
48	cell type composition – spatial clustering	spatial location – tissues
49	gene expression variation – spatial domains	development – tumor microenvironment
50	protein profiling – spatial proteomics	spatial omics – tissues

2.5 Prospective Future Validation

After candidate generation and all filtering thresholds have been fixed, the learned candidate set is frozen. The 50 generated concept pairs are then evaluated against up to 3,000 held-out Europe PMC abstracts published in 2025. The same local Qwen3 sentence-level reviewer is used, so this is a model-labeled prospective literature test rather than an independent expert assessment.

For each generated pair (A, B), define

$$N_{\text{future}}(A, B) = \sum_{D_i \in \mathcal{D}_{\text{future}}} I(D_i, A, B). \tag{13}$$

A generated relation is counted as supported in the primary analysis if at least three distinct 2025 articles contain a sentence labeled as supporting by Qwen3:

$$N_{\text{future}}(A, B) \geq 3. \tag{14}$$

For the 50 learned candidates, the observed prospective support rate at this threshold is

$$\text{PDR} = \frac{1}{50} \sum_{i=1}^{50} \mathbb{I}\left[N_{\text{future}}(A_i, B_i) \geq 3\right]. \tag{15}$$

In this run, 7 of the 50 candidates met the threshold, giving PDR $= 7/50 = 0.14$. This rate is specific to the sampled 2025 Europe PMC corpus and the local Qwen3 reviewer.

This temporal evaluation measures whether relations inferred from historical literature subsequently receive explicit support in future scientific publications, without using future documents during relation learning, candidate generation, or threshold selection.

Among the 50 generated candidates, seven received model-labeled support in the held-out 2025 corpus from at least three distinct articles. Table 3 lists these candidates, their supporting article counts, and the 2024 seed relation from which each was generated. These are Qwen-labeled literature-supported candidates, not expert-confirmed discoveries.

Table 3: Seven generated pairs with support in the held-out 2025 corpus.

Rank	Generated concept pair	Supporting articles	Seed pair
6	cell-cell interactions – spatially resolved transcriptomics	4	development – single-cell resolution
15	spatial omics – tissue architecture	7	throughput – tissues
17	gene expression regulation – transcription factor	6	differentiation – transcriptomics
23	chromatin accessibility – gene regulation	5	spatial omics – tissues
26	gene expression patterns – spatial domains	3	single-cell RNA sequencing – spatial distribution
36	cell types – spatial gene expression	3	single-cell RNA sequencing – spatial distribution
38	gene expression regulation – transcription factors	3	development – tumor microenvironment

3 Limitations

Corpus scale and coverage. The experiment is based on up to 1,000 historical abstracts and up to 3,000 abstracts in each of the 2024 and 2025 corpora. These finite, keyword-retrieved samples do not cover the full biomedical literature. The title/abstract query uses a predefined

union of spatial-biology and gene-regulation terms, so papers using different terminology may be missed. Because the pipeline processes abstracts rather than full text, relevant evidence confined to the Methods, Results, or Discussion may not be retrieved. As a result, both the candidate distribution and the measured support rates may depend on the retrieval query and sample composition.

Model and evidence-review limitations. Qwen3 is used for concept and claim extraction, relation-pattern generation, and sentence-level support classification. This shared model creates dependence between generation and evaluation: its support labels are not independent verification and may reflect model biases, missed context, or errors in concept normalization. MPNet embeddings are fixed and are not fine-tuned or independently validated for this biomedical concept-pair task. Exact concept matching and sentence-level review can also miss paraphrased relations, while an abstract sentence may not contain enough context to determine the strength or direction of a biological claim. The three-article rule reduces reliance on a single record but does not replace expert review or establish that the reports are methodologically independent.

Heuristic novelty criteria. Document frequency, co-occurrence, PMI, and cosine-similarity cutoffs are operational filters applied to a historical corpus of only 1,000 abstracts. Estimates for rare concept pairs can therefore be unstable, and the selected thresholds can materially affect which pairs become seeds or generated candidates. No threshold sensitivity analysis is reported here. A pair that is infrequent or weakly associated in this sample should be described as *under-observed relative to this corpus*, not as globally novel.

Prospective-test interpretation. The result of 7 supported pairs among 50 learned candidates is specific to the retrieved 2025 sample, the chosen support threshold, and the Qwen3 reviewer. It is an exploratory, model-labeled literature-support rate, not a statistically powered estimate of discovery performance. The result does not show that the relations were absent from all prior literature, nor that the model predicted genuinely unknown biology; pretraining exposure to relevant papers cannot be ruled out. Independent expert adjudication and broader, pre-registered corpora are needed before drawing stronger conclusions.

Acknowledgments

This work is supported by Biomedical Computing Platform of National Biomedical Imaging Center, Peking University. We also thank YosTech (`https://yostech.ai`) for providing Codex tokens used during code development.

References

[1] An Yang et al. Qwen3 Technical Report. arXiv:2505.09388, 2025. `https://arxiv.org/abs/2505.09388`.

[2] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and Permuted Pre-training for Language Understanding. In *Advances in Neural Information Processing Systems 33*, 2020. arXiv:2004.09297.

[3] Kenneth Ward Church and Patrick Hanks. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

[4] The Europe PMC Consortium. Europe PMC: a full-text literature database for the life sciences and platform for innovation. *Nucleic Acids Research*, 43(D1):D1042–D1048, 2015. doi:10.1093/nar/gku1061.