CARDS: Citation-Aware Research Diffusion Summarization

Autumn Toney-Wails
UNU-MERIT
toney@merit.unu.edu

Katherine N. Quinn
Georgetown University
kq57@georgetown.edu

Steph Batalis
Georgetown University
sb2132@georgetown.edu

Abstract

Tracing how knowledge diffuses through scientific networks reveals how research builds on prior findings, but doing so at scale remains challenging. We present an LLM-based approach, Citation-Aware Research Diffusion Summarization (CARDS), which describes how knowledge diffuses through the scientific landscape. We trace citations from research funded by the U.S. National Institutes of Health through the landscape quantified by CSET's Map of Science cluster dataset to generate Research Diffusion Pathways, and compare LLM-generated summaries based on (1) cluster-level metadata and (2) citation-aware connections between research clusters. We find that CARDS outperforms cluster-based summaries under both human annotation and LLM-as-a-judge evaluation. Our work outlines a methodology for generating research diffusion paths, and for summarizing them with greater specificity.

1 Introduction

LLM-generated scientific summarization synthesizes information from research publications into concise descriptions of their content. Recent work has shown that incorporating citation context can improve these summaries by providing information about how research is referenced and used by subsequent work (or vice-versa) (Yasunaga et al., 2019; An et al., 2021; Syed et al., 2023; Nguyen et al., 2025a,b; Zhang et al., 2026). We combine these findings with prior work using citation network analysis to trace research diffusion (Chen and Hicks, 2004; Lewison et al., 2005; Liu and Rousseau, 2010; Gao and Guan, 2012; Yue et al., 2022; Aman and Gläser, 2025) and summarize the scientific knowledge underlying diffusion across research areas with scientific community detection and clusters (Boyack and Klavans, 2014; Traag et al., 2019; Rahkovsky et al., 2021).

We present **CARDS** (**C**itation-**A**ware **R**esearch **D**iffusion **S**ummarization), an LLM-based approach for summarizing these connections. Our pipeline has two components: (1) we construct Research Diffusion Pathways (RDPs), sequences of clusters representing distinct scientific areas, defined using citations and semantic similarity; and (2) we summarize the cited and citing publications at each inter-cluster connection (*diffusion point*) to describe the scientific knowledge linking the two research areas (*link summary*). CARDS synthesizes these link summaries into an RDP-level description of how research develops across the pathway.

To evaluate CARDS, we use research funded by the U.S. National Institutes of Health (NIH) as a case study, tracing its diffusion across scientific research areas, with data available on Zenodo at https://doi.org/10.5281/zenodo.23264540. We compare CARDS with a baseline that summarizes the same pathways using cluster-level descriptions rather than citation-level evidence. Using human and LLM-as-a-judge evaluations, we assess linkage coherence, knowledge diffusion specificity, factual consistency, and summary coverage. Both human and LLM-as-a-judge evaluations find CARDS summaries more specific than cluster-based summaries.

2 Experimental Design

Our experimental design consists of four stages: (1) linking NIH-funded publications to research clusters, (2) constructing RDPs that trace citations across clusters, (3) generating RDP summaries, and (4) evaluating a sample of RDP summaries.

2.1 Research Publication Data

We use two publicly available data sources: (1) the NIH RePORTER database[1] and (2) the Center for Security and Emerging Technology (CSET) Map of Science (Quinn et al., 2026).[2] NIH Re-

[1] https://reporter.nih.gov/
[2] https://sciencemap.eto.tech/

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 77–90
November 10, 2026 ©2026 Association for Computational Linguistics

PORTER is an online repository of NIH-funded research projects and their associated research outputs, including publications resulting from individual awards. CSET's Map of Science organizes approximately 146 million scientific articles into 91,594 research clusters using the Leiden community detection algorithm (Traag et al., 2019) over a hybrid network of citation relationships and semantic similarity links. For each research cluster, CSET provides aggregated metadata from the member publications and an LLM-generated (Gemini 3.1 Flash) title and one-sentence summary of the cluster. Using PubMed IDs, we link 2,987,570 NIH-funded publications identified through RePORTER to their corresponding research clusters (49,441 clusters in total). This linkage allows us to identify where NIH-funded research is located within the broader scientific landscape and provides the starting point for tracing its diffusion across research clusters. For our experiments, we rank research clusters by their concentration of NIH-funded publications and select the top 10% as starting points for RDP construction.

2.2 Research Diffusion Pathways

To map how research diffuses across science, we first build Research Diffusion Pathways (RDPs), sequences of research clusters connected through inter-cluster citations. Citation links can exist between clusters (via publications that cite or are cited by other work outside of their cluster) and connect publications across distinct areas of science. We refer to these inter-cluster connections as *diffusion points*, which provide observable instances in which research is cited outside the research area to which it is most closely connected.

We construct RDPs by iteratively following the publications involved in each diffusion point. The citing publications from the preceding diffusion point become the set of publications tracked in the next step. Candidate diffusion points for each subsequent step are ranked by the number of these tracked publications cited by publications in each possible destination cluster. To preserve the temporal direction of diffusion, we additionally require a minimum three-year difference in average publication year between consecutive links. Previously visited clusters are excluded, preventing cycles within a pathway. Figure 1 illustrates an RDP with two diffusion points connecting three clusters.

To construct RDPs of NIH-funded research, we use the clusters identified in Section 2.1 as starting

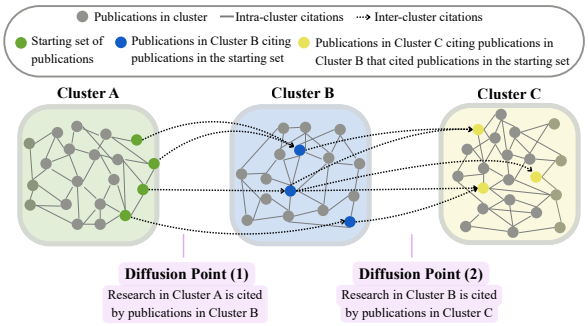


Figure 1: Example of an RDP connecting three clusters with two diffusion points.

points. For each starting NIH cluster, we select a diffusion point that maximizes the number of cited NIH-funded publications. We continue in this manner, selecting the RDP at each iteration which maximizes the number of distinct publications cited at each diffusion point (a greedy RDP), until we have an RDP for each starting cluster with four diffusion points connecting five clusters.

2.3 Research Diffusion Summarization

We test two LLM-based approaches for RDP summary generation that differ in their input: (1) cluster-based summaries and (2) citation-aware summaries (CARDS). By comparing these two approaches, we analyze whether the use of information specific to the diffusion points (i.e., the underlying citation relationships) provides a richer description of how research diffuses across science. We use Gemini 2.5 Flash for all experiments, and all summaries are restricted to 3–5 sentences with `MAX_TOKEN` $= 350$; the full prompts are listed in Appendix A. Figure 2 illustrates the full pipeline for RDP input to summary output, for the baseline cluster-based summary approach and CARDS.

The cluster-based RDP summaries are generated using the one-sentence description for each cluster in CSET's Map of Science (e.g., "Explores molecular mechanisms of cholesterol transport and the involvement of ABC transporters in human health and disease."). Thus, these summaries characterize the general diffusion based on changes in the broader research areas represented by consecutive clusters, without directly observing the publications responsible for the inter-cluster citation links.

CARDS are generated via a two-step summarization pipeline: (1) create a *link summary* for each diffusion point in a given RDP, and then (2) use the link summaries to generate an overall RDP summary. To create a link summary, we identify

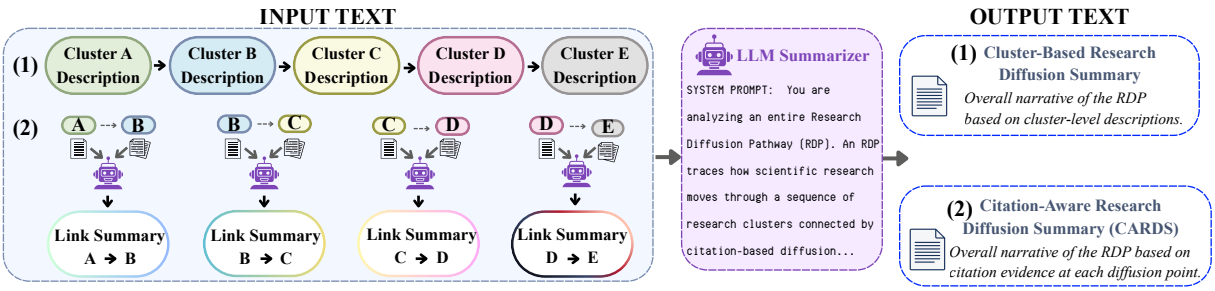


Figure 2: Experimental pipeline for cluster-based diffusion summaries and CARDS.

inter-cluster citation relationships in a diffusion point, preserving their one-to-many structure (i.e., each cited publication is paired with its set of citing publications). We provide the titles and abstracts of these publications to Gemini-2.5-Flash in a prompt to summarize the scientific content connecting the two clusters; we keep the titles but remove the abstracts of citing publications with the fewest citations until the prompt has an estimated 850K tokens to meet context window constraints. Rather than describing the broader research areas represented by the clusters, the link summaries focus specifically on the concepts, methods, findings, or applications associated with the citation relationships through which research crosses cluster boundaries.

2.4 Generative Summary Evaluation

We evaluate a random sample of 81 RDPs and their corresponding summaries with human and LLM-as-a-judge annotation. RDPs span a wide range of scientific research areas, and evaluating the factual consistency of the citation-aware summaries would require annotators to review the titles and abstracts of potentially thousands of publications underlying the diffusion points. We therefore use an automated evaluation to assess factual consistency, while focusing human annotation on two characteristics of the resulting RDP summaries: (1) linkage coherence and (2) knowledge diffusion specificity.

Human Evaluation: For linkage coherence, annotators rated whether each summary provided a coherent and plausible explanation of how research connects across the pathway on a five-point Likert scale (1 = *not coherent*; 5 = *extremely coherent*). For knowledge diffusion specificity, annotators compared the cluster-based summaries and CARDS for the same RDP and selected which more specifically described the concepts, methods, findings, or applications connecting the pathway, or indicated that they were approximately equally specific. Four annotators first evaluated 40 RDPs for linkage coherence, with either the cluster-based or citation-aware summary randomly presented, and then evaluated the remaining 41 RDPs through paired comparisons for knowledge diffusion specificity (a different subset of the sample to avoid annotator priming). For the paired comparisons, summary type was blinded and presentation order randomized. Full annotation instructions are provided in Appendix B.

LLM-as-a-Judge: We complement the human evaluation using DeepEval (Ip and Vongthongsri, 2026), an LLM-based evaluation framework, with GPT-5.4 as the LLM. For hallucination and factual consistency checks we implement the `HallucinationMetric`, which takes in the input text and the generated summary and then provides a score between 0 (hallucinations) to 1 (no hallucinations), and the `SummarizationMetric`, which provides a coverage score between 0 (no coverage) and 1 (full coverage). To run custom evaluations of linkage coherence and knowledge diffusion specificity, we use DeepEval's `GEval` (single input) and `ArenaGEval` (pairwise input) frameworks, respectively. For these custom evaluations, we provide the LLM with evaluation criteria corresponding to the instructions used in the human annotations.

3 Results

We identify 4,826 greedy RDPs originating in NIH-concentrated clusters[3] (118 of the original set of 4,944 were dropped due to incomplete paths); across this set, diffusion points contain a median of 384 citation relationships and 190,876 input tokens. Figure 3 shows three examples of representative RDPs from this set: (1) tracing how research on the immune response to viral infections provided the basis for the later development and production of next-generation CAR T-cell therapies, (2) how

[3]All link summaries and CARDS are available at: https://doi.org/10.5281/zenodo.23264540

CARDS Example #1

This RDP begins by establishing the critical role of tissue-resident memory T (TRM) cells in protection against respiratory viral infections, focusing on mechanisms of their generation, maintenance, and function, including antigen recognition and cytokine signaling. The pathway then progresses to investigate the broader mechanisms of CD8+ T cell differentiation, memory formation, and exhaustion in viral infections and cancer, emphasizing the heterogeneity of T cell subsets, particularly TCF1+ stem-like CD8+ T cells, and the influence of various transcription factors and cytokines. This understanding of T cell dysfunction and exhaustion, including specific transcription factors and epigenetic programs, is subsequently applied to develop strategies for enhancing CAR-T cell efficacy by preventing exhaustion, promoting stemness, and improving persistence through mitochondrial transfer, epigenetic reprogramming, and cytokine optimization. The pathway culminates in the development and evaluation of next-generation CAR T-cell therapies, moving from understanding limitations to designing novel approaches like combining CAR T-cells with small molecule drugs, optimizing manufacturing for less-differentiated cells, and identifying predictive biomarkers for improved clinical application.

CARDS Example #2

This RDP begins with the development and application of isotopic tracer methodologies and analytical techniques to quantify gluconeogenesis, glycogenolysis, and glucose turnover, primarily to elucidate the pathophysiology of diabetes and evaluate therapeutic strategies. This foundational understanding of glucose metabolism and insulin action then transitions into the development of mathematical and computational models of the glucose-insulin system, particularly for simulating postprandial glucose dynamics and designing control algorithms for artificial pancreas systems. Subsequently, these physiological and mathematical models are leveraged to create personalized blood glucose prediction models, integrating machine learning and deep learning techniques for advanced diabetes management. The pathway culminates in the application and optimization of these predictive models within broader diabetes management systems, including early detection, risk assessment, and personalized treatment recommendations, often integrated into mobile health applications and IoT systems.

CARDS Example #3

This Research Diffusion Pathway begins by characterizing the biochemical and functional properties of inositol polyphosphate kinases and phosphatases, establishing their roles in regulating inositol phosphate and phosphoinositide levels. This foundational understanding then progresses to elucidate the broader physiological and pathological roles of phosphoinositides and their metabolizing enzymes in diverse cellular processes like membrane trafficking, signal transduction, and metabolism, linking their dysregulation to diseases such as cancer and neurodegeneration. The pathway further narrows its focus to the phosphoinositide 3-kinase (PI3K)/Akt pathway, investigating its specific isoforms, regulatory mechanisms, and critical involvement in cancer development, progression, and drug resistance, as well as its regulation of cell growth and survival. Ultimately, the pathway culminates in the application of this knowledge to develop and evaluate PI3K/AKT/mTOR pathway inhibitors, such as capivasertib, alpelisib, and ipatasertib, for breast cancer treatment, including identifying predictive biomarkers and understanding mechanisms of therapeutic resistance.

Figure 3: Example CARDS output.

research in liver and kidney function enabled AI-driven disease prediction, and (3) how progression of basic research into single molecular functions led to breakthrough treatments for breast cancer (link summaries in Appendix C).

Figure 4 illustrates how NIH-funded research diffuses across domains. As expected given the NIH's focus on biomedical research, initial clusters are predominantly classified as biology (78.6%). However, after following the RDPs through four diffusion points, the proportion of biology clusters has decreased to 59.5%, while medicine grew from 1.9% to 13.4%, social science from 16.2% to 19.4%, materials science from 0.4% to 2.8%, and computer science from 0.3% to 1.7%. RDPs

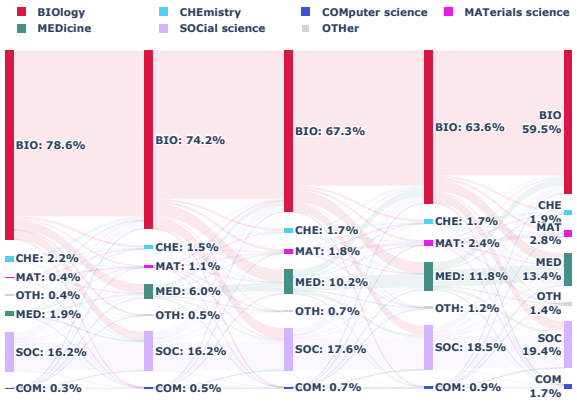


Figure 4: Field composition of the 4,826 NIH RDPs at each of the five cluster positions. Percentages give the fraction of clusters at that position assigned to each field (CSET field classification).

become more diffuse throughout the topical landscape, consistent with knowledge spreading from basic research (biology) to applied (medicine).

3.1 Evaluation

Human annotators find both cluster-based summaries and CARDS to be coherent. CARDS are rated as *Extremely coherent* or *Very coherent* 79.0% of the time, and cluster-based summaries 60.0% of the time (using a binomial test against a 2/5 null probability for random guessing with $N = 40$, obtaining $p < 0.01$ in both instances). Aggregated results are shown in Figure 5.

When comparing between two summaries for the same RDP, human annotators rated CARDS as more specific in describing the flow of information than the cluster-based summaries 70.7% of the time. A binomial test of CARDS-preferred versus all other responses gives $p = 0.017$ (with $N = 41$ majority vote against random guessing). Fleiss' κ is 0.073, indicating slight agreement between annotators. Nearly all disagreements arise when one annotator rates the citation-aware summaries as more specific and another annotator rates both summaries as equivalent.

We apply DeepEval to the full sample of LLM-generated summaries. No instances of hallucinations were detected in either case, and CARDS were uniformly rated as more specific. Mean coherence for the cluster-based summaries and CARDS is 0.880 and 0.901, respectively, indicating high coherence. Mean coverage is 0.953 and 0.719, respectively. This indicates a slightly higher overall coherence from CARDS, however less coverage

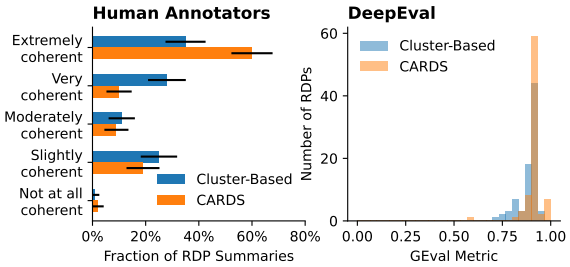


Figure 5: Coherence evaluation of cluster-based RDP summaries and CARDS.

from the input data, likely attributed to the larger input for summarization in the citation-awareness case which must condense numerous papers' titles and abstracts to a few sentences.

4 Discussion and Future Work

The CARDS examples shown in Fig. 3 indicate the type of analysis our proposed pipeline enables. In these three examples, we identify two patterns of scientific development that can be traced and summarized for end-user analysis: the more traditional progression from foundational research to practical applications and the diffusion of knowledge across scientific domains, where research in one field is taken up and extended in another. These examples demonstrate how CARDS can be used to better understand the connections between different areas of research and how scientific knowledge builds upon and extends previous findings.

Research progression from basic, foundational knowledge to application: CARDS Example #1, which starts in research cluster 22564[4] and ends in cluster 4838[5], follows CART T Cell therapy development which, a widely recognized important scientific and medical breakthrough. Along with CARDS Example #3, which starts in cluster 24449[6] and ends in cluster 17417[7] and explores glucose metabolism in the liver and kidneys through to broader health impact such as diabetes, these pathways follow foundational knowledge through to their applications.

Research diffusion *across* domains: CARDS Example #2 traces AI-enabled disease prediction. It begins in a biology-focused area, cluster 37912[8], which establishes scientific principles behind glucose metabolism, continues through social science with cluster 1156[9] that examines control systems and clinical practices for automated insulin delivery, and ends in a computer science with cluster 2002[10] to develop AI systems to predict blood glucose levels and diabetes. This RDP illustrates how knowledge can be discovered in one domain and passed through multiple other domains, where it continues to develop.

Our ongoing work uses RDPs and CARDS to analyze and evaluate the impact of funding portfolios (such as the NIH case study presented here). Specifically, we aim to better understand how funding decisions shape downstream scientific developments and the extent to which foundational research contributes to applications beyond its original scope. Take the CARDS Example #3: funding research on breast cancer alone is unlikely to have led to the development of this class of cancer drugs, because the mechanism they target was initially not known to be connected to breast cancer; it had to be discovered and characterized *before* being connected to breast cancer and leveraged as a therapeutic target.

5 Conclusion

We find that CARDS outperforms cluster-based approaches for summarizing the diffusion of knowledge through the scientific landscape. Human annotators and an LLM judge rate both summary types as coherent, but find CARDS summaries more specific. By incorporating citation-aware features in the summarization process, knowledge diffusing through the scientific landscape can be traced in a more specific manner than with cluster-level descriptions alone. These approaches can help address a range of research questions about how science develops, including how disruptive findings in one domain influence research across other fields, how breakthroughs draw upon knowledge from multiple domains, whether individual discoveries can have outsized impacts on subsequent research, and how long it takes for foundational research to develop into practical applications or contribute to multiple downstream research pathways.

[4] https://sciencemap.eto.tech/cluster/?version=4&cluster_id=22564

[5] https://sciencemap.eto.tech/cluster/?version=4&cluster_id=4838

[6] https://sciencemap.eto.tech/cluster/?version=4&cluster_id=24449

[7] https://sciencemap.eto.tech/cluster/?version=4&cluster_id=17417

[8] https://sciencemap.eto.tech/cluster/?version=4&cluster_id=37912

[9] https://sciencemap.eto.tech/cluster/?version=4&cluster_id=1156

[10] https://sciencemap.eto.tech/cluster/?version=4&cluster_id=2002

References

Valeria Aman and Jochen Gläser. 2025. Investigating knowledge flows in scientific communities: The potential of bibliometric methods. *Minerva*, 63(1):155–182.

Chenxin An, Ming Zhong, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2021. Enhancing scientific papers summarization with citation graph. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 12498–12506.

Kevin W Boyack and Richard Klavans. 2014. Creation of a highly detailed, dynamic, global model and map of science. *Journal of the Association for Information Science and Technology*, 65(4):670–685.

Chaomei Chen and Diana Hicks. 2004. Tracing knowledge diffusion. *Scientometrics*, 59(2):199–211.

Xia Gao and Jiancheng Guan. 2012. Network model of knowledge diffusion. *Scientometrics*, 90(3):749–762.

Jeffrey Ip and Kritin Vongthongsri. 2026. DeepEval: The Open-Source LLM Evaluation Framework.

Grant Lewison, Isla Rippon, and Steven Wooding. 2005. Tracking knowledge diffusion through citations. *Research Evaluation*, 14(1):5–14.

Yuxian Liu and Ronald Rousseau. 2010. Knowledge diffusion through publications and citations: A case study using esi-fields as unit of diffusion. *Journal of the American Society for Information Science and Technology*, 61(2):340–351.

Quoc-An Nguyen, Xuan-Hung Le, Hoang Quynh Le, and Hoang-Quynh Le. 2025a. Beyond the scientific document: A citation-aware multi-granular summarization approach with heterogeneous graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5034–5046.

Quoc-An Nguyen, Xuan-Hung Le, Thi-Minh-Thu Vu, and Hoang-Quynh Le. 2025b. Citearxiv: A citation-enriched dataset and heterogeneous graph-based model for scientific articles summarization. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 3084–3087.

Katherine Quinn, Rebecca Gelles, Jack Le, and Catherine Aiken. 2026. Map of science and map of patent cluster data.

Ilya Rahkovsky, Autumn Toney, Kevin W Boyack, Richard Klavans, and Dewey A Murdick. 2021. Ai research funding portfolios and extreme growth. *Frontiers in Research Metrics and Analytics*, 6:630124.

Shahbaz Syed, Ahmad Dawar Hakimi, Khalid Al Khatib, and Martin Potthast. 2023. Citance-contextualized summarization of scientific papers. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8551–8568.

Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. 2019. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):5233.

Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R Fabbri, Irene Li, Dan Friedman, and Dragomir R Radev. 2019. Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7386–7393.

Zenghui Yue, Haiyun Xu, Guoting Yuan, and Yan Qi. 2022. Modeling knowledge diffusion in the disciplinary citation network based on differential dynamics. *Scientometrics*, 127(12):7593–7613.

Hongpan Zhang, Bhummanat Sittipongpittaya, and Chongzhi Zang. 2026. Geneknow: An auditable ai framework for source-grounded biological evidence synthesis. *bioRxiv*.

A RDP Summary Prompts

Our experiments require three prompt variations: (1) cluster-based RDP summary, (2) link summary, and (3) citation-aware RDP summary.

First, the prompt shown in Figure 6 was used to generate RDP summaries from the ordered cluster-level titles and summaries provided by the CSET Map of Science. The prompt in Figure 7 was used to generate a link summary for each diffusion point in an RDP. The input preserves the citation relationships between cited publications in the source cluster and citing publications in the destination cluster. Lastly, Figure 8 displays the prompt for the CARDS output, taking the summaries generated at the RDPs diffusion points as input.

B Human Annotation Task

Four undergraduate research assistants completed the annotation task during their regular working hours using Qualtrics. Annotators were provided with a brief overview of the project and links to the underlying datasets to provide background on the study and research context, before they began their evaluations. The instructions given in the annotation interface are provided below:

Overall Instructions

In this task, you will evaluate summaries describing pathways of scientific research. Each pathway represents research that connects across multiple areas of scientific activity. The summaries are intended to explain how research connects and develops across these pathways. You will be asked to evaluate summaries based on two criteria:

- **Linkage Coherence:** How coherently the summary explains the connections across the research pathway.
- **Knowledge Diffusion Specificity:** Which of two summaries provides more specific information about the research connecting different parts of the pathway.

There are no right or wrong answers. Please base your ratings on the information presented and your judgment of how well each summary describes the research pathway.

B.1 Linkage Coherence Task

You will be shown one summary of a research diffusion pathway at a time.

Please evaluate whether the summary provides a coherent and plausible explanation of how the research connects across the pathway.

When making your judgment, consider whether:

- the relationships between different areas of research are understandable;
- the summary provides a logical account of how the research connects from one part of the pathway to another; and
- the pathway reads as a connected progression of research rather than as a collection of unrelated research topics.

Annotation Task:

How well does this summary provide a coherent and plausible explanation of how the research connects across the pathway?

Annotation Labels:

1. **Not at all coherent:** The summary does not provide an understandable explanation of how the research connects across the pathway.
2. **Slightly coherent:** Some connections are apparent, but the relationship between parts of the pathway is mostly unclear.
3. **Moderately coherent:** The general connections across the pathway are understandable, although some transitions or relationships are unclear.
4. **Very coherent:** The summary provides a clear and logical explanation of how research connects across most of the pathway.
5. **Extremely coherent:** The summary provides a clear, logical, and well-integrated explanation of how research connects throughout the pathway.

B.2 Knowledge Diffusion Specificity Task

You will now be shown two summaries describing the same research pathway. The summaries were generated using different information about the pathway. Compare the summaries based specifically on how well they describe the research or knowledge connecting different parts of the pathway.

More specific summaries may identify, for example:

- scientific concepts or ideas;

Cluster-Based RDP Summarization Prompt

Prompt

You are analyzing an entire Research Diffusion Pathway (RDP). An RDP traces how scientific research moves through a sequence of research clusters connected by citation-based diffusion.

Below are the research clusters in this RDP in pathway order, from the starting cluster to the final cluster. For each cluster, you are given its cluster title and cluster summary.

Your task is to summarize the scientific trajectory of the entire RDP. Write a concise 3--5 sentence summary explaining how the research focus develops across the ordered sequence of clusters.

Focus on:

- **The starting scientific area.** Identify the main concepts, problems, methods, mechanisms, technologies, or applications represented at the beginning of the pathway.

- **The progression across the pathway.** Explain the major scientific transitions from earlier clusters to later clusters, including changes in topic, method, scale, application, or disciplinary context.

- **Continuity and change.** Identify what scientific ideas or approaches appear to persist across multiple stages and what substantially changes as the pathway progresses.

- **The endpoint.** Describe the broader research area or application represented by the final cluster and how it differs from, or builds conceptually on, the earlier stages.

- **Scientific specificity.** Use concrete scientific terminology from the provided cluster titles and summaries rather than generic descriptions such as "related research," "advances in the field," or "different topics."

- **Evidence only.** Base the summary only on the supplied cluster titles and summaries. Do not invent paper-level citation details, specific mechanisms of transfer, or relationships that are not supported by this information.

Describe the pathway directionally, making clear how the scientific focus changes from the first cluster through the final cluster.

A successful summary should answer:

What scientific trajectory does this RDP represent, and how does the research focus evolve from its starting cluster to its endpoint?

ORDERED CLUSTERS IN THIS RDP:

<cluster_text>

Figure 6: Cluster-Based RDP Summarization Prompt

Link Summarization Prompt

You are analyzing a single link in a Research Diffusion Pathway (RDP). An RDP traces how scientific knowledge moves between clusters of research papers via citations.

The evidence below preserves the actual citation relationships for this link. Each evidence group contains:

- a **CITED SOURCE PAPER** from the source cluster, and

- the **CITING DESTINATION PAPER(S)** from the destination cluster that cite that source paper.

These citation relationships provide evidence about both what scientific knowledge is being transmitted from the source cluster and how that knowledge is being taken up, applied, extended, combined, or reframed in the destination cluster. Your task is to summarize this link only. Write a concise 3--5 sentence summary explaining the research diffusion represented by these citation relationships.

Focus on:

- **What knowledge is being transmitted.** Identify the specific scientific concepts, findings, methods, tools, mechanisms, datasets, technologies, clinical insights, or other contributions in the cited source papers that are carried across the link.

- **How the destination research draws on that knowledge.** Use the citing destination papers to explain how the transmitted knowledge is being used, applied, extended, combined with other ideas, or moved into a different research problem or application.

- **The scientific transition from source to destination.** Make the direction explicit. Prefer formulations such as The destination research draws on..." or The cited source work provides..." when supported by the evidence.

- **The nature of the connection.** When supported by the papers, identify whether the link represents transfer of a method, biological mechanism, empirical finding, conceptual framework, technology, dataset, clinical insight, or another type of scientific contribution.

- **Scientific specificity.** Use concrete scientific terminology rather than generic descriptions such as "related research," "similar methods," "advances in the field," or "a diffusion point."

- **Evidence only.** Base the summary only on the provided cited and citing papers. Do not infer a use or relationship that is not supported by their titles and abstracts.

- **Do not mention paper identifiers.** Paper IDs are provided only to preserve the citation relationships between source and destination papers. Never include paper IDs (for example, identifiers beginning with carticle_) in the summary. Refer to the scientific content of the papers rather than their identifiers.

Describe the actual scientific relationship from the source cluster to the destination cluster.

A successful summary should answer:

What scientific knowledge from the source cluster is crossing this link, and how is the destination-cluster research drawing on or carrying that knowledge forward?

Write the summary as standalone scientific prose. Do not mention paper IDs, evidence-group numbers, source-paper numbers, destination-paper numbers, or other internal identifiers from the input.

CITATION EVIDENCE FOR THIS LINK:

Figure 7: Link Summary Prompt

Citation-Aware RDP Summarization Prompt

Prompt

You are analyzing an entire Research Diffusion Pathway (RDP). An RDP traces how scientific knowledge moves through a sequence of citation-based links between research clusters.

Below are the link-level summaries for one RDP, ordered from the first link in the pathway to the final link. Each link summary describes what scientific knowledge moves from one cluster to the next and how the destination research draws on or carries that knowledge forward.

Your task is to synthesize these link summaries into a single concise summary of the entire RDP. Write a 3--5 sentence summary explaining the scientific trajectory across the pathway as a whole.

Focus on:

- **The starting scientific area.** Identify the main scientific concepts, findings, methods, mechanisms, technologies, or applications represented at the beginning of the pathway.

- **The progression across links.** Explain the major transitions in knowledge as the pathway moves from one cluster to the next. Emphasize how earlier findings, methods, or concepts are carried forward, applied, extended, combined, or reframed at later stages.

- **Continuity and change.** Identify the scientific ideas or approaches that persist across multiple links and the points where the pathway shifts into a different research problem, method, application, or disciplinary context.

- **The endpoint.** Describe the broader scientific area or application represented by the final stage of the pathway and how it relates to the earlier stages.

- **Scientific specificity.** Use concrete scientific terminology from the provided link summaries rather than generic descriptions such as "related research," "advances in the field," or "different topics."

- **Evidence only.** Base the RDP summary only on the supplied link summaries. Do not invent paper-level details, cluster characteristics, citation relationships, or mechanisms of transfer that are not supported by those summaries.

Synthesize the pathway rather than simply listing or repeating the four link summaries. Describe the pathway directionally, making clear how scientific knowledge evolves from the first link through the final link.

A successful summary should answer:

What scientific trajectory does this RDP represent, and how does knowledge evolve as it moves across the sequence of links?

ORDERED LINK SUMMARIES:

<link_summaries>

Figure 8: Citation-Aware Research Diffusion Summarization Prompt

- methods or techniques;
- findings or results;
- data or tools;
- applications; or
- other forms of research knowledge that connect one area of research to another.

Do not select a summary simply because it is longer or contains more detail overall. Focus on whether the additional detail helps explain what research or knowledge connects across the pathway.

Annotation Task:

Which summary more specifically describes the concepts, methods, findings, applications, or other research knowledge connecting different parts of the pathway?

- Summary A is more specific
- The summaries are about equally specific
- Summary B is more specific

C Example RDP Link Summaries

C.1 CARDS Example #1

Link Summary 1: The cited research highlights the critical role of tissue-resident memory T (TRM) cells, particularly in the lung airways, in providing rapid and effective protection against respiratory viral infections. The destination papers draw on this knowledge to investigate the mechanisms governing TRM cell generation, maintenance, and function, emphasizing factors like antigen recognition, cytokine signaling (e.g., IL-2, IL-7, IL-15, TGF-β), and the influence of the tissue microenvironment. This research aims to understand how these factors contribute to the heterogeneity and protective capacity of memory T cell subsets, ultimately informing the development of more effective T cell-based vaccines and immunotherapies for both infectious diseases and cancer.

Link Summary 2: The cited research investigates the complex mechanisms governing CD8+ T cell differentiation, memory formation, and exhaustion in the context of viral infections and cancer. These papers collectively highlight the heterogeneity of CD8+ T cell subsets, the molecular and epigenetic factors that drive their fate decisions, and the impact of the microenvironment on their function and persistence. The destination research draws on this foundational knowledge to explore specific regulatory pathways and therapeutic strategies. For instance, several destination papers utilize the understanding of TCF1+ stem-like CD8+ T cells, a population critical for sustained immune responses and checkpoint blockade efficacy, to investigate factors like TGF-β signaling, IL-15, and various transcription factors (e.g., KLF2, TOX, NR4A3, ID3, BATF, Gfi1, Tle1, PTMA) that influence their maintenance, differentiation, and anti-tumor activity. Other destination papers leverage insights into epigenetic programming, metabolic regulation, and the role of cytokines (e.g., IL-2, IL-10, IL-15, IL-21, IL-33) to develop novel immunotherapeutic approaches, such as targeting specific signaling pathways or using natural products to enhance T cell function and overcome exhaustion in cancer.

Link Summary 3: This RDP link illustrates the transmission of knowledge regarding T-cell dysfunction and exhaustion, particularly in the context of chronic infections and cancer, and its implications for CAR-T cell therapy. The source papers characterize T-cell exhaustion as a state of reduced function, altered transcriptional and epigenetic programs, and highlight the role of specific transcription factors (e.g., TCF1, TOX, BATF, NR4A) and cytokines (e.g., IL-2, TGF-β, IFN-I) in its development and maintenance, as well as the importance of stem-like T-cell populations. The destination papers draw on this knowledge to develop strategies for enhancing CAR-T cell efficacy by preventing or reversing exhaustion, promoting stemness and memory phenotypes, and improving persistence. This includes approaches such as mitochondrial transfer, modulating transcription factors (e.g., FOXO1, BACH2, TET2, NR4A, GSNOR, FOS), epigenetic reprogramming (e.g., targeting SUV39H1, HDACs, TET2, DNMT3A), optimizing cytokine signaling (e.g., IL-10, IL-15, IL-21, CXCL13), and engineering CAR-T cells to resist immunosuppressive microenvironments or improve infiltration.

Link Summary 4: This RDP link illustrates the ongoing efforts to enhance the efficacy and safety of CAR T-cell therapy, particularly in addressing challenges like T-cell exhaustion, tumor microenvironment suppression, and treatment-related toxicities. The source papers contribute foundational knowledge on mechanisms of CAR T-cell dysfunction, strategies to improve CAR T-cell persistence and anti-tumor activity (e.g., through transient "rest," cytokine release, or epigenetic repro-

gramming), and the impact of CAR design and manufacturing protocols on product quality. Destination papers draw on this knowledge to develop and evaluate novel approaches, such as combining CAR T-cell therapy with small molecule targeted drugs, optimizing manufacturing to enrich for less-differentiated T cells, and exploring predictive biomarkers for treatment outcomes and adverse events. This diffusion highlights a scientific transition from understanding the limitations of current CAR T-cell therapies to developing and testing next-generation strategies for improved clinical application.

C.2 CARDS Example #2

Link Summary 1: The cited source papers primarily focus on methods for quantifying gluconeogenesis, glycogenolysis, and glucose turnover in humans using various isotopic tracers (e.g., 2H2O, 13C-labeled glucose, 3H-labeled glucose, 14C-labeled lactate/acetate/alanine, 2H5-glycerol) and techniques like nuclear magnetic resonance spectroscopy and gas chromatography-mass spectrometry. They also investigate the relative contributions of these metabolic pathways to overall glucose production in different physiological states (fasting, postprandial) and in conditions like type 2 diabetes.

The destination research draws on this knowledge by applying and extending these isotopic tracer methodologies and analytical techniques to further investigate glucose metabolism in the context of type 1 and type 2 diabetes, obesity, and pre-diabetes. Specifically, the destination papers use these methods to:

- Assess the mechanisms of fasting and postprandial hyperglycemia in type 2 diabetes, including the roles of increased gluconeogenesis, impaired glycogen synthesis, and altered hepatic autoregulation.
- Quantify insulin sensitivity, beta-cell function, and glucose effectiveness in various diabetic and pre-diabetic populations.
- Evaluate the impact of interventions like caloric restriction and insulin therapy on glucose fluxes and metabolic pathways.
- Develop and validate new mathematical models for estimating glucose metabolism parameters from oral glucose tolerance tests and intravenous glucose tolerance tests.
- Investigate the role of specific organs (liver, kidney, muscle) in glucose homeostasis and their dysfunction in diabetes.

This represents a scientific transition from developing and validating fundamental metabolic measurement techniques to applying these advanced tools to elucidate the complex pathophysiology of diabetes and evaluate therapeutic strategies.

Link Summary 2: The destination research draws on methods and findings related to glucose metabolism and insulin action in both healthy individuals and those with type 1 and type 2 diabetes. Specifically, the cited source papers provide experimental protocols and insights into glucose-induced suppression of glucose production, impaired glucose disposal in type 2 diabetes, the relationship between hepatic glucose production and fasting plasma glucose, and the impact of insulin delivery patterns on carbohydrate tolerance. This knowledge is used by the destination papers to develop and refine mathematical and computational models of the glucose-insulin system, particularly for simulating postprandial glucose dynamics, assessing insulin sensitivity, and designing control algorithms for artificial pancreas systems in diabetes management. The diffusion highlights a transition from physiological studies using tracer techniques and clamp methods to the application and extension of these findings in in silico modeling and control system development for diabetes therapy.

Link Summary 3: The source papers provide a comprehensive foundation in mathematical modeling of glucose-insulin dynamics, including whole-body glucose kinetics, meal absorption patterns, insulin action, and the development of in silico simulators for type 1 and type 2 diabetes. This knowledge is transmitted to the destination research, which focuses on developing and evaluating advanced algorithms for blood glucose prediction and diabetes management. Specifically, the destination papers leverage these physiological and mathematical models to create personalized prediction models, integrate machine learning and deep learning techniques for improved accuracy, and design decision support systems for artificial pancreas applications, often validating their approaches using data generated from the established simulators. This represents a transition from foundational physiological modeling to the application and refinement of these models within predictive and control systems for diabetes.

Link Summary 4: The source cluster introduces various machine learning and deep learning models, including neural networks, support vector machines, and ensemble methods, for predicting blood glucose levels and hypoglycemic/hyperglycemic events in diabetic patients, often incorporating continuous glucose monitoring (CGM) data, meal information, insulin dosages, and physical activity. The destination cluster draws on this knowledge to develop and evaluate new predictive models for diabetes diagnosis, progression, and glucose level fluctuations, frequently employing and comparing different machine learning algorithms, feature selection techniques, and ensemble approaches. This diffusion highlights a transition from foundational glucose prediction models to their application and optimization within broader diabetes management systems, including early detection, risk assessment, and personalized treatment recommendations. The destination research also extends the application of these predictive techniques to different types of diabetes and integrates them into mobile health applications and IoT systems for enhanced patient self-management and clinical decision support.

C.3 CARDS Example #3

Link Summary 1: The cited source papers establish the biochemical and functional characteristics of various inositol polyphosphate kinases and phosphatases, including their substrate specificities, cellular localization, and roles in regulating inositol phosphate and phosphoinositide levels. This knowledge is transmitted to the destination research, which draws on these foundational findings to investigate the broader physiological and pathological roles of phosphoinositides and inositol polyphosphates. Specifically, the destination papers utilize this understanding to explore the involvement of these molecules in diverse cellular processes such as membrane trafficking, cytoskeletal organization, signal transduction, gene expression, and their implications in human diseases like cancer, neurodegeneration, and metabolic disorders. The scientific transition is from identifying and characterizing individual enzymes and their immediate products to elucidating their complex interplay within cellular networks and their impact on overall cell function and disease.

Link Summary 2: The source papers establish the critical roles of phosphoinositides (PIs) and their metabolizing enzymes (kinases and phosphatases) in regulating diverse cellular processes, including membrane trafficking, signal transduction, cell growth, proliferation, and metabolism. They highlight how different PI species are spatially and temporally regulated and how their dysregulation is linked to various human diseases, including cancer and neurodegeneration. The destination research draws on this foundational knowledge to investigate the specific functions of different PI3K isoforms and PI phosphatases in various physiological and pathological contexts. This includes exploring their roles in cancer development and therapy, immune responses, metabolic regulation, and endocytic pathways, often focusing on the therapeutic potential of targeting these enzymes. **Link Summary 3:** The destination research draws on the source cluster's foundational knowledge of phosphoinositide 3-kinases (PI3Ks) and their role in activating Akt (also known as protein kinase B or PKB), a key signaling molecule in various cellular processes. Specifically, the destination papers utilize the understanding of PI3K isoforms, their regulatory subunits, and the generation of phosphoinositide lipids like PI(3,4,5)P3 and PI(3,4)P2, which are crucial for Akt activation and localization. This knowledge is applied to investigate the PI3K/Akt pathway's involvement in cancer development, progression, and drug resistance, as well as its regulation of cell growth, survival, metabolism, and migration. The destination research also explores the isoform-specific functions of Akt and the spatial and temporal control of Akt activation within different cellular compartments, including the nucleus and various membrane structures, to better understand its diverse roles in both physiological and pathological contexts.

Link Summary 4: The cited research establishes the critical role of the PI3K/AKT/mTOR signaling pathway and its components (PI3K, AKT isoforms, PTEN, PDK1) in breast cancer development, progression, and therapeutic resistance. Specifically, it highlights the oncogenic potential of PIK3CA and AKT1 mutations, the distinct functions of Akt isoforms, and the importance of PDK1 in activating downstream kinases. The destination research draws on this knowledge to investigate the efficacy and safety of PI3K/AKT/mTOR pathway inhibitors, such as capivasertib, alpelisib, and ipatasertib, in breast cancer treatment. This includes exploring their use in various breast cancer subtypes, identifying predictive biomarkers for treatment response (e.g., PIK3CA mutations,

PTEN loss, AKT1 E17K mutation), understanding mechanisms of resistance (e.g., mTORC1 activation, feedback loops), and evaluating combination therapies to enhance anti-tumor activity and overcome resistance while managing adverse events like hyperglycemia.