KALIMBA: Knowledge-Assisted Literature Mining for Biological Interaction Analysis

Niloofar Arazkhani[1], Maciej Kotecki[2], Brent Cochran[2], Natasa Miskov-Zivanov[1]

[1]University of Pittsburgh, Pittsburgh, PA, USA, [2] Tufts University, Medford, MA, USA

Correspondence: nmzivanov@pitt.edu

Abstract

The exponential growth of biomedical literature has made manual curation of biological interaction networks increasingly difficult. Existing automated biological interaction extraction systems address the scaling challenge but treat extraction as a final step, delivering structured output with limited or no integrated support for biologists to interactively verify, correct and contextually interrogate extracted interactions against their source evidence within the same environment. We present Knowledge-Assisted Literature Mining for Biological Interaction Analysis (KALIMBA), an end-to-end, human-in-the-loop platform that integrates three complementary extraction methods (NLP-only, LLM-only, and hybrid) alongside expert annotation and evidence-grounded conversational querying through retrieval-augmented generation (RAG) chat module driven by a dual-context prompt, within a single unified workflow. Evaluation on a corpus of 40 signaling-focused papers demonstrates that the LLM-only back-end recovers substantially more interactions than the NLP-only approach. RAG chat evaluation by a domain expert confirms that the conversational module provides scientifically grounded responses that support curation decisions beyond what the structured interaction table alone conveys.

1 Introduction

The curation of large-scale biological knowledge graphs is an important challenge in systems biology, as these structured representations of molecular interactions underpin mechanistic modeling of signaling pathways, disease mechanisms, and drug response Lo Surdo et al. (2017); Milacic et al. (2024); Oughtred et al. (2019); Türei et al. (2021). These knowledge graphs are populated from interactions reported across tens of thousands of primary research articles Szklarczyk et al. (2016). Databases such as SIGNOR Lo Surdo et al. (2017), Pathway-Commons Rodchenkov et al. (2020), and BioGRID Oughtred et al. (2021) represent decades of expert curation, yet the rate of new discoveries consistently outpaces the capacity of manual curation to keep them current.

Significant research effort has been devoted to automating the extraction of biological interactions from the primary literature. Rule-based natural language processing (NLP) systems such as REACH Valenzuela-Escárcega et al. (2015), and INDRA Gyori et al. (2017) have demonstrated the feasibility of extracting structured regulatory interactions from biomedical articles at scale. The INDRA system in particular has established a strong foundation by integrating multiple reading engines and databases, and producing interactions in a standardized, machine-readable format with associated belief scores, which represent the confidence that a relationship is correct.

Despite this progress, existing systems share a common limitation: interaction extraction is their final step and there is no interface for annotating the quality or completeness of individual fields, or for asking follow-up questions about the context of reported findings. This disconnect between automated extraction and domain-expert judgment represents a significant barrier to further adoption of text mining tools in curation workflows.

Large language models (LLMs) offer a compelling opportunity to close the gap simultaneously from both ends. On the extraction side, LLMs provide broader coverage than rule-based NLP systems. On the curation side, the same conversational capabilities that make LLMs effective at extraction also make them natural assistants for post-extraction verification: a domain expert can query the source evidence behind a specific interaction, ask follow-up questions about experimental conditions, and receive grounded, context-aware responses without ever leaving the curation interface. A principled hybrid approach that combines the precision of established NLP systems with the

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 880–889

coverage and conversational capabilities of LLMs, while keeping a domain expert in the loop throughout the entire workflow, has not yet been realized in a single integrated platform.

In this paper, we introduce Knowledge-Assisted Literature Mining for Biological Interaction Analysis (KALIMBA), a human-in-the-loop curation platform that addresses these limitations by unifying automated interaction extraction, expert annotation, and evidence-grounded conversational querying within a single integrated workflow. KALIMBA treats knowledge curation as a collaborative process between automated systems and domain experts, keeping the expert actively engaged throughout the entire workflow.

2 Related work

Biological interaction and relation extraction. Early approaches to automated extraction of biological interactions relied on rule-based pattern matching and dependency parse trees to identify regulatory relationships between named entities in biomedical text. The REACH Valenzuela-Escárcega et al. (2015) system demonstrated that rule-based event extraction could recover mechanistic interactions from biomedical abstracts and full-text articles. INDRA Gyori et al. (2017) built on this foundation by integrating multiple reading engines into a unified assembly pipeline that produces computable statements with associated belief scores, enabling downstream pathway modeling and network assembly.

The emergence of language models has substantially expanded the capabilities of biomedical information extraction. BioBERT Lee et al. (2020) is a domain-specific language model pre-trained on large-scale biomedical corpora that significantly outperforms general-domain BERT Devlin et al. (2019) on biomedical named entity recognition and relation extraction.

Human-in-the-loop curation tools. Human-in-the-loop approaches have been widely applied in biomedical knowledge curation, ranging from active learning frameworks that reduce annotation effort to web-based manual annotation interfaces designed for structured text labeling. BRAT Stenetorp et al. (2012) is an intuitive web-based tool for text annotation supported by NLP technology and has been widely adopted for constructing gold-standard evaluation datasets across a range of biomedical NLP tasks. More recently, OntoChat Zhang et al. (2024) demonstrated that conversational language model agents can meaningfully assist domain experts throughout a knowledge engineering workflow without requiring programming expertise. KALIMBA extends this principle to biological interaction curation, applying conversational LLM assistance to the verification and contextual exploration of automatically extracted interactions grounded in primary literature evidence.

3 Methods

3.1 System Overview

KALIMBA is an end-to-end, human-in-the-loop platform for the automated extraction, structured curation, and interactive verification of biological regulatory interactions from primary biomedical literature. The system comprises two functional layers: an automated literature retrieval and extraction layer supporting three complementary methods, and a human-in-the-loop curation layer integrating expert annotation, RAG-based conversational querying, carefully engineered prompt template, and data import and export. A schematic overview of the full system architecture is presented in Figure 1.

3.2 Literature Retrieval

KALIMBA supports two entry modes that determine how paper content enters the pipeline. In search mode, the user queries PubMed directly through the interface, selects papers from the returned results, and abstracts are cached as part of the search session before extraction begins. In import mode, the user loads a previously extracted BioRECIPE Holtzapple et al. (2024) interaction list directly into the curation layer, bypassing extraction entirely and proceeding immediately to annotation, quality rating, and RAG-based conversational querying.

BioRECIPE is a structured biological interaction representation in a human-readable format, and KALIMBA's import and export including its built-in annotation tool support BioRECIPE format. BioRECIPE has 28 fields, including regulator and regulated entity names, interaction sign, mechanism, cellular compartment, cell line, tissue type, and supporting evidence statement.

For search mode, paper content is retrieved from National Center for Biotechnology Information (NCBI) through a cascading three-step strategy using the Entrez Utilities (E-utilities) RESTful

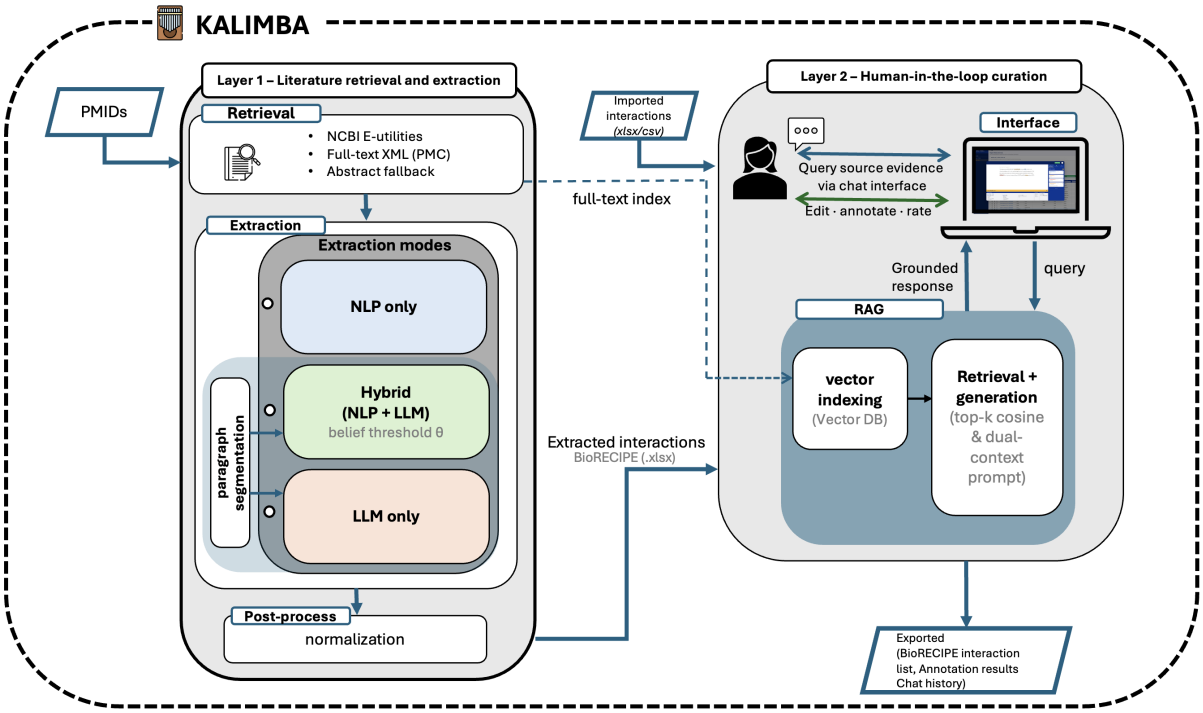

Figure 1: Overview of the KALIMBA framework

API Kans (2025). First, an elink call maps all input PMIDs simultaneously to their corresponding PubMed Central (PMC) identifiers; only PMIDs yielding a valid PMC entry are carried forward. Second, a single batch efetch call downloads full-text XML for all linked articles in one request. Body content is extracted by collecting all <p> elements within the <body> tag and joining them with double-newline separators to preserve paragraph boundaries. Third, a conditional efetch call retrieves PubMed abstracts for papers lacking PMC full-text coverage; this call is skipped entirely when abstracts are already cached from the search session, reducing total API calls to two for the common case.

Full-text content retrieved from PMC serves exclusively as the document corpus for RAG vector indexing (Section 3.7) and is not used as input to the extraction pipeline. Interaction extraction operates uniformly on abstract text for all input papers, regardless of full-text availability. This design choice ensures that extracted evidence statements are grounded in a consistent, universally available text source, preventing systematic differences in extraction quality between papers with and without PMC full-text coverage. In import mode, full-text content is fetched on demand at chat time if no index exists for the imported interaction's source papers, ensuring that RAG-based conversational querying remains available regardless of entry mode.

3.3 Paragraph Segmentation

Prior to extraction using LLM, each abstract is segmented into focused text units to constrain the language model's context window and to ensure that the Statements attribute of each extracted interaction is populated with a precise, traceable passage from the source text. This paragraph-level design serves two purposes: it focuses each extraction call on a narrow, semantically coherent passage, and it guarantees that the verbatim source text used by the annotator in the curation layer (Section 3.6) corresponds exactly to the context seen by the extraction model.

3.4 Interaction Extraction

KALIMBA exposes three configurable extraction methods, selectable at runtime via the interface (Appendix A).

NLP-only mode: Interactions are extracted via the INDRA reading pipeline, which uses NLP readers to process paper contents.

LLM-only mode: Each abstract paragraph produced by the segmentation stage (Section 3.3) is submitted independently to a language model with a structured prompt (Figure 2). KALIMBA is designed to be model-agnostic, supporting both open-

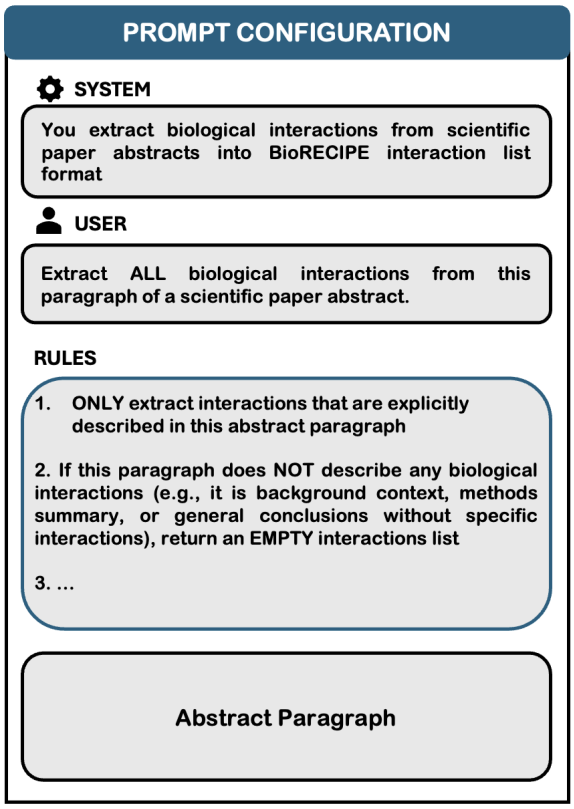

Figure 2: System and user prompt template for paragraph-level extraction.

source models served through the Ollama API and any OpenAI-compatible remote endpoint, allowing researchers to select the language model that best fits their computational resources, data privacy requirements, and performance needs.

Hybrid mode: The rule-based NLP-driven tool is applied first across all input papers. Interactions meeting a configurable belief score threshold (default $\theta = 0.7$) are retained directly. Papers with no NLP output, and papers with interactions that fall entirely below the belief threshold, are passed to the LLM extraction path. If LLM extraction subsequently fails for any paper in this set, the low-confidence interactions are preserved as a fallback rather than discarded. This strategy maximizes recall by using the LLM's broad coverage while preserving the precision of high-confidence traditional NLP extractions, avoiding the need to choose between the two approaches. The balance between NLP and LLM contributions is controlled by the belief threshold θ, which is developed as a continuous slider in the interface (range 0-1) that biologists can adjust before initiating extraction. This design gives biologists direct, interpretable control over the precision-recall trade-off without requiring any knowledge of the underlying implementation.

3.5 Post-Processing

All extracted interactions undergo post-processing normalization, which resolves non-standard field values and recovers a subset of missing fields through controlled vocabulary mapping and external database lookup. Type attributes are mapped to a controlled vocabulary of six categories: protein, gene, chemical, RNA, protein family, and biological process. Biological database identifiers are validated against type-appropriate sources, UniProt UniProt Consortium (2015) for proteins, HGNC Seal et al. (2023) for genes, PubChem Kim et al. (2016) for chemicals, Ensembl Harrison et al. (2024) for RNA, Pfam Mistry et al. (2021) for protein families, and Gene Ontology Gene Ontology Consortium (2019) for biological processes. Subcellular compartments are resolved and standardized against a set of seven canonical cellular locations, each matched with its corresponding Gene Ontology component ID. Mechanisms are similarly standardized by matching extracted values against a controlled vocabulary of 13 terms. Entity names unresolved during extraction are looked up asynchronously against the UniProt REST API (proteins and genes), HGNC (gene symbols), PubChem (chemicals), and the EBI QuickGO service (biological processes). Rows lacking either a regulator or regulated entity name, or missing an evidence statement, are discarded. The Source attribute is set programmatically to reflect the extraction origin of each row.

3.6 Human-in-the-Loop Curation and Annotation

A distinguishing feature of KALIMBA is its integrated human-in-the-loop curation layer, which treats automated extraction not as a final product but as a first-pass draft, subject to expert review and correction. The curation interface is implemented as an in-browser interaction table that uses the BioRECIPE interaction representation format with inline editing, per-row deletion, and column-level text filtering.

Each extracted interaction is linked to the verbatim paragraph text submitted to the extraction model, providing a traceable evidence statement against which automated field values can be verified. Clicking any statement opens the text annotation interface, where the evidence sentence is displayed with span highlights already pre-populated

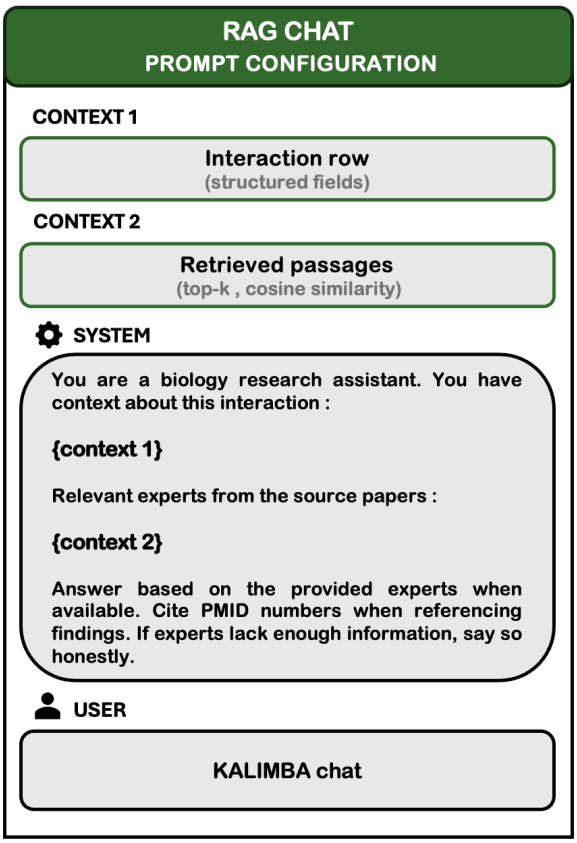

Figure 3: Dual-context prompt structure for RAG-based conversational querying.

from the extracted field values. From there, the biologist can verify whether each highlighted span correctly corresponds to its assigned BioRECIPE attribute, adjust any incorrect highlights, or add new ones for fields that the model left empty.

The interface also supports per-field quality rating and commentary, an overall interaction-level quality score, and a free-text general note field. The resulting structured evaluation data serves as a rich human feedback signal for downstream model evaluation and fine-tuning.

3.7 Retrieval-Augmented Interaction Chat

While interaction databases usually capture rich contextual information such as cell type, tissue, or experimental condition, domain experts often need to return to the primary literature to interrogate the full experimental context behind a specific interaction. KALIMBA addresses this through a tightly integrated RAG chat module that enables domain experts to query the source literature in natural language, anchored to a specific extracted interaction.

During the extraction phase in chat module, full-text content retrieved from PMC is segmented into overlapping 500-word chunks with a 100-word stride, preserving local semantic coherence while enabling retrieval at sub-section granularity. Each chunk is encoded into a dense vector representation using a sentence embedding model and stored in a vector database indexed by cosine similarity.

For papers unavailable in PMC, abstract text is indexed as a fallback. If no pre-built index exists at query time, for instance, when a user initiates chat on an interaction imported from an external file rather than extracted within the current session, full-text fetching and indexing are triggered on demand.

At query time, the user's natural language message is encoded using the same model and the top-k most similar passages (default k = 5) are retrieved by approximate nearest-neighbor search over the session collection. Retrieved passages are injected into the language model's system prompt alongside the structured format with attributes of the selected interaction (Figure 3). This dual-context design grounds responses simultaneously in the structured semantics of the BioRECIPE format (context 1) and in the unstructured experimental narrative of the source paper (context 2), reducing the risk of hallucinated mechanistic claims that are not supported by the cited evidence.

This capability represents a qualitatively different mode of engagement with curated biological knowledge. A researcher inspecting an extracted phosphorylation interaction can immediately ask: "What experimental method was used to validate this interaction?" or " Did the authors note any dose-dependent effects?". These questions can directly help curator in completing and enriching structured interaction records.

3.8 Data Import and Export

To ensure compatibility with a range of systems biology tools Kanehisa et al. (2016); Luo et al. (2026); Milacic et al. (2024); Pratt et al. (2015) and to support KALIMBA as a standalone curation environment, the system includes an import layer accepting BioRECIPE formatted files.

KALIMBA provides three structured export formats. The interaction list export produces a single-sheet Excel workbook with one column per BioRECIPE attribute, suitable for direct use in downstream pathway modeling tools. The annotation export produces a four-sheet workbook capturing span-level annotation details, per-interaction summaries, per-attribute quality ratings, and a full rating matrix. The chat history export produces a

two-sheet workbook containing message-level dialogues and conversation summaries, with each row enriched with the full BioRECIPE context of the discussed interaction.

Methods	Total interactions	Papers	
		covered	no output
NLP-only (INDRA)	392	24	16
LLM-only (LLaMA 3.1)	363	39	1
Hybrid ($\theta = 0.7$)	388	39	1

Table 1: Extraction coverage across methods.

Source	Extraction origin	Interactions	
		count	percentage
LLM	papers INDRA missed	191	49.2%
	low-confidence INDRA, score < 0.7	161	41.5%
INDRA /REACH	high-confidence, score ≥ 0.7, retained	32	8.2%
INDRA /Sparser	high-confidence, score ≥ 0.7, retained	4	1.0%
Total		**388**	**100%**

Table 2: Distribution of hybrid mode interactions.

4 Experiments

4.1 Experimental Setup

Dataset. To evaluate KALIMBA's extraction capabilities, we used a corpus of 40 signaling-focused biomedical papers from high-impact journals including Cell and Nature, published between 2023 and 2026. selected by a domain expert. The corpus is biased toward the RAS and JAK signaling pathways while maintaining diversity across related signaling contexts. Papers involving immune cell signaling were excluded due to the subtleties of cell-type-specific signaling that introduce additional ambiguity beyond the scope of this evaluation.

Models. For the LLM-only and hybrid extraction, we used a locally hosted open-source model (LLaMA 3.1 8B served via the Ollama API) Grattafiori et al. (2024). Although KALIMBA supports remote OpenAI-compatible endpoints such as GPT-4o Hurst et al. (2024), we used LLaMA configuration in this evaluation to avoid API costs and to demonstrate that KALIMBA produces meaningful extraction results without requiring access to closed-source API-dependent models. All experiments were conducted on a MacBook Pro with an Apple M3 Pro chip and 18 GB memory.

Metrics. The 40 papers were processed through all three extraction methods independently. Extraction outputs were compared across methods in terms of paper coverage, total interaction yield, inter-mode agreement, and confidence score distribution. For the RAG chat evaluation, a domain expert curated a set of standardized questions per interaction and rated the system responses on a 0-to-5-star scale, providing qualitative notes for each response.

4.2 Extraction Performance

Table 1 reports the total number of interactions extracted per mode and the number of papers for which at least one interaction was recovered across the full 40-paper corpus. The NLP-only back-end failed to produce any interactions for 16 of the 40 papers. All 16 of these papers were successfully processed by the LLM-only method, demonstrating that the LLM provides substantial coverage recovery for papers outside NLP's effective range. The hybrid mode inherits the LLM's coverage advantage, recovering interactions from 39 of 40 papers while retaining high-confidence NLP extractions where available.

Table 2 reports the source breakdown of interactions in the hybrid output, characterizing the practical contribution of each component.

Only 9.2% of hybrid interactions were contributed directly by high-confidence NLP output. The remaining 90.7% were produced by LLM re-extraction either for papers NLP missed entirely (49.2%) or for papers where NLP's belief scores fell below $\theta = 0.7$ (41.5%). This distribution confirms that the hybrid routing strategy is functioning as designed: NLP's precision is preserved for the interactions it handles confidently, while the LLM fills in coverage for the substantial portion of the corpus that NLP either cannot process or does not assign high confidence scores.

4.3 Confidence and Inter-Mode Agreement

Table 3 reports the confidence score distribution per method. NLP-only scores cluster narrowly between 0.650 and 0.980, reflecting NLP's con-

Method	Mean	Median	Min	Max
NLP-only (INDRA)	0.67	0.65	0.65	0.98
LLM-only (LLaMA 3.1)	0.87	1.00	—	1.00
Hybrid ($\theta = 0.7$)	0.85	0.90	0.00	1.00

Table 3: Confidence score distribution per method.

servative belief scoring system. The LLM-only back-end assigns higher scores on average, but these scores are self-reported and do not guarantee extraction correctness. A limitation directly observed in the RAG chat evaluation, where a highly confident but incorrectly extracted interaction received zero stars across all expert questions.The hybrid mode achieves a higher mean score than NLP-only alone as a direct consequence of the low-confidence NLP interactions are replaced by LLM re-extractions that receive higher confidence scores. Table 4 reports pairwise interaction overlap across the three methods, which show remarkably low agreement, with only 9 interactions agreed upon by all three methods. These 9 interactions include canonical RAS pathway relationships such as MEK→ERK and RASA1→RAS, providing a validity check that all three methods correctly capture well-established signaling interactions. The substantially higher LLM–Hybrid overlap (129 pairs) compared to NLP–LLM overlap (10 pairs) reflects the fact that 90.7% of hybrid interactions are LLM-derived, making these two the most similar pair. The low overall consensus confirms that the three methods are largely complementary rather than redundant, each recovering a distinct subset of the interactions present in the corpus and strengthening the case for the hybrid design.

4.4 RAG Chat Quality Evaluation

To evaluate the RAG-based conversational querying module, a domain expert curated a set of six standardized questions that were defined targeting distinct aspects of the supporting evidence.

Each response was rated by the domain expert on a 0-to-5-star scale and followed by a qualitative curator note where responses were incomplete or incorrect. A total of 47 responses were evaluated across 8 interactions. Table 5 shows that the overall mean rating of 2.64 out of 5 is largely driven down by one incorrectly extracted interaction that received 0.00 across all questions. Excluding it, the

Comparison	Shared pairs	Unique to first	Unique to second
NLP-only vs LLM-only	10	325	231
NLP-only vs Hybrid	44	325	221
LLM-only vs Hybrid	129	231	221
Agreed by all three	9	—	—

Table 4: Pairwise interaction overlap across methods (regulator and regulated name matches, n=40 papers).

adjusted mean rises to 3.27 out of 5 with 83.7% of responses rated 3 stars or above, indicating strong performance for correctly extracted interactions

Qualitative findings. Expert review of curator notes revealed three recurring patterns. First, RAG chat quality is tightly coupled to extraction correctness. The interaction AMP-activated protein kinase → glucose metabolism received a rating of 0.00 across all six questions because the interaction direction was incorrectly extracted. This finding suggests that the chat module itself can serve as a verification tool when responses are inconsistent with biological expectations. This further signals to the expert that the underlying extracted interaction may be incorrect and warrants correction in the annotation layer before further querying.

Second, the system performs strongest on directional and pathway-level questions. Activation vs. inhibition questions received the highest mean rating (3.50) and signaling pathway context questions the second highest (3.12). These question types are most directly answerable from the dual-context prompt that combines the structured BioRECIPE row with the retrieved passages, and the results suggest this design provides reliable grounding for sign and pathway membership queries.

Third, mechanistic and experimental evidence questions are more challenging. Biological mechanism (2.12) and cell line (2.00) questions received the lowest ratings. Expert notes identified two specific failure modes: responses occasionally cited literature referenced within retrieved passages rather than the experimental findings of the paper itself, and responses sometimes provided accurate but incomplete summaries of experimental evidence, omitting key details such as specific phosphorylation site identification, mutational analyses, and

tissue-specific knockout results.

Question type	Rating	
	mean	% rated ≥ 3 stars
Activation vs inhibition	3.50	75.0%
Signaling pathway context	3.12	87.5%
Explicit statement verification	2.75	62.5%
Experimental evidence	2.25	75.0%
Biological mechanism	2.12	62.5%
Cell line / condition	2.00	71.4%
Overall	2.64	72.3%

Table 5: Mean expert ratings of RAG chat responses by question type (n=47, scale 0–5).

Expert Utility Assessment. Following the quantitative rating evaluation, the domain expert was asked three follow-up questions regarding the utility of the RAG chat module for biological curation. The expert confirmed that the chat responses provided meaningful additional context beyond the structured interaction row across all evaluated interactions. Three specific advantages were identified. First, the chat module demonstrated the ability to correct upstream extraction errors for example for the incorrectly extracted interaction AMP-activated protein kinase → glucose metabolism, the expert noted that KALIMBA's responses consistently described the correct interaction direction, effectively flagging the extraction error without explicit prompting. Second, for interactions where the connection type was incorrectly assigned as direct or indirect, the chat module provided evidence that allowed the expert to identify and correct the misclassification. Third, in the cases where the structured interaction row did not include any information on experimental evidence, mechanism, site, cell line, cell type, or tissue type, the chat module provided varying degrees of detail across all of these attributes for every interaction, which the expert described as a significant advantage for curation.

The expert confirmed they would use the KALIMBA chat feature in their own literature review and curation workflow.

5 Conclusion

We presented KALIMBA, a human-in-the-loop curation platform that addresses a persistent gap in biomedical text mining: the disconnect between automated interaction extraction and the expert biological judgment required to verify, correct, and contextually interpret extracted knowledge. KALIMBA integrates three complementary extraction methods, a span-level annotation interface, and a retrieval-augmented conversational querying module within a single unified workflow. Together these capabilities treat biological knowledge curation as a collaborative process between automated systems and domain experts rather than a pipeline that terminates at extraction.

Evaluation on a 40-paper signaling-focused corpus revealed several findings with implications beyond KALIMBA itself. The NLP-only method failed to extract any interactions from 40% of the corpus, a coverage gap that the LLM method recovered entirely, demonstrating that LLM-based extraction provides meaningful complementarity to established NLP systems for recent literature. The hybrid mode successfully combined both approaches, with 90.7% of its output contributed by LLM re-extraction and only 9.3% retained directly from high-confidence NLP output, reflecting the corpus composition bias toward recently published papers underrepresented in NLP's knowledge base. Inter-mode agreement was remarkably low, only 9 interactions were agreed upon by all three methods confirming that the three methods are largely complementary rather than redundant, and that each recovers a distinct subset of the interactions present in the literature. The 9 consensus interactions included canonical RAS pathway relationships such as MEK→ERK and RASA1→RAS, providing validity that all three methods correctly capture well-established signaling interactions.

The RAG chat evaluation demonstrated that conversational querying of source evidence provides grounded responses for correctly extracted interactions, with an adjusted mean rating of 3.27 out of 5 and 83.7% of responses rated 3 stars or above when the one incorrectly extracted interaction is excluded. The finding that chat quality is tightly coupled to extraction quality reinforces the importance of the annotation and correction layer as a prerequisite for productive conversational querying. Together these results position KALIMBA as a practical tool for domain experts who need to extract, verify, and explore biological interactions from the primary literature at scale, without requiring computational expertise or access to commercial API services.

6 Limitations and Future Work

The current evaluation has several limitations that are considered for future work. The corpus of the papers that we used has 40 signaling-focused papers recommended by domain experts, which is very rich in terms of studying the signaling pathways, particularly in the RAS and JACK signals, but it is not constructed as a dataset of gold standard interaction extraction in BioRECIPE format, so the extraction results are reported in terms of coverage not in accuracy related metrics such as precision, recall and F1 against a gold standard dataset. Evaluating the accuracy of extractions using an expanded, more diverse corpus is a priority for our future work. For the RAG chat evaluation, we relied on a single-domain expert who rated 47 responses across 8 interactions using a structured set of question types. In future work, we will add more independent evaluators to strengthen the reliability of these findings and compute an inter-annotator score to generalize the results. To make the system more user-friendly, a formal usability evaluation of the interface with domain expert biologists is planned as future work.

The current prompt template follows a general instruction-based design. Future work will improve extraction performance by incorporating curated few-shot examples into the prompt and evaluating results using commercial models such as GPT-4o.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Gene Ontology Consortium. 2019. The gene ontology resource: 20 years and still going strong. *Nucleic acids research*, 47(D1):D330–D338.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Benjamin M Gyori, John A Bachman, Kartik Subramanian, Jeremy L Muhlich, Lucian Galescu, and Peter K Sorger. 2017. From word models to executable models of signaling networks using automated assembly. *Molecular systems biology*, 13(11):MSB177651.

Peter W Harrison, M Ridwan Amode, Olanrewaju Austine-Orimoloye, Andrey G Azov, Matthieu Barba, If Barnes, Arne Becker, Ruth Bennett, Andrew Berry, Jyothish Bhai, and 1 others. 2024. Ensembl 2024. *Nucleic acids research*, 52(D1):D891–D899.

Emilee Holtzapple, Gaoxiang Zhou, Haomiao Luo, Difei Tang, Niloofar Arazkhani, Casey Hansen, Cheryl A Telmer, and Natasa Miskov-Zivanov. 2024. The biorecipe knowledge representation format. *ACS Synthetic Biology*, 13(8):2621–2624.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. 2016. Kegg as a reference resource for gene and protein annotation. *Nucleic acids research*, 44(D1):D457–D462.

Jonathan Kans. 2025. Entrez® direct: E-utilities on the unix command line. In *Entrez® Programming Utilities Help [Internet]*. National Center for Biotechnology Information (US).

Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A Shoemaker, and 1 others. 2016. Pubchem substance and compound databases. *Nucleic acids research*, 44(D1):D1202–D1213.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Prisca Lo Surdo, Alberto Calderone, Gianni Cesareni, and Livia Perfetto. 2017. Signor: a database of causal relationships between biological entities—a short guide to searching and browsing. *Current protocols in bioinformatics*, 58(1):8–23.

Haomiao Luo, Casey Hansen, Niloofar Arazkhani, Cheryl A Telmer, Difei Tang, Gaoxiang Zhou, Peter Spirtes, and Natasa Miskov-Zivanov. 2026. Violin: A modular framework for scalable reconciliation of heterogeneous interaction graphs. *bioRxiv*, pages 2024–07.

Marija Milacic, Deidre Beavers, Patrick Conley, Chuqiao Gong, Marc Gillespie, Johannes Griss, Robin Haw, Bijay Jassal, Lisa Matthews, Bruce May, and 1 others. 2024. The reactome pathway knowledgebase 2024. *Nucleic acids research*, 52(D1):D672–D678.

Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik LL Sonnhammer, Silvio CE Tosatto, Lisanna Paladin, Shriya Raj, Lorna J Richardson, and 1 others. 2021. Pfam: The protein families database in 2021. *Nucleic acids research*, 49(D1):D412–D419.

Rose Oughtred, Jennifer Rust, Christie Chang, Bobby-Joe Breitkreutz, Chris Stark, Andrew Willems, Lorrie Boucher, Genie Leung, Nadine Kolas, Frederick Zhang, and 1 others. 2021. The biogrid database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science*, 30(1):187–200.

Rose Oughtred, Chris Stark, Bobby-Joe Breitkreutz, Jennifer Rust, Lorrie Boucher, Christie Chang, Nadine Kolas, Lara O'Donnell, Genie Leung, Rochelle McAdam, and 1 others. 2019. The biogrid interaction database: 2019 update. *Nucleic acids research*, 47(D1):D529–D541.

Dexter Pratt, Jing Chen, David Welker, Ricardo Rivas, Rudolf Pillich, Vladimir Rynkov, Keiichiro Ono, Carol Miello, Lyndon Hicks, Sandor Szalma, and 1 others. 2015. Ndex, the network data exchange. *Cell systems*, 1(4):302–305.

Igor Rodchenkov, Ozgun Babur, Augustin Luna, Bulent Arman Aksoy, Jeffrey V Wong, Dylan Fong, Max Franz, Metin Can Siper, Manfred Cheung, Michael Wrana, and 1 others. 2020. Pathway commons 2019 update: integration, analysis and exploration of pathway data. *Nucleic acids research*, 48(D1):D489–D497.

Ruth L Seal, Bryony Braschi, Kristian Gray, Tamsin EM Jones, Susan Tweedie, Liora Haim-Vilmovsky, and Elspeth A Bruford. 2023. Genenames. org: the hgnc resources in 2023. *Nucleic acids research*, 51(D1):D1003–D1009.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Damian Szklarczyk, John H Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T Doncheva, Alexander Roth, Peer Bork, and 1 others. 2016. The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic acids research*, page gkw937.

Dénes Türei, Alberto Valdeolivas, Lejla Gul, Nicolàs Palacio-Escat, Michal Klein, Olga Ivanova, Márton Ölbei, Attila Gábor, Fabian Theis, Dezső Módos, and 1 others. 2021. Integrated intra-and intercellular signaling knowledge for multicellular omics analysis. *Molecular systems biology*, 17(3):MSB20209923.

UniProt Consortium. 2015. Uniprot: a hub for protein information. *Nucleic acids research*, 43(D1):D204–D212.

Marco A Valenzuela-Escárcega, Gus Hahn-Powell, Mihai Surdeanu, and Thomas Hicks. 2015. A domain-independent rule-based framework for event extraction. In *Proceedings of ACL-IJCNLP 2015 system demonstrations*, pages 127–132.

Bohui Zhang, Valentina Anita Carriero, Katrin Schreiberhuber, Stefani Tsaneva, Lucía Sánchez González, Jongmo Kim, and Jacopo de Berardinis. 2024. Ontochat: a framework for conversational ontology engineering using language models. In *European semantic web conference*, pages 102–121. Springer.

A Interface

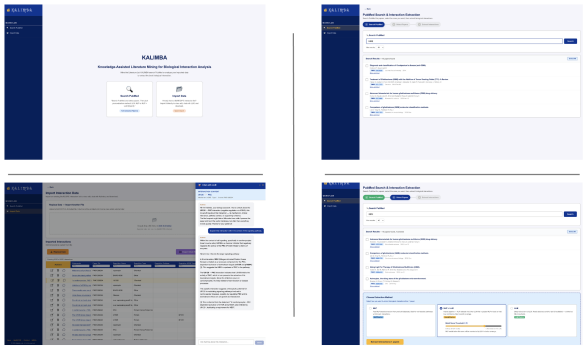

Figure 4: KALIMBA Interface – https://www.boheme.pitt.edu/kalimba