Articles

BioPE-RV: A New Paradigm for Biomedical Relation Extraction Integrating Prompt Engineering and Explicit Rule Verification

DOI:https://dx.doi.org/10.71373/yf07bz53
Submitted 21 December 2025
Accepted 19 January 2026
Published 24 April 2026

Yani Yang[1], Xingrao Li[1], Yongjian Zhou[2], Tingwei Lyu[2], Yuxiao He[2], Jianping Zhou[2],Xuanyu Pan[3], Yanling Hu[2,3,*]

Infectious diseases remain a major public health challenge threatening human health and socioeconomic development. Understanding viral infection mechanisms is therefore critical for effective disease control. Biomedical literature contains a vast number of regulatory relationships between genes and proteins, and accurately extracting these relationships is a fundamental prerequisite for constructing molecular regulatory networks and supporting downstream biological research.

In recent years, large language models have demonstrated considerable potential in biomedical information extraction tasks. However, when applied to regulatory relationship extraction, they still suffer from systematic limitations, including the misclassification of co-expression as regulation, excessive chain causal inference, and insufficient understanding of experimental contexts. These issues substantially compromise the reliability of extracted results.

To address these challenges, this study proposes BioPE-RV, a regulatory relationship extraction framework that integrates large language model–based candidate generation with explicit biological logic verification. The framework first leverages a large language model to generate candidate regulatory relationships from text and subsequently applies multi-level logical constraints—including entity validity, explicit regulatory agency, semantic consistency, causal chain integrity, and experimental context rationality—to iteratively verify and filter candidate results.

Experimental results on COVID-19-related biomedical literature demonstrate that combining generative models with interpretable biological logic constraints significantly improves both the reliability and reproducibility of relationship extraction.

Introduction

Infectious diseases remain a key factor influencing morbidity and mortality in contemporary society, with various emerging major infectious diseases frequently occurring and becoming increasingly severe[1]. Emerging major infectious diseases are characterized by sudden onset, high transmissibility, and uncertain prognosis, capable of causing large numbers of infections or deaths within a short period, severely impacting public health, social stability, and economic development[2]. Therefore, understanding viral infection mechanisms is crucial for suppressing viruses. We can uncover these mechanisms by studying host-pathogen protein-protein interaction networks and host-pathogen gene-gene interaction networks[3]. Network-based approaches effectively reveal disease-gene-protein associations, aiding in identifying key pathogenic genes, understanding underlying regulatory crosstalk driving cellular processes and diseases, and supporting drug design[4,5].Zhang et al. proposed an effective method[6] for studying gene regulatory systems. This approach constructs a comprehensive regulatory network by gathering information from published literature to analyze gene-protein regulatory relationships within the network. The method primarily relies on manual screening of literature from databases like PubMed, presenting three core challenges: First, the vast volume of literature makes manual screening difficult to cover the massive dataset; Second, it involves high subjectivity, as different researchers may apply varying criteria for defining "regulatory relationships," leading to inconsistent results. Third, entity recognition errors occur due to issues like multiple aliases for gene/protein names (e.g., "ACE2" also referred to as "ACEH") and non-standard abbreviations (e.g., 'TMPRSS2' incorrectly written as "TMPRSS").

Traditional automation methods have attempted to enhance efficiency through rule-based templates (such as keyword matching), statistical learning models, and deep learning models [7,8]. However, rule-based models suffer from poor scalability and require significant manual intervention; statistical learning models heavily depend on data quality, with issues like missing data, noise, and outliers potentially compromising predictive capabilities; Deep learning models demand substantial parameters and computational resources, with model performance heavily influenced by the quality and quantity of training data. They are also highly sensitive to parameter settings such as initial weights and learning rates. While pre-trained models like BioBERT[9], SciBERT[10], and PubMedBERT[11] significantly enhance entity recognition and relation extraction performance, they are fundamentally discriminative models with inherent drawbacks: (1) Heavy reliance on large-scale, high-quality supervised labeled data (fine-tuning). When applied to emerging domains like COVID-19, where data evolves rapidly, labeling costs are prohibitively high and inherently delayed; (2) Lack of explicit logical reasoning capabilities. Models primarily classify based on statistical features, struggling to handle complex cross-sentence causal chains; (3) Their 'black-box' nature prevents them from providing justification for judgments (e.g., evidence sentence extraction), diminishing result credibility in biomedical contexts.

Large Language Models (LLMs) offer new possibilities for addressing these challenges through their robust semantic understanding and contextual reasoning capabilities[12,13]. — Existing research demonstrates that LLMs outperform traditional models in biomedical entity recognition and relation

1. Guangxi Medical University School of Information and Management, 530021,Nanning, Guangxi, China.

2. Institute of Life Sciences, Guangxi Medical University, 530021,Nanning, Guangxi, China.

3. School of Basic Medical Sciences, Guangxi Medical University,530021, Nanning, Guangxi, China.

**Corresponding author:*

Yanling Hu, Email: ylhupost@163.com

extraction tasks[14,15]. Without requiring explicit feature engineering, LLMs can capture causal patterns, semantic dependencies, and contextual ontological knowledge. However, general-purpose LLMs still pose risks in biomedical contexts: semantic hallucinations, entity confusion, unstable reasoning, and lack of biological consistency. Therefore, there is an urgent need for a methodological framework that bridges the gap between "general language intelligence" and "specialized biological knowledge extraction."

Prompt engineering is considered a crucial approach to guiding general-purpose LLMs toward reliable scientific modeling[16,17]. Prompts serve as the core medium for interacting with generative AI. As an emerging discipline, prompt engineering optimizes prompts or commands that guide large language models' outputs to accomplish more complex tasks. Its essence lies in the practice of effectively interacting with AI systems to maximize their strengths[18,19,20]. Users can leverage prompt engineering to enhance large models' ability to handle complex problems and significantly improve their performance metrics (e.g., accuracy, logical coherence).

This study proposes the BioPE-RV (Biomedical Prompt Engineering and Rule Verification Framework) to optimize large language models' performance in biomedical knowledge extraction, as illustrated in Figure 1.

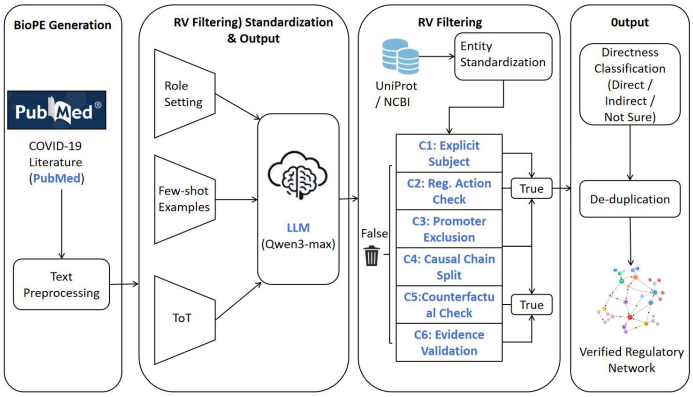

Figure 1. The Overview of BioPE-RV Framework.

Note: The framework consists of two coupled phases: (1) The Generative Phase, where the LLM is guided by Tree-of-Thought (ToT) prompts to extract candidate triples; and (2) The Verification Phase, where explicit biological logic rules (C1-C6) act as a funnel to filter out noise and hallucinations, ensuring the reliability of the constructed COVID-19 regulatory network.

To validate the effectiveness of the proposed BioPE-RV framework in real-world, large-scale, and complex biomedical scenarios, this paper employs COVID-19-related literature as the experimental substrate.

COVID-19 research literature exhibits the following typical characteristics: (1) massive volume and rapid updates; (2) highly diverse representations of gene-protein regulatory relationships; (3) abundant conditional, indirect, and multi-entity coupled regulatory descriptions. These features pose significant challenges to both traditional rule-based methods and supervised models.

This study randomly selected 1,500 COVID-19-related papers from PubMed that provided full text or abstracts, covering multiple research areas such as immune responses, inflammatory pathways, and viral invasion mechanisms. This formed a real-world test dataset characterized by high noise, high complexity, and strong semantic ambiguity, enabling a systematic evaluation of the proposed composite prompt engineering strategy's generalization capability and robustness in complex biomedical texts.

Research Methodology and Design

Overall Framework

This paper proposes a gene/protein regulatory relationship extraction method that integrates candidate generation via large language models with validation through biological logical constraints. It aims to reduce systematic misjudgments in biomedical texts caused by linguistic ambiguity and experimental context complexity.

The overall workflow comprises four main stages: (1) Text preprocessing and paragraph-level segmentation; (2) Candidate regulatory relationship generation using large language models; (3) Multi-layer logical constraint validation based on biological semantics; (4) Standardized output and deduplication of regulatory relationships. This framework positions large language models as candidate relationship generators rather than final decision-makers, thereby enhancing extraction reliability through explicit rule-based constraints.

Data Preparation

Literature Retrieval

In this study, the PubMed database was used for literature retrieval. The search query was formulated as: (COVID-19 OR SARS-CoV-2) AND (promote OR facilitate OR activate OR induce OR stimulate OR response OR recruit OR enrich OR inhibit OR suppress OR degrade OR block OR enhance OR repress OR co-chaperone OR coactivator OR corepressor OR significantly higher), yielding 262,468 results. This query was designed to more precisely identify articles that simultaneously mention COVID-19 (or related terms), gene/protein entities (including host proteins), and regulatory relationship cues—thereby aligning the corpus with downstream large language model (LLM) prompt-engineering needs such as trigger-term coverage, relation-oriented evidence retrieval, and high-recall candidate screening. Full texts were automatically downloaded when *free full text* was available; for non-open-access articles, only abstracts were collected.

Data Preprocessing

When analyzing the retrieved literature and the associated gene–protein data, a critical step is to filter and retain text segments that contain regulatory relation cues. This step improves the efficiency and task relevance of subsequent large language model (LLM) analysis by narrowing the input to relation-bearing evidence, thereby reducing irrelevant context and increasing the effective signal-to-noise ratio for downstream extraction. In this study, we implemented the filtering pipeline in Python. Specifically, we constructed regular-expression patterns to capture tense and derivational variants of regulatory verbs (e.g., *promote*, *promotes*, *promoted*, *promotion*), and covered 18 core categories of regulatory relations (e.g., activation, inhibition, promotion). An illustrative code snippet is provided below:

```
keywords_patterns = [
    r"promot(e|es|ed|ing|ion|ions)?",
    r"inhibit(s|ed|ing|ion|ions)?",
    r"significant(ly)?\s+higher"
    # Patterns for the remaining regulatory relation cues are omitted for brevity.
]
Pattern = re.compile(r"("+"|".join(keywords_patterns) + r")", re.IGNORECASE)
```

Candidate Regulatory Relation Generation

During the candidate generation stage, we leverage a large language model (LLM) to perform inference over paragraph-level biomedical text and to extract potential gene/protein regulatory relationships. Specifically, we adopt a "Few-shot Prompting + Tree-of-Thought (ToT) "strategy to guide the Qwen3-max model in relation extraction. Few-shot prompting refers to exploiting the LLM's in-context learning (ICL) capability, whereby a new task can be addressed by providing a small number of representative examples within the prompt.

Tree of Thought (ToT) is an LLM reasoning framework designed for complex tasks. Its central idea is to decompose a complex problem into hierarchical and iteratively refinable subproblems, explore multiple reasoning branches in a tree-structured manner, and ultimately select the most suitable solution path[21]. Prior studies have shown that Chain-of-Thought (CoT) reasoning can substantially enhance LLM performance on complex reasoning tasks[22]. Compared with directly fine-tuning open-source LLMs, the "Few-shot Prompting + ToT" approach significantly reduces annotation and time costs, and provides a practical foundation for subsequent construction of a knowledge graph or knowledge base[23].

In addition, we constrain the model outputs to a structured JSON format containing the regulator, the regulated entity (target), the regulation type, the directness of regulation, and the supporting evidence, thereby facilitating downstream manual verification.

In this study, we employed a five-example few-shot prompting strategy. The five demonstration examples were curated from relevant literature by two biomedical domain experts, providing high-quality exemplars to guide the LLM in learning the targeted extraction schema and decision patterns. After multiple rounds of experimentation, we developed a prompt template with strong empirical performance, as shown in Table 1 (see Supplementary Note S1 for details).

Table 1 Prompt Template.

Name	Description
Character Design	You are a meticulous biomedical research assistant specialized in extracting "molecular regulatory relationships from scientific literature.
Task Description	You use chain-of-thought reasoning to ensure extraction accuracy, and only extract strictly defined regulatory relationships. You ALWAYS provide COMPLETE evidence sentences that contain both the regulator and target gene/protein names AND the relation word. You do NOT extract signaling pathway relationships. You use 'direct', 'indirect', or 'not sure' for directness based on the available information. You REJECT any evidence that doesn't explicitly demonstrate the regulatory relationship.
Example of Few-shot	"input": "leading to the inability of cGAS-stimulator …signaling.", "output": [{ "regulator": "HAT1", "target":"KPNA2", "relation":"promoted",

"directness":"indirection",

"evidence":"HBV-elevated HAT1 promoted the expression of KPNA2 through modulating acetylation of H4K5 and H4K12 in the system, resulting in nuclear translocation of cGAS."

}]

TOT

1. Entity Identification & Validation:

- Identify all potential gene/protein names in the text
- Verify these names exist in the standard name database
- Exclude non-gene/protein entities (antibodies, viruses, experimental methods, etc.)

2. Relation Detection & Analysis:

- Locate regulatory relationship verbs in the text
- Confirm these verbs are in the predefined regulatory relation list
- Analyze the direction of regulation (who regulates whom)

3. Directness Assessment:

- Determine if it's a direct, indirect, or not sure interaction
- Direct: two molecules interact directly (physical binding, direct modification)
- Indirect: mediated through signaling pathways or other molecules
- Not sure: when the text does not provide clear information about directness

4. Evidence Extraction - CRITICAL STEP:

- Find COMPLETE sentences that support the relationship
- Evidence MUST contain BOTH the regulator and target gene/protein names
- Evidence MUST contain the relation word that describes the regulatory action
- Ensure evidence sentences are grammatically complete and meaningful
- Do NOT use truncated or incomplete sentences
- The evidence sentence MUST explicitly show the relationship between the two entities

5. Result Validation:

- Confirm all extracted relationships match predefined relation types
- Ensure gene/protein names are accurate
- Check that relationships are explicit rather than speculative
- VERIFY that evidence sentences contain both regulator and target AND the relation word
- REJECT any evidence that doesn't clearly demonstrate the regulatory relationship

Output	Format	"regulator": "regulator name",
Constraints		"target": "target name",
		"relation": "relation type",
		"directness": "direct/indirect/not sure",
		"evidence": "COMPLETE evidence sentence containing both gene names AND the relation word"

To mitigate semantic drift and over-generalization, the model is restricted to output only a predefined set of regulatory relation types, namely: promote, facilitate, activate, induce, stimulate, response, recruit, enrich, inhibit, suppress, degrade, block, enhance, repress, co-chaperone, coactivator, corepressor, and significantly higher[24]. Any relation not included in this set is directly filtered out in subsequent stages.

It should be noted that the outputs generated at this stage are candidates only and should not be interpreted as confirmed biological regulatory facts.

Gene and Protein Entity Validation

To prevent non-biological entities from being misidentified as regulatory actors, we constructed a unified, standardized entity lexicon based on UniProt and NCBI resources (see Supplementary Data). Given the naming characteristics of different fields, we adopted differentiated tokenization and normalization strategies to improve coverage of common aliases and abbreviation forms.

A candidate relation is passed to the subsequent logical validation stage only if both the regulator and the regulated entity (target) can be successfully matched to entries in the standardized lexicon.

Biological Logic–Constrained Validation

In the candidate-relation validation stage, we introduce a set of explicit biological logic constraints (C1–C6) (see Supplementary Algorithm S1) to determine whether a candidate constitutes a biologically plausible regulatory relation. These constraints were derived from a systematic analysis of a large number of failure cases and were designed to target common error patterns exhibited by large language models in regulatory relation extraction.

To ensure the effectiveness and generalizability of both the prompt template and the logic rules (C1–C6), we constructed an independent development set. This set comprises 500 randomly sampled COVID-19–related articles retrieved in this study (excluded from the final test set) and manually annotated by two domain experts. Both the prompting strategy and the logic constraints were iteratively refined based on error analysis on the development set. Only after the F1 score on the development set stabilized were the finalized prompts and rules applied to the final test set for evaluation. The biological logic constraints (C1–C6) are listed as follow.

C1 Syntactic Explicitness and Entity Context Constraint

(1) SVO order validation. The evidence sentence must follow the syntactic order “Regulator … Relation word … Target.” The algorithm retrieves the character indices of the entities and the relation cue within the sentence and enforces Index(Regulator) < Index(Relation) < Index(Target). This constraint removes cases arising from passive-voice parsing errors or non-directional co-occurrence statements.

(2) Cell-type context exclusion. To mitigate entity confusion frequently observed in immunology texts, we use regular expressions (e.g., r"\bcd4\+?\s*t\s*cells?\b") to identify and exclude entities occurring in cell-type contexts (e.g., misrecognizing “CD4+ T cells” as the gene “CD4”), thereby preventing cell-level descriptions from being misinterpreted as molecular regulation.

C2 Regulatory Semantic Consistency

Only predefined regulatory actions are accepted (see the REGULATION_RELATIONS dictionary for the controlled vocabulary). The algorithm verifies whether the evidence sentence contains the canonical form of the relation word or its morphological variants (e.g., *promote* → *promoted*, *promotion*). If the “relation type” generated by the model cannot be mapped back to the controlled vocabulary, or if the evidence sentence lacks the corresponding action predicate, the relation is deemed invalid.

C3 Reporter System Exclusion

Biomedical articles frequently use luciferase or reporter-gene assays to characterize promoter activity, which is not equivalent to regulation of endogenous protein expression. This rule maintains a blacklist of terms (e.g., *luciferase*, *reporter*, *promoter-driven*, *construct*). If any blacklist term appears in the evidence sentence, the candidate relation is automatically rejected to avoid misreading experimental readouts as biological facts.

C4 Causal Chain Decomposition

For nested constructions such as “Factor X induces A to inhibit B,” LLMs may incorrectly infer a direct relation “X inhibits B.” This rule uses a regular expression (e.g., r"\binduces\b.+\bto\b.+\binhibit\b") to capture such causal-chain patterns. If the extracted relation skips an intermediate mediator (i.e., directly links X and B), it is considered a violation of direct regulatory logic and is removed.

C5 Counterfactual / Loss-of-Function Constraint

In descriptions of knockdown, blockade, or siRNA perturbation experiments, the reported outcome is often counterfactual (e.g., “Knockdown of Gene A increased Gene B”). To prevent

confusion between experimental intervention and biological function, this rule detects whether the regulator appears as the target of a perturbation operation (e.g., matching *knockdown of <Regulator>*). If a candidate relation falls under a loss-of-function context and is not correctly interpreted by the model as negative regulation or an indirect effect, it is conservatively filtered to reduce false positives.

C6 Recruitment Specificity and Evidence Completeness

(1) Recruitment specificity. For the recruit relation, the algorithm checks whether the target belongs to a set of subcellular localization terms (e.g., *nucleus*, *cytoplasm*, *promoter*). If the target is recognized as a location rather than a molecular entity, the relation is treated as a semantic error and removed.

(2) Evidence completeness. The evidence must be a complete natural-language sentence (length > 10 characters and word count > 5) and must explicitly contain the Regulator, Target, and the relation cue.

Regulation Directness Determination

Based on whether the evidence sentence describes direct physical interaction, transcriptional regulation, or enzymatic modification, each regulatory relation is labeled as direct, indirect, or not sure[24].

Result Deduplication

To ensure consistency of the extracted results, we performed deduplication of regulatory relations using a composite key comprising the regulator, regulated entity (target), regulation type, and the hash value of the evidence sentence.

Experimental Design and Results Analysis

Experimental Environment

All experiments in this study were conducted on a Linux server equipped with a single NVIDIA A800-SXM4-80GB GPU. The hardware configuration includes a 128-core AMD64 (x86_64) CPU and 24 GB of system memory. The software stack is built on Ubuntu 20.04 LTS, with NVIDIA Driver 550, CUDA 12.4, and the PyTorch 2.3 deep learning framework.

We adopted an API-based inference paradigm for LLM reasoning. The large language model used was Tongyi Qianwen (Qwen3-max), accessed via the Alibaba Cloud DashScope platform. The decoding hyperparameters were set to temp--erature = 0.1 and max tokens = 3000. To handle network instability and API rate/throughput constraints, we implemented a robust retry mechanism with up to five retries using an exponential backoff strategy. In addition, multithreaded concurrency (max concurrent requests = 5) was employed to improve the throughput of large-scale text processing.

Experimental Data and Evaluation Metrics

In this study, we randomly selected 1,500 relevant articles retrieved from PubMed. Using paragraphs as the minimal processing unit, we performed length optimization on the text and then fed the resulting paragraphs into the model to generate candidate regulatory relations. The experimental dataset was constructed from publicly available biomedical literature on PubMed, spanning research areas such as immunology, molecular biology, and signal transduction. To ensure the credibility of the evaluation, all paragraphs were independently annotated by two experts with backgrounds in molecular biology, and inter-annotator agreement was assessed using Cohen's kappa ($\kappa = 0.86$) to confirm labeling consistency (see Supplementary Table S1).

For performance evaluation, we used precision, recall, and F1-score as the primary metrics. Precision measures the proportion of true regulatory relations among the extracted results, while recall evaluates the coverage of regulatory facts captured by the method.

Precision (P) = (Number of correctly identified regulatory relations) / (Total number of regulatory relations output by the model)

Recall (R) = (Number of correctly identified regulatory relations) / (Total number of regulatory relations in expert annotations)

$F1\text{-score} = 2 \times P \times R / (P + R)$

Given the semantic complexity and contextual dependence of regulatory relations, we further conducted an error-type–driven qualitative analysis on top of the quantitative metrics to compare how different methods perform under specific semantic scenarios.

Ablation Analysis of Biological Logic Constraints

To quantify the contribution of each biological logic constraint to overall performance, we conducted an ablation study. Specifically, we removed logic rules C1–C6 one at a time and examined the resulting changes in F1-score (as reported in Table 2).

Table 2 Ablation study on biological logic constraints (C1–C6)

Model Variant	Precision (%)	Recall (%)	F1-score(%)	ΔF1
BioPE-RV	**86.29%**	**81.19%**	**83.66%**	-
w/o C1	38.96%	73.17%	50.85%	32.81%
w/o C5	60.00%	80.49%	68.75%	14.91%
w/o C6	65.96%	75.61%	70.46%	13.2%
w/o C2	68.89%	75.61%	72.09%	11.57%
w/o C3	68.63%	85.37%	76.09%	7.57%
w/o C4	70.83%	82.93%	76.40%	7.26%

Note: BioPE-RV denotes the full method incorporating all six layers of biological logic constraints. w/o indicates that the corresponding logic rule (C1–C6) is removed during the inference/validation process. ΔF1 represents the absolute decrease in F1-score (in percentage points) relative to the full model. The results show that C1 (Syntactic Explicitness and Entity Context Constraint) has the largest impact: removing C1 reduces the F1-score by 32.81 percentage points, indicating that this rule plays a decisive role in suppressing LLM hallucinations and preventing false-positive relations.

Removing C1 leads to a collapse in performance (ΔF1 = 32.81 percentage points), with precision dropping to only 38.96%. This confirms that biomedical text contains substantial confounders (e.g., CD4 in "CD4+ T cells"); without C1's explicit context filtering, the LLM is prone to severe entity-level hallucinations and false-positive relations.

Removing C5 decreases the F1-score by 14.91 percentage points, indicating that negative perturbation descriptions such as "knockdown/siRNA" are prevalent in the literature and are easily misinterpreted by the model. C5 effectively separates such experimental interventions from genuine biological regulatory functions, thereby reducing spurious inferences.

Notably, in the ablation study, removing C3 (reporter/promoter-system exclusion) and C4 (causal-chain decomposition) yields higher recall—85.37% and 82.93%, respectively—than the full BioPE-RV model (81.19%). This phenomenon reflects an inherent trade-off introduced by rule-based constraints between noise suppression and information preservation:

(1) The "precision-for-reliability" cost of C3

Biomedical articles frequently rely on reporter systems (e.g., luciferase assays) to quantify regulatory activity. LLMs tend to directly equate such experimental readouts with regulation of endogenous protein expression. When C3 is removed, the model admits a large number of reporter-related statements. Although this increases the number of captured candidate signals (reflected in higher recall), it also introduces substantial false positives by misclassifying experimental assay components as biological entities, causing precision to drop sharply from 86.29% to 68.63%. This demonstrates that C3 functions as a necessary high-selectivity filter: it sacrifices a small amount of recall on borderline cases to ensure that relations retained for downstream knowledge-base integration correspond to biologically grounded facts rather than assay-induced artifacts.

(2) C4 as a safeguard against over-inference

For causal-chain constructions such as "A induces B to inhibit C," models not constrained by C4 tend to infer a direct relation "A inhibits C." While such cross-level indirect links may be accepted by some annotators in certain contexts (leading to a slight increase in recall for *w/o C4*), they represent a logical oversimplification from a rigorous molecular biology perspective. By enforcing causal-chain decomposition, BioPE-RV uses C4 to prevent these shortcut inferences. Although this may marginally reduce coverage of implicit relations, it substantially improves the directness and correctness of extracted relations, yielding a marked gain in precision (+15.46 percentage points).

Overall Performance Comparison Across Methods

To systematically evaluate the effectiveness of different extraction strategies for biomedical regulatory relation extraction, we designed four comparative baselines:

(1) Rule-based

This approach extracts relations using manually defined regular-expression templates and keyword-matching rules. It does not involve LLM inference and relies solely on explicit heuristic rules to determine whether a regulatory relation is present.

(2) Chain-of-Thought (CoT)

This method relies only on the LLM (Qwen3-max) and its semantic understanding and contextual reasoning capabilities. The model is guided by Chain-of-Thought prompting to extract relations step by step, but no post-processing or logical validation using the proposed constraints (C1–C6) is applied.

(3) Few-shot + ToT

This method uses few-shot demonstrations and Tree-of-Thought–style reasoning to guide the LLM to generate candidate regulatory relations, without introducing any subsequent rule-based filtering or biological logic validation.

(4) BioPE-RV

Building on candidate generation via Few-shot + CoT, our method incorporates explicit biological logic constraints to perform multi-stage validation and filtering of candidate relations, reducing common errors such as co-occurrence–driven false positives and over-inference in causal chains.

Notably, we did not include supervised fine-tuning–based models (e.g., BioBERT, PubMedBERT) in the primary comparison. Supervised models typically require thousands of high-quality human-annotated instances to converge and are therefore data-hungry and highly resource-dependent. In contrast, this study targets scenarios in which annotated data are scarce—such as the early stage of emerging disease outbreaks (e.g., COVID-19) or highly specialized subdomains—where the problem is inherently zero-/few-shot. Accordingly, comparing BioPE-RV with other training-free baselines, including CoT-based prompting and rule-based extraction, better highlights the generalization capability and practical utility of our framework under low-resource conditions.

All extracted relations produced by the compared methods were manually validated by two experts with biomedical backgrounds. During validation, the experts independently determined whether each extracted relation constituted a true biological regulatory relationship and whether the associated evidence sentence sufficiently supported the claim. Disagreements were resolved through discussion to reach a final consensus. The resulting manual verification outcomes were treated as the final gold standard for computing precision, recall, and F1-score. The overall results are reported in Table 3 (see Supplementary Table S1 for details).

Table 3 Overall performance comparison of different regulatory relation extraction methods

Prompting strategy	P	R	F1-score
Rule-based	20.9%	4.84%	7.86%
COT	21.5%	75.43%	33.46%
Few-shot+ToT	43.39%	51.61%	47.14%
BioPE-RV	**86.29%**	**81.19%**	**83.66%**

Note: Bolded values indicate the best performance in each column. Rule-based refers to a traditional keyword-matching approach. CoT denotes a baseline strategy using zero-shot Chain-of-Thought prompting without any post-processing rules. Few-shot + ToT uses only few-shot demonstrations and Tree-of-Thought prompting, without

explicit logic-based validation. BioPE-RV is the proposed extraction framework that integrates explicit biological logic constraints. P denotes precision, and R denotes recall.

The results indicate that the rule-based method is inherently limited for this task. Relying solely on handcrafted rules and keyword matching makes it difficult to cover the highly diverse and context-dependent ways in which regulatory relations are expressed in biomedical text. In particular, for cases involving long-range dependencies, indirect regulation, and complex syntactic structures, the rule-based approach tends to miss a substantial number of true relations.

The CoT-based strategy achieves a recall of 75.43%, substantially outperforming the rule-based method; however, its precision is extremely low (21.5%). This suggests that, in the absence of explicit constraints, the LLM tends to over-associate, incorrectly labeling many non-regulatory co-occurrence statements (e.g., cell-type contexts and experimental conditions) as regulatory relations.

After introducing few-shot demonstrations and ToT-style reasoning prompts, overall performance improves markedly, indicating that with in-context examples and multi-step reasoning guidance, the LLM can identify some implicit or non-explicit regulatory relations. Nevertheless, without biologically grounded logic constraints, this approach still produces a non-negligible fraction of semantic-association or co-expression–driven false positives, which limits further gains in precision.

The proposed hybrid extraction framework (BioPE-RV) achieves the best performance across all reported metrics. Building on candidate generation via few-shot prompting and reasoning, BioPE-RV incorporates explicit biological semantic constraints and a relation-validation mechanism, which substantially reduces false positives introduced by unconstrained LLM generation while preserving strong coverage of complex regulatory expressions. Overall, the method attains a more favorable balance between precision and recall, highlighting the benefits of synergistically combining generative modeling with rule-based biological constraints.

Error-Type–Based Analysis of Regulatory Relation Extraction

To systematically analyze the performance differences across methods, we manually reviewed the erroneous outputs produced by the Few-shot + ToT approach. Based on representative failure cases observed in the results, we summarize the major error types as follows.

Co-expression Misinterpreted as Regulation

In biomedical literature, multiple genes or proteins are often reported to be upregulated or downregulated in parallel under the same condition. The Few-shot approach frequently misparses such coordinated expressions as regulatory relations. Our method introduces an explicit co-expression exclusion rule to prevent conditional co-occurrence from being mistaken for regulatory facts, thereby substantially reducing the incidence of this error type.

Confusion Between State Descriptions and Regulatory Actions

Biomedical texts contain numerous state descriptions such as "activation of X" and "phosphorylation of Y." Without constraints, the Few-shot approach tends to equate molecular state changes with regulatory actions on downstream targets. In contrast, our method requires explicit mention of a regulator and an action predicate in the evidence sentence, effectively distinguishing state descriptions from true regulatory relations.

Confusion Between Promoter Binding and Expression Regulation

Promoter binding events and reporter-gene assay results do not necessarily imply genuine gene regulation. However, in such contexts, the Few-shot approach often directly outputs relations such as *promote* or *activate*. By introducing a promoter/reporter-system exclusion rule, our method accepts a relation only when the text explicitly indicates transcriptional or expression-level changes, thereby reducing this class of false positives.

Subject Mismatch in Causal-Chain Descriptions

In signaling and regulatory pathway narratives, causal chains are frequently described in multi-step structures. The Few-shot approach often ignores intermediate regulatory nodes and directly infers cross-level relations. Our method applies a causal-chain decomposition rule to retain only directly stated relations, avoiding overextended causal inference.

Over-inference From Counterfactual Experimental Outcomes

In knockdown or inhibition experiments, observed expression changes are often indirect effects. The Few-shot approach may incorrectly interpret such counterfactual outcomes as direct regulatory conclusions. Our method imposes a counterfactual/loss-of-function constraint and requires explicit positive or negative regulatory statements, thereby mitigating this error type.

Analysis of Error Suppression Effects

Combining the above error-type analysis with the overall experimental results, we observe that the proposed method consistently suppresses multiple high-frequency error modes. The precision gains are particularly pronounced in complex contexts involving co-expression statements, promoter/reporter experiments, and causal-chain inference. These findings suggest that explicitly incorporating biological validity criteria into the extraction pipeline can substantially improve the reliability of regulatory relation extraction while maintaining a reasonable level of recall.

Discussion

In recent years, large language models (LLMs) have demonstrated strong semantic understanding in biomedical text mining, showing clear advantages in parsing complex syntactic structures and performing cross-sentence reasoning. However, our experimental results indicate that relying solely on LLMs for relation extraction still entails systematic risks of

misclassification, largely due to the prevalence of implicit causality, dependence on experimental conditions, and terminological polysemy in biomedical discourse.

Through a systematic analysis of failure cases, we observe that, in the absence of explicit constraints, LLMs tend to misinterpret expression changes, state descriptions, or condition-dependent effects as direct molecular regulatory relationships. This tendency is particularly pronounced in co-expression narratives, promoter/reporter assay contexts, and causal-chain descriptions at the pathway level.

To address these issues, we propose a hybrid extraction framework that introduces explicit biological logic constraints. In this design, the LLM is restricted to the role of a candidate relation generator, while final determination of regulatory relations is performed by a rule-driven semantic validation module. The results demonstrate that this strategy can substantially improve the precision of regulatory relation extraction without significantly sacrificing recall.

Importantly, the introduced logic constraints are not ad hoc heuristics tailored to a specific dataset or writing style; rather, they are derived from a systematic abstraction of common biomedical reporting patterns and recurrent error types. Consequently, the approach is expected to exhibit good transferability in principle and can be applied to different research domains and textual sources.

Nevertheless, the proposed method has limitations. For example, in highly metaphorical descriptions or statements that heavily rely on domain background knowledge, strict logic constraints may filter out some potentially valid regulatory relations. In addition, the current rule set does not yet cover all semantic variants of regulation, leaving room for further expansion and more automated rule induction.

Future work may preserve the biological-logic–constrained framework while exploring automatic induction of constraints and deeper integration with structured biomedical knowledge bases, with the aim of further improving coverage and generalization in regulatory relation extraction.

Conclusion

This study proposes BioPE-RV, a biomedical regulatory relationship extraction framework that integrates large language model–based candidate generation with explicit biological logic verification. By explicitly separating candidate generation from final relationship validation, the framework addresses several persistent challenges in biomedical relation extraction, including co-expression misclassification, overextended chain causal inference, and misinterpretation of experimental contexts.

Experimental results on large-scale COVID-19-related biomedical literature demonstrate that BioPE-RV significantly outperforms both traditional rule-based methods and few-shot + ToT prompting approaches across precision, recall, and F1-score. The improvement is primarily attributed to the introduction of biologically grounded logic constraints, which effectively suppress hallucination-induced false positives while preserving coverage of complex regulatory expressions.

Further error-type analysis shows that BioPE-RV exhibits strong robustness in challenging semantic scenarios such as co-expression descriptions, promoter and reporter system contexts, cascade causal statements, and counterfactual experimental settings. These results indicate that embedding explicit biological reasoning into the relation extraction pipeline is an effective strategy for improving the reliability and reproducibility of large language model–based biomedical information extraction.

Overall, this work demonstrates that large language models, when guided by carefully designed prompt strategies and constrained by explicit domain logic, can serve as reliable tools for complex biomedical knowledge extraction. Future research will focus on extending the biological logic constraint set, exploring automated rule induction, and integrating structured biological knowledge bases to further enhance generalization and scalability.

Reference

1. Standing up to infectious disease[J]. Nat Microbiol, 2019, 4(1): 1.
2. Wei W, Liu Y, Zhou N, et al. Constructing an emergency preparedness evaluation index system for public use during major emerging infectious disease outbreaks: a Delphi study[J]. BMC Public Health, 2023, 23(1): 1109.
3. Khorsand B, Savadi A, Naghibzadeh M. Comprehensive host-pathogen protein-protein interaction network analysis[J]. BMC Bioinformatics, 2020, 21(1): 400.
4. Kim Y, Park JH, Cho YR. Network-Based Approaches for Disease-Gene Association Prediction Using Protein-Protein Interaction Networks[J]. Int J Mol Sci, 2022, 23(13): 7411.
5. Kim D, Tran A, Kim HJ, et al. Gene regulatory network reconstruction: harnessing the power of single-cell multi-omic data[J]. NPJ Syst Biol Appl, 2023, 9(1): 51.
6. Zhang R, Shah MV, Yang J, et al. Network model of survival signaling in large granular lymphocyte leukemia[J]. Proc Natl Acad Sci U S A, 2008, 105(42): 16308-16313.
7. Puccetti G, Giordano V, Spada I, et al. Technology identification from patent texts: a novel named entity recognition method[J]. Technol Forecast Soc Change, 2023, 186: 122160.
8. Li J, Sun A, Han J, et al. A survey on deep learning for named entity recognition[J]. IEEE Trans Knowl Data Eng, 2020, 34(1): 50-70.
9. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining[J]. Bioinformatics, 2020, 36(4): 1234-1240.
10. Cai X, Liu S, Yang L, et al. COVIDSum: A linguistically enriched SciBERT-based summarization model for COVID-19 scientific papers[J]. J Biomed Inform, 2022, 127: 103999.
11. Ying Z, Guanghui X, Xiaoying L, et al. A Multi-Label Text Classifier at Publication Level Based on “PubMedBERT + TextRNN” for Cancer Literature[J]. Stud Health Technol Inform, 2024, 316: 374-375.
12. Zhang Y, Guo S, Cai C, et al. Ponzi scheme detection in smart contracts based on reverse engineering and large language models[J]. Journal of Frontiers of Computer Science and Technology, 2025: 1-14.

13. Tang Y, Tang X, Gao C, et al. Characteristics of interactive hydrological model parameter optimization based on large language models: A case study of HBV and VIC models[J]. Advances in Water Science, 2025: 1-13.

14. Qin M, Feng L, Lu J, et al. ZeroTuneBio NER: A three-stage framework for zero-shot and zero-tuning biomedical entity extraction using large language models and prompt engineering[J]. Comput Methods Programs Biomed, 2025, 272: 109070.

15. Li Y, Viswaroopan D, He W, et al. Improving entity recognition using ensembles of deep learning and fine-tuned large language models: A case study on adverse event extraction from VAERS and social media[J]. J Biomed Inform, 2025, 163: 104789.

16. Ye Q, et al. Prompt engineering a prompt engineer[J]. arXiv preprint, 2023, arXiv:2311.05661.

17. Marvin G, et al. Prompt engineering in large language models[J]. International Conference on Data Intelligence and Cognitive Informatics, 2023: 1-? (Springer).

18. Zhang Y, Wang M. Research on optimization path of environmental map evaluation based on prompts and multimodal large model selection[J]. Geography Teaching, 2025, (13): 20-24.

19. Venerito V, Bilgin E, Iannone F, Kiraz S. AI am a rheumatologist: a practical primer to large language models for rheumatologists[J]. Rheumatology (Oxford), 2023, 62(10): 3256-3260.

20. Meskó B. Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial[J]. J Med Internet Res, 2023, 25: e50638.

21. Wang D, Lu F, Zhang B, et al. A survey of prompt engineering in large language models[J]. Computer Systems & Applications, 2025, 34(01): 1-10.

22. Wei J, Wang XZ, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models[J]. arXiv preprint, 2022, arXiv:2201.11903.

23. Gilardi F, Alizadeh M, Kubli M. ChatGPT outperforms crowd workers for text-annotation tasks[J]. Proc Natl Acad Sci U S A, 2023, 120(30): e2305016120.

24. Hu Y, Gu Y, Wang H, et al. Integrated network model provides new insights into castration-resistant prostate cancer[J]. Sci Rep, 2015, 5: 17280.

Author Contributions

Yani Yang conceived the study, designed the methodology, conducted the experiments, and drafted the manuscript. Xingrao Li and Yongjian Zhou contributed to data preprocessing and experimental implementation. Tingwei Lyu, Yuxiao He, and Jianping Zhou participated in result analysis and interpretation. Xuanyu Pan provided domain-specific biological expertise and assisted in logical rule design. Yanling Hu supervised the study, provided critical revisions, and approved the final manuscript. All authors reviewed and approved the final version of the manuscript.

Funding

This work was supported by the National Key Research and Development Program of China (Grant No. 2023YFC05400), the Guangxi Key Research and Development Program (Grant No. GuiKe 2023AB04032), and the Key Project of the Guangxi Natural Science Foundation (Grant No. 2024GXNSFDA999002), Guangxi College of Artificial Intelligence Postgraduate Innovation Project.

Ethics Statement

This study did not involve human participants, animals, or personal data. All data used in this research were obtained from publicly available biomedical literature. Therefore, ethical approval was not required for this study.

Consent Comments

Not applicable.

Data Availability Statement

The data supporting the findings of this study are derived from publicly available biomedical literature in the PubMed database. All extracted regulatory relationship data and analysis results are available from the corresponding author upon reasonable request.

Supplementary Materials

Additional supplementary information is available for download and review in the supplementary information section located on the right-hand side of this article's HTML page.

Supplementary Note S1: Detailed prompt templates for the LLM.
Supplementary Algorithm S1: Python implementation of the biological logic rules (C1–C6).
Supplementary Table S1: Detailed statistics of the experimental dataset.
Supplementary Data: The gene/protein dictionary.