On the extraction of meaningful RNA interactions from Scientific Publications through LLMs and SPIRES

Emanuele Cavalleri[1], Marco Mesiti[1]

[1]AnacletoLab - Dipartimento di Informatica, Università degli Studi di Milano, Via Celoria 18, Milano

Abstract

Knowledge graphs (KGs) are useful tools to uniformly represent and integrate heterogeneous information about a domain of interest. However, they are inherently incomplete; therefore, new facts should be introduced by extracting them from structured and unstructured data sources. Starting from RNA-KG, the first KG tailored for representing different kinds of RNA molecules that we recently developed, in this paper we evaluate the use of SPIRES for extracting interactions among bio-entities involving RNA molecules from scientific papers guided by the RNA-KG schema. SPIRES is a general-purpose knowledge extraction system for mining information conforming to a specified schema. A customized prompt is generated and submitted to a Large Language Model (LLM) along with a text to extract a set of RDF triples adhering to the schema constraints. The experiments show a high accuracy in extracting interactions from the scientific literature.

Keywords

RNA-based technologies, Knowledge Graphs, RNA-drug discovery, Large Language Models

1. Introduction

The "RNA world" represents a novel frontier for the study of fundamental biological processes and human diseases and is paving the way for the development of new drugs tailored to the patient's biomolecular characteristics. Although scientific data about coding and non-coding RNA molecules are continuously produced and made available from public repositories, they are scattered across different databases and in the scientific literature. A centralized, uniform, and semantically consistent representation of the knowledge on RNA is still lacking. We have recently constructed RNA-KG [1], a knowledge graph integrating biological knowledge about RNA molecules with their functional relationships with genes, proteins, and chemicals and biomedical ontological concepts. RNA-KG includes around 600K nodes and 9M RDF triples representing reliable interactions involving RNA molecules and related biomedical concepts extracted from more than 50 public data sources according to 11 bio-ontologies. RNA-KG is coupled with a meta-graph representing all the possible interactions involving RNA molecules.

SPIRES (Structured Prompt Interrogation and Recursive Extraction of Semantics) [2] is a recently proposed approach to information extraction that exploits Large Language Models (LLMs) [3] to identify instances of a knowledge schema expressed in terms of LinkML [4] starting from plain texts. By identifying and extracting relevant information from an input text, it adopts zero-shot learning to identify and extract relevant entities and relationships among them, which are then normalized and grounded through ontologies and vocabularies. SPIRES is a general-purpose approach that can be used across a variety of domains and does not require specific training/tuning on the considered domain. SPIRES adopts an engineering approach for creating prompts for interacting with an LLM (like GPT [5], Llama 2 [6], Mistral [7], and Zephyr [8]) to improve the quality of the generated responses [9]. In this way, technical challenges for generative AI (e.g., constructing comprehensive real-world knowledge and improving the accuracy of automated responses) can be addressed.

In this paper, we discuss the initial experimental results that we obtained by applying SPIRES in the extraction of interactions among bio-entities involving RNA molecules in the context of the PNRR project "Gene Therapy and Drugs based on RNA Technology". The purpose of the experiments is to show the level of accuracy of the system in extracting interactions from the scientific literature and investigate the possibility of combining RNA-KG with LLMs. Note that the extraction of interactions involving RNA molecules is particularly challenging for two reasons. First, a well-recognized ontology for characterizing non-coding RNA molecules is still lacking, and then different identifiers for representing the same bio-entity are adopted. Even if a more systematic evaluation should be conducted, the initial results are very encouraging.

The paper is structured as follows. Section 2 describes the SPIRES approach and related approaches that integrate LLMs with knowledge data. Section 3 presents the LinkML schema that we have developed for interacting with SPIRES. Section 4 describes the experimental results, while Section 5 reports concluding remarks.

Published in the Proceedings of the Workshops of the EDBT/ICDT 2024 Joint Conference (March 25-28, 2024), Paestum, Italy

emanuele.cavalleri@unimi.it (E. Cavalleri);
marco.mesiti@unimi.it (M. Mesiti)
0000-0003-1973-5712 (E. Cavalleri); 0000-0002-9421-8566 (M. Mesiti)

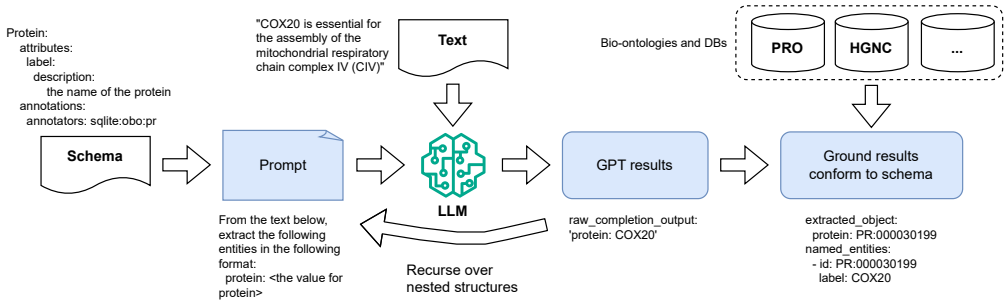


Figure 1: SPIRES workflow.

2. SPIRES and Related Work

The population of a KG by extracting triples from unstructured texts is an interesting research activity and the advent of LLMs has boosted the interpretation of highly technical languages as shown on question-answering benchmarks [10]. However, these techniques have shown different limitations, such as generating incorrect statements due to hallucinations [11] and insensitivity to negations [12], that cannot be tolerated in sensitive domains like precision medicine. SPIRES adopts: *i*) the knowledge schema of a specific domain for the generation of prompts for reducing these drawbacks; and *ii*) bio-ontologies for enhancing the quality of the produced information.

Figure 1 outlines the SPIRES workflow. SPIRES requires the specification of the knowledge schema expressed in LinkML [4] to guide the system in the extraction of knowledge. A LinkML schema contains the classes of entities and relationships among them within the specified domain. Classes can also include attributes (e.g., name, type, and list of synonyms) to enrich entity description. The LinkML schema is automatically processed to generate a list of prompts through which SPIRES interacts with a LLM (e.g., GPT3, GPT4, Llama 2, Mistral, and Zephyr). Each prompt of the list is submitted to the LLM for collecting information that is exploited for completing the following prompt by eventually considering the bio-ontologies (e.g., for changing a protein symbol with the corresponding identifier in an ontology). This refinement recursive process improves the quality of the information gathered through the LLM.

Example 1. *Suppose we wish to extract proteins from a text. A LinkML expression can be generated for describing the class* `Protein` *with its properties and the adopted identification scheme (See Figure 1). A prompt is then generated for this class and used for extracting proteins. However, the result obtained by ChatGPT alone (in this case* `COX20`*) is not compliant with the* `Protein` *class structure. Therefore, SPIRES exploits bio-ontologies (e.g. PRotein Ontology – PRO [13]) to obtain an adequate result.*

Furthermore, in case relationships are identified, SPIRES selectively retains only those aligned with the predefined schema that can be grounded to the Relations Ontology (RO [14]). By exploiting standard identification schemes adopted by the reference bio-ontologies, the system guarantees the generation of triples that can be easily integrated into a biomedical KG.

SPIRES thus creates and refines prompts to maximize the effectiveness of LLMs by exploiting domain knowledge encapsulated through the description of the classes and relationships that we wish to include in the KG.

As outlined in [9], the explicit and structured information contained in KGs can also be used for improving the knowledge awareness of LLMs. KGs have been used: *i*) in the training of the LLM [15, 16]; *ii*) during the inference stage for making available to the LLMs the latest knowledge without retraining [17]; *iii*) to improve the interpretability of LLMs by explaining the facts [18] and by enhancing the reasoning process of LLMs [19]. One of the main disadvantages of solution *i*) is that the enhancement of the knowledge contained in the KG requires a retraining of the model which is a time (and money) consuming activity. For this reason, approaches of solution *ii*) are gaining momentum because they allow the separation of the text space and the knowledge space. In this case, knowledge is injected at the time of inference.

3. The SPIRES Schema for RNA-KG

For the creation of the schema needed for the application of SPIRES, we considered the RNA-KG meta-graph [20] that represents all the kinds of relationships involving RNA molecules in the considered data sources. Starting from it, a UML class diagram was developed that formally describes the schema of the considered domain and can be used for identifying meaningful relationships in the considered domain. Figure 2 shows an excerpt of the generated UML class diagram that consists of four biological and biomedical classes (`miRNA`, `gene`, `protein`, and `disease`) with six kinds of RO relationships.

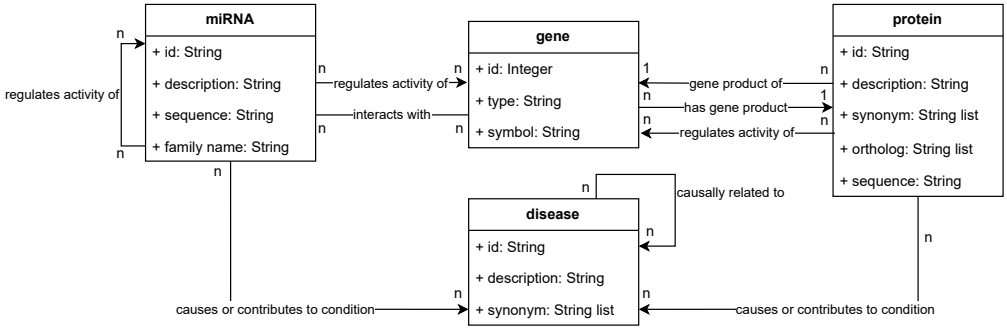


Figure 2: Meta-graph of test to evaluate the capabilities of SPIRES.

`miRNA` molecules are small non-coding RNAs that play a central role in gene expression via interference pathways and their misregulation is associated with several diseases [21]. `miRNA` molecules can generically `interacts with` genes but also more precisely `regulate the activity of` a gene when a miRNA molecule blocks the translation of a gene or promotes the degradation of gene's product. Moreover, miRNA molecules can `regulate the activity of` other miRNAs because they form base-pairing interactions with complementary miRNA molecules according to [22, 23]. The schema also contains the relationships involving genes and proteins. Specifically, the `has gene product` relation and its inverse `gene product of` are used for representing that different proteins are translated from the same gene (i.e. isoforms); while the `regulates activity of` is used for representing that a subclass of the proteins (transcription factors) regulates the activity of genes, promoting or down-regulating their activity acting as enhancers or repressors. Both proteins and miRNAs are connected to the `disease` class by the `causes or contributes to condition` relation. The diagram also contains the main properties that can be associated with these bio-entities (e.g., nucleotide/amino acid sequences, descriptions of molecules/diseases, synonyms).

The proposed UML class diagram was translated into a LinkML schema. Genes are annotated using HGNC [24] IDs. This choice is motivated by the stability of the HGNC IDs even if a gene name or symbol changes. Proteins are grounded to the PRotein Ontology (PRO) while diseases are grounded to both the Monarch Disease Ontology (Mondo [25]) and the Human Phenotype Ontology (HPO [26]). miRNAs were left with no semantic annotation since miRNA labels (e.g., `hsa-let-7b-5p`) and miRBase [27] accession identifiers (`MIMAT0000063`) are CURIE prefixes not included in default SPIRES annotators. We can manually retrieve miRNA molecules from relationships extracted from SPIRES since their labels follow a pattern (for instance, "hsa-" prefix indicates human miRNAs, "mmu-" prefix murine miRNAs, mature miRNA are designated with "miR-" substring whilst "mir" refers to the stem-loop primary transcript). Labels can be then easily translated into miRBase accession identifiers using a look-up table.

Example 2. *A LinkML class used to specify* `causes or contributes to condition` *relationships between proteins and diseases is reported in Listing 1. In the expression, we have to specify the need to extract triples representing relationships between proteins and disease in which the only admitted predicate is* `causes or contributes to condition` *(*`RO:0003302`*). In the expression, samples of the kinds of relationships that we wish to extract are reported. The prompt generated for this class relies on the prompts generated for the classes* `protein` *and* `disease` *and used for the identification of these bio-entities from the scientific literature. Figure 3 shows an output obtained by using SPIRES and the corresponding result obtained by the simple application of ChatGPT. In the SPIRES' output, the extracted interactions are already represented as triples that exploit the required identification scheme. Therefore, checking their presence in RNA-KG and, in case of new triples, their integration is facilitated.*

4. Experimental results

In this section we discuss the experiments that we carried out to evaluate SPIRES for extracting interactions involving RNA molecules. Moreover, we compare SPIRES with ChatGPT (ver. GPT-3.5-turbo), which is the LLM internally integrated in SPIRES, and with Llama 2 (ver. llama-2-70b-chat), another well-known and used LLM.

4.1. Corpus of Annotated Documents

To evaluate the extraction of relations aligned with the meta-graph depicted in Figure 2, we manually selected a

Listing 1: LinkML template for protein-disease interaction.

```
ProteinDiseaseInteraction:
 description: A document that contains protein to
              disease relationships.
 is_a: TextWithTriples
 slot_usage:
   triples:
    range: ProteinToDiseaseRelationship
    annotations:
    prompt: >-
      A semi-colon separated list of protein to
      disease relationships. The relationship
      is "causes or contributes to condition".
      For example:
      DNMT1 causes or contributes to condition
      Alzheimer disease;
      HOXA1 causes or contributes to condition
      Alzheimer disease.
```

corpus of 60 scientific articles gathered from PubMed, ResearchGate, and Google Scholar by specifying keyword-based queries like: “disease”, “comorbidity”, “protein”, “miRNA”, “miRNA regulation”, “gene”. From these documents, we identified paragraphs containing useful information to be extracted (e.g., abstract, discussion, or specific subsections within the domain of interest). In the identification of the paragraphs we have taken into account the following guidelines: *i*) the paragraph should contain different kinds of relations between bio-entities (e.g., “miRNA-interacts with-gene” and “miRNA-regulates activity of-gene”) to evaluate the ability of SPIRES to identify the right relations according to the provided meta-graph; *ii*) the paragraph might also contain irrelevant relationships that should be discarded; *iii*) different identification schemes can be used in the paragraph to check the ability of SPIRES to correctly work with them. Paragraphs have been classified according to the kind of bio-entities that they describe and associated with the list of relationships that should be identified according to the adopted meta-graph. For each kind of bio-entity, the following table shows the number of paragraphs containing relationships involving it (note that a paragraph can contain more than one).

Protein	Disease	miRNA	Gene
44	58	37	21

In the considered paragraphs, we have identified six kinds of interactions among the considered bio-entities (reported in the y-axis of the diagram in Figure 4).

4.2. Accuracy of Interactions extraction

For evaluating the obtained predictions, we have used standard metrics (precision, recall, and F-score) by considering the True Positive (TP), False Positive (FP), and

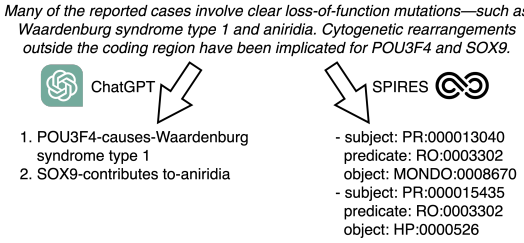


Figure 3: Example of output for SPIRES and ChatGPT.

False Negative (FN) according to the manually tagged paragraphs. Table 1 reports the obtained results for the considered interactions ordered according to the F-score measure. The obtained results indicate a consistent trend where recall tends to be lower than precision due to the prevalence of false negatives over false positives. We think this behavior is due to the difficulty in accurately extracting precise relationships from text, especially in distinguishing specific types of relationships. Furthermore, we observe that disease-disease and miRNA-disease interactions present a high F-score. These kinds of interactions are widely studied in the literature and thus a higher number of publications are available with respect to other interactions (like miRNA-miRNA interactions). Consequently, the abundance of this kind of relationships contributes to a higher true positive rate. Conversely, the F-score for protein-disease relations is notably low because it is influenced by low recall. We noticed that many protein-disease relations are undetected, often because they are expressed in complex ways within the text. For instance, the interchangeable use of symbols like “/” and “,” (e.g., “overexpressions in IL6/MEGF8/RELA, and also TP53 are known to cause osteoporosis”). Additionally, mapping proteins to the PRO proves challenging when textual information is sparse or ambiguously expressed. For instance, the mention of “PMP-22” solely as “myelin protein 22” instead of “peripheral myelin protein 22” (due to assumptions made by authors) can lead to inaccurate grounding. Despite this, precision remains remarkably high and, in the biomedicine context, this is preferable because it prioritizes certainty over ambiguity.

We also compared our results with the average results achieved by SPIRES in other domains. A marginal improvement has been observed in the domain of name entity recognition for chemicals and diseases [2]. We believe that the slightly enhanced accuracy is due to the use of multiple ontology annotators such as PRO for proteins, Mondo and HPO for diseases, and RO for relations.

4.3. Comparison with other LLMs

For assessing the performance of SPIRES with respect to ChatGPT and Llama 2, we focused on a subset of 20

	# Paragraphs	TP	FP	FN	F-score	Precision	Recall
disease-disease	16	54	5	10	0.88	0.92	0.84
miRNA-disease	32	123	20	31	0.82	0.86	0.80
miRNA-miRNA	1	19	1	7	0.82	0.95	0.73
gene-protein	10	52	5	21	0.8	0.91	0.71
miRNA-gene	13	14	3	5	0.78	0.82	0.74
protein-disease	24	42	7	60	0.56	0.86	0.41
Total	(60 texts)	304	41	134	0.76	0.88	0.69

Table 1
Results for named entity recognition evaluation of SPIRES on relations involving protein, miRNA, disease, and gene entities. Grounding was performed against HGNC, PRO, Mondo, HPO, and RO.

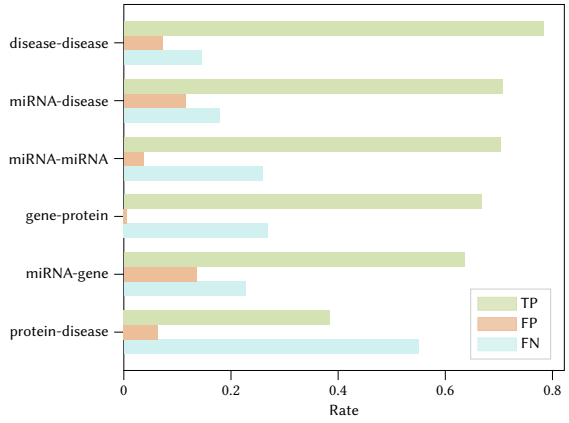
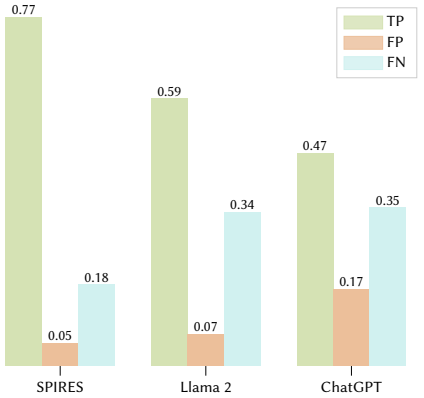


Figure 4: TP, FP, and FN results for evaluation of SPIRES on relations involving protein, miRNA, disease, and gene entities.



	F-score	Precision	Recall
SPIRES	0.86	0.94	0.81
Llama 2	0.74	0.89	0.64
ChatGPT	0.64	0.73	0.57

Figure 5: SPIRES vs Llama 2 vs ChatGPT on 20 texts.

documents where we manually grounded instances and relationships of the extracted triples. For using ChatGPT and Llama 2 we have generated prompts that adhere to the following pattern:

```
extract triples in the form
"subject-relation-object"
within this document: [...]
```

This prompt does not guarantee to obtain the identifiers for the subject and the object of the triples. However, if we try to generate a further prompt with the explicit request of mapping the extracted concepts to appropriate terminologies, both ChatGPT and Llama 2 advise that the provided ontology identifiers are hypothetical and may not correspond to actual ontology identifiers (so, hallucinations can occur in this case). Therefore we decided to substitute the grounding process with our manually curated look-up tables [1].

When using ChatGPT (or Llama 2) alone, we do not have to specify the schema, and results are produced through a single interaction with the user. Avoiding the specification of the schema might be interpreted as an advantage of basic LLMs approaches, but it is not. Indeed, the schema allows us to reduce the relationships to be extracted to only meaningful ones in the considered domain. Finally, no lookup table can be exploited for translating class instance names with the corresponding identifiers in the bio-ontologies (thus requiring a manual identification of the identifiers). All these drawbacks are avoided by the use of SPIRES.

As shown in the bottom part of Figure 5, SPIRES outperforms ChatGPT or Llama 2 alone both in terms of precision and recall. The histogram in Figure 5 points out a high increment in TP rate and a sensible decrease in FP and FN rates when adopting SPIRES instead of ChatGPT or Llama 2 alone for extracting relations that adhere to a specified schema within texts.

5. Concluding remarks

In this paper, we have reported the initial experimentation of the use of SPIRES for extracting triples from

the scientific literature related to RNA molecules by taking advantage of the meta-graph we have realized for the generation of RNA-KG. Even if a more systematic analysis is required, the initial results are quite encouraging. To facilitate the reproducibility of our results, our dataset and the LinkML template can be downloaded from: https://doi.org/10.5281/zenodo.10671796.

As future work, we would like to extend the approach by integrating the entire RNA-KG in different ways. First, we will exploit the RNA-KG triples for enhancing the prompts generated by SPIRES. Moreover, RNA-KG can be used for validating the plausibility of the generated triples by using RNA-KG as a gold standard in the area. Furthermore, we will explore the KG-enhanced LLM inference approaches in combination with SPIRES for further improving the precision of the system by injecting knowledge extracted from RNA-KG at inference time. Finally, we would like to create a web environment for graphically showing to the user the predicted triples directly in the graphical representation of the portion of the knowledge graph that will contain them. The user can thus manually check the proposed triples and provide feedback that will be handled afterward to improve the quality of the predictions.

Acknowledgements

This research was in part supported by the "National Center for Gene Therapy and Drugs based on RNA Technology", PNRR-NextGeneration EU program [G43C22001320007] and in part by the MUSA - Multilayered Urban Sustainability Action - Project, funded by the PNRR-NextGeneration EU program ([G43C22001370007], Code ECS00000037).

References

[1] E. Cavalleri, et al., RNA-KG: An ontology-based knowledge graph for representing interactions involving RNA molecules, 2023. arXiv:2312.00183.

[2] J. H. Caufield, et al., Structured Prompt Interrogation and Recursive Extraction of Semantics (SPIRES): a method for populating knowledge bases using zero-shot learning, Bioinformatics (2024). doi:10.1093/bioinformatics/btae104.

[3] R. Bommasani, et al., On the opportunities and risks of foundation models, CoRR abs/2108.07258 (2021).

[4] S. Moxon, et al., The Linked Data Modeling Language (LinkML): A General-Purpose Data Modeling Framework Grounded in Machine-Readable Semantics, in: Int'l Conf. on Biomedical Ontologies, 2021, pp. 148–151.

[5] OpenAI, Gpt-4 tech. report, 2023. arXiv:2303.08774.

[6] H. Touvron, et al., Llama 2: Open foundation and fine-tuned chat models, 2023. arXiv:2307.09288.

[7] A. Q. Jiang, et al., Mistral 7b, 2023. arXiv:2310.06825.

[8] L. Tunstall, et al., Zephyr: Direct Distillation of LM Alignment, 2023. arXiv:2310.16944.

[9] S. Pan, et al., Unifying large language models and knowledge graphs: A roadmap, 2023. arXiv:2306.08302.

[10] S. Ateia, U. Kruschwitz, Is ChatGPT a Biomedical Expert? – Exploring the Zero-Shot Performance of Current GPT Models in Biomedical Tasks, 2023. arXiv:2306.16108.

[11] Z. Ji, et al., Survey of hallucination in natural language generation, ACM Computing Surveys 55 (2023) 1–38. doi:10.1145/3571730.

[12] A. Ettinger, What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models, Transactions of the Association for Computational Linguistics 8 (2020) 34–48. doi:10.1162/tacl_a_00298.

[13] D. A. Natale, et al., The protein ontology: a structured representation of protein forms and complexes, Nucleic Acids Research 39 (2010). doi:10.1093/nar/gkq907.

[14] C. Mungall, et al., oborel/obo-relations: 2023-08-18 release, 2023. doi:10.5281/zenodo.8263469.

[15] Z. Zhang, et al., ERNIE: Enhanced language representation with informative entities, in: Proc. of Annual Meeting of the Association for Computational Linguistics, 2019, pp. 1441–1451. doi:10.18653/v1/P19-1139.

[16] C. Rosset, et al., Knowledge-aware language model pretraining, CoRR (2020). arXiv:2007.00655.

[17] P. Lewis, et al., Retrieval-augmented generation for knowledge-intensive NLP tasks, in: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, 2020.

[18] M. Danilevsky, et al., A survey of the state of explainable AI for natural language processing, in: Proc. of Int'l Conf. on Natural Language Processing, 2020, pp. 447–459.

[19] B. Y. Lin, X. Chen, J. Chen, X. Ren, KagNet: Knowledge-aware graph networks for commonsense reasoning, 2019. arXiv:1909.02151.

[20] E. Cavalleri, et al., A meta-graph for the construction of an rna-centered knowledge graph, in: Bioinformatics and Biomedical Engineering, Springer, 2023, pp. 165–180. doi:10.1007/978-3-031-34953-9_13.

[21] G. J. Hannon, Rna interference, Nature 418 (2002) 244–251. doi:10.1038/418244a.

[22] L. Guo, et al., miRNA–miRNA interaction implicates for potential mutual regulatory pattern, Gene 511 (2012) 187–194. doi:10.1016/j.gene.2012.09.066.

[23] E. C. Lai, et al., Complementary miRNA pairs suggest a regulatory role for miRNA:miRNA duplexes., RNA 10 (2004) 171–175. doi:10.1261/rna.5191904.

[24] R. L. Seal, et al., Genenames.org: the HGNC resources in 2023, Nucleic Acids Research 51 (2022) D1003–D1009. doi:10.1093/nar/gkac888.

[25] N. A. Vasilevsky, et al., Mondo: Unifying diseases for the world, by the world, medRxiv (2022). doi:10.1101/2022.04.13.22273750.

[26] P. N. Robinson, et al., The human phenotype ontology: A tool for annotating and analyzing human hereditary disease, The American Journal of Human Genetics 83 (2008) 610–615. doi:10.1016/j.ajhg.2008.09.017.

[27] A. Kozomara, et al., miRBase: from microRNA sequences to function, Nucleic Acids Research 47 (2018) D155–D162. doi:10.1093/nar/gky1141.