Research Statement

Qingyun Wang

Scientists are experiencing information overload [4] due to the rapid growth of scientific literature. Beyond this, scientific papers, known as "*Sleeping beauties*", sometimes remain largely unnoticed for long periods before suddenly attracting great attention [19]. Moreover, the process of discovering new scientific hypotheses has remained *slow*, *expensive*, and *highly specialist-dependent*, due to *the increasingly complex experiments*. The recent advancements in large language models (LLMs) raise the prospect that they may be able to solve those problems [16]. Despite their impressive progress, these models often fail to *incorporate domain-specific knowledge effectively* and *support their generated results with evidence*. Furthermore, the knowledge captured in expert-curated databases represents only a small fraction of the entire domain, due to *high annotation cost*. To address this issue and lower the entry barrier for interdisciplinary collaboration, I **develop AI tools to accelerate and democratize the entire research lifecycle for scientists** (Fig 1), from *knowledge acquisition* [6, 11, 17], *hypothesis generation* [15, 20], *multimedia procedure planning* [14] for experiment design, *experiment execution* [5], *conduction to writing* [7, 8, 9, 12] and *evaluating the paper draft* [10]. Unlike AI4Science, which creates AI-driven solutions for scientific challenges, I focus on **AI4Scientists** to *empower scientists with AI tools to enhance their research lifecycle*.

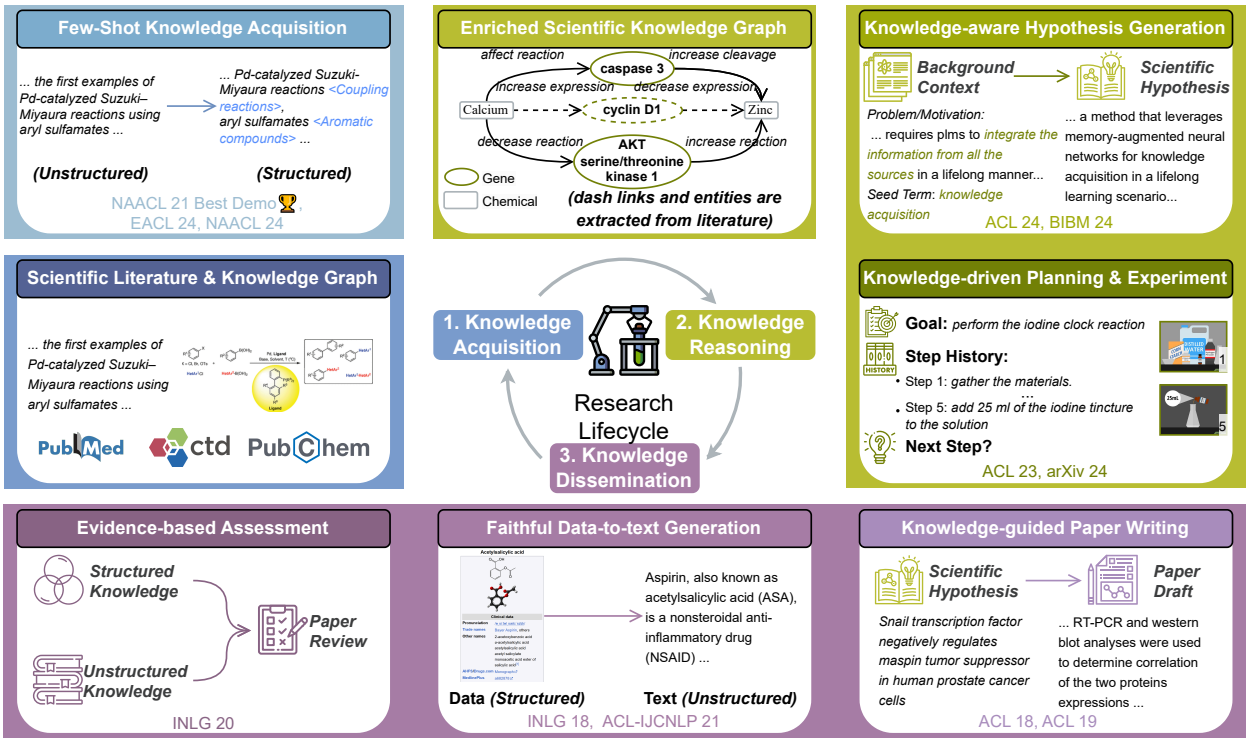


Figure 1: My research developed AI tools that can effectively accelerate the research lifecycle for scientists.

My commitment to developing **AI4Scientists** has led me to focus on **a knowledge lifecycle** with three components:

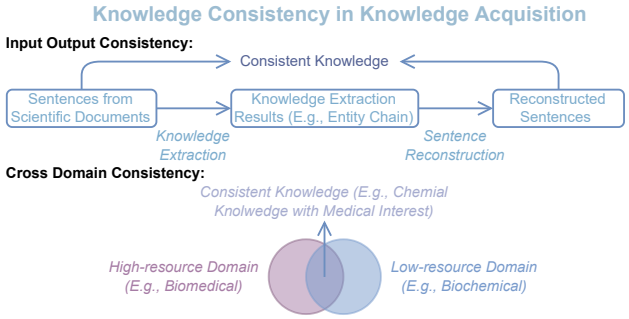
- **Few-shot Scientific Knowledge Acquisition (§11).** Scientific knowledge acquisition is the foundation for many AI4Science applications. To combat COVID-19, my COVID-KG extracted fine-grained multimedia knowledge graphs (KGs) from scientific papers for drug repurposing reports [11]. However, fine-grained information extraction systems usually require large amounts of expert-annotated examples to perform effectively. Therefore, creating methods that can learn effectively from only a few examples, known as *few-shot* approaches, becomes a vital topic in scientific knowledge acquisition. To address this limitation, I proposed to *utilize the knowledge consistency between input text and output knowledge elements* for few-shot fine-grained scientific entity extraction [17]. Furthermore, to reduce the mislabeling in few-shot transfer learning, I built methods to *project the source and target entities into separate feature spaces* [6].
- **Integrating Domain Knowledge with Scientific LLM Reasoning (§2).** Based on the scientific KGs in the previous component, I investigate *scientific hypothesis generation*, as well as *experiment planning and execution*. Simulating the human research process, I propose to *augment LLMs with external heterogeneous KGs from*

previous papers as "inspirations" to generate novel scientific hypotheses and *iteratively boosting the novelty of the generated hypothesis* [15]. Furthermore, to *incorporate external knowledge from both vision and text*, I introduced a new *multimedia procedure learning framework* to produce *visually trackable, inductive, and diverse* task scripts [14]. Finally, using LLM agents, we designed a new autonomous Machine Learning Research with LLMs (MLR-Copilot) [5], to *generate research hypothesis and conduct experiments*.

- **Explainable Scientific Knowledge Dissemination (§3).** To communicate the new idea to readers clearly, I was the first to *introduce scientific paper idea writing assistants* [8, 9], enabling the integration of graph link predictions from both structured and unstructured knowledge into paper draft writing. Another important step in the research lifecycle is evaluating paper quality to prevent distorted scientific dissemination. I built an *explainable paper review generation system* [10], to generate explainable review scores and comments, along with detailed evidence, based on KGs and papers. Beyond the scientific domain, I also want to help people access knowledge more easily. I proposed new *table position attention* [7] and *graph position embeddings* [12] to help models capture structure information of KGs and generate faithful descriptions when describing a KG.

Along this direction, I have published over **24** papers in top venues with over **825** citations and received over **730** stars on GitHub[1] in total, including **11** first-author papers, workshops, and tutorials. My COVID-KG paper has been honorably awarded **NAACL 2021 Best Demo Paper** [11] and widely used by other researchers (downloaded more than 2000 times). My system [10] not only **ranked first** in the Paper with Code websiteat the time of publication but also was **patented during collaborations with Salesforce Research** [13]. Furthermore, the patent has been cited by both IBM and Microsoft. I delivered **tutorials** on AI-assisted research lifecycle at COLING 2024 [16] and EACL 2024 [1], as well as on knowledge-guided language generation at EMNLP 2021 [21]. I lead an **AI4Research** workshop at AAAI 2025 [18] and co-organized a **language + molecule workshop** [2] at ACL 2024 to promote community discussion on the *AI-assisted knowledge-grounded scientific research lifecycle*.

§1 Few-shot Scientific Knowledge Acquisition



Due to the high annotation cost and the rapid pace of scientific discovery, existing expert-curated scientific KGs often miss crucial knowledge elements. Fine-grained information extraction systems are crucial to bridge this gap [11], but enhancing their performance requires many expert-annotated examples, which are limited and expensive. Therefore, developing *few-shot* approaches, that require only a few examples to learn effectively, becomes a vital topic in scientific knowledge acquisition. In my research on few-shot science entity extraction, I have developed methods that leverage **knowledge consistency between input and output** [17], as well as **across domains** [6], which significantly **reduces computational costs and the need for external knowledge**.

Few-shot fine-grained information extraction systems in the scientific domain face two unique problems: *missing mentions* and *incorrect long-tail predictions* due to dense entity coverage in scientific paper sentences and imbalanced entity type distributions. To tackle these issues, inspired by *the knowledge consistency between the input text and structured output*, I pioneeringly proposed a new few-shot entity extraction approach with **a self-validation module** [17] to reconstruct the original sentences based on entity extraction results. Unlike previous dual cycle training frameworks, I introduced *the gumbel-softmax estimator* to **avoid the non-differentiable issue** and reduced the dual cycle training into end-to-end training to **save time and computation resources**. Moreover, I introduced *a new entity decoder contrastive loss* to control the *excessive copying* of entity mention spans from the original sentence. Compared to previous approaches, my new method **does not require any external knowledge or domain adaptive pretraining** and operates with **fewer parameters**, with positive results across the chemical, AI, music, literature, and politics domains. Collaborating with domain experts from Carl R. Woese Institute for Genomic Biology at UIUC, we release *ChemNER+*, a new few-shot fine-grained chemical entity extraction dataset, focusing on the chemical types collected from Wikipedia. As part of an undergraduate student project I mentored, we investigated mislabeling problems in few-shot entity extraction, where previous transfer learning approaches tend to *mislabel source entities as target entities*.

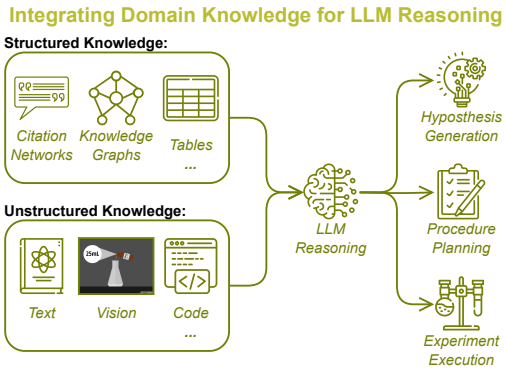
[1] https://github.com/EagleW

To mitigate this issue, we proposed to *project the source entities and target entities* into *separate regions of the feature space* [6]. The proposed framework could **boost the performance of transfer-learning models in other tasks**.

Scientific knowledge acquisition can extend to multimodality. During COVID-19, collaborating with UCLA and Columbia, I created a multimedia knowledge extraction system **COVID-KG** to extract *fine-grained multimedia knowledge elements* from text and diagrams in papers related to COVID-19 [11], containing interactions between genes, diseases, symptoms, and chemicals. Furthermore, I exploited the constructed multimedia KGs for **drug repurposing**. Using COVID-KG, *DARPA biologists were able to suggest three drugs*: Benazepril, Losartan, and Amodiaquine. My system won the **NAACL 2021 Best Demo Award** and received **2000+ downloads**.

§2 Integrating Domain Knowledge with Scientific LLM Reasoning

Knowledge acquisition from massive scientific papers provides many structured knowledge elements. These knowledge elements can enrich expert-curated external knowledge bases, guiding the predictions of new potential research ideas, known as *scientific hypothesis generation*. Previous work on literature-based hypothesis generation has focused on binary link prediction—severely limiting the expressivity of hypotheses. Moreover, those papers rely on expert-calibrated canonicalized KGs, limiting their broader real-world applications. State-of-the-art LLMs (e.g., GPT-4) tend to generate ideas *with overall low technical depth and novelty* due to a lack of domain knowledge. To solve more complex problems, I **enhance LLMs with structured and unstructured domain knowledge**.



My work is founded upon **a brand new research direction, *Scientific Inspiration Machines Optimized for Novelty* (SCIMON)** [15], by *providing descriptions of problem contexts to hypothesis generation*, including current problems, motivations, experimental settings, and constraints. I developed **the first benchmark for this task in natural language processing and biomedical domain**. Simulating the human research process, I *narrowed down the search space of hypothesis generation* and *enriched the background knowledge* by **augmenting LLMs with external heterogeneous KGs from previous papers as "inspirations"** [15] with three types of sources (i.e., semantic neighbors, KG, and citation neighbors). Furthermore, I introduced **a new iterative novelty boosting mechanism** that helps LLMs generate more novel ideas by *explicitly comparing ideas to prior work and updating idea suggestions until sufficient novelty is achieved*, which can improve the novelty by up to 57%. Our domain-agnostic framework can be **applied to other domains** by changing the preprocessing procedure. My work represents **a first step toward evaluating and developing language models that generate new ideas derived from the scientific literature**. Furthermore, we designed a new framework, autonomous Machine Learning Research with LLMs (MLR-Copilot) [5], to enhance machine learning research productivity through the *automatic generation and implementation of research ideas* by *LLM agents*.

To explore ways to **leverage multimedia contextual information** and **assist robots in everyday tasks**, I initiated a pilot study into *generative procedure planning*. Previous generative script learning systems focus solely on text, affected by reporting bias [3] as important details may be omitted in the source text but often implicitly contained in images. To address these problems, I proposed **a new task, Multimedia Generative Script Learning** [14], that requires systems to generate future steps based on the goal and previous steps *with visual scenes depicting their states*. Furthermore, I proposed to encode *visual state changes* through *a selective multimedia encoder*, *transfer knowledge from previously observed tasks* using *a retrieval-augmented decoder*, and further *present distinct information at each step* by optimizing *a diversity-oriented contrastive learning objective*. The proposed framework can be applied to different *AI4Science* tasks, including *automated experiment planning*, *workflow optimization*, etc.

§3 Explainable Scientific Knowledge Dissemination

Scientific knowledge dissemination can be divided into scientific paper writing [8, 9] and scientific paper evaluation [10]. My research has explored **utilizing structured knowledge to provide explainable generation results**.

Inspired by the human writing procedure, I am among the **first researchers** to develop a **virtual scientific research assistant for scientific paper writing** [9], **PaperRobot**, which integrates the generation of new ideas through link prediction into automated paper drafting based on *background KGs and unstructured paper text*. Turing Tests show *PaperRobot* generated paper components are chosen over human-written ones up to *30%*. Another essential step in scientific discovery is evaluating paper quality to *prevent distorted scientific dissemination*. I designed a new **ReviewRobot** [10] based on the *differences between knowledge graphs* to generate review comments, which are *knowledgeable*, being constructive and informative to help improve the paper; and *explainable*, by providing detailed evidence. It achieved *71.4%-100%* accuracy in review score prediction, with *41.7%-70.5%* of its comments rated as valid and constructive by humans, outperforming human-written reviews *20%* of the time.

Beyond the scientific domain, I investigate how NLP models can **faithfully** translate knowledge elements into descriptive text, *helping people access knowledge more easily*. One of my work [7] reveals that 51% of entity attributes in the current English Wikipedia infoboxes are not described in English articles in the Wikipedia dump. Specifically, instead of traditional graph encoders, I focus on **directly incorporating structure information into the structured encoder**. I developed a new *Table Position Self-attention* [7] to capture the inter-dependencies among related slots of Wikipedia infoboxes. To better adapt structure information into transformer-based pretrained language models, I proposed a novel *tree-level embedding* method to capture the inter-dependency structures of the input graph. This new approach achieved the **best performance** for the WebNLG 2017 dataset at that time and was granted a patent [13].

Future Research Agenda

My **long-term vision** is to expand **AI4Scientist** to **democratize scientific discovery** by **equipping machines with the ability to interact dynamically with the human and physical world**. In particular, I will **empower machines to understand and perform reasoning for different modalities of scientific knowledge with human feedback**. Along this path, I plan to continue my research as follows:

- **Scientific Multimodal Foundation Models with Critical Thinking.** Current scientific hypothesis generation systems tend to generate ideas with low technical depth and novelty [15] due to a lack of multimodal reasoning ability. Equipping LLMs with multimodal information can help **mitigate reporting bias** [3] in downstream tasks such as paper review generation, scientific hypothesis generation, and scientific claim verification. I plan to build a new multimodal scientific LLM *to understand formulas, tables, figures, and charts*, since those multimodal knowledge elements are usually complementary to text descriptions. Moreover, most of today's multimodal LLMs *ignore structural information*. For example, LLMs usually take in molecules as linearized sequences, losing structural information. I plan to *design a plug-and-play module to capture structural information*. Finally, my proposed framework will be able to **dynamically extract and integrate new multimodal knowledge elements without additional training**. This approach can *reduce computational costs* and *enhance the explainability* of the results. Simultaneously, the new framework will have the ability to **quantify uncertainty when reasoning** to reduce hallucination. I am passionate about working with researchers in **biology**, **physics**, **mathematics**, **chemistry**, and **social science** to improve the ability of LLMs to understand scientific concepts.
- **Reliable Scientific Foundation Models.** Since the primary function of the AI4Scientist is to help scientists, it is crucial to **align the model to provide reliable and safe outputs**. In addition, while many research articles, especially preprints, provide new perspectives for researchers, they duplicate findings, spread misinformation, and show disagreements among themselves [11, 16]. To reduce inconsistency and increase interpretability, I plan to continue my research on integrating domain knowledge with LLMs by **incorporating neurosymbolic frameworks for scientific reasoning**. Specifically, by building on my current work in scientific knowledge acquisition and future work in multimodal foundation models for scientific discovery, I will first construct a multimodal domain-specific KG with minimal supervision to understand the corresponding background corpus. I then aim to develop an AI system that can not only *provide supporting evidence and corresponding logic chain* but also *align its internal reasoning procedure with the provided explanation*. The proposed system can **utilize both structured and unstructured knowledge as well as logic rules among knowledge elements** to *produce trustworthy and explainable results* as well as *verify the feasibility of scientific claims*. I am excited to work with researchers in **neurosymbolic AI** to improve the trustworthiness and explainability of AI models.
- **Scientific Research Agents with Physical World Interactions.** Existing scientific discovery procedures are time-consuming, labor-intensive, inefficient, and expensive. Therefore, I envision developing a human-in-the-

loop **self-driving laboratory** that can complete the scientific research lifecycle through **interactions with the physical world**, such as a robotic laboratory. To achieve this goal, I plan to design a powerful embodied agent that can *start with a high-level description of scientific questions and decompose them into sub-tasks*, including *knowledge retrieval*, *hypothesis generation*, *experiment planning*, *conducting simulations and experiments*, and *results analysis*. Furthermore, I aim to develop experiment agents that can **follow safety guidelines to conduct experiments** by analyzing visual, textual, and audio information. At the same time, such agents will be able to **update its internal knowledge based on human feedback and experiment results**. For example, in **chemistry**, machines will have the ability to automatically design and produce new drugs and enzymes based on existing literature and experimental feedback. The human, experimental, and literature feedback will be used as rewards to train a new **human-in-the-loop reinforcement learning framework**, which can **leverage small datasets in closed-loop discovery platforms**. Such systems can help *democratize biological/material synthesis and characterization*, and *achieve unbiased data collection and analysis*. I will collaborate with researchers in **robotics**, **medicine**, **biology**, and **chemistry** to explore possible applications.

Collaboration and Funding. I am fortunate to collaborate with more than 15 professors from top universities and research institutes, such as Columbia University, University of Washington, University of California Los Angeles, University of North Carolina at Chapel Hill, etc. I also have had the fortune to work closely with researchers in fields outside of computer science, including chemistry, biology, healthcare, education, and agriculture, to conduct interdisciplinary research. As I expand my research scope in the AI-assisted scientific research lifecycle, I envision collaborating with experts in many fields, including natural language processing, machine learning, data mining, computer vision, social science, robotics, HCI, biology, chemistry, etc. During my Ph.D. study, I was supported by NSF, DARPA, the Department of Education, the Department of Energy, and AFRI. I have also contributed to the writing of winning grant proposals, including idea generation, method design, idea illustration, and visual aid creation, such as for the DARPA ITM project and NSF MMLI project. I will continue to actively seek funding opportunities in the future from multiple funding agencies (e.g., DARPA, NSF, NIH) and industry partners.

References

[1] Carl Edwards, **Qingyun Wang**, and Heng Ji. Language + molecules. In *EACL Tutorial*, 2024. URL https://aclanthology.org/2024.eacl-tutorials.3.

[2] Carl Edwards, **Qingyun Wang**, Manling Li, Lawrence Zhao, Tom Hope, and Heng Ji, editors. *Proceedings of the 1st Workshop on Language + Molecules (L+M 2024)*, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.langmol-1.0.

[3] Jonathan Gordon and Benjamin Van Durme. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, page 25–30. Association for Computing Machinery, 2013. ISBN 9781450324113. URL https://doi.org/10.1145/2509558.2509563.

[4] Esther Landhuis. Scientific literature: Information overload. *Nature*, 535(7612):457–458, 2016. URL https://www.nature.com/articles/nj7612-457a.

[5] Ruochen Li, Teerth Patel, **Qingyun Wang** (mentor), and Xinya Du. Mlr-copilot: Autonomous machine learning research based on large language models agents. *arXiv preprint arXiv:2408.14033*, 2024. URL https://arxiv.org/pdf/2408.14033.

[6] Hongyi Liu, **Qingyun Wang** (mentor), Payam Karisani, and Heng Ji. Named entity recognition under domain shift via metric learning for life sciences. In *NAACL*, 2024. URL https://aclanthology.org/2024.naacl-long.1.

[7] **Qingyun Wang**, Xiaoman Pan, Lifu Huang, Boliang Zhang, Zhiying Jiang, Heng Ji, and Kevin Knight. Describing a knowledge base. In *INLG*, 2018. URL https://aclanthology.org/W18-6502.

[8] **Qingyun Wang***, Zhihao Zhou*, Lifu Huang, Spencer Whitehead, Boliang Zhang, Heng Ji, and Kevin Knight. Paper abstract writing through editing mechanism. In *ACL*, 2018. URL https://aclanthology.org/P18-2042.

[9] **Qingyun Wang**, Lifu Huang, Zhiying Jiang, Kevin Knight, Heng Ji, Mohit Bansal, and Yi Luan. PaperRobot: Incremental draft generation of scientific ideas. In *ACL*, 2019. URL https://aclanthology.org/P19-1191.

[10] **Qingyun Wang**, Qi Zeng, Lifu Huang, Kevin Knight, Heng Ji, and Nazneen Fatema Rajani. ReviewRobot: Explainable paper review generation based on knowledge synthesis. In *INLG*, 2020. URL https://aclanthology.org/2020.inlg-1.44.pdf.

[11] **Qingyun Wang**, Manling Li, Xuan Wang, Nikolaus Parulian, Guangxing Han, Jiawei Ma, Jingxuan Tu, Ying Lin, Ranran Haoran Zhang, Weili Liu, Aabhas Chauhan, Yingjun Guan, Bangzheng Li, Ruisong Li, Xiangchen Song, Yi Fung, Heng Ji, Jiawei Han, Shih-Fu Chang, James Pustejovsky, Jasmine Rah, David Liem, Ahmed ELsayed, Martha Palmer, Clare Voss, Cynthia Schneider, and Boyan Onyshkevych. COVID-19 literature knowledge graph construction and drug repurposing report generation. In *NAACL Demonstrations* ***(Best Demo Award)***, 2021. URL https://aclanthology.org/2021.naacl-demos.8.pdf.

[12] **Qingyun Wang**, Semih Yavuz, Xi Victoria Lin, Heng Ji, and Nazneen Rajani. Stage-wise fine-tuning for graph-to-text generation. In *ACL-IJCNLP Student Research Workshop*, August 2021. URL https://aclanthology.org/2021.acl-srw.2.

[13] **Qingyun Wang**, Semih Yavuz, Xi Lin, and Nazneen Rajani. Structured graph-to-text generation with two step fine-tuning. In *US Patent US11727210B2*, 2022. URL https://patents.google.com/patent/US11727210B2/.

[14] **Qingyun Wang**, Manling Li, Hou Pong Chan, Lifu Huang, Julia Hockenmaier, Girish Chowdhary, and Heng Ji. Multimedia generative script learning for task planning. In *ACL Findings*, 2023. URL https://aclanthology.org/2023.findings-acl.63.

[15] **Qingyun Wang**, Doug Downey, Heng Ji, and Tom Hope. SciMON: Scientific inspiration machines optimized for novelty. In *ACL*, 2024. URL https://aclanthology.org/2024.acl-long.18.

[16] **Qingyun Wang**, Carl Edwards, Heng Ji, and Tom Hope. Towards a human-computer collaborative scientific paper lifecycle: A pilot study and hands-on tutorial. In *LREC-COLING Tutorial*, May 2024. URL https://aclanthology.org/2024.lrec-tutorials.10.

[17] **Qingyun Wang**, Zixuan Zhang, Hongxiang Li, Xuan Liu, Jiawei Han, Huimin Zhao, and Heng Ji. ChemFINESE: Validating fine-grained few-shot entity extraction through text reconstruction. In *EACL Findings*, 2024. URL https://aclanthology.org/2024.findings-eacl.1.

[18] **Qingyun Wang**, Wenpeng Yin, Lifu Huang, Fung May, Xinya Du, Carl Edwards, and Tom Hope, editors. *Proceedings of the 2nd AI4Research: Towards a Knowledge-grounded Scientific Research Lifecycle (AI4Research@AAAI 2025)*, 2025. URL https://aaai.org/conference/aaai/aaai-25/workshop-list/.

[19] Anthony FJ Van Raan. Sleeping beauties in science. *Scientometrics*, 59(3):467–472, 2004. URL https://link.springer.com/article/10.1023/B:SCIE.0000018543.82441.f1.

[20] Kexuan Xin, **Qingyun Wang**, Junyu Chen, Pengfei Yu, Huimin Zhao, and Heng Ji. Gene-metabolite association prediction with interactive knowledge transfer enhanced graph for metabolite production. In *BIBM 2024*, 2024. URL https://arxiv.org/pdf/2410.18475.

[21] Wenhao Yu, Meng Jiang, Zhiting Hu, **Qingyun Wang**, Heng Ji, and Nazneen Rajani. Knowledge-enriched natural language generation. In *EMNLP Tutorial*, 2021. URL https://aclanthology.org/2021.emnlp-tutorials.3.