Modern Natural Language Processing Techniques for Scientific Web Mining: Tasks, Data, and Tools

Xuan Wang, Hongwei Wang, Heng Ji, Jiawei Han
University of Illinois at Urbana-Champaign
{xwang174,hongweiw,hengji,hanj}@illinois.edu

ABSTRACT

This tutorial targets researchers and practitioners who are interested in natural language processing (NLP) technologies for scientific web mining. Exploring the vast amount of rapidly growing scientific literature available on the web is highly beneficial for scientific discovery. However, scientific web mining is particularly challenging due to the lack of specialized domain knowledge in natural language context, complex sentence structures in scientific writing, and multi-modal representations of scientific knowledge.

This tutorial presents a comprehensive overview of recent research and development on using NLP techniques for scientific web mining, focusing on the biomedical and chemistry domains. First, we introduce the motivation and unique challenges of web mining in the scientific domains. Then we discuss a set of methods that perform effective information extraction (named entity recognition, relation extraction, and event extraction), information retrieval (textual evidence retrieval, cross-modal molecule retrieval, and chemical reaction tracking) from scientific literature, and their applications on reaction prediction. Finally, we conclude our tutorial by demonstrating, on real-world datasets (COVID-19 and organic chemistry literature), how the information can be extracted and retrieved, and how they can assist further exploratory analysis. We also discuss the emerging research problems and future directions of using NLP techniques for scientific web mining.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; **Web mining**; • **Computing methodologies** → **Natural language processing**; **Information extraction**; • **Applied computing** → *Life and medical sciences.*

KEYWORDS

Natural Language Processing, Information Retrieval, Information Extraction, Biomedical Web Mining, Chemistry Web Mining

ACM Reference Format:
Xuan Wang, Hongwei Wang, Heng Ji, Jiawei Han, University of Illinois at Urbana-Champaign . 2018. Modern Natural Language Processing Techniques for Scientific Web Mining: Tasks, Data, and Tools. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/1122445.1122456

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

1 WHY IS THIS TUTORIAL IMPORTANT?

Exploring the vast amount of rapidly growing scientific literature available on the web is highly beneficial for scientific discovery. However, scientific web mining is particularly challenging due to a lack of specialized domain knowledge in natural language context, complex sentence structures in scientific writing, and multi-modal representations of scientific knowledge. Recent years have witnessed a growing interest in bringing natural language processing (NLP) and text mining for scientific knowledge discovery. For example, chemists and computer scientists have developed a software platform that uses NLP techniques to translate the organic chemistry literature directly into editable code, which significantly accelerate the progress of AI-driven chemical synthesis and molecule discovery [14]. Unfortunately, there is no existing tutorial covering the topics on using NLP techniques for scientific web mining to enable and accelerate scientific discovery. This tutorial bridge this gap and talks about the recent trends and emerging challenges in this topic, representative tools and learning resources to obtain ready-to-use models, and how related models and techniques benefit scientific web mining applications.

2 TOPICS COVERED IN THIS TUTORIAL

This tutorial presents a comprehensive overview of recent research and development on using NLP techniques for scientific web mining, focusing on the biomedical and chemistry domains. We first present our vision on using NLP techniques for scientific web mining to enable and accelerate scientific discovery. Then we introduce three major modules of our tutorial: (1) scientific information extraction, (2) scientific information retrieval, and (3) system demonstrations and future directions.

2.1 Scientific Information Extraction

We first introduce the background of information extraction from scientific literature. Then we discuss three major tasks: (1) named entity recognition, (2) relation extraction, and (3) event extraction. **Named Entity Recognition (NER)** is a fundamental step in scientific literature analysis to build AI-driven systems for molecular discovery, synthetic strategy designing, and manufacturing [2, 7, 18, 19, 31]. Previous scientific NER studies are mostly focused on a few coarse-grained entity types (i.e., genes, chemicals, and diseases) [6, 10, 28]. However, scientific literature analysis needs fine-grained entities to provide a wide range of information for scientific discovery. One major challenge for comprehending the fine-grained scientific entities is the urgent need of domain-specific background knowledge, because there are a number of

domain-specific explained common expressions, acronyms, and abbreviations that are difficult for the model to understand. We discuss recent methods that incorporate domain knowledge for fine-grained scientific NER [12, 32]. Another major challenge is that the need for dozens to hundreds of fine-grained entity types makes consistent and accurate annotation difficult even for crowds of domain experts. Apart from human annotation, domain-specific ontologies and knowledge bases can be easily accessed, constructed, or integrated, which makes distant supervision realistic for fine-grained scientific NER. We discuss recent methods for fine-grained scientific NER with distant supervision [23, 24, 27].

Relation Extraction is important for real-world scientific applications as clinical decision support [4] and question answering [1], biological pathway and network analysis [15], integrative biology [16], biocuration [17, 29, 30] and pharmacovigilance [5]. A major challenge is that authors tend to compose long and complex sentences in scientific papers, where the entity mentions under the same relationship are usually located far away from each other within the sentence. We first discuss recent methods that use Abstract Meaning Representation (AMR) to compress the wide context to uncover a clear semantic structure for each complex sentence [12, 32]. Supervised relation extraction can extract limited types of relations and cannot be easily extended to new relation types. On the other hand, open relation extraction requires no pre-specified relation types but aims to extract all the relation tuples from a corpus. We then discuss open relation extraction methods that conduct meta-pattern discovery over long and complex sentences in scientific literature [13, 26].

Event Extraction is more complex than entity and relation extraction because it involves extraction of event triggers and all participants (arguments). First, compared with general natural language texts, sentences from scientific papers usually possess wider contexts between knowledge elements. Moreover, comprehending the fine-grained scientific entities and events urgently requires domain-specific background knowledge. In this tutorial, we will present recent event extraction methods to tackle these two challenges by compressing wide context to uncover a clear semantic structure for each complex sentence [32], and constructing the sentence-level knowledge graph from an external knowledge base and using it to enrich the semantic parsing graph to improve the model's understanding of complex scientific concepts [12, 32].

2.2 Scientific Information Retrieval

To use the extracted information for a better web search, we discuss three major tasks for scientific information retrieval: (1) textual evidence discovery, (2) cross-modal molecule retrieval with natural language queries, and (3) molecule representation learning.

Textual Evidence Discovery aims to automatically retrieve evidence sentences given an user-input query. Textual evidence discovery is an important but underexplored problem in scientific text mining. With the rapid development of high-throughput technologies, scientists can generate many scientific hypotheses in silico within a short period. Experimental validation of these scientific hypotheses is the most conclusive validation procedure in scientific research. However, the high cost incurred in performing the experimental validation dictates that the in silico-generated hypotheses must be carefully prioritized to select the most confident ones for validation. To prioritize the hypotheses before the experimental validation, scientists often look into the scientific literature for previously uncovered textual evidence that partially validates the hypotheses at hand. Traditional literature search engines (e.g., PubMed) are designed for document retrieval and do not allow direct retrieval of specific statements. Some of these statements may serve as textual evidence that is key to hypothesis generation and new finding validation. We discuss recent work on textual evidence mining from scientific literature [22, 25].

Cross-Modal Molecule Retrieval aims to discover new properties and applications of different molecules to accelerate discovery in medicine and science. Existing databases contain tens of millions of molecules; PubChem [8, 9] alone has 110 million compounds. Many information retrieval tools for chemistry rely on queries based on natural language descriptions of the molecules and existing chemical reactions. Hundreds of millions of possible molecules cannot all possibly undergo laboratory experimentation and be given attention by experts in order to create a description. To address this issue, it is critical to retrieve molecules directly from natural language descriptions. We discuss recent work that retrieve molecules directly from natural language descriptions and allow newly discovered molecules to be easily integrated into the proposed information retrieval framework [3]. This framework also allows for semantic-level search between natural language descriptions and molecules as well as for query expansion within traditional chemistry information retrieval systems.

Molecule Representation Learning aims to embed molecules into a real vector space. Chemists usually use IUPAC nomenclature, molecular formula, structural formula, skeletal formula, etc., to represent molecules in chemistry literature. However, such representations are initially designed for human readers rather than computers. To facilitate machine learning algorithms understanding and making use of molecules, molecule representation learning is proposed to map molecules into a low-dimensional real space and represent them as dense vectors. The learned vectors (a.k.a. embeddings) of molecules can benefit information retrieval and text mining technologies for chemistry [11]. We discuss recent work that use chemical reactions to assist learning molecule representations and improving their generalization ability [20].

2.3 System Demos and Future Directions

Finally, we conclude our tutorial by demonstrating, on real-world datasets (COVID-19 and organic chemistry literature), how the information can be extracted and retrieved, and how they can assist further exploratory analysis. We first introduce a novel and comprehensive knowledge discovery framework, COVID-KG [21], that extracts fine-grained multimedia knowledge elements (entities, relations and events) from scientific literature. We exploit the constructed multimedia knowledge graphs (KGs) for question answering and report generation, using drug repurposing as a case study. COVID-KG also provides detailed contextual sentences, subfigures and knowledge subgraphs as evidence. All of the data, KGs, reports[1], resources and shared services are publicly available[2]. We then introduce a chemistry reaction tracker system, built on the

[1]Video: http://159.89.180.81/demo/covid/Covid-KG_DemoVideo.mp4
[2]System page: http://blender.cs.illinois.edu/covid19/

advanced information extraction and information retrieval methods discussed above, that tracks chemistry research publications related to particular organic chemical reactions based on users' queries. This chemistry reaction tracker system demonstrates the strong ability of modern NLP techniques on tracking the rapidly growing scientific literature available on the web. Last, we discuss the emerging research problems and future directions of using NLP techniques for scientific web mining.

3 RELEVANCE TO THEWEBCONF

This tutorial is highly relevant to TheWebConf on the topic of web search and mining, focusing on an interdisciplinary direction that bridge web mining with scientific discovery in biomedicine and chemistry. The tutorial instructors has rich experience in delivering tutorials in major NLP (ACL, EMNLP, and NAACL), AI (AAAI and IJCAI), web mining (SIGIR and WWW), data mining (KDD and ICDM), and database (SIGMOD and VLDB) conferences (see Section 10 for more details).

4 TUTORIAL OUTLINE

This **lecture-style** tutorial presents a systematic overview of recent advances in NLP for scientific web mining. This tutorial is expected to be **1.5 hours** in duration. The contents are outlined below.

- Introduction [20min]
  - Motivations [5min]
  - Overview of Using NLP Techniques for Scientific Web Mining [10min]
  - Unique Challenges of Scientific Domains [5min]
- Scientific Information Extraction [20min]
  - Fine-grained Scientific Named Entity Recognition (NER) [10min]
    - Knowledge-Enhanced Fine-Grained Scientific NER
    - Ontology-guided Fine-Grained Scientific NER
  - Relation Extraction from Scientific Literature [10min]
    - Abstract Meaning Representation Guided Biomedical Relation Extraction
    - Meta-Pattern Guided Open Relation Extraction
- Scientific Information Retrieval [30min]
  - Textual Evidence Discovery from Scientific Literature [10min]
    - What is textual evidence discovery?
    - Textual Evidence Discovery in COVID-19 Literature
  - Cross-Modal Molecule Retrieval with Natural Language Queries [10min]
    - Multi-Modal Embedding to Represent Chemical Entities
    - Cross-Modal Molecule Retrieval
  - Chemical-Reaction-Aware Molecule Representation Learning [10min]
    - What is Molecule Representation Learning (MRL)?
    - Chemical-Reaction-Aware MRL
- System Demonstrations and Future Directions [20min]
  - System Demonstrations [10min]
    - COVID-KG: Multi-media Knowlege Extraction in COVID-19 Literature
    - ReactionTracker: Query-guided Reaction Tracking in Organic Chemistry Literature
  - Research Problems and Future Directions [10min]

5 TARGETED AUDIENCE

This tutorial is intended for researchers and practitioners in natural language processing, information retrieval, data mining, text mining, graph mining, machine learning, and their applications to other domains. While the audience with a good background in the above areas would benefit most from this tutorial, we believe the material to be presented would give general audience and newcomers a complete picture of the current work, introduce important research topics in this field, and inspire them to learn more. Our tutorial is designed as self-contained, so no specific background knowledge is assumed of the audience. However, it would be beneficial for the audience to know about basic deep learning technologies, pre-trained word embeddings (e.g. Word2Vec) and language models (e.g. BERT) before attending this tutorial. We will provide a reading list of background knowledge to the audience on our tutorial website.

6 PREVIOUS EDITIONS

The proposed tutorial is considered a **cutting-edge** tutorial that introduces the recent advances in an emerging area of using NLP techniques for scientific web mining. The presented topic has not been covered by previous ACL, EMNLP, NAACL, EACL, COLING, KDD, or TheWebConf tutorials in the past four years. This tutorial has not been presented elsewhere. We estimate that at least 60% of the works covered in this tutorial are from researchers other than the tutorial instructors.

7 TUTORIAL MATERIALS

We will provide attendees a **website** (and a link from the TheWebConf'22 tutorial website) and upload our tutorial materials (outline, slides, references, reading list, and pre-recorded videos) there. All the materials will be openly available and there is no copyright issue. Standard equipment will be enough for our tutorial and we will bring our own laptop and a wireless pointer.

8 VIDEO TEASER

A video teaser of our tutorial is available on YouTube: https://youtu.be/WopPIXm4ZeE.

9 ORGANIZATION DETAILS

This tutorial will be delivered both **in person** and **online** (e.g., via Zoom) during the conference. We will also provide **pre-recorded videos** as a backup plan that overcomes the potential occurrence of technical problems. We will release our tutorial website and all the materials one week in advance of the tutorial.

10 TUTORIAL INSTRUCTORS

Xuan Wang is a Ph.D. candidate at Computer Science Department, University of Illinois at Urbana-Champaign. Her research focuses on mining and constructing structured knowledge from massive unstructured corpora with minimum human supervision, emphasizing applications to biological and health sciences. She received M.S degree in Statistics and M.S. degree in Biochemistry from University of Illinois at Urbana-Champaign in 2017 and 2015, respectively, and B.S. degree in Biological Science from Tsinghua University in 2013. She is the recipient of YEE Fellowship Award in 2020-2021.

Hongwei Wang is a postdoctoral researcher at Computer Science Department, University of Illinois Urbana-Champaign. His research interests include machine learning and data mining, particularly in graph representation learning mechanisms, algorithms, and their applications in real-world data mining scenarios such as knowledge graphs, recommender systems, social networks, and sentiment analysis. He received Ph.D. degree from Department of Computer Science, Shanghai Jiao Tong University in 2018, and B.E. degree from ACM Class, Shanghai Jiao Tong University in 2014. He was a postdoctoral researcher at Computer Science Department, Stanford University, from 2019 to 2021. He was one of the recipients of 2020 CCF (China Computer Federation) Outstanding Doctoral Dissertation Award and 2018 Google Ph.D. Fellowship.

Heng Ji is a Professor at Computer Science Department of University of Illinois Urbana-Champaign, and an Amazon Scholar. She received her B.A. and M. A. in Computational Linguistics from Tsinghua University, and her M.S. and Ph.D. in Computer Science from New York University. Her research interests focus on Natural Language Processing, especially on Multimedia Multilingual Information Extraction, Knowledge Base Population and Knowledge-driven Generation. She was selected as "Young Scientist" and a member of the Global Future Council on the Future of Computing by the World Economic Forum in 2016 and 2017. The awards she received include "AI's 10 to Watch" Award by IEEE Intelligent Systems in 2013, NSF CAREER award in 2009, Google Research Award in 2009 and 2014, IBM Watson Faculty Award in 2012 and 2014 and Bosch Research Award in 2014-2018, Amazon AWS Award in 2021, ACL2020 Best Demo Paper Award, and NAACL2021 Best Demo Paper Award. She has coordinated the NIST TAC Knowledge Base Population task since 2010. She has served as the Program Committee Co-Chair of many conferences including NAACL-HLT2018. She is elected as the North American Chapter of the Association for Computational Linguistics (NAACL) secretary 2020-2021. Additional information is available at https://blender.cs.illinois.edu/hengji.html.

Jiawei Han is Michael Aiken Chair Professor, Department of Computer Science, University of Illinois at Urbana-Champaign. His research areas encompass data mining, text mining, data warehousing, and information network analysis, with over 1000 research publications. He is Fellow of ACM, Fellow of IEEE, and received numerous prominent awards, including ACM SIGKDD Innovation Award (2004) and IEEE Computer Society W. Wallace McDowell Award (2009). He delivered 50+ conference tutorials or keynote speeches (e.g., SIGKDD 2017-2021 tutorials and WSDM 2018 keynote).

The presenters of this tutorial have given the following tutorials at leading international conferences and venues in the past:

- **Xuan Wang**:
  - IEEE-BigData'19: Taming Unstructured Big Data: Automated Information Extraction from Massive Text
- **Heng Ji**:
  - AAAI'22: Deep Learning on Graphs for Natural Language Processing.
  - KDD'21: Deep Learning on Graphs for Natural Language Processing.
  - IJCAI'21: Deep Learning on Graphs for Natural Language Processing.
  - SIGIR'21: Deep Learning on Graphs for Natural Language Processing.
  - EMNLP'21: Knowledge-Enriched Natural Language Generation.
  - ACL'21: Event-Centric Natural Language Processing.
  - NAACL'21: Deep Learning on Graphs for Natural Language Processing.
  - AAAI'21: Event-Centric Natural Language Understanding.
  - CCL'18 and NLP-NADB'18: Multi-lingual Entity Discovery and Linking.
  - ACL'18: Multi-lingual Entity Discovery and Linking.
  - SIGMOD'16: Automatic Entity Recognition and Typing in Massive Text Data.
  - ACL'15: Successful Data Mining Methods for NLP.
  - ACL'14: Wikification and Beyond: The Challenges of Entity and Concept Grounding.
  - NLPCC'14: Wikification and Beyond: The Challenges of Entity and Concept Grounding.
  - COLING'12: Temporal Information Extraction and Shallow Temporal Reasoning.
- **Jiawei Han** (Recent Courses/Tutorials Mostly Related to This Tutorial):
  - AAAI'22: Pre-Trained Language Representations for Text Mining.
  - ICDM'21: Automated Taxonomy Discovery and Exploration"
  - KDD'21: On the Power of Pre-Trained Text Representations: Models and Applications in Text Mining
  - Winter School Short Course, "From Unstructured Text to TextCube: Automated Construction and Multi-dimensional Exploration", 6th Int. Winter School on Big Data (BigDat'2020), Ancona, Italy, Jan. 2020
  - KDD'20: Embedding-Driven Multi-Dimensional Topic Mining and Text Analysis
  - VLDB'19: TextCube: Automated Construction and Multi-dimensional Exploration
  - KDD'19: Constructing and Mining Heterogeneous Information Networks from Massive Text
  - Summer school short course, "Mining Structured Knowledge from Massive Text Data: A Data-Driven Approach", 3rd ACM European Summer School in Data Science, Athens, Greece, July 2019
  - KDD'18: Towards Multidimensional Analysis of Text Corpora
  - KDD'17: Mining Entity-Relation-Attribute Structures from Massive Text Data
  - SIGMOD'17: Building Structured Databases of Factual Knowledge from Massive Text Corpora
  - WWW'17: Constructing Structured Information Networks from Massive Text Corpora

ACKNOWLEDGMENTS

This material is based upon work supported in part by the Molecule Maker Lab Institute: An AI Research Institutes program supported by NSF under Award No. 2019897, US DARPA KAIROS Program No. FA8750-19-2-1004, U.S. DARPA AIDA Program No. FA8750-18-2-0014, Air Force Nos. FA8650-17-C-7715 and FA8750-20-2-10002, SocialSim Program No. W911NF-17-C-0099, and INCAS Program

No. HR001121C0165, and National Science Foundation IIS-19-56151, IIS-17-41317, and IIS 17-04532. Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and do not necessarily represent the views, either expressed or implied, of DARPA or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

REFERENCES

[1] Yonggang Cao, Feifan Liu, Pippa Simpson, Lamont Antieau, Andrew Bennett, James J Cimino, John Ely, and Hong Yu. 2011. AskHERMES: An online question answering system for complex clinical questions. *Journal of biomedical informatics* 44, 2 (2011), 277–288.

[2] A Filipa de Almeida, Rui Moreira, and Tiago Rodrigues. 2019. Synthetic organic chemistry driven by artificial intelligence. *Nature Reviews Chemistry* 3, 10 (2019), 589–604.

[3] Carl Edwards, ChengXiang Zhai Zhai, and Heng Ji. 2021. Text2Mol: Cross-Modal Molecule Retrieval with Natural Language Queries. In *Proceedings of the 2021 Empirical Methods in Natural Language Processing*.

[4] Carol Friedman, George Hripcsak, Lyuda Shagina, and Hongfang Liu. 1999. Representing information in patient reports using natural language processing and the extensible markup language. *Journal of the American Medical Informatics Association* 6, 1 (1999), 76–87.

[5] Rave Harpaz, Alison Callahan, Suzanne Tamang, Yen Low, David Odgers, Sam Finlayson, Kenneth Jung, Paea LePendu, and Nigam H Shah. 2014. Text mining for adverse drug events: the promise, challenges, and state of the art. *Drug safety* 37, 10 (2014), 777–790.

[6] Jiayuan He, Dat Quoc Nguyen, Saber A Akhondi, Christian Druckenbrodt, Camilo Thorne, Ralph Hoessel, Zubair Afzal, Zenan Zhai, Biaoyan Fang, Hiyori Yoshikawa, et al. 2020. Overview of ChEMU 2020: named entity recognition and event extraction of chemical reactions from patents. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 237–254.

[7] Jingshan Huang, Fernando Gutierrez, Dejing Dou, Judith A Blake, Karen Eilbeck, Darren A Natale, Barry Smith, Yu Lin, Xiaowei Wang, Zixing Liu, et al. 2015. A semantic approach for knowledge capture of microRNA-target gene interactions. In *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 975–982.

[8] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. 2019. PubChem 2019 update: improved access to chemical data. *Nucleic acids research* 47, D1 (2019), D1102–D1109.

[9] Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A Shoemaker, et al. 2016. PubChem substance and compound databases. *Nucleic acids research* 44, D1 (2016), D1202–D1213.

[10] Martin Krallinger, Florian Leitner, Obdulia Rabal, Miguel Vazquez, Julen Oyarzabal, and Alfonso Valencia. 2015. CHEMDNER: The drugs and chemical names extraction challenge. *Journal of cheminformatics* 7, 1 (2015), S1.

[11] Martin Krallinger, Obdulia Rabal, Analia Lourenco, Julen Oyarzabal, and Alfonso Valencia. 2017. Information retrieval and text mining technologies for chemistry. *Chemical reviews* 117, 12 (2017), 7673–7761.

[12] Tuan Lai, Heng Ji, ChengXiang Zhai, and Quan Hung Tran. 2021. Joint Biomedical Entity and Relation Extraction with Knowledge-Enhanced Collective Inference. *arXiv preprint arXiv:2105.13456* (2021).

[13] Qi Li, Xuan Wang, Yu Zhang, Fei Ling, Cathy Wu H, and Jiawei Han. 2018. Pattern Discovery for Wide-Window Open Information Extraction in Biomedical Literature. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 420–427.

[14] S Hessam M Mehr, Matthew Craven, Artem I Leonov, Graham Keenan, and Leroy Cronin. 2020. A universal system for digitization and automatic execution of the chemical synthesis literature. *Science* 370, 6512 (2020), 101–108.

[15] Alexander Nikitin, Sergei Egorov, Nikolai Daraselia, and Ilya Mazo. 2003. Pathway studio – the analysis and navigation of molecular networks. *Bioinformatics* 19, 16 (2003), 2155–2157.

[16] Dietrich Rebholz-Schuhmann, Anika Oellrich, and Robert Hoehndorf. 2012. Text-mining solutions for biomedical research: enabling integrative biology. *Nature Reviews Genetics* 13, 12 (2012), 829.

[17] Fabio Rinaldi, Simon Clematide, Hernani Marques, Tilia Ellendorff, Martin Romacker, and Raul Rodriguez-Esteban. 2014. OntoGene web services for biomedical text mining. *BMC bioinformatics* 15, 14 (2014), S6.

[18] Damian Szklarczyk, John H Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T Doncheva, Alexander Roth, and Peer Bork. 2017. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic acids research* 45, D1 (2017), D362–D368.

[19] Damian Szklarczyk, Alberto Santos, Christian von Mering, Lars Juhl Jensen, Peer Bork, and Michael Kuhn. 2015. STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data. *Nucleic acids research* 44, D1 (2015), D380–D384.

[20] Hongwei Wang, Weijiang Li, Xiaomeng Jin, Kyunghyun Cho, Heng Ji, Jiawei Han, and Martin D Burke. 2021. Chemical-Reaction-Aware Molecule Representation Learning. *arXiv preprint arXiv:2109.09888* (2021).

[21] Qingyun Wang, Manling Li, Xuan Wang, Nikolaus Parulian, Guangxing Han, Jiawei Ma, Jingxuan Tu, Ying Lin, Haoran Zhang, Weili Liu, et al. 2020. COVID-19 literature knowledge graph construction and drug repurposing report generation. *arXiv preprint arXiv:2007.00576* (2020).

[22] Xuan Wang, Yingjun Guan, Weili Liu, Aabhas Chauhan, Enyi Jiang, Qi Li, David Liem, Dibakar Sigdel, John Caufield, Peipei Ping, and Jiawei Han. 2020. EVIDENCEMINER: Textual Evidence Discovery for Life Sciences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Online, 56–62. https://doi.org/10.18653/v1/2020.acl-demos.8

[23] Xuan Wang, Vivian Hu, Xiangchen Song, Shweta Garg, Jinfeng Xiao, and Jiawei Han. 2021. ChemNER: Fine-Grained Chemistry Named Entity Recognition with Ontology-guided Distant Supervision. In *Proceedings of the 2021 Empirical Methods in Natural Language Processing*.

[24] Xuan Wang, Xiangchen Song, Bangzheng Li, Kang Zhou, Qi Li, and Jiawei Han. 2020. Fine-Grained Named Entity Recognition with Distant Supervision in COVID-19 Literature. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 491–494.

[25] Xuan Wang, Yu Zhang, Aabhas Chauhan, Qi Li, and Jiawei Han. 2020. Textual Evidence Mining via Spherical Heterogeneous Information Network Embedding. In *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 828–837.

[26] Xuan Wang, Yu Zhang, Qi Li, Yinyin Chen, and Jiawei Han. 2018. Open Information Extraction with Meta-pattern Discovery in Biomedical Literature. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 291–300.

[27] Xuan Wang, Yu Zhang, Qi Li, Xiang Ren, Jingbo Shang, and Jiawei Han. 2019. Distantly supervised biomedical named entity recognition with dictionary expansion. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 496–503.

[28] Taiki Watanabe, Akihiro Tamura, Takashi Ninomiya, Takuya Makino, and Tomoya Iwakura. 2019. Multi-Task Learning for Chemical Named Entity Recognition with Chemical Compound Paraphrasing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 6244–6249. https://doi.org/10.18653/v1/D19-1648

[29] Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. 2013. PubTator: a web-based text mining tool for assisting biocuration. *Nucleic acids research* 41, W1 (2013), W518–W522.

[30] Thomas C Wiegers, Allan Peter Davis, and Carolyn J Mattingly. 2014. Web services-based text-mining demonstrates broad impacts for interoperability and process simplification. *Database* 2014 (2014).

[31] Boya Xie, Qin Ding, Hongjin Han, and Di Wu. 2013. miRCancer: a microRNA–cancer association database constructed by text mining on literature. *Bioinformatics* 29, 5 (2013), 638–644.

[32] Zixuan Zhang, Nikolaus Parulian, Heng Ji, Ahmed Elsayed, Skatje Myers, and Martha Palmer. 2021. Fine-grained Information Extraction from Biomedical Literature based on Knowledge-enriched Abstract Meaning Representation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 6261–6270.