AI for Science in the Era of Large Language Models

Zhenyu Bi[1], **Minghao Xu**[2], **Jian Tang**[2], **Xuan Wang**[1]
[1]Department of Computer Science, Virginia Tech, USA
[2]Mila - Quebec AI Institute, Canada
[1]{zhenyub,xuanw}@vt.edu,
[2]minghao.xu@mila.quebec, [2]tangjianpku@gmail.com

Abstract

The capabilities of AI in the realm of science span a wide spectrum, from the atomic level, where it solves partial differential equations for quantum systems, to the molecular level, predicting chemical or protein structures, and even extending to societal predictions like infectious disease outbreaks. Recent advancements in large language models (LLMs), exemplified by models like ChatGPT, have showcased significant prowess in tasks involving natural language, such as translating languages, constructing chatbots, and answering questions. When we consider scientific data, we notice a resemblance to natural language in terms of sequences – scientific literature and health records presented as text, bio-omics data arranged in sequences, or sensor data like brain signals. The question arises: Can we harness the potential of these recent LLMs to drive scientific progress? In this tutorial, we will explore the application of large language models to three crucial categories of scientific data: 1) textual data, 2) biomedical sequences, and 3) brain signals. Furthermore, we will delve into LLMs' challenges in scientific research, including ensuring trustworthiness, achieving personalization, and adapting to multi-modal data representation.

1 Tutorial Content

The impressive capabilities of Artificial Intelligence (AI) within the realm of science span a wide spectrum, from the atomic level, where it attempts to solve partial differential equations for quantum systems, to the molecular level, where it accurately predicts the structures of chemicals and proteins, and extends even further, encompassing societal predictions like forecasting infectious disease outbreaks (Zhang et al., 2023a). Amidst this landscape of possibilities, recent advancements in large language models (LLMs), notably exemplified by models like ChatGPT[1], have risen to the forefront, demonstrating significant proficiency in tasks tied to natural language. These tasks include language translation, constructing chatbots, and answering questions (Yang et al., 2023).

Interestingly, when we turn our attention to scientific data, we discover a striking resemblance to natural language in terms of sequences. Scientific literature and health records are laid out as textual narratives, bio-omics data takes the form of molecular sequences, and even sensor data like brain signals is inherently sequential (Wang et al., 2021a; Thirunavukarasu et al., 2023). This observation prompts a compelling question: Can we leverage the potential of these advanced LLMs to propel scientific advancement?

In this tutorial, we embark on a journey to explore precisely this intersection—the fusion of cutting-edge large language models with scientific inquiry. Our exploration zooms in on three pivotal categories of scientific data: 1) textual data (Alsentzer et al., 2019; Singhal et al., 2022; Beltagy et al., 2019; Lee et al., 2020; Gu et al., 2021; Alrowili and Vijay-Shanker, 2021; Yasunaga et al., 2022), 2) biomedical sequences (Ji et al., 2021; Zvyagin et al., 2022; Fishman et al., 2023; Dalla-Torre et al., 2023; Nguyen et al., 2023; Yamada and Hamada, 2022; Yang et al., 2022; Chen et al., 2022; Zhang et al., 2023b; Rives et al., 2021; Bepler and Berger, 2021; Brandes et al., 2022; Madani et al., 2023; Lin et al., 2023; Zheng et al., 2023; Xu et al., 2023), and 3) brain signals (Wang et al., 2022a; Wang and Ji, 2022; Tang et al., 2023). By drawing inspiration from the transformative capabilities of LLMs, we seek to unravel novel understanding and innovation within each domain.

As we move forward, we further discuss the intricate challenges that accompany the infusion of AI into scientific research. The foundation of trustworthiness stands tall—how do we ensure the reliability of AI-enhanced scientific insights? The concept of personalization emerges as a critical

[1] https://chat.openai.com/chat

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts, pages 32–38
November 12–16, 2024. ©2024 Association for Computational Linguistics

consideration, urging us to tailor LLMs to the specific contours of scientific investigation. Furthermore, the multi-dimensional nature of scientific data beckons us to master the art of handling data representations that span across various modalities.

2 Tutorial Type

This is a **cutting-edge** tutorial, bridging the gap between the NLP community and AI for Science.

3 Target Audience and Prerequisites

This tutorial is intended for researchers and practitioners in natural language processing, machine learning, and their applications to science 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 the general audience and newcomers a complete picture of the important research topics in AI for science with large language models. 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 the basics of deep learning technologies and pre-trained language models (e.g., Transformer (Vaswani et al., 2017), BERT (Kenton and Toutanova, 2019), GPT (Brown et al., 2020), and T5 (Raffel et al., 2020)) before attending this tutorial. We will provide a reading list of background knowledge on our tutorial website.

4 Tutorial Outline

This tutorial is expected to be **3 hours** in duration plus a **30-minute break** in between. The contents are outlined below.

4.1 Background and Motivation [20 min]

We will first introduce the background knowledge of LLMs and the big picture of AI for Science. Then we will motivate the following topics of LLMs for science on three pivotal categories of scientific data: 1) textual data, 2) biomedical sequences, and 3) brain signals.

4.2 LLMs on Scientific Textual Data [40 min]

First, we introduce LLMs in the realm of scientific textual data, which encompasses diverse domains like biomedical literature (Beltagy et al., 2019; Lee et al., 2020; Gu et al., 2021; Alrowili and Vijay-Shanker, 2021; Yasunaga et al., 2022) and electronic health records (Alsentzer et al., 2019; Singhal et al., 2022). This form of scientific textual data closely mirrors the fundamental structure of large language models. It finds extensive utility across science and healthcare, facilitating tasks such as extracting valuable information and responding to queries. The applicability spans a multitude of areas, underpinning scientific and healthcare endeavors for information extraction (Wang et al., 2021b; Zhong et al., 2023; Wang et al., 2022b) and question-answering (Krithara et al., 2023).

4.3 LLMs on Biomedical Sequences [60 min]

Next, we extend the application of LLMs to the intricate realm of biological sequence data, where a rich landscape of possibilities emerges. Within this domain, we shift our focus to three distinct yet interwoven categories of biological sequences:

DNA sequences: From the blueprint of life, we draw inspiration as we delve into works such as (Ji et al., 2021), (Zvyagin et al., 2022), (Fishman et al., 2023), (Dalla-Torre et al., 2023), and (Nguyen et al., 2023). These pioneering endeavors pave the way for unraveling the secrets encrypted within the very essence of organisms. The DNA LLMs have a wide application in downstream tasks such as predicting regulatory elements for enhancers, promoters, epigenetic marks, and splice sites from DNA sequences (Grešová et al., 2023; Dalla-Torre et al., 2023).

RNA sequences: Navigating the intricate world of gene expression, we embrace the innovative contributions outlined in (Yamada and Hamada, 2022), (Yang et al., 2022), (Chen et al., 2022), and (Zhang et al., 2023b). These strides empower us to decode the symphony of biological processes orchestrated by RNA. The RNA LLMs have a wide application in RNA structure and function prediction (Yamada and Hamada, 2022; Zhang et al., 2023b), RNA-protein interaction prediction (Chen et al., 2022), and cell type annotation (Yang et al., 2022).

Protein sequences: Venturing into the complex realm of proteins, we are guided by luminous works like (Rives et al., 2021), (Bepler and Berger, 2021), (Brandes et al., 2022), (Madani et al., 2023), (Lin et al., 2023), (Zheng et al., 2023), and (Xu et al., 2023). These endeavors illuminate the path to unraveling the intricate choreography of molecular functions and interactions. The protein LLMs have a wide application in functional protein generation

(Leinonen et al., 2004) and protein structure prediction (Suzek et al., 2015).

Within these domains, the transformative capabilities of LLMs manifest in a myriad of high-impact downstream applications. From predicting molecular structures to forecasting molecule interactions, and from unraveling molecule functions to drawing poignant associations with disease progression processes, LLMs stand as beacons of innovation, guiding us towards a deeper comprehension of life's building blocks.

4.4 LLMs on Brain Signals [30 min]

Last, we delve into the fascinating realm of applying LLMs to the realm of brain signals. In this section, we start with the introduction of a pioneering pre-trained brain signal representation model, as detailed in (Wang et al., 2022a). Building upon this foundation, we further introduce an exciting topic of open-vocabulary brain-to-text translation (Wang and Ji, 2022; Tang et al., 2023). This intriguing endeavor involves training translation models to automatically decipher the intricate contents of individuals' thoughts, offering a captivating glimpse into the potential convergence of technology and cognitive processes.

4.5 Future Research Directions [30 min]

As a conclusion, we will take a closer look at the challenges that come with using AI in scientific research. One big challenge is making sure that the scientific insights enhanced by AI are reliable and trustworthy, including model explainability and interpretability, model robustness to adversarial attacks, model bias towards different populations, and data privacy issues. We also think about the idea of personalization, which means adjusting LLMs to fit the specific needs of different personalized data. For example, there is a large individual variance in brain signals when different people are thinking of the same word under the same context. Instead of using one LLM to fit everyone, can we construct personalized LLMs based on different brain patterns for different people? And since scientific information can be very varied, we learn how to handle different types of data in a skillful and effective way. For example, Google has announced Med-PaLM-2 (Singhal et al., 2023) that integrates image, text, and genome data in the electronic health record, declaring an expert-level ability for medical question answering. Can we develop more effective and efficient methods to integrate multi-modal and multi-omic LLMs into one powerful unified LLM?

5 Others People's Work

We will include a broad spectrum of other people's work that consists of **more than 60%** of the tutorial content (see References).

6 Diversity Consideration

We will discuss large language models scaled up to various scientific domains and various data formats (textual data, biomedical sequences, and brain signals). Our instructors consist of PhD students (Zhenyu Bi and Minghao Xu), junior faculty (Xuan Wang, Assistant Professor), and senior faculty (Jian Tang, Associate Professor). Our instructors also came from diverse geographical locations (Zhenyu Bi and Xuan Wang from Virginia Tech in the US, and Minghao Xu and Jian Tang from Mila - Quebec AI Institute in Canada). We plan to involve inclusive topics, accessible materials, diverse instructors, flexible formats, and targeted outreach to ensure a broad and varied audience engagement.

7 Reading List

We will provide a reading list of background knowledge on our tutorial website. A preliminary reading list can be found as the References.

8 Tutorial Presenters

Zhenyu Bi is a Ph.D. student in the Computer Science Department at Virginia Tech. His research area lies in the field of natural language processing, emphasizing real-world applications of Large Language Models. He is mainly interested in information extraction with weak supervision, especially text mining and event extraction; as well as fact-checking and trustworthy NLP. He received an M.S. degree in Intelligent Information Systems from Carnegie Mellon University in 2023, a B.S. degree in Cognitive Science, and a B.S. Degree in Computer Science from the University of California, San Diego in 2021.

Minghao Xu is a Ph.D. student at Mila - Quebec AI Institute, Canada. His research interests mainly lie in protein function understanding and protein design. He aims to understand diverse protein functions with joint guidance from protein sequences, structures, and biomedical text, especially boosted by large-scale multi-modal pre-training. He is also pursuing structure- and sequence-based protein design via generative AI, geometric deep learning and dry-wet experiment closed looping. He has given an Oral presentation at the main conference of ICML'23.

Jian Tang is an Associate Professor at Mila - Quebec AI Institute, Canada. His long-term interests focus on understanding the language of life (DNA, RNAs, and Proteins) with generative AI and geometric deep learning, with applications in biomedicine and synthetic biology. His group has developed one of the first open-source machine learning frameworks on drug discovery, TorchDrug (for small molecules) and TorchProtein (for proteins), and developed the first diffusion models for 3D molecular structure generation, GeoDiff (among the 50 most cited AI paper in 2022). He has given a few tutorials at international AI and data mining conferences including KDD 2017, AAAI 2019, AAAI 2022.

Xuan Wang is an Assistant Professor in the Computer Science Department at Virginia Tech. Her research focuses on natural language processing and text mining, emphasizing applications to science and healthcare domains. Her current projects include NLP and text mining with extremely weak supervision; text-augmented knowledge graph reasoning; fact-checking and trustworthy NLP, AI for science; and AI for healthcare. She received a Ph.D. degree in Computer Science, an M.S. degree in Statistics, and an M.S. degree in Biochemistry from the University of Illinois Urbana-Champaign in 2022, 2017, and 2015, respectively, and a B.S. degree in Biological Science from Tsinghua University in 2013. She has delivered tutorials in IEEE-BigData 2019, WWW 2022, and KDD 2022.

9 Estimated Audience Size

This is a cutting-edge tutorial that introduces new frontiers in the intersection of NLP and AI for Science. The presented topic has not been covered by ACL/EMNLP/NAACL/EACL/COLING tutorials in the past four years. It is hard to give an estimate of audience size given no similar tutorials have been delivered before. A rough estimate would be around **tens to hundreds of participants**.

10 Preferred Venues

We prefer the following venues for this tutorial: 1) ACL, 2) EMNLP, and 3) NAACL.

11 Technique Requirement

Standard equipment will be enough for our tutorial and we don't have specific requirements. We will bring our own laptop and a wireless pointer.

12 Presentation Materials

We will provide tutorial materials (e.g., tutorial slides and relevant list of papers) **one month** prior to the date of the tutorial. The tutorial materials will be **publically available** for open access.

13 Ethics Statement

Ethical quandaries frequently confront technological advancements, especially when it comes to dual-use scenarios where an innovation can bring both advantages and disadvantages. The tutorial introduces IE technologies, where the distinction between beneficial and detrimental employment predominantly hinges on data usage. Employing this technology responsibly necessitates the lawful and ethical acquisition of input text collections and other forms of input.

Regulations and standards establish a legal framework to ensure appropriate data utilization, granting individuals the right to request the removal of their data. In the absence of such regulations, the ethical responsibility falls upon technology practitioners to uphold righteous data use. Moreover, biases can infiltrate training and evaluation data, potentially diminishing system accuracy for underrepresented groups or in novel domains. This bias can result in performance disparities based on attributes like ethnicity, race, and gender.

Additionally, systems trained on specific data can experience degradation when confronted with new, dissimilar data. This accentuates the need to thoughtfully contemplate matters of fairness and generalizability when employing IE technologies with particular datasets.

To guarantee the conscientious application of dual-use technology, a comprehensive approach involves prioritizing ethical considerations as foundational principles during every phase of system design. Transparency and interpretability should remain paramount across data, algorithms, models, and functionality within the system. Public verification and auditing can be facilitated by making software open source. Furthermore, strategies to safeguard marginalized groups should be explored as a part of ethical technology deployment.

Acknowledgement

Our work is sponsored by the Commonwealth Cyber Initiative and a generous gift from the Amazon + VT Center for Efficient and Robust ML.

References

Sultan Alrowili and K Vijay-Shanker. 2021. Biomtransformers: building large biomedical language models with bert, albert and electra. In *Proceedings of the 20th workshop on biomedical language processing*, pages 221–227.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, WA Redmond, and Matthew BA McDermott. 2019. Publicly available clinical bert embeddings. *NAACL HLT 2019*, page 72.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. 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)*, pages 3615–3620.

Tristan Bepler and Bonnie Berger. 2021. Learning the protein language: Evolution, structure, and function. *Cell systems*, 12(6):654–669.

Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. 2022. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jiayang Chen, Zhihang Hu, Siqi Sun, Qingxiong Tan, Yixuan Wang, Qinze Yu, Licheng Zong, Liang Hong, Jin Xiao, Tao Shen, et al. 2022. Interpretable rna foundation model from unannotated data for highly accurate rna structure and function predictions. *bioRxiv*, pages 2022–08.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Hassan Sirelkhatim, Guillaume Richard, et al. 2023. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, pages 2023–01.

Veniamin Fishman, Yuri Kuratov, Maxim Petrov, Aleksei Shmelev, Denis Shepelin, Nikolay Chekanov, Olga Kardymon, and Mikhail Burtsev. 2023. Gena-lm: A family of open-source foundational models for long dna sequences. *bioRxiv*, pages 2023–06.

Katarína Grešová, Vlastimil Martinek, David Čechák, Petr Šimeček, and Panagiotis Alexiou. 2023. Genomic benchmarks: a collection of datasets for genomic sequence classification. *BMC Genomic Data*, 24(1):25.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. 2021. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. 2023. Bioasq-qa: A manually curated corpus for biomedical question answering. *Scientific Data*, 10(1):170.

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

Rasko Leinonen, Federico Garcia Diez, David Binns, Wolfgang Fleischmann, Rodrigo Lopez, and Rolf Apweiler. 2004. Uniprot archive. *Bioinformatics*, 20(17):3236–3237.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130.

Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos Jr, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. 2023. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, pages 1–8.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Callum Birch-Sykes, Michael Wornow, Aman Patel, Clayton Rabideau, Stefano Massaroli, Yoshua Bengio, et al. 2023. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *arXiv preprint arXiv:2306.15794*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Baris E Suzek, Yuqi Wang, Hongzhan Huang, Peter B McGarvey, Cathy H Wu, and UniProt Consortium. 2015. Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932.

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G Huth. 2023. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, pages 1–9.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature Medicine*, pages 1–11.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Benyou Wang, Qianqian Xie, Jiahuan Pei, Zhihong Chen, Prayag Tiwari, Zhao Li, and Jie Fu. 2021a. Pre-trained language models in biomedical domain: A systematic survey. *ACM Computing Surveys*.

Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. 2022a. Brainbert: Self-supervised representation learning for intracranial recordings. In *The Eleventh International Conference on Learning Representations*.

Xuan Wang, Vivian Hu, Minhao Jiang, Yu Zhang, Jinfeng Xiao, Danielle Cherrice Loving, Heng Ji, Martin Burke, and Jiawei Han. 2022b. Reactclass: Cross-modal supervision for subword-guided reactant entity classification. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 844–847. IEEE.

Xuan Wang, Vivian Hu, Xiangchen Song, Shweta Garg, Jinfeng Xiao, and Jiawei Han. 2021b. Chemner: fine-grained chemistry named entity recognition with ontology-guided distant supervision. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Zhenhailong Wang and Heng Ji. 2022. Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5350–5358.

Minghao Xu, Xinyu Yuan, Santiago Miret, and Jian Tang. 2023. Protst: Multi-modality learning of protein sequences and biomedical texts. *arXiv preprint arXiv:2301.12040*.

Keisuke Yamada and Michiaki Hamada. 2022. Prediction of rna–protein interactions using a nucleotide language model. *Bioinformatics Advances*, 2(1):vbac023.

Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. 2022. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence*, 4(10):852–866.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *arXiv preprint arXiv:2304.13712*.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. Linkbert: Pretraining language models with document links. In *Association for Computational Linguistics (ACL)*.

Xuan Zhang, Limei Wang, Jacob Helwig, Youzhi Luo, Cong Fu, Yaochen Xie, Meng Liu, Yuchao Lin, Zhao Xu, Keqiang Yan, et al. 2023a. Artificial intelligence for science in quantum, atomistic, and continuum systems. *arXiv preprint arXiv:2307.08423*.

Yikun Zhang, Mei Lang, Jiuhong Jiang, Zhiqiang Gao, Fan Xu, Thomas Litfin, Ke Chen, Jaswinder Singh, Xiansong Huang, Guoli Song, et al. 2023b. Multiple sequence-alignment-based rna language model and its application to structural inference. *bioRxiv*, pages 2023–03.

Zaixiang Zheng, Yifan Deng, Dongyu Xue, Yi Zhou, Fei Ye, and Quanquan Gu. 2023. Structure-informed language models are protein designers. *bioRxiv*, pages 2023–02.

Ming Zhong, Siru Ouyang, Minhao Jiang, Vivian Hu, Yizhu Jiao, Xuan Wang, and Jiawei Han. 2023. Reactie: Enhancing chemical reaction extraction with weak supervision. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12120–12130.

Maxim Zvyagin, Alexander Brace, Kyle Hippe, Yuntian Deng, Bin Zhang, Cindy Orozco Bohorquez, Austin Clyde, Bharat Kale, Danilo Perez-Rivera, Heng Ma, et al. 2022. Genslms: Genome-scale language models reveal sars-cov-2 evolutionary dynamics. *bioRxiv*, pages 2022–10.