Language + Molecules

Carl Edwards and **Qingyun Wang** and **Heng Ji**
University of Illinois Urbana-Champaign
{cne2, qingyun4, hengji}@illinois.edu

1 Description

Climate change, access to food and water, pandemics— these words, when uttered, immediately summon to mind global challenges with possible disastrous outcomes. The world faces enormous problems in the coming decades on scales of complexity never-before-seen. To address these issues, developing scientific solutions which are scalable, flexible, and inexpensive is critical. Further, we need to develop these solutions quickly. Broadly speaking, chemistry can provide molecular solutions to many of these problems: breakthrough drugs (e.g., kinase inhibitors (Ferguson and Gray, 2018)), materials (e.g., organic photovoltaics (Kippelen et al., 2009)), and chemical processes. The extremely large search spaces in which these solutions exist make AI tools critical for finding them. Of particular note, multimodal models combining language with molecules are poised to be a critical tool for discovering these solutions (Zhang et al., 2023). In this tutorial, we will discuss the role which natural language processing can play in discovering and accelerating solutions to global problems via the broad chemistry domain.

One of the first questions that probably comes to mind is why we would want to integrate natural language with molecules. Succinctly, combining these types of information has the possibility to accelerate scientific discovery. As motivating scenarios, imagine a future where a doctor can receive a novel, patient-specific drug necessary to treat an ailment just by writing a few sentences describing the patient's symptoms (also taking into account their genotype, phenotype, and medical history). Or, imagine a scientist tackling challenging problems by designing a molecule satisfying desired functions (e.g., antimalarial or a photovoltaic) rather than its structure or low level properties (e.g., solubility). Controlling molecules and drug design in such a high-level manner has potential to be hugely impactful, but it requires a method of abstract description; luckily, humans have already developed one: natural language.

In recent months, because of this potential impact, significant attention and growth has occurred in scientific NLP and AI research, including integration of molecules with natural language and multimodal AI for science/medicine ((Zhang et al., 2023) Section 10.3.3, (Wang et al., 2023)). We believe a sufficient amount of work has now been done, along with significant interest generated, to propose an **Introductory to NLP** (yet still **Cutting-Edge**) tutorial on "Language + Molecules". This tutorial is designed to require no knowledge and will enable participants to begin exploring relevant and impactful research. Since most relevant work is still cutting-edge, this will broaden the community's understanding of the associated challenges, methodologies, and goals in multimodal molecule-language models. We will present an interactive hands-on example and release accompanying relevant code and resources. The tutorial will precede and prepare the way for the Language+Molecules workshop later in the year at ACL.

2 Outline [180 min.]

Applying language models to the scientific domain is becoming increasingly popular due to its potential impact for accelerating scientific discovery (Hope et al., 2022). Beyond extracting information from scientific literature, NLP has the possibility to increase control of the scientific discovery process, which can be achieved through multimodal representations and generative language models.

2.1 Background [60 min.]

Scientific Information Extraction [15 min.]

To start, we will provide a high-level overview on traditional NLP tasks used for scientific discovery (e.g., named entity recognition, entity linking, and relation extraction), as well as recent domain-specific LLMs designed for superior performance

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Tutorial Abstracts, pages 14–20

on scientific tasks (Beltagy et al., 2019).

What is a molecule? [15 min.]

Half of the title is molecules, but what is one? We will start from scratch and discuss what a molecule actually is, including the basic constituents of molecules, atoms and bonds, and how they essentially form graph structures. Then, we will focus on molecular string languages, which are a key building block for chemical language models. We will discuss tradeoffs of these languages (Grisoni, 2023; Weininger, 1988; O'Boyle and Dalke, 2018; Krenn et al., 2020; Cheng et al., 2023). Krenn et al. (2020) proposes a formal grammar approach, which may particularly interest the ACL community.

Molecule Design using Language Models [15]

Now that we know what a molecule is, we will overview recent work applying NLP techniques to these molecular languages with impressive results. These molecular LLMs are trained with adapted pre-training techniques from (natural) language models to learn molecule representation from large collections of molecule strings (Frey et al., 2022; Chithrananda et al., 2020; Ahmad et al., 2022; Fabian et al., 2020; Schwaller et al., 2021; NVIDIA Corporation, 2022; Flam-Shepherd and Aspuru-Guzik, 2023; Tysinger et al., 2023). Applications include molecule and material generation, property prediction, and protein binding site prediction.

Drug Discovery–A Brief Primer [15 min.]

Ok, so NLP is being used for molecules now. What can we do with it?—here, we present a brief overview of drug discovery–an important but challenging problem. Historically, molecular discovery has commonly been done by humans who design and build individual molecules, but this can cost over a billion dollars and take over ten years (Gaudelet et al., 2021). We'll discuss a little of the process here, including non-NLP deep learning methods, so that we know how to improve it.

2.2 Integrating Language with Molecules [95]

What does natural language have to offer? [15]

At least at first, integrating languages and molecules seems like an odd idea. Here, we'll start an interactive discussion with the audience on what they think potential benefits might be. We'll make sure to mention the following major advantages, as discussed in the recent survey (Zhang et al., 2023):

1. **Generative Modeling**: One of the largest problems in current LLMs—-hallucination—becomes a strength for discovering molecules with high-level functions and abstract properties. In particular, language is compositional by nature (Szabó, 2020; Partee et al., 1984; Han et al., 2023), and therefore holds promise for composing these high-level properties (e.g., antimalarial) (Liu et al., 2022).
2. **Bridging Modalities**: Language can serve to "bridge" between modalities for scarce data.
3. **Domain Understanding**: Grounding language models into external real world knowledge (here, molecular structures) can improve understanding of unseen molecules and advance many emerging tasks, such as experimental procedure planning, which use LLMs as scientific agents.
4. **Democratization**: Language enables scientists without computational expertise to leverage advances in scientific AI.

Do I want multimodality? [5 min.]

An important, yet often overlooked, question in multimodal NLP is to ask: do I need multimodality? For example, if one wants to extract reactions from the literature, a text-to-text model (Vaucher et al., 2020) might be sufficient. However, editing a drug with high-level instructions requires language (Liu et al., 2023a; Fang et al., 2023). Here, we will dive into this question and discuss example scenarios with the audience for how to answer it.

2.2.1 Integrating Modalities [30 min.]

Ok, we've decided we want or need multimodality. Next, we need to discuss how people are currently tackling this-we'll start with two primary methods, bi-encoder models and joint representation models.

Bi-Encoder Models (and beyond) Bi-encoder models consist of an encoder branch for text and a branch for molecules. They have the advantage of not requiring direct, early integration of the two modalities, allowing existing single-modal models to be integrated. Representative examples we will discuss include Text2Mol (Edwards et al., 2021), CLAMP (Seidl et al., 2023), and BioTranslator (Xu et al., 2023). Generally, bi-encoder models are effective for cross-modal retrieval (Edwards et al., 2021; Su et al., 2022; Liu et al., 2022; Zhao et al., 2023b), but they may also be integrated into molecule (Su et al., 2022; Liu et al., 2022) and protein (Liu et al., 2023b) generation frameworks.

We'll talk about all these tasks, applications, and return to some important motivations (e.g., bridging modalities).

Joint Molecule-Language Models Joint models, on the other hand, seeks to model interactions between multiple modalities inside the same network to allow fine-grained interaction. We will discuss encoder-only models (Zeng et al., 2022), encoder-decoder models (Christofidellis et al., 2023), and decoder-only models (Liu et al., 2023c).

Model Differences: We will answer important questions such as: Which model should I use? What tasks can each do? Tasks include retrieval (Edwards et al., 2021), "translation" between molecules and language (Edwards et al., 2022a), editing molecules (Liu et al., 2022), and chemical reaction planning (Vaucher et al., 2020, 2021).

An Interactive Example - Targeting Microtubules for Cancer Treatment [20 min.]

At this point, there's been a lot of ideas thrown around. We'll consolidate them by exploring an interactive example of language-enabled molecule design using Google Colab.

We will focus on microtubules for the example. These cellular structures play an important role in many processes such cell growth and division, and mutations can be oncogenic (Mukhtar et al., 2014; Wattanathamsan and Pongrakhananon, 2022). In modern medicine, tumors such as pancreatic cancer are commonly treated by microtubule-targeting drugs such as paclitaxel (Albahde et al., 2021). In our example, we will explore creating new drugs with this function using natural language instructions, which may be useful in cases of paclitaxel resistance (Kavallaris, 2010). Our hands-on example will consist of three components:

1. **Language-enabled Drug Design:** Participants will explore inputs to language→molecule models to generate candidate drugs which target microtubules.
2. **Language-Guided Assay Testing:** Here, participants will test their proposed drugs in an assay. We will follow (Seidl et al., 2023), where natural language descriptions are used for assay predictions.
3. **Interaction Prediction:** Finally, we will test if proposed drugs bind with beta-tubulin using Autodock Vina, a well established docking program (Trott and Olson, 2010), via DockString (García-Ortegón et al., 2022).

Applications [25 min.] Here, we will discuss important applications to improve cross-discipline communication, including drug discovery (Mukhtar et al., 2014; Ferguson and Gray, 2018), organic photovoltaics (Kippelen et al., 2009), and catalyst discovery for renewable energy (Zitnick et al., 2020).

2.3 Recent Trends and Conclusion [25 min.]

Instruction-Following Molecular Design [10]

In the last year, instruction-following language models (Wei et al., 2021) have surged in popularity. Following this trend, training methodologies and datasets have recently emerged to allow language models to follow instructions related to molecule properties (Liang et al., 2023; Fang et al., 2023; Zeng et al., 2023; Zhao et al., 2023a). We will give a brief overview of this new line of work.

LLMs as Scientific Agents [5 min.] Further, we'll focus on recent work which looks to control experiments with language models (Boiko et al., 2023) and to create tools for enabling domain-specific capabilities in general language models (Bran et al., 2023; Liu et al., 2023a).

Conclusion [10 min.] We will discuss the key difficulties in the molecule-language domain that need to be addressed by the research community to allow similar progress to the vision-language domain. This includes 1) data scarcity due to domain expertise requirements, 2) addressing inconsistency when training on scientific literature, 3) improved methods for integrating geometric structures into LLMs, and 4) developing better evaluation metrics for chemical predictions without real-world experiments.

3 Logistics and Details

Diversity Considerations For this tutorial, our team originates from geographically distant countries and has varying level of seniority, including two PhD students and a full professor, The team includes a female researcher. This tutorial will augment a workshop on "Language + Molecules" to be held at a the ACL conference, which already has confirmed speakers and organizers with diversity in geography, ethnicity, and gender. This tutorial will strongly promote academic diversity, since it requires combining the specialties of chemists, physicians, pharmacists, computational linguists, and machine learning researchers. Further, this tutorial will promote the usage of NLP in high-impact

areas, ranging from drug discovery to organic photovoltaics. The methods we will introduce are language-agnostic. All tutorial materials (slides, example, reading list) will be shared to reach such a diverse audience.

Target Audience and Background We will target this tutorial at NLP researchers with no knowledge of chemistry or molecules– thus, we will provide an extensive discussion of background material. However, we will assume that the target audience is familiar with modern NLP methods including training deep neural network-based language models (e.g., BERT). We anticipate an audience size of 75-150 researchers. We will discuss relevant background for applying NLP to molecules and important applications in chemistry.

Reading List

- Molecule Representations and Language Models: (Weininger, 1988; Krenn et al., 2020; Cheng et al., 2023; Chithrananda et al., 2020; Ahmad et al., 2022; Tysinger et al., 2023)
- Molecule-Language Modeling: (Edwards et al., 2021; Zhao et al., 2023b; Zeng et al., 2022; Edwards et al., 2022b; Zhao et al., 2023a; Su et al., 2022; Liu et al., 2022, 2023c; Xu et al., 2023; Liu et al., 2023a; Luo et al., 2023)
- Applications: (Jordan and Roughley, 2009; Mukhtar et al., 2014; Kippelen et al., 2009)
- LLMs as Scientific Agents: (Boiko et al., 2023; Bran et al., 2023; Castro Nascimento and Pimentel, 2023; White et al., 2023)
- Survey: (Zhang et al., 2023) Section 10.3.3

We won't require reading these beforehand to ensure the tutorial is introductory.

Breadth of Tutorial Papers in the reading list were created by a diverse set of authors and include other disciplines. Specifically, only 2 papers and a survey from the instructors will be covered.

Ethical Considerations

Broader Impacts Our tutorial will have potential broader impacts: 1) It will help ACL researchers to better understand the research goals and constraints in chemical sciences, allowing them to do more impactful research there. 2) Studying language models in the context of non-human languages can help develop an understanding of their workings; due to our own personal linguistic biases, human researchers often misattribute abilities to language models. This is particularly relevant for developing new methodologies which are applicable to low-resource human languages. 3) It will promote further research in text-based molecule generation, with potential to enable a large shift in chemistry research so that custom molecules can be developed for each application or patient.

Ethical Concerns Like most methodologies reliant on LLMs, there may be biases learned by the model due to its large-scale training data. In this domain, these biases may affect what type of molecules are generated. Thus, any molecules or drugs discovered should be strictly evaluated by standard clinical processes before being considered for human or medicinal use. Another risk is that potentially dangerous molecules may be discovered. However, knowledge of dangerous molecule's existence and structure is generally not harmful due to the requisite technical knowledge and laboratory resources required for synthesis. Overall, we believe these downsides are outweighed by the benefits to the research and pharmaceutical communities.

3.1 Tutorial Presenters

Carl Edwards is a Ph.D. student in the Computer Science Department at UIUC. Broadly, he is interested in information extraction, information retrieval, text mining, representation learning, AI4Science, and multimodality. Particularly, he is interested in applying these to the scientific domain to accelerate scientific discovery. His work focuses on integrating natural language and molecules, especially using multimodal representations.

Qingyun Wang is a Ph.D. student in computer science at UIUC. His research lies in NLP for scientific discovery. Recently, he works on extracting reaction information from scientific literature. He served as a PC member in conferences including ICML, ACL, ICLR, NeurIPS, etc. His work was recognized in the first Alexa Prize competition and by the NAACL-HLT 2021 Best Demo Award. He has presented a tutorial at EMNLP 2021.

Heng Ji is a professor at the Computer Science Department of UIUC, and Amazon Scholar. She is a leading expert on multimodal multilingual information extraction, including NLP for Science with a particular interest in leveraging NLP for drug discovery. She has coordinated the NIST TAC Knowledge Base Population task since 2010. She has served as the PC Co-Chair of many conferences including NAACL-HLT2018 and AACL-IJCNLP2022 and has presented many tutorials. She was elected as NAACL secretary 2020-2023.

Acknowledgements

This work is supported by the Molecule Maker Lab Institute: an AI research institute program supported by NSF under award No. 2019897, and by the DOE Center for Advanced Bioenergy and Bioproducts Innovation, U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under Award Number DESC0018420. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied of, the National Science Foundation, the U.S. Department of Energy, and the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2022. Chemberta-2: Towards chemical foundation models. *arXiv preprint arXiv:2209.01712*.

Mugahed Abdullah Hasan Albahde, Bulat Abdrakhimov, Guo-Qi Li, Xiaohu Zhou, Dongkai Zhou, Hao Xu, Huixiao Qian, and Weilin Wang. 2021. The role of microtubules in pancreatic cancer: Therapeutic progress. *Frontiers in Oncology*, 11:640863.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Daniil A Boiko, Robert MacKnight, and Gabe Gomes. 2023. Emergent autonomous scientific research capabilities of large language models. *ArXiv preprint*, abs/2304.05332.

Andres M Bran, Sam Cox, Andrew D White, and Philippe Schwaller. 2023. ChemCrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*.

Cayque Monteiro Castro Nascimento and André Silva Pimentel. 2023. Do large language models understand chemistry? a conversation with chatgpt. *Journal of Chemical Information and Modeling*, 63(6):1649–1655.

Austin H Cheng, Andy Cai, Santiago Miret, Gustavo Malkomes, Mariano Phielipp, and Alán Aspuru-Guzik. 2023. Group selfies: a robust fragment-based molecular string representation. *Digital Discovery*.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2020. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*.

Dimitrios Christofidellis, Giorgio Giannone, Jannis Born, Ole Winther, Teodoro Laino, and Matteo Manica. 2023. Unifying molecular and textual representations via multi-task language modelling. *arXiv preprint arXiv:2301.12586*.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022a. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 375–413, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022b. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 375–413, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Carl Edwards, ChengXiang Zhai, and Heng Ji. 2021. Text2Mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 595–607, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Benedek Fabian, Thomas Edlich, Héléna Gaspar, Marwin Segler, Joshua Meyers, Marco Fiscato, and Mohamed Ahmed. 2020. Molecular representation learning with language models and domain-relevant auxiliary tasks. *arXiv preprint arXiv:2011.13230*.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. 2023. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. *arXiv preprint arXiv:2306.08018*.

Fleur M Ferguson and Nathanael S Gray. 2018. Kinase inhibitors: the road ahead. *Nature reviews Drug discovery*, 17(5):353–377.

Daniel Flam-Shepherd and Alán Aspuru-Guzik. 2023. Language models can generate molecules, materials, and protein binding sites directly in three dimensions as xyz, cif, and pdb files. *arXiv preprint arXiv:2305.05708*.

Nathan Frey, Ryan Soklaski, Simon Axelrod, Siddharth Samsi, Rafael Gómez-Bombarelli, Connor Coley, and Vijay Gadepally. 2022. Neural scaling of deep chemical models.

Miguel García-Ortegón, Gregor N. C. Simm, Austin J. Tripp, José Miguel Hernández-Lobato, Andreas Bender, and Sergio Bacallado. 2022. Dockstring: Easy molecular docking yields better benchmarks for ligand design. *Journal of Chemical Information and Modeling*, 62(15):3486–3502.

Thomas Gaudelet, Ben Day, Arian R Jamasb, Jyothish Soman, Cristian Regep, Gertrude Liu, Jeremy BR Hayter, Richard Vickers, Charles Roberts, Jian Tang, et al. 2021. Utilizing graph machine learning within drug discovery and development. *Briefings in bioinformatics*, 22(6):bbab159.

Francesca Grisoni. 2023. Chemical language models for de novo drug design: Challenges and opportunities. *Current Opinion in Structural Biology*, 79:102527.

Chi Han, Jialiang Xu, Manling Li, Yi R. Fung, Chenkai Sun, Tarek Abdelzaher, and Heng Ji. 2023. Lm-switch: Lightweight language model conditioning in word embedding space. *ArXiv preprint*, abs/2305.12798.

Tom Hope, Doug Downey, Oren Etzioni, and Weld. 2022. A computational inflection for scientific discovery. *ArXiv preprint*, abs/2205.02007.

Allan M Jordan and Stephen D Roughley. 2009. Drug discovery chemistry: a primer for the non-specialist. *Drug discovery today*, 14(15-16):731–744.

Maria Kavallaris. 2010. Microtubules and resistance to tubulin-binding agents. *Nature Reviews Cancer*, 10(3):194–204.

Bernard Kippelen et al. 2009. Organic photovoltaics. *Energy & Environmental Science*, 2(3):251–261.

Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. 2020. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024.

Youwei Liang, Ruiyi Zhang, Li Zhang, and Pengtao Xie. 2023. Drugchat: Towards enabling chatgpt-like capabilities on drug molecule graphs.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Anima Anandkumar. 2022. Multi-modal molecule structure-text model for text-based retrieval and editing. *arXiv preprint arXiv:2212.10789*.

Shengchao Liu, Jiongxiao Wang, Yijin Yang, Chengpeng Wang, Ling Liu, Hongyu Guo, and Chaowei Xiao. 2023a. Chatgpt-powered conversational drug editing using retrieval and domain feedback. *arXiv preprint arXiv:2305.18090*.

Shengchao Liu, Yutao Zhu, Jiarui Lu, Zhao Xu, Weili Nie, Anthony Gitter, Chaowei Xiao, Jian Tang, Hongyu Guo, and Anima Anandkumar. 2023b. A text-guided protein design framework. *ArXiv preprint*, abs/2302.04611.

Zequn Liu, Wei Zhang, Yingce Xia, Lijun Wu, Shufang Xie, Tao Qin, Ming Zhang, and Tie-Yan Liu. 2023c. Molxpt: Wrapping molecules with text for generative pre-training. *arXiv preprint arXiv:2305.10688*.

Yizhen Luo, Kai Yang, Massimo Hong, Xingyi Liu, and Zaiqing Nie. 2023. Molfm: A multimodal molecular foundation model. *arXiv preprint arXiv:2307.09484*.

Eiman Mukhtar, Vaqar Mustafa Adhami, and Hasan Mukhtar. 2014. Targeting microtubules by natural agents for cancer therapy. *Molecular cancer therapeutics*, 13(2):275–284.

NVIDIA Corporation. 2022. Megamolbart v0.2.

Noel O'Boyle and Andrew Dalke. 2018. Deepsmiles: an adaptation of smiles for use in machine-learning of chemical structures.

Barbara Partee et al. 1984. Compositionality. *Varieties of formal semantics*, 3:281–311.

Philippe Schwaller, Daniel Probst, Alain C Vaucher, Vishnu H Nair, David Kreutter, Teodoro Laino, and Jean-Louis Reymond. 2021. Mapping the space of chemical reactions using attention-based neural networks. *Nature Machine Intelligence*, 3(2):144–152.

Philipp Seidl, Andreu Vall, Sepp Hochreiter, and Günter Klambauer. 2023. Enhancing activity prediction models in drug discovery with the ability to understand human language. *ArXiv preprint*, abs/2303.03363.

Bing Su, Dazhao Du, Zhao Yang, Yujie Zhou, Jiangmeng Li, Anyi Rao, Hao Sun, Zhiwu Lu, and Ji-Rong Wen. 2022. A molecular multimodal foundation model associating molecule graphs with natural language. *ArXiv preprint*, abs/2209.05481.

Zoltán Gendler Szabó. 2020. Compositionality.

Oleg Trott and Arthur J Olson. 2010. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461.

Emma P Tysinger, Brajesh K Rai, and Anton V Sinitskiy. 2023. Can we quickly learn to "translate" bioactive molecules with transformer models? *Journal of Chemical Information and Modeling*, 63(6):1734–1744.

Alain C Vaucher, Philippe Schwaller, Joppe Geluykens, Vishnu H Nair, Anna Iuliano, and Teodoro Laino. 2021. Inferring experimental procedures from text-based representations of chemical reactions. *Nature communications*, 12(1):2573.

Alain C Vaucher, Federico Zipoli, Joppe Geluykens, Vishnu H Nair, Philippe Schwaller, and Teodoro Laino. 2020. Automated extraction of chemical synthesis actions from experimental procedures. *Nature communications*, 11(1):3601.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. 2023. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60.

Onsurang Wattanathamsan and Varisa Pongrakhananon. 2022. Emerging role of microtubule-associated proteins on cancer metastasis. *Frontiers in Pharmacology*, 13:935493.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

David Weininger. 1988. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36.

Andrew D White, Glen M Hocky, Heta A Gandhi, Mehrad Ansari, Sam Cox, Geemi P Wellawatte, Subarna Sasmal, Ziyue Yang, Kangxin Liu, Yuvraj Singh, et al. 2023. Assessment of chemistry knowledge in large language models that generate code. *Digital Discovery*, 2(2):368–376.

Hanwen Xu, Addie Woicik, Hoifung Poon, Russ B Altman, and Sheng Wang. 2023. Multilingual translation for zero-shot biomedical classification using biotranslator. *Nature Communications*, 14(1):738.

Zheni Zeng, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2022. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nature communications*, 13(1):862.

Zheni Zeng, Bangchen Yin, Shipeng Wang, Jiarui Liu, Cheng Yang, Haishen Yao, Xingzhi Sun, Maosong Sun, Guotong Xie, and Zhiyuan Liu. 2023. Interactive molecular discovery with natural language. *arXiv preprint arXiv:2306.11976*.

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

Haiteng Zhao, Shengchao Liu, Chang Ma, Hannan Xu, Jie Fu, Zhi-Hong Deng, Lingpeng Kong, and Qi Liu. 2023a. Gimlet: A unified graph-text model for instruction-based molecule zero-shot learning. *bioRxiv*, pages 2023–05.

Wenyu Zhao, Dong Zhou, Buqing Cao, Kai Zhang, and Jinjun Chen. 2023b. Adversarial modality alignment network for cross-modal molecule retrieval. *IEEE Transactions on Artificial Intelligence*.

C Lawrence Zitnick, Lowik Chanussot, Abhishek Das, Siddharth Goyal, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Thibaut Lavril, Aini Palizhati, Morgane Riviere, et al. 2020. An introduction to electrocatalyst design using machine learning for renewable energy storage. *arXiv preprint arXiv:2010.09435*.