Application of Large Language Models in Chemistry Reaction Data Extraction and Cleaning

Xiaobao Huang*
xhuang2@nd.edu
University of Notre Dame
Notre Dame, IN, USA

Mihir Surve*
msurve@nd.edu
University of Notre Dame
Notre Dame, IN, USA

Yuhan Liu*
yliu57@nd.edu
University of Notre Dame
Notre Dame, IN, USA

Tengfei Luo
tluo@nd.edu
University of Notre Dame
Notre Dame, IN, USA

Olaf Wiest
owiest@nd.edu
University of Notre Dame
Notre Dame, IN, USA

Xiangliang Zhang
xzhang33@nd.edu
University of Notre Dame
Notre Dame, IN, USA

Nitesh V. Chawla
nchawla@nd.edu
University of Notre Dame
Notre Dame, IN, USA

Abstract

Chemical reaction data has existed and still largely exists in unstructured forms. But curating such information into datasets suitable for tasks such as yield and reaction outcome prediction is impractical via manual curation and not possible to automate through programmatic means alone. Large language models (LLMs) have emerged as potent tools, showcasing remarkable capabilities in processing textual information and therefore could be extremely useful in automating this process. To address the challenge of unstructured data, we manually curated a dataset of structured chemical reaction data to fine-tune and evaluate LLMs. We propose a paradigm that leverages prompt-tuning, fine-tuning techniques, and a verifier to check the extracted information. We evaluate the capabilities of various LLMs, including LLAMA-2 and GPT models with different parameter counts, on the data extraction task. Our results show that prompt tuning of GPT-4 yields the best accuracy and evaluation results. Fine-tuning LLAMA-2 models with hundreds of samples does enable them and organize scientific material according to user-defined schemas better though. This workflow shows an adaptable approach for chemical reaction data extraction but also highlights the challenges associated with nuance in chemical information. We open-sourced our code at GitHub.

CCS Concepts

• **Information systems** → **Language models**; • **Computing methodologies** → **Information extraction**; • **Applied computing** → **Chemistry**.

*All authors contributed equally to this research.

CIKM '24, October 21–25, 2024, Boise, ID, USA

ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3679874

Keywords

LLM, Chemistry Reaction Data Extraction, Data Mining, Data-Driven Chemistry

ACM Reference Format:
Xiaobao Huang, Mihir Surve, Yuhan Liu, Tengfei Luo, Olaf Wiest, Xiangliang Zhang, and Nitesh V. Chawla. 2024. Application of Large Language Models in Chemistry Reaction Data Extraction and Cleaning. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3627673.3679874

1 Introduction

The combination of data-driven methods and chemistry [2–4, 7, 8, 12, 19] has achieved outstanding progress in recent years. The explosion in the use of machine learning (ML) techniques has increased the demand for large, well-organized datasets[15]. However, in the domain of chemistry, the challenge for a lot of predictive model design has been around the lack of structured data itself. "Data points" in chemistry are often scattered in literature within tables, figures, and free text and generally have a lot of contextual information to be inferred. Therefore, traditional data extraction methods lead to several complications and mistakes. This is also reflected in the United States Patent and Trademark Office (USPTO) dataset [14], extracted from granted and applied patents from 1976-2016. The dataset has about three million reactions, represented in the Simplified Molecular Input Line Entry Specification (SMILES) format, with mined freetext and some abstractions from it. Such data presents a lot of opportunities for thorough analysis as well as predictive modeling if structured properly. For example, in several instances, the freetext contains characterization information (obtained through methods like Nuclear Magnetic Resonance [NMR], mass spectrometry) and experimental method information (synthesis conditions, chemical process results), which could be crucial for representing reactions and understanding outcomes related to them.

Large Language Models (LLMs), such as the generative pretrained transformer (GPT)-4, have recently sparked a lot of scientific and

general interest, indicating that they have significant potential to solve this challenge. While traditional manual extraction and curation of unstructured text into structured datasets could take months, LLMs are a strong alternative that might effectively speed this process up.

In this study, we will probe the applicability of LLMs in extracting useful information from such unstructured/unorganized data sources in chemistry, with the goal of obtaining organized information hierarchies for downstream modeling. The contribution of the paper can be summarized as follows:

- We propose a workflow that combines prompt-tuning and fine-tuning techniques for LLMs in chemical reaction information extraction and a fact-based verifier to evaluate the information within the extraction.
- We manually curate a dataset from over 800 classes representing the majority of examples in the USPTO database and extracted accurate, structured chemical reaction information in JavaScript Object Notation (JSON) format.
- We design prompts for extracting structured information out of this dataset, and compare the same with using part of the dataset for finetuning.
- We evaluate the performance of these LLMs in reaction information extraction tasks with natural language processing (NLP) metrics and fact-based verifier metrics.

2 Related work

In recent years, ML models explicitly designed for direct property prediction have been increasingly integrated into the early stages of drug discovery and design workflows [1, 16, 21]. The efficacy of these models, however, depends on the availability of large training datasets from organized databases. For example, in experimental organic synthesis, the traditional methods are typically non-uniform and highly context-dependent, which will lead to a loss of information when translated into a structured, tabular format [10].

LLMs, which employ semantic links across varied lengths of natural language sequences [11], show great potential for overcoming these constraints. The emergent capability of artificial "general intelligence" was initially demonstrated with OpenAI's GPT-3.5 and GPT-4 [6], and this led to a large number of proprietary as well as open source efforts to train powerful LLMs. Parallelly, there has also been a lot of work on trying to most effectively use LLM inferences through prompt tuning. More recently, prompts and architectures based on the utilization of LLMs as agents with tools have led to the creation of further possibilities for LLM usage without explicit retraining or fine-tuning. But as LLM fine-tuning methods become more optimal and available, adapting them to different domains and workflows is even more facilitated, giving rise to a multitude of LLM-based strategies for various tasks.

To extract information from scientific journals, Zheng et al. [23] developed a prompt-engineering method known as ChemPrompt w/ ChatGPT. They primarily focus on structuring text into tables, generating semi-structured summaries, and compiling information from the pretraining corpus. Huang and Cole [9] fine-tuned a BERT model by training it on battery publications to improve a database of NLP-extracted battery data. Despite breakthroughs in data transformation, it is still challenging to consistently turn unstructured data into structured representations. Dunn et al. [5] and Walker et al. [22] have proven the use of GPT-based models, which are iteratively fine-tuned to effectively organize data from scholarly papers into JSON format. However, such workflows have not been implemented for organic reactions, which involve several components like reactant(s), product(s), conditions and procedures. With the USPTO database as a proof of concept, we aim to evaluate the efficacy of LLMs for such a task..

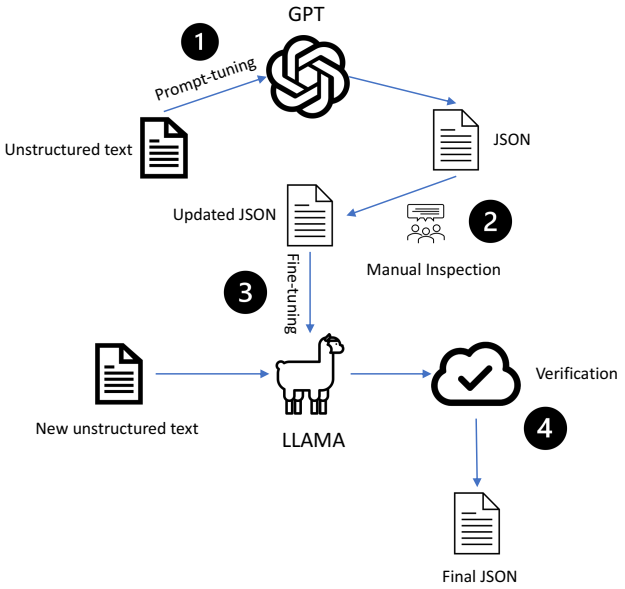

Figure 1: Our workflow for chemical reaction data extraction and cleaning. They are separated into four stages: prompt-tuning, manual inspection, fine-tuning, and verification

3 Methodology

3.1 Workflow of the Extraction

As shown in Figure 1, we separate our workflow into four distinct stages: prompt-tuning, manual inspection, fine-tuning, and verification. In the initial prompt-tuning stage, we leverage expert-curated prompts tailored for GPT models to distill essential information from unstructured text. This step significantly streamlines subsequent manual inspection, as many tasks, such as extracting chemical formulas and reactant names, can be efficiently handled by state-of-the-art LLMs. The subsequent stage involves the JSON output generated by the prompt-tuned LLMs being manally scrutinized, particularly focusing on detecting instances of hallucination or misinterpretation, especially within experimental procedures. Next, the human-inspected JSON data is employed to fine-tune open-sourced LLMs. Following fine-tuning, the refined model can continuously process new unstructured data and proceed to the information verification stage.

3.2 Dataset Curation

Firstly, we used the USPTO database for the curation of the reaction extraction dataset. We used the RXNFP [20] that classified the reactions in USPTO into 834 classes. As a proof of concept, we sample one example from each of 834 classes for maximal coverage. The procedure part of each example, after combination with our

deliberately designed prompt was prompted to GPT-4 (❶, Figure 1). We specify the reaction labels to be JSON files. The JSON format includes sections for reactants, spectators (solvents), products, yield, procedures, total reaction time, and product information. The reactants, spectators, and products sections contain information about the chemicals involved in the reaction, including their names, amounts, moles, and SMILES notation. The yield section indicates the reaction yield. The procedures section includes details about the procedure type, chemicals involved, description, temperature, and time. The total_time field indicates the total time for the reaction. The product information section specifies qualitative or quantitative data about the products, such as color, physical state, or analytical details from spectroscopic methods like NMR or LCMS. "N/A" was used to handle any missing information Following this, we manually inspected these 834 data points, including the reactant, product, their amount, reaction procedures and reaction labels (❷, Figure 1).

3.3 Prompt Design

Prompt design is requirement dependent and subjective. Based on our setting, we performed three different levels of prompt design, where almost no prompt, a slightly designed prompt, or a fully designed prompt is given to the LLM. The 'almost no prompt' refers to the prompt design that only instructs the LLM to format the unstructured data into the JSON format without specifying what exactly the JSON will look like. The slightly designed prompt is generated with the help of GPT, with a poorly structured and unexplained JSON format provided. Lastly, an expert-curated prompt is generated with the inspection of the result from a slightly designed prompt and chemist expertise. The design is based on the insight from both iterative LLM prompting experiments and chemists experienced with chemical names and lab procedures. The prompt does include the full JSON schema mentioned in the previous section with an explanation, but also provides guidelines for handling missing information and information mismatches between the structured and unstructured parts. The model is instructed to interpolate between different parts of the given data and use reasoning about chemicals without adding any extra information to the formatted JSON. If information is missing, "N/A" should be used as a placeholder. Finally, the model is instructed to output only the formatted JSON and nothing else.

3.4 Fine-tuning

For finetuning (❸, Figure 1), we split the curated dataset with a ratio of 8:2 for training and testing purposes and use Low-rank adaptation (LoRA) for efficient parameter fine-tuning. We chose 7b, 13b Llama-2, and GPT-3.5-turbo models for fine-tuning. We finetuned the 7b and 13b Llama-2 models on 2 A100 GPUs for 200 epochs within 12 hours, and the GPT-3.5-turbo by calling the OpenAI API. The JSON outputs are treated as labels for the finetuning process.

3.5 Verification

In addition to the extraction workflow, we add a fact-based verifier to double check the extracted information. Such an approach serves a dual purpose: a chemistry-oriented sanity check to the workflow and a quantitative estimation of the data quality itself. The verification stage (❹, Figure 1) consisted of three checks: SMILES validity, number of moles and yield verification. based on RDKit (an open-source cheminformatics toolkit)-based tools, the inputs of which were planned by a GPT model when prompted with the extracted data. The workflow is structured as follows. The output extracted from the LLM is structured into inputs of SMILES, weights, and number of moles. The toolkit first checks if the SMILES for each reactant and product are valid. If they are, the SMILES strings are used to calculate molecular weights through RDKit, which allows for a quantitative check over the number of moles. If these calculated values deviate more than 20 percent from the reported values, the molecule is flagged. The mole values are also used to compute a yield estimation metric to flag out unrealistically high reaction yields which have either been misreported or wrongly extracted. This simply done by comparing the product moles against the limiting reagent. The metrics used for verification are further highlighted in Section 4.1

4 Experiment

Our workflow integrates both LLMs and the verifier. Therefore, we measure the capability of the LLMs under different settings and LLMs with the verifier to demonstrate the effectiveness of our workflow in reaction information extraction.

4.1 Evaluation Metrics

Traditional evaluation metrics in the NLP area are employed in our study. However, they are not sufficient to evaluate the quality of the extraction in our case. Therefore, we employ the fact-based verifier metrics to better evaluate the quality of the extractions.

1. SacreBLEU: We use SacreBLEU [18] as a reference-based evaluation metric for machine translation. SacreBLEU computes a score based on the n-gram overlap between the machine-generated translations and one or more reference translations. The higher the SacreBLEU score, the better the translation quality, indicating a higher similarity between the machine-generated text and the reference reaction information.

2. BLEU-1: We utilize the BLEU-1 [17], which considers only unigrams (individual words) for evaluation. BLEU-1 is particularly useful because it provides a simple and automated way to measure the quality of special words such as compounds and operations in the reaction data. Even though it cannot capture higher-level linguistic phenomena, it is still valuable in providing a quick and informative evaluation of reaction information.

3. ROUGE: We employ ROUGE [13] as an evaluation metric that is commonly used in NLP. ROUGE evaluates the overlapping between the generated text and the reference text. In essence, ROUGE scores help assess how well a machine-generated summary or translation captures the important phrases or concepts present in the reference text, which is reaction information in our case.

4. Fact-based verifier Metrics: We choose the proportion of valid SMILES, correctly verified mole values, and correctness of average yield estimation as three evaluation metrics in the verifier. These metrics evaluate the factual correspondence of the information within JSON files. Metric A refers to the proportion of SMILES strings in the extraction that are valid and can be parsed through RDKit. Metric B is the proportion of correctly reported molar values. Values of calculated number of moles deviating more than 20% from their reported counterparts, as well as values with unreported components (SMILES, weight or number of moles), are flagged

Tuning Type	Model&Prompt type	Metrics				
		SacreBLEU	BLEU-1	ROUGE-1	ROUGE-2	ROUGE-L
Prompt-tune	LLAMA2-7b No-prompt	14.86	0.07	0.10	0.07	0.09
	LLAMA2-13b No-prompt	13.88	0.03	0.50	0.37	0.40
	LLAMA2-70b No-prompt	10.38	0.03	0.35	0.23	0.26
	LLAMA2-13b Moderate	24.50	0.03	0.53	0.33	0.43
	LLAMA2-70b Moderate	17.20	0.03	0.48	0.29	0.41
	LLAMA2-13b Expert	47.41	0.03	0.63	0.43	0.55
	LLAMA2-70b Expert	50.00	0.03	0.66	0.48	0.59
	GPT3.5-turbo Expert	75.22	0.16	0.83	0.73	0.74
	GPT4.0 Expert	**98.62**	**0.45**	**0.98**	**0.98**	**0.98**
Fine-tune	GPT3.5-turbo	<u>84.14</u>	<u>0.20</u>	<u>0.89</u>	<u>0.83</u>	<u>0.85</u>
	LLAMA2-7b	41.36	0.05	0.62	0.58	0.58
	LLAMA2-13b	77.50	0.04	0.83	0.75	0.77

Table 1: Comparison of models with different parameters and tuning methods under BLEU and ROUGE metrics. Bold and underline are the best and the second best result, respectively, across each metric

Tuning Type	Model&Prompt type	Metrics		
		A	B	C
Prompt-tune	GPT4.0 Expert	**0.98**	0.59	0.56
Fine-tune	GPT3.5-turbo	0.98	0.26	0.22
	LLAMA2-7b	0.96	**0.62**	**0.60**
	LLAMA2-13b	0.97	0.56	0.52

Table 2: Verification scores of different model types. Metric A: Proportion of valid SMILES; Metric B: Proportion of correctly verified mole values; Metric C: Average yield estimation correctness

with a score of 0, and this is averaged across all molecules. The mole values can also be used to compute a yield estimation metric to flag out unrealistically high reaction yields which have either been misreported or wrongly extracted. This is done by comparing the product moles to the limiting reagent, and is simply expressed as a 0/1 output per reaction. This averaged across the test set is mentioned as Metric C.

4.2 Evaluation of LLMs

The models listed include different versions of GPT (3.5 and 4.0) as well as LLAMA2 with varying parameters. Performance varies across models and metrics, indicating that different configurations and tuning methods have significant impacts on model performance. Based on our experiments in Table 1, prompt-tuning GPT-4 is the state-of-the-art model. In the fine-tuning models, GPT-3.5 shows promising performance across all metrics among all of the fine-tuning models. LLAMA2 models have relatively lower scores compared to GPT models, but they also show improvements after fine-tuning and prompt-tuning. Fine-tuning with only hundreds of datapoints could improve the performance of the model by at least 200% and yield well-structured text. More parameters do not guarantee better performance. Generally, fine-tuning and prompt-tuning improve model performance across all models and metrics. While GPT3.5-turbo and GPT-4 perform well overall, it's essential to consider the cost for the vast amount of literature text in the chemistry domain. Therefore, we leverage GPT models with prompt-tuning to generate a small amount of initial JSON for manual inspection, then proceed to fine-tune locally deployed Llama-2 models for future extensive corpus of chemistry literature extraction. Certainly, local computational resources and task requirements also require careful consideration.

4.3 Evaluation of LLMs with Verifier

Values computed in the verification stage are reported in Table 2, averaged across all test samples. While the SMILES scores (Metric A) are clear indicators of data quality, low scores with correctly verified mole values (Metric B) could be both due to a higher proportion of incomplete information or hallucinated/missing information in the extraction. The yield estimations (Metric C) are conditional on the availability of all the information in the first two metrics, justifying their lower scores. GPT-4 seems to provide the best performance, but the finetuned LLAMA variants also largely pick the correct information from the text without hallucinating. Using such a verifier in tandem with an extraction workflow could be vital for creating datasets reliably while ensuring quality and correctness. Using a fact-based metric could have the disadvantage of not clearly delineating between the quality of the data and the quality of the extraction though. As mentioned, a low score on these metrics could be due to data quality issues or incorrect extractions. Therefore, more detailed metrics need to be ideated to gauge model performance better with respect to this task.

5 Conclusion

Our study explores the use of LLMs for extracting chemical reaction data from unstructured texts using the USPTO reaction dataset. We prompt and fine-tune models (GPT-3.5-turbo, GPT-4.0, and Llama 2) to refine data, assessing performance with metrics like SacreBLEU, BLEU, and ROUGE. Extracted text items are verified for accuracy. Our results indicate that LLMs, when fine-tuned with sufficient samples, can effectively organize scientific content into user-defined schemas. The workflow's simplicity and accessibility enable the conversion of unstructured scientific texts into structured databases, balancing commercial LLM performance with accuracy. However, data extraction in specialized domains like chemical reactions remains complex, requiring further studies on evaluation methods and data quality.

Acknowledgments

This work was supported by National Science Foundation through the NSF Center for Computer-Assisted Synthesis (C-CAS), under grant number CHE-2202693.

References

[1] Kamal Choudhary, Brian DeCost, Chi Chen, Anubhav Jain, Francesca Tavazza, Ryan Cohn, Cheol Woo Park, Alok Choudhary, Ankit Agrawal, Simon JL Billinge, et al. 2022. Recent advances and applications of deep learning methods in materials science. *npj Computational Materials* (2022).

[2] Connor W Coley, William H Green, and Klavs F Jensen. 2018. Machine learning in computer-aided synthesis planning. *Accounts of chemical research* 51, 5 (2018), 1281–1289.

[3] Connor W Coley, Wengong Jin, Luke Rogers, Timothy F Jamison, Tommi S Jaakkola, William H Green, Regina Barzilay, and Klavs F Jensen. 2019. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chemical science* 10, 2 (2019), 370–377.

[4] Connor W Coley, Dale A Thomas III, Justin AM Lummiss, Jonathan N Jaworski, Christopher P Breen, Victor Schultz, Travis Hart, Joshua S Fishman, Luke Rogers, Hanyu Gao, et al. 2019. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* 365, 6453 (2019), eaax1566.

[5] John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S Rosen, Gerbrand Ceder, Kristin A Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature Communications* (2024).

[6] OpenAI et. al. 2024. GPT-4 Technical Report. (2024).

[7] Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang Zhang, et al. 2023. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems* 36 (2023), 59662–59688.

[8] Zhichun Guo, Chuxu Zhang, Wenhao Yu, John Herr, Olaf Wiest, Meng Jiang, and Nitesh V Chawla. 2021. Few-shot graph learning for molecular property prediction. In *Proceedings of the web conference 2021*. 2559–2567.

[9] Shu Huang and Jacqueline M. Cole. 2022. BatteryBERT: A Pretrained Language Model for Battery Database Enhancement. *Journal of Chemical Information and Modeling* (2022).

[10] Kevin Maik Jablonka, Qianxiang Ai, Alexander Al-Feghali, Shruti Badhwar, Joshua D Bocarsly, Andres M Bran, Stefan Bringuier, L Catherine Brinson, Kamal Choudhary, Defne Circi, et al. 2023. 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon. *Digital Discovery* (2023).

[11] Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences* (2023).

[12] Steven M Kearnes, Michael R Maser, Michael Wleklinski, Anton Kast, Abigail G Doyle, Spencer D Dreher, Joel M Hawkins, Klavs F Jensen, and Connor W Coley. 2021. The open reaction database. *Journal of the American Chemical Society* 143, 45 (2021), 18820–18826.

[13] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *ACL*.

[14] Daniel Mark Lowe. 2012. Extraction of chemical structures and reactions from the literature. (2012).

[15] Yihong Ma, Xiaobao Huang, Bozhao Nan, Nuno Moniz, Xiangliang Zhang, Olaf Wiest, and Nitesh V. Chawla. 2024. Are we Making Much Progress? Revisiting Chemical Reaction Yield Prediction from an Imbalanced Regression Perspective. In *WWW*.

[16] Osvaldo N Oliveira Jr and Maria Cristina F Oliveira. 2022. Materials discovery with machine learning and knowledge discovery. *Frontiers in chemistry* (2022).

[17] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

[18] Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *WMT*.

[19] Mandana Saebi, Bozhao Nan, John E Herr, Jessica Wahlers, Zhichun Guo, Andrzej M Zurański, Thierry Kogej, Per-Ola Norrby, Abigail G Doyle, Nitesh V Chawla, et al. 2023. On the use of real-world datasets for reaction yield prediction. *Chemical science* 14, 19 (2023), 4997–5005.

[20] 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* (2021).

[21] Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, et al. 2019. Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery* (2019).

[22] Nicholas Walker, John Dagdelen, Kevin Cruse, Sanghoon Lee, Samuel Gleason, Alexander Dunn, Gerbrand Ceder, A. Paul Alivisatos, Kristin A. Persson, and Anubhav Jain. 2023. Extracting Structured Seed-Mediated Gold Nanorod Growth Procedures from Literature with GPT-3. (2023).

[23] Zhiling Zheng, Oufan Zhang, Christian Borgs, Jennifer T. Chayes, and Omar M. Yaghi. 2023. ChatGPT Chemistry Assistant for Text Mining and the Prediction of MOF Synthesis. *Journal of the American Chemical Society* (2023).