Harnessing Instruction-Tuned Large Language Models to Mine Structured Omics Data for Predicting Chemical Toxicity

Yufan Liu,a Guoping Lian,a,b Tao Chena\*

aSchool of Chemistry and Chemical Engineering,University of Surrey, Stag Hill, Guildford GU2 7XH, UK

bUnilever R&D Colworth, Unilever, Sharnbrook MK44 1LQ, UK

\* t.chen@surrey.ac.uk

Abstract

Chemical safety and toxicology are important considerations in designing safer and sustainable products and processes. Omics technologies, including transcriptomics, proteomics, and metabolomics, provide crucial insights into chemical toxicity by identifying molecular-level changes post-chemical exposure and elucidating regulatory pathways. Despite the vast literature on this topic, there's a lack of comprehensive datasets detailing chemical perturbations and their outcomes. A tool that can efficiently and accurately extract structured data from scientific literature is needed. Large Language Models (LLMs) like GPT-4 offer the potential for efficient information retrieval from intricate texts. However, optimising their factuality and desired behaviour often requires labour-intensive human feedback. Addressing this, our work introduces a semi-automated pipeline for structured information extraction from voluminous literature. Initially, literature that contain any type of omics in the title or abstract and mention pathway analysis in the text were obtained from PubMed. Subsequently, GPT-4 was employed to extract data points including omics type, perturbation, perturbation type and study results, from selected literature abstracts in a zero-shot manner. After manual corrections, this data served to fine-tune the GPT-3.5-turbo model. This fine-tuned model then processed a new batch of abstracts, with its output validated by GPT-4. Discrepancies were manually reconciled, and the consolidated data was used to further fine-tune the GPT-3.5-turbo model. Following an iterative process of reconciliation and fine-tuning, the resulting model demonstrated high accuracy and alignment in extracting structured data from literature with minimal human intervention, which holds the potential to accelerate knowledge transformation. Additionally, we present a structured dataset encapsulating omics type, perturbations, perturbation types, results, etc., that can be used for future omics studies.

**Keywords**: Chemical safety, omics technologies, Large Language Models (LLMs), Structured information extraction, fine-tuning GPT.

* 1. Introduction

In the field of chemical safety and toxicology, the need to develop safer and more sustainable products and processes is increasingly recognised (Anastas, 2016). The emergence of omics technologies, including transcriptomics, proteomics, metabolomics, and other related disciplines, has significantly enhanced our understanding of the mechanisms underlying chemical toxicity. These technologies provide detailed insights into the molecular-level changes that transpire following chemical exposure, thus elucidating on the complex regulatory pathways within cells (Liu et al., 2023).

The regulatory pathways of a cell are a complex network involving a myriad of biomolecules, each possessing distinct physicochemical properties and engaging in complex, non-linear interactions. Single-omics techniques can measure biomolecules of a specific type, providing a fragmented view of these pathways. Multi-omics approaches enable a more holistic understanding of pathway’s response to chemical exposure (Canzler *et al.*, 2020).

Transcriptomics, which focuses on the comprehensive detection of RNAs in the cell, primarily identifies pathway responses through the differential expression of a known set of target genes. Proteomics offers insights into the proteins is particularly valuable for understanding toxicopathic effects and responses, leading to the development of protein-centric adverse outcome pathways (AOPs) and predictive models of toxicological pathways (Madeira and Costa, 2021). Phosphoproteomics, which aims to map and quantify protein phosphorylation, plays a vital role in deciphering the intracellular signalling networks that respond to cellular stresses (Titz *et al.*, 2014). Metabolomics, focusing on a wide array of chemically heterogeneous molecules, has emerged as a pivotal tool in toxicological studies, examining metabolic alterations in cells under chemical exposures (Olesti *et al.*, 2021). On this basis, the fields of toxicogenomics, toxicoproteomics, toxicometabolomics and other fields have been developed to use omics technology to predict toxicity.

Despite these advancements, a significant gap remains in the availability of datasets that encompass the information about chemical perturbation and corresponding real omics data and detailed information about regulatory pathways. Existing databases, such as DrugMatrix, Open TG-GATEs and L1000 (Liu *et al.*, 2023), provide limited information, predominantly in gene expression. While the literature on the use of omics data to explore chemical toxicity mechanisms is growing, the effective extraction of information from these complex datasets is a formidable challenge. The development of large language models (LLMs) like GPT has facilitated the efficient extraction of information from complex, unstructured text sources (Zhu *et al.*, 2023). However, a critical challenge remains in evaluating the accuracy of the outputs generated by LLMs and in ensuring their human alignment. This alignment is crucial to guarantee that the model's outputs are not only precise but also consonant with human values and intentions. Conventionally, such evaluations necessitate intricate training procedures or labour-intensive manual annotation of samples, which are time-consuming (Fu *et al.*, 2023). Therefore, a more expedient and efficient approach is imperative.

In this context, we propose a semi-automated method that enables effective information extraction with minimal human intervention. Our approach has led to the extraction of 10 critical data points, including omics layer, species, biological sample, differentially expressed genes (DEGs), proteins (DEPs), metabolites (DEMs), and regulatory pathways for a range of chemicals. This dataset is poised to be a valuable resource for downstream analysis, offering the potential to unravel the intricate relationships between these data points and to further our understanding of chemical toxicity mechanisms.

* 1. Methods and results
     1. Corpus retrieval

Our methodology commenced with the stringent selection of articles from a comprehensive corpus. Criteria for inclusion mandated the presence of specific omics-related terms (proteomics, metabolomics, transcriptomics, phosphoproteomics, multi-omics) in either the title or abstract. Additionally, articles were required to contain the phrase "pathway analysis" within their main text. Exclusion criteria were set to omit articles with any mention of "plant" in the title or abstract, reflecting our study's focus on non-plant organisms. From the PMC database (as of 25th October 2023), we retrieved 13,102 articles in XML format. The retrieved articles were pre-processed to keep the articles with ‘article-type’ of ‘research-article’.

A diagram of a scientific research process

Description automatically generated

Figure 1 The overview of information extraction

* + 1. GPT-4 extraction in zero-shot prompting

GPT-4 API was used for extracting structured information from 150 randomly chosen articles. This process was executed in a zero-shot manner with a structured prompt: ‘Based on the title and abstract of the scientific research article, please extract the following details:’.

The prompt was designed to extract key details including (1) Target article: This question is set to know whether this article contains the information we want before extracting information; (2) Type of omics: Different omics layers contain different layers of molecular information; (3) Specific primary perturbation: It is important to know the aim of the article, which is the primary perturbation that causing omics change; (4) Perturbation type: Figuring out what the type of the primary perturbation can help us to do downstream analysis; (5) Comparison group: With this data point, the settings of control group can be acquired; (6) Biological sample: Different biological samples with difference cells composition can convey different information. (7) Species: The same treatment can act differently one different animal; (8) Differentially Expressed Genes (DEGs); (9) Differentially Expressed Proteins (DEPs); (10) Differentially Expressed Metabolites (DEMs). (11) Pathway Enrichment Results: Results in regulatory can help us to understand the mechanism of cellular or tissue’s responses to stimulation.

The prompts that describe each task are shown in Table 1. To avoid hallucination, the description of each prompt is ended with “*If the information isn't available, respond with ‘Not specified’*”.

Table 1 Extracted data points and the description of prompt.

|  |  |
| --- | --- |
| Data points | Prompt |
| Target article | Is the article focused on omics analysis comparing two conditions (e.g., disease vs. control, treated vs. untreated, treatment A vs. treatment B)? |
| Type of omics | What omics data layers are examined in this research? Identify all layers, such as genomics, transcriptomics, proteomics, phosphoproteomics, and metabolomics. |
| Perturbation | Define the main factor causing the observed omics changes. |
| Perturbation type | Characterize the nature of the main cause. Options include 'disease', 'monomolecular substance', 'Complex mixtures/substance', 'physical processes', 'genetic modifications', 'development', 'resistance', and 'other'. |
| Comparison group | Identify the reference or control group against which the primary perturbation was studied. If not clearly stated or if the reference group lacks any specific condition, answer 'Control'. |
| Biological sample | Specify the biological samples used in the research, such as cells, plasma, or urine |
| Species | Specify clearly the species the study was conducted on, such as Homo sapiens, mouse, rat. |
| DEGs | List only the names of the specific differentially expressed genes, e.g., ATP2A1. |
| DEPs | List only the names of the specific differentially expressed proteins, e.g. Paralemmin-3/A6NDB9. |
| DEMs | List only the names of the specific differentially expressed metabolites. |
| Pathways | List the biological pathways (excluding individual genes or transcription factors) impacted by the primary perturbation. |

Extracted data underwent initial manual validation and formatting, with semicolons delineating multiple comparison groups. The validation classified each answer into three classes, a) Correct: the information is correct and the format is desired; b) Correct but not aligned: Correct information location but not in desired format; c) Wrong: wrong information extracted. The GPT-4 model achieved 83% to 100% in 11 tasks (Figure 2).

Figure 2 The accuracy of zero-shot prompting in extracting information from 150 articles.

* + 1. Fine-tuning GPT model

Two fine-tuned GPT-3.5-turbo models (gpt-3.5-turbo-1106) were developed: one for classification and another for information extraction.

* + - 1. Fine-tuned model for target article identification

The classification model was trained using 150 positive articles from the initial step and 50 negative articles unrelated to omics analysis. The positive examples were labeled with ‘True’, and negative as ‘False’. The prompt was structured as *“You are an expert in Bioinformatics. You are to determine if the study focused on omics analysis comparing two conditions (e.g., disease vs. control, treated vs. untreated, treatment A vs. treatment B) based on article title and abstract provided.”* Which is followed by the title and the abstract. The True/False labels were used as expected output to fine tune the model.

* + - 1. Fine-tuned model for information extraction

The preliminary model for information extraction was trained using pairs of titles, abstracts, and their respective extracted information, post manual correction. The prompt used for training contains two parts. The first part is system message, which tells the system the details about the task of extraction, including the return format, JSON, and the 10 data points with the descriptions. The second part is user message, which contains the title and the abstract of the article. The ten time points were used as output to fine-tune the model.

* + 1. Validation and correction

After the development of the fine-tuned models for classification and information extraction, these models were employed to process 200 articles. Zero-shot prompting was also employed for this determination. Comparisons were made between the results from the fine-tuned models and GPT-4, with any discrepancies being manually reviewed.

Of these, 169 articles were identified as target papers and used for further information extraction. The results from the fine-tuned models underwent validation using zero-shot prompting GPT-4. The prompt was structed to have two parts, the first part, system message, starts with “Your task is to determine if the student's answers is correct or not.

To solve the problem do the following:

- First, work out your own answers to the questions.

- Then compare your solution to the student's solution and evaluate if the student's solution is correct or not.” Also indicated the desired answer is “Correct/Wrong”. The second part user massage was structured to include the instruction of desired data points extraction, title, abstract, and “student solutions”.

The results that were regarded as a wrong answer were validated manually. To understand the performance of both the GPT-4 and fine-tuned GPT-4 model, the output from two model were assessed comprehensively.

Following manual corrections, the consolidated results were used to train two additional fine-tuned models. Step 2.3.1- 2.4 were repeated on both the second and third batch, comprising 200 articles each. This semi-automated fine-tuning process reduced the workload by 72% compared to manually assessing all data points.

* + 1. Large-scale information extraction

After three iterations of fine-tuning, the models achieved accuracies of 97% in classification and average accuracies of 97% in 10 tasks in information extraction. The classification model was then applied to the remaining articles in the pool, identifying those relevant for further information extraction. Ultimately, this process yielded data on 1063 monomolecular substances and 758 complex mixtures/substances, alongside corresponding differentially expressed genes, proteins, metabolites, and regulatory pathways.

b)

a)

Figure 3 a) The performance of GPT-4 in validating fine-tune model; b) The performance of fine-tuned models in ten tasks of information extraction in three iterations.

* 1. Conclusions

Our study has successfully demonstrated the integration of advanced computational methods with omics data analysis in chemical safety and toxicology. By developing a semi-automated pipeline using GPT-4 and fine-tuned models, we have streamlined the extraction of critical data from extensive scientific literature, achieving high accuracy in both classification and information extraction. Our findings pave the way for future research in chemical toxicity and safer product development, showcasing the significant role of artificial intelligence in advancing the field of toxicological studies.

Acknowledgement

Y. Liu received a PhD studentship co-funded by Unilever and the University of Surrey.

References

Anastas, N. D. (2016) ‘Connecting toxicology and chemistry to ensure safer chemical design’, *Green Chemistry*, 18(16), pp. 4325–4331. doi: 10.1039/c6gc00758a.

Canzler, S. *et al.* (2020) ‘Prospects and challenges of multi-omics data integration in toxicology’, *Archives of Toxicology*. Springer, pp. 371–388. doi: 10.1007/s00204-020-02656-y.

Fu, J. *et al.* (2023) ‘GPTScore: Evaluate as You Desire’.

Liu, A. *et al.* (2023) ‘Using chemical and biological data to predict drug toxicity’, *SLAS Discovery*, 28(3), pp. 53–64. doi: 10.1016/j.slasd.2022.12.003.

Liu, Y., Lian, G. and Chen, T. (2023) ‘A novel multi-omics data analysis of dose-dependent and temporal changes in regulatory pathways due to chemical perturbation: a case study on caffeine’, *Toxicology Mechanisms and Methods*, 0(0), pp. 1–12. doi: 10.1080/15376516.2023.2265462.

Madeira, C. and Costa, P. M. (2021) ‘Proteomics in systems toxicology’, *Advances in Protein Chemistry and Structural Biology*, 127, pp. 55–91. doi: 10.1016/bs.apcsb.2021.03.001

Olesti, E. *et al.* (2021) ‘Approaches in metabolomics for regulatory toxicology applications’, *Analyst*, 146(6), pp. 1820–1834. doi: 10.1039/d0an02212h.

Titz, B. *et al.* (2014) ‘Proteomics for systems toxicology’, *Computational and Structural Biotechnology Journal*, 11(18), pp. 73–90. doi: 10.1016/j.csbj.2014.08.004.

Zhu, Y. *et al.* (2023) ‘Large Language Models for Information Retrieval: A Survey’, pp. 1–26.