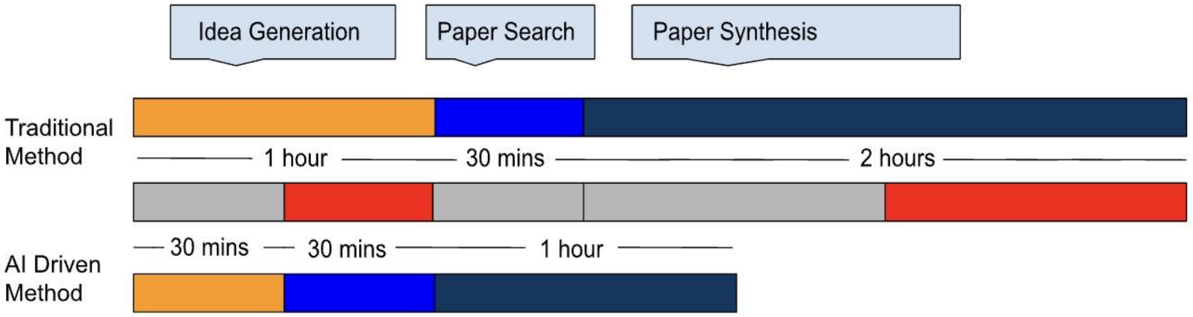
Sciencia
CONSULTING

Content Brief Development: Achieve a 25% Efficiency Boost Using AI Powered PDF Synthesis Tools

Abstract: Doing research for blog posts typically takes an excessive amount of time to read and synthesize recent scientific literature, and ensure information is up to date and accurate. Sciencia was tasked, with limited time and resources, to develop a blog post regarding multi-omic data and advancements made in high complexity data analytics with AI and ML tools. In order to achieve this goal, we analyzed the utility of the tool ChatDoc to extract relevant information from literature in the development of a content brief on streamlining multi-omic data interpretation using AI. A procedure was developed around the combination of ChatGPT and ChatDoc that resulted in a streamlined content brief synthesis saving several hours in research time.

Method: Several phases of content brief development were tested and optimized including idea generation, paper discovery, and paper synthesis with the objective of streamlining the time saving value of ChatDoc into an efficient and reproducable process.

Figure 1: The timeline provides a visualization of content brief development, comparing traditional methods against the AI driven process. The AI driven process saves 30 minutes of the hour previously spent on idea generation and an hour of the two hours previously spent on Paper Synthesis.



Idea Generation (Time 30 mins): ChatGPT was used to generate and structure ideas for the content brief. Think of a set of open-ended questions or prompts related to your chosen topic. These questions will serve as the starting point for your conversation with ChatGPT. Engage in a conversation with ChatGPT: Use the prompts you prepared to start a conversation with ChatGPT. Pose your questions one at a time, allowing the model to provide detailed responses.

You can have a back-and-forth conversation, asking follow-up questions to delve deeper into specific subtopics. The accuracy of the information here does not matter and simply will serve as a tool to guide the conversation in ChatDoc which should provide more up to date and accurate information. Finally ask ChatGPT to generate a blog post outline using the topics discussed.

Paper Search (Time: 30 mins) The research papers were sourced through the utilization of the Google Scholar search engine, employing relevant keywords aligned with the content brief. To ensure the attainment of high-quality and pertinent articles via Google Scholar quickly and efficiently, it is imperative to adhere to effective search practices. One should strive to refine search queries by employing specific key terms, and making use of quotations when necessary, enhancing the relevance of the results. Additionally, making use of the advanced search features offered by Google Scholar enables the narrowing down of results based on fields, publication years, and the exclusion of certain terms. Furthermore, exploring the related articles section and leveraging the "Cited by" feature facilitates the discovery of supplementary sources closely associated with the original query, as well as more recent research on the same subject matter. Lastly, it is vital to evaluate the credibility and relevance of the sourced material by assessing factors such as article quality, author expertise, journal reputation, peer review, citation count, and the alignment of content with the research topic.

Figure 2: A sample of 13 papers were found using google scholar and uploaded to ChatDoc. ChatDoc used 9 of the 13 papers in its response to guided questions to complete the content brief.

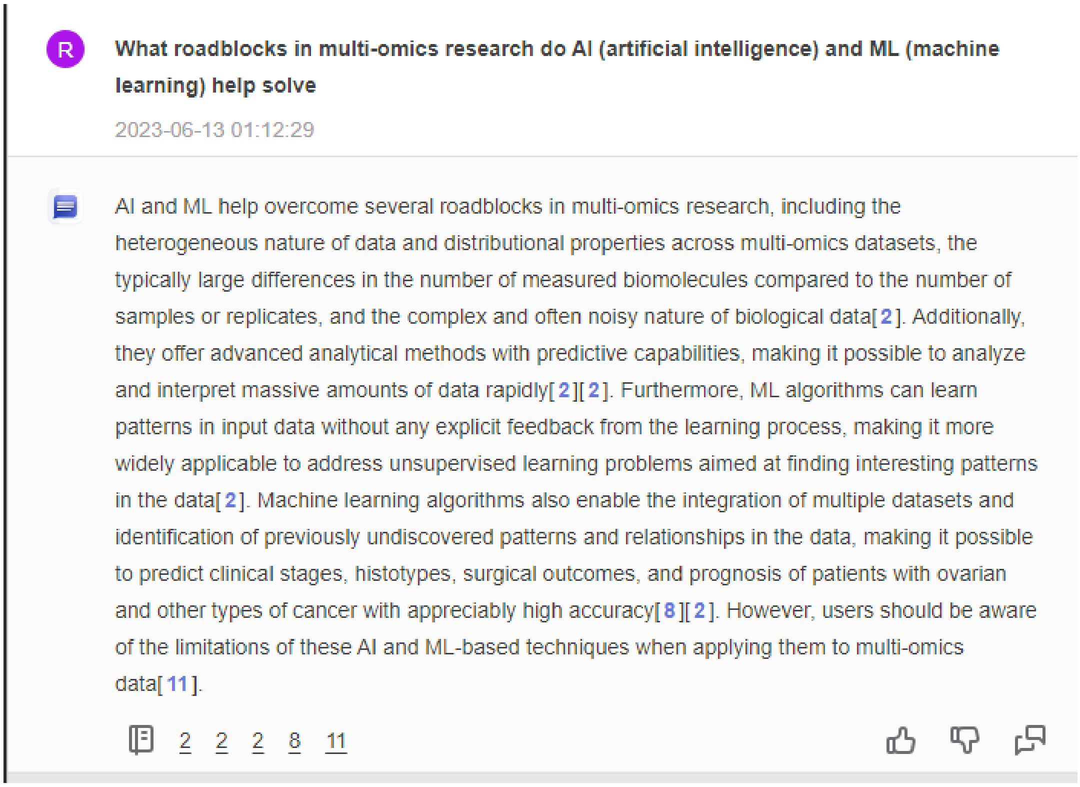
Keywords: Multi-Omics, AI, Artificial Intelligence, ML, Machine Learning

Search Time: ~15 minutes

Title	Journal	Citations	Year Published	ChatDoc Folder	Used in Responses
Making multi-omics data accessible to researchers	Scientific Data	105	2019	✔	✔
Practicing precision medicine with intelligently integrative clinical and multi-omics data analysis	Human Genomics	55	2020	✔	✔
Artificial Intelligence (AI)-Based Systems Biology Approaches in Multi-Omics Data Analysis of Cancer	Frontiers in Oncology	48	2020	✔	❌
Integrated Multi-Omics Analyses in Oncology: A Review of Machine Learning Methods and Tools	Frontiers in Oncology	138	2020	✔	✔

Gut microbiome-mediated epigenetic regulation of brain disorder and application of machine learning for multi-omics data analysis	Genome	19	2021	✔	✔
A New Era of Neuro-Oncology Research Pioneered by Multi-Omics Analysis and Machine Learning	Biomolecules	10	2021	✔	✔
Multi-Omics and Artificial Intelligence-Guided Data Integration in Chronic Liver Disease: Prospects and Challenges for Precision Medicine	OMICS	5	2022	✔	❌
Machine Learning: A New Prospect in Multi-Omics Data Analysis of Cancer	Frontiers in Genetics	23	2022	✔	✔
Integration of artificial intelligence and multi-omics in kidney diseases	Seminars in Cancer Biology	2	2022	✔	✔
Multi-omics approaches for biomarker discovery in early ovarian cancer diagnosis	eBioMedicine	21	2022	✔	❌
Missing data in multi-omics integration: Recent advances through artificial intelligence	Frontiers in Artificial Intelligence	1	2023	✔	✔
Editorial: Advances in methods and tools for multi-omics data analysis	Frontiers in Molecular Biosciences	0	2023	✔	✔
Dealing with dimensionality: the application of machine learning to multi-omics data	Bioinformatics	3	2023	✔	❌

Synthesis: (Time 60 minutes) Very general questions were first asked to ChatDoc to fill out the outline generated by ChatGPT (Example Question Below). When key terms came up that could have value for deeper understanding (example: problems with Heterogeneous Data), ChatDoc was prompted to expand more on these ideas. Flexibility with the outline is key in achieving good results, when new ideas pop up that were not initially present in the outline, new directions for the content brief may be found. Due to poor writing quality, ChatGPT was used to rewrite and condense responses into the final content brief.



R

What roadblocks in multi-omics research do AI (artificial intelligence) and ML (machine learning) help solve

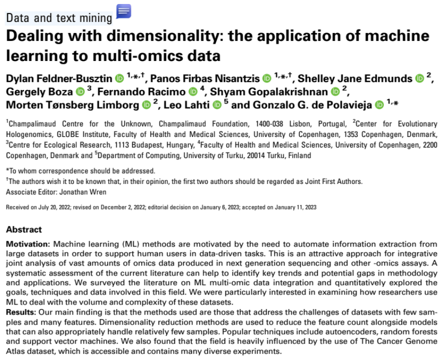
2023-06-13 01:12:29

AI and ML help overcome several roadblocks in multi-omics research, including the heterogeneous nature of data and distributional properties across multi-omics datasets, the typically large differences in the number of measured biomolecules compared to the number of samples or replicates, and the complex and often noisy nature of biological data[2]. Additionally, they offer advanced analytical methods with predictive capabilities, making it possible to analyze and interpret massive amounts of data rapidly[2][2]. Furthermore, ML algorithms can learn patterns in input data without any explicit feedback from the learning process, making it more widely applicable to address unsupervised learning problems aimed at finding interesting patterns in the data[2]. Machine learning algorithms also enable the integration of multiple datasets and identification of previously undiscovered patterns and relationships in the data, making it possible to predict clinical stages, histotypes, surgical outcomes, and prognosis of patients with ovarian and other types of cancer with appreciably high accuracy[8][2]. However, users should be aware of the limitations of these AI and ML-based techniques when applying them to multi-omics data[11].

Overcoming Challenges Associated With ChatDoc

AI & Prompting Bias:

Problem: The AI's selection of papers to generate a response is biased towards those that align closely with the prompt, even if other papers in the folder contain valuable information. The criteria for paper selection are unclear, potentially leading to arbitrary exclusions.



Data and text mining

Dealing with dimensionality: the application of machine learning to multi-omics data

Dylan Feldner-Busztin [1,*,†], Panos Firbas Nisantzis [1,*,†], Shelley Jane Edmunds [2], Gergely Boza [3], Fernando Racimo [4], Shyam Gopalakrishnan [2], Morten Tønsberg Limborg [2], Leo Lahti [5] and Gonzalo G. de Polavieja [1,*]

[1]Champalimaud Centre for the Unknown, Champalimaud Foundation, 1400-038 Lisbon, Portugal, [2]Center for Evolutionary Hologenomics, GLOBE Institute, Faculty of Health and Medical Sciences, University of Copenhagen, 1353 Copenhagen, Denmark, [3]Centre for Ecological Research, 1113 Budapest, Hungary, [4]Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen, Denmark and [5]Department of Computing, University of Turku, 20014 Turku, Finland

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Jonathan Wren

Received on July 20, 2022; revised on December 2, 2022; editorial decision on January 6, 2023; accepted on January 11, 2023

Abstract

Motivation: Machine learning (ML) methods are motivated by the need to automate information extraction from large datasets in order to support human users in data-driven tasks. This is an attractive approach for integrative joint analysis of vast amounts of omics data produced in next generation sequencing and other -omics assays. A systematic assessment of the current literature can help to identify key trends and potential gaps in methodology and applications. We surveyed the literature on ML multi-omic data integration and quantitatively explored the goals, techniques and data involved in this field. We were particularly interested in examining how researchers use ML to deal with the volume and complexity of these datasets.

Results: Our main finding is that the methods used are those that address the challenges of datasets with few samples and many features. Dimensionality reduction methods are used to reduce the feature count alongside models that can also appropriately handle relatively few samples. Popular techniques include autoencoders, random forests and support vector machines. We also found that the field is heavily influenced by the use of The Cancer Genome Atlas dataset, which is accessible and contains many diverse experiments.

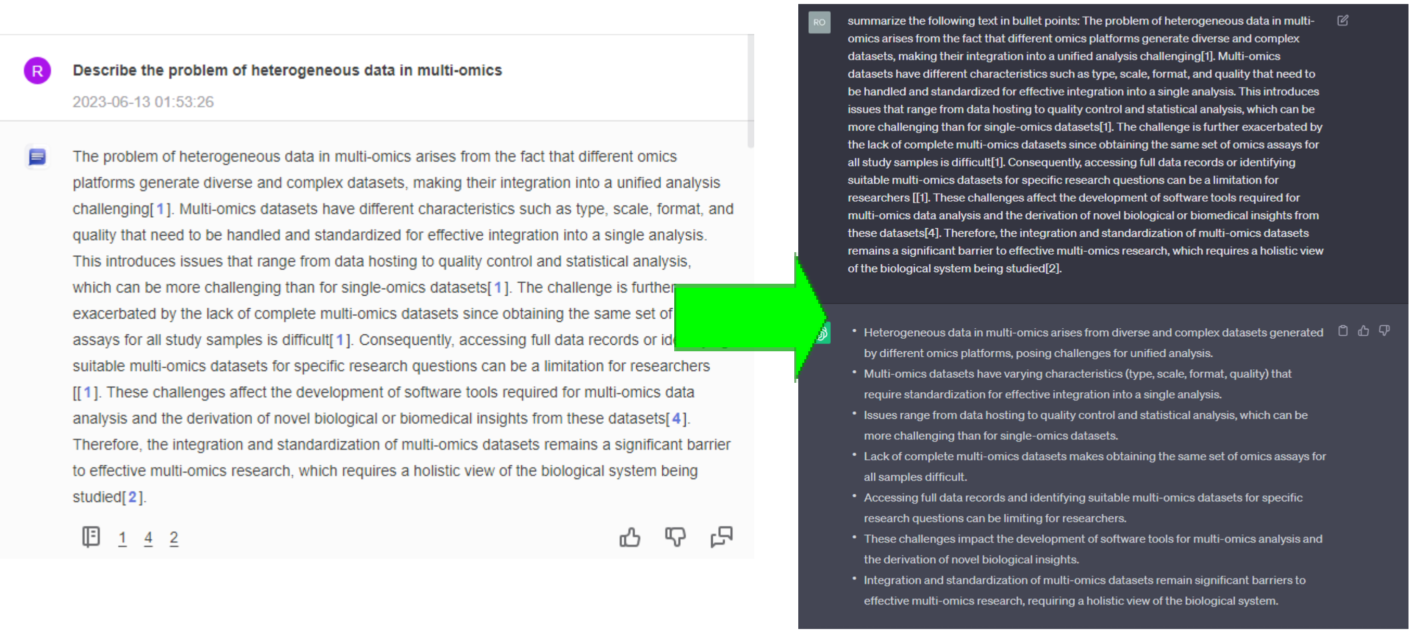
ChatDoc failed to draw from papers with relevant information. For example, a relevant paper identifying specifically the “dimensionality” of multi-omics data as a roadblock was ignored instead using less informative terms such as “complexity” of the data used in other papers. This was possibly due to oversaturation of papers with a potential response to the prompt leading to an arbitrary sampling process.

Solutions: To ensure inclusion of specific papers, exclusively prompt the AI using the desired paper when necessary. Further, prioritize quality over quantity of paper in the sampling selection.

Response Quality:

Problem: AI responses are often linguistically low in quality, often just regurgitating information. Additionally, responses are often not concise.

Solution: To improve response quality, condense answers into bullet points using ChatGPT and request tone changes or further rewriting.



Describe the problem of heterogeneous data in multi-omics

2023-06-13 01:53:26

The problem of heterogeneous data in multi-omics arises from the fact that different omics platforms generate diverse and complex datasets, making their integration into a unified analysis challenging[1]. Multi-omics datasets have different characteristics such as type, scale, format, and quality that need to be handled and standardized for effective integration into a single analysis. This introduces issues that range from data hosting to quality control and statistical analysis, which can be more challenging than for single-omics datasets[1]. The challenge is furthe
exacerbated by the lack of complete multi-omics datasets since obtaining the same set of
assays for all study samples is difficult[1]. Consequently, accessing full data records or id
suitable multi-omics datasets for specific research questions can be a limitation for researchers [[1]. These challenges affect the development of software tools required for multi-omics data analysis and the derivation of novel biological or biomedical insights from these datasets[4]. Therefore, the integration and standardization of multi-omics datasets remains a significant barrier to effective multi-omics research, which requires a holistic view of the biological system being studied[2].

summarize the following text in bullet points: The problem of heterogeneous data in multi-omics arises from the fact that different omics platforms generate diverse and complex datasets, making their integration into a unified analysis challenging[1]. Multi-omics datasets have different characteristics such as type, scale, format, and quality that need to be handled and standardized for effective integration into a single analysis. This introduces issues that range from data hosting to quality control and statistical analysis, which can be more challenging than for single-omics datasets[1]. The challenge is further exacerbated by the lack of complete multi-omics datasets since obtaining the same set of omics assays for all study samples is difficult[1]. Consequently, accessing full data records or identifying suitable multi-omics datasets for specific research questions can be a limitation for researchers [[1]. These challenges affect the development of software tools required for multi-omics data analysis and the derivation of novel biological or biomedical insights from these datasets[4]. Therefore, the integration and standardization of multi-omics datasets remains a significant barrier to effective multi-omics research, which requires a holistic view of the biological system being studied[2].

- Heterogeneous data in multi-omics arises from diverse and complex datasets generated by different omics platforms, posing challenges for unified analysis.
- Multi-omics datasets have varying characteristics (type, scale, format, quality) that require standardization for effective integration into a single analysis.
- Issues range from data hosting to quality control and statistical analysis, which can be more challenging than for single-omics datasets.
- Lack of complete multi-omics datasets makes obtaining the same set of omics assays for all samples difficult.
- Accessing full data records and identifying suitable multi-omics datasets for specific research questions can be limiting for researchers.
- These challenges impact the development of software tools for multi-omics analysis and the derivation of novel biological insights.
- Integration and standardization of multi-omics datasets remain significant barriers to effective multi-omics research, requiring a holistic view of the biological system.

ChatDoc is a valuable time-saving tool for extracting information from individual PDFs

Synthesizing research papers through AI tools is a new and rapidly improving time saving mechanism that may cut hours from research time. However, one must be aware of any limitations. These tools may result in overall lower comprehension of papers and the loss of nuanced idea generation and interpretations that may come from reading through literature. Analyzing multiple PDFs simultaneously increases time savings but amplifies the risks. Nonetheless, the time-saving capabilities of ChatDoc remain unmatched.

Future Directions

The digital tool landscape is rapidly evolving with tools emerging for almost any use case. Further research should be done regularly to keep up to date with future synthesis tools and updates to current ones. Additionally, alternatives to google scholar were not assessed and new and emerging digital tools may improve the quality of research papers obtained. Digital tools such as Elicit may aid in efficiently finding recent, relevant and reputable papers that may otherwise be missed. New tools should constantly be assessed for ease of use and quality of outputs, and new workflows should be developed around these tools to generate the best content most efficiently.