Plan Overview

A Data Management Plan created using DMPonline

Title: Developing an agent framework for biomedical data analysis and autonomous experimental design

Creator: Jingyu Sun

Principal Investigator: Jingyu Sun

Data Manager: Jingyu Sun

Project Administrator: Jingyu Sun

Affiliation: University of Manchester

Template: University of Manchester Generic Template

ORCID iD: 0009-0007-7184-949X

Project abstract:

This project aims to develop an AI agent for biomedical data analysis and autonomous experimental design. The system will rely on large language model (LLM) APIs provided by established vendors, including OpenAI, Google, and Anthropic, as backbone models for reasoning, planning, and orchestration. These models will not be trained or fine-tuned from scratch as part of the project. Instead, they will be used to support the execution of predefined analytical workflows and decision-making steps within the agent framework.

The project does not involve the use of sensitive personal data, patient-identifiable information, or unique restricted datasets. In addition, the LLM components are intended only to drive the agent’s execution of user-defined analysis tasks and experimental design logic; they will not be granted automatic access to raw datasets or unrestricted data ingestion capabilities. Any interaction with external LLM APIs will therefore be limited, controlled, and subject to appropriate data minimization and de-identification procedures.

ID: 200852

Start date: 02-03-2026

End date: 31-08-2026

Last modified: 20-04-2026

Copyright information:

Developing an agent framework for biomedical data analysis and autonomous experimental design

Manchester Data Management Outline

1. Will this project be reviewed by any of the following bodies (please select all that apply)?

- None of the above

2. Is The University of Manchester collaborating with other institutions on this project?

- Yes - Part of a collaboration and owning or handling data

Yes. This project involves collaboration between The University of Manchester and The University of Melbourne as part of a joint PhD programme. The project is led by a student enrolled under this joint arrangement.
The biological data used for analysis are provided by The University of Melbourne. The University of Manchester’s role is limited to technical support for the development of the AI agent and related computational methods. The University of Manchester is not responsible for the further collection, expansion, or generation of the underlying biological data.

3. What data will you use in this project (please select all that apply)?

- Re-use existing data (please list below)

Existing biological / biomedical research data provided by The University of Melbourne will be re-used in this project for analysis within a joint PhD collaboration. The University of Manchester will not acquire new raw data and will provide technical support only for the development and evaluation of the AI agent and associated computational workflows.

4. Where will the data be stored and backed-up during the project lifetime?

- P Drive (postgraduate researchers and students only)

5. If you will be using Research Data Storage, how much storage will you require?

- < 1 TB

6. Are you going to be receiving data from, or sharing data with an external third party?

- Yes

Yes. Existing biological research data will be received from The University of Melbourne, Australia, as part of a joint PhD collaboration. The University of Manchester will use these data only for the purposes of technical development and evaluation of the AI agent and associated computational workflows, and will not be responsible for further collection or expansion of the underlying data.
In addition, limited de-identified and data-minimised information may be shared with external commercial AI API providers, such as OpenAI, Google, and Anthropic, for the purpose of supporting the agent's reasoning and task orchestration functions. These models will not be given unrestricted access to raw datasets, and only the minimum necessary non-sensitive information will be transmitted.

7. How long do you intend to keep your data for after the end of your project (in years)?

- 0-4 years

Guidance for questions 8 to 13

Highly restricted information defined in the Information security classification, ownership and secure information handling SOP is information that requires enhanced security as unauthorised disclosure could cause significant harm to individuals or to the University and its ambitions in respect of its purpose, vision and values. This could be: information that is subject to export controls; valuable intellectual property; security sensitive material or research in key industrial fields at particular risk of being targeted by foreign states. See more examples of highly restricted information .

If you are using 'Very Sensitive' information as defined by the Information Security Classification, Ownerships and Secure Information Handling SOP, please consult the Information Governance Office for guidance.

Personal information, also known as personal data, relates to identifiable living individuals. Personal data is classed as special category personal data if it includes any of the following types of information about an identifiable living individual: racial or ethnic origin; political opinions; religious or similar philosophical beliefs; trade union membership; genetic data; biometric data; health data; sexual life; sexual orientation.

Please note that in line with data protection law (the UK General Data Protection Regulation and Data Protection Act 2018), personal information should only be stored in an identifiable form for as long as is necessary for the project; it should be pseudonymised (partially de-identified) and/or anonymised (completely de—identified) as soon as practically possible. You must obtain the appropriate ethical approval in order to use identifiable personal data.

8. What type of information will you be processing (please select all that apply)?

- No confidential or personal data

9. How do you plan to store, protect and ensure confidentiality of any highly restricted data or personal data (please select all that apply)?

- Not applicable

10. If you are storing personal information (including contact details) will you need to keep it beyond the end of the project?

- Not applicable

11. Will the participants' information (personal and/or sensitive) be shared with or accessed by anyone outside of the University of Manchester?

- Not applicable

12. If you will be sharing personal information outside of the University of Manchester will the individual or organisation you are sharing with be outside the EEA?

- Not applicable

13. Are you planning to use the personal information for future purposes such as research?

- No

14. Will this project use innovative technologies to collect or process data?

- Yes, and innovative technologies will not collect or process personal data (please list the innovative technologies below)

Yes. This project will use innovative AI-based technologies to support the processing of research data, specifically an AI agent framework built on third-party large language model (LLM) APIs provided by vendors such as OpenAI, Google, and Anthropic. These technologies will be used to support reasoning, task orchestration, and the execution of predefined analytical workflows in biomedical research.

The project does not involve training foundation models from scratch, and the LLM components will not be used to autonomously access or ingest raw datasets. Their role is limited to supporting controlled, predefined analysis and experimental design tasks. The project is not expected to process personal

data through these external AI systems; where external API use is required, only the minimum necessary, de-identified, and non-sensitive information will be shared.

15. Who will act as the data custodian for this study, and so be responsible for the information involved?

Hongpeng Zhou

16. Please provide the date on which this plan was last reviewed (dd/mm/yyyy).

2026-03-30

Project details

What is the purpose of your research project?

The purpose of this research project is to develop and evaluate an AI agent that can support biomedical data analysis and autonomous experimental design. The project aims to investigate how large language model–based agents can be used to assist with structured analytical workflows, scientific reasoning, and decision support in biomedical research settings.

The study will focus on designing an agent framework that can execute predefined analysis tasks, coordinate computational tools, and help generate or refine experimental plans based on user-defined objectives. The research also aims to assess the feasibility, reliability, and practical utility of such an agent for supporting biomedical researchers in routine data analysis and hypothesis-driven experimental planning.

What policies and guidelines on data management, data sharing, and data security are relevant to your research project?

This project has no external funder. It will follow the University of Manchester’s Research Data Management Policy and associated procedures, the University’s Data Protection Policy, relevant Information Governance and information security guidance, and the University’s guidance on the responsible use and secure development of AI systems. Where applicable, collaboration, data sharing, or data processing agreements with The University of Melbourne and external AI API providers will also be followed. ICO guidance on Data Protection Impact Assessments will be considered where relevant.

Responsibilities and Resources

Who will be responsible for data management?

Day-to-day responsibility for data management in this project will lie with the PGR student, **Jingyu Sun**, who is leading the project and will be responsible for the secure handling, organisation,

documentation, and appropriate use of project data in line with University policies and procedures.

What resources will you require to deliver your plan?

The project will require secure institutional data storage, access-controlled computing facilities, and standard research software support for data handling and analysis. In addition, a modest budget is required for the use of external AI APIs (e.g. OpenAI, Google, and Anthropic), which will be covered through the PGR student’s individual Research Training Support Grant (RTSG) budget. No substantial additional infrastructure or specialist data management resources are anticipated beyond existing University systems and standard project-level support.

Data Collection

What data will you collect or create?

This project will use existing biological data provided by The University of Melbourne, primarily pooled CRISPR screening data relating to lipid accumulation. No new raw biological data will be collected by The University of Manchester as part of this project.

The project may also generate derived research materials, including processed analysis outputs, result tables, code, workflow records, and technical documentation created during the development and evaluation of the AI agent. The project is not expected to involve sensitive personal data.

How will the data be collected or created?

The primary biological data will be supplied by The University of Melbourne and will consist of existing pooled CRISPR screening data relating to lipid accumulation. The University of Manchester will not collect new raw biological data.

Derived data will be created through controlled computational analysis of the supplied datasets, including preprocessing, statistical analysis, workflow execution, and evaluation of the AI agent. This may generate processed outputs, summary tables, code, configuration files, and technical documentation. External LLM APIs will be used only to support predefined agent functions and will not autonomously collect or ingest raw data.

Documentation and Metadata

What documentation and metadata will accompany the data?

The data will be accompanied by README files, sample metadata sheets, and methodological documentation describing who created the data, when they were created, what each file contains, how the data were generated or processed, and under what conditions they can be accessed. Variable definitions, sample identifiers, file formats, software used, processing steps, and key analytical

assumptions will also be recorded. Metadata will be stored in README documents, CSV/TSV sample sheets, and version-controlled workflow records. Where possible, non-proprietary formats and community-conventional pooled CRISPR screen file structures compatible with tools such as MAGeCK will be used.

Ethics and Legal Compliance

How will you manage any ethical issues?

No specific ethical issues are currently anticipated. The project does not involve human participants, sensitive personal data, or new data collection by The University of Manchester. Existing biological research data will be provided by The University of Melbourne, and any use of external AI APIs will be limited to de-identified, non-sensitive, and data-minimised information. If the project scope changes, appropriate ethical review will be sought.

How will you manage copyright and Intellectual Property Rights (IPR) issues?

Copyright and IPR will be managed by ensuring that existing biological data provided by The University of Melbourne are used only under the agreed terms of the joint PhD collaboration and are not redistributed without permission. Any third-party software, API services, documentation, or code used in the project will be used in accordance with their relevant licence terms and conditions. Project-generated outputs, including code, documentation, and derived analysis files, will be handled in line with applicable institutional and collaborative arrangements on ownership, access, and reuse.

Storage and backup

How will the data be stored and backed up?

Project data will be stored on secure, access-controlled University-approved storage systems and accessed only by authorised project personnel. Backups will rely on the standard backup and recovery arrangements provided through University-managed systems. Data will not be stored on unsecured personal devices or shared through unapproved platforms, and any working copies will be kept to a minimum and managed securely.

How will you manage access and security?

Access to the data will be restricted to authorised project personnel only and managed on a need-to-know basis using University-approved access controls. Data will be stored in secure institutional systems with appropriate authentication and permissions, and only the minimum necessary de-identified, non-sensitive information will be shared with external AI API providers where required. Raw datasets will not be made openly accessible to external services, and data handling will follow relevant University policies on information governance, data protection, and information security.

Selection and Preservation

Which data should be retained, shared, and/or preserved?

Data to be retained and preserved will include the final derived analysis outputs, processed result files, code, workflow configurations, and accompanying documentation/metadata necessary to understand and reproduce the research. Existing raw biological data provided by The University of Melbourne will be retained or preserved only in line with the terms under which they are supplied and any applicable collaborative arrangements. Data shared with others will be limited to materials that can be made available lawfully and appropriately, such as non-sensitive derived outputs, documentation, and code, while any restricted source data will remain access-controlled.

What is the long-term preservation plan for the dataset?

The long-term preservation plan is to retain the key derived research outputs, including processed results, code, workflow files, and essential documentation/metadata, in appropriate University-approved storage or repository systems for the required retention period. Preservation of the original biological source data will remain subject to the arrangements under which they are provided by The University of Melbourne. Any data retained long term will be preserved in formats that support future access and interpretation, with access controls maintained where necessary.

Data Sharing

How will you share the data?

Data will be shared in a controlled and proportionate way. Where appropriate, non-sensitive derived outputs, supporting documentation, and code may be shared through suitable academic or institutional channels, subject to any collaborative, copyright, confidentiality, or data access restrictions. Raw biological data provided by The University of Melbourne will not be shared onward by The University of Manchester unless this is permitted under the relevant agreements and governance arrangements.

Are any restrictions on data sharing required?

Yes. Restrictions on data sharing are required because the primary biological data are provided by The University of Melbourne and may only be used and shared in accordance with the relevant collaborative and access arrangements. In addition, any onward sharing must take account of confidentiality, intellectual property, and data protection requirements, and only non-sensitive derived outputs or documentation will be shared where appropriate.