AI Curation Methods for NASA Scientific Data

Walter Alvarado[1], Charlotte Nelson[2], Harlan Phillips[3], Erick Watanabe[2], Lauren Sanders[1], Hari Parthasarathy[3], Samrawit Gebre[1], Sigrid Reinsch[1], and Sylvain Costes[1]

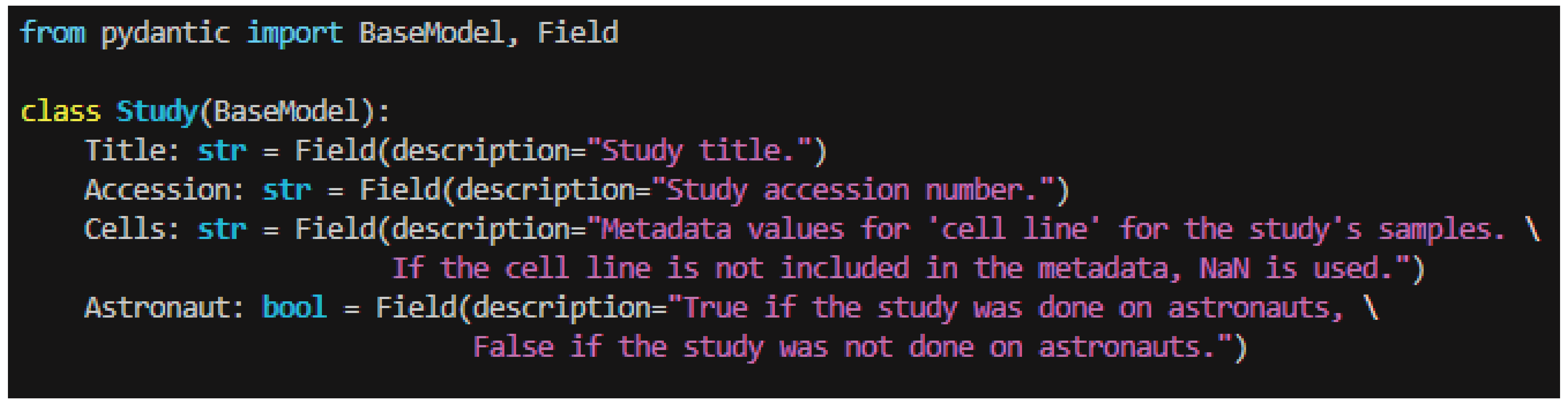
[1]NASA Ames Research Center, Mountain View, CA, [2]Blue Marble Space Institute of Science, Seattle, WA, [3]University of California, Berkeley, Berkeley, CA

Abstract

The AI for Curation project aims to integrate advanced LLM models into various aspects of our data curation workflow, enhancing the efficiency and accuracy from submission to user interaction. This initiative will impact multiple touchpoints, including data ingestion, curation processes, and user engagement with our curated datasets, affecting various centers, domains, and a broad user base. First, we are developing tools capable of parsing incoming data across all formats and utilizing LLMs to convert unstructured data into structured, standardized formats. This process streamlines curation by converting data into community-standard formats like ISA-Tab, significantly reducing the time and effort required by curators. This allows curators to allocate more resources to scientific analysis rather than data formatting tasks. Second, we are implementing AI/ML models to automate and enhance the accuracy of data validation and verification. These models ensure that the curated data adheres to high standards of quality and reliability, benefiting researchers and data users by providing them with rigorously verified datasets. Finally, we are developing a conversational agent (chatbot) that interfaces with our extensive repository of curated scientific studies on the Open Science Data Repository (OSDR). The chatbot enhances data discoverability by assisting users in navigating the knowledge base and referencing relevant studies. This improvement in accessibility makes scientific data more available to the community, thereby promoting the principles of open science.

Structured Outputs from LLM

In our efforts to enhance the AI for Curation project, we've developed advanced tools for automated tagging that significantly streamline the curation process. These tools are adept at identifying patterns within diverse datasets, automatically suggesting context-relevant metadata tags. This capability not only facilitates the initial organization of newly ingested data but also allows for the retrospective curation and categorization of existing datasets. Once data is aligned with community standards, these tools apply contextual tags to meticulously organize the information, rendering it AI-ready for subsequent training processes. For example, they can generate specific tags for studies involving astronauts or particular experimental conditions, ensuring a high level of precision and relevance in data handling. This method greatly enhances the accessibility and utility of the curated data, supporting more efficient research and analysis.



```python
from pydantic import BaseModel, Field

class Study(BaseModel):
    Title: str = Field(description="Study title.")
    Accession: str = Field(description="Study accession number.")
    Cells: str = Field(description="Metadata values for 'cell line' for the study's samples. \
                       If the cell line is not included in the metadata, NaN is used.")
    Astronaut: bool = Field(description="True if the study was done on astronauts, \
                         False if the study was not done on astronauts.")
```

Figure 1. Pydantic, a Python library for data validation and settings management, facilitates the creation of schemas that ensures responses from large language models (LLMs) conform to a specified structure. Here, a Pydantic class is used to define the structure for automated tagging of OSDR studies, illustrating how data is organized and standardized for AI processing.

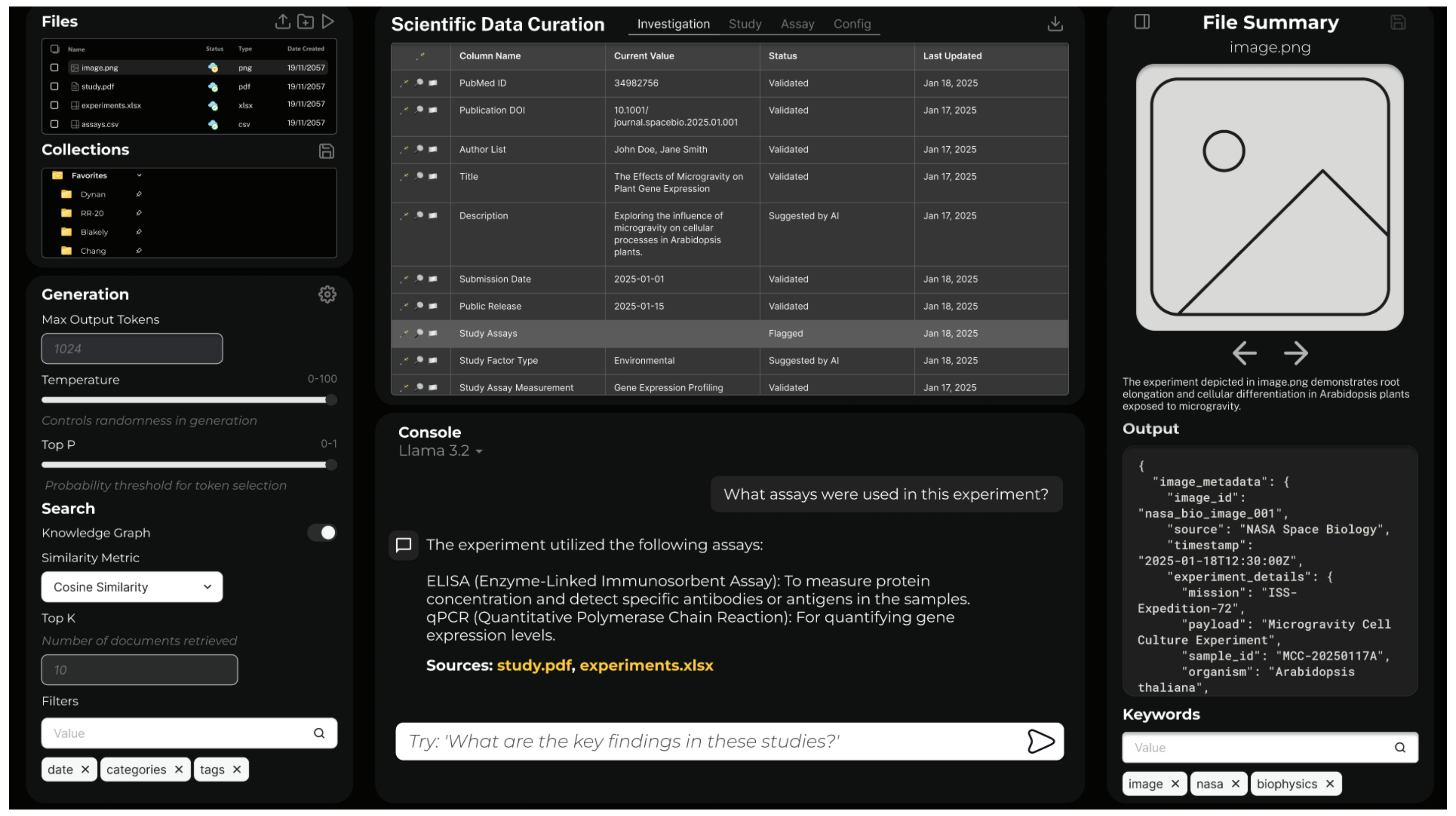
Data Alignment & Verification

Generative AI can be applied throughout the entire curation pipeline, offering numerous use cases, especially in data curation. One of the initial challenges addressed was the alignment of data received from researchers. Data often arrives in various formats and with disparate terminologies. To tackle this, we have developed tools that map and align these terms to standardized ontologies. For instance, while fields like "sex" and "genotype" may be consistent, others may use different terms for the same concepts.

Combining programmatic techniques with generative AI provides a robust toolkit for curating datasets. For example, we utilize similarity metrics such as the Levenshtein distance to match strings that are similar, handling a significant portion of the workload. For terms that use different words but are semantically similar, we leverage generative AI to understand the context and find the appropriate matches, enhancing the precision and utility of our curated datasets.

To further streamline the data alignment process, our graphical user interface (GUI) provides a suite of intuitive tools. These GUI capabilities are integral to ensuring that the data not only meets standardization requirements but also aligns with the users' specific research needs and contexts.

Users can choose from predefined templates, which promote consistency across different datasets. Once a template is selected, the system automatically identifies and matches the fields according to the template structure. This automated matching significantly reduces the time and effort required to align data.



Name	Status	Type	Date Created
image.png		png	19/11/2057
study.pdf		pdf	19/11/2057
experiments.xlsx		xlsx	19/11/2057
assays.csv		csv	19/11/2057

Generation

Max Output Tokens: 1024

Temperature 0-100
Controls randomness in generation

Top P 0-1
Probability threshold for token selection

Search

Knowledge Graph

Similarity Metric: Cosine Similarity

Top K
Number of documents retrieved
10

Filters

Scientific Data Curation

Column Name	Current Value	Status	Last Updated
PubMed ID	34982756	Validated	Jan 18, 2025
Publication DOI	10.1001/ journal.spacebio.2025.01.001	Validated	Jan 17, 2025
Author List	John Doe, Jane Smith	Validated	Jan 17, 2025
Title	The Effects of Microgravity on Plant Gene Expression	Validated	Jan 17, 2025
Description	Exploring the influence of microgravity on cellular processes in Arabidopsis plants.	Suggested by AI	Jan 17, 2025
Submission Date	2025-01-01	Validated	Jan 18, 2025
Public Release	2025-01-15	Validated	Jan 18, 2025
Study Assays		Flagged	Jan 18, 2025
Study Factor Type	Environmental	Suggested by AI	Jan 18, 2025
Study Assay Measurement	Gene Expression Profiling	Validated	Jan 17, 2025

Console
Llama 3.2

What assays were used in this experiment?

The experiment utilized the following assays:

ELISA (Enzyme-Linked Immunosorbent Assay): To measure protein concentration and detect specific antibodies or antigens in the samples.
qPCR (Quantitative Polymerase Chain Reaction): For quantifying gene expression levels.

Sources: study.pdf, experiments.xlsx

Try: 'What are the key findings in these studies?'

File Summary
image.png

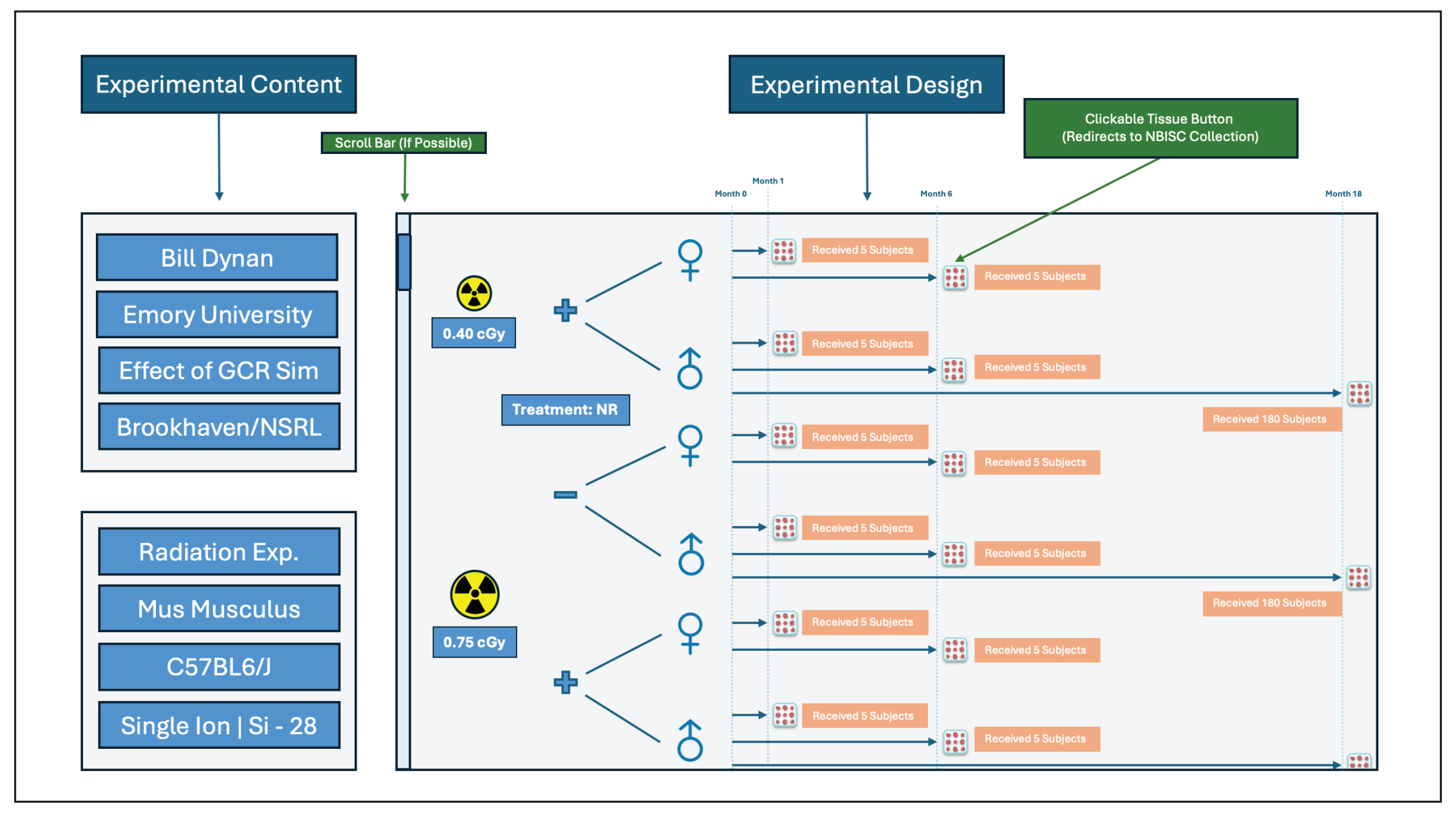
The experiment depicted in image.png demonstrates root elongation and cellular differentiation in Arabidopsis plants exposed to microgravity.

Output

```
{
  "image_metadata": {
    "image_id":
"nasa_bio_image_001",
    "source": "NASA Space Biology",
    "timestamp":
"2025-01-18T12:30:00Z",
    "experiment_details": {
      "mission": "ISS-
Expedition-72",
      "payload": "Microgravity Cell
Culture Experiment",
      "sample_id": "MCC-20250117A",
      "organism": "Arabidopsis
thaliana",
```

Keywords

Auto-Generated Graphical Abstracts



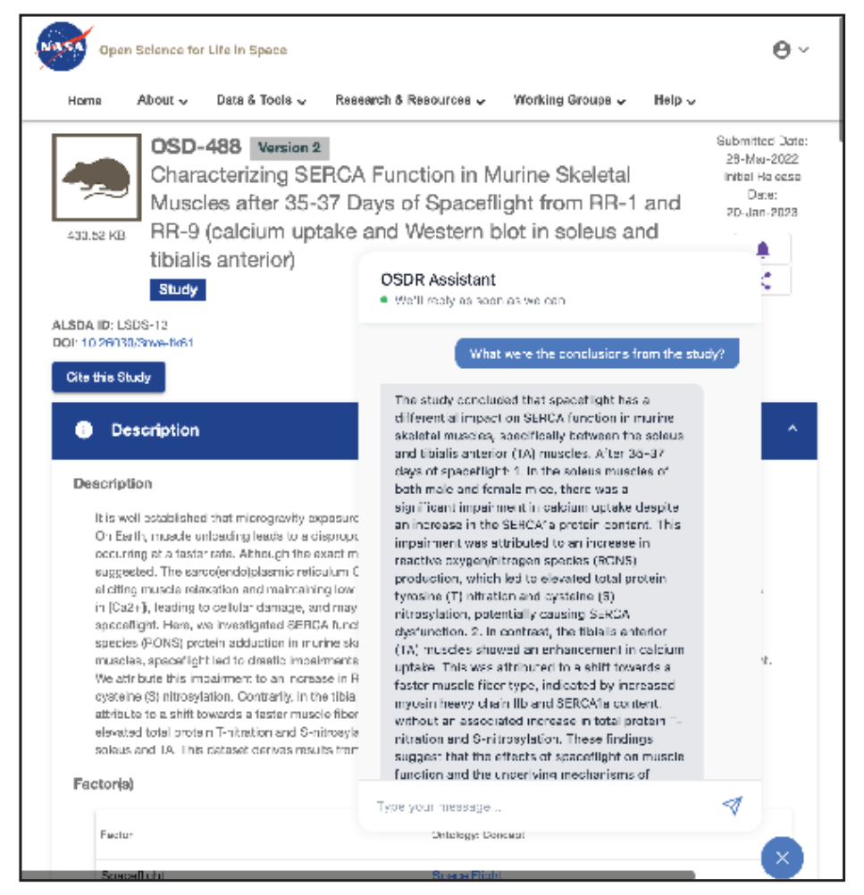
To enhance the presentation and understanding of experimental data, we levearge generative AI to create graphical abstracts. Users can manually input experimental parameters using a user-friendly interface, ensuring direct control over the data visualization process. Additionally, the system supports file imports, as well as intelligent extraction techniques using generative AI to parse and interpret data directly from scientific papers. This capability automates the visualization of complex experiments, transforming dense textual information into clear, informative graphical abstracts.

Enhancing Study Discoverability

We've implemented a chatbot to improve how researchers and users interact with our data. The chatbot integrates directly with our databases, pulling information in real-time to ensure that responses are always up-to-date and accurate. This facilitates dynamic interaction based on the latest data available. It also employs natural language processing to summarize intricate research findings or explain complex scientific concepts, making the information more accessible and easier to understand for a broader audience.

Additionally, the chatbot incorporates a Retrieval-Augmented Generation (RAG) model, which enhances its ability to generate precise and contextually relevant responses. The RAG model first retrieves relevant information from a vast database of scientific literature before producing a coherent reply, thereby improving the chatbot's effectiveness in navigating and discovering scientific studies. These integrated features ensure that critical information is not only more discoverable but also more understandable and secure within the scientific community.

This chatbot initiative aligns with our commitment to open science, facilitating broader access to knowledge and fostering a collaborative environment for innovation.



OSD-488 Version 2
Characterizing SERCA Function in Murine Skeletal Muscles after 35-37 Days of Spaceflight from RR-1 and RR-9 (calcium uptake and Western blot in soleus and tibialis anterior)
Study

433.52 KB

Submitted Date: 28-Mar-2022
Initial Release Date: 20-Jan-2023

ALSDA ID: LSDS-13
DOI: 10.26030/3rwe-tk61

Cite this Study

Description

Description

It is well established that microgravity exposure
On Earth, muscle unloading leads to a disprop
occurring at a faster rate. Although the exact m
suggested. The sarco(endo)plasmic reticulum C
eliciting muscle relaxation and maintaining low
in [Ca2+], leading to cellular damage, and may
spaceflight. Here, we investigated SERCA func
species (RONS) protein adduction in murine ske
muscles, spaceflight led to drastic impairments
We attribute this impairment to an increase in R
cysteine (S) nitrosylation. Contrarily, in the tibia
attribute to a shift towards a faster muscle fiber
elevated total protein T-nitration and S-nitrosyla
soleus and TA. This dataset derives results from

Factor(s)

Factor	Ontology Concept
Spaceflight	Space Flight

OSDR Assistant
We'll reply as soon as we can

What were the conclusions from the study?

The study concluded that spaceflight has a differential impact on SERCA function in murine skeletal muscles, specifically between the soleus and tibialis anterior (TA) muscles. After 35-37 days of spaceflight: 1. In the soleus muscles of both male and female mice, there was a significant impairment in calcium uptake despite an increase in the SERCA1a protein content. This impairment was attributed to an increase in reactive oxygen/nitrogen species (RONS) production, which led to elevated total protein tyrosine (T) nitration and cysteine (S) nitrosylation, potentially causing SERCA dysfunction. 2. In contrast, the tibialis anterior (TA) muscles showed an enhancement in calcium uptake. This was attributed to a shift towards a faster muscle fiber type, indicated by increased myosin heavy chain IIb and SERCA1a content, without an associated increase in total protein T-nitration and S-nitrosylation. These findings suggest that the effects of spaceflight on muscle function and the underlying mechanisms of

Type your message ...

References

1. Colvin, S., Jolibois, E., Ramezani, H., Garcia Badaracco, A., Dorsey, T., Montague, D., Matveenko, S., Trylesinski, M., Runkle, S., Hewitt, D., & Hall, A. (2024). Pydantic (Version 2.9.0) https://github.com/pydantic/pydantic
2. Gao, Yunfan, et al. "Retrieval-augmented generation for large language models: A survey." arXiv preprint arXiv:2312.10997 (2023).
3. Dubey, Abhimanyu, et al. "The llama 3 herd of models." arXiv preprint arXiv:2407.21783 (2024).
4. Data are courtesy of the NASA Open Science Data Repository (OSDR). NASA Ames Research Center. (2024).

We extend our thanks to the Ames Life Sciences Data Archive (ALSDA), NASA Biological Institutional Scientific Collection (NBISC), and NASA GeneLab, integral projects of the Open Science Data Repository (OSDR), for their invaluable contributions to this project. Additionally, we are grateful to the AI for Life Science (AI4LS) group for their innovative approaches that have significantly enhanced our data analysis processes. Our appreciation also goes out to the Analysis Working Group (AWG), a team of dedicated volunteers, whose expertise and commitment have been instrumental in refining our research outputs.

SILICON VALLEY
AMES RESEARCH CENTER