NASA GENELAB - KNOWLEDGE GRAPH FABRIC ENABLES DEEP BIOMEDICAL ANALYSIS OF MULTI-OMICS DATASETS
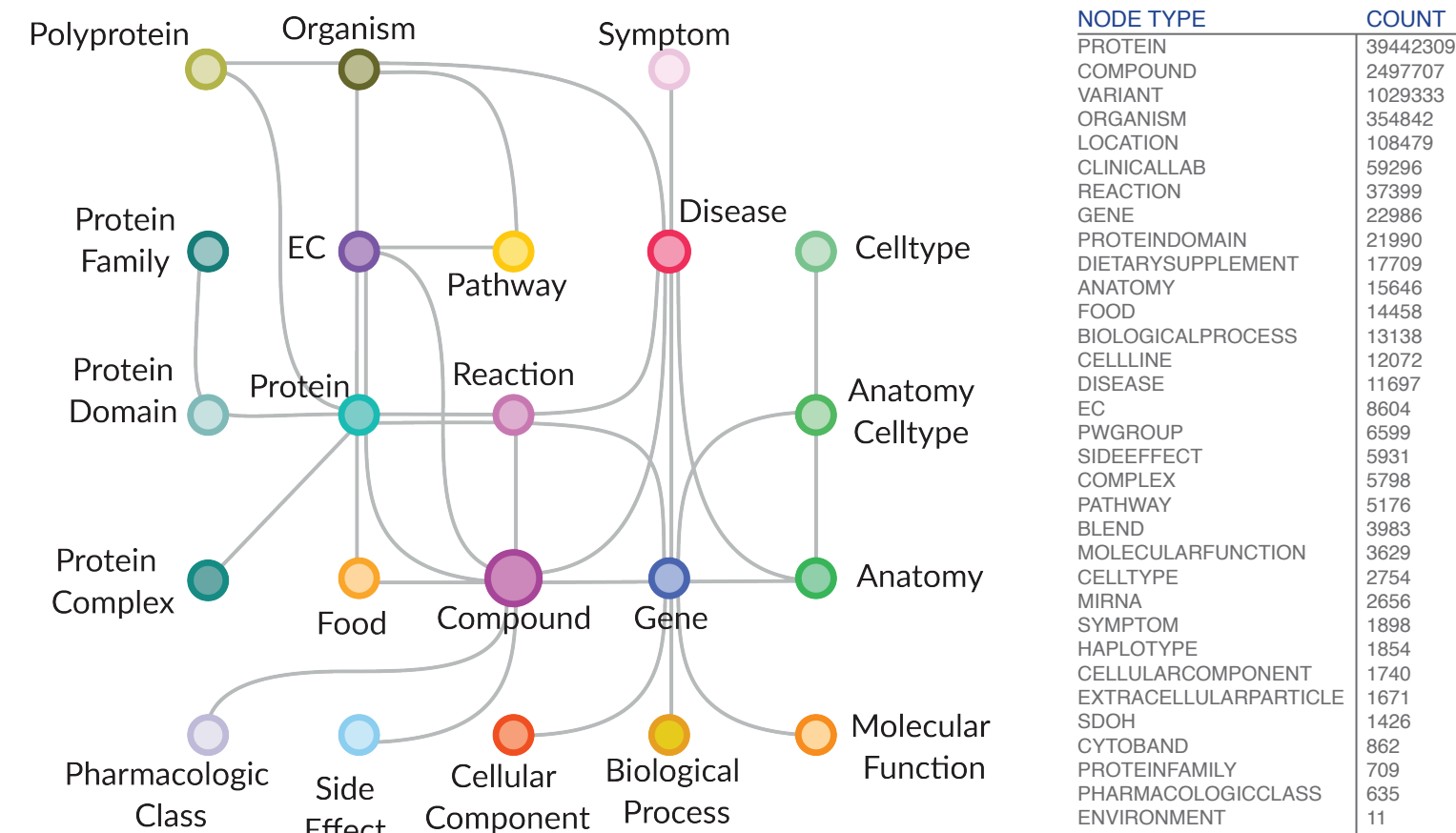
C. A. Nelson[1], P. W. Rose[2], K. Soman[3], L. M. Sanders[1], S. G. Gebre[4], S. V. Costes[4], S. E. Baranzini[3]

[1]Blue Marble Space Institute of Science, NASA Ames Research Center, Moffett Field, CA 94035, USA [2]San Diego Supercomputer Center, University of California, San Diego, La Jolla, CA 92093, USA [3]Department of Neurology, Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA 94158, USA [4]Space Biosciences Division, NASA Ames Research Center, Moffett Field, CA 94035, USA

INTRODUCTION

The limited number of astronauts and human samples from long-duration space missions pose significant challenges for studying the health risks associated with spaceflight and developing new treatments. As a result, much of our understanding of the biological impact of space travel relies on samples from model organisms. NASA GeneLab, integrated into Open Science Data Repository (OSDR) is a centralized multi-omics resource containing almost 1000 datasets from over 500 space-related studies from human and model organism samples. Previous studies have demonstrated that human phenotypes and physiological changes caused by spaceflight can be identified by connecting gene expression data from model organisms flown in space to a biomedical knowledge graph (SPOKE)[1]. In this work, we present a data fabric connecting OSDR datasets to SPOKE that empowers biomedical analyses through the GeneLab visualization portal. This collaboration is funded by NSF's Proto-OKN program.

SPOKE KNOWLEDGE GRAPH

NODE TYPE	COUNT
PROTEIN	39442309
COMPOUND	2497707
VARIANT	1029333
ORGANISM	354842
LOCATION	108479
CLINICALLAB	59296
REACTION	37399
GENE	22986
PROTEINDOMAIN	21990
DIETARYSUPPLEMENT	17709
ANATOMY	15646
FOOD	14458
BIOLOGICALPROCESS	13138
CELLLINE	12072
DISEASE	11697
EC	8604
PWGROUP	6599
SIDEEFFECT	5931
COMPLEX	5798
PATHWAY	5176
BLEND	3983
MOLECULARFUNCTION	3629
CELLTYPE	2754
MIRNA	2656
SYMPTOM	1898
HAPLOTYPE	1854
CELLULARCOMPONENT	1740
EXTRACELLULARPARTICLE	1671
SDOH	1426
CYTOBAND	862
PROTEINFAMILY	709
PHARMACOLOGICCLASS	635
ENVIRONMENT	11

The Scalable Precision Medicine Open Knowledge Engine (SPOKE) is a heterogeneous population-level knowledge graph. SPOKE is rebuilt on a weekly basis – pulling data from 60+ foundational databases. It currently contains over 43 million nodes of 33 types and 184 million edges of 116 types.

SPOKE-OSDR DATA FABRIC
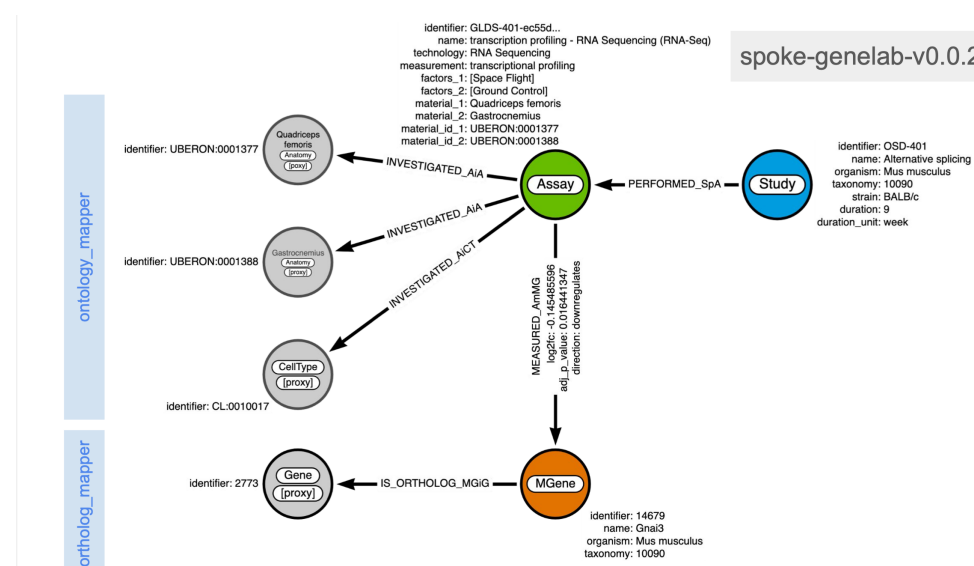
The GeneLab data was extracted and reformatted into a graph representation to create the GeneLab knowledge graph (GeneLab KG). The architecture of the GeneLab KG was designed to encompass all meta and experimental data. This was achieved by pulling data with the OSDR (GeneLab) API and restructuring the information into node and edge files. These files were then imported into a Neo4j database using kg-import. Since the initial focus of the fabric was the integration of all transcriptomics datasets, model organism gene nodes (MGene) were connected to Assay nodes with the MEASURED_AmMG edge. The corresponding p-value, log2 fold change, and direction were embedded into each edge. MGene nodes were then connected to proxy SPOKE Gene nodes, which enabled the coupling of the two knowledge graphs in the fabric.

## SPOKE ANALYSES IN GENELAB VISUALIZATION PORTAL
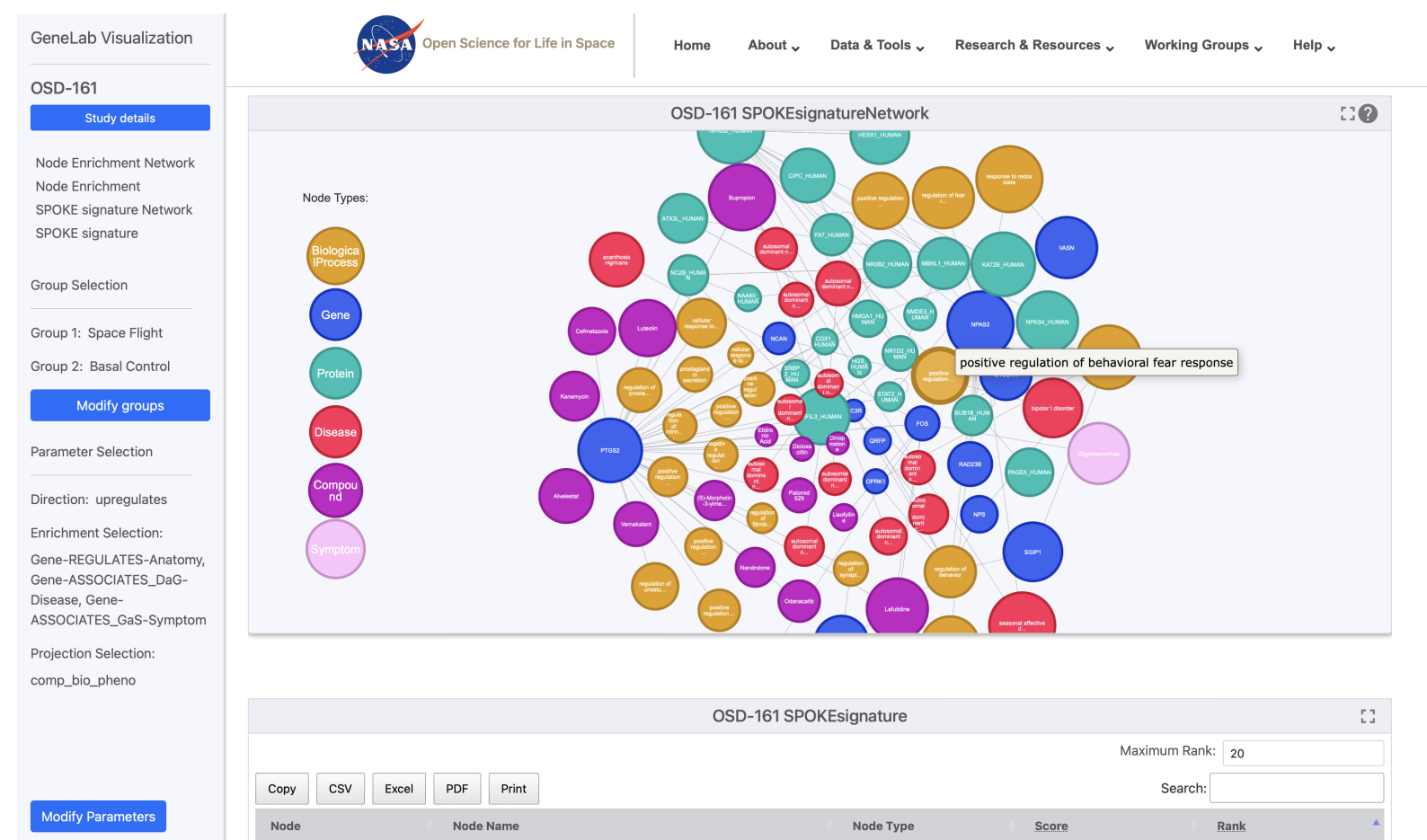
An API and user interface were developed to increase the fabric's accessibility. Through the GeneLab Visualization Portal, researchers can now generate SPOKE Signatures and examine node enrichment for the GeneLab transcriptomics datasets.The above figure shows the SPOKE Signature generated using transcriptomic data from mouse adrenal glands (OSD-161; Space

NODE ENRICHMENT
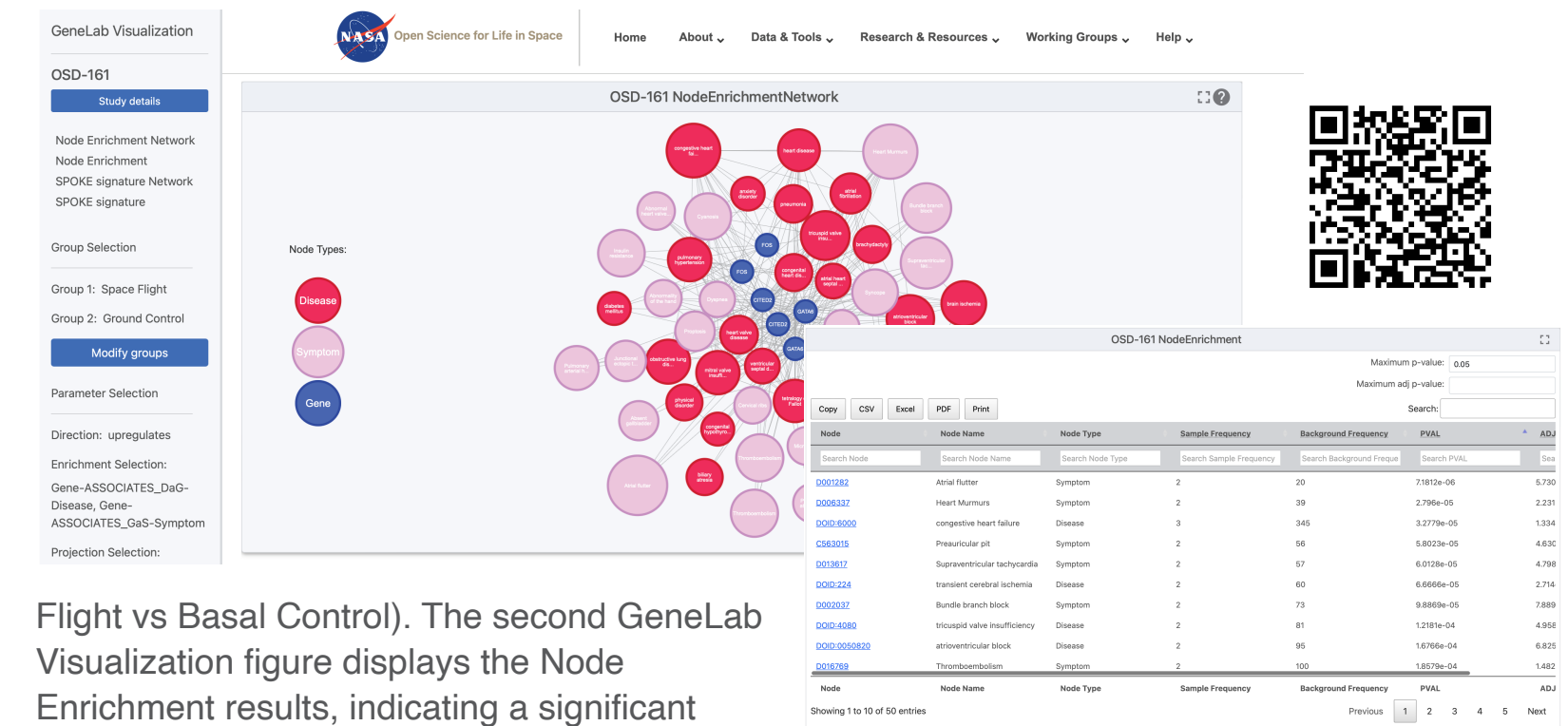
Flight vs Basal Control). The second GeneLab Visualization figure displays the Node Enrichment results, indicating a significant increase in heart-related issues among mice during space flight.Both analyses suggest that the space flight mice may experience a fight-or-flight response.

RESULTS & CONCLUSION
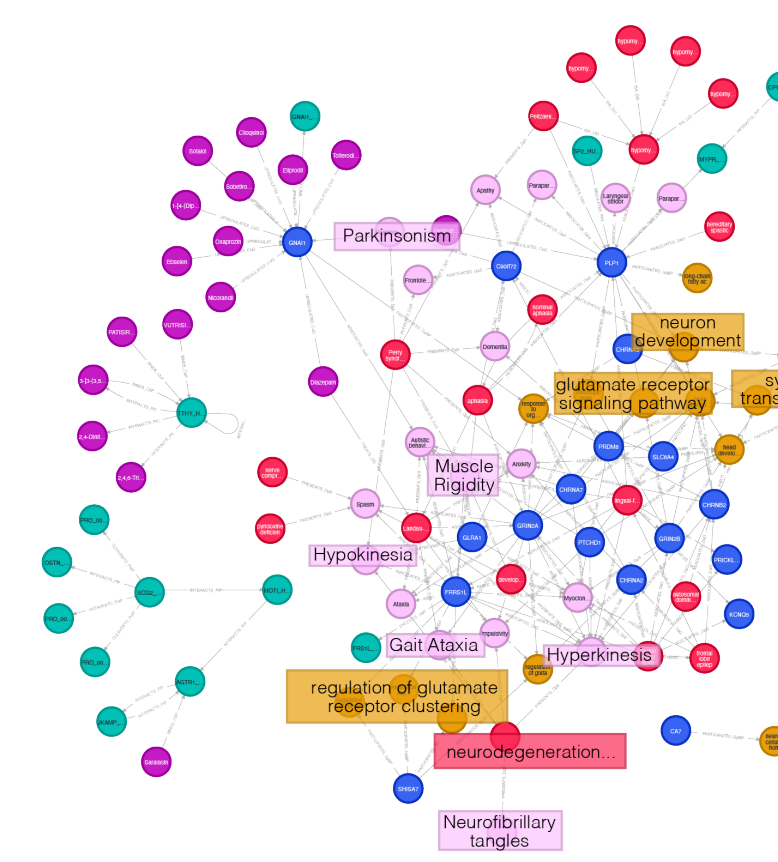
To examine the API's output, the results were compared to a recent spatial transcriptomics study that identified neurodegeneration-associated pathways in cortical neurons and potential similarities with diseases such as Parkinson's disease [2]. Both the SPOKE Signature and Node Enrichment tools in the GeneLab visualization portal returned nodes related to Parkinson's disease such as Parkinsonism, Tremor, Muscle Spasticity, Gait Ataxia, Hyperkinesis, Hypokinesia, Muscle Rigidity, and Neurofibrillary tangles (Figure 3). These early results suggest the GeneLab-SPOKE data fabric is a powerful tool for extracting biomedical insights from spaceflight datasets.

REFERENCES

[1] Nelson, C. A. et al (2021) Life 11.1: 42. [2]Masarapu, Y. et al (2024) Nature Communications 15.1 4778.