Machine Learning, Artificial Intelligence and Data Modeling for the Next Decade of Space Biology Research and Astronaut Health Support

- Lauren M. Sanders, Blue Marble Space Institute of Science, Space Biosciences Division, NASA GeneLab, NASA Ames Research Center, (530) 409-2174, lauren.m.sanders@nasa.gov
- Ryan T. Scott, KBR, Space Biosciences Division, NASA Ames Life Sciences Data Archive, NASA Ames Research Center, (650) 316-9901, ryan.t.scott@nasa.gov
- Sylvain V. Costes, Space Biosciences Division, NASA Ames Research Center, (650) 604-5343, sylvain.v.costes@nasa.gov

Authors from Workshop on AI and Modeling for Space Biology: NASA Ames, June 2021

Erik Antonsen, Baylor College of Medicine, erik.l.letonsen@nasa.gov
Joel Babdor, University of California San Francisco, joel.babdor@ucsf.edu
Sergio E. Baranzini, University of California San Francisco, Sergio.Baranzini@ucsf.edu
Richard Barker, University of Wisconsin-Madison, dr.richard.barker@gmail.com
Afshin Beheshti, KBR-NASA Ames Research Center, afshin.beheshti@nasa.gov
Daniel C. Berrios, KBR-NASA Ames Research Center, daniel.c.berrios@nasa.gov
Stuart Chalk, University of North Florida, schalk@unf.edu
Guillermo M. Delgado-Aparicio, Georgia Tech, gmo.delgado@gatech.edu
Hector Garcia Martin, Lawrence Berkeley National Lab, hgmartin@lbl.gov
Benjamin S Glicksberg, Icahn School of Medicine at Mount Sinai, benjamin.glicksberg@msm.edu
Casey S. Greene, University of Colorado School of Medicine, casey.s.greene@cuanschutz.edu
Melissa Haendel, University of California San Francisco, melissa@tislab.org
Arif A. Hamid, University of Minnesota, arifhamid.DA@gmail.com
Jaden J A Hastings, Weill Cornell Medicine, jah4006@med.cornell.edu
Philip Heller, San José State University, philip.heller@sjsu.edu
Adrienne Hoarfrost, USRA-NASA Ames Research Center, adrienne.l.hoarfrost@nasa.gov
Daniel Jamieson, Biorelate, dan@biorelate.com
Katelyn J. Jarvis, University of California San Francisco, kjjarvis@ucdavis.edu
John Kalantari, Mayo Clinic, kalantari.john@mayo.edu
Kia Khezeli, Mayo Clinic, khezeli.kia@mayo.edu
Svetlana V. Komarova, McGill University, svetlana.komarova@mcgill.ca
Matthieu Komorowski, Imperial College London, matthieu.komorowski@gmail.com
Prachi Kothiyal, SymbioSeq LLC-NASA Johnson Space Center, prachi.kothiyal@nasa.gov
Graham Mackintosh, BAERI, NASA Ames Research Center, graham.mackintosh@nasa.gov
Ashish Mahabal, California Institute of Technology, ashish@caltech.edu
Uri Manor, Salk Institute for Biological Studies, umanor@salk.edu
Mona Matar, NASA Glenn Research Center, mona.matar@nasa.gov
George I. Mias, Michigan State University, george.mias@gmail.com
Jack Miller, KBR-NASA Ames Research Center, j_miller@lbl.gov
Jerry G. Myers, Jr., NASA Glenn Research Center, jerry.g.myers@nasa.gov
Charlotte Nelson, University of California San Francisco, charlotte.nelson@ucsf.edu
Jonathan Oribello, Blue Marble Space Institute of Science-NASA Ames Research Center, jonathan.d.oribello@nasa.gov
Seung-min Park, Stanford University School of Medicine, sp293@stanford.edu
Patricia Parsons-Wingerter, NASA Glenn Research Center, patricia.a.parsons-wingerter@nasa.gov
R. K. Prabhu, USRA, NASA Glenn Research Center, raj.prabhu@nasa.gov
Amina Ann Qutub, University of Texas, San Antonio and UT Health Sciences, amina.qutub@utsa.edu
Jon Rask, NASA Ames Research Center, jon.c.rask@nasa.gov
Robert J. Reynolds, Mortality Research & Consulting, Inc., rreymonds@mortalityresearch.com
Amanda Saravia-Butler, Logyx-NASA Ames Research Center, amanda.m.saravia-butler@nasa.gov
Suchi Saria, Johns Hopkins University and Bayesian Health, ssaria@cs.jhu.edu
Aenor J. Sawyer, University of California San Francisco, aenor.sawyer@ucsf.edu
Nitin Kumar Singh, Jet Propulsion Laboratory, nitin.k.singh@jpl.nasa.gov
Frank Soboczenski, King College London, frank.soboczenski@kcl.ac.uk
Michael Snyder, Stanford University School of Medicine, mpsnyder@stanford.edu
Karthik Soman, University of California San Francisco, karthik.soman@ucsf.edu
Corey A. Theriot, UTMB, NASA Johnson Space Center, corey.theriot@nasa.gov
David Van Valen, California Institute of Technology, vanvalen@caltech.edu
Kasthuri Venkateswaran, Jet Propulsion Laboratory, kasthuri.j.venkateswaran@jpl.nasa.gov
Liz Warren, Center for the Advancement of Space in Science, lwarren@issnationallab.org
Liz Worthey, University of Alabama Birmingham, eworthey@peds.uab.edu
Jason H. Yang, Rutgers New Jersey Medical School, jason.y@rutgers.edu
Marinka Zitnik, Harvard University, marinka@hms.harvard.edu

Endorser: Susana B. Zanello, KBR, NASA Johnson Space Center, susana.b.zanello@nasa.gov
Abstract
We propose a ten-year research campaign to maximally adopt artificial intelligence (AI) and machine learning (ML) capabilities in space biology research and the design of spaceflight health systems. Fully leveraging AI/ML functionalities will enable generating and analyzing large amounts of data while requiring limited human time, and increasing knowledge gain of biological spaceflight effects. Further, integration of AI/ML into support systems for spacecraft, ecosystem, and astronaut health will increase the power of adverse event prediction and mitigation.

Introduction
Machine learning (ML) trains a mathematical model on sample data to predict an outcome of interest, and lies at the core of artificial intelligence (AI) algorithms[1]. AI/ML streamlines myriad aspects of our lives on Earth, from purchase recommendations to Google search results to self-driving cars. From a market size value in 2021 of US$ 93 billion, the revenue growth forecast for AI/ML in 2028 is nearly a trillion, at US$ 997 billion[2]. Notwithstanding continued capital growth and refinement expected in this relatively new scientific realm, biomedical research today has already successfully exploited AI/ML technology to empower scientists[3] and physicians[4] to learn from high dimensional datasets. For example, AI/ML toolsets allow researchers to predict disease onset[5], survival[6], drug response[7], or protein structure[8] and to classify microbe strain[9], disease stage[10], and disease diagnosis[11]. Such modeling ability is uniquely suited to research on effects and mitigation of spaceflight hazards, a field that necessarily generates small, sparse, and heterogeneous datasets. The stressors experienced by humans during spaceflight are profound yet incompletely understood. Further, as humanity travels beyond low Earth orbit (LEO) and into deep space, investment in appropriately autonomous, Earth-independent medical, biomonitoring, and science research systems are needed. Crew performance prediction, clinical decision support including prophylactic recommendation systems, and early detections of diseases are all areas that AI/ML can support.

Here we propose a research campaign for the next decade supporting a large-scale project centered in the NASA Biological and Physical Sciences Division (BPS), but drawing on far-reaching collaborations and expertise involving computer science, technology developers, engineering, operations and medicine. The purpose of this campaign is to enable the adoption and usage of AI/ML methods to model and maximally automate 1) space biology research, 2) spacecraft and habitat ecosystem health, and 3) both risk prediction and support for astronaut health. We recommend addressing the following four goals, in an order which represents the continuum from basic scientific research to human health support and risk mitigation (Figure 1). Each goal leverages and builds upon the previous:

1. (Earth-based) Establish Open Science cloud-based data sharing mechanisms for space biomedical-relevant research data and AI readiness, via these subgoals:
   a. Make all space biomedical data FAIR (findable, accessible, interoperable, reusable)[12] in a maximally open-access archive (e.g., omics, phenotypic, imaging, video, behavioral, telemetry, pertinent mission data), helping to ensure biomedical evidence is reproducible and transparent (because the greatest potential for knowledge gain comes from data reuse, data should be as open as possible, closed as necessary; metadata can always be open[13]).
b. Develop knowledge discovery tools (meta-analysis, visualization, AI/ML modeling portals) to enable involvement from a broadly diverse field of researchers and citizen scientists.

c. Migrate analyses and data into a data management environment (DME) with quality control tools, to streamline collection of trustworthy and curated data from investigators and missions (both ground and space). The DME will eventually be geared for collecting space-generated data from deep space communications networks. Digital research notebooks for continuous data collection will be essential for ground and space research.

2. *(Earth-based)* Identify existing AI/ML algorithms available for modification/adaptation and develop corresponding novel methodologies to fill gaps in current research capabilities (e.g., methods capable of working with low sample number, such as few-shot learning[14], since space biology experiments cannot fly large numbers of samples);

3. *(Mission)* Develop verified and validated AI/ML-powered automated research facilities and analysis platforms on board ISS, Cubesats, Lunar laboratories[15], and deep space missions, leveraging deep space and cloud DME network data connections;

4. *(Mission)* Develop an AI/ML powered continuous space health monitoring system with human and environmental sensors, building on research data and employing predictive modeling, for a ‘Precision Space Health’ astronaut health system for spacecraft and habitats beyond LEO missions when Earth-based support becomes limited.

This Research Campaign white paper reports the findings from the NASA AI and Modeling in Space Biology Workshop held in June 2021, which assembled a large cohort of external-to-NASA subject matter experts to develop recommendations for the next decade. Two manuscripts represent companion articles to this Research Campaign white paper, and are invited review-recommendation submissions to Nature Machine Intelligence. Scott et al., 2021 is available as an unpublished pre-print[16] and presents required biomonitoring technology, biomarker science, spacecraft hardware, intelligent software, and streamlined data management for a Precision Space Health system. Sanders et al. 2021 is available as an unpublished pre-print[17] and presents self-driving labs, data management systems, tailored space biology AI algorithms, digital twins, and cellular engineering efforts towards enabling deep space life science research. The proposed timeline and budget for this Research Campaign is shown in **Figure 2**.

**Goal 1:** Open Science, Data Management, and Modeling Environment

The first step in effectively deploying computational modeling techniques such as AI/ML is the standardization and open-source availability of experimental data and analysis pipelines. Structured data repositories[18–22] and scientific knowledge discovery infrastructure[23–25] accelerate research, improve data AI-readiness, and improve reproducibility through community engagement and data sharing[26–29].

In response to the 2013-2022 Decadal Survey, NASA released its GeneLab Open Science Strategic Plan in 2014 and has since developed NASA GeneLab, a public data repository for FAIR and Open genomics and multi-omics data for space biology. GeneLab and its associated Analysis Working Groups have resulted in tremendous knowledge gain[30–39].

In order to continue moving forward and enable broadscale analytics through data integration, all phenotypic data (e.g., biochemical, behavioral, imaging, video) and mission data (e.g., telemetry, radiation, acoustics, vibrational, space habitat, vehicle, CO2, O2, humidity) must also adhere to FAIR and Open[40] principles of operation. Further, there are hundreds of thousands of unprocessed biospecimens[34] which offer a considerable Open Science opportunity to better understand radiation, microgravity, and other
spaceflight hazards through data mining, knowledge extraction, and boosting statistical power by increasing data volume (key for AI methods).

To enable maximal scientific return, we recommend the establishment of a centralized Open Science Data Management Environment (DME) and accompanying data analysis/modeling portal designed to be open to the broader space biology research community, with investigator participation incentivized through granting structures[17]. Worldwide engagement with such a DME will accelerate Open Science for space biology and facilitate widespread data reuse and knowledge discovery[41]. The DME will support the integration of scientific, mission-based, and environmental data from Earth- and space-based experiments with the goal of accurately modeling predictions of spaceflight effects on life.

The DME will include data ingestion and semantic data retrieval tools[42], ontological mapping capabilities, digital research notebooks[43], meta-analysis tools[44], and data visualization and modeling tools. In order to decipher cause and effects and establish advanced AI models for Space Biology, quantitative[45] (e.g., -omics, physiological, biochemical, imaging) and qualitative (e.g., behavioral) data from the same missions must be integrated together with environmental/telemetry mission data and experimental factors. Although data intake (from both Earth and space experiments) will necessarily be manual at first, investment in space data specific AI/ML data management will eventually facilitate automated data ingestion, standardization, and processing. Ultimately, cloud-based compute storage and analysis resources will allow for two-way communication between the DME and research and research/clinical activities in space; so that Earth-based data can support predictive modeling during missions venturing farther into deep space. To avoid unnecessary up/downlink of large datasets, the DME system can take advantage of ML-based data compression[46], on-board processing[47], few-shot learning which requires few samples for inference[48], and federated learning (FL) which trains a global model of separate datasets by sharing only model weights instead of data[49]. FL is also used in settings where sharing a dataset is impossible due to privacy constraints, due to its compatibility with differential privacy[50,51] and homomorphic encryption[52]. Therefore FL may be useful to train models on human astronaut data, which would greatly enhance our ability to derive human-relevant knowledge from existing spaceflight data[53].

The need for our proposed DME and its functionalities are well documented in the recently submitted Life and Physical Sciences 2023-2032 Decadal Topical White Papers: investment in pivoting from natural language term indexing to graph-based data storage and linking and semantic data retrieval[54], fundamental automated support for Open Science repository capture of experimental data at multiple stages of the investigation lifecycle[55], and development of Open Science cloud-based experiment workspaces for data storage[55].

Goal 2: Algorithm Repurposing and Development

Recent advances in AI/ML methods have revolutionized analysis, prediction and diagnostic capabilities in the biomedical and clinical fields[6–11]. In parallel, AI/ML-driven methodologies have shown great potential for knowledge discovery and clinical support in the field of space biology[31,56–58]. However, irrespective of these advances, the space biology field faces particular challenges for AI/ML implementation in the near and long term. For example, small sample sizes[59] and high feature counts (particularly with regards to -omics datasets) can lead to overfitting the model, which then exhibits poor performance when deployed on unseen data. There also are often heterogeneous, sparse or missing (not at random) data, which degrades performance of both traditional and AI/ML modeling. There are also complexities in translating findings from model organisms to human astronauts.

For the space biology field to take full advantage of AI/ML approaches, we must invest in identifying, adapting, and developing context appropriate methods that perform well within the constraints of space biological and biomedical data. In order to maximally repurpose existing methods, we can learn from the
rare diseases field which faces many of the same challenges. That domain often has small patient populations, spread across many institutions, and the data are intrinsically heterogeneous and noisy[60], but have leveraged AI/ML for specific tasks[61,62]. Other examples of key ML methods involve integration with knowledge graphs which help extract relationships from unstructured data[31,63,64], and transfer learning, which creates inferences on smaller datasets by ‘transferring knowledge’ from larger, better-characterized datasets (with proven ability for model-to-human transference)[65,66].

Other existing ML methodologies generally hold great promise for space biology research, but currently are based around algorithms designed for Earth-based purposes[67]. This leaves a significant capability gap, and misses a key opportunity for designing innovative, next-generation computational infrastructure for in-flight learning, analysis, and inference on real-time data. Therefore, we recommend investment in researching current methodologies that can be adapted or repurposed for space biology data analysis[67]; and developing novel methods to meet the challenges of the research field and future automated astronaut health support systems. Methods will be continually improved based upon their performance in real-world settings.

**Goal 3: In-flight AI/ML-powered Research Facilities and Analysis Platforms**

The Biological and Physical Sciences 2023-2032 Decadal Survey call received an overwhelming submission of topical proposals for Earth- and space-based experiments needed to fill fundamental gaps in our knowledge of spaceflight effects on life systems[45,68–71]. Space-based research has certain advantages over Earth-based analog experiments, but many researchers opt to test their hypotheses in Earth-based systems due to the expense and limitations of space-flown experiments. Experiments performed in space necessarily have a low sample size, few replicates, and significant inter-experiment variability. Space-flown biological research is also slow and expensive due to the manual labor required, the limited amount of crew time available for each experiment, and the lack of longitudinal measurements on animal models.

We envision a future in which an astronaut researcher on the Lunar surface[15] can program a hypothesis into an automated experimental platform, and return days later to view the analyzed results of the original and follow-up hypotheses which were automatically generated and tested by learning from the original experiment. The experimental data will have been processed and automatically linked to the cloud-based space biology DME, along with experimental metadata and model weights and biases. At the same time, an Earth-based researcher can test the same factors in the Martian atmosphere via a telemetry-based autonomous satellite payload. Citizen scientists can analyze the data from both experiments and test new hypotheses via the public data analysis portal.

This new scientific vision would require improvements in several areas (i.e., robotics, hardware, compute, communications). Here, we focus on work needed to achieve in-flight, automated, self-driving experimental systems ("self-driving labs") (Figure 3). The experimental workflow must be woven together with closed-loop AI systems for learning and hypothesis generation. Many components necessary to achieve such independence already exist on Earth. For instance, many biological processes have been automated through technology and microfluidics advances[30,72–79]; several versions of automated closed-loop discovery systems have been developed[80–87]; and the ML field of lifelong learning has enabled neural networks to constantly learn from new information[88–91]. However, a closed-loop spaceflight-hardened research platform that requires limited human intervention is
not yet reality. We therefore recommend dedicated research and development in automated in-flight laboratory systems to allow space biology studies to continue when on the Moon, Mars or uncrewed spacecraft. Longitudinal, broad-spectrum measurements of living systems beyond LEO will be essential to enable accurate prediction of spaceflight effects on life across diverse extraterrestrial environments. One example of an ML-based predictor is a "digital twin", which can model biological functionality via mathematical modeling[92–94], and has been successfully developed on the microbial and single-organ levels[95–98]. A real-time research data stream from diverse space environments would aid development of digital twins for research organisms and ultimately humans, to be able to predict outcomes under various conditions.

Currently, conducting biological experiments in space requires significant human input, just as the current astronaut health support system, developed for LEO missions, depends heavily on mission control center real-time communication with crew[16]. However, as humanity ventures further into deep space, a paradigm shift is needed toward AI-powered, self-driving experimentation, as well as crew-centered health decision making with AI support (see Goal 4).

**Goal 4: AI/ML for Health Monitoring and Precision Space Health**

Ultimately, space biological research aims to better characterize spaceflight risks to human health and inform risk mitigation and countermeasure strategies. On future deep space missions, crew-centric medical decisions will be best empowered by a precision health system informed by automated biomedical and environmental monitoring[16,99]. Precision health approaches provide genetic, epigenetic and physiological baseline measurements for individual patients or crew members, with the ability to detect potentially pathogenic deviations from those baselines and predict outcomes, to allow for preventative action rather than reactive treatment[100–107].

AI/ML technology is well suited to support such a system's automated assessment and prediction needs. While in varying stages of maturity and clinical adoption, Earth-based precision medicine approaches have demonstrated efficient use of AI/ML techniques for pathology detection and outcome prediction[108–119]. We recommend developing an AI/ML powered multi-layered approach for monitoring individual astronauts as well as spacecraft and habitable environments. The approach would have three layers[17]: the first layer would be non-invasive, continuous, sensor-based environmental monitoring of physical, chemical and biological components of the spacecraft or habitat environment. AI models should be developed and employed to integrate environmental data streams, extrapolate risk prediction, anomaly/fault detection, and provide visual readouts for astronauts. The second layer would involve non-invasive collection of physiological measurements from wearables and point-of-care devices. The third layer would include the collection of molecular and genetic biomarkers and other measurements, from methods such as blood draws, saliva sampling, or smart toilets[120].

The data streams from these three layers of monitoring would then be woven together into an AI-driven, proactive Precision Space Health system, ensuring that astronaut health is predictive, preventative, participatory, and personalized[121]. Integrating such an approach with the “digital twin” technology will also be essential here. The health system would also draw on predictive modeling from Earth- and space-based research experiments, captured in the cloud-based DME. Therefore, we recommend a structured research program into the development of an AI-powered Precision Space Health system with the capabilities to assess and aggregate multiparametric, disparate data; to analyze those data and provide actionable health decision support for the crew medical officer; to adapt to changes in the environment and to baseline changes in crew health through a dynamic learning system.
References


21. Gilmore RO, Adolph KE. Video can make behavioural science more reproducible. Nat Hum Behav. 2017;1. https://doi.org/10.1038/s41562-017-0128


https://doi.org/10.1093/bioinformatics/bty884


