Biological Research and Space Health Enabled by Machine Learning to Support Deep Space Missions

Ryan T. Scott

Lauren M. Sanders, Ph.D.

ASGSR2023, Nov 18, 2023

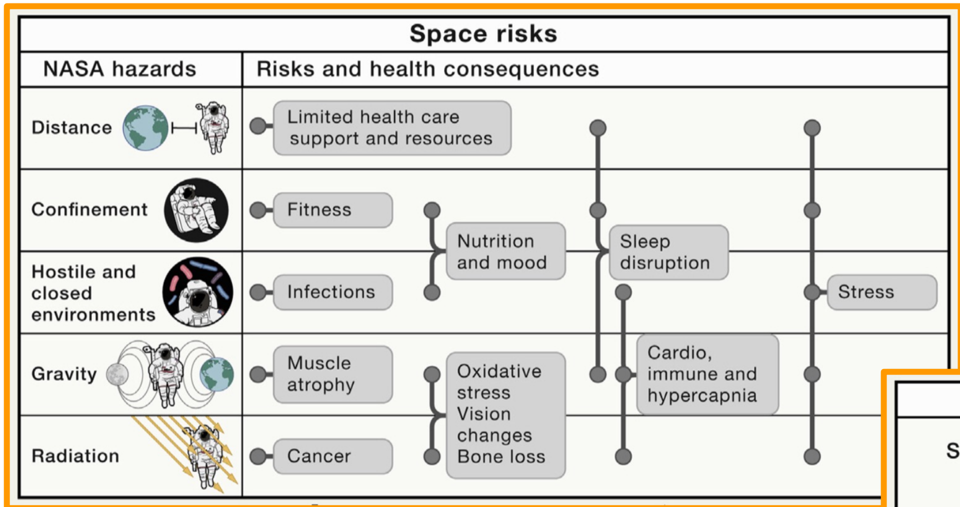
NASA Ames Research Center

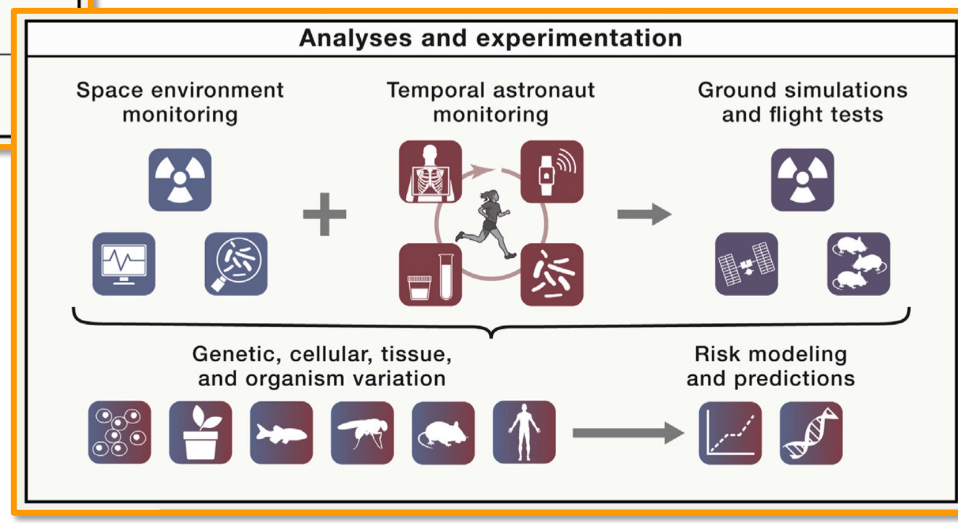
KBR & Blue Marble Space Institute of Science

National Aeronautics and Space Administration

NASA

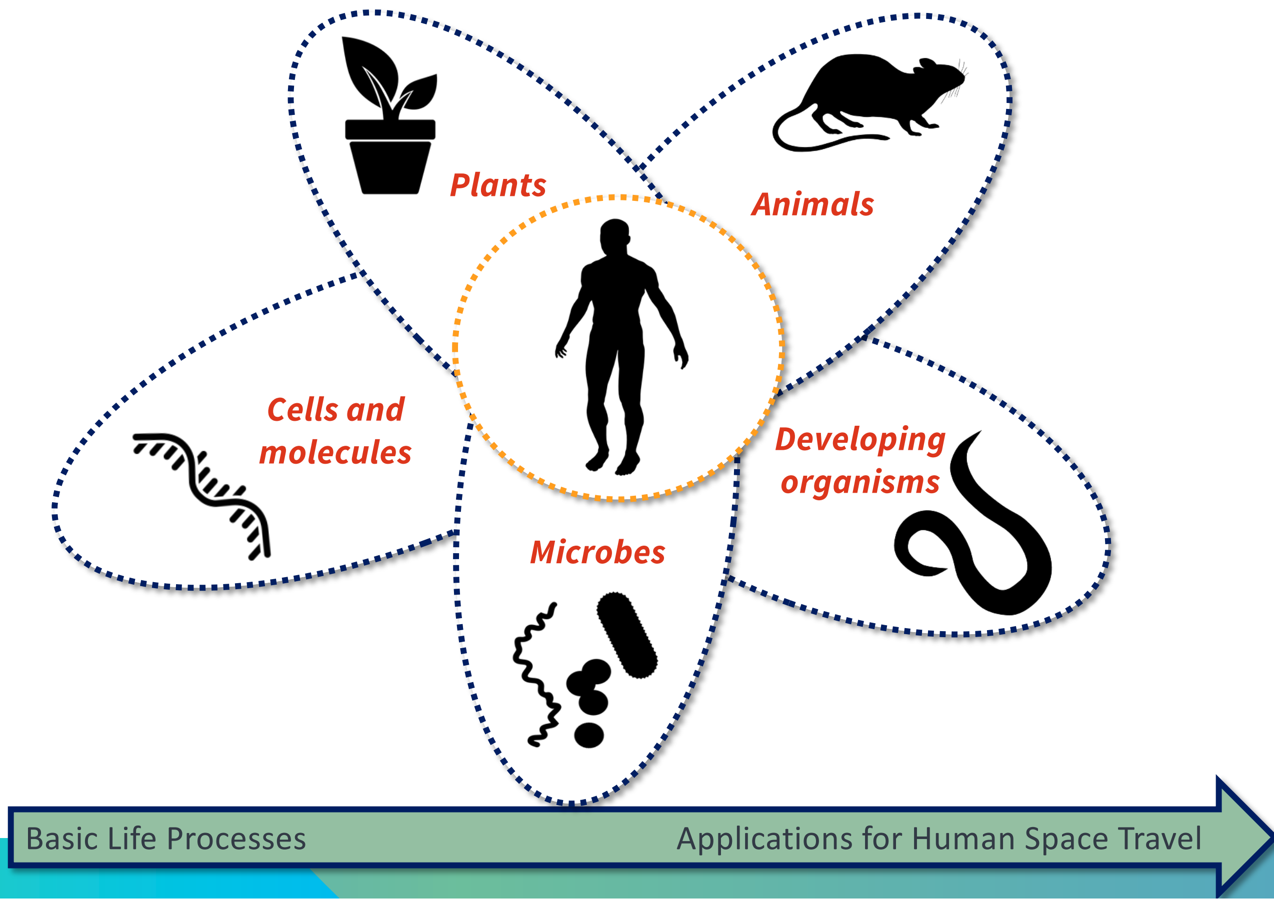
Space Travel Introduces Risks for Living Systems





Space Biology Research

Characterizing the response of living systems to spaceflight



433
Studies

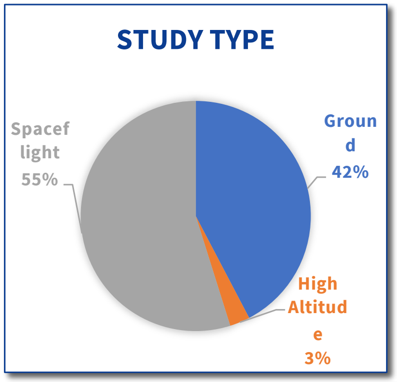
819
Datasets

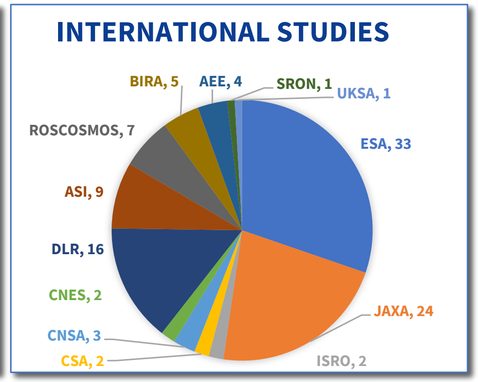
45
Species

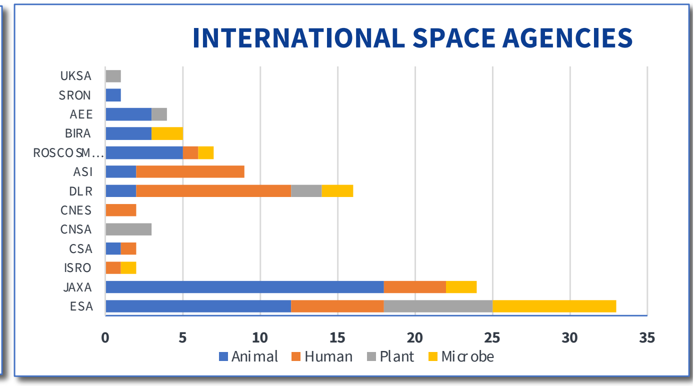
>25
Assays

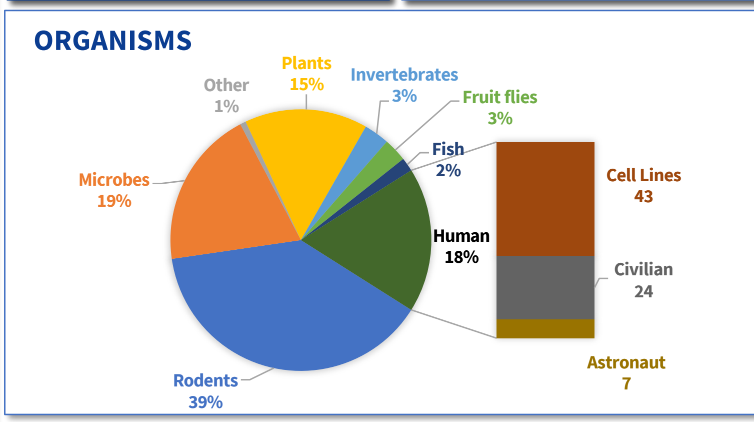
>150TB
Data

NASA Open Science Data Repository









Civilian and Astronaut	Bed Rest, Spaceflight, Mars simulation
Cell Lines	Radiation (Ground), Simulated uG, Spaceflight, Parabolic Flight

- **Spaceflight and space-relevant** biological datasets
- **FAIR data:** Findable, Accessible, Interoperable, Reusable

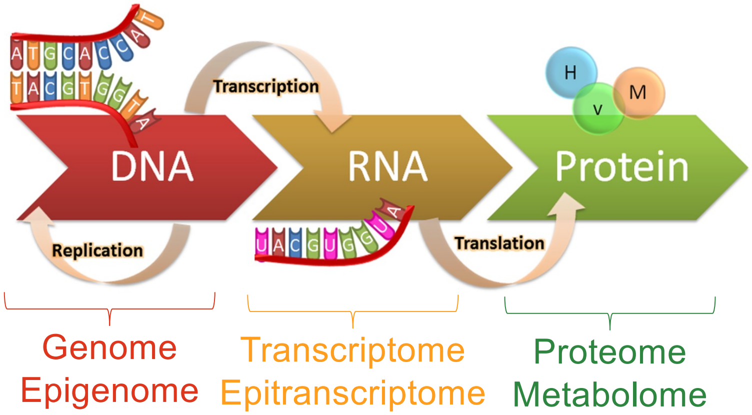
osdr.nasa.gov/bio

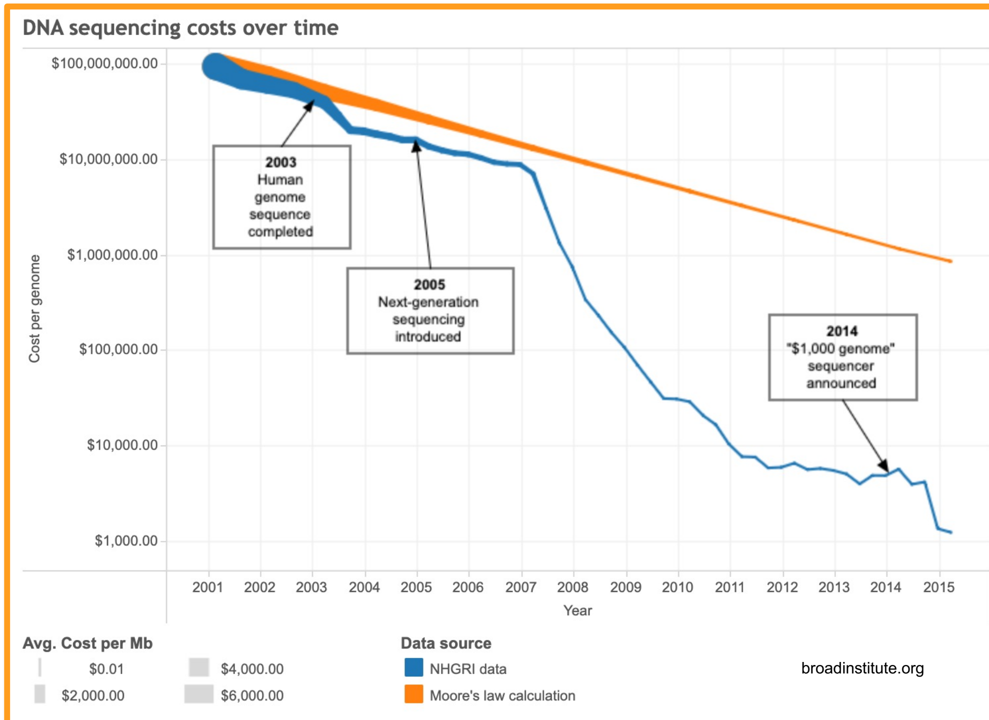
*Data originates from Gene Expression Omnibus (GEO) and EMBL's European Bioinformatics Institute (EMBL-EBI) repositories.

The Challenges of High **Complexity**, High **Dimensionality**, Low **Sample Size** Data

Paradigm Shift in Biological Data Generation

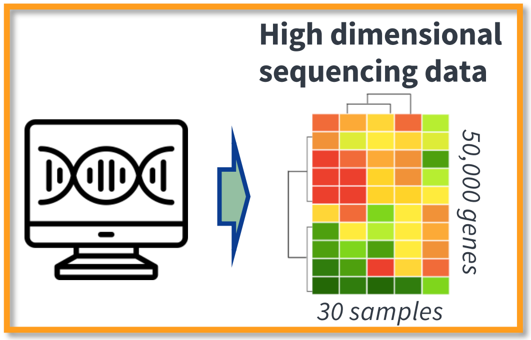
- **Traditional molecular biology** studies a few genes or proteins at a time through
- **High-throughput sequencing ('omics)** gives a readout of the entire genome in a cell or tissue sample





Decline in real costs compared to expected declines based on Moore's Law.
Trend line: Cost per human genome. Line width: Cost per megabase (Mb)
(*Data: NHGRI https://www.genome.gov/27541954/dna-sequencing-costs-data/*)

Space biological data analysis challenges



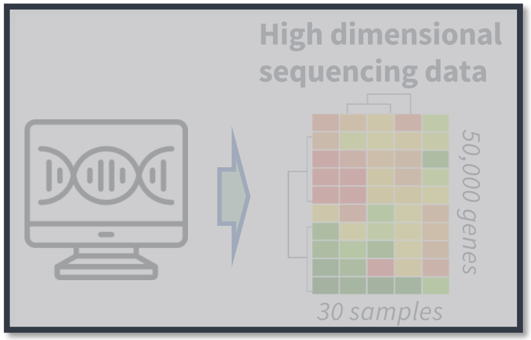
Space Biological Data Challenges

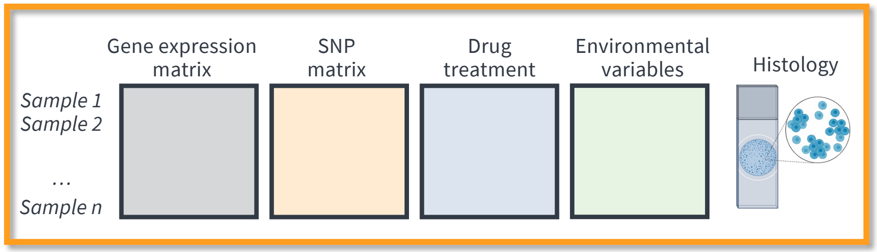
- **Small sample *n***
- **High feature count**
- Heterogeneous data
- Sparse data
- Transfer from model to human

Space biological data analysis challenges

Space Biological Data Challenges

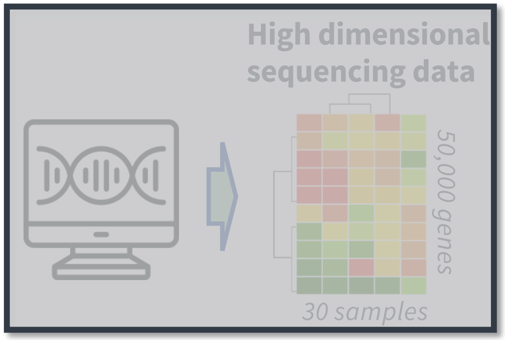
- Small sample *n*
- High feature count
- **Heterogeneous data**
- **Sparse data**
- Transfer from model to human

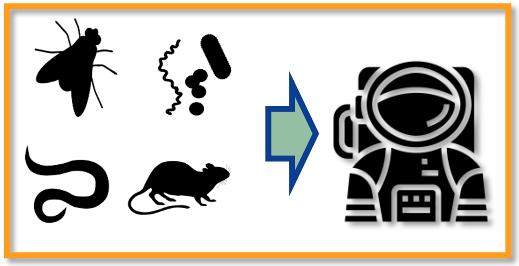




Space biological data analysis challenges

Space Biological Data Challenges
- Small sample *n*
- High feature count
- Heterogeneous data
- Sparse data
- **Transfer from model to human**





Complex (Biological) Systems need Complex Models

Multiple approaches for characterizing patterns in biology

STATISTICAL METHODS

- Draw conclusions from observed data (**inference**)
- Assume specific data distributions
- Examples: hypothesis testing, correlative analysis

MACHINE LEARNING

- Learn from data to make predictions on unseen data (**prediction**)
- Able to model nonlinear relationships without assuming a data distribution
- Examples: classification, regression, clustering

Bzdok et al., *Nature Methods* 2018

ML Learns and Predicts Complex Biological Phenomena

> J Thorac Imaging. 2020 Nov 1;35(6):361-368. doi: 10.1097/RTI.0000000000000544.

A Novel Machine Learning-derived Radiomic Signature of the Whole Lung Differentiates Stable From Progressive COVID-19 Infection: A Retrospective Cohort Study

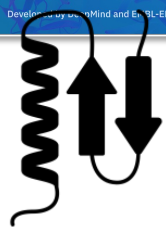
Liping Fu [1], Yongchou Li [2], Aip

> Sci Data. 2021 Apr 29;8(1):121. doi: 10.1038/s41597-021-00900-3.

COVID-CT-MD, COVID-19 computed tomography scan dataset applicable in machine learning and deep learning

Parnian Afshar [1], Shahin Heidarian [2], Nastaran Enshaei [1], Farnoosh Naderkhani [1], Moezedin Javad Rafiee [3], Anastasia, Konstantinos N Plataniotis [7], Arash

AlphaFold Protein Structure Database



> PLoS One. 2013 Apr 30;8(4):e61318. doi: 10.1371/journal.pone.0061318. Print 2013.

Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties

Michael P Menden [1], Francesco Iorio, Mathew Garnett, Ultan McDermott, Cyril H Benes, Pedro J Ballester, Julio Sae

6th International Conference on Smart Computing and Communications, ICSCC 2017, 7-8 December 2017, Kurukshetra, India

Lung Cancer Detection using CT Scan Images

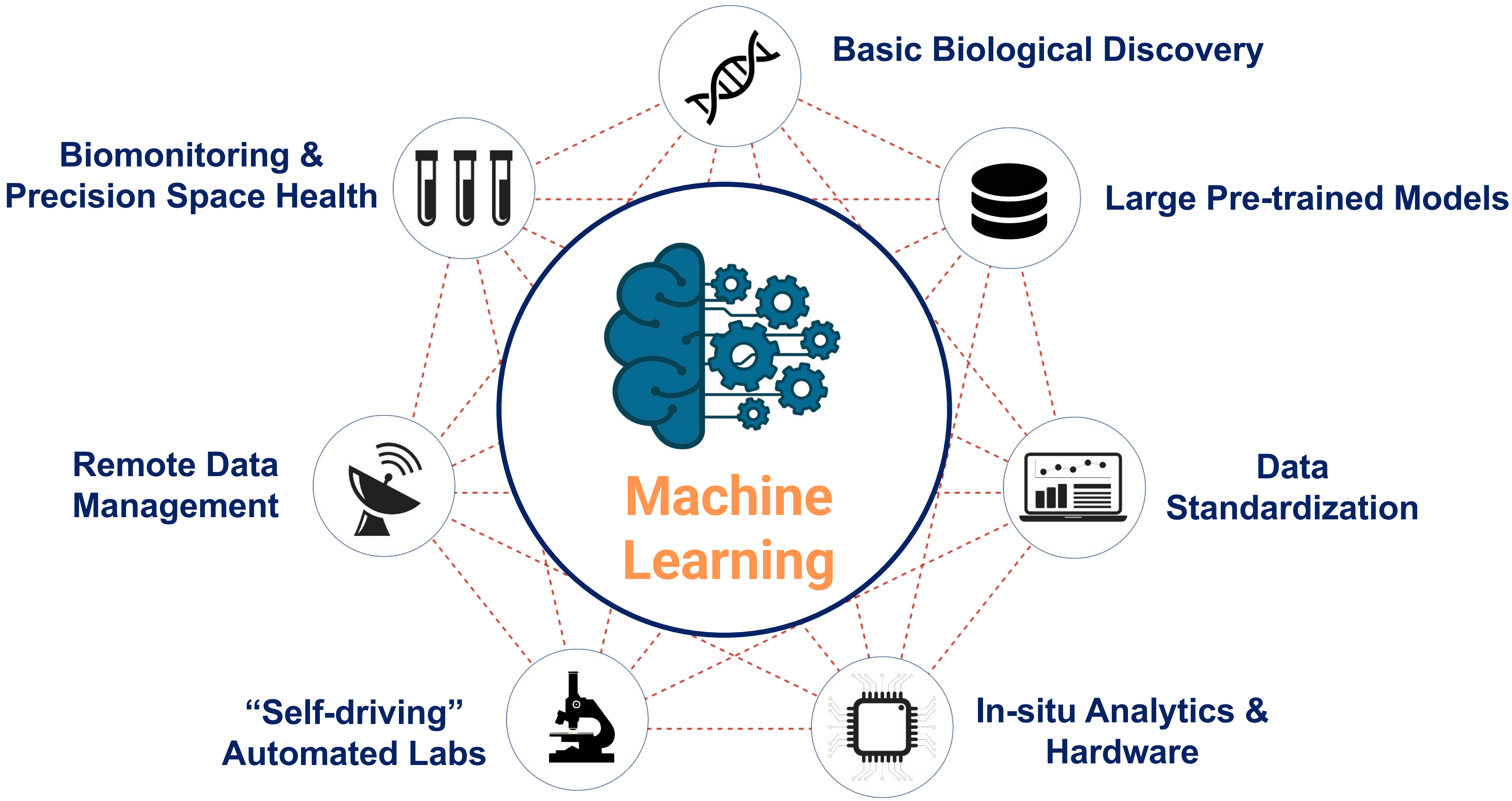
Suren Makaju[a], P.W.C. Prasad

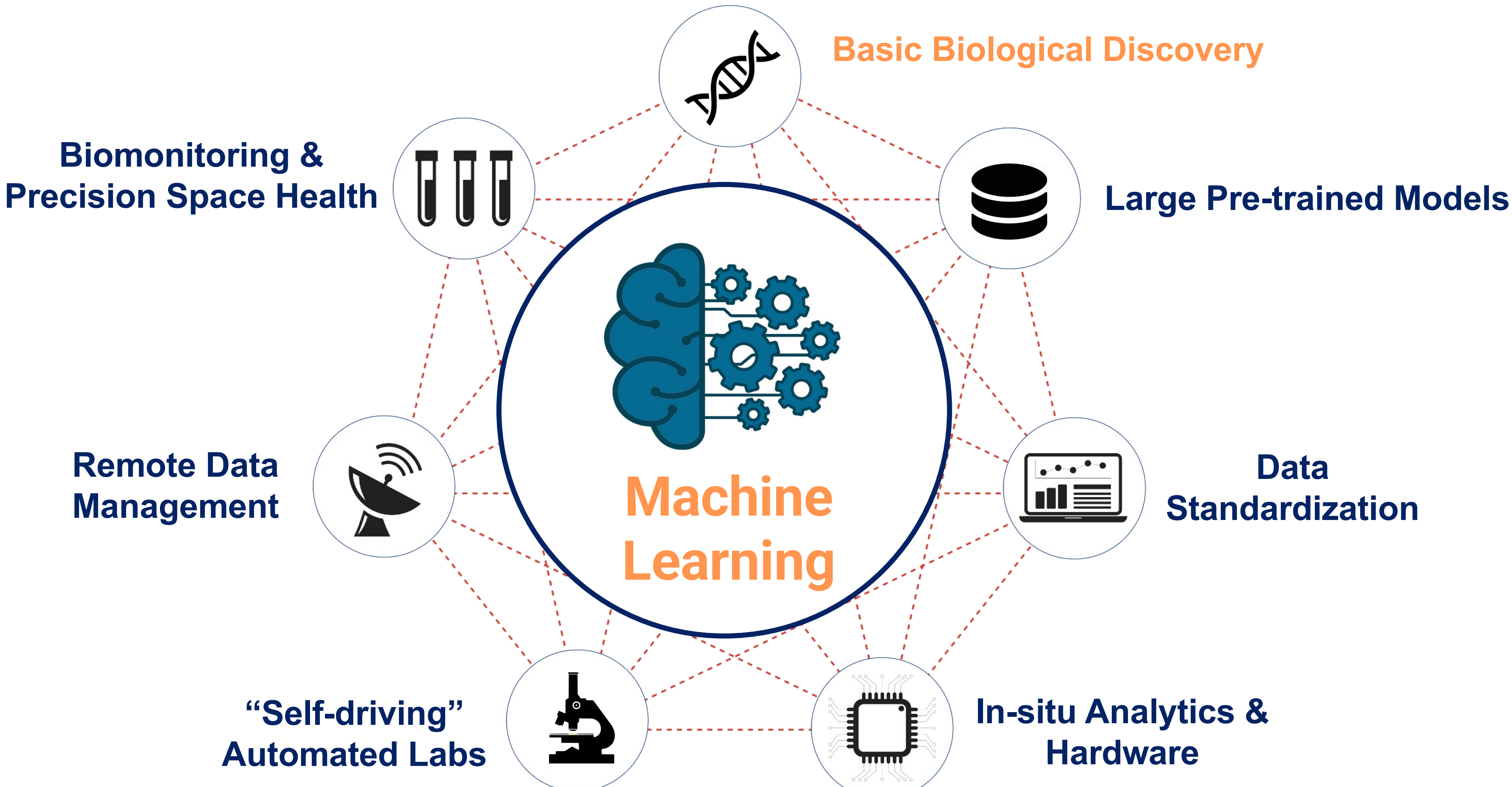
Review > Iran J Public Health. 2017 Feb;46(2):165-172.

Improving the Prediction of Survival in Cancer Patients by Using Machine Learning Techniques: Experience of Gene Expression Data: A Narrative Review

Azadeh Bashiri [1], Marjan Ghazisaeedi [1], Reza Safdari [1], Leila Shahmoradi [1], Hamide Ehtesham [1]

The opportunity: adapt ML principles to power knowledge discovery and address key challenges in space biological research

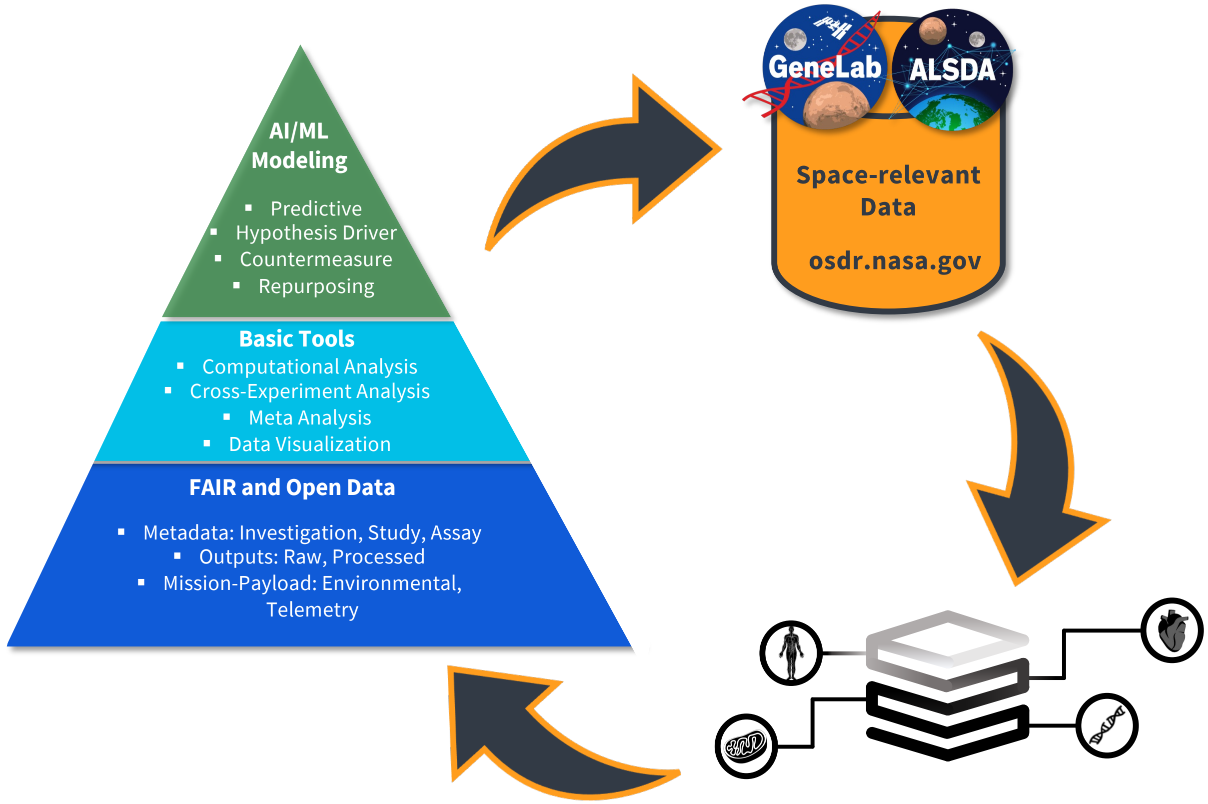

Basic Biological Discovery
Biomonitoring &
Precision Space Health
Large Pre-trained Models
Machine
Learning
Remote Data
Management
Data
Standardization
“Self-driving”
Automated Labs
In-situ Analytics &
Hardware



Basic Biological Discovery
Biomonitoring & Precision Space Health
Large Pre-trained Models
Machine Learning
Remote Data Management
Data Standardization
"Self-driving" Automated Labs
In-situ Analytics & Hardware

AI for Life in Space working group: AI4LS

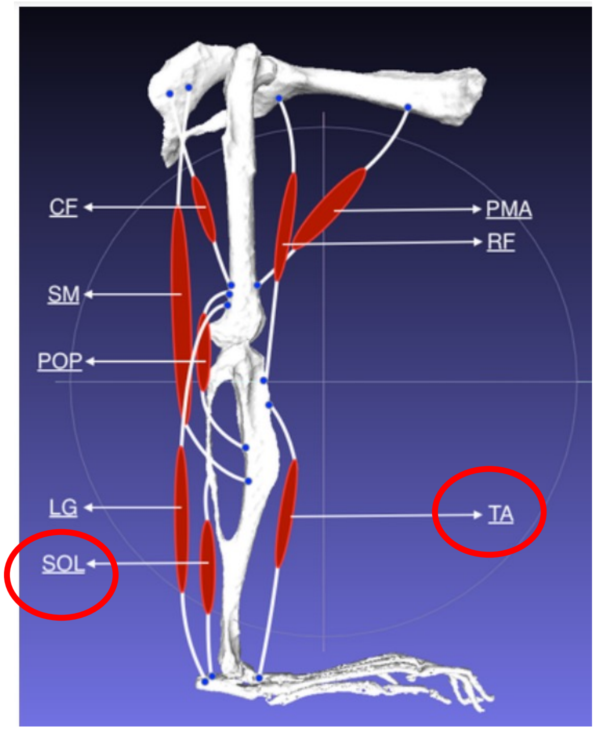
Leveraging ML and AI methods to model space biology data from the NASA Open Science Data Repository: **NASA GeneLab** (omics) and **NASA Ames Life Sciences Data Archive** (ALSDA; phen-omics) to better understand the complex effects of spaceflight on living systems across hierarchical biological levels.



Explainable ML to Interrogate the Molecular Underpinnings of **Spaceflight Muscle Atrophy**

Spaceflight Changes Muscles at the Cellular Level

SOLEUS MUSCLE
- "slow-twitch" muscle



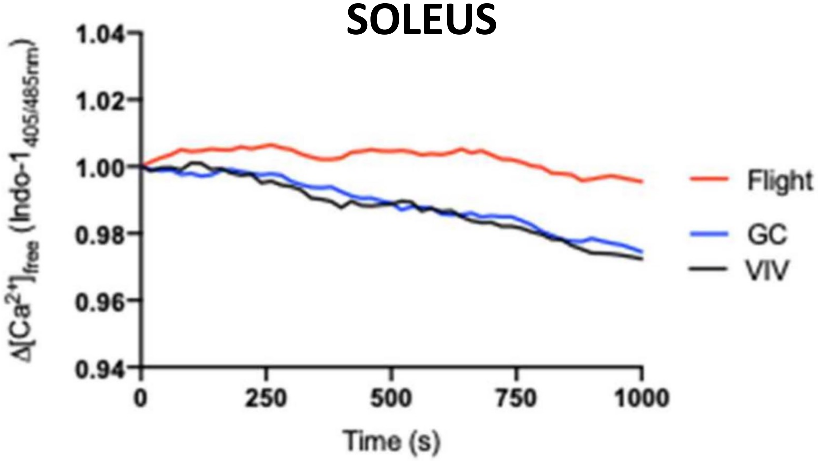
TIBIALIS ANTERIOR MUSCLE
- "fast-twitch" muscle

Jensen et al. 2020

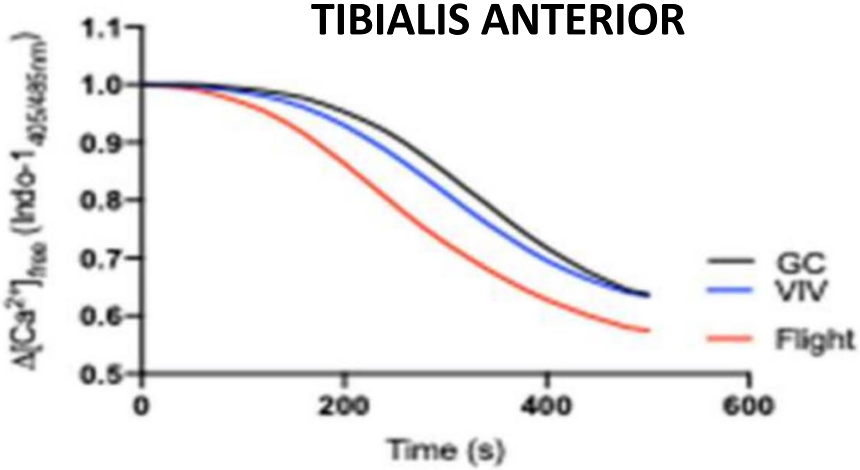
Spaceflight Changes Muscles at the Cellular Level

*Muscle cells take in **calcium** for normal contraction*

SOLEUS



TIBIALIS ANTERIOR



In spaceflown mice,

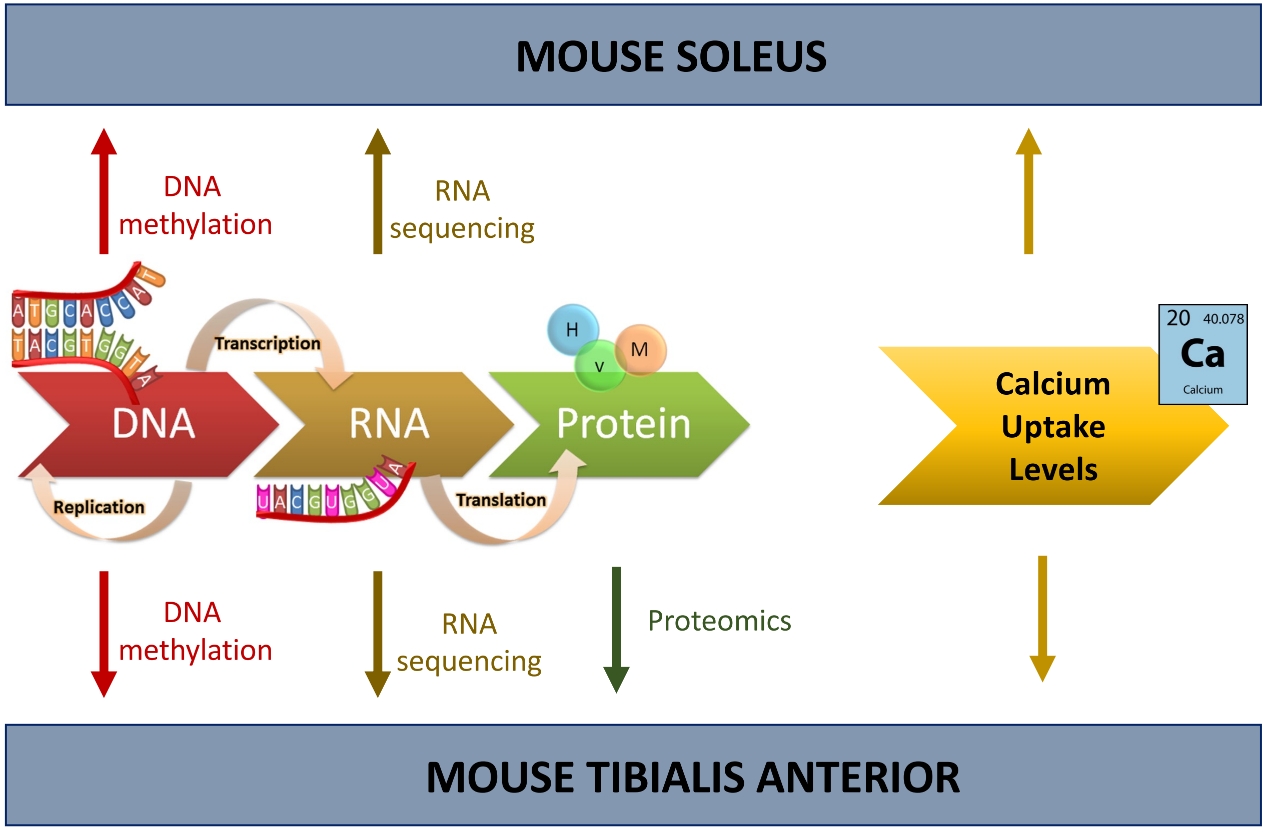
calcium uptake efficiency *decreases* in **soleus** muscle...

but *increases* in **tibialis anterior** muscle!

GOAL

- Train a machine learning model to learn the relationship between calcium reuptake levels and molecular changes within the cell
- Interrogate the model to identify molecular predictors (biomarkers) of calcium reuptake changes in spaceflight

OSDR Datasets:
OSD-104
OSD-105
OSD-488



Data Pipeline and Model Training

- Data preprocessing
- Multi modal data integration
- Dimensionality reduction and feature selection

→

Train ML model to predict calcium levels using molecular data

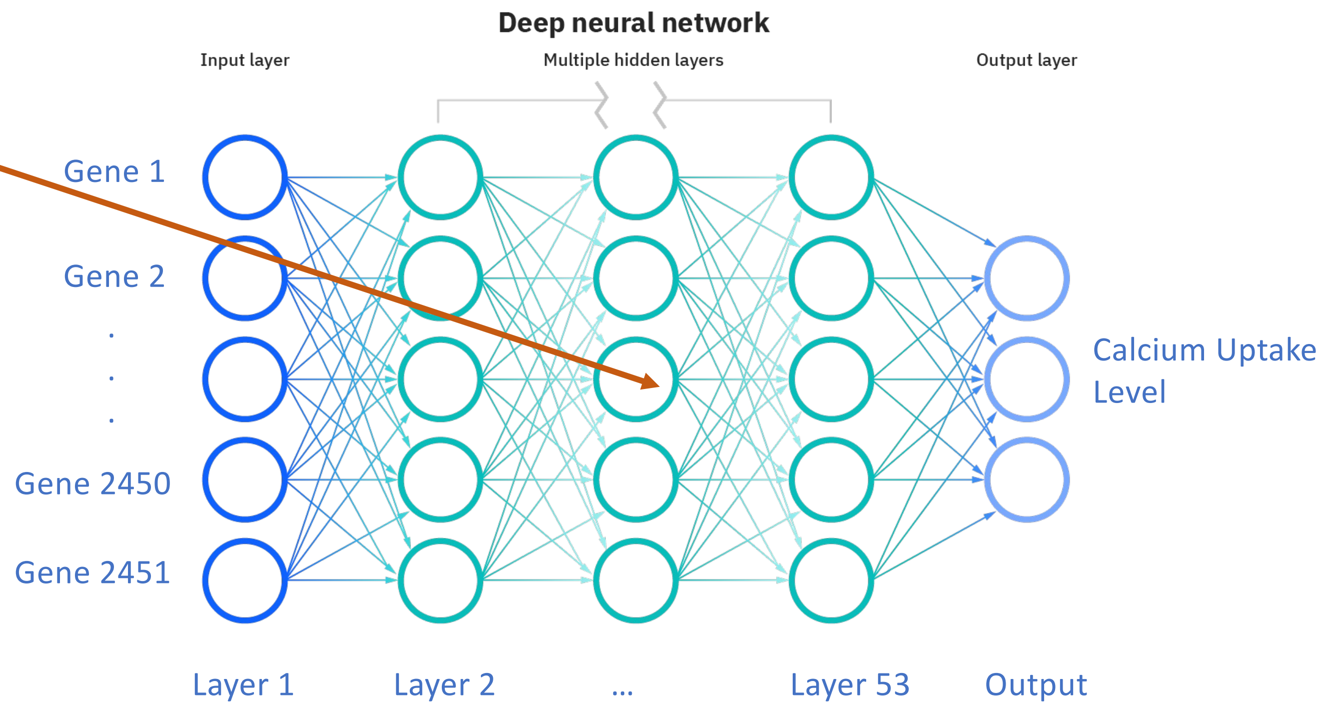
→

Identify predictive features (biomarkers)

- Leave-one-out-cross-validation (LOOCV) due to low sample sizes (n=12)
- Model performance calculated as an average across LOOCV runs
- Top features across all LOOCV runs evaluated as biomarkers

Explainable ML for Biomedical Research

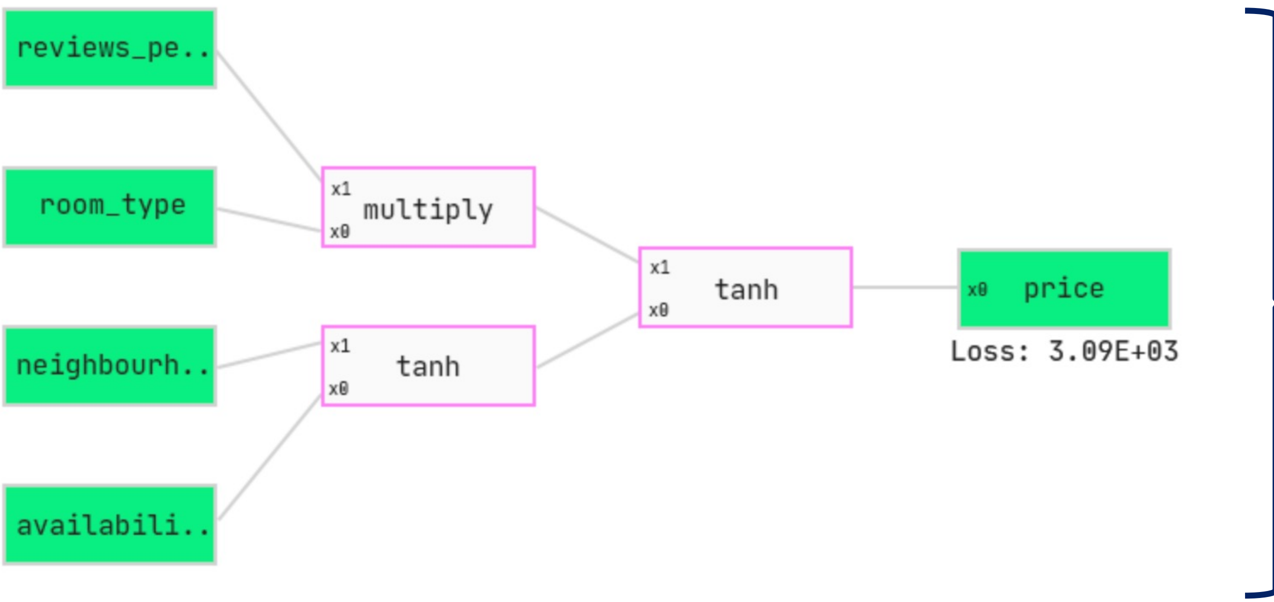
What is the biological interpretation of this intermediate value in the 32nd layer?



21

QLattice Symbolic Regression Machine Learning Algorithm

Interpretable computational graphs represent mathematical relationships



A computational graph predicting short term rental prices from data

Novel Biomarkers for Calcium Uptake Changes in Muscle

b

Feature	Models (n)
Acyp1 (proteomics)	89
Rps7 (proteomics)	27
Cct6a (proteomics)	5
Glt28d2 (RNA-seq)	4

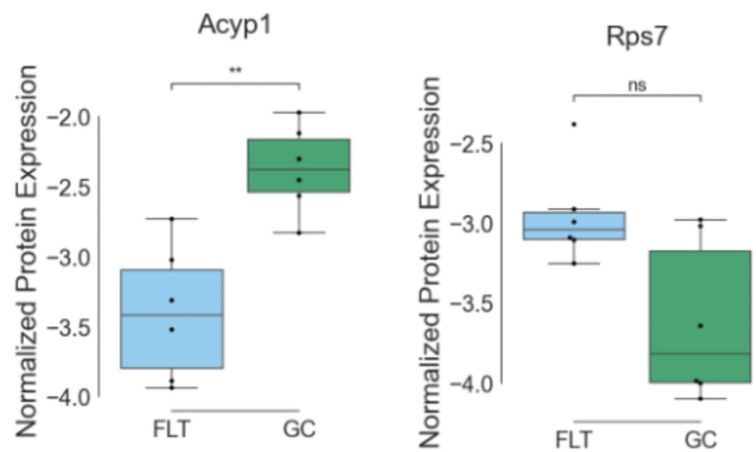
c

T1 CV R^2	0.894
T10 CV R^2	0.711
RNA-seq features (n)	12
Proteomic features (n)	38

- Top Biomarkers for Calcium Uptake in Tibialis Anterior Muscle:
  1. **Acyp1**
     - <u>Inhibits</u> calcium transport in *fast-twitch* muscles (tibialis)
     - <u>Enhances</u> calcium transport in *slow-twitch* muscles (soleus)
  2. **Rps7**
     - Downregulated by *nitrosative stress* which decreases calcium transport

TIBIALIS ANTERIOR

Increased Calcium Transport Efficiency in Spaceflight



Decreased **Acyp1** in flight allows increased calcium transport…

…and increased **Rps7** in flight shows low nitrosative stress.

Novel Biomarkers for Calcium Uptake Changes in Muscle

b

Feature	Models (n)
Acyp1 (proteomics)	89
Rps7 (proteomics)	27
Cct6a (proteomics)	5
Glt28d2 (RNA-seq)	4

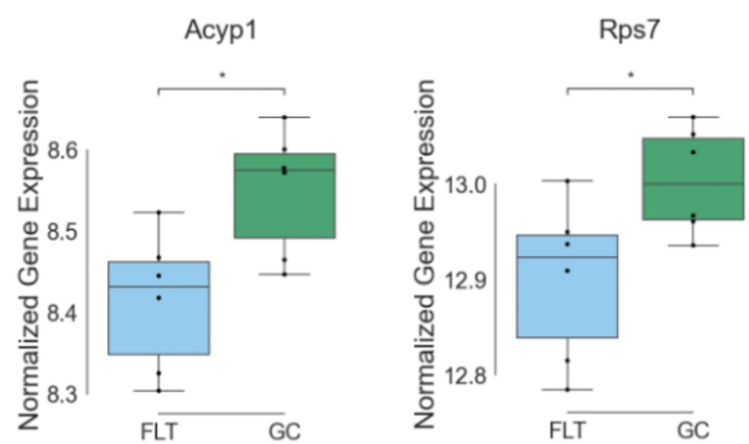
c

T1 CV R^2	0.894
T10 CV R^2	0.711
RNA-seq features (n)	12
Proteomic features (n)	38

- Top Biomarkers for Calcium Uptake in Tibialis Anterior Muscle:
  1. **Acyp1**
     - Inhibits calcium transport in *fast-twitch* muscles (tibialis)
     - Enhances calcium transport in *slow-twitch* muscles (soleus)
  2. **Rps7**
     - Downregulated by nitrosative stress which decreases calcium transport

SOLEUS

Decreased Calcium Transport Efficiency in Spaceflight

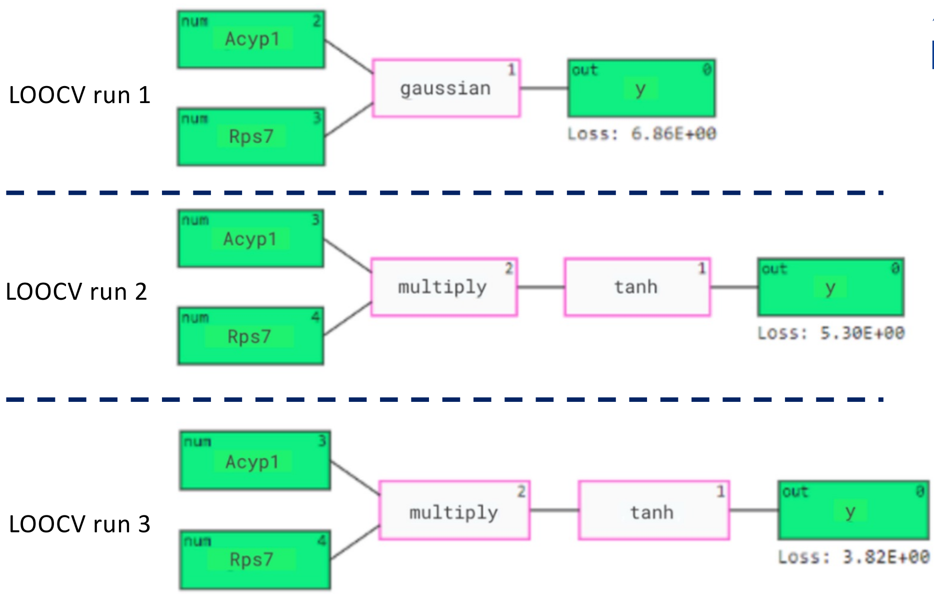


Decreased **Acyp1** in flight means decreased calcium transport...

...and decreased **Rps7** in flight shows high nitrosative stress.

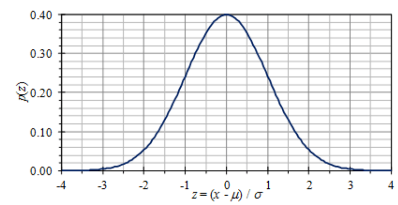
Symbolic Regression Identifies Biomarker Relationships

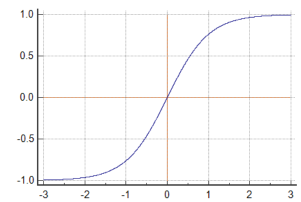
Biological features operate in interconnected networks



➢ QLattice identifies mathematical relationships between **Acyp1** and **Rps7**

- Gaussian
- Multiply, then tanh



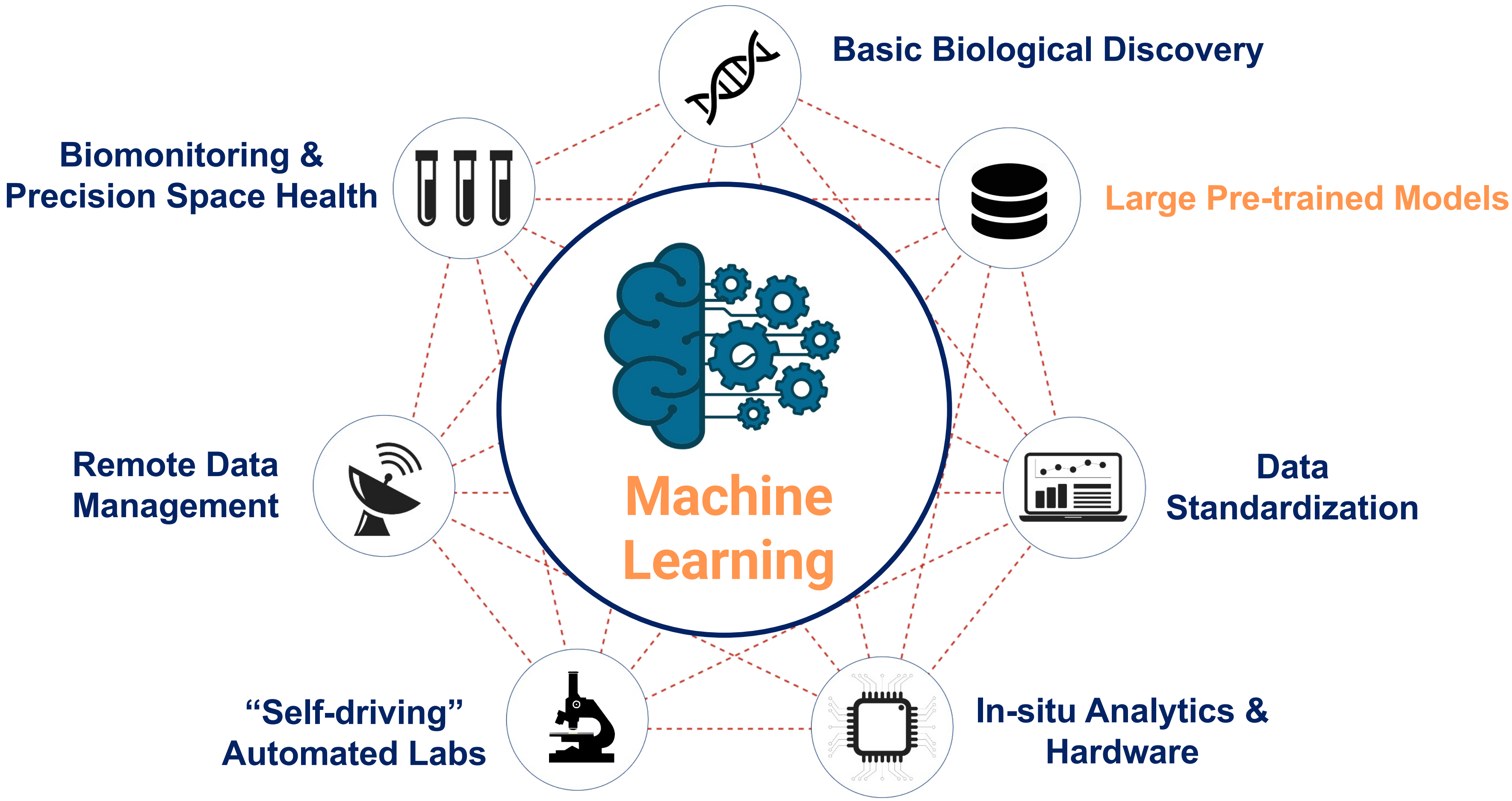


Summary of Outcomes

Explainable ML** to Interrogate the Molecular Underpinnings of **Spaceflight Muscle Atrophy

- Explainable ML methods can provide insight to complex biological relationships

- Explainable ML analysis of multi-modal biological datasets from the NASA Open Science Data Repository resulted in:
  - Novel biomarkers
  - Mathematical relationships between biomarkers
  - Consistency with previous biological knowledge
  - Starting point for new investigations

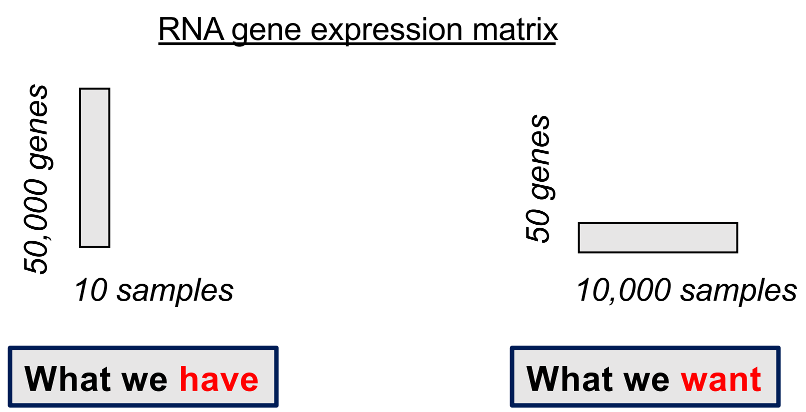

Basic Biological Discovery
Biomonitoring &
Precision Space Health
Large Pre-trained Models
Machine
Learning
Remote Data
Management
Data
Standardization
"Self-driving"
Automated Labs
In-situ Analytics &
Hardware

Large Pre-trained Models to Connect Biomedical Knowledgebases with **Small Spaceflight Datasets**

Biological Data and the Curse of HDLSS

- High-throughput sequencing provides a readout of the molecular makeup of cells
  - Genome sequencing (3 billion nucleotides: A,C,G,T)
  - Gene expression sequencing (text readout converted to numerical values: ~50,000 genes)

- This leads to high dimensionality and low sample size (HDLSS)

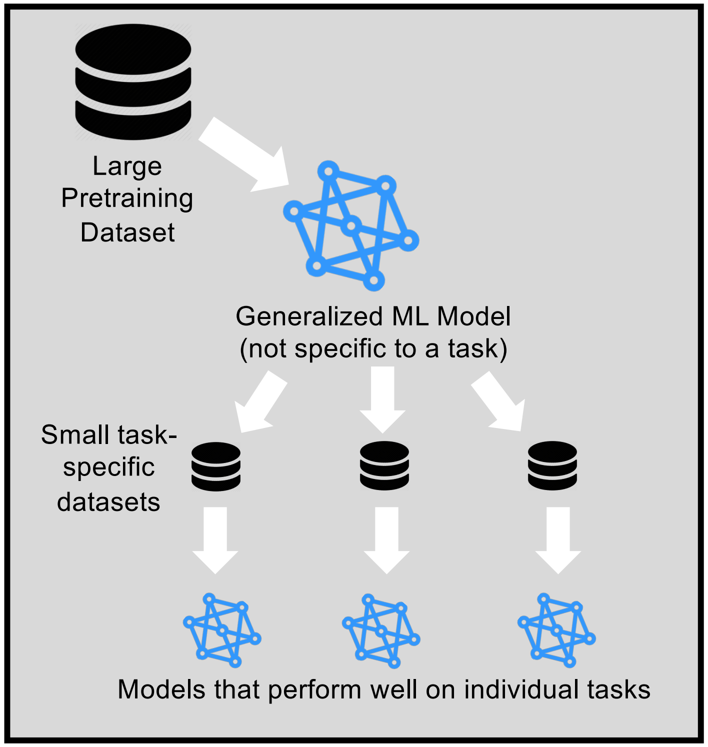


Transfer Learning: Pretrained Models

- General knowledge about a particular domain is useful for any specific task within that domain

<u>Strategy:</u>

- **Pre-train** a model on a large training dataset in the desired domain
- For other tasks ("downstream tasks") in the domain, start with the pre-trained model and **fine-tune**
  - General knowledge carries over and does not need to be re-learned
  - Fine tuning requires fewer samples than training from scratch



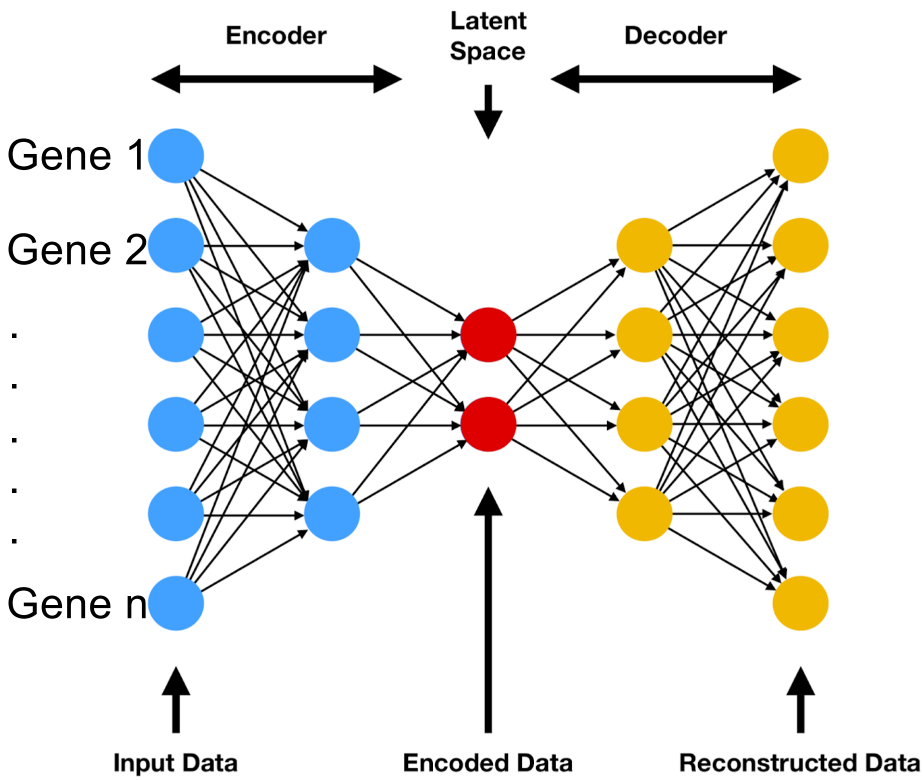
Pretrained Model for Gene-Gene Interaction Networks

- **Overview:**
  - Pre-train a model on a huge gene activity dataset to learn the relationships between all genes in human (or mouse) cells in general
  - Fine-tune for specific tasks on smaller datasets
    - Example: identify gene-gene network dysregulation in spaceflight compared to ground control samples

- **Pretraining dataset: *recount3***
  - 750,000+ publicly available, uniformly processed human and mouse RNA sequencing samples
  - Captures a huge amount of biological complexity and variability

recount3

Gene-Gene Interaction Model

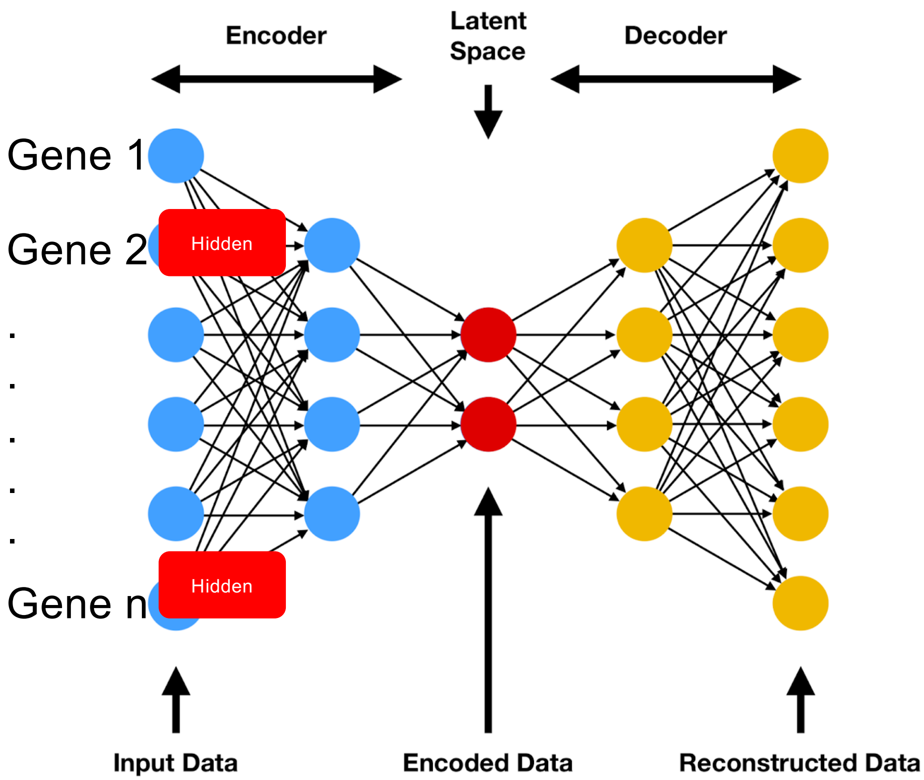


(Flores 2019)

- **Deep learning model architecture:** scBERT: encoder-decoder
- Self-supervised pre-training on massive amounts of unlabeled data

Yang et al., *Nat Mach Intel* 2022

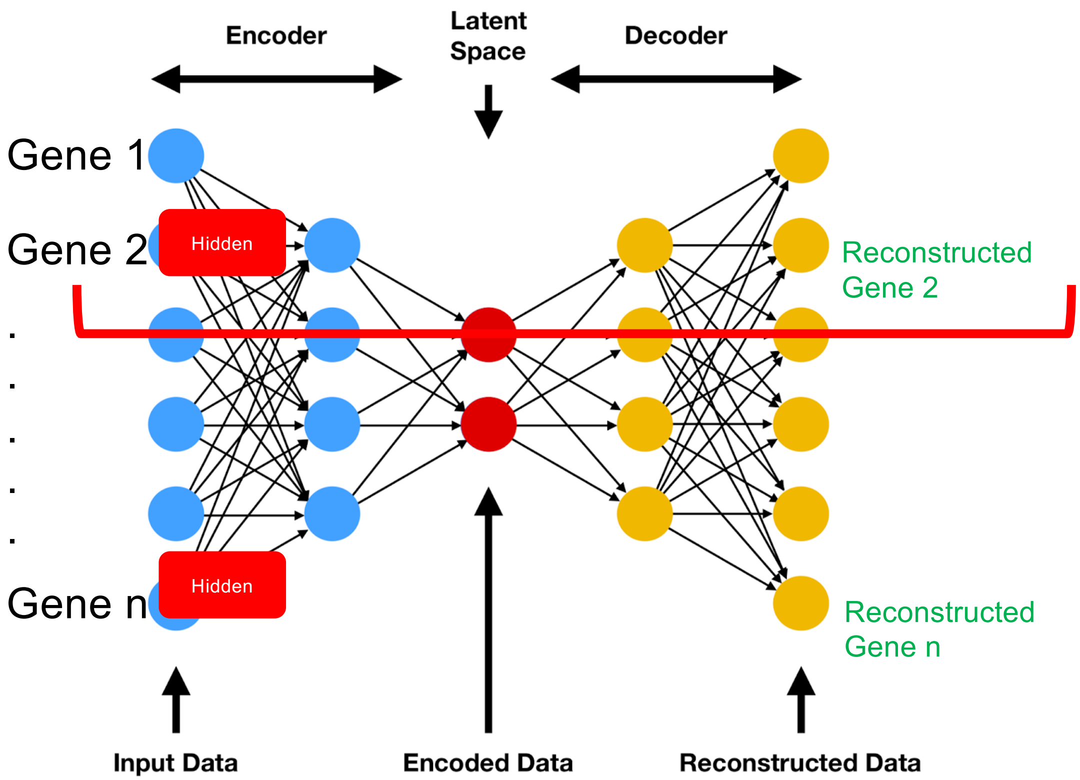
Masking Values for Training



- Randomly mask (hide) expression values of some input genes

Yang et al., *Nat Mach Intel* 2022

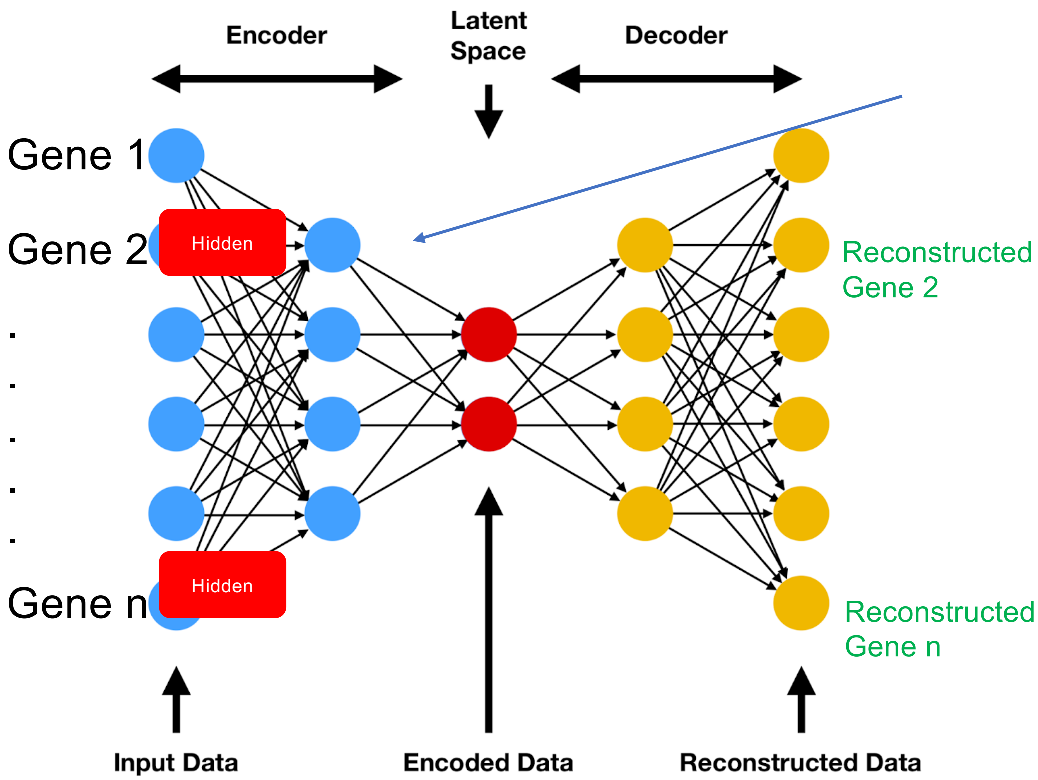
Minimize Error



- **Context Learning:** Train model to use the values of other, non-masked gene expression values to reconstruct the masked values
- **Minimize the difference** between the reconstructed and original (hidden) expression values

Yang et al., *Nat Mach Intel* 2022

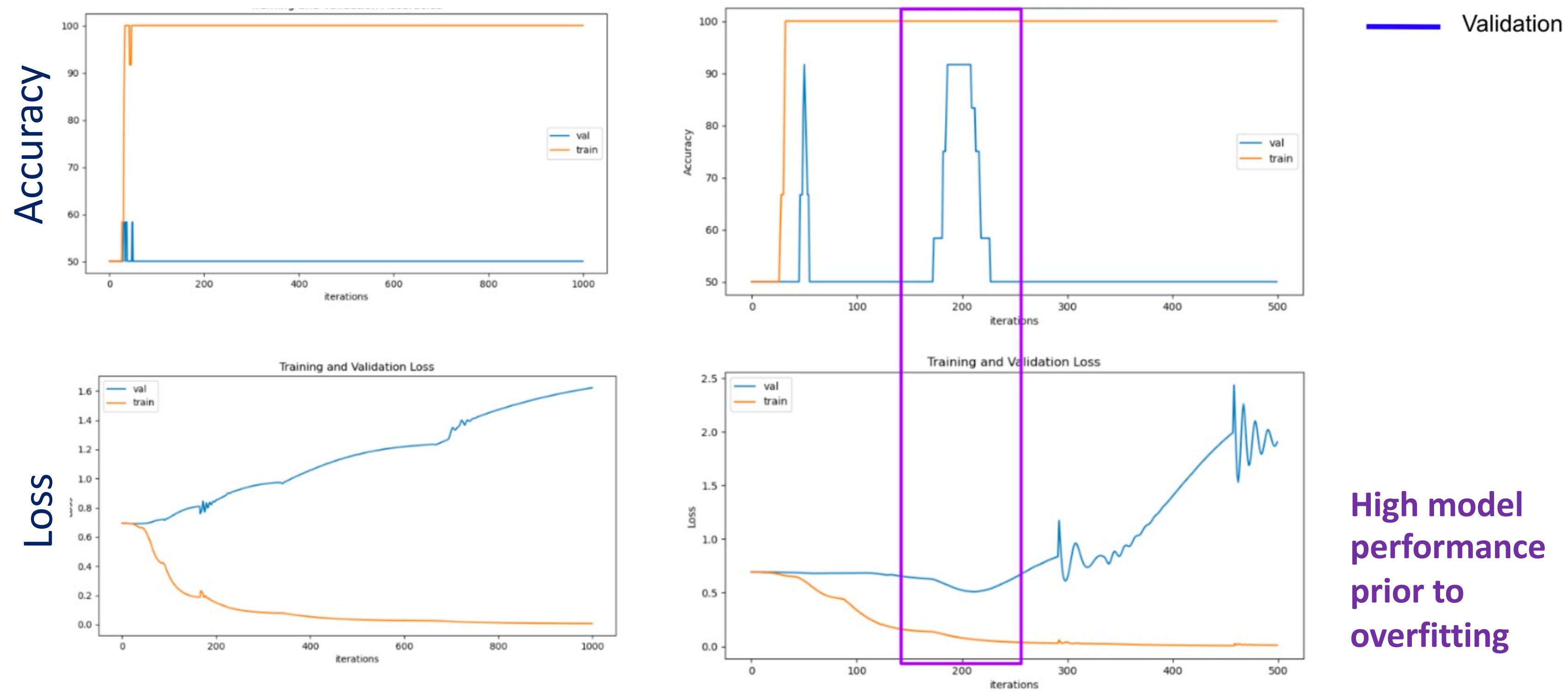
Learn General Gene-Gene Knowledge



- In the process, **general knowledge** about gene-gene interactions is learned and stored in the encoder weights
- The pre-trained weights are a good starting point for gene-gene interaction tasks in general, and they can then be **fine-tuned** to specific downstream tasks

Transfer learning outperforms traditional training

Traditionally Trained Model | **Transfer Learning Model**

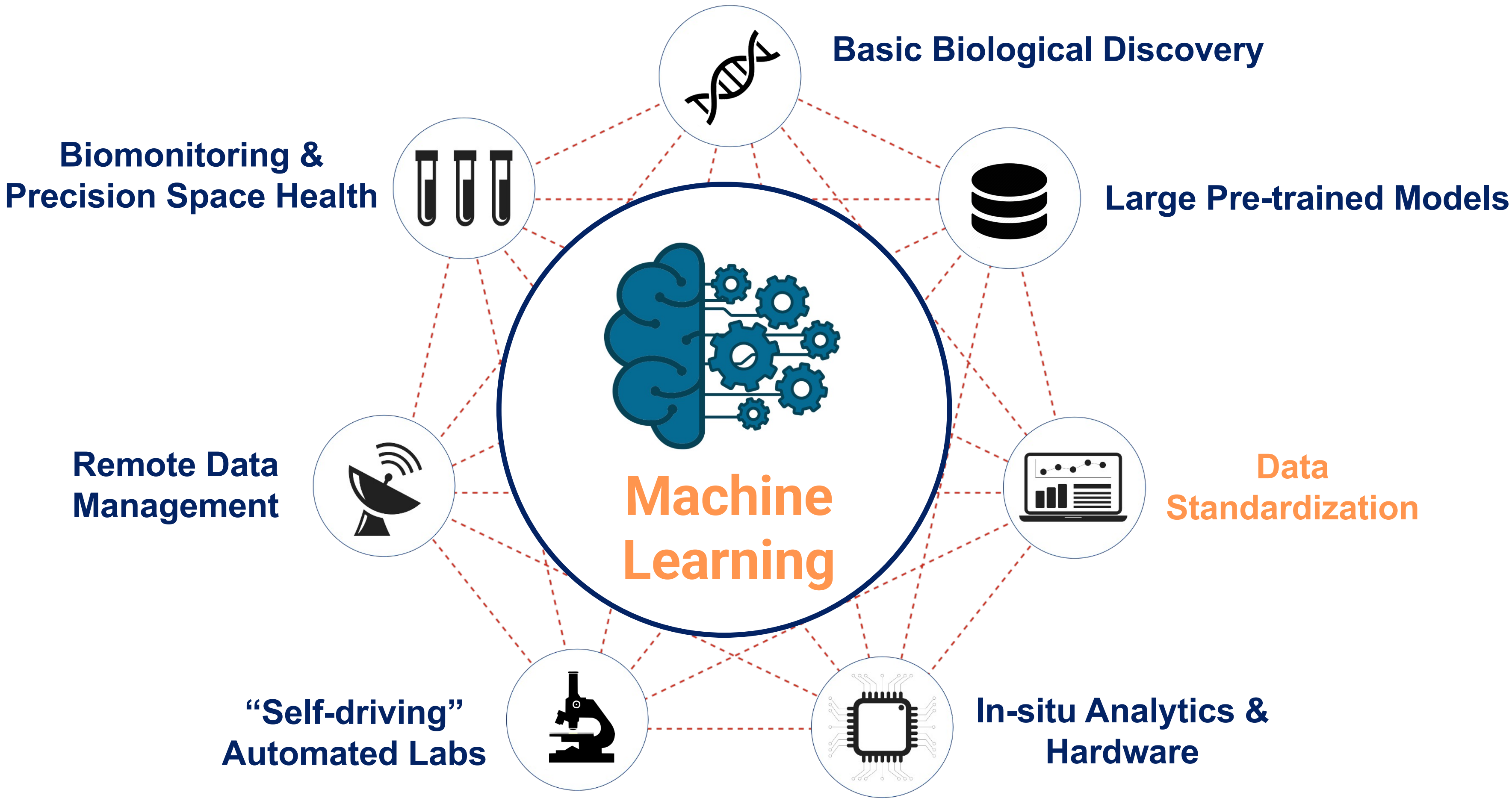


High model performance prior to overfitting

Summary of Outcomes

Large Pre-trained Models *to Connect Biomedical Knowledgebases with* ***Small Spaceflight Datasets***

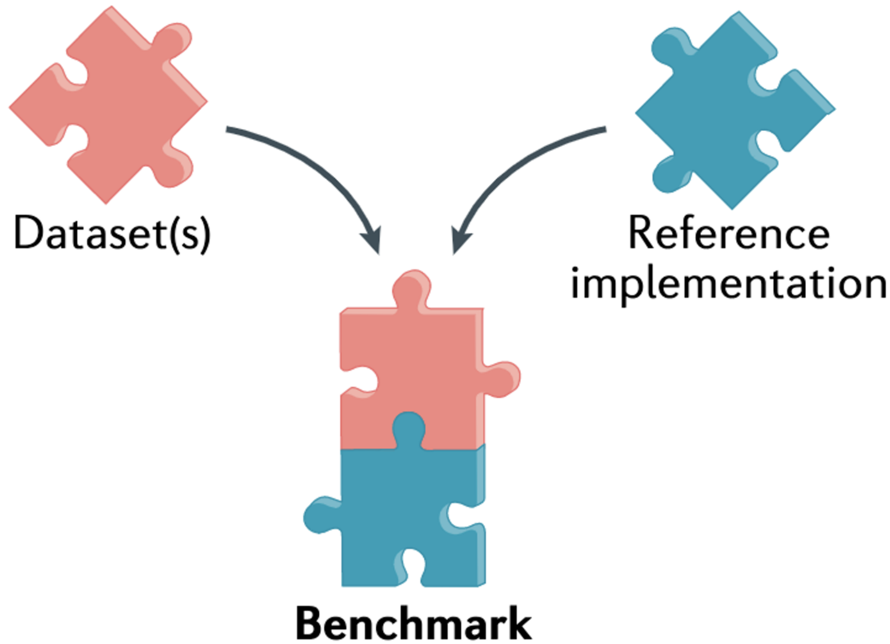
- Pretrained a large encoder model to learn gene-gene interaction networks in general
- Tested the trained model on a downstream, supervised task using a tiny space biology dataset
- Pretrained model outperforms traditionally trained model
- Future vision: “model zoo” of many pretrained models available to the space biology research community


Basic Biological Discovery
Biomonitoring & Precision Space Health
Large Pre-trained Models
Machine Learning
Remote Data Management
Data Standardization
"Self-driving" Automated Labs
In-situ Analytics & Hardware

A Suite of **Standardized** and **ML-ready** Training Datasets for Space Biology

Benchmark Datasets for Space Biology

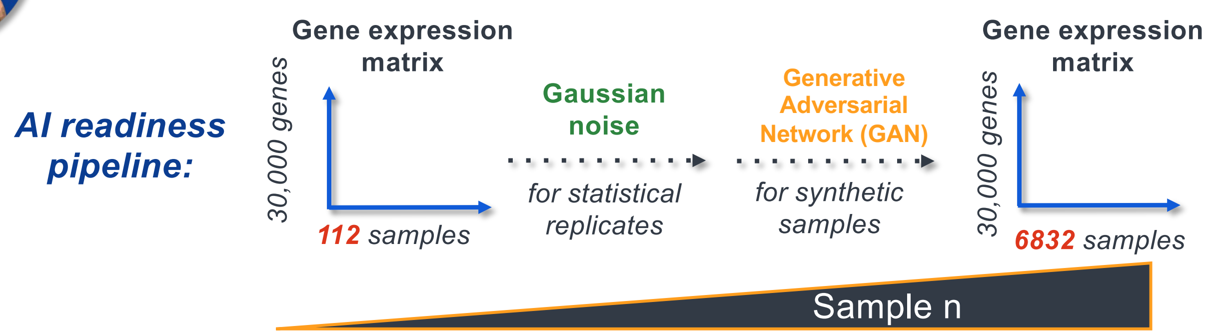
- **Scientific ML benchmarking**—Best ML algorithm for this problem
- **Application benchmarking**—Algorithm performance
- **System benchmarking**—Hardware and software architecture



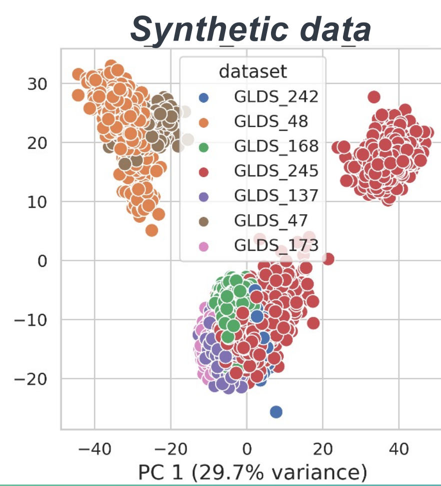
Thiyagalingam et al, *Nature Reviews* 2022

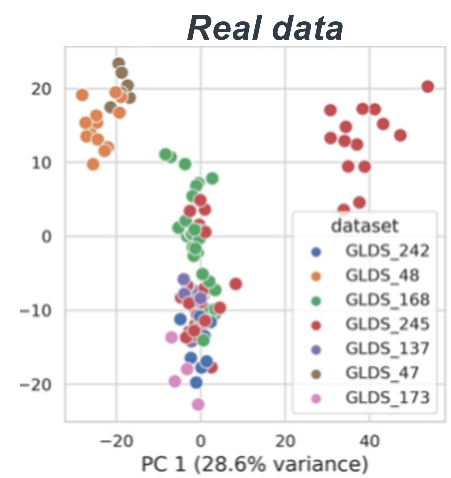
RNA sequencing benchmark dataset

Scientific motivation: Effects of spaceflight on mouse liver health



GeneLab Data Set	Tissue	Spaceflight Mission
47	liver	RR1 CASIS
48	liver	RR1 NASA
168	liver	RR1 NASA RR3 CASIS
137	liver	RR3 CASIS
173	liver	STS-135
242	liver	RR9
245	liver	RR6



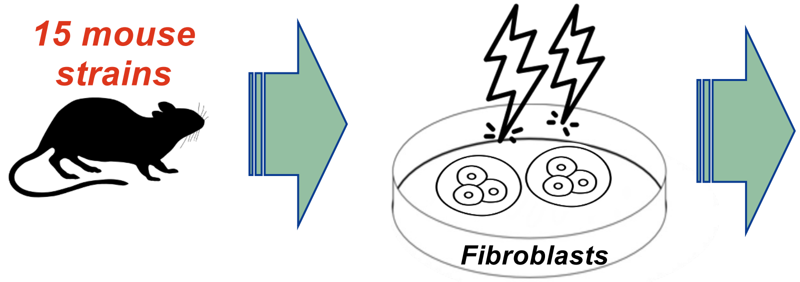


Fluorescence microscopy benchmark dataset

Scientific motivation: simulated space radiation causes cellular DNA damage

15 mouse strains

Simulated galactic cosmic rays



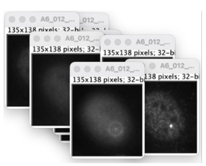
- 53PB1+ immunocytochemistry
- High-throughput imaging
- Automated quantification of 53 PB1+ radiation-induced foci

Penninckx et al. *Radiation Research* 2019

aws

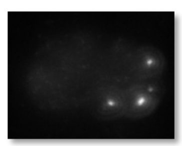
AI readiness pipeline:

DIP lib



maximum intensity projection

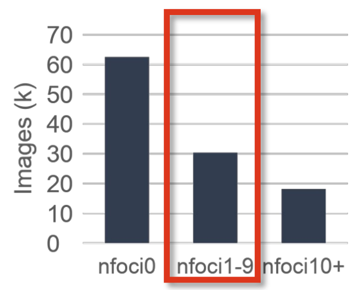
16-bit conversion



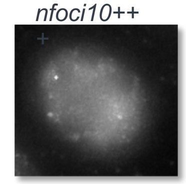
Raw Dataset (n = 94,193): *32-bit Z stacks (9 indices)*

Max Intensity Dataset (n=94,193): *16-bit single-index TIFFs*

automatically estimated nfoci



nfoci10++



Labels:

- *nfoci* (number of 53BP1+ DNA damage foci)
- *Radiation* (X-ray, ^{56}Fe…)

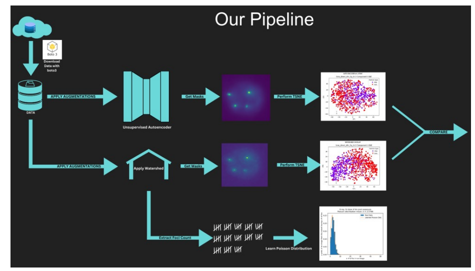
Cloud-based ML-ready Data Increase Scientific Community Engagement

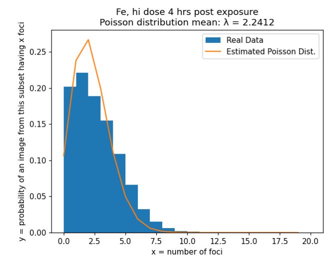
UC Irvine CS175: "Project in Artificial Intelligence" Senior Course

- BPS Microscopy benchmark dataset formed the basis for UCI CS175 senior ML projects
- Real-world data and scientific problems inspired the students to generate creative solutions
- ML-ready dataset allowed students to spend time on ML rather than preprocessing

- 9 teams focused on a variety of scientific questions:
  - Supervised classification
  - Unsupervised learning
  - Self-supervised learning
  - Segmentation and detection
  - Graph neural networks
  - Generating synthetic data

Image Segmentation and Foci Detection





Opportunities and Applications for **Machine Learning** to Support **Deep Space Exploration**

NASA Mission Goals: Deep Space Exploration

Deep space exploration challenges

- Distance from earth
- High latency communications
- Data bandwidth and power constraints
- Infrequent resupply
- Inability to evacuate
- Limited crew time

Moon to Mars Missions: Human and Biological Sciences Goal: Advance understanding of how biology responds to the environments of Moon, Mars, and deep space to advance fundamental knowledge, support safe, productive human space missions and reduce risks for future exploration.

Required: maximally autonomous and automated systems for science and health data collection, analysis, and real-time decision-making

Moon to Mars Mission Goals can be supported by current terrestrial capabilities in AI and ML

AI/ML Architecture to Support Deep Space Mission Goals

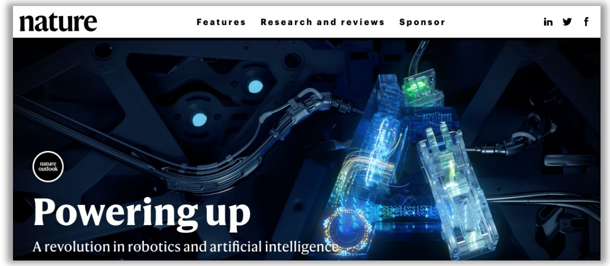
nature machine intelligence

Review article https://doi.org/10.1038/s42256-023-00617-5

Biomonitoring and precision health in deep space supported by artificial intelligence

Received: 23 December 2021

Ryan T. Scott[1,52], Lauren M. Sanders[2,52], Erik L. Antonsen



nature machine intelligence

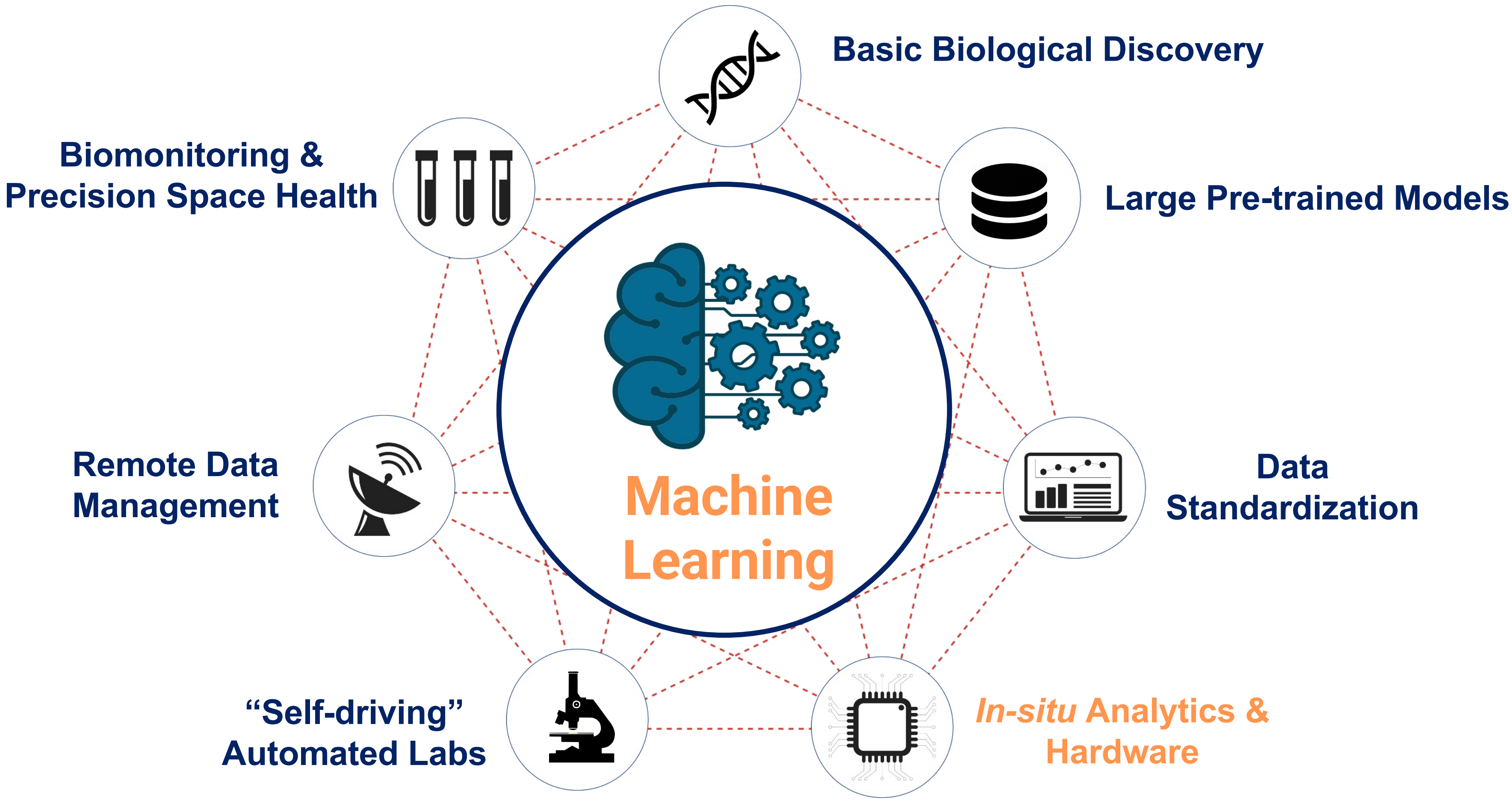
Review article https://doi.org/10.1038/s42256-023-00618-4

Biological research and self-driving labs in deep space supported by artificial intelligence

Received: 23 December 2021

Lauren M. Sanders[1,52], Ryan T. Scott[2,52], Jason H. Yang[3], Amina Ann Qutub[4],

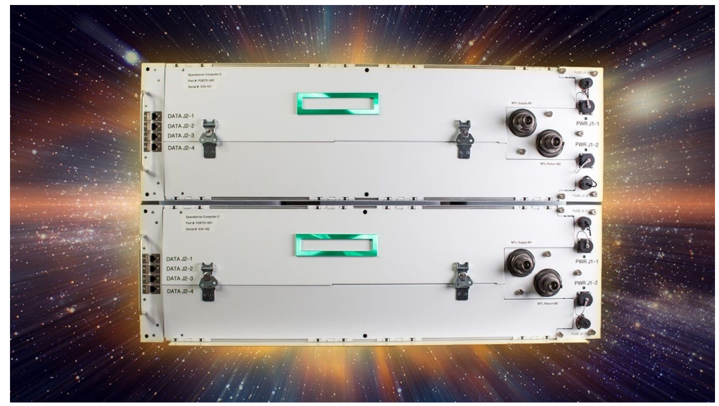
Stanford University · SJSU San José State University · Anschutz · The University of Texas at San Antonio UTSA · UCSF · Berkeley Lab Lawrence Berkeley National Laboratory · Mayo Clinic · Baylor College of Medicine · Icahn School of Medicine at Mount Sinai · Salk Institute for Biological Studies · Medical School University of Minnesota Driven to Discover · Johns Hopkins University · Rutgers New Jersey Medical School · Google · Harvard University · KBR · Michigan State University · NASA · Caltech · UAB The University of Alabama at Birmingham · UNF University of North Florida · Weill Cornell Medicine

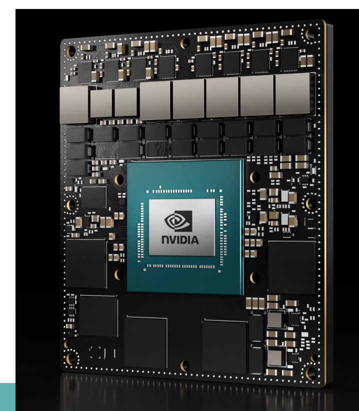

Basic Biological Discovery
Biomonitoring &
Precision Space Health
Large Pre-trained Models
Machine
Learning
Remote Data
Management
Data
Standardization
“Self-driving”
Automated Labs
In-situ Analytics &
Hardware

Hardware for ML-enabled *in-situ* Data Collection and Analysis

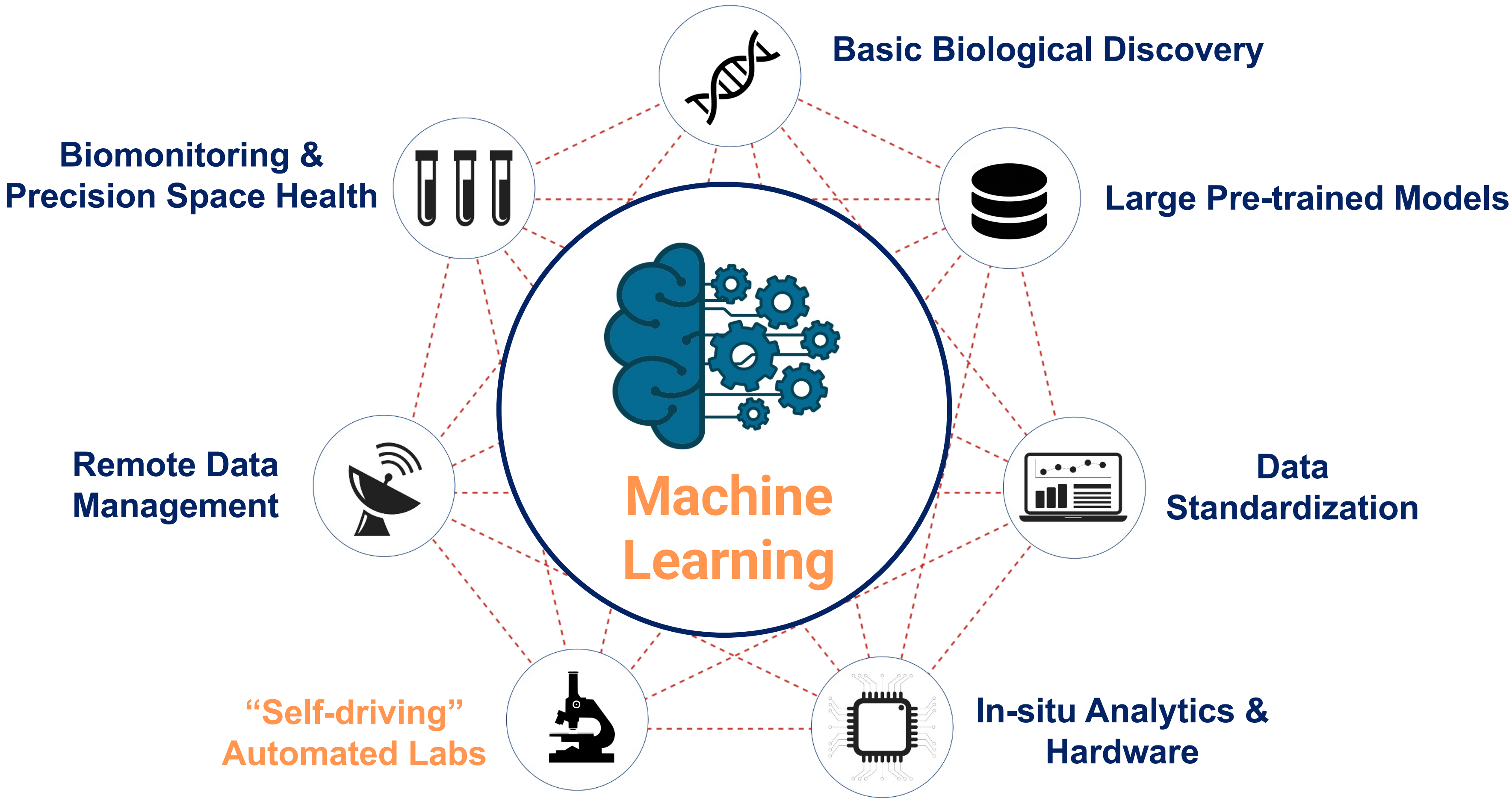
Category	Technology	Relevance to spaceflight
In situ capabilities: small footprint and resilient to environmental factors (radiation, acceleration, vibration)	Neuromorphic processors Edge computing[44]	Space-borne computing with very low power, little or no cooling, high efficacy for AI algorithms and resilience to radiation[146,147] Process and analyse data collected in deep-space missions on board for input to the PSH system

HPE's Spaceborne Computer



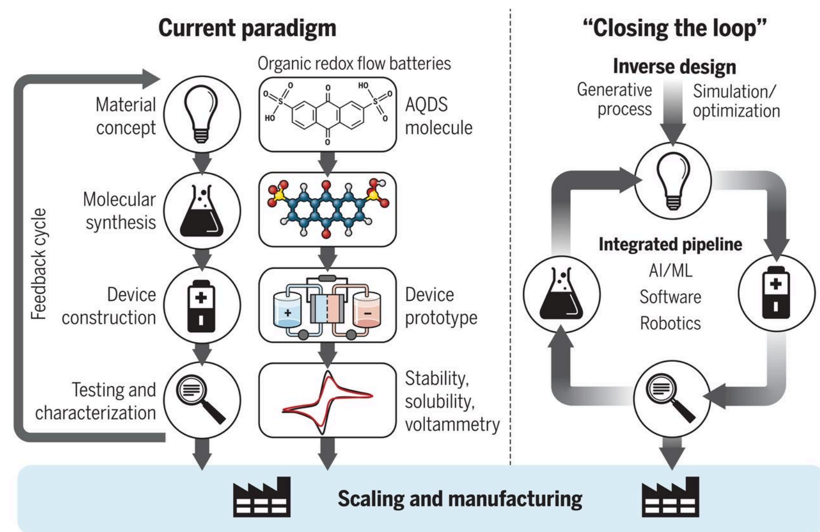


NVIDIA Jetson edge AI and robotics platform

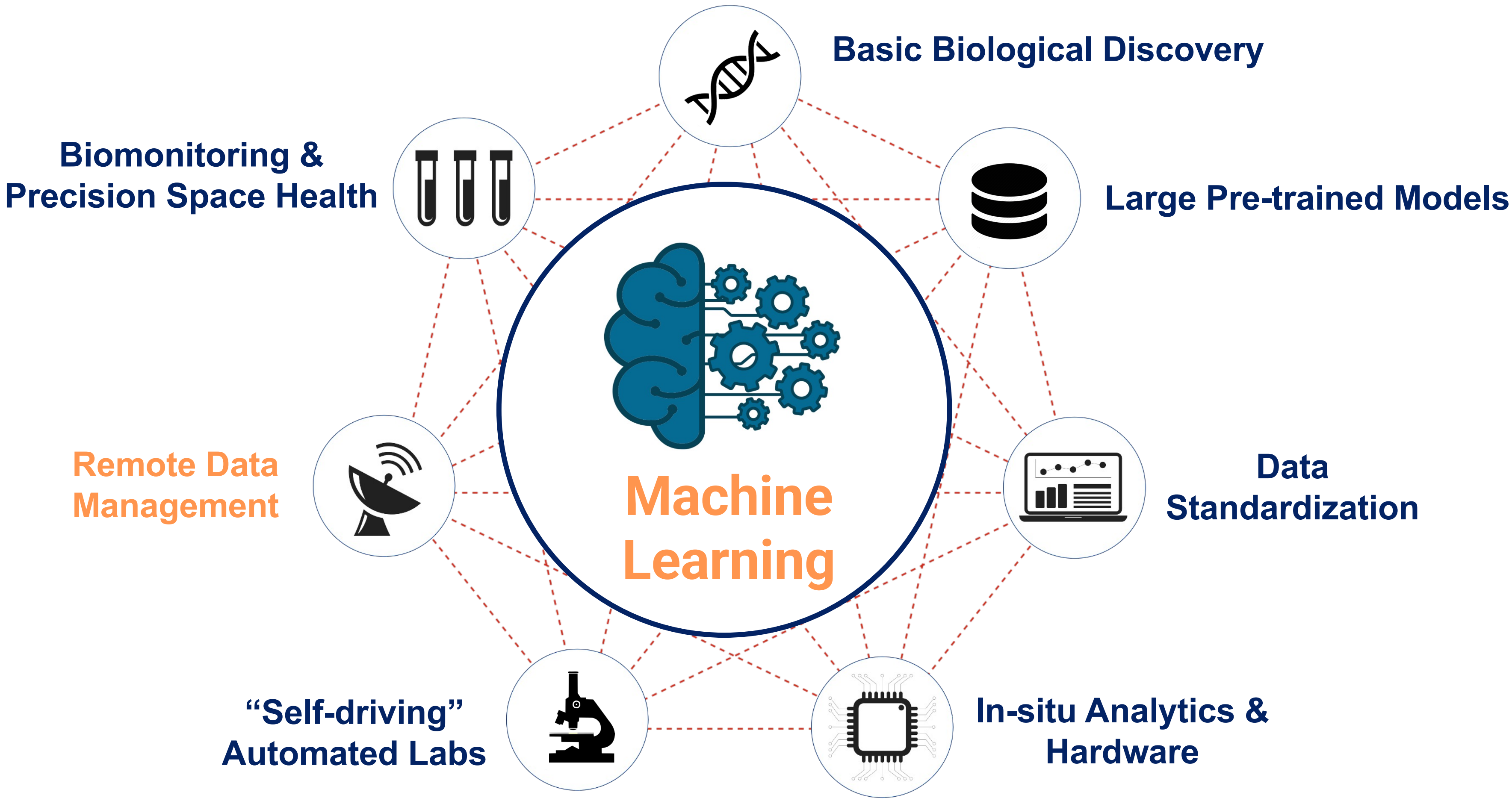

Basic Biological Discovery
Biomonitoring &
Precision Space Health
Large Pre-trained Models
Machine
Learning
Remote Data
Management
Data
Standardization
"Self-driving"
Automated Labs
In-situ Analytics &
Hardware

ML-enabled "Self-Driving" Laboratories

Category	Technology	Relevance to spaceflight
In situ data analysis	Active learning[107]	Train and deploy a model, which continuously monitors and retrains itself with self-assessments and regular human inspection

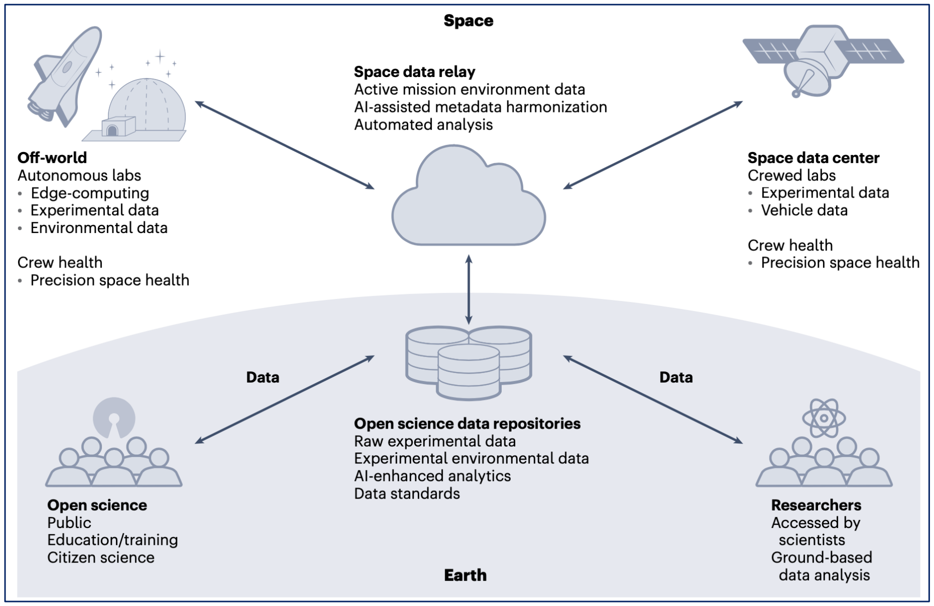


- "Self-driving" automated laboratory capabilities enable *in-situ* data collection
- Active learning & edge computing would allow *in-situ* data analysis


Basic Biological Discovery
Large Pre-trained Models
Data Standardization
In-situ Analytics & Hardware
"Self-driving" Automated Labs
Remote Data Management
Biomonitoring & Precision Space Health
Machine Learning

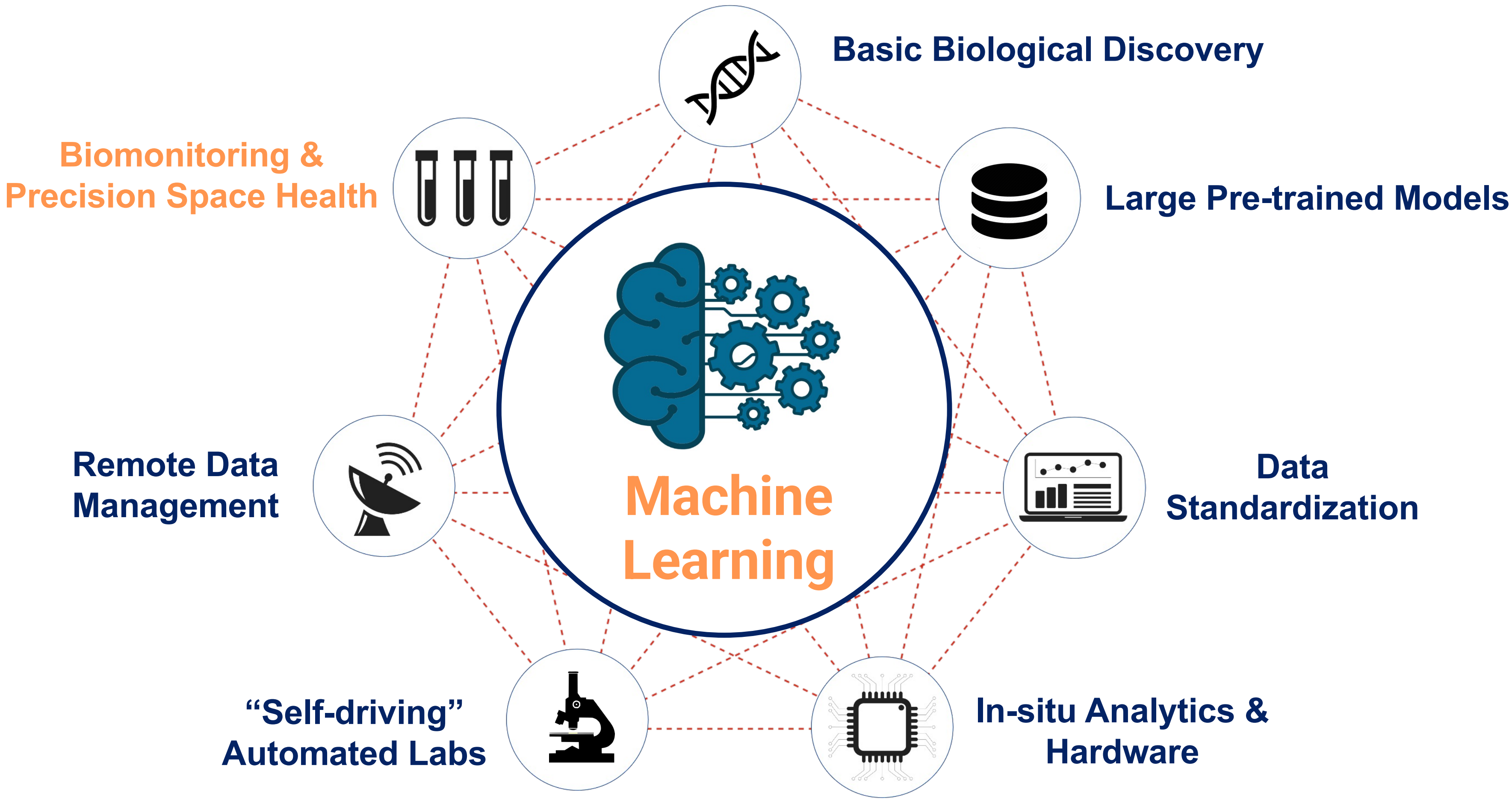
ML Approaches to Support Remote Data Management

Category	Technology	Relevance to spaceflight
Limiting data transfer to Earth	Federated learning[116]	Train a model on data collected in a deep-space mission and on Earth-based data for stronger inference
Distilling and maximizing computing needs in space	Transfer learning Dimensionality reduction[148] TinyML[149] Few-shot learning[150]	Train large models on Earth and deploy on data collected in-flight Identify key features to reduce data size Prune large neural networks to deploy on spacecraft or habitats with operational constraints Learn from few data points by leveraging contextual information



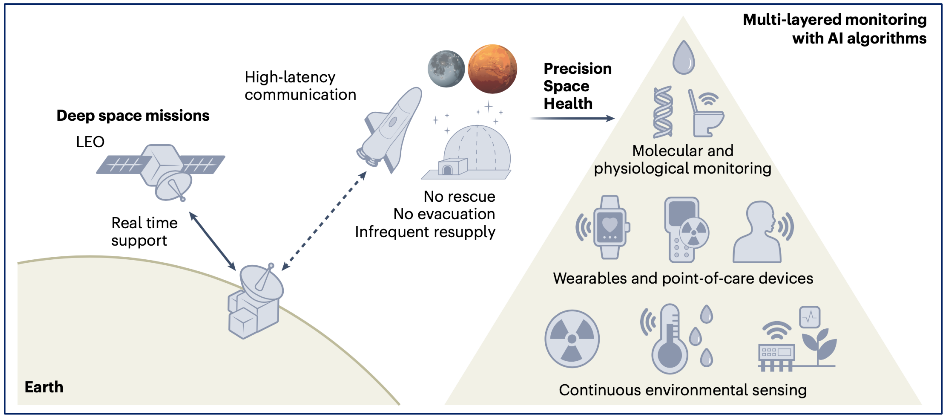
- ML methods such as federated learning, transfer learning, and few-shot learning support deep space data transfer

Scott et al., *Nature Machine Intelligence* 2023
Sanders et al., *Nature Machine Intelligence* 2023


Basic Biological Discovery
Biomonitoring &
Precision Space Health
Large Pre-trained Models
Machine
Learning
Remote Data
Management
Data
Standardization
"Self-driving"
Automated Labs
In-situ Analytics &
Hardware

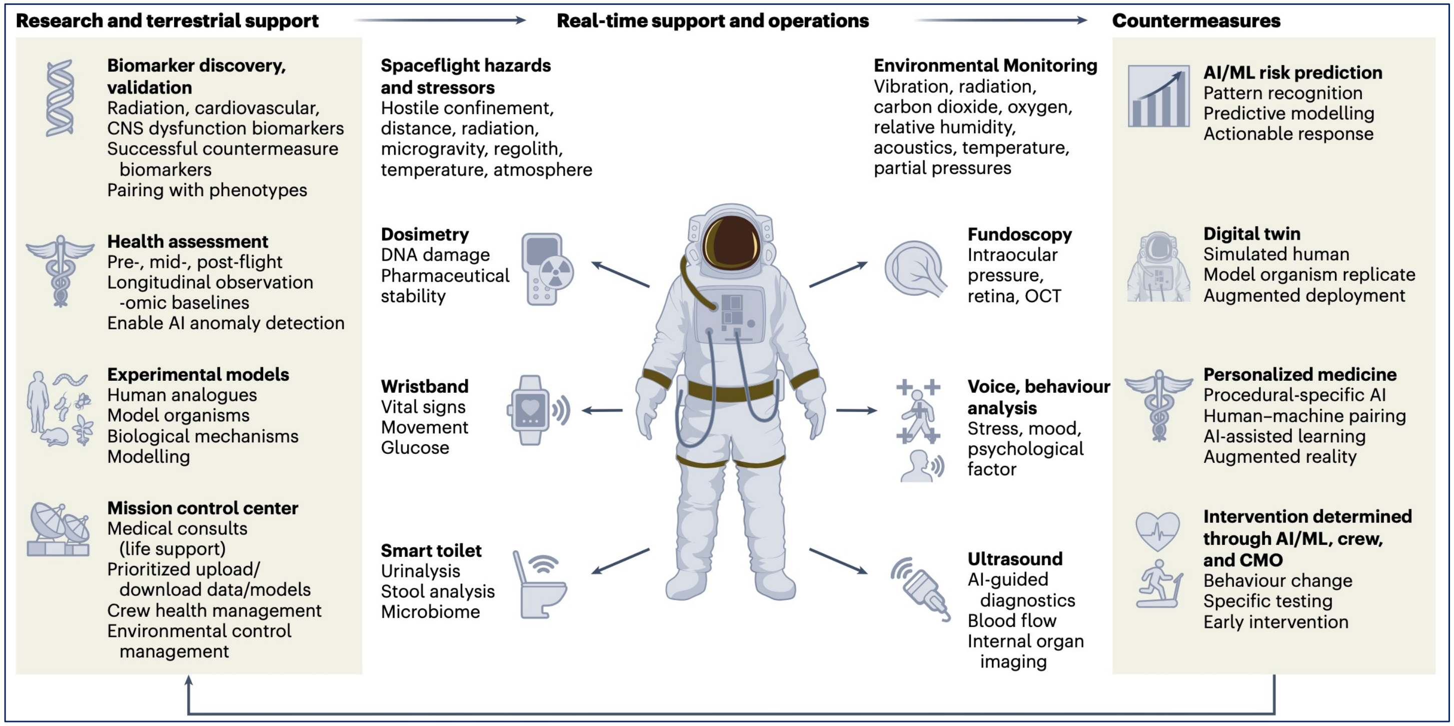
ML-enabled Biomonitoring Approaches

Category	Technology	Relevance to spaceflight
Methods to train on data that differ from inferencing context	Translation[152,153]	For example, train on radiation exposure data in animals and predict radiation risks for human crew members
Methods for when inferencing data are extremely different (for example, outliers) from training data	Generalization: Risk extrapolation[154,155] Domain invariant representation learning[154,155]	Prediction in a situation where an astronaut's biosensor data are outliers compared to the terrestrial clinical data used for model training
Methods for when inferencing data are persistently different from training data	Adaptation[156]	For example, adapting a model trained using terrestrial electrocardiogram data to a 'new normal' of electrocardiogram readings from astronauts whose heart physiology has changed in spaceflight



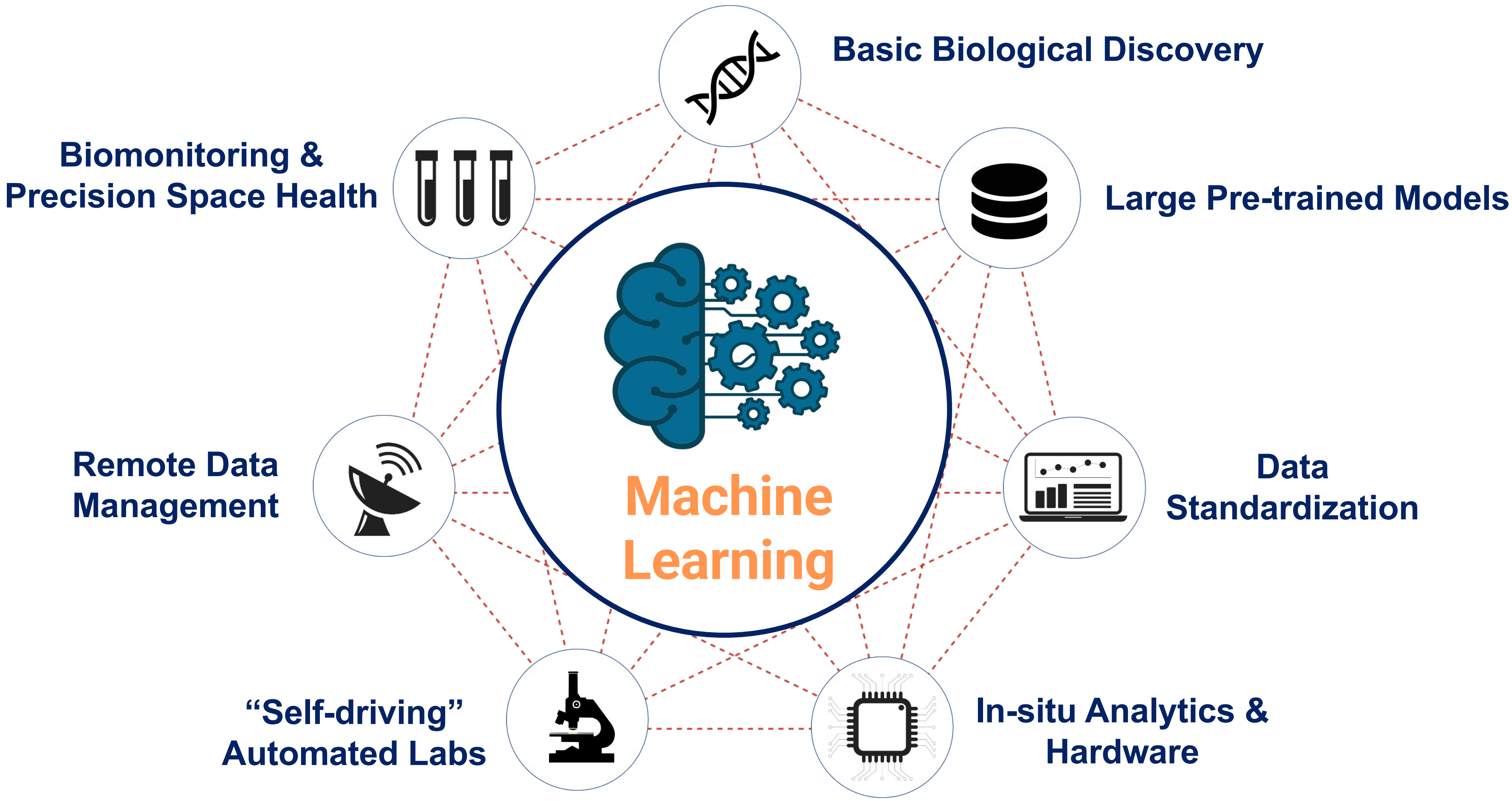
- Multi-layered monitoring of spacecraft and habitats & *in-situ* computing capabilities for real-time recommendations

ML-aided Precision Space Health System



- ML to support modeling, prediction and recommendations for a Precision Space Health system for real-time decision-making in deep space

Scott et al., *Nature Machine Intelligence* 2023
Sanders et al., *Nature Machine Intelligence* 2023


Basic Biological Discovery
Biomonitoring & Precision Space Health
Large Pre-trained Models
Machine Learning
Remote Data Management
Data Standardization
"Self-driving" Automated Labs
In-situ Analytics & Hardware

Acknowledgements

AI for Life in Space
- Kevin Li
- Riya Desai
- Adrienne Hoarfrost
- James Casaletto
- Nadia Ahmed
- Sylvain Costes

Radiation Biophysics Lab
- Egle Cekanaviciute
- Connie Pasternak
- Sylvain Costes

Open Science for Life in Space Teams

GeneLab ALSDA NBISC

AI4LS
Artificial Intelligence for Life in Space

2021 AI/ML Workshop Participants

Open Science Analysis Working Group Members

Compute

NASA Center for Climate Simulation Science Managed Compute Environment
- Aaron Skolnik
- Andre Avelino Paniagua
- Ellen Salmon
- Daniel Duffy

NASA

Support

- NASA Space Biology Program
- NASA Science Mission Directorate
- NASA Human Research Program
- NASA Biological and Physical Sciences
- NASA Postdoctoral Program