**Topical: Gaining Insights from Omics Data Using Artificial Intelligence in Space Crop Production**

*Primary Author:* Marshall Tabetah, +237-675352086, Plane Space Design, met4t@virginia.edu

*Co-Author:* Angela Price, IBM

**Introduction**

Data-driven, artificial intelligence (AI) powered systems can play a valuable role in space crop production and analysis. AI and automation can serve as an augmentation of space crew abilities to increase the efficiency of plant maintenance tasks and deliver enhanced insight into plant health and growth patterns in space. This in turn can reduce the amount of time required for a space crew to maintain a plant growth environment and the plants. The level of autonomy that can exist will depend on the overall intelligence and value of the insight generated from the data collected and on the abilities of the automation systems and robotics of the growth environment. The most critical component of any successful AI-powered system lies within the underlying data that it learns from. There is need for a highly organized plant knowledge repository that contains baseline plant species information, healthy growth patterns, unhealthy growth patterns and other anomalies that enable the AI and automation systems to detect the state of plant health and perform actions when intervention is necessary to ensure optimal growth conditions or to perform research related analysis.

There is a critical need for real-time and on-site execution of AI related functionalities which involves space-viable, robust and resilient communication protocol mechanisms, sensor networks, plant sample extraction and on-demand computing infrastructure. Given the existence of mass and power restrictions, miniaturization, modularity, and power usage efficiency of these components are necessary attributes in spaceflight conditions.

There are many different research objectives, most of them prioritized in the 2018 midterm assessment of the implementation of the Decadal Survey of Life and Physical Sciences Research at NASA [[[[1]](#endnote-1)](#NASEM)], relevant to space crop production that a well-designed, data driven, AI-powered system can enable. Some of the most significant areas include: early stress detection in plants; determining the relation between genetic engineering of plants and their adaptation to biotic and abiotic stresses; characterization and monitoring of plant microbes in order to identify microbes that are pathogenic or beneficial; diagnosis of infectious plant disease; physiological changes experienced by plants during growth; and optimization of crop selection with respect to a given microecosystem in the space habitat. Some types of omics data can be collected via environmental sensors, but some types of data can only be collected via plant sample extraction, e.g. data of the genome, transcriptome, and proteome of plants, and of microbes. This paper provides a review of the state-of-the-art implementation of artificial intelligence techniques in revealing meaningful insights and predictions from plant omics data to enable fulfillment of, but not limited to, the listed objectives; and aims to address any necessary developments (highlighted in italic) both on the data and the algorithms that will improve the level of insight and prediction.

1. **Early Stress Detection; Optimization of Crop Selection**.

Detecting plant stress from abiotic factors such as environmental temperatures or pH of growth medium, and from biotic factors such as microbial pathogenic infections during crop growth cycles is needed to ensure an optimal environment for targeted crop production objectives. Early identification and detection of plant stressors is very valuable as it enables corrective and preventive actions to be taken before the appearance of visual symptoms. Notable early stress detection techniques in plants are hyperspectral imaging (HSI) and positron emission tomography (PET). In [[[[2]](#endnote-2)](#Kumar)], genes related to water stress have been detected using machine learning and HSI. Ref. [[[[3]](#endnote-3)](#Galieni)] highlights PET as the key quantitative functional imaging technique that provides a time-dynamic non-disruptive information of the modifications of functional mechanisms and transport flows in the vascular system in response to biotic and abiotic stress.

In exploring different growth conditions, plants can be subjected to varying permutations of environmental factors such as different water and nutrient concentrations or other explicitly orchestrated conditions. For example, in [[[[4]](#endnote-4)](#Pandey)], 60 maize and 60 soybean plants are subjected to varying levels of either water deficiency or nutrient limitation stress, with the goal of creating a wide range of variation in the chemical properties of plant leaves. The plant chemical traits were determined from HSI data. In [[[[5]](#endnote-5)](#Feng)], HSI has been combined with machine learning to study the physiological mechanisms of okra’s response to salt stress for selection purposes, i.e. selecting the breeds that show fewer insufficiencies and fewer deficiencies in physiology. Hyperspectral imaging is, thus, deemed a high-throughput phenotyping technology given the potency to provide a variation in chemical properties of plants as the levels of growth supplies are varied. [***Objective 1***] ***Hyperspectral imaging methods should be developed that select wavelength bands that contain information relevant to detection of stress and disease***, as in [[[[6]](#endnote-6)](#Nagasubramanian)]. Given the need for a specific subset of wavelength bands that contain pertinent data for processing, computing and storage efficiency can be optimized. ***There is also need to develop the knowledge repository of spectral signatures to crop stressors.***

1. **Genetic Adaptation of Plants to Stressors in Space; the Importance of the Environmental Context**.

Plant seeds can be genetically engineered to produce plants that can adapt to potential stressors in spaceflight conditions. Given that single-gene expression differences can have strong effects on physiological response, it is necessary to analyze and understand the data collected from experiments conducted in space conditions which pertain to the plant genotype and the response to specific stressors. In a study with the aim to reveal how gene expression determines physiological response in microgravity, samples of Arabidopsis thaliana were grown from its seedlings aboard the ISS and compared to ground analogs, in [[[[7]](#endnote-7)](#Kruse)]. Results of transcriptomic and proteomic analyses revealed that microgravity-induced differences in the regulation of transcripts did not match the trends in the differential expressions in the proteome: some proteins were more expressed in microgravity when the associated transcripts were downregulated, while other proteins were less expressed when the associated transcripts were upregulated. The gene expression here inadequately explains the protein expressions. It is also demonstrated in [[[[8]](#endnote-8)](#doAmaral)] that the gene expression, determined through genomic or transcriptomic analyses, does not adequately account for physiological function in plants. It turns out that the higher levels of organization affect the lower levels, with the result that responses to environmental changes and stimuli do affect gene expression. Environmental changes are causative to molecular interactions in plants both at the intracellular and intercellular levels, and to systemic responses that involve several organs and tissues. [***Objective 2***] ***The increase in system complexity from responses to environmental conditions can be accounted for by performing integral analyses of omics data types that convey information about the different levels of organization involved***. The various observations show that one single genotype can result in multiple different phenotypes when cultivated under different environmental conditions, and the environmental context in which the data are collected should be noted. [***Objective 3***] ***The environmental context should be part of the space plant knowledge repository; and the context should be part of the metadata when the experimental data is archived.***

1. **Integration of Multiple Data Types**.

Data of different levels (multi-omics data) from samples of the same cohort can be integrated in parallel, or in a hierarchy [[[[9]](#endnote-9)](#Wu)]. In parallel integration, the first stage of analysis is to select features that have important implications in each single-omics dataset analyzed independently; the next stage is to combine the selected features from the different levels to determine the phenotypic outcome. Hierarchical integration, on the other hand, incorporates prior knowledge of regulatory relationship among different platforms of omics data in the integration procedure; the features in a lower-level of organization are seen to influence the features in a higher level of organization. [***Objective 4***] ***Data integration approaches should account for the environmental effects on plant biology and the interaction of different levels of organization*. *There is need to conduct experiments that will provide adequate knowledge of regulatory relationships among different levels of omics data in unexplored space environments of interest****.*

1. **Application of Machine Learning Algorithms in Analyzing Omics Data**.

Machine learning (ML), a common artificial intelligence data approach, can be used to connect genotypes to various phenotype levels – the biochemical level and the macroscopic level – in different environmental conditions. The biochemical level is described by the analysis of omics data, which represent genomics, epigenomics, protein binding to DNA, transcriptomics, proteomics, and metabolomics. Traditional statistical approaches have difficulty scaling to multiple variables, and rely on assumptions about the data-generating systems. In comparison, ML algorithms formulate the rules from the data instead of relying on assumptions, and are often characterized by a combination of various features. The ability of machine learning to combine disparate pieces of input features is particularly useful when the data originate from different measurements [[[[10]](#endnote-10)](#Mahood)].

Machine learning can be used to predict the activity, regulation, and function of genes and gene products, as well as to predict higher-order interactions such as subcellular localization and protein-protein interactions [[[[11]](#endnote-11)](#vanDijk)]. ML has unique capabilities in the quantification of relevant plant traits. ML can overcome limitations to infer gene regulatory networks in traditional-based approaches which cannot properly deal with complex non-linear and higher-order dependencies between expression levels [[11](#vanDijk)]. Some examples are the inference of a dynamic gene regulatory network related to nitrogen use efficiency in Arabidopsis [[[[12]](#endnote-12)](#Varala)], and of regulatory networks coordinating timing and rate of gene expression in response to environmental signals in rice [[[[13]](#endnote-13)](#Wilkins)].

[***Objective 5***] ***The effects of the environmental conditions and environmental stimuli in new plant space experiments will have to be captured in machine learning models***. In some scenarios, the AI-powered systems in space will initially rely on ML algorithms that are developed without prior knowledge of the space environment. This implies the unavailability of data for training a supervised learning model. An unsupervised learning approach in which the environmental conditions and stimuli, captured from sensors, will be among input variables is needed. Once the dependence of the phenotypic outcomes on the input variables are learned from the unsupervised model, these relationships can be used as labeled data for training supervised machine learning models. As a result, the effects of the environmental conditions (pressure, temperature, microgravity or altered gravity) and environmental stimuli (high energy radiation events) in space can be captured in the ML models.

***Performance of Machine Learning Models***. A number of measures must be taken for machine learning to be implemented successfully at scale. In supervised learning models, a predictive model is optimized by fitting its parameters to perform well on labeled training data, consisting of inputs and corresponding known outputs. The resulting models can then make predictions for novel test data. Care should be taken to avoid overfitting, situations in which the model performs well on training data but does not generalize well to novel data. Feature-based ML models, such as random forest (RF) and support vector machines (SVMs), use other models such as image processing in a first step to extract features, and then use machine learning for decision-making. The performance is thus limited to precision in feature extraction. Deep learning (DL) and artificial neural networks (ANNs) overcome the performance limitations of ML algorithms by integrating both feature extraction and decision-making into a single model. While a few dozen to a few hundred input training data points can be sufficient for feature-based methods, DL and ANN algorithms typically require thousands to millions of input data points for accurate model development [[10](#Mahood)]. Implementation of DL or ANN can, thus, be computationally demanding. In space applications where excessive power consumption is prohibitive, the implementation of ANN or DL algorithms may surpass the limitations in computing and power requirements. [***Objective 6***] ***Solutions that expand the utilization of computing resources beyond those located in the space platform will be needed to accommodate complex and demanding computing workloads.***

1. **Monitoring of Microbial Activity and Function**

Microbes that constitute important microbiomes of plants, as well as those that cause disease in plants should be monitored and controlled as a preventive contribution to plant health and for diagnosis of disease. Knowledge of the early signs of harmful pathogens can help in creating mitigation strategies that neutralize the adverse effects of harmful pathogens before they become further exacerbated and negatively impact crop production. An indicator of microbial activity and function is the amount of transcript sequences from the microorganisms in a microbiome [[[[14]](#endnote-14)](#Lucaciu)]. A Next Generation Sequencing (NGS) technology has been used to characterize the plant microbiome and screen for pathogens aboard the International Space Station (ISS), in VEGGIE tests on different plantings of red romaine lettuce [[[[15]](#endnote-15)](#Khodadad)]. The diverse microbial communities were identified as potentially non-pathogenic to humans, and the cultured leafy greens were deemed a safe supplement to the diet of astronauts. In [[[[16]](#endnote-16)](#Haveman)], the microbiome of red romaine lettuce grown in ISS-like ground control conditions is characterized using a nanopore sequencing platform. As in plants, ML and DL can be applied to molecular data of the microbiome to obtain meaningful insights: see [[[[17]](#endnote-17)](#Deng)] on the use of DL to predict the plant phenotype from the microbiome data; and see [[[[18]](#endnote-18)](#Thompson)] on deciphering the functional relationship between specific microbes and ecosystem properties through the integration of DNA sequencing technology and machine learning. A major limitation in the characterization of the plant microbiome, however, is that the plant microbial repository which is used to predict the metagenome, metatranscriptome, or the metaproteome is currently deficient in certain areas – with the need to expand and refine databases which combine species information with environmental and plant data [[14](#Lucaciu)], and the need for a comprehensive peptide reference database [[[[19]](#endnote-19)](#Levy)]*.* [***Objective 7***]***Several experiments will have to be conducted in order to expand and refine the microbiome repository***.

1. **Computing Hardware**

There have been ongoing efforts in the miniaturization of DNA/RNA sequencing technologies. The nanopore sequencing platform mentioned in [[16](#Haveman),[[[20]](#endnote-20)](#StahlRommel)] is pocket-sized and weighs less than 100 g. The on-site sequencing capabilities have alleviated the requirement to return samples to Earth for analysis in culture-dependent methods that bear a number of limitations, with the most significant being bias towards the detection of culturable organisms and the inherent delay between sample collection and ground-based analysis. The sequenced data is downlinked to Earth for further analyses as opposed to analyzing on-site. [***Objective 8***] ***There is, thus, need to integrate machine learning capabilities in the miniaturized platforms for the analyses of the sequenced data onboard spacecraft (or space habitat), e.g. by leveraging edge computing****.*

The omics data that is collected for analysis may not fit in the total local storage of the space platform. [***Objective 9***] ***This may warrant the need to develop a means to partition the data among remote memory blocks which are connected to the local memory in a distributed memory network****.* Capabilities for the submission of multiple jobs to several processors or nodes have to be developed in space platforms. The computing platform should leverage remote computing nodes that implement algorithms that are synchronized across the network via a primary orchestration node.

**References**

1. **National Academies of Sciences, Engineering, and Medicine** **2018**. *A Midterm
Assessment of Implementation of the Decadal Survey on Life and Physical
Sciences Research at NASA*. Washington, DC: The National Academies Press.
<https://doi.org/10.17226/24966>**.** [↑](#endnote-ref-1)
2. **Kumar** P, Eriksen RL, Simko I and Mou B (**2021**) Molecular Mapping of Water-Stress Responsive Genomic Loci in Lettuce (Lactuca spp.) Using Kinetics Chlorophyll Fluorescence, Hyperspectral Imaging and Machine Learning. Front. Genet. 12:634554. doi:[10.3389/fgene.2021.634554](https://doi.org/10.3389/fgene.2021.634554) [↑](#endnote-ref-2)
3. **Galieni** A, D’Ascenzo N, Stagnari F, Pagnani G, Xie Q and Pisante M (**2021**) Past and Future of Plant Stress Detection: An Overview From Remote Sensing to Positron Emission Tomography.Front. Plant Sci. 11:609155. doi: [10.3389/fpls.2020.609155](https://doi.org/10.3389/fgene.2021.634554) [↑](#endnote-ref-3)
4. **Pandey** P, Ge Y, Stoerger V and Schnable JC (**2017**) High Throughput In vivo Analysis of Plant Leaf Chemical Properties Using Hyperspectral Imaging. Front. Plant Sci. 8:1348. [doi: 10.3389/fpls.2017.01348](https://doi.org/10.3389/fpls.2017.01348) [↑](#endnote-ref-4)
5. **Feng** X., Zhan Y., Wang Q., Yang X., Yu C., Wang H., Tang Z-Y., Jiang D., Peng C. and He1 Y., Hyperspectral imaging combined with machine learning as a tool to obtain high-throughput plant salt-stress phenotyping, The Plant Journal, (**2020**), 101, 1448–1461. [doi: 10.1111/tpj.14597](https://doi.org/10.1111/tpj.14597) [↑](#endnote-ref-5)
6. **Nagasubramanian** K., Jones S., Singh A.K., Sarkar S., Singh A. and Ganapathysubramanian B., Plant disease identification using explainable 3D deep learning on hyperspectral images, Plant Methods (**2019**) 15:98. <https://doi.org/10.1186/s13007-019-0479-8> [↑](#endnote-ref-6)
7. **Kruse** C.P.S., Meyers A.D., Basu P., Hutchinson S., Luesse D.R. and Wyatt S.E., Spaceflight induces novel regulatory responses in Arabidopsis seedling as revealed by combined proteomic and transcriptomic analyses, BMC Plant Biology (**2020**) 20:237. <https://doi.org/10.1186/s12870-020-02392-6> [↑](#endnote-ref-7)
8. **do Amaral** MN and Souza GM (**2017**) The Challenge to Translate OMICS Data to Whole Plant Physiology: The Context Matters. Front. Plant Sci. 8:2146. [doi: 10.3389/fpls.2017.02146](https://doi.org/10.3389/fpls.2017.02146) [↑](#endnote-ref-8)
9. **Wu** C., Zhou F., Ren J., Li X., Jiang Y. and Ma S., A Selective Review of Multi-Level Omics Data Integration Using Variable Selection, *High-Throughput* **2019**, *8*, 4. [doi:10.3390/ht8010004](https://doi.org/10.3390/ht8010004) [↑](#endnote-ref-9)
10. **Mahood**, E. H., L. H. Kruse, and G. D. Moghe. **2020**. Machine learning: A powerful tool for gene function prediction in plants. *Applications in Plant Sciences* 8(7): e11376. [doi:10.1002/aps3.11376](https://doi.org/10.1002/aps3.11376) [↑](#endnote-ref-10)
11. **van Dijk** A.D.J., Kootstra G., Kruijer W., and de Ridder D., Machine learning in plant science and plant breeding, iScience 24, 101890, January 22, **2021**. <https://doi.org/10.1016/j.isci.2020.101890> [↑](#endnote-ref-11)
12. **Varala** K., Marshall-Colón A., Cirrone J., Brooks M.D., Pasquino A.V., Léran S., Mittal S., Rock T.M., Edwards M.B., Kim G.J., Ruffel S., McCombie W.R., Shasha D., and Coruzzi G.M.,Temporal transcriptional logic of dynamic regulatory networks underlying nitrogen signaling and use in plants, May **2018**. Proceedings of the National Academy of Sciences 115(25):201721487. <http://dx.doi.org/10.1073/pnas.1721487115> [↑](#endnote-ref-12)
13. **Wilkins** O., Hafemeister C., Plessis A., Holloway-Phillips M-M, Pham G.M., Nicotra A.B., Gregorio G.B., Jagadish S.V.K., Septiningsih E.M., Bonneau R., and Purugganan M., EGRINs (Environmental Gene Regulatory Influence Networks) in Rice That Function in the Response to Water Deficit, High Temperature, and Agricultural Environments, Plant Cell **2016** Oct; 28(10): 2365-2384. <https://dx.doi.org/10.1105/tpc.16.00158> [↑](#endnote-ref-13)
14. **Lucaciu** R, Pelikan C, Gerner SM, Zioutis C, Köstlbacher S, Marx H, Herbold CW, Schmidt H and Rattei T (**2019**) A Bioinformatics Guide to Plant Microbiome Analysis. Front. Plant Sci. 10:1313. [doi: 10.3389/fpls.2019.01313](https://doi.org/10.3389/fpls.2019.01313) [↑](#endnote-ref-14)
15. **Khodadad**, C.L.M.; Hummerick, M.E.; Spencer, L.E.; Dixit, A.R.; Richards, J.T.; Romeyn, M.W.; Smith, T.M.; Wheeler, R.M.; Massa, G.D. Microbiological and Nutritional Analysis of Lettuce Crops Grown on the International Space Station. Front. Plant Sci*.* **2020**, *11*. <https://doi.org/10.3389/fpls.2020.00199> [↑](#endnote-ref-15)
16. **Haveman** N.J., Khodadad C.L.M, Dixit A.R., Louyakis A.S., Massa G.D., Venkateswaran K. and Foster J.S. Evaluating the lettuce metatranscriptome with MinION sequencing for future spaceflight food production applications, npj Microgravity (**2021**) 7:22. <https://doi.org/10.1038/s41526-021-00151-x> [↑](#endnote-ref-16)
17. **Deng** Z., Zhang J., Li J. and Zhang X., Application of Deep Learning in Plant–Microbiota Association Analysis, Front. Genet., 08 October **2021** <https://doi.org/10.3389/fgene.2021.697090> [↑](#endnote-ref-17)
18. **Thompson** J, Johansen R, Dunbar J, Munsky B (**2019**) Machine learning to predict microbial community functions: An analysis of dissolved organic carbon from litter decomposition. PLoS ONE 14(7): ): e0215502. <https://doi.org/10.1371/journal.pone.0215502> [↑](#endnote-ref-18)
19. **Levy**, A., Conway, J. M., Dangl, J. L., and Woyke, T. (**2018**). Elucidating bacterial
gene functions in the plant microbiome. *Cell Host Microbe* 24, 475–485. <https://doi.org/10.1016/j.chom.2018.09.005> [↑](#endnote-ref-19)
20. **Stahl-Rommel** S., Jain M., Nguyen H.N., Arnold R.R., Aunon-Chancellor S.M., Sharp G.M., Castro C.L., John K.K., Juul S., Turner D.J., Stoddart D., Paten B., Akeson M., Burton A.S. and Castro-Wallace S.L., Real-Time Culture-Independent Microbial Profiling Onboard the International Space Station Using Nanopore Sequencing, *Genes* **2021**, *12(1)*, 106. <https://doi.org/10.3390/genes12010106> [↑](#endnote-ref-20)