npj | microgravity

Article

Published in cooperation with the Biodesign Institute at Arizona State University, with the support of NASA

https://doi.org/10.1038/s41526-025-00525-5

GLARE: discovering hidden patterns in spaceflight transcriptome using representation learning

Check for updates

DongHyeon Seo[1], Hunter F. Strickland[2,3], Mingqi Zhou[3], Richard Barker[4], Robert J. Ferl[3,5], Anna-Lisa Paul[3,6] & Simon Gilroy[7]

Spaceflight studies present novel insights into biological processes through exposure to stressors outside the evolutionary path of terrestrial organisms. Despite limited access to space environments, numerous transcriptomic datasets from spaceflight experiments are now available through NASA's GeneLab data repository, which allows public access, encouraging further analysis. While various computational pipelines and methods have been used to process these transcriptomic datasets, learning-model-driven analyses have yet to be applied to a broad array of such spaceflight-related datasets. In this study, we present an open-source pipeline, GLARE: GeneLAb Representation learning pipelinE, which consists of training different representation learning approaches from manifold learning to self-supervised learning that enhance the performance of downstream analytical tasks. We illustrate the utility of GLARE by applying it to gene-level transcriptional values from the results of the CARA spaceflight experiment, an Arabidopsis root tip transcriptome dataset that spanned light, dark, and microgravity treatments. We show that GLARE not only substantiated the findings of the original study concerning cell wall remodeling but also revealed additional patterns of gene expression affected by the treatments, including evidence of hypoxic response. This work suggests there is great potential to supplement the insights drawn from initial studies on spaceflight omics-level data through further machine-learning-enabled analyses.

Spaceflight studies present unprecedented insights into biological processes through exposure to unique environmental stressors that have not been experienced by any form of life on Earth. In response to the spaceflight environment, organisms initiate specific transcriptional responses to novel conditions. Thus, one key to understanding how biology responds to spaceflight stressors like microgravity, radiation, and hypoxia is through transcriptomic analysis to study the gene expression profiles that drive physiological adaptation triggered by the spaceflight environment[1]. Space-related transcriptional studies have now also broadened into multi-omic spaceflight investigations that are well-suited to multiple rounds of analysis facilitated by the publicly available datasets in the NASA GeneLab database.

The importance of studying plant biology specifically in space has been identified both for exploring the fundamental responses of biology to the spaceflight environment and at a very practical level for developing bio-regenerative life support systems for long-term space exploration[2,3]. Understanding of transcriptomic and physiological changes elicited in plants by spaceflight conditions through analyzing transcriptional and other –omic patterns is therefore a focus of much current plant space biology experimentation (e.g.,[4,5]). For example, the CARA (Characterizing Arabidopsis Root Attraction) experiment was designed to compare the spaceflight transcriptome responses between different genotypes of *Arabidopsis thaliana's* root tips under various conditions[6]. This experiment explored the patterns of gene expression from root tip cells in the spaceflight environment on the International Space Station (ISS) with comparable ground controls, and the lighting sub-environments among three different genotypes. While these kinds of experiments in plant space biology have provided many key insights, they have so far largely relied upon the primary transcriptomic analysis of the original research team. To provide a pipeline that can be applied to increase the depth of transcriptomic analyses for previous and future spaceflight experiments, we introduce GLARE:

[1]Information Sciences Institute, University of Southern California, Marina del Rey, CA, USA. [2]Plant Molecular and Cellular Biology Program, University of Florida, Gainesville, FL, USA. [3]Department of Horticultural Sciences, University of Florida, Gainesville, FL, USA. [4]Blue Marble Space Institute of Science, Seattle, WA, USA. [5]Office of Research, University of Florida, Gainesville, FL, USA. [6]Interdisciplinary Center for Biotechnology Research, University of Florida, Gainesville, FL, USA. [7]Department of Botany, University of Wisconsin-Madison, Madison, WI, USA. e-mail: sgilroy@wisc.edu

GeneLab Representation learning pipelinE. This pipeline consists of multiple methods of data visualization and projection, representational learning, and post hoc analyses, provided in a pre-built series of modules targeted for use with datasets from the Genelab Data System (GLDS). We show the utility of GLARE by applying it to the CARA dataset (OSD-120) and conducting an in-depth analysis to illustrate how applying novel machine learning methods to transcriptomic datasets extends insights beyond the original transcriptomic analysis of the data. Further, we also show its generalizability through broader application to a suite of other datasets from GLDS (i.e., OSD-217; OSD-406; OSD-427).

Our analysis pipeline applies state-of-the-art representation learning models to find underlying patterns in the FPKM values (fragments per kilobase of transcript per million mapped fragments) that are proportional to the abundance of each locus's transcript. The application of representation learning models enhances the characterization of data points by capturing meaningful latent features, thereby improving downstream tasks such as clustering. This approach facilitates the grouping of genes with similar attributes, offering the possibility to reveal new and significant biological insights[7] and providing a robust foundation for analytical methodologies, such as further investigation of the effects of spaceflight on, e.g., phytohormone signaling and associated physiological phenotypes[8–10]. Moreover, considering that the CARA experiment also utilized lighting sub-environments, we can shine further light on the potential spaceflight effects and interactions with this factor that may have been overlooked in past studies. Overall, the GLARE method provides insights to better understand plant behavior in the spaceflight environment based on its endogenous and exogenous cues.

Results

Overview of GLARE (The GeneLAb Representation learning pipelinE)

We introduce GLARE, an analytical pipeline that combines data projection with representational learning and clustering to aid in discovering cryptic patterns within datasets deposited at the GenneLab data repository. We first present a broad overview of the pipeline, followed by a more detailed discussion of the results obtained from applying this approach to the CARA and other OSD datasets in the following sections.

We start by presenting a key first step, a verification study where we perform prediction tasks on restructured data. The verification study serves as a validity check for the approach prior to deploying the full GLARE. If the data did not exhibit any learnable and distinctive patterns between the experiment settings that we wanted to compare against, as revealed by making poor predictions on the test set in this verification study, then applying unsupervised methods on that data of interest would be ineffective, as the extracted representations would not capture meaningful latent information. The results of this verification study are detailed below, and the architectural specifications and implementation details are provided in the "Methods" section 'Verification Study'. Based on the positive results from this initial analysis, representation learning models are then implemented. These models can then empower researchers to move beyond conventional dimensionality reduction techniques in their omics-focused research, such as reliance on Principal Component Analysis (PCA) or t-distributed Stochastic Neighbor Embedding (t-SNE). This approach enables the extraction of data representations using a trained learning-based model, thereby allowing the exploration of latent structures to unveil the hidden patterns inherent within the dataset.

GLARE incorporates Sparse Autoencoder (SAE), a deep-learning-based model that enables efficient data compression while preserving salient features. In this way, it can capture intricate hierarchical structures within the data by simultaneously learning both compressed data representation and the features necessary for reconstruction[11,12]. To further enhance the utility of these representations for downstream tasks such as clustering, we introduce an additional self-supervised learning step, where the SAE model is pre-trained on unlabeled high-throughput single-cell data as a pretext task. We then take the pre-trained weights to fine-tune SAE using our normalized counts data to build Fine-Tuned SAE (FT-SAE). Researchers can readily select an appropriate single-cell dataset for this pre-training step depending on the target GLDS dataset to which they wish to apply GLARE. The suggested self-supervised learning step offers several advantages, including the augmentation of feature granularity and the incorporation of cellular-level insights, thereby enhancing the fidelity and relevance of the learned representations for downstream analyses[13]. The results of this representation learning-based approach are discussed in detail below, with specifications of its implementation provided in the Methods Section 'Learning Data Representations'.

After retrieving the learned data representation, GLARE employs an ensemble clustering scheme to improve upon the commonly used application of single clustering approaches. Ensemble clustering offers several advantages over relying on a single clustering algorithm. Notably, when working with complex data such as representations retrieved from a fine-tuned sparse autoencoder, ensemble clustering can effectively address inherent complexities to capture hidden patterns and discover biologically meaningful clusters[14]. GLARE adopts Evidence Accumulation Clustering (EAC)[15] as its ensemble clustering method, integrating three base clustering algorithms: Gaussian Mixture Models (GMM)[16], Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)[17], and Spectral clustering[18]. By merging the clustering outcomes based on distinct statistical backgrounds through consensus voting, followed by hierarchical clustering with average linkage on the generated co-association matrix, we can mitigate the biases and noises inherent in each base clustering method to create more robust and reliable clustering results. In addition to obtaining consensus cluster labels using EAC, researchers can leverage GLARE results from three base clustering algorithms to get unique clusters for each gene by retrieving the intersected cluster from its respective cluster assignments.

Lastly, we demonstrate the full utility of GLARE using the results derived from the ensemble clustering on learned data representation and perform post-pipeline analysis. We perform Gene Ontology (GO) analysis, in conjunction with clustering results, a widely used approach to find the functional significance of co-expressed genes in the clusters and provide a comprehensive understanding of the biological functions and processes underlying the observed gene expression patterns. Moreover, the prediction task that we undertook for the verification study serves as the foundation for feature explanation analysis, enabling the incorporation of feature importance explanation schemes, such as SHapley Additive exPlanations (SHAP)[19]. The feature importance values can reveal, e.g., within CARA, which genotypes and light conditions contributed the most to the predictions overall, as well as provide more insights into specific genes of interest. Combining these insights with the clustering results from the GLARE should substantially empower researchers in general to see new patterns in their omics-level data. The illustration of the overall pipeline and details of GLARE are shown in Fig. 1.

Verification study

The following sections present case study results using the CARA dataset (OSD-120) to demonstrate GLARE's analytical capabilities from initial data processing through downstream biological interpretation. The verification study starts with restructuring the typical GLDS processed normalized counts data from a wide format to a long format, a process often called 'data melting'[20]. This is a necessary process because many key elements of the experiment design are often combined into the text of the sample label when data is uploaded into a repository such as GeneLab. The data melting transformation involves reorganizing this uncategorized data so that each experimental factor (such as genotype, spaceflight versus ground control, and lighting regime) becomes a separate categorical variable, or a label, instead of embedding them in a complex column heading. This approach allows each aspect of the experiment environment to be explicitly indicated through the labels, facilitating subsequent automated analyses[21]. After melting the data, GLARE then trains a machine learning classification model on the restructured data using the concatenated feature vectors as the input matrix, and the new labels that represent the experiment environment

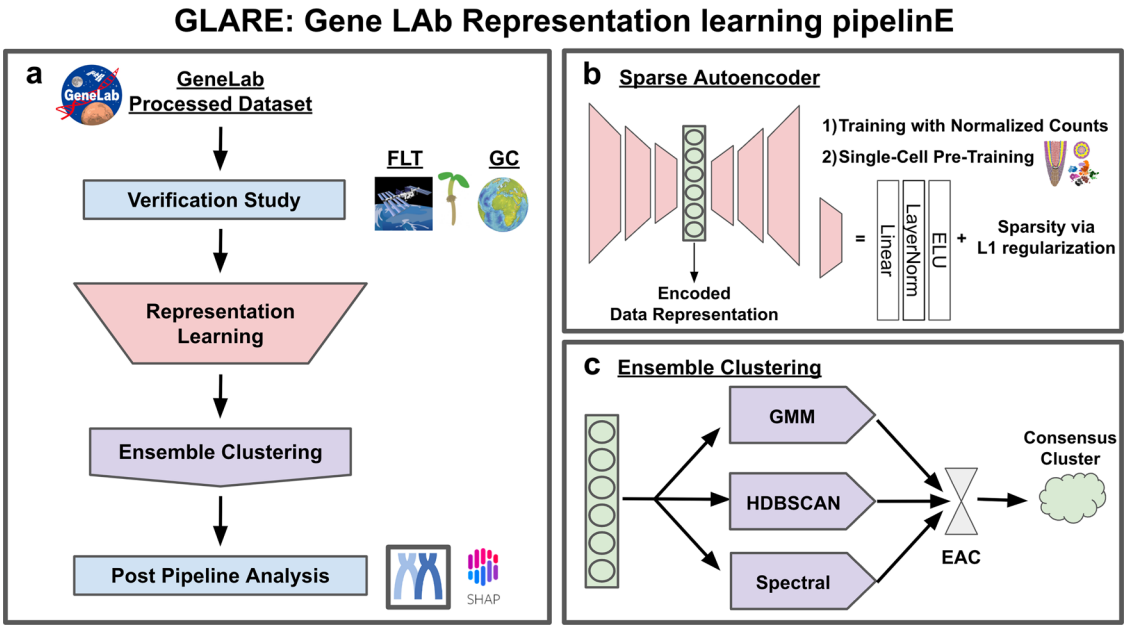


Fig. 1 | Overall pipeline of GLARE: Gene LAb Representation learning pipelinE. **a** Illustration of GLARE, starting with a verification study followed by representation learning and ensemble clustering. GLARE provides implementation of representation learning models, including the state-of-the-art SAE model pre-trained with high-throughput single-cell data. Retrieved data representation is then processed through ensemble clustering to find the hidden patterns within the data. Results from the verification study and ensemble clustering are then used for post-pipeline analysis. **b** Model architecture illustration of the employed SAE for both training with and without pre-training. **c** Ensemble clustering using three base clustering algorithms based on different statistical methodologies. Evidence accumulation clustering is used to derive consensus clusters from these algorithms.

as target labels. High predictive performance, such as accurately assigning a sample to flight versus ground control, would indicate that learnable patterns are indeed present in the original data, thus motivating the use of unsupervised representation learning techniques from GLARE.

Figure 2 shows an illustration of how we restructured the CARA dataset (OSD-120) through data melting, and the performance of the subsequent classification using XGBoost[22]. XGBoost was chosen through an ablation study using the CARA data (detailed in Supplementary Table 1). We compared prediction performances across multiple data melting models on a held-out test set of the restructured CARA data, which is presented in Table 1. This test set is a random subset of the data that was set aside during training to be used for evaluating the model's performance on unseen data. Data models that were tested include our 'base model', where we have location labels indicating if the experiments were performed in space (ISS) or on the ground (KSC). This model has 18 continuous features (e.g., genotype or lighting regime) and one label of flight versus ground. We also performed additional data melting to extract other experiment settings, such as 'Genotype' and 'Light condition', adding these additional labels for the modeling to use as further categorical predictors. We found that our base data model without other additional categorical variables yields the highest performance with ~91% test accuracy on predicting if the experiments were done in space or on the ground based on the normalized counts of FPKM values.

The ability of this model to make accurate predictions indicates there are latent patterns within the data. Encouraged by the outcome of this verification study, that a machine learning approach should be able to extract potentially novel features from the CARA spaceflight dataset, we applied GLARE. Prior to analysis with the full pipeline, we implemented a cluster-based outlier detection (detailed in Supplementary Fig. 1) to enhance the data quality. This preprocessing step identified and eliminated three anomalous genes (AT1G0759, AT3G41768, ATMG00020), thereby ensuring the robustness of our downstream analyses. In the following sections, we present an evaluation of generated data representations from using GLARE on the CARA dataset and subsequent analysis results.

Data representation evaluation

In this section, we compare data representations from different algorithms that could be retrieved from GLARE. Figure 3 shows visualizations of each of the representations of spaceflight (FLT) and the paired ground control (GC) from CARA data using PCA, t-SNE, Uniform Manifold APproximation (UMAP), SAE, and FT-SAE. Data representation from SAE and FT-SAE has n-dimensions depending on the number of neurons on the bottleneck layer. This value was determined through hyperparameter tuning, where various values of n were tested and evaluated based on their performance. The optimal value of $n = 16$ was selected based on the best reconstruction error observed during this tuning process. In order to present these SAE and FT-SAE data as figures, dimensionality was reduced to 2 using t-SNE. All other data representations from PCA, t-SNE, and UMAP already have a 2-dimensional matrix. On initial inspection, the PCA representations are highly condensed in a single region of the map, while t-SNE and UMAP representations exhibit a more widespread distribution. On the other hand, SAE and FT-SAE representations show more cluster-forming shapes for their t-SNE coordinates, where the locally condensed points are separated from others.

Although such an initial, qualitative visual analysis provides a useful overview of each technique's output, GLARE provides for a more quantitative interpretation. Table 2 shows this next element of the analysis, evaluating these data representations using multiple quantitative measures by examining local neighborhood preservation and performance on downstream tasks by calculating: (1) trustworthiness score, which measures how well the local structure of the high-dimensional data is preserved in the lower-dimensional representation by penalizing unexpected neighbors[23], (2) retention of class-discriminative information through k-nearest neighbors (KNN) classifier accuracy with cross-validation[24], and (3) quality of emergent clusters measured by the Silhouette Score[25]. Among the data representations that could be retrieved from GLARE, we compare all the methods that perform a non-linear transformation to the original dataset, leaving out PCA. To perform KNN classification and Silhouette score evaluation, we generated pseudo-labels ($n = 15$) from simple k-means clustering, i.e., we used k-means clustering to automatically categorize data

a

Loci	FLT_col_Light	FLT_col_Dark	GC_col_Light	…
AT1G01010	###	###	###	###
AT1G01020	###	###	###	###
AT1G01030	###	###	###	###

Loci	col_Light	col_Dark	…	WS_Dark	**Location**
AT1G01010	###	###	###	###	FLT
AT1G01010	###	###	###	###	GC
AT1G01020	###	###	###	###	FLT

b

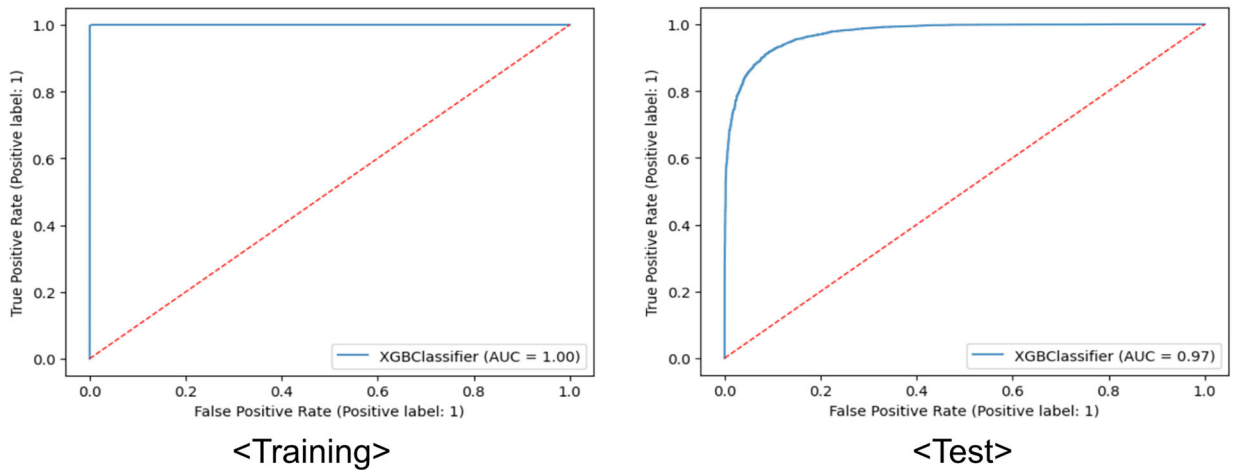


Fig. 2 | Data melting for data restructuring and supervised learning prediction performances. a Illustration showing the restructuring process of our base data model, where we make the experiment location a categorical variable (right-hand table, column label `Location'). Raw FPKM numerical data (denoted as ###) as reads per locus (e.g., AT1G01010, AT1G01020 across all ~25,000 genes in the Arabidopsis genome) for each experimental sample (e.g., FLT sample of Columbia ecotype grown in the light, FLT_col_Light, or GC sample of Columbia ecotype grown in the light, GC_col_Light) is shown in the left hand table. Each gene has one row in the data matrix, but each column label encodes multiple experimental factors (flight vs ground, genotype, and lighting regime). After restructuring (right table), each gene has two instances, one from the flight and one from the ground sample, separately. This facilitates subsequent analysis, asking, in this case, if flight versus ground is a key discriminator within the dataset. **b** Receiver Operating Characteristic (ROC) curves using the training and test datasets on the best-performing data model, which is the base model. Blue line represents the XGBoost classifier, showing the ratio of true positives to false positives in the model predictions from the training data (left) and the test set (right). Red line is the random chance baseline. FLT spaceflight, GC ground control.

Table 1 | Classification performances on held-out test set using XGBoost on data from different data models (with ±standard deviation via 5-fold cross-validation)

Data model	Test accuracy ↑	F1-score ↑	ROC-AUC ↑
Base data melting model	**91.29 ± 0.26**	**0.913 ± 0.002**	**0.975 ± 0.001**
+Genotype data melting	77.09 ± 0.14	0.770 ± 0.002	0.865 ± 0.001
+Light data melting	83.18 ± 0.07	0.831 ± 0.001	0.919 ± 0.001
+Genotype & Light data melting	68.69 ± 0.14	0.686 ± 0.001	0.772 ± 0.001

F1-score: The harmonic mean of precision (avoidance of false positives) and recall (avoidance of false negatives), ROC-AUC: Area under the Receiver Operating Characteristic curve, summarizing true positive vs. false positive trade-off. Test accuracy is the % correctness of predictions for classifying a sample as spaceflight or ground control in the test set using each data model.
Bold values indicate the best-performing results for each column, corresponding to individual evaluation metrics. Metrics marked with '↑' indicate that higher values correspond to better performance

into 15 groups. We trained a KNN classifier ($k = 500$ with cosine distance) on the training dataset with these pseudo-labels and measured prediction accuracy on the held-out test set using 5-fold cross-validation. Silhouette scores were computed using the same cluster labels.

Our results show that FT-SAE outperforms others on both of these measures, suggesting its robust capability to generate meaningful and discriminative representations. On the other hand, SAE shows the highest trustworthiness score for both FLT and GC. This result shows the advantages of non-linear, sparse representation for complex biological datasets and its effectiveness in preserving local data structures. While FT-SAE retains a relatively high trustworthiness score with >0.8[26], it has the lowest among others. This is likely due to the model's adaptation to features from the high-throughput single-cell dataset used in pre-training, which may not perfectly align with the local structure of the dataset used in the fine-tuning process. Specifically, the increased resolution of the single-cell data can induce clustering patterns that are not present in the CARA data, leading to discrepancies in how the model captures and represents the underlying structures of the dataset[27]. However, this trade-off appears to be beneficial overall, with FT-SAE showing better performance in downstream tasks.

Overall, this analysis indicates FT-SAE is our method of choice for further analyses based on its KNN and Silhouette scores, which rank top of all techniques, and its acceptably high trustworthiness value. This suggests that the use of non-linear, sparse representations from SAE, along with the introduction of the high-throughput single-cell dataset during pre-training, enhances the model's ability to capture local and global structures in the data. In this analysis, we focused on comparing the data representation from different microgravity environments, FLT and GC, based on the verification study results. Although the additional factors, such as light environment and genotype, were not explicitly separated during training the representation learning models, they remain present in the dataset (as in Fig. 2a) and can contribute to the learned latent structure. Therefore, we designed the post-pipeline analysis described below to allow us to investigate the sub-patterns

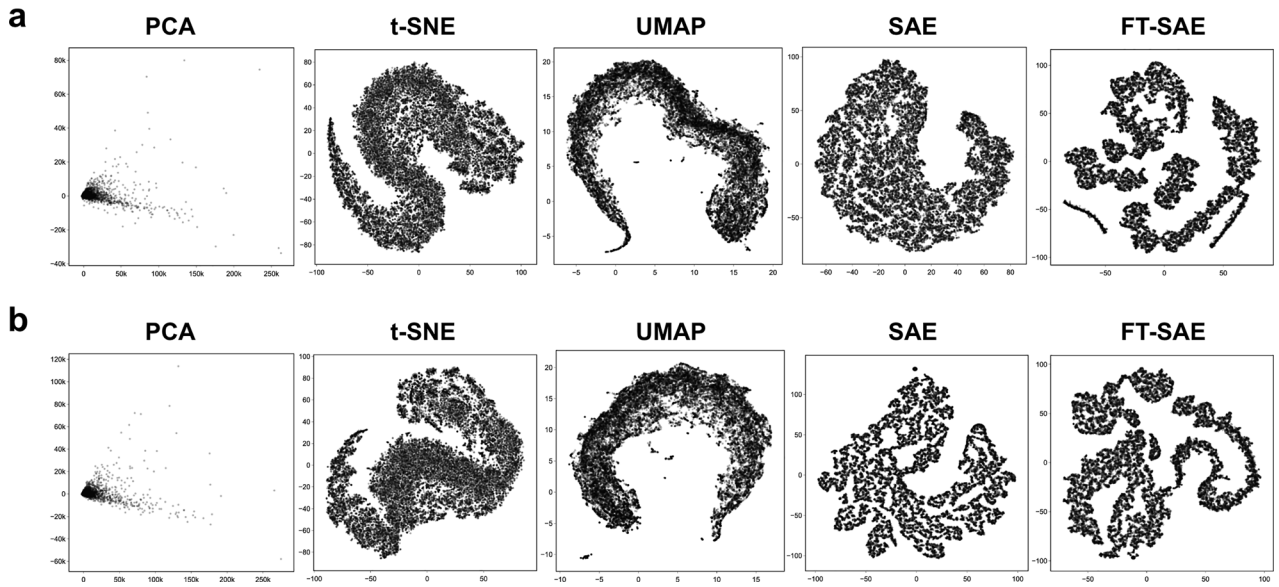


Fig. 3 | Comparison of data representations retrieved from GLARE. PCA, t-SNE, UMAP, SAE, and FT-SAE from left to right for both **a** Flight data (Upper-panel) and **b** Ground Control data (Lower-panel). t-SNE was used for the visualization of 16-dimensional data representation from SAE and FT-SAE.

Table 2 | Comparison of various evaluation metrics on raw data and data representations

		Evaluation metrics		
Environment	**Data representations**	**Trustworthiness score ↑**	**KNN accuracy ↑**	**Silhouette score ↑**
FLT	Raw Data	-	65.96 ± 0.39	0.5107
	t-SNE	0.942	67.67 ± 0.55	0.3842
	UMAP	0.943	81.36 ± 0.53	0.3985
	SAE	**0.967**	90.66 ± 0.36	0.4959
	FT-SAE	0.888	**96.08 ± 0.33**	**0.5635**
GC	Raw Data	-	62.48 ± 0.78	0.5047
	t-SNE	0.946	64.04 ± 0.82	0.3743
	UMAP	0.949	80.14 ± 0.54	0.3791
	SAE	**0.951**	93.80 ± 0.56	0.5634
	FT-SAE	0.870	**94.45 ± 0.45**	**0.5886**

FT-SAE shows the highest KNN accuracy (with ±standard deviation via 5-fold cross-validation) and the highest Silhouette score while having a lower trustworthiness score compared to others for both FLT and GC. The trustworthiness score is calculated with the raw data and the transformed data representation; therefore, it is not applicable for the raw data alone and is left blank.
Bold values indicate the best-performing results for each column, corresponding to individual evaluation metrics. Metrics marked with '↑' indicate that higher values correspond to better performance

associated with these variables. Moreover, GLARE's flexibility allows the user to reconfigure the dataset to focus on particular experimental factors of interest, enabling the discovery of both dominant and subtle biological patterns.

Ensemble clustering

Figure 4 shows ensemble clustering results using GLARE on the best-performing data representation of the CARA dataset, FT-SAE, with t-SNE used only for the purpose of visualization of the 16-dimensional data representation. We show individual clustering results from the base clustering algorithms we considered, GMM, HDBSCAN, and Spectral clustering, along with a final consensus cluster through evidence accumulation clustering (EAC)[15]. We note that GMM and spectral clustering require a user-defined cluster level. These were set to 20 and 25, respectively, for FLT and 25 and 20 for GC, driven by results from previous studies[28,29]. HDBSCAN defines its own cluster number.

Spaceflight. Clustering of the FLT dataset resulted in the identification of 20, 13, and 25 clusters for GMM, HDBSCAN, and Spectral clustering, respectively. GMM clusters had two large clusters, each containing 7623 and 5778 genes, with most of the other clusters having sizes of 300–1000 genes. HDBSCAN identified a smaller number of clusters, where most of the clusters had 1000–2500 genes. Spectral clusters had the most consistent cluster sizes compared to GMM and HDBSCAN, with most of the clusters having 1000–1300 genes. These results highlight how the precise nature of clusters is different depending on the clustering approach taken. Each clustering strategy has distinct strengths. GMM works well when the data does not have well-defined boundaries, where HDBSCAN is useful for datasets with noise and outliers, and spectral clustering is highly suited for data with non-linear manifold structures[30]. In order to leverage all of these advantages to a robust and reliable analysis of CARA data representation, we combined all three approaches via ensemble clustering through consensus voting[31]. These ensemble clusters exhibited diverse characteristics, having clusters with a size of <1000 genes to two large clusters, finding patterns in local structure, each containing 7627 and 4715 genes, similar to clusters identified by GMM. The number of clusters and size of remaining clusters, ranging from 1000 to 2500 genes, likely reflect the outputs of HDBSCAN and Spectral clusters, finding patterns throughout the global structure[32]. We evaluated the cluster quality using domain-driven metrics, measuring the average percentage of differentially expressed genes (DEGs) per valid cluster (excluding cluster size of >3000). Notably, ensemble clustering produced the highest

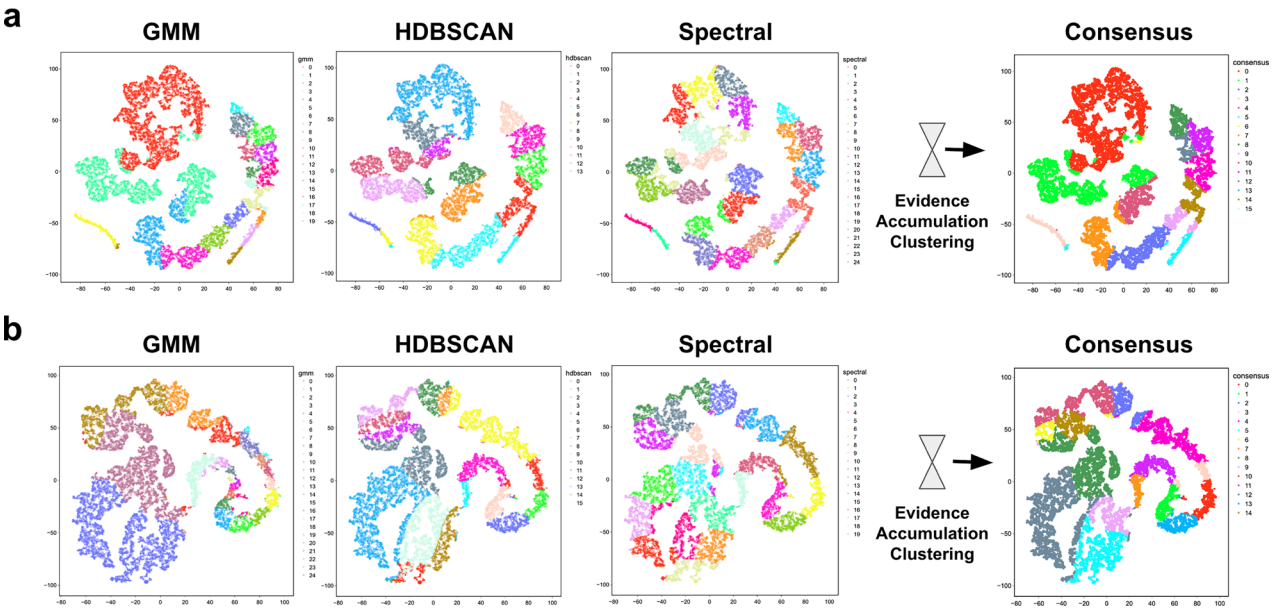


Fig. 4 | Ensemble clustering via EAC. Results from base clustering algorithms, GMM, HDBSCAN, and Spectral clustering, are shown starting from left to right for both **a** FLT FT-SAE representation (Upper-panel) and **b** GC FT-SAE representation (Lower-panel). t-SNE was used for the visualization of 16-dimensional data representation from FT-SAE. EAC results are shown at the right, with FLT having 16 consensus cluster labels and 15 consensus cluster labels for GC (depicted as different colors).

average percentage at 14.95%, outperforming individual methods such as baseline k-means (10.46%), GMM (12.93%), HDBSCAN (7.48%), and Spectral clustering (7.46%). This higher enrichment of DEGs highlights the likely biological relevance and robustness of the consensus approach (Ensemble clustering DEG proportions are detailed in Supplementary Data 3).

Ground control. Clustering of the GC dataset resulted in the identification of 25, 15, and 20 clusters for GMM, HDBSCAN, and Spectral clustering, respectively. Despite the slight change in the number of clusters, the qualitative characteristics of the results remained largely consistent with those obtained from FLT. Specifically, GMM results revealed two large clusters, each containing 7445 and 6157 genes for GC, along with most of the other clusters having lesser sizes of 200–900 genes, highlighting local patterns. Similarly, HDBSCAN and spectral clusters had a comparable, consistent number of genes as FLT clusters, finding patterns throughout the global structure. Ensemble clustering demonstrated similar outcomes to FLT as well, exhibiting a diverse range of gene counts within each cluster.

These identified clusters then form the starting point for an in-depth post-GLARE analysis for biological function as described for the CARA dataset in the following sections. However, GLARE is not limited to analyzing the CARA dataset, and to demonstrate that GLARE is a generally applicable workflow, we applied it to several GLDS datasets beyond CARA (OSD-120), including OSD-217[33], OSD-406[34], and OSD-427[35]. These datasets were all from experiments using Arabidopsis to take advantage of the pre-trained SAE developed for the CARA analysis described above. The notebook showing how to apply GLARE to these datasets is shared in the code repository (https://github.com/OpenScienceDataRepo/Plants_AWG/tree/main/Manuscript_Code/glare). The final ensemble clustering results from OSD-217, OSD-406, and OSD-427 are presented in Supplementary Figs. 2, 3, and 4.

Post-pipeline analysis: gene ontology analysis

For the post-pipeline analysis after retrieving the clustering results, we use the Metascape platform (http://metascape.org), which integrates various functional annotation databases[36] to perform GO enrichment analysis. We take the clusters from EAC on FT-SAE and process them through Metascape after excluding clusters with extreme sizes, as this tool can only take gene lists of less than 3000 counts for the enrichment analysis. Specifically, we exclude the two large clusters for both the FLT and GC datasets, along with one small cluster comprising only 2 genes in the FLT dataset, which leaves us 13 clusters for both the FLT and GC datasets. GO analysis on these clusters revealed various groups of ontology terms, including cellular metabolic processes, oxidative phosphorylation, light response and signaling, and vesicle-mediated transport (Supplementary Data 1).

The prior study on the CARA dataset[6] found that genes associated with cell wall metabolism seemed most prevalent among the differentially expressed genes. Our analysis demonstrates that both FLT and GC clusters, which contain a high proportion of differentially expressed genes, are enriched in terms related to "Protein synthesis, Energy metabolism, and Stress response pathways" and "Cell wall biogenesis and Intracellular transport". These results highlight pathways related to environmental adaptation and emphasize the likely critical role of cell wall metabolism identified in the CARA dataset. Moreover, we found unique hypoxia response-related clusters that were only prevalent in the FLT results and that were not highlighted in the original analysis of the CARA data. Root zone hypoxia is predicted to occur in spaceflight as a loss of buoyancy-driven convection in microgravity should limit oxygen resupply to intensely respiring tissues (e.g.,[37]). However, transcriptional fingerprints of hypoxia response in plants in spaceflight have often proven elusive. We therefore concentrated the focus of the rest of our analysis on this hypoxia-related cluster.

In Fig. 5, we show a heatmap for the FPKM values for the genes within the main hypoxia-related cluster, investigating all combinations of experimental factors (Fig. 5a), GO analysis results for this cluster using Metascape[3,6] (Fig. 5b), and Stress Knowledge Map (SKM)[38] centered around the Transcription Factors (TFs) in the cluster (Fig. 5c). The Stress Knowledge Map (SKM; https://skm.nib.si/) is a curated resource offering two types of knowledge graphs on plant molecular interactions and stress signaling[38]. We used one of these, the Comprehensive Knowledge Network (CKN), to gain insights into potential stress signaling and associated plant biological processes around our genes of interest. The map in Fig. 5c was drawn with the five TFs that we found in the 43 gene hypoxia-related cluster: *DREB2A*, *RHL41/ZAT12*, *MYC2*, *RRTF1/ERF109*, and *STZ/ZAT10*. The CKN map shows an intricate network of TFs and their interactions in the context of

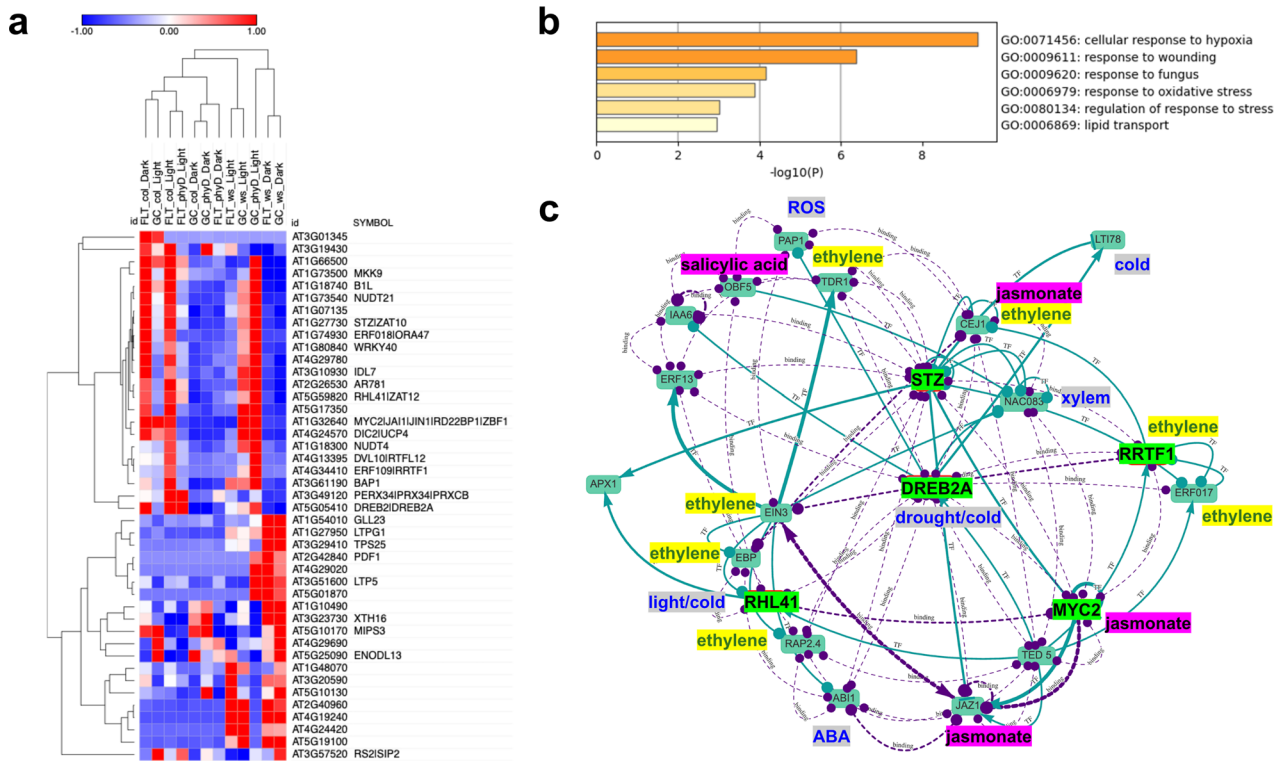


Fig. 5 | Analysis of hypoxia-related cluster found in the FLT clustering result. **a** Heatmap of normalized FPKM values on the cluster. **b** Enriched GO terms on the cluster from Metascape. **c** Stress Knowledge Map (SKM) on five Transcription Factors (TFs; green highlight) in hypoxia-related cluster: *`DREB2A'*, *`RHL41/ZAT12'*, *`MYC2'*, *`RRTF1/ERF109'*, and *`STZ/ZAT10'*. Purple lines: physical interaction (binding), teal lines: TF interactions. Annotations of relationships to Abscisic acid (ABA), Ethylene, Jasmonate, Light/cold, Reactive oxygen species (ROS), Salicylic acid, and Xylem are from manual inspection of each gene's annotated roles in The Arabidopsis Information Resource (https://www.arabidopsis.org).

stress response mechanisms and related signaling pathways with other genes such as *HY5/TED5*, *ABI1*, and *JAZ1*. Inspection of this network reveals ethylene as a likely important player in this response. GO analysis of the network for biological function (Supplementary Data 2) indicates elements of defense, water stress, and cold response may also be important features for further study.

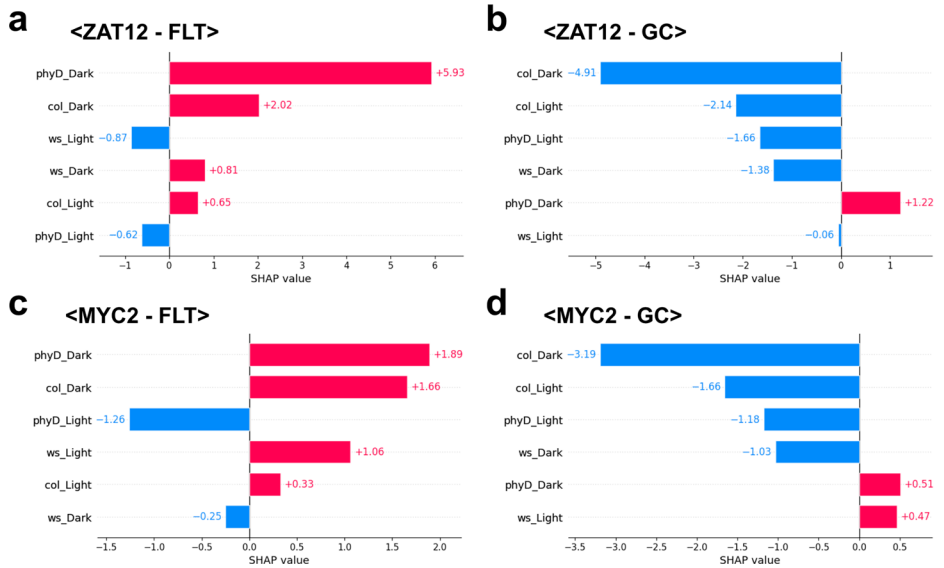
Additionally, we identified that the hypoxia-related cluster had the highest percentage of differentially expressed genes within the cluster (DEG proportions in Supplementary Data 3). These genes were then used for cell-type prediction using scPlantDB[39] to highlight potentially responding cells in different root organ groups, taking advantage of the single-cell pre-trained model FT-SAE of GLARE. We perform this cell-type prediction for all the clusters as well (Supplementary Data 3). The hypoxia-related cluster had more than half of its differentially expressed genes identified in the lateral root cap and columella of the root cap, while also appearing in phloem parenchyma and xylem. These results reveal a distinct enrichment of hypoxia response-related genes in root cap tissues, while also highlighting the involvement of the root vasculature, suggesting a potentially coordinated response to low oxygen conditions across root structures.

Post-pipeline analysis: SHAP analysis

Up to this point, our analysis has been directed toward uncovering distinct patterns, specifically between FLT and GC, by generating separate data representations for clustering and GO analysis. However, we chose CARA as a dataset to interrogate in depth due to the multiple experimental factors within the experiment's design. Therefore, after identifying the patterns within the FLT data using GLARE, particularly those related to hypoxia, we used this newly identified cluster to evaluate the effect of varying light conditions on different genotypes in each location. We took the found TFs within the hypoxia-related cluster and applied SHAP analysis to quantify feature contribution, to define which experimental conditions had the most effect in classifying this pattern within the data between FLT and GC. SHAP analysis provides a way to understand the impact of each feature on the model's predictions, enabling better model transparency and insights into the underlying relationships within the data[19]. Higher positive SHAP scores reflect features contributing more to this discrimination within the dataset to designate a sample to FLT, while negative values reveal factors that have a negative impact on the FLT assignment, i.e., reveal the data as GC.

In Fig. 6e, we show the difference in SHAP values for the TFs within the hypoxia-related cluster. Among these five TFs, *ZAT12* has the largest aggregate difference in SHAP values between FLT and GC, and *MYC2* the smallest. In Fig. 6a–d, we show SHAP local bar plots for *ZAT12* and *MYC2* explaining the feature importance representing different environmental conditions and genotypes. For both *ZAT12* and *MYC2*, the *PHYD* mutants and Col-0 genotype in the dark setting had the most contribution to model prediction in FLT (Fig. 6a, c). On the other hand, both the dark and light settings for the Col-0 genotype had the most contribution to model prediction in GC (Fig. 6b, d). The large difference in aggregated SHAP value between FLT and GC for *ZAT12* suggests that the relative importance and contributions of these features vary between the FLT and GC. In contrast, the contributions for *MYC2* appear more consistent and stable across both FLT and GC classifications. This analysis using SHAP values provides deeper insight into the differentiation between environmental conditions and genotypes at different locations, solely based on gene expression data.

Lastly, in Fig. 7, we present summary SHAP plots on these features, varying light conditions on different genotypes, to offer a more comprehensive understanding of feature contribution across the entire dataset. We can observe features with different degrees of impact on the model's prediction from the SHAP value scatterplot in Fig. 7a. For example, for the WS genotype in a light setting, the majority of the data aligns with positive SHAP values, supporting FLT classification, whereas under dark conditions, the trend is reversed. The beeswarm plot (Fig. 7b) illustrates the distribution of



e

TFs	Difference in SHAP value between FLT / GC
ZAT12	16.842
ZAT10	16.472
ERF109	13.964
DREB2A	9.606
MYC2	9.51

Fig. 6 | SHAP analysis on Transcription Factors (TFs) in the hypoxia-related cluster. A positive SHAP value (Red color) means that the feature value made a greater contribution than others in classifying the gene as FLT, while a negative SHAP value (Blue color) suggests they had more contribution in GC classification. **a** *ZAT12* - FLT **b** *ZAT12* - GC **c** *MYC2* - FLT **d** *MYC2* - GC. **e** Summary of the difference in SHAP value between FLT and GC for the 5 TFs in hypoxia.

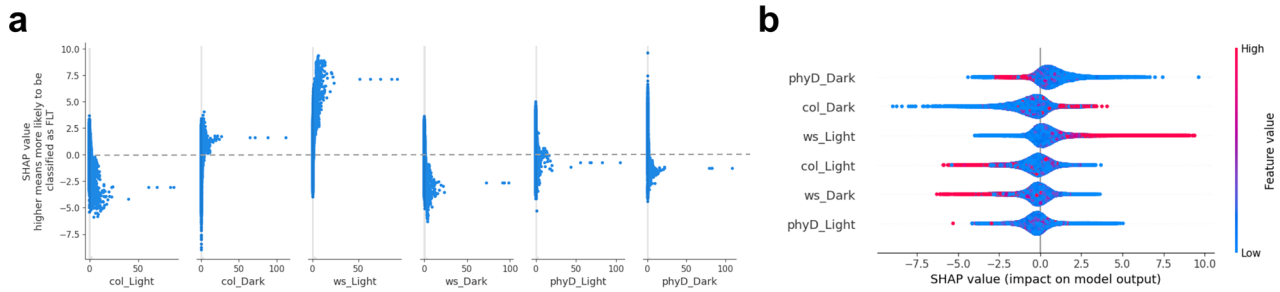


Fig. 7 | SHAP value distribution for each treatment. Comparing SHAP values from a classification using the XGBoost on the restructured CARA dataset. **a** The summary SHAP value scatterplot for each feature displays the distribution of SHAP values alongside raw feature values. **b** The summary SHAP beeswarm plots, where features are ordered by their importance (measured by mean absolute SHAP values), with the most impactful features appearing at the top. The color bar represents raw feature values. Both plots present the same information.

SHAP values for each feature as well. The color gradient from blue to red represents the feature value (FPKM values), with blue indicating low expression and red indicating high expression. Figure 7b illustrates that *PHYD* mutants in a dark setting have the highest effect on the classification with longer tails towards positive SHAP value, while most of the high FPKM values have negative SHAP value. Suggesting that high expression levels from *PHYD* mutants in dark settings decrease the likelihood of FLT classification. Similarly, the Col genotype in a dark setting has tails toward negative SHAP values, while most of the high FPKM values have positive SHAP values. These observations emphasize the presence of intricate interactions between gene expressions, reflecting the complexity of the transcriptome data and the underlying biological mechanisms.

Such SHAP analysis offers a unique perspective on the patterns within the dataset, particularly when the data comprises various environmental settings as features. Through analyzing the variations and similarities in SHAP values, researchers can identify genes that are sensitive to complex environments based on genotype-dependent patterns in the data.

Discussion

In this study, we present a deep-learning-based analysis pipeline for spaceflight biology research, GLARE. GLARE incorporates widely adopted methods such as autoencoders, clustering, and SHAP analysis, yet its novelty lies in the integration of these components into a unified and extensible pipeline tailored for transcriptomic analysis. Although GLARE should be applicable to any transcriptomic datasets, the approach is well-suited to address major challenges in the field of space biology arising from various reasons, such as the limited availability of experimental samples, complex experiment design, and subtle transcriptional responses that are difficult to identify using conventional approaches. GLARE offers a blueprint for discovering latent biological signals in complex spaceflight datasets, serving as a foundational model for future spaceflight -omics analysis pipelines.

We chose to demonstrate the use of GLARE with a previously analyzed dataset, the CARA experiment (OSD-120)[6] which allows for an investigation of the overall utility of the pipeline itself and a comparison with the prior findings. For analysis of the root samples in the CARA spaceflight data, we pre-trained the model using high-throughput plant root single-cell data and fine-tuned it using the restructured CARA data, along with ensemble clustering, to identify hidden patterns in the spaceflight transcriptome. For other spaceflight datasets, such as whole seedlings, shoot tissues, microbe, animal tissues, or cell types, matching pre-training datasets to the particular experimental design would similarly add significant depth to the analyses.

After applying the full pipeline, the addition of a post-pipeline analysis employed select bioinformatics tools and incorporated post hoc interpretation of sub-patterns associated with experimental factors other than spaceflight by applying SHAP analysis. Such analyses confirmed previously observed spaceflight-related patterns in the data, such as clusters of transcripts involved in cell wall remodeling and vesicle-mediated transport.

Critically, this approach also revealed new features, notably a molecular signature associated with the hypoxia response in the spaceflight samples that was otherwise challenging to extract from the plant transcriptomic dataset. Moreover, our analyses revealed that this cryptic signature was dependent on experimental conditions such as plant genotype and lighting regime. For example, Fig. 6 shows that SHAP analysis of the 5 signature spaceflight-related, hypoxia response-related transcription factors identified in this study highlights the complex interplay of genotype and lighting conditions, demonstrating the value of the application of the machine learning-powered GLARE in uncovering subtle molecular signaling and hidden biological patterns.

We presented this in-depth post-pipeline analysis as a guideline to fully utilize the output of GLARE, yet researchers can also leverage their preferred analytics tools when applying GLARE to their datasets to uncover patterns. To this end, we actively encourage contributions and novel suggestions through our open science repository (https://github.com/OpenScienceDataRepo/Plants_AWG/tree/main/Manuscript_Code/glare). Its open-source nature means researchers can readily adapt GLARE to other datasets from GeneLab to reinforce their initial studies and expand on such computational findings. Nonetheless, limitations of our current approach and the recent rapid advancement in the machine learning field warrant future work on GLARE. Single-cell transcriptomes are becoming a powerful tool to dissect biological responses, and we have capitalized upon this potential by using single-cell datasets in a pre-training step for the FT-SAE model. This single-cell resolution provides key new insight, but a potential distribution shift during training through a domain gap between the pre-training dataset (single-cell gene expression) and the target dataset (bulk RNA-seq) may introduce reduced performance in downstream tasks and bias in further fine-tuning steps[40]. While our application of an adapter layer before the fine-tuning step and induced sparsity from SAE can contribute to lessening this negative effect[41,42], we mark this challenge as an important direction for future development of GLARE, such as implementing the use of domain adaptation techniques[40].

Despite this intrinsic limitation, integrating single-cell datasets has begun to be widely adopted for their advantage in providing nuanced insights at the cellular level. Indeed, transformer-based foundation models for single-cell multi-omics have been suggested[43], which offer the potential to generate synthetic data or for gene network inference. Our future vision for GLARE is to extend beyond autoencoder-based models to add more advanced self-supervised representation learning models, such as the contrastive learning methods that are well-used in the field of computer vision and natural language processing[44], to enhance robustness for smaller datasets with fewer features and help address distribution shifts. Additionally, causal representation learning methods offer an exciting avenue to discover the causal structure within the data and study the relationship between particular genes[45,46].

Methods

GeneLab Data System and data entries

The Genelab Data System (GLDS) is a public, space-related -omics data repository, which curates data from a wide variety of species and experimental spaceflight conditions[47]. GLDS obtains spaceflight-related –omics datasets from multiple locations, such as the Gene Expression Omnibus (GEO), European Bioinformatics Institute (EBI), publications directly, and others[47]. This data is then cataloged with the relevant metadata, such as protocols, payload numbers, and experimental variables, and made available as an Open Science Dataset (OSD) in NASA's Open Science Data Repository (OSDR).

The CARA dataset (OSD-120; https://osdr.nasa.gov/bio/repo/data/studies/OSD-120) was chosen from the GLDS to use as a test case for analysis using GLARE due to its many experimental conditions. The CARA experiments were conducted with three ecotypes/genotypes of *Arabidopsis thaliana*: wild-type Wassilewskija (WS), wild-type Columbia-0 (Col-0), and a mutant in the *PHYTOCHROME D* gene in the Col-0 background (*PHYD*)[6]. Briefly, these genotypes were planted on gel media in Petri dishes and grown in either ambient light conditions or in the dark on the ISS for 11 days; Parallel controls were performed on the ground. After the 11 days, germinated seedlings were photographed and collected into Kennedy Space Center Fixation Tubes (KFTs;[48]) containing RNAlater. Seedlings preserved in RNAlater were returned to Earth frozen, and then the roots were dissected into the last 2 mm of the tip for the light-grown plants and the last 1 mm for the dark-grown plants. RNA was extracted and sent to the Interdisciplinary Center for Biotechnology Research (ICBR), University of Florida, for RNA sequencing using a NextSeq 500 system, producing ~40 million paired-end reads per sample. Finally, these paired-end reads were mapped to the TAIR10 *A. thaliana* reference genome using Spliced Transcripts Alignment to a Reference (STAR) software, and differential expression was performed using the Cufflinks tool[49,50].

Verification study

A verification study was performed on the CARA dataset (OSD-120) to ensure that learnable patterns indeed exist within this spaceflight transcriptome dataset before using it with the GLARE. To enable this analysis, we restructured the original OSD-120 dataset from a wide format to a long format, a process often called 'data melting'[20]. We extracted feature vectors representing each experimental condition and restructured the numerical FPKM data to be indexed by both gene locus and individual experimental factor labels rather than just by gene locus (illustrated in Fig. 2a). Essentially, restructured data, X_{melted}, becomes organized to facilitate analysis across multiple experimental dimensions while containing the same information as the original data, X. This original data has 36 numerical features as it contains experiment results from two locations (spaceflight versus ground), under two different light conditions (light versus dark), and for three genotypes (Ws, Col-0, and *PHYD* mutants), with three replicate samples of each. The modeling for classification tasks was performed using the Python libraries 'sklearn' and 'xgboost', applying a set of supervised learning models, such as logistic regression, random forest, KNN classifier, support vector machine, and XGBoost. The ablation study to choose the best-performing model is detailed in Supplementary Table 1. The held-out test set was 20% of the CARA dataset chosen at random via 5-fold cross-validation and not used in model training.

High-dimensional data analysis

Overview. Statistical methods have been widely integrated into the bioinformatics pipelines in multi-omics studies for analyzing the data. Specifically, due to multi-omics datasets having complex data topology, dimension reduction and clustering are two commonly used techniques for further investigation[51]. GLARE capitalizes upon such approaches. For example, Principal Component Analysis (PCA) and Factor Analysis are methods with widespread application for dimensionality reduction[52]. After achieving a statistical representation of the dataset with these dimensionality reduction techniques, clustering methods are utilized to group similar representations to uncover underlying patterns within the dataset. Among these, K-means and hierarchical clustering are featured as two of the most favored methodologies[53].

Learning data representations

While PCA is popularly used for its simplicity and efficiency, it has its limits for losing essential non-linear features through linear embedding, which often degrades the clustering quality[54]. Several alternative methods that do not only rely on data point distribution but also leverage latent data structures via learned representations have shown advantages in handling biological data, thereby enhancing clustering precision[55]. These alternatives to PCA include t-distributed Stochastic Neighbour Embedding (t-SNE), a non-linear dimensionality reduction technique particularly adept at preserving local structures within high-dimensional data, and Uniform Manifold APproximation (UMAP)[56,57], a manifold learning approach that efficiently captures complex relationships within the data.

However, alternative deep-learning-based approaches for obtaining data representations have been largely neglected in the field of plant space biology, despite the advantage of their ability to capture contextual

information from the non-linear mappings. Specifically, this approach of capturing contextual information through complex, higher-level features is known as representation learning or feature extraction[58]. Therefore, along with PCA, t-SNE, and UMAP, we have investigated the application of Sparse Autoencoder (SAE) as one of the representation learning methods in GLARE. SAE is an unsupervised learning algorithm based on a neural network that aims to learn an approximation of the identity function that represents the data. Unlike standard autoencoders, SAEs incorporate sparsity constraints during training, keeping most neurons inactive for a given input. The model is trained through an encoding-decoding process with sparsity constraints applied to the encoding phase, promoting only a small subset of neurons to be activated, allowing the discovery of the unique underlying structure in the data[59]. The sparsity can be implemented through regularization techniques such as L1-regularization. The decoder then reconstructs the original data from the encoded representation by optimizing a reconstruction loss. This optimization ensures that the model captures the most essential patterns in the data representation. While autoencoders are more commonly used for reconstructing the original input data to generate new output data, prior studies show autoencoders as a representation learning approach to learn meaningful latent representations that work favorably in the context of multi-omics datasets[60].

Our implementation of SAE is constructed with a sequence of building blocks, each comprising a Linear layer followed by LayerNorm and Exponential Linear Unit (ELU) activation. We chose to add the LayerNorm block to improve convergence and stable optimization, considering that most GLDS data consists of multiple experimental results from different environment settings[61]. Towards this matter, we employ ELU activation as well[62]. We use three of these building blocks for the encoder and three blocks for the decoder to make the SAE. We choose L1-regularization over L2-regularization to induce sparsity effectively and for better implicit feature selection, addressing the sparse and heterogeneous nature of normalized counts of FPKM values[63]. The sparsity penalty parameter for L1-regularization was set to $1e-5$, determined through empirical experimentation for hyperparameter tuning. The model training is optimized using mean squared error loss for reconstruction, Adam optimizer[64] with weight decay, early stopping, and gradient clipping to address exploding gradients and ensure stable training. These hyperparameters were tested to find optimal parameter sets, which are described in our shared code repository (https://github.com/OpenScienceDataRepo/Plants_AWG/tree/main/Manuscript_Code/glare).

Prior to training SAE using the target GLDS data, GLARE implements a pre-training step that complements the representations from the model by incorporating detailed single-cell transcriptome profiling. Similar to our approach, such building of foundation models pre-trained with high-throughput single-cell data has demonstrated great utility in a diverse array of tasks in the life science field, including pattern recognition by incorporating foundational knowledge of the data[65]. Specifically, a single-cell root transcriptome dataset from Shulse et al.[28] was used for our analysis, as the CARA dataset is drawn from root tip samples. Researchers can select an appropriate single-cell dataset to pre-train the SAE relevant to the GLDS data they plan to study. Such datasets are publicly available in the Single Cell Expression Atlas (https://www.ebi.ac.uk/gxa/sc/home), which includes data from 21 different species. Following the pre-training step, the SAE model is initialized with the learned weights and subsequently fine-tuned using the target GLDS data. We maintain the original model structure and introduce adapter layers atop the main model to adjust varying dimensions between the single-cell matrix and our data appropriately, ensuring seamless integration into our SAE framework. Finally, after the full model training, we extract data representation from the bottleneck layer between the encoder and decoder using this optimized model.

Upon employing multiple approaches to obtain data representation, evaluating these representations is critical to understanding the strengths and limitations of various data representation techniques. Prior research has used several evaluation techniques to assess the fidelity between data representations and the original dataset and the quality of the data representation structure, so we have used these methods in the development of the GLARE approach. Trustworthiness score measures the preservation of local topological structure in the data and is widely applied to evaluate the fidelity and faithfulness of the learned representation, testing its ability to maintain local structure and inherent relationships[23,66]. On the other hand, evaluating the efficacy of these data representations in potential downstream tasks, such as classification or clustering, provides crucial insights into their practical utility and generalizability. K-Nearest Neighbors (KNN) classifier can be utilized to test the quality and separability of the data structure in reduced dimension by assessing the class-discriminative information[24]. Furthermore, the Silhouette score is widely used to check the insights into clustering performance and compactness of the data representation[25]. These metrics would help determine when it's appropriate to use the data representations for specific downstream tasks.

Clustering data representations

Within the clustering paradigm, several alternative methods to K-means exist for the effective organization of these representations, and we have explored their application as part of the GLARE. Among these, Gaussian Mixture Models (GMM) with the Expectation-Maximization (EM) algorithm offer a probabilistic framework, wherein each cluster is represented by a Gaussian distribution, facilitating more nuanced cluster assignments[16]. Density-based clustering methods have gained attention with respect to their ability to detect clusters of arbitrary shapes and sizes, thus overcoming some of the limitations associated with distance-based methods[67]. Notably, an extension of this approach, Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), utilizes a hierarchical approach to density-based clustering to robustly identify clusters at multiple levels with varying densities[17]. Additionally, spectral clustering presents an alternative approach, leveraging the eigenstructure of the similarity matrix to partition the data into clusters, thereby offering an effective means of characterizing complex structures within the dataset[18].

Ensemble clustering is an additional powerful technique that combines these multiple clustering solutions to obtain consensus clusters that are more robust and accurate. Several ensemble clustering methods have been proposed, including Evidence Accumulation Clustering (EAC)[15], which accumulates evidence from different base clustering algorithms to build a co-association matrix. Applying hierarchical clustering to this matrix derives a final consensus clustering result. Other notable examples include the HyperGraph-Partitioning Algorithm (HGPA)[68], which derives consensus clustering through a partitioning hypergraph where each base clustering set is a hyperedge in a hypergraph, with vertices representing data points. These ensemble techniques have demonstrated their utility in various domains, such as bioinformatics, text mining, and computer vision, where data is often high-dimensional, noisy, and complex[31]. Therefore, they are strong candidates for equivalent analyses of the often highly complex structures that make up plant transcriptomics datasets. Hence, these approaches are incorporated into the GLARE.

Data availability

The transcriptome datasets utilized in this study (OSD-120, OSD-217, OSD-406, and OSD-427) can be found on the NASA Open Science Data Repository (https://osdr.nasa.gov/bio/repo/). High-throughput single-cell transcriptome data (https://doi.org/10.1016/j.celrep.2019.04.054) used for pre-training the sparse autoencoder can be accessed at (https://www.ebi.ac.uk/gxa/sc/experiments/E-CURD-5/downloads) in the Normalized counts files folder.

Code availability

The underlying code for the presented analysis pipeline for this study can be found on this publicly available GitHub repository (https://github.com/OpenScienceDataRepo/Plants_AWG/tree/main/Manuscript_Code/glare).

Received: 24 January 2025; Accepted: 31 August 2025;
Published online: 28 October 2025

References

1. Mustroph, A. et al. Cross-kingdom comparison of transcriptomic adjustments to low-oxygen stress highlights conserved and plant-specific responses. *Plant Physiol.* **152**, 1484–1500 (2010).
2. Rutter, L. et al. A new era for space life science: international standards for space omics processing. *Patterns* **1** (2020).
3. Fu, Y. et al. How to establish a bioregenerative life support system for long-term crewed missions to the moon or Mars. *Astrobiology* **16**, 925–936 (2016).
4. Paul, A.-L., Zupanska, A. K., Schultz, E. R. & Ferl, R. J. Organ-specific remodeling of the arabidopsis transcriptome in response to spaceflight. *BMC plant biology* **13**, 112 (2013).
5. Villacampa, A. et al. From spaceflight to Mars g-levels: adaptive response of A. thaliana seedlings in a reduced gravity environment is enhanced by red-light photostimulation. *Int. J. Mol. Sci.* **22**, 899 (2021).
6. Paul, A. -L. et al. Genetic dissection of the Arabidopsis spaceflight transcriptome: are some responses dispensable for the physiological adaptation of plants to spaceflight? *PLoS ONE* **12**, e0180186 (2017).
7. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868 (1998).
8. Abts, W., Vandenbussche, B., De Proft, M. P. & Van de Poel, B. The role of auxin-ethylene crosstalk in orchestrating primary root elongation in sugar beet. *Front. Plant Sci.* **8**, 444 (2017).
9. Ferl, R. J. & Paul, A. -L. The effect of spaceflight on the gravity-sensing auxin gradient of roots: GFP reporter gene microscopy on orbit. *npj Microgravity* **2**, 1–9 (2016).
10. Iqbal, N. et al. Ethylene role in plant growth, development and senescence: interaction with other phytohormones. *Front. Plant Sci.* **8**, 475 (2017).
11. Ng, A. et al. Sparse autoencoder. *CS294A Lect. Notes* **72**, 1–19 (2011).
12. Ranzato, M. et al. Sparse feature learning for deep belief networks. *Adv. Neural Inf. Process. Syst.* **20** (2007).
13. Kiselev, V. Y., Yiu, A. & Hemberg, M. scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods* **15**, 359–362 (2018).
14. Monti, S., Tamayo, P., Mesirov, J. & Golub, T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.* **52**, 91–118 (2003).
15. Fred, A. L. & Jain, A. K. Combining multiple clusterings using evidence accumulation. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**, 835–850 (2005).
16. Reynolds, D. Gaussian mixture models. In *Encyclopedia of biometrics*, 827–832 (Springer, 2015).
17. Campello, R. J., Moulavi, D. & Sander, J. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* 160–172 (Springer, 2013).
18. Ng, A., Jordan, M. & Weiss, Y. On spectral clustering: analysis and an algorithm. *Adv. Neural Inf. Process. Syst.* **14** (2001).
19. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* Vol. 30 (ed. Guyon, I. et al.) 4765–4774 (Curran Associates, Inc., 2017).
20. Wickham, H. Reshaping data with the reshape package. *J. Stat. Softw.* **21**, 1–20 (2007).
21. Xu, C., Tao, D. & Xu, C. A survey on multi-view learning. Preprint at https://arxiv.org/abs/1304.5634 (2013).
22. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, 2016).
23. Venna, J. & Kaski, S. Neighborhood preservation in nonlinear projection methods: an experimental study. In *International Conference on Artificial Neural Networks* 485–491 (Springer, 2001).
24. Van Der Maaten, L. et al. Dimensionality reduction: a comparative review. *J. Mach. Learn. Res.* **10**, 13 (2009).
25. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
26. Lee, J. A. et al. *Nonlinear Dimensionality Reduction*, Vol. 1 (Springer, 2007).
27. Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22**, 1345–1359 (2009).
28. Shulse, C. N. et al. High-throughput single-cell transcriptome profiling of plant cell types. *Cell Rep.* **27**, 2241–2247 (2019).
29. Shahan, R. et al. A single-cell Arabidopsis root atlas reveals developmental trajectories in wild-type and cell identity mutants. *Dev. Cell* **57**, 543–560 (2022).
30. Xu, D. & Tian, Y. A comprehensive survey of clustering algorithms. *Ann. Data Sci.* **2**, 165–193 (2015).
31. Vega-Pons, S. & Ruiz-Shulcloper, J. A survey of clustering ensemble algorithms. *Int. J. Pattern Recognit. Artif. Intell.* **25**, 337–372 (2011).
32. Jain, A. K. Data clustering: 50 years beyond k-means. *Pattern Recognit. Lett.* **31**, 651–666 (2010).
33. Zhou, M., Sng, N. J., LeFrois, C. E., Paul, A. -L. & Ferl, R. J. Epigenomics in an extraterrestrial environment: organ-specific alteration of DNA methylation and gene expression elicited by spaceflight in Arabidopsis thaliana. *BMC Genom.* **20**, 1–17 (2019).
34. Califar, B. et al. Shared metabolic remodeling processes characterize the transcriptome of Arabidopsis thaliana within various suborbital flight environments. *Gravit. Space Res* **9**, 13–30 (2021).
35. Paul, A. -L., Haveman, N., Califar, B. & Ferl, R. J. Epigenomic regulators elongator complex subunit 2 and methyltransferase 1 differentially condition the spaceflight response in Arabidopsis. *Front. Plant Sci.* **12**, 691790 (2021).
36. Zhou, Y. et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* **10**, 1523 (2019).
37. Porterfield, D. M. The biophysical limitations in physiological transport and exchange in plants grown in microgravity. *J. Plant Growth Regul.* **21**, 177–190 (2002).
38. Bleker, C. et al. Stress knowledge map: A knowledge graph resource for systems biology analysis of plant stress responses. *Plant Commun.* **5**, 100920 (2024).
39. He, Z. et al. scplantdb: a comprehensive database for exploring cell types and markers of plant cell atlases. *Nucleic Acids Res.* **52**, D1629–D1638 (2024).
40. Li, J., Yu, Z., Du, Z., Zhu, L. & Shen, H. T. A comprehensive survey on source-free domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **46**, 5743–5762 (2024).
41. Malik, B., Ramesh Kashyap, A., Kan, M.-Y. & Poria, S. UDAPTER - efficient domain adaptation using adapters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2249–2263 (Association for Computational Linguistics, Dubrovnik, Croatia, 2023).
42. Parhi, R. & Nowak, R. D. Deep learning meets sparse regularization: a signal processing perspective. *IEEE Signal Process. Mag.* **40**, 63–74 (2023).
43. Cui, H. et al. scgpt: toward building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods* **21**, 1470–1480 (2024).
44. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning* 1597–1607 (PMLR, 2020).
45. Komanduri, A., Wu, X., Wu, Y. & Chen, F. From identifiable causal representations to controllable counterfactual generation: A survey on causal generative modeling. *Transactions on Machine Learning Research* (2024).
46. Schölkopf, B. et al. Toward causal representation learning. *Proc. IEEE* **109**, 612–634 (2021).
47. Ray, S. et al. GeneLab: Omics database for spaceflight experiments. *Bioinformatics* **35**, 1753–1759 (2018).

48. Ferl, R. J. et al. The performance of ksc fixation tubes with RNALater for orbital experiments: a case study in ISS operations for molecular biology. *Adv. Space Res.* **48**, 199–206 (2011).
49. Dobin, A. et al. Star: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
50. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with tophat and cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
51. Rappoport, N. & Shamir, R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Res.* **46**, 10546–10562 (2018).
52. Zeng, I. S. L. & Lumley, T. Review of statistical learning methods in integrated omics studies (an integrated information science). *Bioinform. Biol. Insights* **12**, 1177932218759292 (2018).
53. Hulot, A., Chiquet, J., Jaffrézic, F. & Rigaill, G. Fast tree aggregation for consensus hierarchical clustering. *BMC Bioinform.* **21**, 1–12 (2020).
54. Gan, G., Ma, C. & Wu, J. *Data Clustering: Theory, Algorithms, and Applications* (SIAM, 2020).
55. Karim, M. R. et al. Deep learning-based clustering approaches for bioinformatics. *Brief. Bioinform.* **22**, 393–415 (2021).
56. Maaten, L. v. d. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
57. McInnes, L., Healy, J., Saul, N. & Großberger, L. Umap: Uniform manifold approximation and projection. *J. Open Source Softw* **3**, 861 (2018).
58. Aljalbout, E., Golkov, V., Siddiqui, Y., Strobel, M. & Cremers, D. Clustering with deep learning: Taxonomy and new methods. Preprint at https://arxiv.org/abs/1801.07648 (2018).
59. Makhzani, A. & Frey, B. K-sparse autoencoders. Preprint at https://arxiv.org/abs/1312.5663 (2013).
60. Chaudhary, K., Poirion, O. B., Lu, L. & Garmire, L. X. Deep learning–based multi-omics integration robustly predicts survival in liver cancer. *Clin. Cancer Res.* **24**, 1248–1259 (2018).
61. Ba, J. L., Kiros, J. R. & Hinton, G. E. Layer normalization. Preprint at https://arxiv.org/abs/1607.06450 (2016).
62. Clevert, D.-A., Unterthiner, T. & Hochreiter, S. Fast and accurate deep network learning by exponential linear units (elus). In *International Conference on Learning Representations* (2016).
63. Ng, A. Y. Feature selection, l 1 vs. l 2 regularization, and rotational invariance. In *Proc. Twenty-First International Conference on Machine Learning* 78 (ACM, 2004).
64. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations* (2015).
65. Hao, M. et al. Large-scale foundation model on single-cell transcriptomics. *Nat. Methods* **21**, 1481–1491 (2024).
66. Hinton, G. E. & Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science* **313**, 504–507 (2006).
67. Ester, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, vol. 96, 226–231 (AAAI Press, 1996).
68. Strehl, A. & Ghosh, J. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* **3**, 583–617 (2002).

Acknowledgements

The CARA experiment was supported by grant number GA-2013-104, Center for Advancement of Science in Space, to A.-L. Paul (PI) and R.J. Ferl (CoI). We gratefully acknowledge support from NASA 80NSSC19K0132, 80NSSC19K0126, and 80NSSC21K0577 to S.G. The authors would like to acknowledge the sequencing and bioinformatics services provided by the Interdisciplinary Center for Biotechnology Research's (ICBR) Gene Expression (RRID: SCR_019145), NextGen Sequencing (RRID: SCR_019152), and Bioinformatics (RRID: SCR_019120) cores.

Author contributions

D.H.S. and R.B. conceived of the study and fundamental design. D.H.S. and H.F.S. contributed to model testing, data analysis, and figure preparation. D.H.S., H.F.S., M.Z., R.B., A.-L.P., R.J.F., and S.G. contributed to manuscript preparation. All authors contributed to the manuscript review and editing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41526-025-00525-5.

Correspondence and requests for materials should be addressed to Simon Gilroy.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025