Machine learning ensemble identifies distinct age-related response to spaceflight in mammary tissue

James A. Casaletto[1], Tyler Zhao[2], Jay Yeung[2], Amaan Ansari[2], Ayush Raj[2], Arnav Mishra[2], Amber Fry[2], Kathryn Sun[2], Sofia Lendahl[2], Willy Guan[2], Abigail Lee[2], Lauren M. Sanders[3], Sylvain V. Costes[3]

[1] Blue Marble Space Institute of Science at NASA Ames Research Center
[2] SAAS Data Science, University of California, Berkeley
[3] NASA Ames Research Center

Abstract

Spaceflight presents unique environmental stressors, such as microgravity and radiation, that significantly affect biological systems at the molecular, cellular, and organismal levels. Female astronauts, in particular, face an increased risk of developing breast cancer due to exposure to ionizing radiation and other spaceflight-related factors. Age also plays a crucial role in the mammary gland's response to these stressors, with younger organisms generally exhibiting more efficient response mechanisms than older ones. In this study, we utilized an ensemble of machine learning algorithms to analyze gene expression profiles from mammary tissue of young and old female mice exposed to spaceflight to predict age (old vs young) and condition (spaceflight vs ground control). Using the genes our ensemble identified as most predictive, we investigate the molecular pathways involved in spaceflight-related health risks, particularly in the context of breast cancer and cardiovascular health. We identified age-dependent differences in the gene expression profiles of spaceflight-exposed mice compared to ground control. Specifically, younger mice exhibited enriched pathways related to cellular structure, while older mice showed activation of pathways involved in cortisol synthesis and muscle contraction. All mice responded to spaceflight with evidence of elevated lipid metabolic function. These findings highlight the critical role of age in modulating the response to spaceflight-induced stress and suggest that these molecular pathways may contribute to differential outcomes in tissue homeostasis, cardiovascular and metabolic disorders, and breast cancer susceptibility.

Introduction

Spaceflight exposes living organisms to a unique set of environmental challenges, including microgravity, radiation, and altered gas composition, which can significantly impact biological systems at the molecular, cellular, and organismal levels. Several systems have been

shown to be impacted in both male and female organisms, including the cardiovascular[1], musculoskeletal[2], immune[3], neurologic[4], hepatic[5], and ophthalmologic[6] systems to name a few. Female astronauts in particular have an increased risk of breast cancer due to exposure to galactic cosmic radiation[7]. Ionizing radiation is one of the few exposures known to increase breast cancer risk[8], and microgravity – another spaceflight stressor – has been shown to exacerbate breast cancer[9]. Furthermore, spaceflight disrupts circadian rhythms, and consequent lower levels of melatonin reduces its efficacy in inhibiting cancer cells[10,11]. Mammary cellular response to spaceflight has been shown to differ with age, as younger organisms typically exhibit more efficient cellular repair and adaptive mechanisms than their older counterparts[12]. However, adolescent murine mammary glands exposed to ionizing radiation show increased activation of mammary stem cell and Notch signaling pathways, heightened mammary repopulating activity, and an increased propensity to develop estrogen receptor (ER) negative tumors[13]. The Childhood Cancer Survivor Study indicates that breast cancer risk is highest in young women treated for Hodgkin's lymphoma, but it is also increased in those who received moderate-dose chest radiation for other pediatric or young adult cancers[14]. Female astronauts are at a higher risk of developing spaceflight-related pathologies than their male counterparts, with age being a contributing factor to this increased vulnerability.

Bioinformatic approaches have been used to study the effect of spaceflight on health. Many methods in bioinformatics, such as genome-wide association studies (GWAS) and differential gene expression analysis (DGEA), leverage statistical hypothesis testing as a mechanism to discover new insights. Machine learning (ML) has been leveraged but to a much lesser extent[15]. Machine learning and artificial intelligence (AI) models have been getting more complex, are trained on larger datasets, and are run on faster processors – all accelerating the rate of adoption of AI/ML techniques in more domains including bioinformatics[16]. Researchers typically need a large number of samples to build an accurate AI/ML model, particularly when using high-dimensional data such as gene expression data as predictors[17]. To mitigate this, researchers employ some form of feature selection – a broad collection of techniques that reduces the dimensionality of the feature space[18]. Filtering methods such as coefficient of variation and feature correlation to target are examples of feature selection techniques. Traditional machine learning algorithms such as support vector machines, single-layer perceptrons, and linear regression may be considered weak learners in the context of high-dimensional datasets. But leveraged together in an ensemble, such weak learners can achieve excellent performance[19].

The use of machine learning to study spaceflight-induced changes in mammary gene expression can offer valuable insights into the mechanisms of breast cancer development[20]. In this study, we examine the gene expression profiles from a controlled experiment in which mammary tissue of young and old mice were exposed to spaceflight. We used traditional machine learning methods including random forest, logistic regression, support vector machine, and the single-layer perceptron. We observed distinct age-dependent differences in the molecular pathways activated by spaceflight as opposed to ground control. Specifically, in younger mice, gene expression was enriched for pathways related to focal adhesion and extracellular matrix structure, while older mice exhibited activation of pathways associated with cortisol synthesis and muscle contraction. These findings suggest that age plays a critical role in

modulating the response to spaceflight-induced stress, potentially contributing to differential outcomes in tissue growth, immune response, and disease susceptibility.

Data

In this section, we discuss the NASA Open Science Data Repository (OSDR) dataset we used (OSD-511) and how we transformed it in preparation for our machine learning experiments.

OSD-511 dataset

In the Rodent Research Reference Mission 1 (RRRM-1), a total of forty female BALB/cAnNTac mice were sent to the International Space Station (ISS) to investigate the impact of age on spaceflight outcomes. The study included 21 younger mice (aged 10–12 weeks) and 22 older mice (aged 32 weeks). Among the younger mice (YNG), 5 were flown in space, 8 were kept in standard cages, and 8 were housed in vivarium cages. For the older mice (OLD), 5 were flown in space, 7 were housed in flight hardware, and 10 in vivarium cages. After 40 days in space, the mice were safely returned to Earth, given 2 days to recover (Live Animal Return, LAR), and then euthanized. Each group of space-flown (FLT) mice had corresponding control groups (GC), including flight hardware controls (10 mice housed under conditions similar to the flight hardware), basal controls (10 mice euthanized one day post-launch), and vivarium controls (10 mice housed in standard vivarium conditions). All mice were weighed at euthanasia. The dataset contains ribo-depleted total RNA-seq data from mammary glands dissected from six LAR groups (young and old mice in flight, ground, and vivarium conditions). Mus musculus Spliced Transcripts Alignment to a Reference (STAR) and RNA-Seq by Expectation-Maximization (RSEM) references were generated using STAR (version 2.7.10a) and RSEM (version 1.3.1), mapped to Ensembl release 107, genome version GRCm39.This data is available in the Open Science Data Repository[21] as dataset OSD-511[22]. OSDR normalized the RSEM count data using the median of ratios method from DESeq2. All 3 datasets (RSEM, STAR, and normalized RSEM) are published with OSD-511. In this research, we combine the vivarium and ground control groups into a single group called “non-flight”, comprising a total of 33 non-flight mice and 10 flight mice. The principal components analysis (PCA) plots of the data are shown in Figure 1.

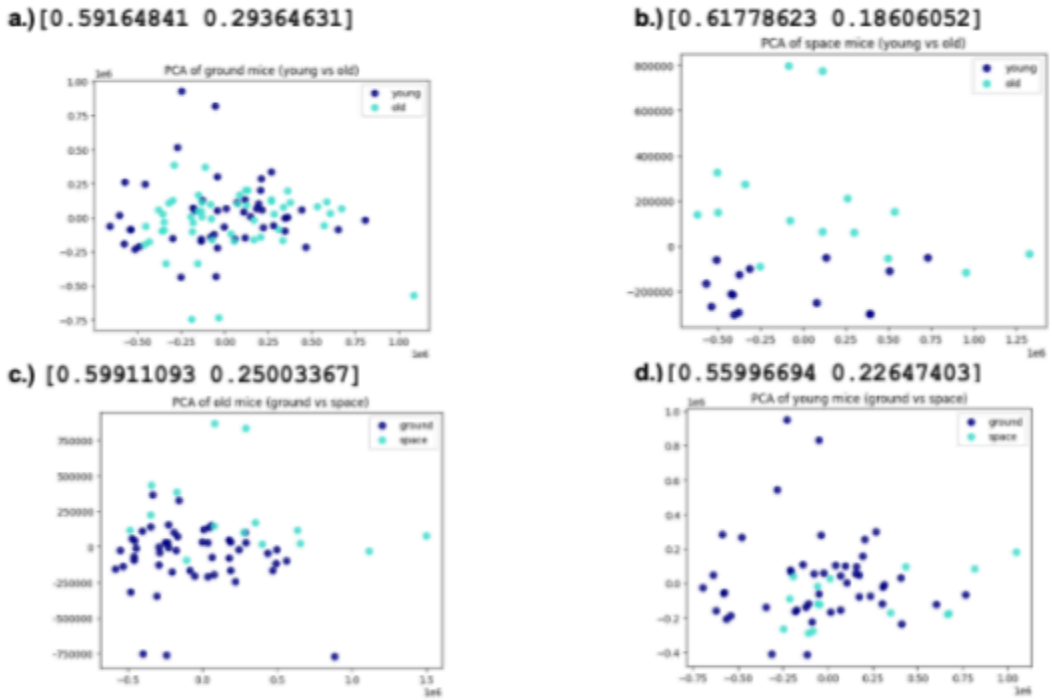

Figure 1. PCA plots colorized by age for ground mice (on the left) and space mice (on the right)

Based on the PCA plots of Figure 1a and 1d, we observe that the overall distinction in gene expression response to spaceflight is not well defined. Similarly, the marginal distributions of ground mice and young mice gene expression in Figures 1b and 1f also do not separate distinctly. However, the response in gene expression to spaceflight from the marginal distribution of old mice in Figure 1e is slightly more distinct, and the difference in gene expression response between young mice and old mice is very apparent in Figure 1c. This suggests that predicting condition for old mice and predicting age for space mice should be easier for a model to learn than the counterpart experiments, at least when using these two principal components as predictors.

Data transformations

Because machine learning model performance improves with more data points, we used data augmentation to increase the size of the dataset. We combined the RSEM and STAR data sets by creating 2 data points per biological sample: one for the RSEM quantification and one for the STAR quantification. This increased the size of our dataset by a factor of 2. Because machine learning model performance improves with fewer dimensions, we performed four filtering methods to reduce the dimensionality of the dataset: we kept differentially expressed genes between groups (FLT vs GC and YNG vs OLD) using an alpha=0.5 level of significance; genes with a coefficient of variation above a 0.5 threshold; only genes that code for proteins; and genes which were not missing counts for any of the samples. Some machine learning algorithms require standardizing the data prior to building the model to put all the feature values

on the same scale. These algorithms include logistic regression and single-layer perceptron. We used the `StandardScaler` method from `scikit-learn` to put all feature values in units of z-scores. The other algorithms in our ensemble (random forest and support vector machine) do not require standardization.

Methods

In this section, we describe the ensemble of ML algorithms we leveraged, how we derived feature importance from the trained models, and how we combined and filtered the results of the models to form a final set of gene results from our experiments. Figure 2 shows the graphical summary of the methods we employed. We adopt the syntax GROUP:target to denote the experiment where GROUP represents the subsets ground control (GC) and spaceflight (FLT) or the subsets young mice (YNG) and old mice (OLD), and the target represents the binary class (age or condition) which the machine learning model is trained to predict.

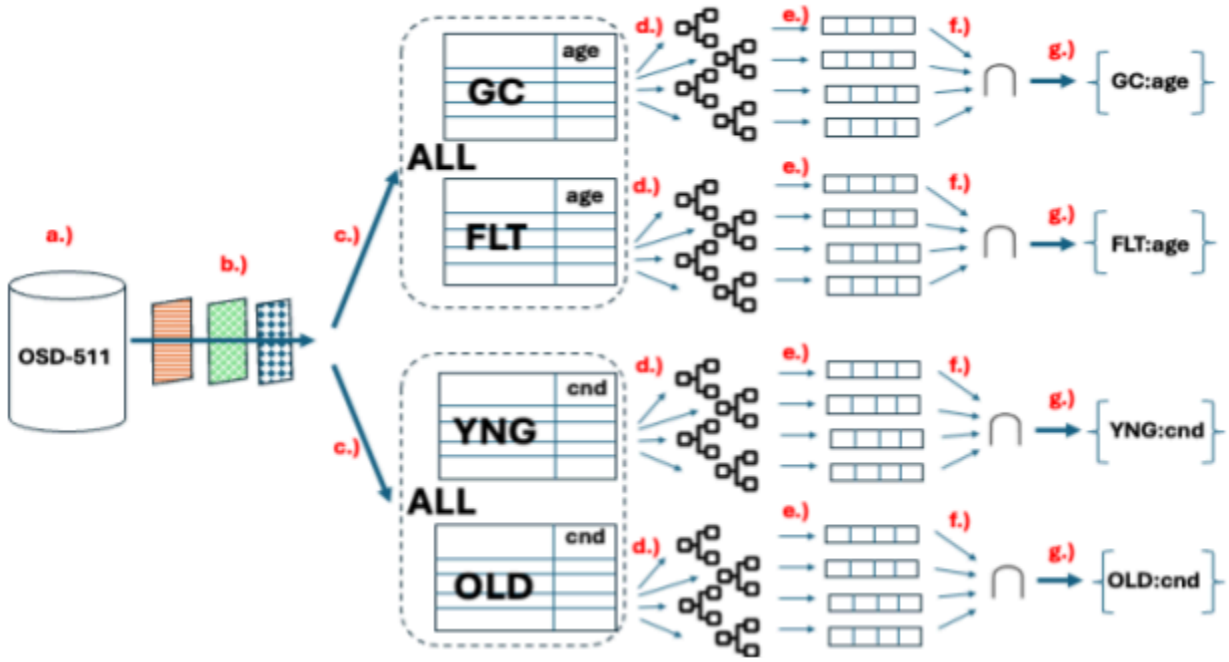

Figure 2. Graphical summary of the methods employed in this research. **2a.** The OSD-511 dataset contains RNA-seq data for mouse mammary tissue. **2b**. The data were filtered to reduce dimensionality. **2c**. Data were divided into GC and FLT groups to predict age and divided into YNG and OLD groups to predict condition. **2d.** Each subset of data was used to build 4 models in the ensemble. **2e.** Each model generated a list of genes most predictive of the target. **2f.** Those genes were majority-intersected to yield the intermediate results in **2g**.

Figure 2 shows all the steps in the pipeline to produce the intermediate result set of genes which were further processed as described in Figure 3.

Algorithms

We leveraged four supervised machine learning models on the gene expression data to predict labels associated with each sample. These models were trained and tested to classify binary labels (spaceflight vs ground control and old vs young): random forest (RF), logistic regression (LR), single-layer perceptron (SLP), and support vector machine (SVM). These models were selected to capture a broad range of classification patterns, including non-linear decision boundaries (RF, LR) and linear relationships (SLP, SVM).

Random forest is one of the most commonly used machine learning algorithms for both classification and regression for its ease of use, explainability, and overall good performance[23]. RF is itself an ensemble algorithm as a collection of decision trees - weak learners which when combined together create accurate predictions of targets based on numerical or class-based features. We used the `scikit-learn` implementation of RF as a binary classifier with a maximum depth of 4, and all other hyperparameters were set to their defaults. Logistic regression, despite its name, is a binary classification algorithm which provides a probability for the binary target outcome based on a set of discrete or continuous features[24]. As such, it has a built-in metric which can be leveraged for confidence in the prediction. And because it does use regression, there are model coefficients associated with the features that may be used for feature importance. We used the `scikit-learn` implementation of LR as a binary classifier with all default values for the hyperparameters. The single-layer perceptron was developed in the 1950's by Frank Rosenblatt and is the most basic form of neural network[25]. The input features are weighted in a linear combination which can either be sent through a sigmoidal activation function for binary classification or through a linear activation function for regression. Feature importance is conveniently derived directly from the feature weights which makes the SLP an easy-to-interpret ML algorithm. We used the `scikit-learn` implementation of SLP as a binary classifier with all default hyperparameter values. The support vector machine was created by Hava Siegelmann and Vladimir Vapnik as a margin-based classifier using so-called support vectors to separate classes in the feature space[26]. In addition to linearly separating the feature space, a kernel transformation can be leveraged to do nonlinear separation. Feature importance is derived directly from the coefficients of the support vectors of linear kernels. We used the `scikit-learn` implementation of the linear SVM with all default hyperparameter values.

All four models were trained using a train/test split of 75/25 and validated using the `scikit-learn` implementation of k-fold cross validation (k=10). We deployed the four classification algorithms as binary classifiers in two experiments: predicting age (OLD/YNG) and predicting condition (FLT/GC) using gene expression data as predictors. We deployed the linear regression model to predict weight at euthanasia. After training each model, we identified the features most predictive of the classes using the two methods described in the Feature Importance section.

Per-model feature importance

In this research, we combined multiple machine learning algorithms into a single ensemble to predict either condition (ground control vs spaceflight) or age (young vs old). We used two approaches from the `scikit-learn` package for feature importance: cross-validation feature importance and permutation feature importance. The cross-validation feature importance approach divides the dataset into k folds, trains the model on 1 fold and tests on the other k-1 folds[27]. This is repeated over all k folds to train k models. The `scikit-learn` implementation of cross-validation returns an estimator from which feature importances are assigned to each feature in the model. We calculate the mean value of the feature importances for all features across the models from the k folds to assign a final importance metric to each feature. Sorting the features in descending order based on their mean feature importance, we took the top 30 features most predictive of the target for each model. The permutation feature importance approach randomly permutes the values of each feature, one by one, and calculates the model performance before and after having done so[28]. The degree to which the model performance changes is assigned to each feature as its importance metric. We used the `scikit-learn` implementation of permutation feature importance, setting the `scoring` metric to `accuracy` for the binary classification methods. The other option for the scoring metric is F1-score which is particularly well-suited for imbalanced datasets. Our classes were exactly balanced between old (n=10) and young (n=10) and between spaceflight (n=10) and ground control (n=10), so we chose accuracy. Each of these feature importance methods generated a set of up to `n=30` genes.

We then used union and intersection set operations to combine these gene sets together into a single set of genes as described in the next section.

Per-experiment ensemble voting

The primary method of combining results from a machine learning ensemble is by majority vote[29]. For each machine learning experiment, we used union and intersection set operations to combine the results of the two feature importance approaches for each model into a single set of genes as follows. First, for each model, we performed a union of the set of results from each of the two feature importance methods to get a set of up to 60 genes. Then we performed an intersection across the 4 model unions whereby we included a gene if it belonged to at least 3 union sets (3 is the simple majority of 4).

We then use the difference set operation to identify the final set of genes most predictive of the labels as described in the next section.

Final gene set formulation

To formulate the final set of genes most predictive of a target (age or condition) for a given subset (YNG/OLD or FLT/GC) of mice, we removed those genes which are generally predictive of the target regardless of their subset. In this way, we identify the marginal set of genes that are uniquely predictive in that subset. For example, in the experiment in which we predict age, we ran 3 experiments: one in which we used only ground control samples to predict

age (GC:age), one with only spaceflight samples to predict age (FLT:age), and one with all the samples combined to predict age (ALL:age). Each of these 3 experiments gave a set of gene results as previously described. In Figure 3, we describe how we formulate our final set of gene results for analysis.

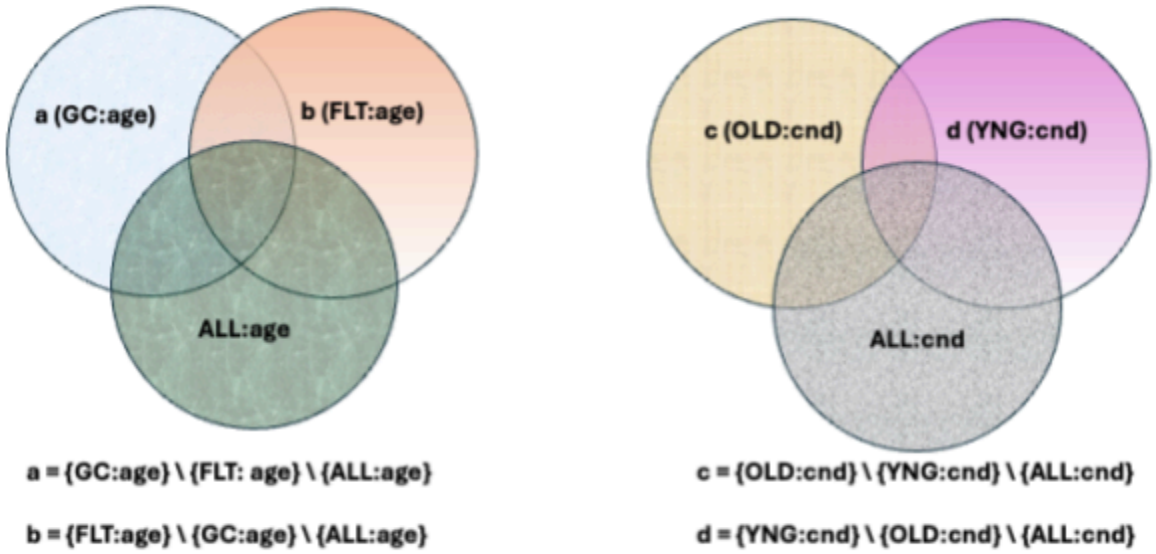

Figure 3. Venn diagrams depicting set difference operations to identify genes uniquely predictive of the target for a given subset of mice.

In Figure 3a, we remove ALL:age genes and FLT:age genes that intersect with GC:age to obtain those genes that uniquely predict age for ground control mice. These genes are represented by the light blue part of the Venn diagram on the left. Similarly, in Figure 3b, we remove ALL:age genes and GC:age genes that intersect with FLT:age genes to obtain those genes that uniquely predict age for spaceflown mice. These genes are represented by the light orange part of that Venn diagram on the left. Using the same logic, we obtain those genes that uniquely predict condition for old mice in the yellow textured part of the Venn diagram on the right. And finally, we obtain those genes that uniquely predict condition for young mice in the pink part of the Venn diagram on the right. These set operations yielded the final gene results we discuss in the next section.

Results

In this section, we discuss the final results of our four experiments: predicting age for ground control samples, predicting age for spaceflight samples, predicting condition for old samples, and predicting condition for young samples.

Model performance

The tables in Figure 4 display the performance of each of the 4 classification models in the ensemble for each of the four experiments.

a.) predicting age

Model	Accuracy train | test | validate	Experiment
Random forest	0.99 | 0.89 | 0.87	FLT:age
	0.97 | 0.91 | 0.94	GC:age
Support vector machine	0.97 | 0.80 | 0.87	FLT:age
	0.95 | 0.87 | 0.93	GC:age
Logistic regression	0.96 | 0.90 | 0.89	FLT:age
	0.93 | 0.89 | 0.91	GC:age
Single-layer perceptron	0.91 | 0.77 | 0.79	FLT:age
	0.89 | 0.82 | 0.75	GC:age

b.) predicting condition

Model	Accuracy train | test | validate	Experiment
Random forest	0.96 | 0.96 | 0.89	OLD:cnd
	0.97 | 0.97 | 0.96	YNG:cnd
Support vector machine	0.96 | 0.94 | 0.85	OLD:cnd
	0.94 | 0.93 | 0.94	YNG:cnd
Logistic regression	0.96 | 0.96 | 0.91	OLD:cnd
	0.95 | 0.95 | 0.94	YNG:cnd
Single-layer perceptron	0.89 | 0.76 | 0.61	OLD:cnd
	0.88 | 0.81 | 0.76	YNG:cnd

Figure 4. Accuracy of each of the 4 classification models for a.) experiment predicting age (old or young) for those mice in the spaceflight and ground control groups and for b.) experiment predicting condition (spaceflight or ground control) for those mice in the old and young groups.

As shown in Figure 4, the test and validate accuracies are, on average, sub-optimal despite the relatively high training accuracies. Moreover, the held-out test and validation accuracies for the experiments which predict the age of spaceflight mice are not as good as those for predicting age with ground control mice. Similarly, the accuracy of predicting condition for old mice is mostly lower than the accuracy of predicting condition for young mice. This contradicts our analysis based on the PCA plots and suggests that the separation of those subsets in high-dimensional space is nonlinear. Indeed, the single-layer perceptron and support vector machine – both of which were configured with linear decision boundaries – are the weakest learners in the ensemble.

Most predictive genes

In this section we discuss the genes most predictive of the labels for each experiment. Table 1 lists the genes most predictive of the label for each of the four experiments.

Predict age for ground control mice	Predict age for spaceflight mice	Predict condition for young mice	Predict condition for old mice
4921522P10Rik ATP8 Acsl1 Acta1 Adamts12 Angptl8 Anxa8 Areg Bhlhe40 Ccdc121rt1 Ccn6 Ccnd2 Cdcp2 Ckm Col6a3 Cyp2j7 Dpt Epha10 Fabp4 Gm16401 Gm17271 Gm49359 Gm5581 Greb1 Hk2 Krt71 Krt79 Krt87 Mc4r Mcm6 Mmd2 Mucl1 Mup14 Myh1 Myh2 Nnat Npr3 Nup58 Or1ad1 Pdc Pde4dip Pdk4 Plaur Ppp1r14bl Ppp1r3b Psg26 Ptpn5 Ptx3 Serpine1 Sh2d1b2 Snorc Sprr2a2 Stoml3 Timd6 Tkt Ttr Ubtfl1 Ush2a Vkorc1 mt-Atp8 mt-Nd3	A530016L24Rik ATP6 Acss2 Actn3 Adcy10 Aspg Bltp3a C3 Casq1 Ccn5 Ccr3 Csn2 Defb50 Dgat2 Dnah8 Dsc3 Erdr1y Fbxw13 Ghrhr Gm20431 Gpam Lpin1 Me1 Mlxipl Mmp20 Mogat2 Mss51 ND4L Nppc Odad4 Pgd Plin2 Ppp3r2 Pvalb S100a8 Satb1 Scd2 Sh2d5 Slc27a1 Stfa3 Tet2 Tmc3 Tns1 Vsig8 Wap mt-Nd2 mt-Nd4	A930009A15Rik Aadacl3 Adm2 Adra2b Agbl4 Alb Amy2a1 Ankrd23 Ankrd53 Anks1b Cacna1s Ccer2 Cct8l1 Cfd Cldn34c4 Cldn9 Cngb1 Cstdc5 Cyp11b1 Cypt2 Dmkn Dnah14 Dsg1a Dsg3 E030025P04Rik ENSMUSG00000072693 Fat2 Fcgbp Fcnb Fgf17 Flg2 Folh1 Fzd9 Gcm2 Gm10874 Gm2174 Gm5901 Gp6 Gpr149 Gpr63 Gsn Hbq1a Htr1b Igfbpl1 Igfn1 Kcna4 Krt10 Lrrc24 Mroh2b Mucl2 Myh6 Myl2 Myl3 Ngp Nmrk2 Npbwr1 Or13a18 Or5k1 Pabpc1 Pde6g Pdilt Pdzrn4 Peg10 Perp Ptma Rcor2 Rph3a Rspo1 Sdr16c5 St6gal2 Tcap Teddm1b Tlx1 Tmem89 Trim58 Vmn1r53 Vwa5b1 Zscan10	4931429P17Rik Aacs Adamts1 Ahnak Aldh3a1 Apoe Apold1 Col4a1 Col6a1 Cst6 Cycs Etfb Fabp3 Flna Gm4847 Gpr158 H2bc24 Ido1 Iqgap3 Krt4 Krt86 Krtap1-5 Lep Mt1 Myom1 ND6 Obscn Olfm4 Pfkfb3 Pi16 Ppia Rbp1 Sdr16c6 Sncg Spmip5 Tchh Tnxb Tpm2 Tuba1b Unc80 Vmn1r34 Vmn1r35

Table 1. List of genes most predictive of the target for the given subset.

The genes listed in Table 1 constitute the final results of our method that result from the set operations portrayed in Figure 3. In Figure 5, we show the distribution of the gene expression for the entire RNA-seq dataset.

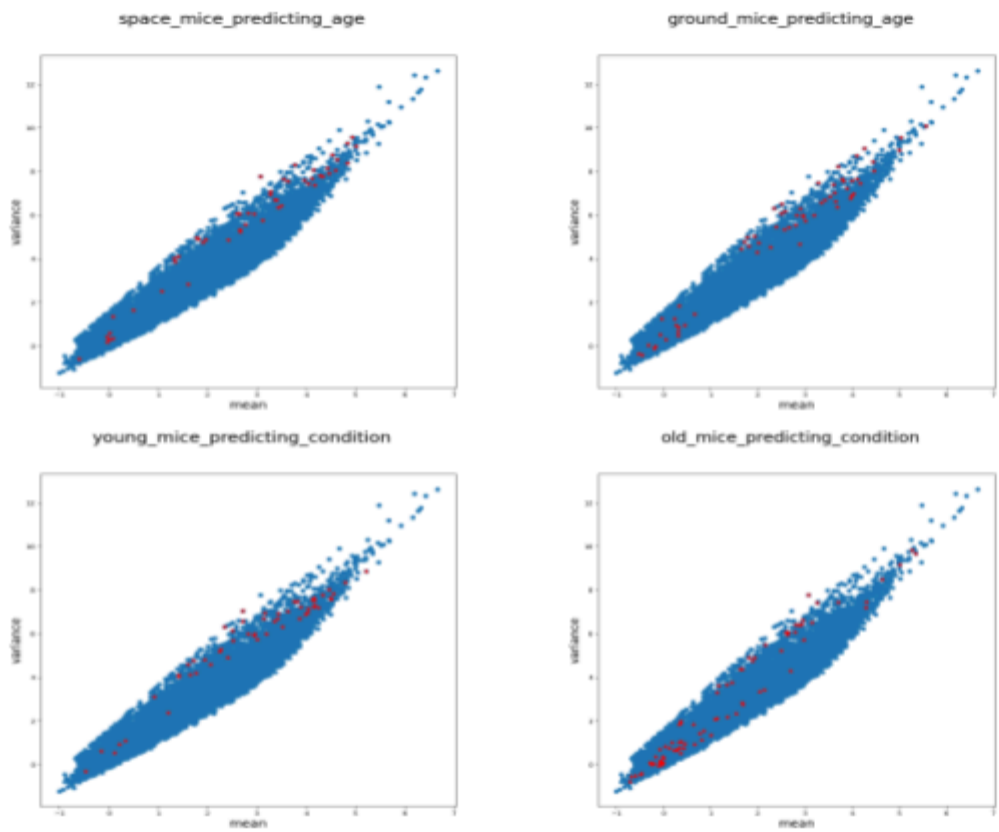

Figure 5. Scatter plots of variance versus mean for each of the 4 experiments. Each point in the plot represents the mean and variance of a single gene's expression. There is one blue point for each gene in the dataset we used for training, and the red points are those identified

As shown in Figure 5, the distribution of the genes across the spectrum of expression identified by our machine learning ensemble is approximately uniform. That is, the machine learning algorithms do not portray any bias based on the magnitude or shape of the gene distributions. This indicates that the models and their ensemble are not vulnerable to the heteroskedastic nature of the gene expression data.

Pathway enrichment analysis

We submitted our lists of most predictive features to ShinyGO version 0.81 – an online pathway enrichment analysis tool (http://bioinformatics.sdstate.edu/go/) – using the KEGG pathway databases[30], FDR cutoff of 0.05, minimum pathway size of 2, and displayed the top 5 most enriched pathways. The results of these analyses are captured in Figure 6.

a.) SPACE:AGE

Pathway	FDR q-value
Pyruvate metabolism	0.001
PPAR signaling	0.001
Fat digestion and absorption	0.003
Carbon metabolism	0.031
Glyoxylate and dicarboxylate metabolism	0.031

c.) OLD:CONDITION

Pathway	FDR q-value
Cardiac muscle contraction	0.012
Hypertrophic cardiomyopathy	0.012
Dilated cardiomyopathy	0.012
Cortisol synthesis and secretion	0.043
Adrenergic signaling	0.043

b.) GC: AGE

No significant enrichment found!

c.) YNG:CONDITION

Pathway	FDR q-value
ECM-receptor interaction	0.006
Focal adhesion	0.006

Figure 6. Pathway enrichment analyses for the experiment which uses a.) space mice to predict age, b.) ground control mice to predict age, c.) old mice to predict condition, and d.) young mice to predict condition.

There is a distinct trend among the pathways enriched for each experiment. The genes most predictive of age for space-flown mice are largely involved in lipid metabolism and energy production, but there were no significantly enriched pathways from the genes most predictive of age for ground mice. The genes most predictive of condition for old mice are mostly responding to cardiovascular stress, whereas the genes most predictive of condition for young mice are responsible for cellular shape and structure.. The next section explores in more detail the genes that enriched these pathways.

Gene function

We used the NCBI gene database to determine the function of each of the genes involved in the pathways of Figure 6. The genes are listed in Figure 7, along with a brief functional annotation derived from the National Center for Biotechnology Information (NCBI) gene database[31].

a.) space mice target age

Gene	Function
Acss2	Catalyzes energy and lipid synthesis
Adcy10	Regulates energy metabolism
C3	Eliminates pathogens reduces inflammation
Dgat2	Catalyzes triglyceride synthesis
Dnah8	Involved in cellular transport
Glu1	Maintains homeostasis of glucose
Gpam	Synthesizes glycerolipids in mitochondria
Lipin1	Regulates lipid metabolism
Me1	Synthesizes fatty acids
Mogat2	Metabolizes lipids
Pgd	Catalyzes decarboxylation of phosphogluconate for metabolism
Plin2	Promotes formation of lipid droplets and regulates lipid storage
Ppp3r2	Regulates protein phosphatase activity in stress response
Scd2	Synthesizes fatty acids
Slc27a1	Transports long-chain fatty acids into cells
mt-Atp6	Transports electrons and ions across mitochondria membrane

b.) old mice target condition

Gene	Function
Adra2b	Encodes adrenergic receptor in stress response
Cacna1s	Creates calcium channel subunit for muscle contraction
Cyp11b1	Synthesizes corticosteroids in response to stress
Fzd9	Involved in WNT signaling for tissue homestasis
Kcna4	Involved in potassium channels in muscles
Myh6	Produces alpha-myosin for cardiac muscle
Myl2 Myl3	Encodes myosin for cardiac muscle contraction

c.) young mice target condition

Gene	Function
Col4a1	Synthesizes collagen for basement membranes, cellular structure, and connective tissues.
Col6a1	
Col6a2	
Flna	Maintains cellular structure, shape, and movement
Tnxb	Regulates assembly and organization of collagen fibers

Figure 7. List of genes from enriched pathways and their function.

The three tables in Figure 7 list the genes and a brief description of their associated function. Overall, the genes most predictive of age for mice flown in space relate to lipid and energy metabolism, however there were no enriched pathways for the genes that predict age for ground control mice. The genes most predictive of condition for old mice relate to cardiovascular activity in response to stress, whereas the genes most predictive of condition for young mice are largely related to cellular and extracellular shape, structure, and integrity.

Discussion

In this study, we used a novel approach combining results from an ensemble of machine learning models to predict condition (spaceflight or ground control) and age (young or old) using features derived from gene expression data. The results revealed distinct gene expression signatures that differentiate both age and exposure to spaceflight in mice, revealing some of the molecular mechanisms that may underpin the effects of spaceflight, aging, and their potential impact on diseases such as breast cancer.

The machine learning models designed to predict the age of spaceflown mice from gene expression data identified a set of genes that are involved in critical biological processes such as lipid metabolism, inflammation, and cellular signaling. These processes are known to change with age, and several of the identified genes are involved in pathways that may be more pronounced in older organisms, including energy storage, oxidative stress response, and

cellular maintenance[32]. For example, genes like *Gpam*, *Lipin1*, and *Me1*, which are involved in lipid metabolism, reflect changes in how older mice manage energy stores and cellular integrity[33]. Lipid metabolism has been linked to cancer progression, as tumors often alter lipid pathways to meet their energy demands[34]. Moreover, hyperlipidemia – the excessive metabolism of lipids – is positively associated with cardiovascular disease[35]. The expression of *Plin2* and *Scd2* – genes involved in fat storage and metabolism – could also have implications for cancer, as the dysregulation of these metabolic pathways is often observed in breast cancer cells where altered lipid metabolism contributes to tumor growth and metastasis[34,36]. Furthermore, *mt-Atp6*, a mitochondrial gene involved in energy production, highlights a potential link between mitochondrial dysfunction and cancer: mitochondrial changes, including those associated with aging, can promote an environment conducive to tumorigenesis[37,38].

When predicting condition in old mice, our ensemble identified a distinct set of genes involved in stress response and cardiovascular function. For example, genes such as *Cacna1s* and *Kcna4*, which regulate calcium and potassium ion channels, are essential for maintaining cardiac muscle function which is subject to microgravity-induced atrophy[39,40]. Calcium signaling also plays a well-established role in breast cancer progression: altered calcium homeostasis can influence cancer cell migration, proliferation, and invasion[41]. Changes in these genes in space-flown old mice could therefore indicate broader disruptions that may affect not only cardiovascular health but also predispose the individual to malignant cancers where ion channel dysfunction is often implicated[42]. The differential expression of genes in space-flown old mice may also reflect maladaptive responses to the reduced mechanical loading in space. Notably, *Cyp11b1*, which regulates corticosteroid production, is linked to stress responses. Chronic stress has been shown to influence cancer progression through the activation of stress-related signaling pathways, including those mediated by glucocorticoids[43]. Furthermore, spaceflight-induced changes in cardiovascular function and overall tissue homeostasis, particularly in aging organisms, could influence the expression of genes involved in cell motility and cytoskeletal integrity including *Myh6*, *Myl2*, and *Myl3*. These alterations may be linked to systemic changes that impact other tissues, potentially influencing breast cancer progression indirectly[44–46].

For young mice, the most predictive genes for distinguishing spaceflight from ground control conditions were primarily involved in extracellular matrix (ECM) remodeling and cytoskeletal function. These genes play crucial roles in maintaining tissue structure and function, particularly in response to mechanical stresses. The differential expression of ECM-related genes, such as *Col4a1* and *Col6a1*, likely reflects changes in the mechanical environment caused by spaceflight, where tissue remodeling occurs in response to altered gravity[47]. Dysregulated ECM remodeling is a hallmark of cancer progression as the ECM not only provides structural support but also regulates cellular behaviors such as migration, invasion, and metastasis[48]. Additionally, *Flna* and *Tnxb*, which are involved in cytoskeletal structure and cellular signaling, likely reflect how spaceflight-induced stress impacts the cellular architecture of tissues, particularly in the context of growth and adaptation to the microgravity environment. Alterations in the cytoskeleton can impact cell shape, adhesion, and migration, processes that are crucial for cancer cell invasion and metastasis[49]. Thus, the altered expression of these genes could not only highlight the adaptive responses to spaceflight but

also raise concerns about how similar mechanisms may be exploited in the development of cancer.

Conclusion

Our study underscores the importance of gene expression profiling in understanding how spaceflight affects age-related changes in biological systems. Our research finds that both younger and older mice have different predispositions towards spaceflight-associated pathologies including cardiovascular disease, metabolic disorders, and cancer. We demonstrate the efficacy of using machine learning techniques to find biomedical signals in high-dimensional gene expression data using a relatively small number of samples. The different gene expression signatures associated with age and spaceflight reveal insights into the physiological and molecular mechanisms that underlie the aging process, the effects of spaceflight, and their potential implications for disease susceptibility. The identification of age-related and condition-specific genes provides a deeper understanding of how spaceflight-induced stressors impact tissue function, cellular signaling, and homeostasis, particularly in the context of aging. The dysregulation of ECM remodeling, cytoskeletal function, and stress response pathways observed in radiation-exposed mice, especially those involving calcium signaling, lipid metabolism, and hormonal regulation, may mirror processes observed in cancer biology. Given that breast cancer is influenced by many of these same pathways, the spaceflight-induced changes in gene expression could provide valuable insights into how environmental factors may alter disease susceptibility and progression. These findings offer valuable information for further studies into the impact of spaceflight on female astronaut health, reiterates well-established roles between aging and disease, and provides a straightforward machine learning approach to leverage on a vast array of unexplored data.

References

1. Hughson, R. L., Helm, A. & Durante, M. Heart in space: effect of the extraterrestrial environment on the cardiovascular system. *Nat. Rev. Cardiol.* **15**, 167–180 (2018).
2. Comfort, P. *et al.* Effects of Spaceflight on Musculoskeletal Health: A Systematic Review and Meta-analysis, Considerations for Interplanetary Travel. *Sports Med.* **51**, 2097–2114 (2021).
3. Crucian, B. E. *et al.* Immune System Dysregulation During Spaceflight: Potential Countermeasures for Deep Space Exploration Missions. *Front. Immunol.* **9**, 1437 (2018).
4. Van Ombergen, A. *et al.* The effect of spaceflight and microgravity on the human brain. *J. Neurol.* **264**, 18–22 (2017).

5. Beheshti, A. *et al.* Multi-omics analysis of multiple missions to space reveal a theme of lipid dysregulation in mouse liver. *Sci. Rep.* **9**, 19195 (2019).
6. Mao, X., Stanbouly, S., Holley, J., Pecaut, M. & Crapo, J. Evidence of Spaceflight-Induced Adverse Effects on Photoreceptors and Retinal Function in the Mouse Eye. *Int. J. Mol. Sci.* **24**, 7362 (2023).
7. Kumar, K. *et al.* Simulated Galactic Cosmic Radiation Exposure-Induced Mammary Tumorigenesis in ApcMin/+ Mice Coincides with Activation of ERα-ERRα-SPP1 Signaling Axis. *Cancers* **16**, 3954 (2024).
8. Helm, J. S. & Rudel, R. A. Adverse outcome pathways for ionizing radiation and breast cancer involve direct and indirect DNA damage, oxidative stress, inflammation, genomic instability, and interaction with hormonal regulation of the breast. *Arch. Toxicol.* **94**, 1511–1549 (2020).
9. Nassef, M. Z. *et al.* Short-Term Microgravity Influences Cell Adhesion in Human Breast Cancer Cells. *Int. J. Mol. Sci.* **20**, 5730 (2019).
10. Bartsch, C., Bartsch, H. & Peschke, E. Light, melatonin and cancer: current results and future perspectives 1. *Biol. Rhythm Res.* **40**, 17–35 (2009).
11. Malhan, D., Schoenrock, B., Yalçin, M., Blottner, D. & Relógio, A. Circadian regulation in aging: Implications for spaceflight and life on earth. *Aging Cell* **22**, e13935 (2023).
12. Arnold, C., Casaletto, J. & Heller, P. Spaceflight Disrupts Gene Expression of Estrogen Signaling in Rodent Mammary Tissue. *Med. Res. Arch.* **12**, (2024).
13. Tang, J. *et al.* Irradiation of Juvenile, but not Adult, Mammary Gland Increases Stem Cell Self-Renewal and Estrogen Receptor Negative Tumors. *Stem Cells* **32**, 649–661 (2014).
14. Mertens, A. C. *et al.* Cause-Specific Late Mortality Among 5-Year Survivors of Childhood Cancer: The Childhood Cancer Survivor Study. *JNCI J. Natl. Cancer Inst.* **100**, 1368–1379 (2008).
15. Larrañaga, P. *et al.* Machine learning in bioinformatics. *Brief. Bioinform.* **7**, 86–112 (2006).

16. Auslander, N., Gussow, A. B. & Koonin, E. V. Incorporating Machine Learning into Established Bioinformatics Frameworks. *Int. J. Mol. Sci.* **22**, 2903 (2021).
17. Feldner-Busztin, D. *et al.* Dealing with dimensionality: the application of machine learning to multi-omics data. *Bioinformatics* **39**, btad021 (2023).
18. Jovic, A., Brkic, K. & Bogunovic, N. A review of feature selection methods with applications. in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* 1200–1205 (IEEE, Opatija, Croatia, 2015). doi:10.1109/MIPRO.2015.7160458.
19. Rincy, T. N. & Gupta, R. Ensemble Learning Techniques and its Efficiency in Machine Learning: A Survey. in *2nd International Conference on Data, Engineering and Applications (IDEA)* 1–6 (IEEE, Bhopal, India, 2020). doi:10.1109/IDEA49133.2020.9170675.
20. Strube, F. *et al.* Alteration of Cytoskeleton Morphology and Gene Expression in Human Breast Cancer Cells under Simulated Microgravity. *Cell J.* **22**, 106–114 (2020).
21. Sanders, L. M. *et al.* Celebrating 30 years of access to NASA Space Life Sciences data. *GigaScience* **13**, giae066 (2024).
22. Galazka, J. M. *et al.* Transcriptional profiling of mammary glands from mice flown on the RRRM-1 mission. NASA GeneLab https://doi.org/10.26030/WDPR-VA45 (2022).
23. Biau, G. & Scornet, E. A random forest guided tour. *TEST* **25**, 197–227 (2016).
24. DeMaris, A. & Selman, S. H. Logistic Regression. in *Converting Data into Evidence* 115–136 (Springer New York, New York, NY, 2013). doi:10.1007/978-1-4614-7792-1_7.
25. Singh, J. & Banerjee, R. A Study on Single and Multi-layer Perceptron Neural Network. in *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)* 35–40 (IEEE, Erode, India, 2019). doi:10.1109/ICCMC.2019.8819775.
26. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
27. Wong, T.-T. & Yeh, P.-Y. Reliable Accuracy Estimates from *k* -Fold Cross Validation. *IEEE Trans. Knowl. Data Eng.* **32**, 1586–1594 (2020).

28. Altmann, A., Toloşi, L., Sander, O. & Lengauer, T. Permutation importance: a corrected feature importance measure. *Bioinformatics* **26**, 1340–1347 (2010).
29. Dietterich, T. G. Ensemble Methods in Machine Learning. in *Multiple Classifier Systems* vol. 1857 1–15 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2000).
30. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
31. Sayers, E. W. *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **50**, D20–D26 (2022).
32. Hamsanathan, S. & Gurkar, A. U. Lipids as Regulators of Cellular Senescence. *Front. Physiol.* **13**, 796850 (2022).
33. Houtkooper, R. H. *et al.* The metabolic footprint of aging in mice. *Sci. Rep.* **1**, 134 (2011).
34. Fu, Y. *et al.* Lipid metabolism in cancer progression and therapeutic strategies. *MedComm* **2**, 27–59 (2021).
35. Alloubani, A., Nimer, R. & Samara, R. Relationship between Hyperlipidemia, Cardiovascular Disease andStroke: A Systematic Review. *Curr. Cardiol. Rev.* **17**, e051121189015 (2021).
36. Broadfield, L. A., Pane, A. A., Talebi, A., Swinnen, J. V. & Fendt, S.-M. Lipid metabolism in cancer: New perspectives and emerging mechanisms. *Dev. Cell* **56**, 1363–1393 (2021).
37. Smith, A. L. M. *et al.* Age-associated mitochondrial DNA mutations cause metabolic remodeling that contributes to accelerated intestinal tumorigenesis. *Nat. Cancer* **1**, 976–989 (2020).
38. Picca, A. *et al.* The contribution of mitochondrial DNA alterations to aging, cancer, and neurodegeneration. *Exp. Gerontol.* **178**, 112203 (2023).
39. Patel, O. V., Partridge, C. & Plaut, K. Space Environment Impacts Homeostasis: Exposure to Spaceflight Alters Mammary Gland Transportome Genes. *Biomolecules* **13**, 872 (2023).
40. Vernice, N. A., Meydan, C., Afshinnekoo, E. & Mason, C. E. Long-term spaceflight and the cardiovascular system. *Precis. Clin. Med.* **3**, 284–291 (2020).

41. Zheng, S., Wang, X., Zhao, D., Liu, H. & Hu, Y. Calcium homeostasis and cancer: insights from endoplasmic reticulum-centered organelle communications. *Trends Cell Biol.* **33**, 312–323 (2023).
42. Parthasarathi, K. T. S. *et al.* In Silico Analysis of Ion Channels and Their Correlation with Epithelial to Mesenchymal Transition in Breast Cancer. *Cancers* **14**, 1444 (2022).
43. Chiriac, V.-F., Baban, A. & Dumitrascu, D. L. Psychological stress and breast cancer incidence: a systematic review. *Clujul Med. 1957* **91**, 18–26 (2018).
44. Mehta, L. S. *et al.* Cardiovascular Disease and Breast Cancer: Where These Entities Intersect: A Scientific Statement From the American Heart Association. *Circulation* **137**, (2018).
45. Liu, Y. *et al.* LncRNA TCL6 regulates miR-876–5p/MYL2 axis to suppress breast cancer progression. *Transl. Oncol.* **53**, 102210 (2025).
46. Cherukuri, S. P. *et al.* Breast Cancer and the Cardiovascular Disease: A Narrative Review. *Cureus* **14**, e27917 (2022).
47. Tian, J., Pecaut, M. J., Slater, J. M. & Gridley, D. S. Spaceflight modulates expression of extracellular matrix, adhesion, and profibrotic molecules in mouse lung. *J. Appl. Physiol.* **108**, 162–171 (2010).
48. Winkler, J., Abisoye-Ogunniyan, A., Metcalf, K. J. & Werb, Z. Concepts of extracellular matrix remodelling in tumour progression and metastasis. *Nat. Commun.* **11**, 5120 (2020).
49. Li, X. & Wang, J. Mechanical tumor microenvironment and transduction: cytoskeleton mediates cancer cell invasion and metastasis. *Int. J. Biol. Sci.* **16**, 2014–2028 (2020).