

Gravitational and Space Research

Deep Cross-Organism Generalization of the Physiological Effects of Spaceflight from Mammalian Model Organisms to Humans

Konstantinos I. Adamopoulos^{1,2}, Lauren M. Sanders², Adrienne Hoarfrost³, Sylvain V. Costes^{4*}

¹National Technical University of Athens, School of Electrical and Computer Engineering, Biomedical Engineering Laboratory, Heron Polytechniou 9, Zografou, 15780 Athens, Greece; ²Blue Marble Space Institute of Science, Space Biosciences Division, NASA Ames Research Center, Moffett Field, CA 94043, USA; ³Department of Marine Sciences, University of Georgia, Athens, GA, USA; ⁴NASA Space Biosciences Division, NASA Ames Research Center, Moffett Field, CA 94035, USA

Abstract

*The forthcoming human deep space exploration missions necessitate a thorough understanding of the impact of spaceflight conditions on human physiological systems. The NASA Open Science Data Repository (OSDR; <https://osdr.nasa.gov/bio/>) serves as a valuable resource, housing data derived from model organisms and human experiments conducted in spaceflight and terrestrial microgravity analogues. Machine Learning applications could maximize the use of existing data to understand and ultimately counteract physiological abnormalities during long-term missions. In our present study, we identified enriched terms and pathways associated with significantly dysregulated genes within each species and across orthologous counterparts. We also generated AI-ready merged meta-datasets comprised of musculoskeletal tissues from *Mus musculus* and *Homo sapiens* organisms. We then applied a series of supervised Machine Learning models to classify genes that were significantly over-expressed and under-expressed. Subsequently, we explored the utility of Transfer Learning in this domain by pretraining a model on the larger *Mus musculus* merged dataset and then refining it on the smaller *Homo sapiens* dataset. This approach showcases the potential of Transfer Learning in providing an insight into the effective transfer of information from model organisms to humans, offering a robust framework for advancing research in space biology and developing countermeasures for long-duration space exploration.*

Keywords

NASA GeneLab • Spaceflight • Microgravity • Space Biology • Microarrays • Differential Gene Expression • Machine Learning; Classification • Musculoskeletal System;

Introduction

The National Aeronautics and Space Administration (NASA) Open Science Data Repository (OSDR; <https://osdr.nasa.gov/bio/>) is an interactive, open-access omics database that empowers users to upload, download, share, store, and analyze data from spaceflight and all space-relevant experiments and studies (Ray et al., 2019; Ramachandran et al., 2021; Berrios et al., 2021). NASA aims to enable the return of humans to the Moon and prepare the first crewed missions to Mars. There are intricate hazards to astronauts' health caused by the existing environmental stressors (e.g., altered gravity, exposure to radiation, confinement, hostile and closed environments, changes in diet and distance) during spaceflight (Overbey et al., 2024; Mason et al., 2024; Rutter et al., 2024). Studies have shown that altered gravitational conditions have detrimental effects on both human and mouse physiology, particularly on the musculoskeletal system (Morey-Holton, 2003; Adamopoulos et al., 2021). Musculoskeletal tissue is profoundly influenced by its mechanical loading environment, a fact underscored by observed skeletal muscle atrophy and bone mass loss. Microgravity also causes a decrease

in bone formation by osteoblasts and an increase in osteolytic functions of osteoclasts (Allen et al., 2009; Gambari et al., 2017; Afshinnekoo et al., 2020; Chakraborty et al., 2021; Man et al., 2022; Blottner et al., 2022; Reynolds et al., 2022).

Afshinnekoo et al. highlighted that the biological responses during spaceflight arise from six fundamental molecular and cellular features, including oxidative stress, DNA damage, mitochondrial dysregulation, epigenetic changes, and alterations of telomere length (Afshinnekoo et al., 2020).

There are still many unknown mechanisms regarding how human physiology is affected during long-term deep space exploration missions (Liphardt et al., 2023). It is a significant challenge to accurately map animal findings to humans, as most data originates from model organisms. Nonetheless, rodent models are indispensable for advancing our understanding of these effects, given their established use and physiological similarities in simulating microgravity conditions (Moyer et al., 2016; Cahill et al., 2021; Ronca and Lowe, 2022).

Some stressors (e.g., confinement) are comparatively straightforward to replicate, while altered gravity and radiation

[†]Corresponding author: Sylvain V. Costes
E-mail: sylvain.v.costes@nasa.gov

are considerably more challenging to simulate by spaceflight analogues on Earth. Microgravity analogues, such as random positioning machines, clinostats, short-term free-fall intervals during parabolic flight, head-down bed rest, hindlimb unloading rodent models, and rotating wall vessels, have been devised to simulate infrequent altered gravity conditions. Although the analogues often fall short of fully capturing the nuanced effects of spaceflight. These limitations underscore the need for continued refinement and innovation in analogue development (Morey-Holton and Globus, 2002; Brungs et al., 2016; Kiss et al., 2019; Afshinnekoo et al., 2020).

In our present study, we initially conducted an extensive differential expression analysis of OSDR data from *Mus musculus* and *Homo sapiens* samples exposed to spaceflight stressors, aiming to detect commonly dysregulated orthologous genes, and subsequently identify enriched terms and pathways related to the significantly dysregulated genes within each species and across orthologous genes. Following this analysis, we constructed AI-ready datasets for *Mus musculus* and *Homo sapiens*, derived from cross-platform microarray experiments, including the fundamental features of gene expression under microgravity exposure. These datasets were identified by compiling an exhaustive list of microgravity microarray experiments from OSDR. Subsequently, employing a variety of machine learning models, we classified clusters of genes exhibiting differential expression levels under altered gravitational conditions, leveraging the aforementioned gene features. Lastly, we demonstrated the efficacy of transfer learning by initially pretraining our model on the larger mouse dataset to acquire gene features affected by microgravity exposure in general, followed by refinement on the smaller human dataset, enabling us to leverage information from model organism studies to gain insight into microgravity effects on humans.

Methods

Data Selection

The overall study workflow is presented in Figure 1. Firstly, we systematically identified all the available microarray datasets and studies housed in NASA Open Science Data Repository (OSDR) by applying the search filters of “Microarray”, “*Homo sapiens* AND Rodent, *Mus musculus*” as “Assay Type” and “Organisms” respectively. We focused on studies and experiments meeting the following inclusion criteria: 1. Derived from musculoskeletal tissue of *Homo sapiens* or *Mus musculus* organisms, given the abundance of comparable tissues within these species, 2. Implemented a case-control study design, 3. Utilized untreated samples. This term refers to samples that have not been subjected to experimental interventions, such as drug treatments. Conducted actual or simulated microgravity vs control group comparisons. The

selected datasets are as follows. For *Mus musculus* OSD-111 (<https://doi.org/10.26030/9580-9n52>, soleus and extensor digitorum longus), OSD-228 (<https://doi.org/10.26030/bfm-x866>, soleus), OSD-227 (<https://doi.org/10.26030/vk0m-0558>, soleus and gastrocnemius), OSD-135 (<https://doi.org/10.26030/rjyq-x751>, longissimus dorsi and tongue), and OSD-21 (<https://doi.org/10.25966/c36b-3g68>, gastrocnemius, and all parts of calf) and for *Homo sapiens* OSD-51 (<https://doi.org/10.26030/3w73-jn41>, soleus and vastus lateralis), OSD-195 (<https://doi.org/10.26030/r6bv-rk07>, vastus lateralis), and OSD-370 (<https://doi.org/10.26030/zyy0-6497>, vastus lateralis) (Berrios et al., 2021).

Data Preprocessing

The data preprocessing and analysis of differential expression were performed using the R programming language within the R Bioconductor environment. Following retrieval of the raw microarray expression data for each dataset from NASA OSDR, we performed normalization utilizing affy, oligo (for Affymetrix microarray experiments), or limma (for Agilent) packages (Gautier et al., 2004; Carvalho and Irizarry, 2010; Ritchie et al., 2015). Outlier samples were identified and removed through Quality Control visualization methods, including diagnostic plots such as PCA plots, boxplots of array intensity distributions, density of raw intensities plots, and MA plots. Official Gene Symbols, as designated by the Human Genome Organization Gene Nomenclature Committee, were employed as the identifiers across experiments and platforms.

Differential Expression Analysis

The Differentially Expressed Genes (DEGs) between the microgravity and control group were identified using the limma R package (Ritchie et al., 2015). The Benjamini-Hochberg method was used to adjust the p-value (Benjamini and Hochberg, 1995). In particular, the identification of DEGs was achieved through the following criteria: 1. Adjusted p-value threshold of below 0.05, and 2. $|\log_2FC|$ (absolute value of logarithm to the base 2 of fold change) threshold above 1. Furthermore, Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways enrichment analyses were performed using the bioinformatics resource system DAVID (Database for Annotation, Visualization and Integrated Discovery) (Huang et al., 2009; Sherman et al., 2022).

AI-Ready Merged Datasets

Data obtained from the screening of Differentially Expressed Genes (DEGs) across individual datasets and tissues were integrated into new data frames, resulting in two merged meta-datasets, each corresponding to a species. These meta-datasets included adjusted p-value, Gene Symbol, and

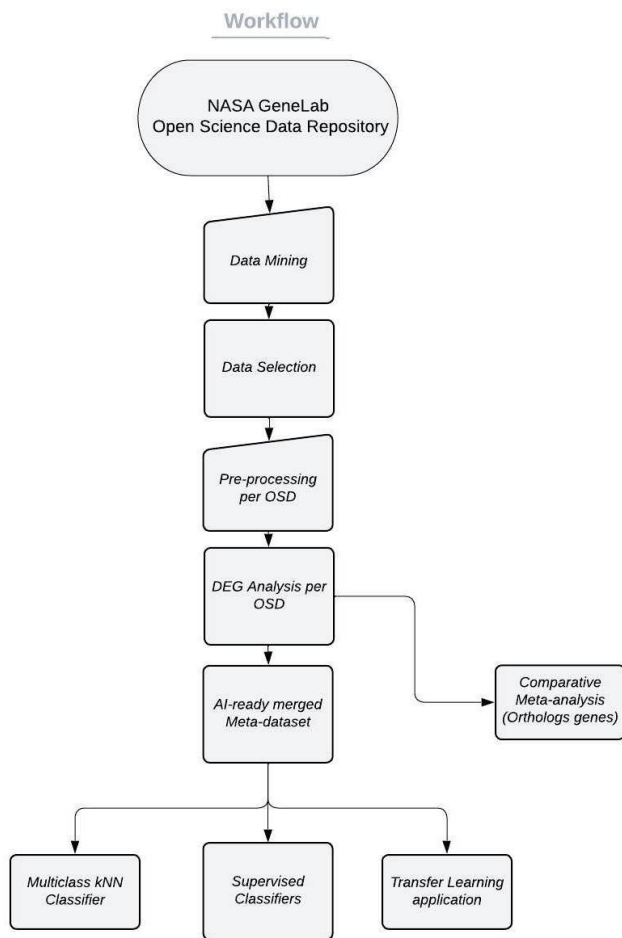


Figure 1. Overall study workflow.

log2FC value as primary features. Subsequently, the following features were added to investigate as predictors in our Machine Learning models: Gene Ontology ID, gravitational method (actual or simulated), gravitational exact method (e.g., spaceflight, bed rest analogue, hindlimb suspension etc.), type of tissue (e.g., soleus, vastus lateralis, gastrocnemius, etc.), microarray platform (GPL6480 Agilent, GPL17692 Affymetrix etc.), Chromosome annotation, and time of exposure in microgravity (in hours). To address imbalanced datasets, resampling techniques such as oversampling of minority classes and the Synthetic Minority Over-sampling Technique (SMOTE) algorithm were applied (Ramachandran et al., 2021). Categorical features were encoded into numeric representations using the OneHotEncoder method from the scikit-learn library (Pedregosa et al., 2011). These structured, preprocessed, and feature-enriched meta-datasets collectively define what is termed an “AI-ready dataset,” optimized for training robust and interpretable Machine Learning models.

Machine Learning Analysis

After data preparation and pre-processing procedures, we proceeded to train multiple supervised Machine Learning models aimed at developing effective classifiers for gene expression levels. Logistic Regression, Decision Tree, Support Vector Machine (SVM), Random Forest, Naïve Bayes, and Artificial Neural Network Classifiers were employed to classify whether a gene was overexpressed or underexpressed. Additionally, we conducted a multi-class classification using K-Nearest Neighbors to predict whether a gene is over-expressed, under-expressed, or non-significantly differentially expressed in microgravity. In the context of our Transfer Learning approach, we initially pretrained our model on mouse datasets and then fine-tuned it using a comparatively smaller human target dataset. To reduce dimensionality and facilitate analysis, Principal Component Analysis (PCA) was applied to both the mouse and human datasets, reducing the feature dimensionality of each dataset to the top 100 principal components, which were then used as input to the classifiers. The performance of these classifiers was evaluated and compared using metrics such as accuracy, precision, and recall. Following hyperparameter tuning of the Neural Network architecture, we incorporated three layers with activation functions: ReLU, SELU, and Sigmoid, respectively. The Nadam optimizer and binary cross-entropy loss function were employed during compilation. Five-fold cross-validation ($k = 5$) was implemented for robust evaluation. Our dataset was partitioned into train and test sets using an 80-20% split ratio. We utilized a linear kernel for the SVM model, configured the Random Forest classifier with seventy estimators, and applied Gaussian Naïve Bayes to the mouse dataset, while employing the Bernoulli Naïve Bayes model for the human dataset.

Machine Learning models were implemented using the Python programming language, with Jupyter Notebook serving as the web-based interactive computing platform (Van Rossum and Drake, 2009). The TensorFlow library and Keras were employed to execute the Neural Networks (Developers, 2024). Furthermore, the scikit-learn and seaborn libraries were utilized to generate classification reports, confusion matrices, and heatmaps. Pandas, NumPy, and Matplotlib were also utilized for data processing, numerical computations, and data visualization tasks, respectively (Hunter, 2007; Harris et al., 2020; Waskom, 2021).

Results

Enriched terms and pathways of significantly dysregulated genes in *Mus musculus*

We present below (Figures 2 and 4) heatmaps depicting the differential expression profiles of genes identified through a

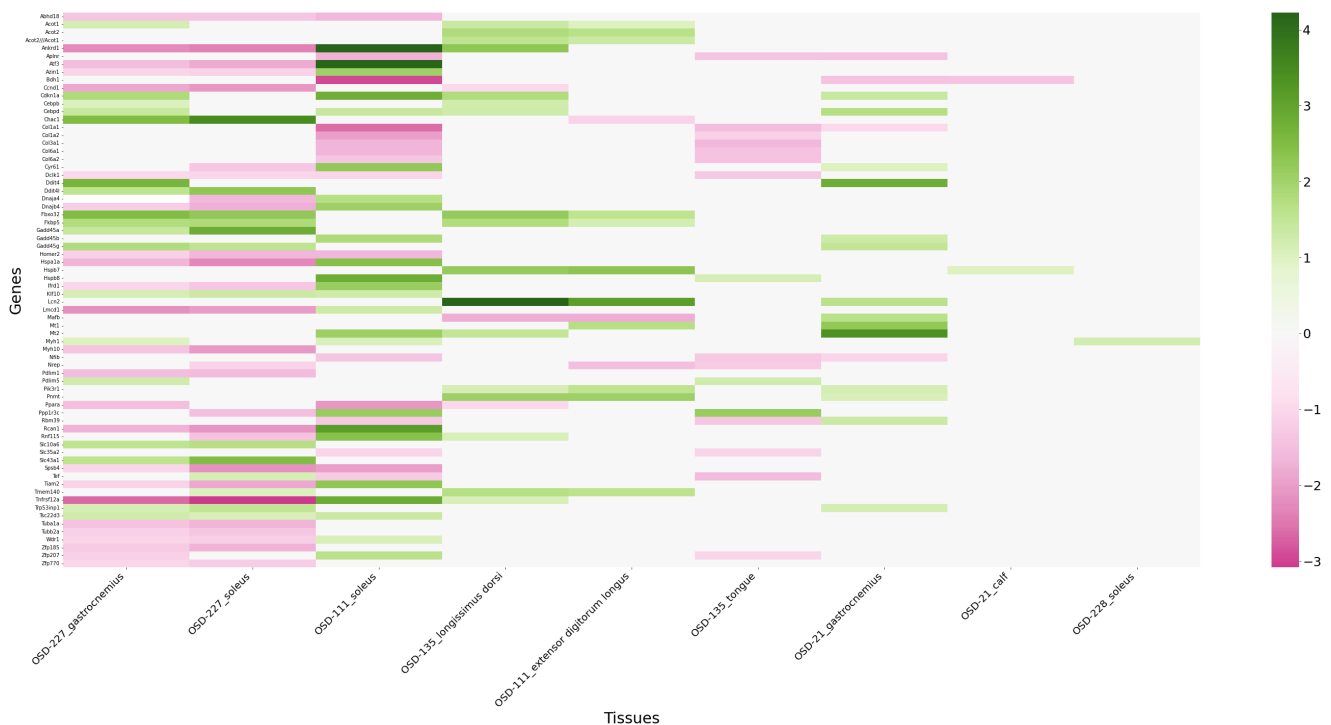


Figure 2. Heatmap of significantly dysregulated genes in *Mus musculus*. Genes that are shaded in dark green indicate high up-regulation with a logFC value above 1.5, while genes shaded in light green indicate up-regulation with a logFC value above 1. Genes with a logFC value between -1 and 1 are non-significantly dysregulated. Conversely, genes shaded in dark red represent highly down-regulated genes with a logFC below -1.5. The analyses and results depicted in the Heatmap were conducted with consideration only for the Top 50 differentially expressed genes per study or dataset.

DEG analysis of *Mus musculus* and *Homo sapiens* datasets, comparing microgravity conditions with controls. The genes exhibiting upregulation across multiple *Mus musculus* tissues include *Acot1*, *Acot2*, *Cdkn1a*, *Cebpb*, *Cebpd*, *Ddit4*, *Ddit4l*, *Fbxo32*, *Fkbp5*, *Gadd45a*, *Gadd45b*, *Gadd45g*, *Hspb7*, *Hspb8*, *Klf10*, *Lcn2*, *Mt1*, *Mt2*, *Myh1*, *Pdlim5*, *Pik3r1*, *Pnmt*, *Slc10a6*, *Slc43a1*, *Tmem140*, *Trp53inp1*, and *Tsc22d3*. Enrichment analysis revealed that *Gadd45g*, *Gadd45a*, *Gadd45b*, *Cdkn1a*, and *Pik3r1* were associated with various KEGG pathways, including FoxO signaling pathway, endometrial cancer, melanoma, non-small cell lung cancer, glioma, pancreatic cancer, chronic myeloid leukemia, colorectal cancer, small cell lung cancer, breast cancer, gastric cancer, hepatocellular cancer, cellular senescence, and Epstein-Barr virus infection. Additionally, biological processes enriched included positive regulation of the p38 MAPK and JNK cascade, negative regulation of protein kinase activity, and positive regulation of the apoptotic process involving genes *Gadd45g*, *Gadd45a*, *Gadd45b*, *Lcn2*, and *Trp53inp1*.

The downregulated genes observed across multiple *Mus musculus* tissues, including *Bdh1*, *Col1a1*, *Ppara*, *Col1a2*, *Spsb4*, *Aplnr*, *Col6a1*, *Col3a1*, *Homer2*, *Abhd18*, *Nfib*, *Col6a2*, *Dclk1*, *Ccnd1*, *Myh10*, *Nrep*, *Pdlim1*, *Tuba1a*,

Tubb2a, *Zfp185*, and *Zfp770*, were associated with several enriched terms and pathways. These pathways include protein digestion and absorption (involving *Col1a1*, *Col1a2*, *Col3a1*, *Col6a1*, and *Col6a2*), focal adhesion (involving *Col1a1*, *Col1a2*, *Col6a1*, *Col6a2*, and *Ccnd1*), ECM-receptor interaction (involving *Col1a1*, *Col1a2*, *Col6a1*, and *Col6a2*), AGE-RAGE signaling (involving *Col1a1*, *Col1a2*, *Col3a1*, *Ccnd1*), pathways in diabetic complications and diabetic cardiomyopathy (involving *Col1a1*, *Col1a2*, *Col3a1*, and *Ppara*), human papillomavirus infection and the PI3K-Akt signaling pathway (involving *Col1a1*, *Col1a2*, *Col6a1*, *Col6a2*, and *Ccnd1*). Moreover, enriched biological processes include blood vessel development (involving *Col1a1*, *Col1a2*, *Col3a1*, and *Aplnr*), and neuron migration (involving *Col3a1*, *Dclk1*, *Myh10*, *Tuba1a*, and *Tubb2a*). Molecular functions such as extracellular matrix structural constituent conferring tensile strength, platelet-derived growth factor binding, nucleotide binding (involving *Dclk1*, *Myh10*, *Tuba1a*, *Tubb2a*), and metal ion binding (involving *Pdlim1*, *Col1a1*, *Col1a2*, *Col3a1*, *Ppara*, *Tuba1a*, *Tubb2a*, *Zfp185*, and *Zfp770*) were also enriched.

Moreover, nineteen genes exhibited a dynamic behavior, displaying upregulation in certain tissues while being

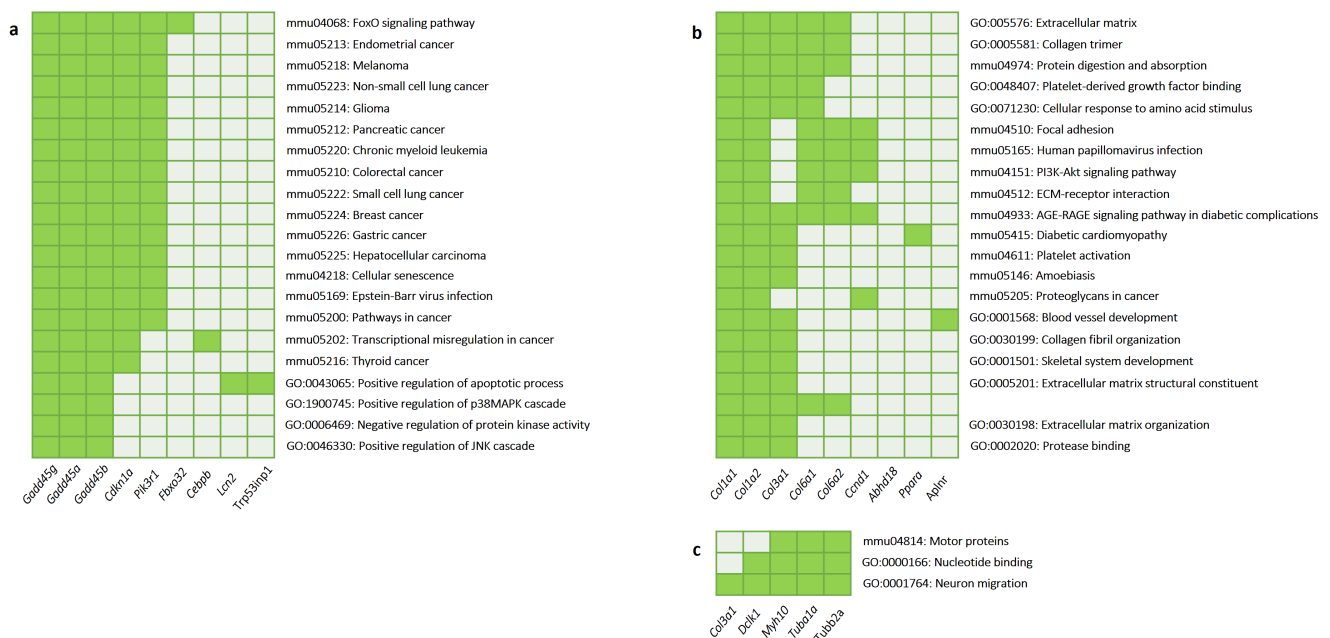


Figure 3. Enriched terms and pathways of dysregulated genes in *Mus musculus*. Panel a presents the enriched terms of upregulated genes, while panels b and c display the enriched terms of downregulated genes. Corresponding enriched Gene Ontology terms and KEGG pathways are depicted on the right side of each panel, with associated genes shown at the bottom. The dark green color indicates the correlations between genes and enriched terms or pathways.

downregulated in others, thus lacking consistent expression patterns. Among these genes, *Mafk*, *Tef*, *Atf3*, and *Ankrd1* were identified as being associated with positive regulation of transcription from RNA polymerase II promoters, as well as terms related to identical protein binding and DNA binding enrichment.

Enriched terms and pathways of significantly dysregulated genes in *Homo sapiens*

Differential expression analysis in *Homo sapiens* revealed common upregulation of *CALM1*, *CALM2*, *CXCL14*, *MATN2*, and *EEF1A1* across OSD-51 soleus and vastus lateralis tissues. The aforementioned genes were associated with the enhancement in the molecular function of protein kinase binding. Subsequently, enriched terms related to cellular components included cytosol, cytoplasm, nucleus, and extracellular region. In OSD-195 and OSD-370 vastus lateralis tissues, the upregulation of *CHAD*, *TSPYL2*, *CHRND*, and *RRAD* was observed, while *NME9*, *CCDC39*, and *KCNJ16* exhibited significant upregulation specifically in OSD-370. *CALM1*, *CALM2*, *EEF1A1*, *RRAD*, *CHRND*, and *KCNJ16* were identified in the plasma membrane-enriched term. *CALM1*, *CALM2*, and *KCNJ16* were also associated with gastric acid secretion and the voltage-gated potassium channel complex. Furthermore, *CALM1* and *CALM2* were implicated in various pathways, including melanogenesis, cellular senescence, and

long QT syndrome, according to DAVID UniprotKB Keywords. *CXCL14* and *MATN2* genes were linked to kidney aging, while *CALM1*, *CALM2*, and *MATN2* were correlated with Type 2 Diabetes and edema according to the Genetic Association Database (GAD) (Becker et al., 2004). Lastly, dysregulation of *CALM1*, *CALM2*, and *CASQ2* genes was associated with ventricular tachycardia, according to DisGeNET (Bauer-Mehren et al., 2010).

On the other hand, a significant downregulation was observed in twenty-one genes, namely *ABP3*, *COLQ*, *MYOZ2*, *CASQ2*, *ACOT11*, *CNNM4*, *PPP1R1C*, *MYL12A*, *EFR3B*, *PLIN5*, *TMEM108*, *DNAJA4*, *TP53INP2*, *NMRK2*, *HMGCS2*, *SLC26A9*, *KCNC1*, *KLHL34*, *KLHL40*, *COQ10A*, and *CA14*. This under-expression was accompanied by enrichment in cellular component terms such as Z-disc, cytosol, and cytoplasm. Biological processes enriched in the analysis included responses to temperature, abiotic stimuli, and responses to stress, while molecular function enrichment revealed a predominance of protein binding, with fifteen out of twenty-one downregulated genes associated with this function. *HMGCS2*, *FABP3*, and *PLIN5* were implicated in the enriched PPAR signaling pathway. *DNAJA4*, *EFR3B*, *COLQ*, and *PPP1R1C* were correlated with body height according to GAD, while *CASQ2* and *MYOZ2* were associated with hypertrophic cardiomyopathy according to DisGeNET.

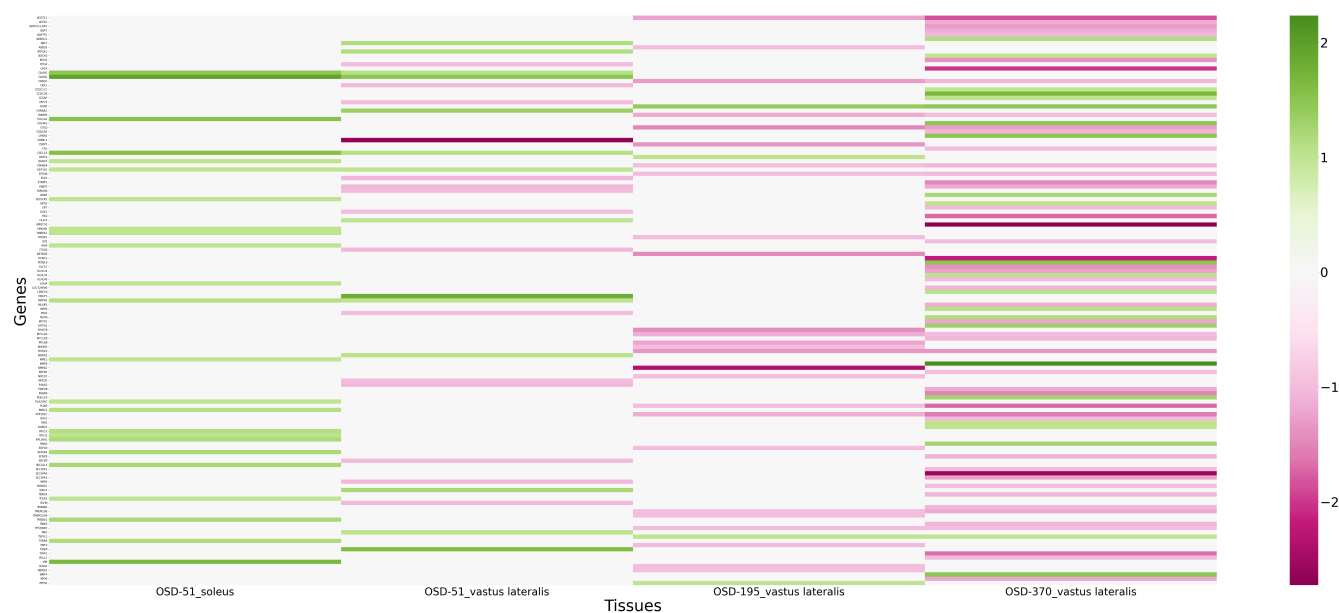


Figure 4. Heatmap of significantly dysregulated genes in Homo sapiens. Genes that are shaded in dark green indicate high up-regulation with a logFC value above 1.5, while genes shaded in light green indicate up-regulation with a logFC value above 1. Genes with a logFC value between -1 and 1 are considered to be non-significantly dysregulated. Conversely, genes shaded in dark red represent highly down-regulated genes with a logFC below -1.5. The analyses and results depicted in the Heatmap were conducted with consideration only for the Top 50 differentially expressed genes per study or dataset.

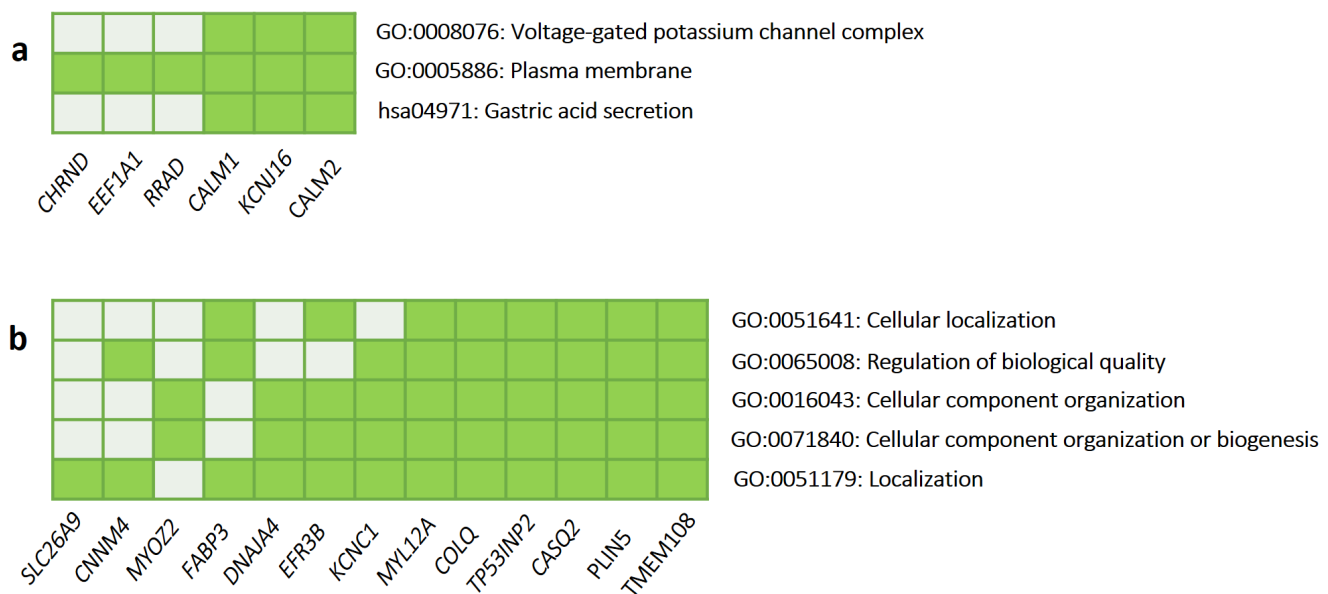


Figure 5. Enriched terms and pathways of dysregulated genes in Homo sapiens. Panel a presents the enriched terms of upregulated genes, while panel b displays the enriched terms of downregulated genes. Corresponding enriched Gene Ontology terms and KEGG pathways are depicted on the right side of each panel, with associated genes shown at the bottom. The dark green color indicates the correlations between genes and enriched terms or pathways.

Comparative analysis of orthologous genes

We present a comparative meta-analysis of orthologous genes between *Homo sapiens* and *Mus musculus* organisms, aiming to identify common dysregulations across species and potentially conserved biological pathways. A total of eight orthologous genes were found to be upregulated, namely *ARRDC4*, *ART3*, *CHRNA1*, *DDIT4*, *RPL12*, *TXNIP*, *WNT4*, and *ZFP36*. Among these, the cytoplasm was identified as the cellular component related to six out of eight upregulated orthologous genes (*DDIT4*, *WNT4*, *ZFP36*, *ARRDC4*, *RPL12*, and *TXNIP*). Furthermore, *DDIT4*, *ZFP36*, *ARRDC4*, *RPL12*, and *TXNIP* were associated with the molecular function of protein binding, while *DDIT4* and *ZFP36* are related to 14-3-3 protein binding molecular function. *DDIT4* and *WNT4* were implicated in the mTOR signaling pathway and the biological process of neuron differentiation. Additionally, *DDIT4* and *TXNIP* were linked to rheumatoid arthritis according to DisGeNET.

Conversely, a total of nineteen orthologous genes were found to be downregulated, namely *ACSS1*, *BDH1*, *CASQ2*, *CNNM4*, *CSRP3*, *DNAJA4*, *EFR3B*, *FABP3*, *KCNC1*, *MLXIPL*, *MTFP1*, *MYL12A*, *NMRK2*, *P4HA2*, *PAQR9*, *SLC16A1*, *TBRG4*, *TGFBI*, and *VGLL2*. Downregulated genes such as *BDH1*, *CNNM4*, *NMRK2*, *P4HA2*, and *SLC16A1* were correlated with the enriched cellular component term of intracellular membrane-bounded organelle. *BDH1*, *ACSS1*, and *TBRG4* were associated with the mitochondrial matrix, while *CASQ2*, *CSRP3*, and *MYL12A* were related to the Z-disc. *MLXIPL*, *CSRP3*, and *SLC16A1* were correlated with the enriched term of glucose homeostasis, while *CSRP3* and *NMRK2* were associated with negative regulation of myoblast differentiation. Subsequently, *CASQ2* and *CSRP3* were related to the enriched term of cardiac muscle contraction, while *VGLL2* and *CSRP3* were associated with skeletal muscle tissue development. According to the GAD Database, several genes, including *ART3*, *EFR3B*, *CHRNA1*, *CSRP3*, *FABP3*, *KCNC1*, *SLC16A1*, *TXNIP*, and *TGFBI*, were correlated with Type 2 Diabetes, while *BDH1*, *ACSS1*, *SLC16A1*, and *TBRG4* were related to acquired immunodeficiency syndrome progression. Lastly, *TXNIP* and *TGFBI* were associated with nodular glomerulosclerosis and diabetic nephropathy according to DisGeNET.

Machine Learning applications

After the identification of dysregulated genes, two AI-ready merged datasets were produced, one for each species, integrating essential features such as Benjamini-Hochberg adjusted p-value and binary logarithm of Fold Change value for differential gene expression between microgravity conditions and controls, Official Gene Symbol, , Gene Ontology ID, gravitational method (i.e. actual or simulated gravity), gravitational exact method (e.g. spaceflight, bed

rest analogue, hindlimb suspension etc.), specific type of musculoskeletal tissue (e.g. longissimus dorsi, extensor digitorum longus soleus, vastus lateralis, gastrocnemius etc.), Chromosome annotation, hours of exposure in microgravity. Following data preparation, we employed various supervised Machine Learning models to develop efficient classifiers for gene expression levels. Specifically, classifiers including Logistic Regression, Decision Trees, Support Vector Machines, Random Forests, Naïve Bayes, and Neural Networks were utilized for classifying over-expressed and under-expressed gene clusters. Subsequently, we conducted multi-class classification using the K-Nearest Neighbors (KNN) approach to predict whether a gene is over-expressed, under-expressed, or non-significantly differentially expressed. Ultimately, we utilized a Transfer Learning methodology to explore effectiveness of transferring knowledge from *Mus musculus* organism to *Homo sapiens*. To achieve this, our model was pretrained on the mouse merged dataset, followed by fine-tuning on the human target merged dataset. The performance of these classifiers was evaluated and compared based on accuracy, recall, and precision metrics. Relevant classification reports and confusion matrices are provided (Figures 6, 8).

In summary, concerning the *Mus musculus* organism, the Random Forest and Neural Network classifiers showcase superior performance, achieving accuracy rates of 77% and 79%, respectively. The Support Vector Machine (SVM) and Logistic Regression classifiers also exhibit satisfactory performance, with accuracy rates of 76.1% and 76.9%, respectively. Meanwhile, the Decision Tree and Gaussian Naïve Bayes classifiers show lower accuracy rates, at 73%. The overall accuracy of the KNN multi-class model stands at 74%. It seems that the classifier performs best for class 1 (not significantly dysregulated class), achieving an F1 score of 78%. KNN classifier also demonstrates better performance for class 0 (under-expressed genes) at 76%, compared to class 2 (over-expressed genes) at 69% (Figures 6, 8).

Feature importance rankings can provide valuable insights into the factors that influence a classifier's predictions, aiding in feature selection and interpretation. Here, we present the feature importance plots of Random Forest and Decision Tree models applied to the *Mus musculus* dataset (Supplementary file). We observe that both classifiers share certain important features, including exposure time in microgravity, soleus as a specific type of musculoskeletal tissue, and a portion of Gene Ontology (GO) terms. Among these, GO:0008201 corresponds to the biological process' response to oxygen levels', GO:0044822 refers to the molecular function' protein domain specific binding', GO:0004842 represents the molecular function' ubiquitin-protein transferase activity', and GO:0005515 indicates the molecular function of 'protein binding'.

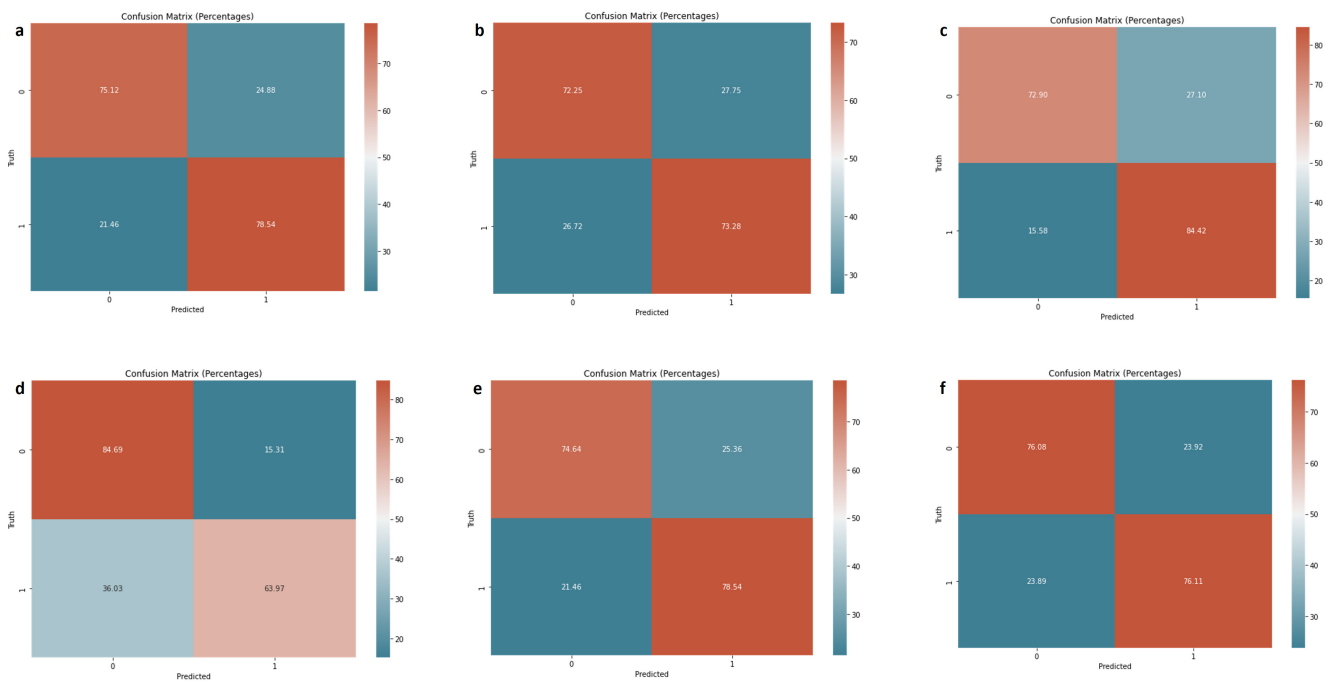


Figure 6. Confusion Matrices of *Mus musculus*' classifiers. a) Logistic Regression, b) Decision Tree, c) Neural Network, d) Gaussian Naive Bayes, e) Random Forest, f) SVM Support Vector Machine, SVC (kernel = linear). The y-axis corresponds to the true expression level, while the x-axis corresponds to the predicted expression level. Class 0 represents down-regulated genes, and class 1 represents up-regulated genes.

Regarding the *Homo sapiens* organism, the SVM classifier demonstrates satisfactory performance, achieving an accuracy rate of 72%. The Random Forest, Bernoulli Naïve Bayes, and Logistic Regression classifiers exhibit accuracy rates of 66%, 69%, and 66%, respectively. Conversely, the Decision Tree and Neural Network classifiers display comparatively lower accuracy rates, at 59% and 62%, respectively. The overall accuracy of the KNN multi-class model stands at 69%. The classifier exhibits an F1 score of 69% in predicting class 0 (under-expressed genes), while demonstrating better performance for class 1 (not significantly dysregulated class) at 67%, compared to class 2 (over-expressed genes) at 53% (Figures 7, 8).

To improve the performance of our Neural Network classifier, initially yielding 62% accuracy on the human dataset, we investigated the application of Transfer Learning techniques. Specifically, we pre-trained our model on the larger mouse dataset to assimilate common features before refining it with the smaller human dataset. Following this, our Transfer Learning model exhibited an improved accuracy of 71% (Figure 8c). Notably, the recall score was higher (at 80%) for the under-expressed class compared to precision, while the precision score was superior (at 78%) for the over-expressed class compared to recall. Figure 9 shows confusion matrices for KNN multi-classifiers and Transfer Learning.

Discussion

Enriched terms and pathways associated with significantly dysregulated genes were identified within each species. Regarding *Mus musculus*, the upregulation of genes across multiple tissues involved in resistance to apoptosis, dysregulation of cell proliferation, and altered signaling pathways suggests a common theme of disrupted cell cycle control, increased survival signals, and cancer development and progression. The downregulated genes were linked to the development of diabetic complications. Specifically, dysregulated AGE-RAGE signaling is linked to chronic inflammation, endothelial cell dysfunction, and oxidative stress in diabetes. Furthermore, focal adhesion disruptions, alterations in ECM-receptor and HPV interactions, and abnormal activation of the PI3K-Akt signaling pathway are often correlated with cancer development.

As far as *Homo sapiens* is concerned, the over-expressed genes are involved in melanogenesis, cellular senescence, long QT syndrome, and diabetes. Type 2 Diabetes is related to defective tissue repair and systemic complications such as cardiovascular arrhythmias and ventricular tachycardia. On the other hand, under-expressed genes were associated with hypertrophic cardiomyopathy, impaired response to temperature and abiotic stresses, while dysregulation of the

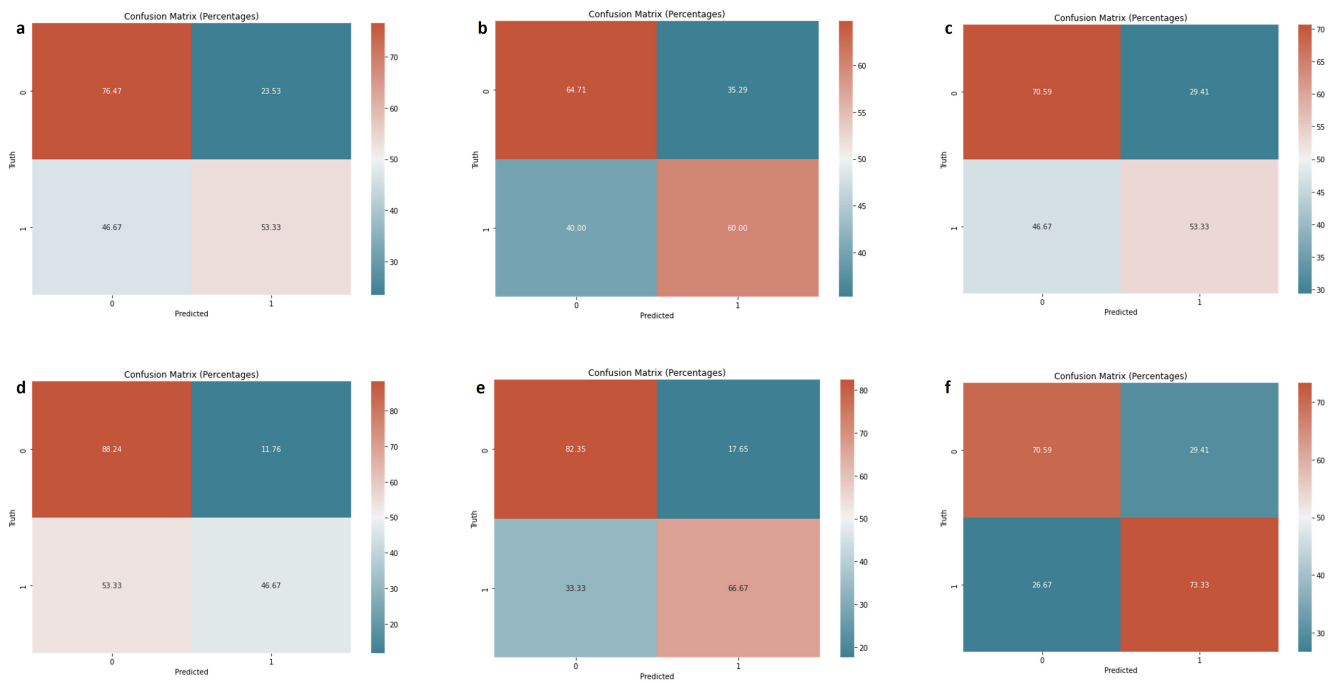


Figure 7. Confusion Matrices of Homo sapiens' classifiers. a) Logistic Regression, b) Decision Tree, c) Neural Network, d) Gaussian Naive Bayes, e) Random Forest, f) SVM Support Vector Machine, SVC (kernel = linear). The y-axis corresponds to the true expression level, while the x-axis corresponds to the predicted expression level. Class 0 represents down-regulated genes, and class 1 represents up-regulated genes.

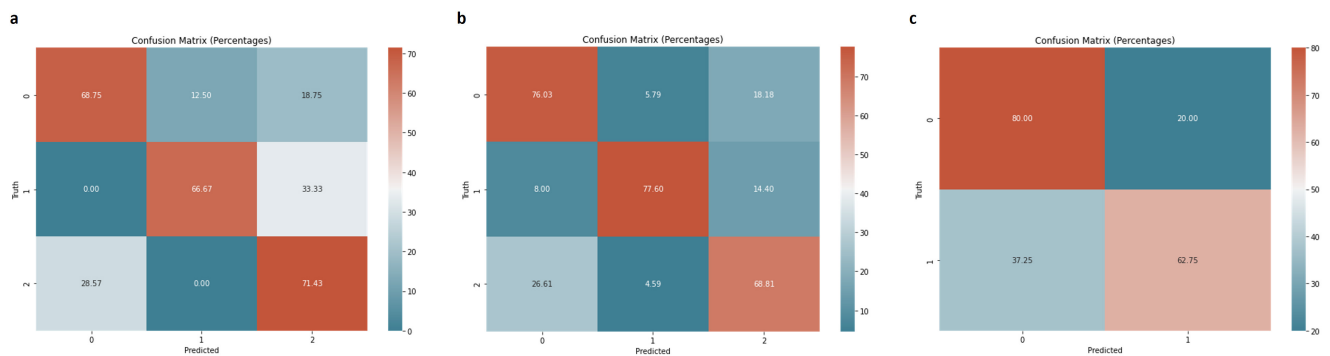


Figure 8. Classification reports of a) Mus musculus' and b) Homo sapiens' classifiers, c) Classification Report of Transfer Learning application. The classification report for our Transfer Learning model reveals an enhanced accuracy of 71% compared to the NN classifier on the Homo sapiens dataset. The performances of these classifiers were evaluated and compared based on accuracy, recall, and precision metrics.

PPAR signaling pathway affects cell metabolism, energy homeostasis, and immune response inhibition.

Following the comparative meta-analysis of orthologous genes between human and mouse organisms, we identified common dysregulations across species, leveraging the high conservation of orthologous genes across evolutionary time. The upregulated orthologous genes were primarily involved in cellular growth, metabolic regulation, and immune responses, while the downregulation of nineteen orthologous genes was linked to kidney-related complications, Type 2 Diabetes, and pathways such as glucose metabolism

dysfunction and insulin resistance. The identification of these dysregulated orthologous genes holds potential for elucidating conserved biological pathways such as impaired muscle and skeletal development, cardiovascular functions, metabolic homeostasis, and inflammatory and stress response.

Following the identification of enriched terms and pathways for the dysregulated genes, we constructed AI-ready merged datasets, one per species, to implement multiple supervised Machine Learning models and develop efficient classifiers. Random Forest and Neural Network classifiers showcase higher performance on the *Mus musculus* dataset, achieving

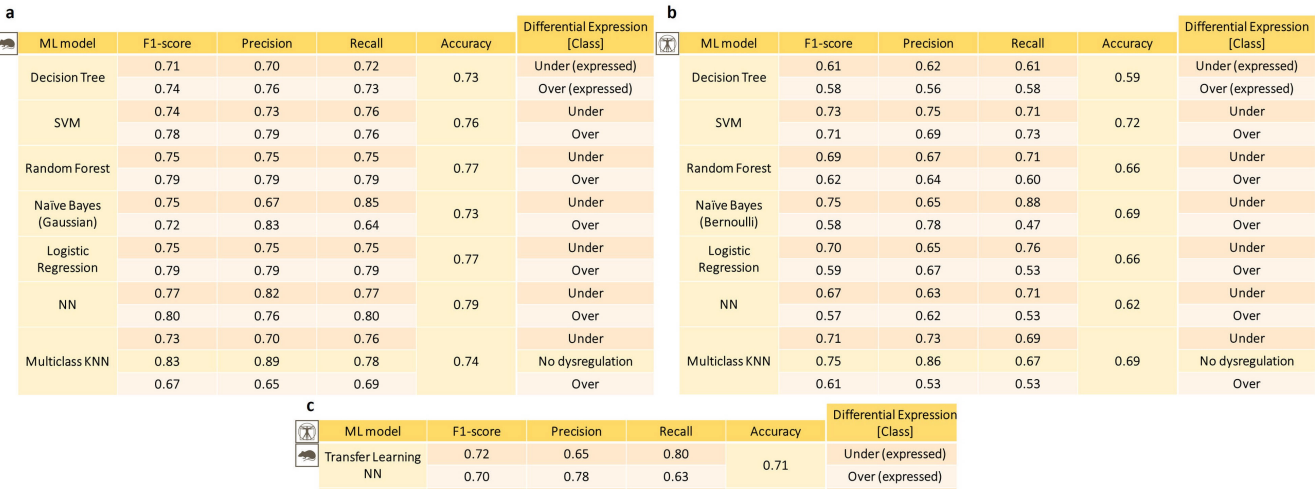


Figure 9. Confusion Matrices of KNN multi-classifiers and Transfer Learning application. a) KNN multi-classifier on human dataset, b) KNN multi-classifier on mouse dataset, class 0 represents down-regulated genes, class 1 represents non-significantly dysregulated genes, and class 2 represents up-regulated genes. c) Confusion matrix of Transfer Learning application, class 0 represents down-regulated genes, and class 1 represents up-regulated genes. The y-axis corresponds to the true expression level, while the x-axis corresponds to the predicted expression level.

accuracy rates of 77% and 79%, respectively. The overall accuracy of the KNN multi-class model stands at 74%, performing best in predicting the class of the not significantly dysregulated class, achieving an F1 score of 78%. On the other hand, only the SVM classifier demonstrates a satisfactory performance on *Homo sapiens* dataset, achieving an accuracy rate of 72%. The evaluation metrics presented indicate superior performance of the classifiers on the mouse dataset compared to the human dataset. Notably, the human dataset constitutes only approximately 10% of the data when compared to the mouse dataset, which likely contributes to the inferior performance of its classifiers. Therefore, we explored the application of Transfer Learning since we aimed to enhance the performance of the Neural Network classifier, initially achieving 62% accuracy on *Homo sapiens* data. We pre-trained our model using mouse data and subsequently fine-tuned it using human data. Our model was first pretrained on the mouse dataset to learn features that are common to that data modality, and then refined using a smaller human target dataset. This approach led to a notable improvement in accuracy (at 71%).

Conclusion

In conclusion, this study highlights the significant biological implications of altered gravitational conditions on gene expression and the dysregulation of pathways. The findings in *Mus musculus* indicate a predisposition to cancer, disrupted tissue homeostasis, and an increased risk of inflammation, oxidative stress, and diabetic complications.

Similarly, in *Homo sapiens*, the results highlight systemic metabolic disruptions, and cardiovascular vulnerabilities, including arrhythmias and impaired tissue repair mechanisms. Comparative analysis across species reveals conserved pathways affecting muscle and skeletal integrity, metabolic homeostasis, and inflammatory responses. These biological implications emphasize the necessity for targeted countermeasures to mitigate systemic and long-term health risks associated with altered gravitational conditions during extended spaceflight. Additionally, this study highlights the potential of leveraging Machine Learning and Transfer Learning methodologies to overcome the challenges of analyzing species-specific datasets in the context of space research. The application of Transfer Learning provides valuable insights into bridging interspecies gaps by facilitating the effective transfer of knowledge from model organisms to humans, addressing the limitations posed by the scarcity of human data. To further advance the field, future research should prioritize standardizing designs to minimize heterogeneity, expanding the diversity and size of datasets, incorporating longitudinal studies to capture temporal dynamics, and leveraging advanced computational methodologies such as multi-omics integration, Transfer Learning, and explainable AI models.

Data availability

Data used in the preparation of the current research article were obtained from the NASA GeneLab Open Science Data Repository (OSDR; osdr.nasa.gov/bio).

(Accession codes, DOIs): (OSD-111, <https://doi.org/10.26030/9580-9n52>), (OSD-228, <https://doi.org/10.26030/bfm-x866>), (OSD-227, <https://doi.org/10.26030/vk0m-0558>), (OSD-135, <https://doi.org/10.26030/rjyq-x751>), (OSD-21, <https://doi.org/10.25966/c36b-3g68>), (OSD-51, <https://doi.org/10.26030/3w73-jn41>), (OSD-195, <https://doi.org/10.26030/r6bv-rk07>), (OSD-370 (<https://doi.org/10.26030/zyy0-6497>)).

Code availability

The code for the bioinformatics methods can be made available upon reasonable request to the corresponding author.

Acknowledgments

This work was supported by both the Fulbright Foundation in Greece and the Fulbright Program fellowship.

Author Disclosure Statement

No competing financial interests exist.

Author Contributions

Conceptualization, K.I.A., L.M.S., A.H. and S.V.C.; methodology, K.I.A. and S.V.C.; investigation K.I.A., L.M.S., A.H. and S.V.C.; writing—original draft preparation, K.I.A. and L.M.S.; data curation, K.I.A.; formal analysis, K.I.A., L.M.S., A.H. and S.V.C.; algorithms implementation, K.I.A. and L.M.S.; writing—review and editing, K.I.A., L.M.S., A.H. and S.V.C.; supervision, S.V.C.; project administration, S.V.C.; All authors have read and agreed to the published version of the manuscript.

Competing Interests

The authors declare no competing interests.

References

- Adamopoulos K, Koutsouris D, Zaravinos A, Lambrou GI (2021) Gravitational influence on human living systems and the evolution of species on Earth. *Molecules* 26
- Afshinnikoo E, et al. (2020) Fundamental biological features of spaceflight: Advancing the field to enable deep-space exploration. *Cell* 183:1162–84
- Allen DL, et al. (2009) Effects of spaceflight on murine skeletal muscle gene expression. *Journal of Applied Physiology* 106:582–95
- Bauer-Mehren A, Rautschka M, Sanz F, Furlong LI (2010) DisGeNET: a Cytoscape plugin to visualize, integrate, search and analyze gene–disease networks. *Bioinformatics* 26:2924–26
- Becker KG, Barnes KC, Bright TJ, Wang SA (2004) The genetic association database. *Nature Genetics* 36:431–32
- Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57:289–300
- Berrios DC, Galazka J, Grigorev K, Gebre S, Costes SV (2021) NASA GeneLab: interfaces for the exploration of space omics data. *Nucleic Acids Research* 49:D1515–D1522
- Blottner D, et al. (2022) Reciprocal Homer1a and Homer2 isoform expression is a key mechanism for muscle soleus atrophy in spaceflown mice. *International Journal of Molecular Sciences* 23
- Brungs S, et al. (2016) Facilities for simulation of microgravity in the ESA ground-based facility programme. *Microgravity Science and Technology* 28:191–203
- Cahill T, et al. (2021) Mammalian and invertebrate models as complementary tools for gaining mechanistic insight on muscle responses to spaceflight. *International Journal of Molecular Sciences* 22
- Carvalho BS, Irizarry RA (2010) A framework for oligonucleotide microarray preprocessing. *Bioinformatics* 26:2363–7
- Chakraborty N, et al. (2021) Gene-metabolite networks associated with impediment of bone fracture repair in spaceflight. *Computational and Structural Biotechnology Journal* 19:3507–20
- Chawla N, Bowyer K, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. *ArXiv abs/1106.1813*
- Developers T (2024) TensorFlow. Preprint at <https://doi.org/10.5281/zenodo.12726004>
- Gambara G, et al. (2017) Gene expression profiling in slow-type calf soleus muscle of 30 days space-flown mice. *PLoS One* 12:e0169314
- Gautier L, Cope L, Bolstad BM, Irizarry RA (2004) affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20:307–15
- Harris CR, et al. (2020) Array programming with NumPy. *Nature* 585:357–62
- Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* 4:44–57
- Hunter JD (2007) Matplotlib: a 2D graphics environment. *Computing in Science and Engineering* 9:90–95
- Kiss JZ, Wolverton C, Wyatt SE, Hasenstein KH, van Loon JWA (2019) Comparison of microgravity analogs to spaceflight in studies of plant growth and development. *Frontiers in Plant Science* 10
- Liphardt AM, Fernandez-Gonzalo R, Albracht K, Rittweger J, Vico L (2023) Musculoskeletal research in human space flight – unmet

- needs for the success of crewed deep space exploration. *Nature Partner Journals Microgravity* 9:9
22. Man J, Graham T, Squires-Donnelly G, Laslett AL (2022) The effects of microgravity on bone structure and function. *Nature Partner Journals Microgravity* 8:9
23. Mason CE, et al. (2024) A second space age spanning omics, platforms, and medicine across orbits. *Nature*
24. Morey-Holton ER (2003) 9 - The impact of gravity on life. In *Evolution on Planet Earth*, LJ Rothschild, AM Lister (eds), pp 143–59. London: Academic Press.
25. Morey-Holton ER, Globus RK (2002) Hindlimb unloading rodent model: technical aspects. *Journal of Applied Physiology* 92:1367–77
26. Moyer EL, et al. (2016) Evaluation of rodent spaceflight in the NASA animal enclosure module for an extended operational period (up to 35 days). *Nature Partner Journals Microgravity* 2:16002
27. Overbey EG, et al. (2024) The Space Omics and Medical Atlas (SOMA) and international astronaut biobank. *Nature*
28. Pedregosa F, et al. (2011) Scikit-learn: machine learning in Python. *Journal of Machine Learning Research* 12:2825–30
29. Ramachandran R, Bugbee K, Murphy K (2021) From open data to open science. *Earth and Space Science* 8:e2020EA001562
30. Ray S, et al. (2019) GeneLab: omics database for spaceflight experiments. *Bioinformatics* 35:1753–1759
31. Reynolds RJ, et al. (2022) Validating causal diagrams of human health risks for spaceflight: An example using bone data from rodents. *Biomedicine* 10
32. Ritchie ME, et al. (2015) Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* 43:e47
33. Ronca AE, Lowe MG (2022) Rodents as a model for research in space. In *Handbook of Space Pharmaceuticals*, YV Pathak, M Araújo dos Santos, L Zea (eds), pp 679–700. Cham: Springer International Publishing.
34. Rutter LA, et al. (2024) Astronaut omics and the impact of space on the human body at scale. *Nature Communications* 15:4952
35. Sherman BT, et al. (2022) DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Research* 50:W216–21
36. Van Rossum G, Drake FL (2009) *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace
37. Waskom M (2021) seaborn: statistical data visualization. *Journal of Open Source Software* 6:3021