Heliyon 10 (2024) e40772

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Proteomics and machine learning: Leveraging domain knowledge for feature selection in a skeletal muscle tissue meta-analysis

Alireza Shahin-Shamsabadi *, John Cappuccitti

Evolved.Bio, 280 Joseph Street, Kitchener, Ontario, Canada

ARTICLE INFO

Keywords:
Machine learning
Proteomics
Domain knowledge
Feature selection
Skeletal muscle tissue

ABSTRACT

Omics techniques, such as proteomics, contain crucial data for understanding biological processes, but they remain underutilized due to their high dimensionality. Typically, proteomics research focuses narrowly on using a limited number of datasets, hindering cross-study comparisons, a problem that can potentially be addressed by machine learning. Despite this potential, machine learning has seen limited adoption in the field of proteomics. Here, skeletal muscle proteomics datasets from five separate studies were combined. These studies included conditions such as *in vitro* models (both 2D and 3D), *in vivo* skeletal muscle tissue, and adjacent tissues such as tendons. The collected data was preprocessed using MaxQuant, and then enriched using a Python script fetching structural and compositional details from UniProt and Ensembl databases. This was used to handle high-dimensional and sparsely labeled dataset by breaking it down into five smaller categories using cellular composition information and then training a Random Forest model for each category separately. Using biological context for interpreting the data resulted in improved model performance and made tailored analysis possible by reducing the dimensionality and increasing signal-to-noise ratio as well as only preserving biologically relevant features in each category. This integration of domain knowledge into data analysis and model training facilitated the discovery of new patterns while ensuring the retention of critical details, often overlooked when blind feature selection methods are used to exclude proteins with minimal expressions or variances. This approach was shown to be suitable for performing diverse analyses on individual as well as combined datasets within a broader biological context, ultimately leading to the identification of biologically relevant patterns. Besides from generating new biological insights, this approach can be used to perform tasks such as biomarker discovery, cluster analysis, classification, and anomaly detection more accurately, but incorporation of more datasets is needed to further expand the computational capabilities of such models in clinical settings.

1. Introduction

Proteomics, the large-scale study of proteins and their interactions, plays a pivotal role in understanding biological processes in both *in vivo* and *in vitro* environments. In contrast to genomics, which focuses on the static nature of genes, proteomics sheds light into the dynamic functionalities of proteins. This covers a spectrum of cellular functions, such as signal transduction, structural support, and metabolic regulation [1]. Innovations in mass spectrometry have increased the prevalence of proteomics studies, producing

* Corresponding author. Evolved.Bio, 280 Joseph Street, Kitchener, Ontario, N2G4Z5, Canada.
E-mail address: alireza@itsevolved.com (A. Shahin-Shamsabadi).

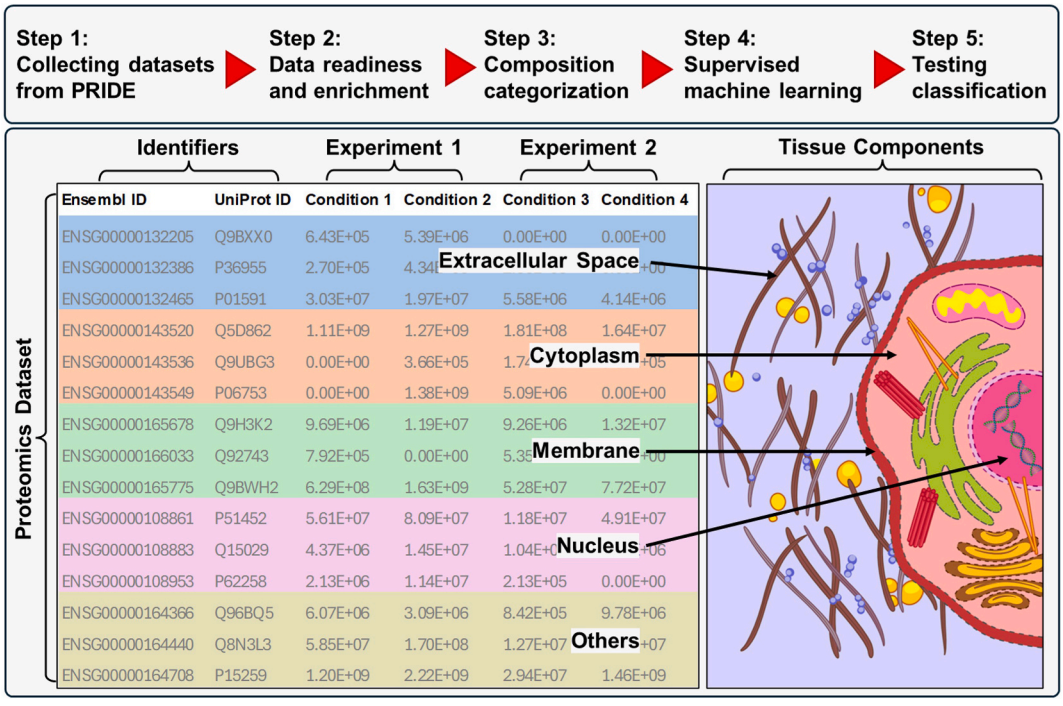
https://doi.org/10.1016/j.heliyon.2024.e40772
Received 5 May 2024; Received in revised form 22 October 2024; Accepted 27 November 2024
Available online 29 November 2024

extensive data with intrinsic heterogeneity. Such data offers unparalleled insights into cellular behavior, potentially enabling new therapeutic strategies, especially in the case of intricate tissues and cell cultures [2]. Many of these datasets are archived in public repositories, such as the Proteomics Identification Database (PRIDE). These datasets not only stand as valuable research assets for re-analysis but also hold potential to create a more comprehensive perspective on protein expression and functionality across various conditions [3,4]. Merging datasets from diverse publications could represent a significant stride towards this goal and by reducing individual study biases, a more integrated and holistic biological viewpoint can be established [5].

Proteomic analysis of skeletal muscle tissue stands as one of the most complex yet pivotal undertakings in molecular biology. Skeletal muscle, the largest tissue in the human body, exhibits a high degree of complexity. It comprises a diverse array of cell types, including distinct populations of multi-nucleated muscle fibers, each characterized by a unique metabolic and contractile profile [6]. Such inherent heterogeneity, along with its dynamic reactivity to external stimuli, heightens the challenge of its proteomic characterization. Nevertheless, proteomics is proven to be an invaluable tool to unravel the subtle details of muscle physiology including intricate molecular processes that guide muscle development and regeneration [7]. Moreover, proteomics analyses of altered proteins have helped in discovering potential biomarkers associated with muscular disorders, such as different muscular dystrophies [8–10].

Machine learning holds great potential to revolutionize proteomics research by enabling the identification of intricate patterns, robust classification of data, development of predictive models, and the discovery of novel biomarkers [11]. A variety of machine learning algorithms, including Random Forests, have found applications in such use cases within high-dimensional and high-throughput omics datasets [12,13]. Yet, the incorporation of machine learning into proteomics is still in its early stages, especially in high-dimensional scenarios where conventional feature selection methods often falter. Notably, as of 2022, a mere 2 % of proteomics publications have integrated machine learning approaches [14]. However, it's worth noting that this field is rapidly evolving, and the integration of machine learning in proteomics research has been growing as well. This significant underutilization has fueled a growing interest in exploring alternative strategies to effectively navigate these complexities. Several other fields grapple with similar issues, including time-consuming data collection processes, poor data quality, and limited sample sizes. These limitations can lead to the underperformance of machine learning models, often due to the mismatch between the high dimensionality of the feature space and the low number of samples available [15].

In this investigation, five projects from the PRIDE database, representing 12 distinct conditions of skeletal muscle tissue and its adjacent tissues, such as tendon and musculotendinous junction, were selected and combined. These datasets had clear sample annotations but lacked any enrichment strategies. To overcome the inherent high dimensionality of proteomics datasets, a novel methodology leveraging machine learning was introduced. This approach involved segmenting the proteomic data using cellular composition information obtained from public databases UniProt and Ensembl, based on the known biological principles that function often correlates with cellular location and composition (**Schematic 1**). Subsequently, diverse analytical techniques, including principal component analysis (PCA), linear discriminant analysis (LDA), correlation network analysis, and feature importance analysis, were applied to the segmented data. This ultimately enabled the creation of cellular composition-specific machine learning models. A



Schematic 1. Current study's workflow for proteomics dataset enrichment and categorization using cellular composition information instead of feature selection for supervised machine learning.

refined analytical framework was developed by replacing traditional machine learning feature selection techniques, which often neglect the biological significance of individual proteins, with this domain knowledge-driven approach that revealed the delicate roles of individual proteins correlated to their physical location in the tissues. The presence of conditions exhibiting significant physiological similarities led to challenges in classification, and, while feature selection could not address these, incorporating cellular composition information significantly enhanced the resolution of this framework.

2. Methods

2.1. Proteomics data collection and pre-processing

Raw proteomics files detailing skeletal muscle tissue under various conditions were sourced from different experiments deposited on the PRIDE database. Each experiment encompassed multiple conditions including different *in vivo* derived samples, as well as two-dimensional (2D) and three-dimensional (3D) *in vitro* models. Multiple replicates for each condition were present. All the.raw files underwent re-analysis in MaxQuant with a false discovery rate (FDR) set at 0.01 for peptides and proteins, peptide-spectrum match (PSM) at 0.01, and a stipulation of 2 unique peptide essential for protein identification to facilitate the extraction of intensity values for each protein. For data processing, detected intensities and protein identifiers (namely Gene_Name, Major_Protein_IDs, and Protein_Names) for each condition and its replicates were stored.

To establish comparability between distinct projects, protein IDs were standardized. A Python script was used to retrieve new IDs for each protein from the Ensembl (https://useast.ensembl.org/index.html) and UniProt (https://www.UniProt.org/) databases. For Ensemble IDs, the script made requests to the Ensembl REST API (http://rest.ensembl.org) cross-referencing for *Homo sapiens* while for UniProt IDs, the script used the Mygene library (an interface to the https://mygene.info/web service) to fetch the relevant information. The program subsequently replaced the original three identifier columns with new ones labeled Ensembl_ID and UniProt_ID. The script's logic was iterative, transitioning from Gene_Name to Major_Protein_IDs and finally to Protein_Names to acquire one unique Ensembl ID for each protein. After the initial hit for each protein, the script would proceed to the next protein on the list. If a protein had multiple identifiers, the algorithm would navigate through them in sequence, proceeding to the next protein once the first Ensembl_ID was secured. These Ensembl IDs then served as a gateway to fetch UniProt IDs. In case no Ensembl or UniProt IDs were found for a protein, it was removed from the entire project along with its expression values for all conditions. This process was executed individually for each dataset. After standardizing the IDs across the experiments, the datasets were combined using UniProt IDs into one big dataset, streamlining subsequent analyses and the training of the machine learning model. In instances where an ID was absent in a particular condition, its expression values were allocated a zero indicating no expression of that protein in that condition. The script incorporated checks to ensure that no ID recurred within the combined dataset.

2.2. Domain knowledge implementation for categorization

To address the high dimensionality of the combined dataset, characterized by the presence of thousands of proteins across a limited number of conditions and replicates, a Python script was employed for data enrichment. This script retrieved cellular composition information for proteins from UniProt. To do this, the function made an HTTP request to the UniProt database, specifically requesting the text-based record for the given UniProt ID. Once the record was fetched, the function parsed the text to identify lines that indicate subcellular location for each protein. An additional column, termed "Category", was introduced to the combined dataset and was populated with the primary cellular composition information for each protein. Leveraging this enriching information, the combined dataset was divided into five distinct categories. While each of these categorical datasets retained all conditions and replicates from different experiments, the proteins were exclusive to their respective categories. The categories were "Extracellular Space", "Membrane", "Cytoplasm", and "Nucleus". A fifth category, labeled "Others", included proteins not classified into any of the previously specified categories. The combined dataset and the newly generated categorical datasets underwent further analyses in the subsequent section.

2.3. Analytical evaluation of proteomics data

PCA and LDA were used to represent and visualize high-dimensional proteomics datasets within a 2D space. The aim was to retain local structures and relationships between different conditions by identifying potential clustering patterns. The necessary libraries, including Pandas, NumPy, Matplotlib, and scikit-learn (sklearn), were employed. The Python script began by standardizing the protein intensity values using sklearn's StandardScaler. This standardization ensured uniform scales across intensities, eliminating biases introduced by varying magnitudes. Standardization transformed the data to have a mean of zero and a standard deviation of one by subtracting the mean and dividing by the standard deviation for each protein's intensity value. This scaling allowed for proper comparison of measurements across different experimental conditions.

After standardization, PCA was applied by initializing an object from sklearn.decomposition.PCA with n_components = 2. The script calculated the covariance matrix of the standardized data and performed eigenvalue decomposition to extract eigenvalues and corresponding eigenvectors, revealing the principal directions of the data. The eigenvalues were sorted in descending order, and the top two were selected to define the 2D space for projection. For LDA, sklearn.discriminant_analysis.LinearDiscriminantAnalysis was used. LDA identified directions that maximally separated different conditions, considering information from all classes simultaneously. LDA aimed to find directions that maximized the ratio of between-class variance to within-class variance, resulting in optimal

separation of conditions. The first two dimensions provided by the LDA output were used for visualization. These dimensions resulted from solving the generalized eigenvalue problem inherent to LDA, where the largest eigenvalues correspond to the axes with the greatest separability. Scatter plots were created for both PCA and LDA results with each condition's markers color-coded and shaped differently.

Next, a Python script was written to analyze the distribution of individual proteins' importances in defining each experimental condition. First, expression value intensities were standardized by scaling the expression values to a mean of zero and a standard deviation of one. This normalization was crucial to ensure that features with larger scales did not dominate the importance calculations. Then, for each protein, expression values were averaged across replicates of the condition to minimize experimental bias. A Random Forest Regressor was trained for each condition, using the protein expression intensities of all other conditions as features and the target condition as the response. The Random Forest algorithm quantified the importance of each protein by measuring how much each protein (as a feature) decreased impurity in a decision tree within the forest. This importance indicated how crucial a protein was in predicting the overall expression profile of each specific condition compared to others. These importances were visualized using a "Box Plot", illustrating the distribution of importance values for proteins across conditions on the y-axis. This analysis was repeated for each category and the combined dataset. The script used the scikit-learn library to standardize expression intensities and perform the Random Forest analysis. The Matplotlib library was used to generate the box plots.

The overall variability in protein expressions within each condition was analyzed using a Python script with Pandas for data manipulation and Matplotlib for visualization. For both the combined dataset and individual categories, the script calculated the average protein expression and standard deviation for each unique condition. The data were then standardized by subtracting the mean value and dividing by the overall standard deviation. The results were visualized using "Violin Plots", illustrating the distribution of expression values. These plots juxtaposed expression trends within specific categories against those of the combined dataset, allowing for comparison of protein expressions between different conditions and the effect of cellular composition categorization.

The differences in protein expression patterns between different conditions within each category, relative to the combined dataset, were investigated using "Volcano Plots". The Python script compared the mean expression levels of all proteins within a category under each specific condition against their mean expression across all other conditions. This approach provided a meaningful assessment of differences between conditions by focusing on general category-specific trends, contrasting with comparing individual protein expressions across the dataset, which could only show isolated behaviors of individual proteins. Statistical significance of these differences was determined using Welch's *t*-test for calculating p-values, and the magnitude of expression shifts was quantified through log2 fold changes. The volcano plots visually represented these category-wide comparisons, with log2 fold changes on the x-axis and -log10(p-values) on the y-axis. This approach highlighted significant expression differences that were condition-specific within each category, providing a comprehensive view of the combined dataset.

The relationships between different conditions were further analyzed using "Correlation Network Analysis". A custom Python script was developed to visualize the intricate relationships among different conditions. Utilized libraries included Pandas for data manipulation, Networkx for graph creation, and Matplotlib for visualization. First, mean values of replicates within each condition were calculated and standardized similar to previous methods. Then, the pairwise correlations between the conditions were calculated using the Pearson correlation coefficient to measure the linear dependence between each pair of conditions. Finally, graphs where nodes represented conditions and edges represented correlations surpassing a defined threshold (0.6 in this case) were plotted. The resulting networks, colored to differentiate between conditions, were displayed through six different plots, each providing a unique perspective on the inter-relationships between conditions within both the combined dataset and individual categories. The network graphs were drawn using a spring layout for better visualization.

A "Correlation Analysis" of protein expression intensity values across conditions within both the combined dataset and individual categories was performed. The script utilized Pandas for data manipulation and Seaborn and Matplotlib for graphical representations. The analysis began by calculating the mean values of replicates for each condition, followed by normalization using StandardScaler to ensure comparability across conditions. Subsequently, a correlation matrix was computed for the combined dataset and individual categories to capture the degree of linear relationship between the intensity values across different conditions. This involved calculating pairwise correlation coefficients for all conditions to clarify how mean protein profiles varied between conditions. The resultant correlation matrices were visually presented using heatmaps, displaying the correlation coefficients within the theoretical range of −1 to 1. These heatmaps provided a summary of how protein expressions co-varied across different conditions within the combined dataset and individual categories.

2.4. Machine learning model development

To develop a model capable of classifying conditions, a custom Python script was implemented using a Random Forest classifier for machine learning purposes. This script was applied to the combined proteomics dataset to analyze and visualize the classification accuracy for various experimental conditions. First, preprocessing was performed by dropping non-numerical columns, including Ensembl_ID, UniProt_ID, and Category. Then, the data was transposed and standardized using StandardScaler to ensure all features had equal influence on the model. Labels (conditions) were extracted subsequently. A K-Fold Cross Validation strategy with 3 splits (equal to the lowest number of replicates among different conditions) was implemented to evaluate the performance of the Random Forest classifier on the dataset, enhancing the validity of the results. The data was partitioned into training and testing sets, with 30 % of the data reserved for testing to evaluate the model's performance on unseen data. To find the optimal hyperparameters, a Grid Search was performed across a predefined set of hyperparameters, including "n_estimators" ([50, 100, 200]), "max_depth" ([None, 10, 20, 30]), "min_samples_split" ([2,5,10]), and "min_samples_leaf" ([1,2,4]). The search used a 3-Fold Cross Validation to ensure proper

assessment across different data partitions. The best hyperparameters and the highest accuracy were reported. Once the classifier was trained and validated, the predicted tissue conditions were obtained. The accuracy score and classification report, containing precision, recall, and f1-score for each label, were used for evaluation. Additionally, a confusion matrix was generated to provide insights into the classifier's true positive, false positive, false negative, and true negative predictions, displayed as a heatmap using the Seaborn library.

In the next step, a RandomForest-based feature selection was performed to reduce the dimensionality of the proteomics dataset, and the same Random Forest classifier model was deployed on the selected features from the larger combined dataset using the same metrics and outputs to evaluate the model. Feature selection was performed using another Random Forest classifier with 100 trees as the estimator to rank the importance of features in the combined proteomics dataset. The Random Forest model's built-in mechanism for calculating the importance of each feature, based on how much a feature reduced impurity in the decision trees, was utilized. Using the SelectFromModel utility from Scikit-learn, a Random Forest model was fit to the data, identifying the most significant features based on the importance assigned by the model. Features with importance scores above the defined threshold, calculated as the mean of all feature importances, were selected to retain only those features that contributed significantly to the model's decision-making process. The resultant mask of selected features was applied to extract the relevant columns.

To further investigate the impact of feature selection on model performance, an ablation study was conducted. Instead of using the mean of all feature importances as the criterion for feature selection, different proportions of features were retained based on their importance scores, specifically the top 25 %, 50 %, and 75 % of features. These subsets of features were used to train and evaluate the Random Forest classifier to analyze how varying the number of selected features affected the model's accuracy, precision, recall, and f1-score. Additionally, memory usage and processing time were measured using the memory_profiler library and the time module to track the resources consumed during the training and evaluation of the Random Forest classifier for each subset.

In addition to the RandomForest-based feature selection, two other dimensionality reduction techniques based on PCA and LDA were employed. For PCA, the proteomics dataset was first standardized using StandardScaler to ensure all features were on the same scale. PCA was then applied to the scaled data, and the number of components was determined adaptively by analyzing the explained variance ratio. A cutoff point was established where the difference in explained variance between consecutive components fell below a threshold of 0.001, or at most 50 % of the total components. Feature importance was calculated based on the absolute sum of loadings for the selected components, normalized to sum to 1. Features with importance scores above the mean were retained. For LDA, the data was similarly standardized, and the number of components was set to the minimum of either one less than the number of unique classes or the number of features. LDA was applied using the Singular Value Decomposition (SVD) solver, and feature importance was derived from the absolute sum of LDA coefficients across all discriminant functions, again normalized to sum to 1. The optimal number of features was determined by selecting those that cumulatively explained 95 % of the variance. Reduced set of features from both PCA and LDA methods were then used with the same Random Forest classifiers with similar hyperparameter tuning, K-fold, and train-test splitting.

As an alternative to feature selection that disregarded proteins' functions and biological roles, Random Forest classifiers with similar hyperparameter tuning, K-fold, and train-test splitting were trained individually on each of the five cellular composition categories, generated in previously using the information fetched from Uniprot database. Accuracy scores, classification reports, and heatmap confusion matrices facilitated the comparison between the classifier models for these new categories and the models previously developed using the combined dataset or its RandomForest-, PCA-, and LDA-based feature-selected versions.

In the next step, a pipeline was developed to test the previously trained random forest model using new proteomic datasets from new experiments not included in the model's training. These datasets were from two separate experiments, one with conditions similar to some used in training (3D *in vitro* models of skeletal muscle tissue, with and without differentiation) and another with conditions unlike those in training (left ventricle and atria from *in vivo* heart samples). The pre-trained random forest model, loaded from the

Table 1
List of PRIDE datasets used for training the machine learning model. Data was extracted from 5 experiments with 12 conditions and a total of 52 replicates.

PRIDE accession number	Experimental detail	Condition, Number of Replicates
PXD006182 (Murgia et al. [16])	Single fast/slow muscle fibers from young/old male subjects[a]	Old Slow Fiber, 4 Old Fast Fiber, 4 Young Slow Fiber, 4 Young Fast Fiber, 4
PXD020192 (Di Meo et al. [17])	Muscle tissue from healthy adults	Skeletal Muscle, 2
PXD033025 (Mengeste et al. [18])	2D *in vitro* muscle models with and without electrical stimulation	2D Stimulated, 7 2D Unstimulated, 7
PXD011919 (Mills et al. [19])	3D *in vitro* muscle models with and without differentiation using primary myoblasts	3D Differentiated, 4 3D Undifferentiated, 4
PXD029307 (Karlsen et al. [20])	Tendon, myotendinous junction, and combination of muscle and tendon from healthy young males	Tendon, 4 Musculotendinous Junction, 4 Skeletal Muscle and Tendon, 4

[a] This study presented multiple repetitions for each fiber type: slow, slow 2A, fast 2A, fast 2X, and fast 2A/2X from each individual subject. For the sake of class balance, only the slow and fast 2A fibers were included here. In instances where multiple measurements for a particular fiber type from the same subject were documented, the median value was used to represent that subject.

previous code, was employed to predict the likelihood of each new test condition being similar to the training conditions. The similarity between test and training conditions was assessed using the predictive probabilities generated by the random forest model, without direct comparison between the training and new test datasets. These probabilities indicated the likelihood of each new test condition aligning with the characteristics of each training condition, as learned during training. The model outputted probabilities for

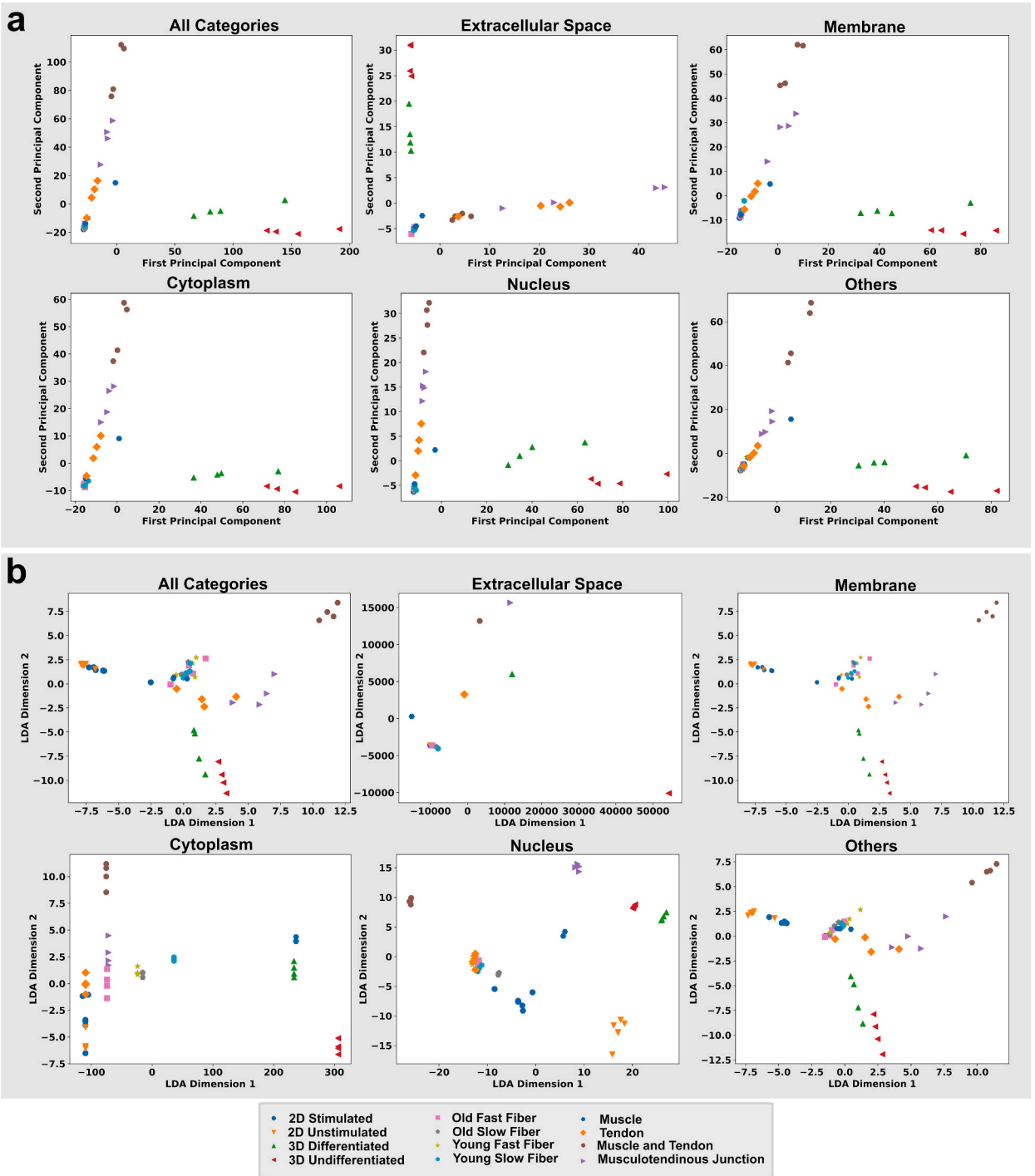


Fig. 1. Visualization of dimensionality reduction and cluster formation across distinct categories within the proteomics dataset via **a)** PCA and **b)** LDA graphs. Consistently, conditions with more similar biological properties were clustered closer in most categories when analyzed using PCA and LDA. However, the distinctions between conditions became more emphasized when individual categories based on cellular composition were subjected to LDA transformations, and to some degree in PCA.

each potential category, serving as a quantitative measure of similarity, where higher probability indicated greater similarity to a specific training condition's profile. For each test condition, the top two most similar training conditions were identified based on these probability predictions, with probabilities converted into percentages to facilitate interpretation.

3. Results and discussion

Proteomics datasets are valuable collections of information relating to tested conditions. However, the current scope of analysis often harnesses only a small portion of what these datasets have to offer, typically through manual access to a few online repositories and datasets. In the current study, publicly available proteomic datasets from a few different experiments encompassing skeletal muscle tissue [16,17], its 2D [18] and 3D [19] *in vitro* models, and anatomically adjacent tissues, such as tendons and musculotendinous junctions [20], were sourced from the PRIDE database (Table 1). These were subsequently reanalyzed using MaxQuant software with uniform parameters. In order to detect a proper array of proteins and thereby test the impending platform's capability with ultra-high-dimensional datasets, the protein FDR was set at 0.01 and relied on two unique peptides for protein identification. This was in alignment with conservative approaches used for high-confidence protein detection [21,22]. Following the analysis of.raw files, the detected protein intensities were directly employed without additional processing.

Subsequently, a Python script was tailored to perform data preprocessing. Initially, this procedure used gene names, major protein IDs, and protein names as identified by MaxQuant to extract Ensembl (https://useast.ensembl.org/index.html) and UniProt (https://www.UniProt.org/) IDs for each protein, streamlining protein annotations. The subsequent step merged proteomics datasets from different experiments, each containing one or more separate conditions and their replicates, into a unified dataset using UniProt IDs as references. Then, the script retrieved cellular composition information for every protein from the UniProt database, appending it to the final column of the combined dataset as "Category" (Table S1). These compositions included "Extracellular Space", "Membrane", "Cytoplasm", and "Nucleus". Proteins that deviated from these categories, or for which no category was detected, were cataloged under "Others". A total of 7145 proteins were included in the combined dataset with 410, 1810, 1913, 1250, and 1551 proteins in Extracellular Space, Membrane, Cytoplasm, Nucleus, and Others categories, respectively. UniProt database was selected for this purpose for its detailed and specific subcellular localization annotations compared to other online databases, as well as seamless integration of its interface into Python pipelines.

To gain initial insights into the complex, high-dimensional proteomics datasets, dimensionality reduction techniques PCA and LDA were employed for both combined dataset as well as each individual cellular composition category. These methods projected the data onto 2D spaces while preserving the intrinsic relationships between different experimental conditions (Fig. 1). First, the expression intensity values of the proteins were standardized to allowed subsequent dimensionality reduction techniques to operate on a comparable basis across all conditions from different studies. However, it is important to note that while this method normalized the scale of data, it couldn't eliminate biases from systematic errors or discrepancies inherently present in the experimental procedures.

Both PCA and LDA are linear transformation methodologies used for dimensionality reduction, each with its own strengths. PCA, an unsupervised method, is suitable for revealing global relationships between conditions and was used to visualize the high-dimensional structure of the data within each cellular composition category. This helped assess whether intra-category variability could be effectively captured in reduced dimensionality compared to the combined dataset. On the other hand, LDA, a supervised method, excels in optimal class separability, making it better suited for classification purposes. Using PCA and LDA together provided a comprehensive analysis of the dataset's structure and the impact of categorization on model performance. This approach of visualizing high-dimensional data in a condensed space has proven effective in revealing patterns, clusters, or distinctions among conditions or groups [23,24]. By using both methods in tandem, the balance between accurately representing the original data and rendering it into an understandable 2D form was preserved, laying the groundwork for uncovering subtle yet significant patterns in proteomics data that could shed light on the molecular underpinnings of the conditions.

The distinctions between PCA and LDA across different categories and the combined dataset were noticeable (Fig. 1). The PCA graphs (Fig. 1a) displayed similar clustering patterns in the combined dataset and all categories, except for the Extracellular Space, which showed slight differences. Undifferentiated and differentiated 3D conditions formed their own clusters in all six plots, while 2D cultures (with and without electrical stimulation), muscle, and individual muscle fibers formed another cluster. The musculotendinous junction, combined muscle and tendon, and tendon formed a loose cluster. In the Extracellular Space, the muscle condition from this cluster was positioned closest to the cluster of individual muscle fibers, whereas, in all other plots, muscle was the furthest away from them. In contrast, the LDA graph showed distinct clustering patterns in each plot (Fig. 1b). In the combined dataset, as well as in the Membrane and Others categories, each of the 2D and 3D cultures formed their own clusters. The individual fiber types, muscle, and tendon formed a dense cluster, with the musculotendinous junction positioned close to it and the muscle and tendon group away from it. In the Extracellular Space category, 2D cultures and individual muscle fibers formed a dense cluster. The 3D differentiated condition, muscle, tendon, muscle and tendon, and musculotendinous junction formed a loose cluster close to it, but the 3D undifferentiated condition was positioned far from the rest. The Cytoplasm and Nucleus categories also showed distinct cluster formations. Overall, LDA achieved a clearer distinction between different categories.

Proteomics datasets typically feature a configuration where the number of variables (e.g., total protein number, 7145 in this case) significantly exceeds the number of samples (e.g., specific conditions, 12 conditions from 5 different experiments, with a total of 52 technical and biological replicates). Handling datasets with low sample-to-variable ratios often requires data reduction for meaningful interpretation, and statistical methodologies like PCA and LDA have been employed to treat proteins as collective biomarkers or classification agents while ignoring their individual roles [25,26]. The novel approach of protein classification used here, which integrates cellular composition and physical location of proteins, offers several advantages. It enhances the sensitivity and specificity of

analyses and provides valuable insights into the potential origin and causative factors of a given condition, as physical proximity often defines protein interactions and synergistic effects between proteins. This is especially true when large numbers of datasets representing different conditions from different experiments are combined and compared. As more conditions from different studies are combined, more sophisticated paradigms, such as domain adaptation PCA [27] that introduce elements of supervision, can be used to further clarify the distinctions between different conditions.

Since proteomics datasets from various experiments were combined to facilitate inter-condition comparisons, the number of proteins with non-zero expression intensities in some experiments/conditions was significantly lower than the total number of proteins present in the combined dataset. Evaluating protein importance in defining each condition was essential. First, the expression values were scaled to a mean of zero and a standard deviation of one to ensure that features with larger scales did not dominate the importance calculations. Average protein values for each condition, accounting for its replicates, were then computed. Subsequently, a Random Forest Regressor was employed to calculate the importance of each protein for defining its corresponding condition. In training the Random Forest Regressor, protein expression values from all conditions except the target were used as features, while the mean expression values for the target condition served as the response. The model calculated the importance of each protein by measuring how much it reduced impurity in decision trees within the forest, averaged across all trees. This importance reflected each

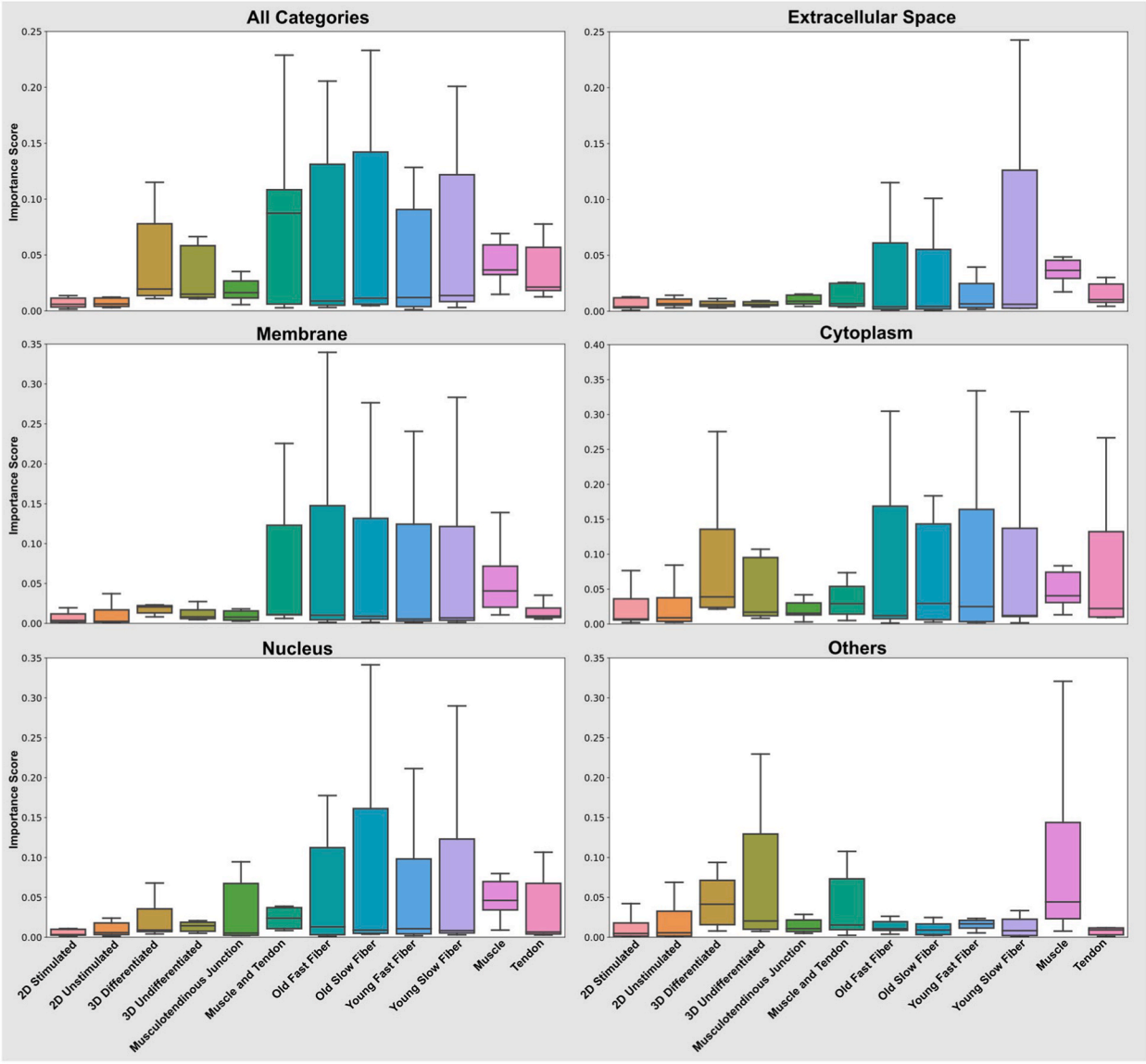


Fig. 2. Distribution of protein importance scores across different conditions, illustrating the variability and relative significance of individual proteins in defining each condition, in each cellular composition category. Box plots showed the spread and variability of importance scores for all proteins within each condition, highlighting the diverse proteomic profiles that characterize each condition.

protein's ability to improve model accuracy, independent of its expression level. To prevent bias from dominant proteins, only a random subset of proteins was considered at each split in the decision trees. This process provided a measure of each protein's contribution, highlighting the complex pattern of protein expressions that characterizes each condition. The outcomes were visually represented using box plots (Fig. 2), showing the distribution of protein importances per condition. These plots provided a deeper understanding of the distribution and variability of these importances across conditions. Expansive boxes and extended whiskers indicated a greater number of proteins with non-zero importance scores, implying a complex proteomics profile. This approach is particularly relevant here, where proteomics datasets from multiple separate experiments are combined and compared. Unlike other studies that typically have a control condition against which other experimental conditions are compared, here, each condition is compared to all others. If new datasets are included, the importance of each protein in defining its condition will change, as the newly added conditions might be similar to or different from the existing ones.

The total number of proteins in each category was significantly smaller than the combined dataset, and variations in importances showed distinct patterns when either different conditions within a specific category were compared with each other or when the same condition was considered in different categories. For example, differentiated 3D muscle models had larger boxes in the combined dataset, but their Extracellular Space, Membrane, and Nucleus, showed much smaller boxes and whiskers, indicating higher variablity in this condition originating from its cytoplasm. When this 3D model was compared with human muscle conditions in different categories, the same patterns were observed. Interestingly, none of the 2D models showed such similarity to the human muscle condition. Although individual proteins causing these similarities or differences need to be identified to more accurately compare different conditions with each other, this could be a testiment that 3D models are better representatives of physiological conditions, as it has

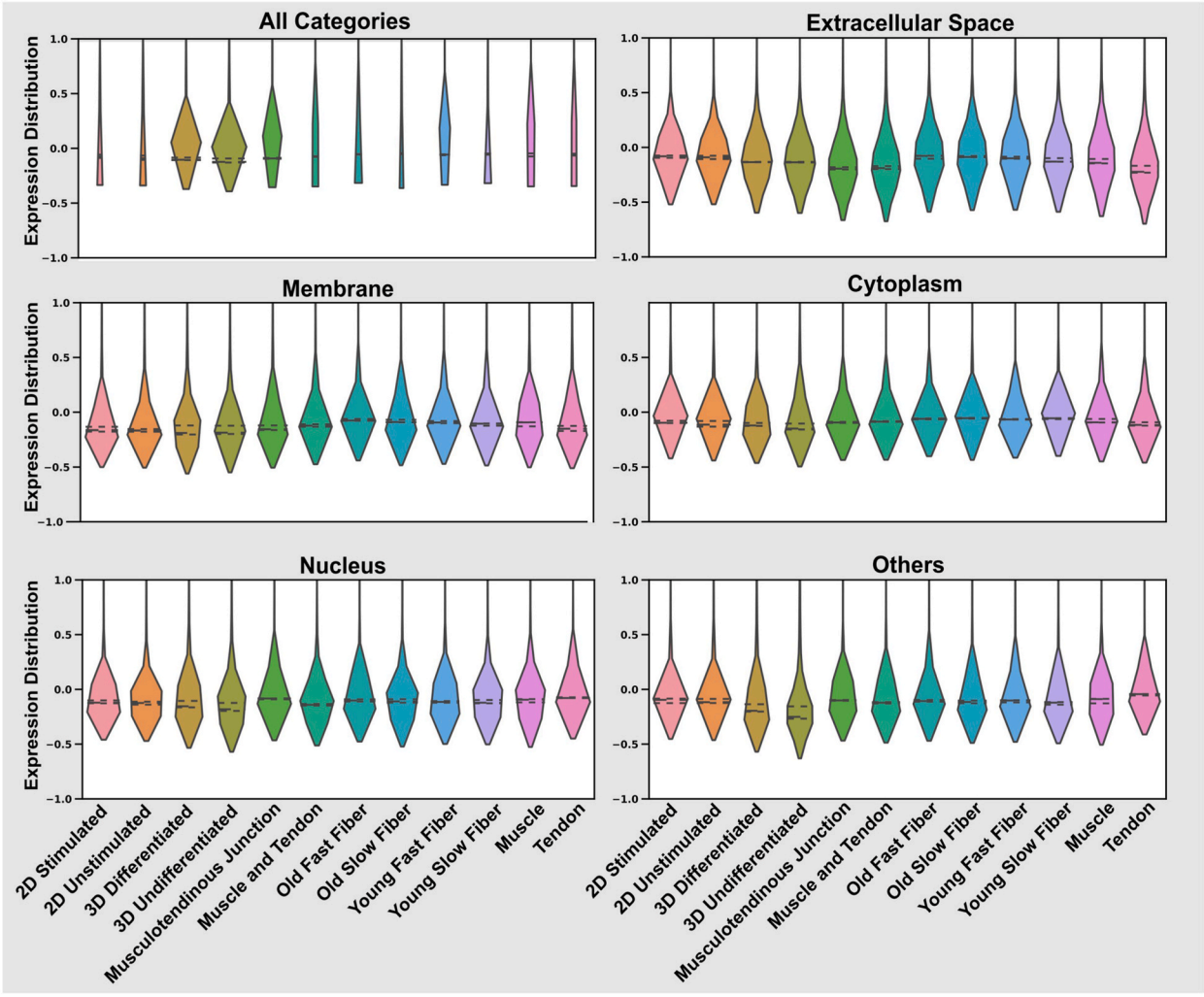


Fig. 3. Visualization of distribution of protein expression intensity values in combined dataset and individual categories. The visualization scope is narrowed for the violin plots for clarity, and the deviation lines are abbreviated, underscoring the violin's core shape. The unique contours of the violins clearly highlight disparities between conditions within a category and draw contrasts for identical conditions across diverse categories. These plots illustrated distribution of expression values of proteins using the standardized values.

been shown before [28,29].

The effective utilization of proteomics datasets, and broadly speaking, omics methodologies, demands data-driven strategies to identify significant features thinly dispersed within such high dimensional data. Several statistical algorithms have been used to highlight the significance of features in these datasets, primarily to facilitate data dimensionality reduction for enhanced classification, diagnosis, and prognosis [30,31]. The methodology presented here, assessing protein importances across different experiments/conditions and different cellular composition categories, not only harnesses these capabilities but, when merged with the physiological understanding of protein functions, offers a more comprehensive perspective on proteomics data. Crucially, this approach avoids the pitfall of discarding proteins with lower expression values that may still be instrumental in characterizing the state of the target conditions. Furthermore, when combining multiple datasets, this method becomes a valuable tool in better understanding various tissues and conditions, spanning diverse physiological and patholigical conditions and their models. At the same time, when comparing different experiments with each other, the potential effect of different experimental conditions should also be taken into account as direct comparisons might not always be possible.

While the box plots showed significant differences in the relative importance of proteins between conditions within each category compared to the combined dataset, understanding the disparities in protein expression levels across conditions within each category provided an additional distinct perspective (Fig. 3). This was achieved using violin plots for a comprehensive visualization of protein expression intensity distributions. The Python script first computed the mean and standard deviation for each condition, standardizing the data. Positive values indicated expression intensity levels higher than the overall mean for a condition, while negative values indicated levels lower than the mean. The violin plots detailed the distribution of protein expressions, with the form of the violin mirroring the kernel density estimation of the dataset. Width variations in the violins reflected data density, with broader segments indicating higher data densities and narrower portions indicating lower densities. The violin's core represented the dataset's median, and key quartiles, including the 25th, median, and 75th percentiles, were included within each violin. Interestingly, when the data was analyzed by cellular composition category, differences between conditions became less noticeable compared to the combined dataset. Violins that were narrow in the combined dataset became thicker within each category (Fig. 3), indicating more similar expression patterns within each category. In contrast, considering the importance distribution of proteins (Fig. 2), differences between conditions became more pronounced in each category, suggesting that each protein might play different roles in each condition, even with similar expression patterns.

Differences between conditions across categories were analyzed using volcano plots, in order to visualize the significance and magnitude of these differences (Fig. 4). The mean protein expressions of individual conditions were compared against the pooled

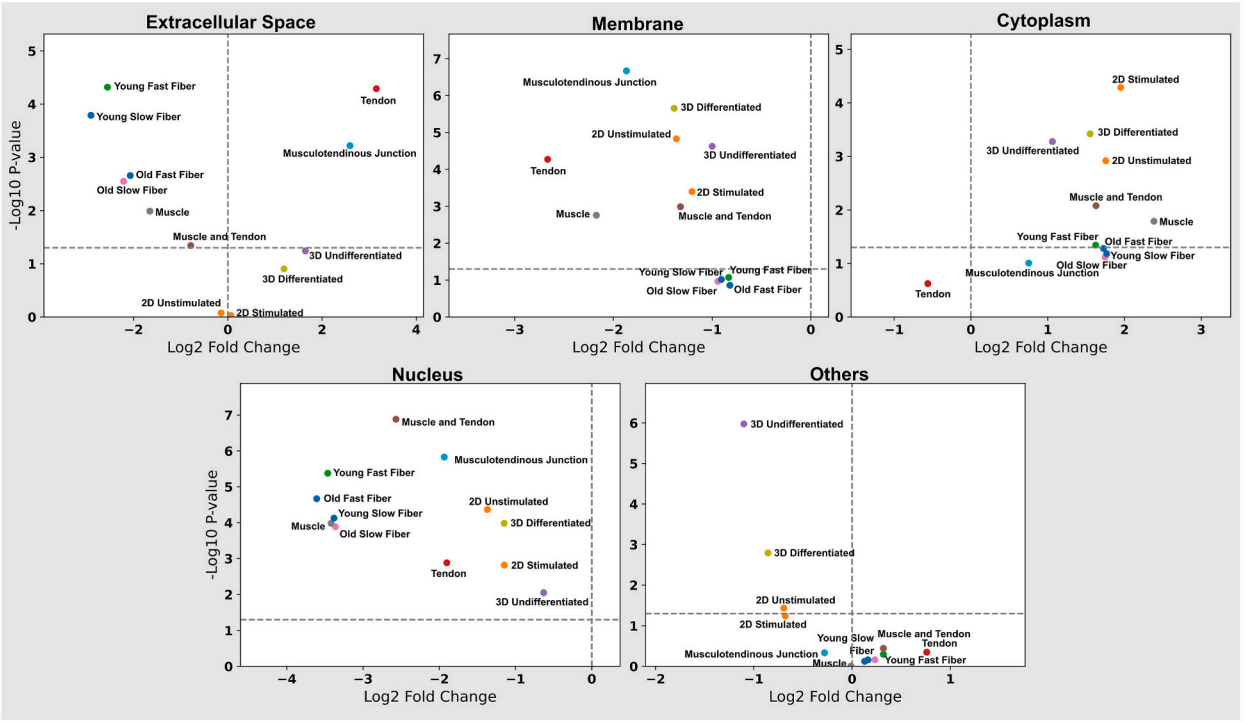


Fig. 4. Volcano plots juxtaposing individual categories against the aggregate of all other ones. Each visualization contrasts the mean protein expression value of a given category against the aggregated mean of all other categories for respective conditions. Such representations explains distinctions between conditions within varied categories. Notably, the disparities between slow and fast fibers in both young and elderly individuals became more pronounced in their Extracellular Space and Nucleus categories. Conversely, in other categories, these differences become narrower, with these four conditions appearing nearly indistinguishable. Each point in the volcano plot represents the outcome of a separate comparison.

average of all other conditions within the category. For each category, the Python script segmented the proteomics data into two groups, one with all proteins belonging to the category and the other with all other proteins. An independent Welch's *t*-test was performed between these groups for each condition, calculating a p-value. Subsequently, the log2 fold change was calculated to reveal the relative shift in overall protein expression between conditions within the selected category and the rest. In the resulting volcano plots, the x-axis represented the log2 fold change, where values above zero indicated higher similarity in protein expression within the category, and values below zero indicated divergence. The y-axis showed the negative logarithm (base 10) of the p-value, with lower p-values (higher statistical significance) positioned higher on the graph. Data points above the horizontal dashed line indicated conditions with p-values below the 0.05 threshold for significance, while the vertical dashed line at $x = 0$ marked neutral expression differentiation. This framework provided a comparative overview of protein expression variations between conditions within each category. Each point represented the difference between a condition's expression and the average of all others in the category. Conditions that clustered closely were more likely to have similar protein expression profiles. This approach allowed direct comparison of multiple experiments and conditions, extending the traditional use of volcano plots, which typically compare only two conditions. For instance, data representing the Muscle condition and its Fiber types (old vs. young, fast vs. slow) from two separate experiments were closely aligned across all five categories due to their similar biological contexts. In contrast, 2D and 3D models, which were less accurate representatives of *in vivo* muscle samples, were positioned further apart. Similarly, conditions like Tendon and Musculotendinous Junction, from another experiment, were still positioned closer to Muscle, reflecting their anatomical proximity and higher physiological similarities.

Different conditions were further analyzed using correlation network analysis (Fig. 5) to supplement the analysis performed with volcano plots using a network-based visualization. The Python script conducted the correlation network analysis on proteomics datasets by first preprocessing the data through averaging replicates for each condition and then standardizing the protein expression values. A Pearson correlation matrix was then computed, capturing the linear relationships between pairs of experimental conditions. The Pearson correlation coefficient, ranging from −1 to 1, measures the linear dependence between two variables, 1 indicates a perfect positive correlation, 0 indicates no correlation, and −1 indicates a perfect negative correlation. From this matrix, a network graph was

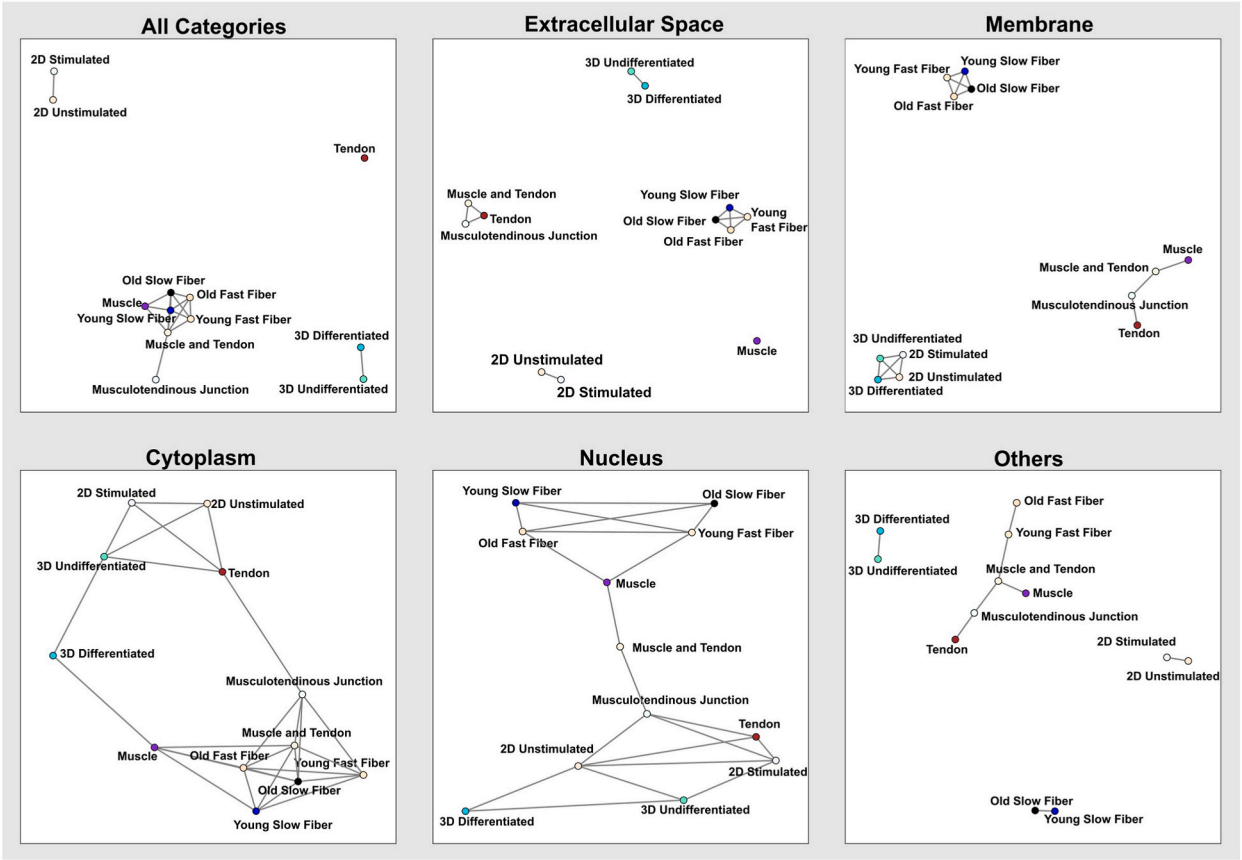


Fig. 5. Correlation network analysis for different conditions in the combined dataset as well as individual categories. In each network graph, conditions are represented as nodes and their significant interrelations, determined by a threshold of 0.6, are visualized as edges connecting the nodes. Compared to other types of analyses performed in this study, this analysis showed greater sensitivity to cross-experiment variations. Some clusters consisted of conditions from the same experiments rather than clustering conditions with more pronounced physiological or anatomical similarities.

constructed where each condition was represented as a node. Significant correlations, determined by a threshold of 0.6, were visualized as edges connecting the nodes. This graphical representation allowed for intuitive insights into the relationships between conditions, showing strong positive or negative correlations within the dataset. Similar correlation network analysis has traditionally been employed in numerous studies, primarily to show positive or negative correlations between genes, proteins, and metabolites, facilitating biomarker discovery and pathway analysis [32,33]. In this context, this analysis was harnessed to highlight correlations between distinct conditions rather than individual proteins. This approach offered a powerful tool for a holistic understanding of broader conditions using proteomics datasets collected from different experiments to minimize the effect of variability coming from different experimental conditions.

The combined dataset's correlation network presented distinct clusters (Fig. 5). Muscle tissue, including fibers from young and old subjects, formed a cohesive cluster. The mixed muscle-tendon tissue correlated strongly with muscle conditions, while the musculotendinous junction, although within this cluster, was distantly connected. *in vitro* 2D and 3D models formed isolated clusters without connections, and tendon remained uncorrelated with other conditions. For the Extracellular Space, muscle fibers clustered together with robust correlations, while muscle tissue lacked strong links with other conditions. Tendon, mixed muscle-tendon, and musculotendinous junction coalesced into a cluster with interwoven connections. *in vitro* 2D and 3D models retained their individual clusters. In the Membrane category, the musculotendinous junction was the central connection within the tendon and muscle cluster, and *in vitro* 2D and 3D models formed a unified cluster. The Cytoplasm and Nucleus categories had one major cluster with two discernible subclusters, one of which centered on muscle, its fiber types, and tendon, and the other on *in vitro* models. Both subclusters showed strong internal correlations. Finally, the Others category exhibited unique clustering, with *in vitro* models and slow muscle fibers as distinct units, while other conditions loosely connected into a singular cluster.

After combining various proteomics datasets, a total of 7145 proteins (features) were compiled across five studies/experiments and 12 conditions (labels). A classifier machine learning model was constructed using the Random Forest Library and a K-fold cross-validator with a K-value of 3 from SciKit-Learn. To identify the optimal hyperparameters for this model, a grid search paired with 3-fold cross-validation was used. The model was then trained using the best parameters on 70 % of the data, reserving the remaining 30 % for testing. The evaluation of this model was visualized through a heatmap confusion matrix, and metrics including accuracy, precision, recall, and f1-score were calculated, all showing low performances (Table 2). Subsequently, feature selection was performed on the combined dataset using SciKit-Learn's feature selection library, specifically employing a RandomForest to rank feature importance based on impurity reduction in decision trees. From the total number of proteins, 1075 were identified as significant. Despite the substantial reduction in the number of features, the model's ability to classify the conditions saw only a marginal enhancement, with accuracy, precision, and recall remaining low. Given the total of 12 conditions, a random selection would yield an 8.3 % probability of accurately choosing a condition, but the model's accuracy saw a modest enhancement of 6 % following the feature selection process (Table 2). The model faced challenges classifying specific conditions, including muscle, fiber types, and the musculotendinous junction apparent in the confusion matrices of the models (Fig. 6a and b).

An ablation study was also conducted to investigate the impact of feature selection on the model's performance. The top 25, 50, and 75 % of features, based on their importance scores, were retained and used to train and evaluate the same Random Forest classifier (Table 2). Model accuracy, precision, recall, and f1-score were compared across these subsets. Additionally, memory usage and processing time were measured for each subset. The results showed a moderate decrease in memory usage (from 282 MB to 274 MB) and processing time (from 67 to 63 s) when reducing the feature set from 75 % to 25 %, indicating some improvements in computational efficiency. These reductions in computational requirements could be beneficial for cross-validation and hyperparameter tuning, especially with much larger datasets. However, the ablation study revealed that while reducing the number of features generally improved computational efficiency, the improvements in model accuracy were not significant, with the 75 % group even underperforming the combined dataset. This underscores the importance of balanced and targeted feature selection and categorization to optimize both accuracy and computational efficiency. On the other hand, performing a similar ablation study on each of the cellular

Table 2
Accuracy, Precision, Recall, and f1-scores for the developed Random Forest Classifiers used with the entire dataset, different feature selection methods, ablation study, as well as cellular composition categorization. Weighted averages are reported.

Proteomics Dataset	Accuracy	Precision	Recall	F1-score
Entire Dataset	**0.50**	**0.44**	**0.50**	**0.46**
Feature Selection				
RandomForest	0.56	0.47	0.56	0.49
PCA	0.50	0.47	0.50	0.48
LDA	0.50	0.43	0.50	0.45
Ablation Study				
With 75 % feature retention	0.38	0.41	0.38	0.38
With 50 % feature retention	0.50	0.48	0.50	0.49
With 25 % feature retention	0.56	0.48	0.56	0.50
Categories				
Extracellular Space	0.75	0.71	0.75	0.72
Membrane	0.62	0.65	0.62	0.62
Cytoplasm	0.75	0.75	0.75	0.73
Nucleus	0.69	0.67	0.69	0.65
Others	0.81	0.78	0.81	0.79

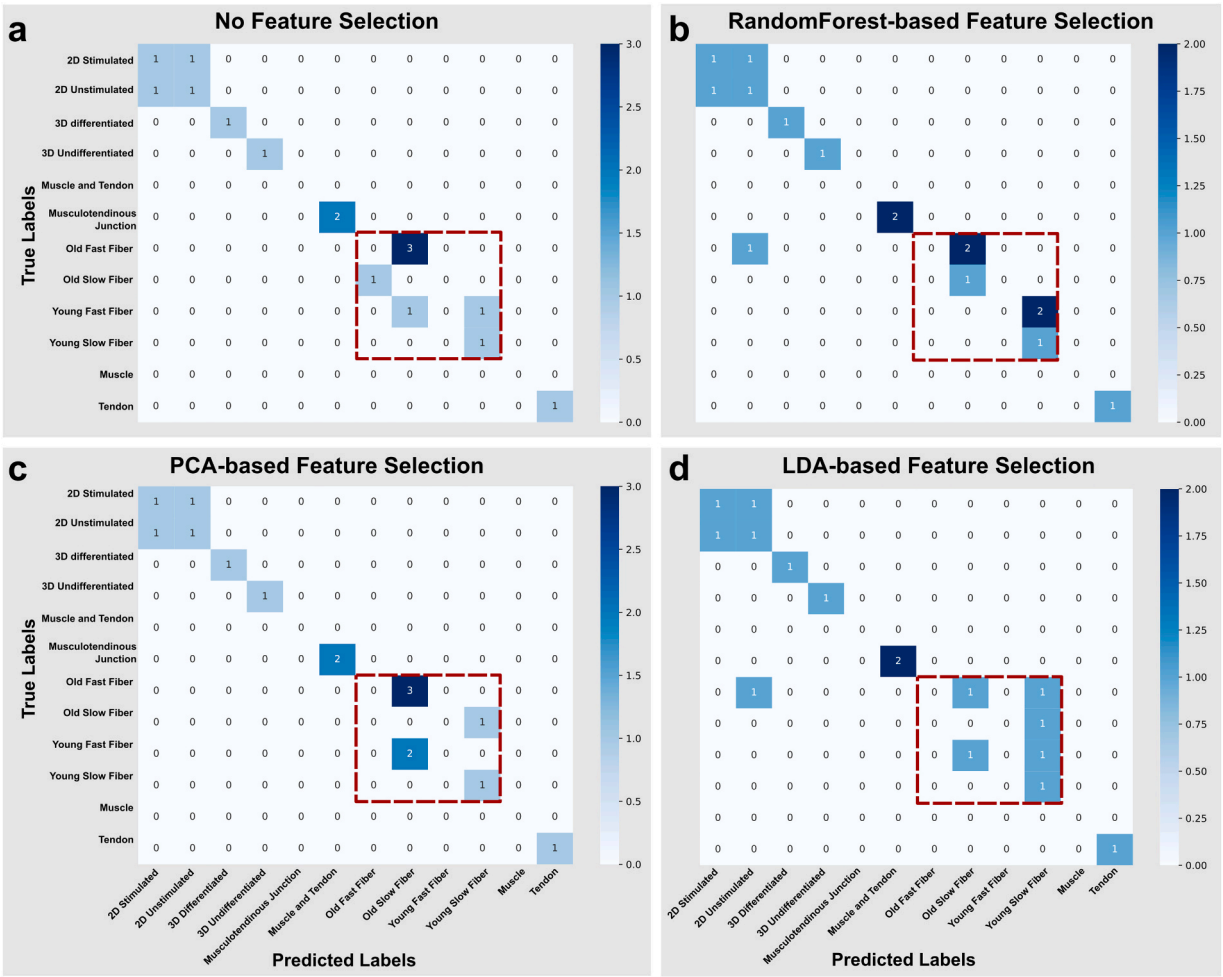


Fig. 6. Heatmap representations of the confusion matrices for Random Forest Classifier machine learning models trained on **a)** the entire dataset, and feature selected versions of dataset through **b)** RandomForest, **c)** PCA, and **d)** LDA methods. Even with a notably reduced protein numbers using different feature selection methods, there was no marked enhancement in the model's predictive abilities for different conditions. The optimal hyperparameters identified in both scenarios included: max_depth: None, min_samples_leaf: 1, min_samples_split: 2, and n_estimators: 100.

composition categories themselves would not be biologically meaningful. These categories represent biological sections of the tissue where proteins are physically co-located and might function as parts of an integrated system. Removing or combining proteins from these categories could potentially disrupt the biological context that the method described here specifically aims to preserve.

RandomForest-based feature selection was used in this study as it provides an intuitive measure of feature importance, robustly handles large datasets to prevent overfitting, and captures both linear and non-linear relationships, making it a robust choice for feature selection in complex datasets. This made it particularly effective for identifying significant features in proteomics data, balancing dimensionality reduction with the retention of crucial information for accurate classification [34,35]. However, it proved to be ineffective in this case, and two alternative methods based on PCA and LDA were also used. These methods offer different advantages compared to Random Forest. PCA can effectively reduce dimensionality by capturing the maximum variance in the data, although it primarily captures linear relationships. LDA, as a supervised method, explicitly considers class labels, enhancing its discriminative power by maximizing the separation between classes. PCA identified 3949 proteins to be significant, while LDA selected 4881 significant proteins. Both methods employed adaptive thresholding based on cumulative explained variance. Despite the different number of features selected, both significantly higher than the RandomForest-based method, when used with the same Random Forest classifier, neither improved the model's performance compared to the original RandomForest-based feature selection (Table 2, Fig. 6c and d). This lack of performance improvement suggested that the challenge in classifying the conditions might lie in the inherent complexity of the proteomics data rather than the feature selection method used. It's worth mentioning that while other feature selection methods (such as Filter and Wrapper approaches) could be explored here, the selection of PCA, LDA, and RF as representatives of different feature selection paradigms (unsupervised, supervised, and embedded approaches, respectively), was sufficient to demonstrate that biologically-informed categorization outperforms traditional feature selection approaches in this

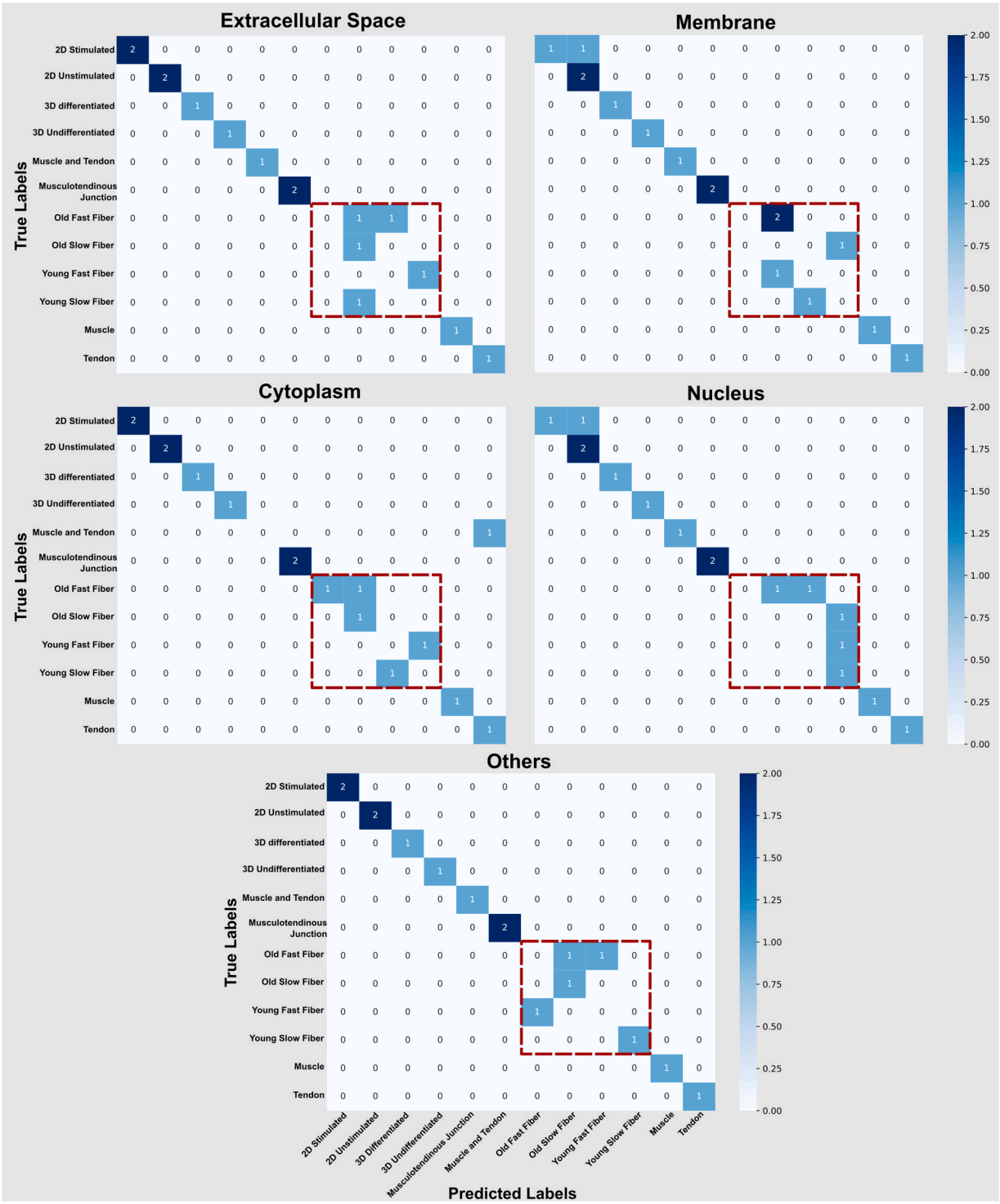


Fig. 7. Confusion matrices for the Random Forest Classifier models tailored for each cellular composition category. For all categories, excluding the Nucleus, the optimal hyperparameters were: max_depth: None, min_samples_leaf: 1, min_samples_split: 2, and n_estimators: 50. For the Nucleus category, the best hyperparameters identified were: max_depth: None, min_samples_leaf: 1, min_samples_split: 2, and n_estimators: 200. Dashed lines show models' struggles with classifying different muscle fiber types.

context.

The combined dataset included a high number of proteins, showcasing considerable variations in expression levels across conditions (Fig. 3). The feature selection methods did not significantly enhance the model's accuracy. These feature selection approaches often prioritize proteins based on the magnitude of their variations among conditions, which might not be optimal since some proteins with modest variations could still possess critical biological significance. While algorithms leveraging domain knowledge are being developed for more precise feature elimination, constructing separate Random Forest classifiers for each cellular composition category emerged as the preferred approach. This strategy ensured the retention of all proteins within each category, regardless of their expression levels or variation across conditions. Using the same Random Forest classifier model led to a noticeable increase in accuracy and other metrics across all categories, with improvements ranging between 12 % and 31 % (Table 2).

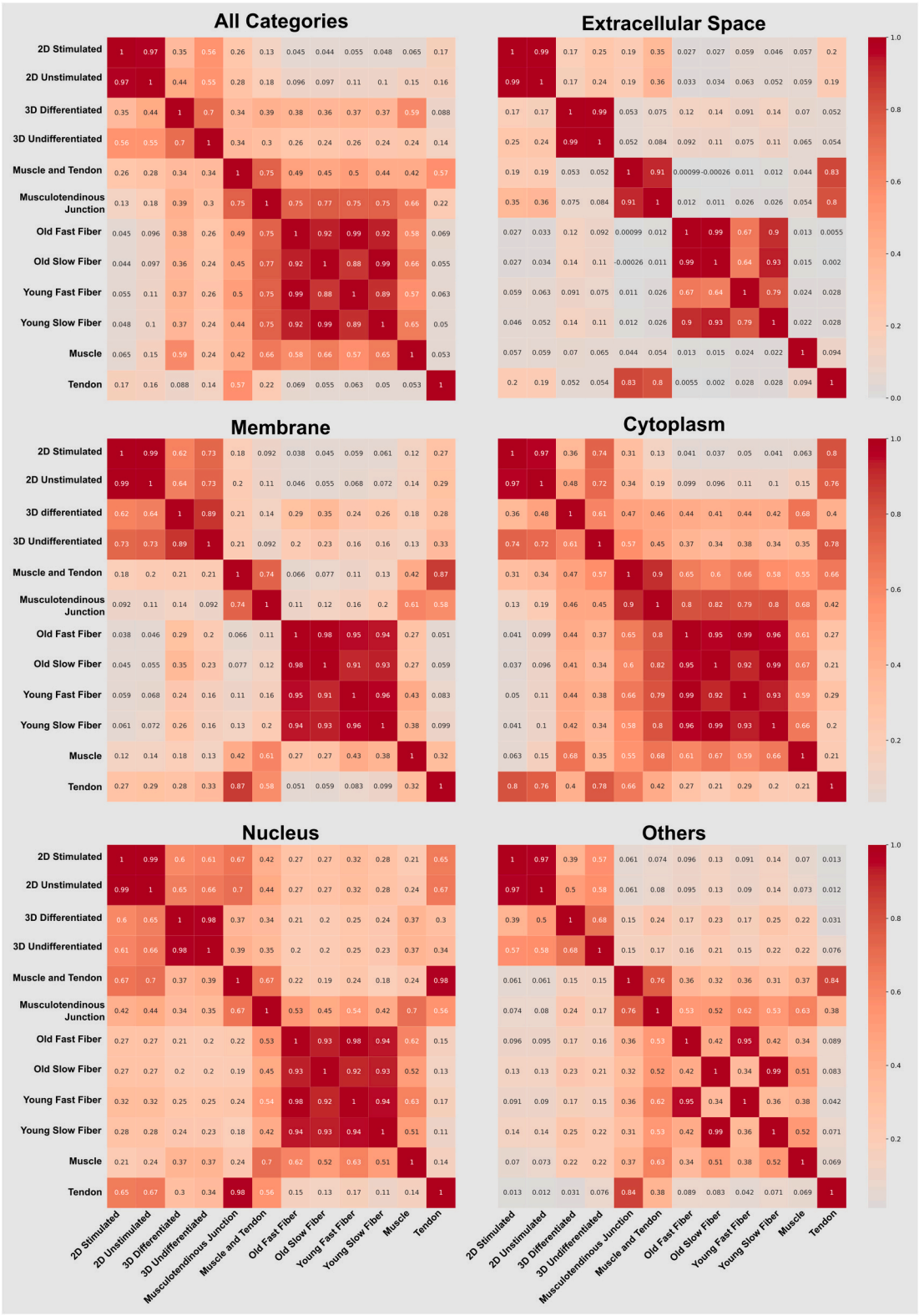
The mismatch between model parameters and sample size, specifically when data acquisition from multiple sources is performed, usually results in subpar performance of the models, making data quantity governance necessary [15]. The potential benefits of incorporating physical laws and prior knowledge to perform informed machine learning and to deal with insufficient training data has been already shown [36,37]. Other examples of incorporating domain knowledge into feature selection in fields such as materials science include multi-layer feature selection method integrating weighted expert knowledge for materials property modeling [38] as well as embedding materials domain knowledge with feature selection to deal with highly correlated features in selection process [39]. These approaches demonstrated improved prediction accuracy and interpretability across multiple material datasets. Interestingly it is also possible to combine such methods with the biologically informed categorization introduced here to further improve the models' performance.

Despite enhancements in the performance of the Random Forest Classifier models in these cellular composition categories, they continued to struggle with accurately classifying different fiber types and distinguishing them in all categories (Fig. 7), likely due to pronounced similarities among some conditions. Additionally, isolating individual fibers from muscle tissue might have excluded other tissue components, particularly those associated with the extracellular space, potentially influencing the model's outcomes. The limited number of datasets and their sourcing from different studies might have introduced variability and noise as well. Enhancing the model's performance may require training on additional datasets. Using pure intensities from MaxQuant, while streamlining the process, could be supplemented with parameters like retention time, fragmentation intensity, ion mobility, and protein detectability for more precise outcomes. Despite concerns about reproducibility with Random Forest models, hyperparameter tuning and setting a random seed were incorporated to mitigate such issues. With the limited number of studies, $K = 3$ was used, but a larger K value in K-fold cross-validation could improve the model. Additionally, robust non-linear modeling techniques such as Gradient Boosting Machines and ensemble techniques to stack multiple models could be considered to develop better-performing models.

Despite advancements in the Random Forest classifier models' capacity to distinguish between different conditions after categorization, the models faced challenges in accurately categorizing certain conditions (Fig. 7). This could be attributed to the high similarity levels among these conditions. To further investigate this aspect, a correlation analysis was conducted to assess the correlation in protein expression intensity values across conditions within both the combined dataset and individual categories. The correlation matrix was visually represented using a heatmap, displaying coefficients ranging from -1 to 1 (Fig. 8). The heatmaps provided a comprehensive perspective on overall protein expression behavior in the proteomics dataset. Notably, the observed correlations were predominantly positive, suggesting a general trend of protein expressions across the analyzed conditions. These heatmaps highlighted inherent similarities in each of the 2D and 3D *in vitro* culture conditions and among different fiber types, with varying degrees of similarity between different categories. For example, despite originating from separate studies, Muscle and Fiber types exhibited high correlation coefficients, indicating a strong linear relationship in their protein expression patterns. Additionally, the 3D differentiated model showed a higher correlation coefficient with the Muscle condition compared to the 3D undifferentiated one. Such findings underscored the pronounced similarities and differences across various categories, consistent with the classifier model's difficulties in accurately categorizing some conditions.

Performance of the trained model in classifying datasets not used during its training was evaluated with datasets that included four conditions from two different experiments (Table 3). The first experiment used myogenic progenitors derived from induced pluripotent stem cells (iPSCs) and created 3D *in vitro* models using a blend of fibrinogen and Matrigel as the ECM [40]. Both differentiated and undifferentiated samples were examined through proteomics analysis. In training of the model however, the 3D *in vitro* models were constructed with commercially available Pax7 positive primary human myoblasts with Collagen I and Matrigel as ECM [19]. In the second experiment, biopsies from the right atria and left ventricle of men undergoing mitral valve surgery were tested [41], but no heart tissue datasets were used in the model's training. The analysis was conducted with a Python script employing the pre-trained Random Forest model to classify test conditions by matching them with the top two most similar training conditions and calculating similarity probabilities as percentages (Table 4). This process involved computing mean expression levels for condition across replicates before feeding the test data into the Random Forest model.

The Random Forest model accurately identified 3D *in vitro* models from the test dataset as matching those from the training dataset, either as the top choice or the second option in almost all categories. The accuracy of classification and the similarity scores were higher for the categories compared to the combined dataset. However, the similarity scores were relatively low (20–40 %), possibly due to variations in experimental conditions as well as the less mature state of iPSC-derived *in vitro* skeletal muscle models compared to those derived from primary cells. Differentiating iPSCs into mature skeletal muscle has proven challenging, with iPSC-derived models tending to resemble fetal tissues more closely rather than adult ones [42]. For heart tissues, the left ventricle and atria, where the training dataset lacked similar samples, the model often identified the musculotendinous junction, a condition related to skeletal muscle, as the most similar. This choice reflects the known similarities between skeletal muscle and heart tissues [43]. Surprisingly, the similarity scores calculated for heart tissue conditions were higher (25–55 %) than those of the 3D models. Given the complex nature of



(caption on next page)

Fig. 8. Correlation analysis heatmaps for protein expressions across combined dataset and individual categories. Each heatmap is color-coded to visually denote the degree of linear correlation between protein expressions, with correlation coefficients ranging from −1 (indicative of a perfect inverse correlation) to 1 (indicative of a perfect positive correlation). The analysis showed overall similarities and differences in protein expression patterns among different conditions. The correlation matrices underscored the challenges faced by the classifier model in accurately categorizing conditions with pronounced similarities, aiding the interpretation of classification results.

Table 3
List of PRIDE datasets used for testing the machine learning model, but not in its training. Data was extracted from 2 experiments with 4 conditions and a total of 19 replicates.

PRIDE accession number	Experimental detail	Condition, Number of replicates
PXD045145 (van der Wal et al. [40])	3D *in vitro* models of skeletal muscles made with iPSC derived myogenic progenitors	3D model without differentiation, 3 3D model with differentiation, 3
PXD008722 (Linscheid et al. [41])	Heart biopsies from male subjects	Left ventricle, 7 Left atria, 6

Table 4
Classification of test conditions compared to training conditions through pre-trained Random Forest model. The color green represents correct classification as first or second candidate, orange and red represent somewhat similar or unsimilar first candidates, respectively.

Test condition	Cellular composition category	Candidate 1, Score (%)	Candidate 2, Score (%)
3D model without differentiation	All categories	3D Undifferentiated, 27	Musculotendinous Junction, 25
	Extracellular Space	3D Undifferentiated, 30	Tendon, 20
	Membrane	3D Undifferentiated, 34	Muscle and Tendon, 21
	Cytoplasm	3D Undifferentiated, 38	Musculotendinous Junction, 20
	Nucleus	3D Undifferentiated, 32	3D Differentiated, 20
	Others	3D Differentiated, 26	Muscle and Tendon, 24
3D model with differentiation	All categories	3D Undifferentiated, 27	Musculotendinous Junction, 25
	Extracellular Space	3D Differentiated, 24	3D Undifferentiated, 24
	Membrane	3D Undifferentiated, 29	3D Differentiated, 24
	Cytoplasm	3D Undifferentiated, 34	3D Differentiated, 24
	Nucleus	3D Differentiated, 29.5	3D Undifferentiated, 22.5
	Others	3D Differentiated, 30	Musculotendinous Junction, 24
Heart left ventricle	All categories	Musculotendinous Junction, 30	3D Undifferentiated, 21
	Extracellular Space	Musculotendinous Junction, 50	Tendon, 14
	Membrane	Musculotendinous Junction, 40	3D Undifferentiated, 18
	Cytoplasm	Musculotendinous Junction, 43	3D Undifferentiated, 19
	Nucleus	3D Undifferentiated, 27	3D Differentiated, 24.5
	Others	Musculotendinous Junction, 26	3D Differentiated, 25
Heart left atria	All categories	Musculotendinous Junction, 40	3D Undifferentiated, 21
	Extracellular Space	Musculotendinous Junction, 54	Tendon, 12
	Membrane	Musculotendinous Junction, 36	3D Undifferentiated, 18
	Cytoplasm	Musculotendinous Junction, 45	3D Undifferentiated, 17
	Nucleus	3D Undifferentiated, 27.5	3D Differentiated, 24
	Others	Musculotendinous Junction, 26	3D Differentiated, 25

the data, the similarities among conditions, and the limited datasets used for training, the Random Forest model's performance was enhanced by incorporating cellular composition as a method inspired by domain knowledge, an alternative to traditional feature selection techniques, but the model's effectiveness could be further improved by training it with larger numbers of datasets.

Analysis of proteomics workflows presents considerable challenges due to the intricate experimental processes and the inherent ambiguities in the generated data. Such complexities, coupled with high dimensionality of datasets, necessitate the use of informatic strategies, predominantly machine learning methodologies. Although machine learning can be incorporated at different steps of the proteomic analysis process, its applications are largely focused into two primary domains, direct application to mass spectral peaks, and to proteins identified via sequence dataset searches. Historically, most machine learning applications in proteomics have belonged to the former [11,44]. Recent trends, however, have seen a shift towards using protein abundances for sample classification based on tissue and cell types. Nevertheless, even when pronounced differences between conditions exist, certain ambiguities in classification persist due to the inherent intricacies of proteomic studies [45]. This study underscored such ambiguities, especially when evaluating

conditions with high biological similarities. Although the merits of combining and repurposing multiple proteomic datasets to craft a more comprehensive view of human physiology have been considered before, proper tools for performing such analyses are yet to be developed [46–48]. This study demonstrated that heterogeneity originating from different experimental conditions can compromise model efficacy. Such challenges may intensify as higher numbers of datasets from different experimental conditions are combined. This shows the importance of proper standardization across all experimental and analytical steps to mitigate biases arising from methodological differences and developing custom algorithms for their analyses. In summary, this study highlighted that, despite challenges of integrating machine learning and proteomic datasets, considerable progress is achievable by adding domain-specific physiological insights to enhance machine learning models. Further improvements could be explored by training a diverse suite of machine learning algorithms beyond the Random Forest Classifier used here, tailoring the choice to the specific characteristics of the proteomic datasets in question.

4. Conclusions

The current study examined the interplay between machine learning and proteomics, focusing on various *in vitro* and *in vivo* skeletal muscle conditions as well as adjacent tissues like tendons. Proteomics datasets were sourced from publicly accessible repositories, and a series of Python scripts were developed to fetch information from diverse online platforms to enhance the datasets. This added information facilitated the categorization of proteins according to their cellular composition. Various analytic techniques, including PCA, LDA, feature importance analysis, and correlation networks were used on both the combined dataset and distinct cellular composition categories. The preliminary application of a Random Forest Classifier model on the combined datasets highlighted certain limitations in model accuracy. Nonetheless, the consideration of cellular composition categories provided pivotal insights that might have remained undiscovered if only the combined dataset was analyzed. Incorporating more studies, especially those with similar experimental procedures, can further improve such models' efficacy. These findings emphasized the important role of machine learning in deciphering proteomics datasets and advocated the need for a broader investigative scope instead of a narrow focus on individual studies, especially when complemented by domain-specific expertise, to improve understanding of complex biological systems.

CRediT authorship contribution statement

Alireza Shahin-Shamsabadi: Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Funding acquisition, Conceptualization. **John Cappuccitti:** Writing – review & editing, Validation, Project administration, Funding acquisition.

Data and code availability

Data associated with this study is included as supplementary information. Further requests for resources from the current study should be directed to and will be fulfilled by the lead contact, Alireza Shahin-Shamsabadi (alireza@itsevolved.com). The Python scripts supporting the findings of this study are available on GitHub at https://github.com/Evolved-Bio/MLProteomicsCellCompoFeatureSelection.

Declaration of use of generative AI and AI-assisted technologies

During the preparation of this work, the authors used Open AI's ChatGPT in order to grammatically edit the manuscript and edit the Python scripts. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

The authors declare no conflict of interest. All expenses are covered by the authors' institution, Evolved.Bio.

Acknowledgements

This work was partially performed as part of activities in the DaRMoD program from the Vector Institute for Artificial Intelligence. The authors would like to thank Michael Joseph and Sajjad Pakdamansavoji for their support.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e40772.

References

[1] J. Peng, S.P. Gygi, Proteomics: the move to mixtures, J. Mass Spectrom. 36 (10) (2001) 1083–1091.
[2] J.R. Yates, 3rd, the revolution and evolution of shotgun proteomics for large-scale proteome analysis, J. Am. Chem. Soc. 135 (5) (2013) 1629–1640.
[3] L. Martens, et al., PRIDE: the proteomics identifications database, Proteomics 5 (13) (2005) 3537–3545.
[4] Y. Perez-Riverol, et al., The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences, Nucleic Acids Res. 50 (D1) (2022) D543–d552.
[5] M. Vaudel, et al., Exploring the potential of public proteomics data, Proteomics 16 (2) (2016) 214–225.
[6] M. Gonzalez-Freire, et al., The human skeletal muscle proteome project: a reappraisal of the current literature, J Cachexia Sarcopenia Muscle 8 (1) (2017) 5–18.
[7] K. Ohlendieck, Skeletal muscle proteomics: current approaches, technical challenges and emerging techniques, Skeletal Muscle 1 (1) (2011) 6.
[8] K. Ohlendieck, Proteomic identification of biomarkers of skeletal muscle disorders, Biomarkers Med. 7 (1) (2013) 169–186.
[9] V. Corasolla Carregari, et al., Proteomics of muscle microdialysates identifies potential circulating biomarkers in facioscapulohumeral muscular dystrophy, Int. J. Mol. Sci. 22 (1) (2020).
[10] P. Dowling, A. Holland, K. Ohlendieck, Mass spectrometry-based identification of muscle-associated and muscle-derived proteomic biomarkers of dystrophinopathies, J. Neuromuscul. Dis. 1 (1) (2014) 15–40.
[11] A.L. Swan, et al., Application of machine learning to proteomics data: classification and biomarker identification in postgenomics biology, OMICS 17 (12) (2013) 595–610.
[12] N. Pudjihartono, et al., A review of feature selection methods for machine learning-based disease risk prediction, Front Bioinform 2 (2022) 927312.
[13] B. Mirza, et al., Machine learning and integrative analysis of biomedical big data, Genes 10 (2) (2019).
[14] H. Desaire, E.P. Go, D. Hua, Advances, obstacles, and opportunities for machine learning in proteomics, Cell Rep Phys Sci 3 (10) (2022).
[15] Y. Liu, et al., Data quantity governance for machine learning in materials science, Natl. Sci. Rev. 10 (7) (2023) nwad125.
[16] M. Murgia, et al., Protein profile of fiber types in human skeletal muscle: a single-fiber proteomics study, Skeletal Muscle 11 (1) (2021) 24.
[17] A. Di Meo, et al., Proteomic profiling of the human tissue and biological fluid proteome, J. Proteome Res. 20 (1) (2021) 444–452.
[18] A.M. Mengeste, et al., Insight into the metabolic adaptations of electrically pulse-stimulated human myotubes using global analysis of the transcriptome and proteome, Front. Physiol. 13 (2022) 928195.
[19] R.J. Mills, et al., Development of a human skeletal micro muscle platform with pacing capabilities, Biomaterials 198 (2019) 217–227.
[20] A. Karlsen, et al., The proteomic profile of the human myotendinous junction, iScience 25 (2) (2022) 103836.
[21] Y. Zhang, et al., ProteinInferencer: confident protein identification and multiple experiment comparison for large scale proteomics projects, J. Proteonomics 129 (2015) 25–32.
[22] K.D.B. Anapindi, et al., Peptide identifications and false discovery rates using different mass spectrometry platforms, Talanta 182 (2018) 456–463.
[23] G.T. Reddy, et al., Analysis of dimensionality reduction techniques on big data, IEEE Access 8 (2020) 54776–54788.
[24] N. Sharma, K. Saroha, A novel dimensionality reduction method for cancer dataset using PCA and Feature Ranking, in: 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2015.
[25] A.M. Rodríguez-Piñeiro, F.J. Rodríguez-Berrocal, M. Páez de la Cadena, Improvements in the search for potential biomarkers by proteomics: application of principal component and discriminant analyses for two-dimensional maps evaluation, J. Chromatogr., B: Anal. Technol. Biomed. Life Sci. 849 (1–2) (2007) 251–260.
[26] D.L. Sampson, et al., A comparison of methods for classifying clinical samples based on proteomics data: a case study for statistical and machine learning approaches, PLoS One 6 (9) (2011) e24973.
[27] E.M. Mirkes, et al., Domain adaptation principal component analysis: base linear method for learning with out-of-distribution data, Entropy 25 (1) (2022).
[28] J. Wang, et al., Engineered skeletal muscles for disease modeling and drug discovery, Biomaterials 221 (2019) 119416.
[29] A. Khodabukus, Tissue-Engineered skeletal muscle models to study muscle function, plasticity, and disease, Front. Physiol. 12 (2021) 619710.
[30] M. Lualdi, M. Fasano, Statistical analysis of proteomics data: a review on feature selection, J. Proteonomics 198 (2019) 18–26.
[31] Z. Shi, et al., Feature selection methods for protein biomarker discovery from proteomics or multiomics data, Mol. Cell. Proteomics 20 (2021) 100083.
[32] P. Langfelder, S. Horvath, WGCNA: an R package for weighted correlation network analysis, BMC Bioinf. 9 (2008) 559.
[33] H. Kitagawa, et al., Phenotyping analysis of the Japanese Kampo medicine maoto in healthy human subjects using wide-targeted plasma metabolomics, J. Pharm. Biomed. Anal. 164 (2019) 119–127.
[34] M. Hauskrecht, et al., Feature selection and dimensionality reduction in genomics and proteomics, in: W. Dubitzky, M. Granzow, D. Berrar (Eds.), Fundamentals of Data Mining in Genomics and Proteomics, Springer US, Boston, MA, 2007, pp. 149–172.
[35] A. Destrero, et al., Feature selection for high-dimensional data, Comput. Manag. Sci. 6 (1) (2009) 25–40.
[36] Z. Hao, et al., Physics-informed machine learning: a survey on problems, methods and applications, arXiv preprint arXiv:2211.08064 (2022).
[37] L.v. Rueden, et al., Informed machine learning – a taxonomy and survey of integrating prior knowledge into learning systems, IEEE Trans. Knowl. Data Eng. 35 (1) (2023) 614–633.
[38] Y. Liu, et al., Multi-layer feature selection incorporating weighted score-based expert knowledge toward modeling materials with targeted properties, Advanced Theory and Simulations 3 (2) (2020) 1900215.
[39] Y. Liu, et al., Feature selection method reducing correlations among features by embedding domain knowledge, Acta Mater. 238 (2022) 118195.
[40] E. van der Wal, et al., Highly contractile 3D tissue engineered skeletal muscles from human iPSCs reveal similarities with primary myoblast-derived tissues, Stem Cell Rep. 18 (10) (2023) 1954–1971.
[41] N. Linscheid, et al., Quantitative proteomics of human heart samples collected in vivo reveal the remodeled protein landscape of dilated left atrium without atrial fibrillation, Mol. Cell. Proteomics 19 (7) (2020) 1132–1144.
[42] F. Iberite, E. Gruppioni, L. Ricotti, Skeletal muscle differentiation of human iPSCs meets bioengineering strategies: perspectives and challenges, NPJ Regen Med 7 (1) (2022) 23.
[43] C. Lindskog, et al., The human cardiac and skeletal muscle proteomes defined by transcriptomics and antibody-based profiling, BMC Genom. 16 (1) (2015) 475.
[44] R. Bouwmeester, et al., The age of data-driven proteomics: how machine learning enables novel workflows, Proteomics 20 (21–22) (2020) e1900351.
[45] T. Claeys, et al., Machine learning on large-scale proteomics data identifies tissue and cell-type specific proteins, J. Proteome Res. 22 (4) (2023) 1181–1192.
[46] K. Verheggen, L. Martens, Ten years of public proteomics data: how things have evolved, and where the next ten years should lead us, EuPA Open Proteomics 8 (2015) 28–35.
[47] Y.K. Paik, et al., Toward completion of the human proteome parts list: progress uncovering proteins that are missing or have unknown function and developing analytical methods, J. Proteome Res. 17 (12) (2018) 4023–4030.
[48] S. Adhikari, et al., A high-stringency blueprint of the human proteome, Nat. Commun. 11 (1) (2020) 5301.