Machine Learning Identifies Plasma Proteomic Signatures of Descending Thoracic Aortic Disease

Amanda Momenzadeh PharmD[1,2,3], Simion Kreimer PhD[2,3], Dongchuan Guo PhD[6], Matthew Ayres[2], Daniel Berman MD[3,5], Kuang-Yuh Chyu MD[3], Prediman K Shah MD[3], Dianna Milewicz MD PhD[6], Ali Azizzadeh MD, Jesse G. Meyer PhD[1,2,3]*, Sarah Parker PhD[2,3,4]*
*Co-Corresponding authors

[1] Department of Computational Biomedicine, Cedars Sinai Medical Center, Los Angeles, California, USA
[2] Advanced Clinical Biosystems Research Institute, Cedars Sinai Medical Center, Los Angeles, California, USA
[3] Schmidt Heart Institute, Cedars Sinai Medical Center, Los Angeles, California, USA
[4] Department of Biomedical Sciences, Cedars Sinai Medical Center, Los Angeles California, USA
[5]Cedars-Sinai Imaging Department, Cedars Sinai Medical Center, Lost Angeles, California, USA
[6] Department of Internal Medicine, McGovern Medical School, University of Texas Health Science Center, Houston, Texas

* Correspondence to Jesse Meyer (email: jesse.meyer@cshs.org) or Sarah Parker (email: sarah.parker@cshs.org)

Abstract
Descending thoracic aortic aneurysms may go undetected until severe and catastrophic complications such as rupture or dissection occur. Few clinical indicators exist to screen for aneurysms or predict risk of dissection or rupture. This study generated a plasma proteomic dataset from 150 patients with descending thoracic aortic disease and 52 controls to identify proteomic signatures capable of differentiating descending thoracic aortic disease from controls, as well as aneurysm from descending aortic dissections. Of the 1,468 peptides and 195 proteins quantified, 853 peptides and 99 proteins were significantly different between disease and controls. Using machine learning to classify disease from control, the highest precision-recall area under the curve (PR AUC) was achieved using significantly different proteins (PR AUC 0.99), followed by significant peptides (PR AUC 0.96). Despite no statistically significant protein signatures between aneurysm and dissection, use of all proteins was able to modestly classify between the disease states (PR AUC 0.77). To address correlation between proteins, a disease versus control classifier was optimized using only seven unique protein clusters, which achieved comparable performance to models trained on all/significant proteins (PR AUC 0.90). Model interpretation with permutation importance revealed that proteins differentiating disease and control function in coagulation, protein-lipid complex remodeling, and acute inflammatory response.

Introduction

Thoracic aortic aneurysms arise due to dysregulated growth and remodeling of the aorta in the segment spanning from the aortic root to the diaphragm[1], which predispose the vessel wall to dissection and rupture. Aortic dissections occur when there is a loss of integrity, also described as a tear, in the intimal layer of the blood vessel. This tearing results in the formation of a 'false lumen' in the medial layer of the vessel, allowing aberrant blood flow patterns, risk of aneurysm formation/rupture, risk of thrombosis, and reduced perfusion of downstream tissues, all of which are associated with substantial morbidity and mortality[2]. The incidence of thoracic aortic aneurysms and thoracic aortic aneurysm dissections (TAAD) has risen over the past several decades, with an approximate doubling of new events between 1982[3] and 2006[4] (from 5.9 to 10.1 events per 100,000 individuals, respectively). Increased incidence may be attributed to a combination of improved diagnosis as well as increased prevalence of risk-factors such as atherosclerosis, hypertension, and an aging population[4]. TAADs are further categorized by anatomical region into ascending aneurysms and Stanford type A dissections or descending and Stanford type B dissections. To date, there are a large proportion of TAAD studies focused on ascending disease, but differences in etiology and other aspects of descending disease warrant focused attention on mechanisms and biomarkers for disease cases in this specific region.

Descending TAAD can be driven by both syndromic (e.g., hereditary genetic conditions such as familial TAAD, Ehlers-Danlos and Marfan syndromes) and non-syndromic/sporadic (as yet undescribed genetic causes, atherosclerosis, and hypertension)[1]. Syndromic causes of descending TAAD are rare, and a majority of descending TAAD events occur absent any *a priori* indicators of patient risk. While imaging is a highly effective means of detecting and diagnosing TAAD, the low prevalence of TAAD in the general population renders the cost-benefit ratio of such a screening approach prohibitive. Circulating biomarkers capable of detecting the presence of descending thoracic aneurysm and risk for type B dissection would provide a valuable and cost-effective tool to screen for risk and flag individuals from the general population for more detailed follow up and diagnosis. Since many times Type B dissections occur absent of predisposing aneurysm formation[5], determining whether there are distinguishing biomarkers for these unique type B cases will be an important consideration. Dissection-specific biomarkers could also assist in evaluating the progression of descending aneurysms (DTAA) and predicting likelihood for imminent dissection risk.

A number of studies have explored and examined possible biomarkers for thoracic aortic aneurysms and dissections and are the subject of a recent and thorough review[6]. A smaller handful of studies focused specifically on descending thoracic aortic disease[7]. Among markers studied thus far, many have shown preliminary promise including d-dimer, matrix metalloproteinases, certain collagen chains, smooth muscle cell proteins, and various inflammatory markers such including the somewhat general inflammatory marker CRP. Although no circulating biomarker panels are used clinically to diagnose descending aortic disease, interest remains strong in identifying plasma protein signatures with adequate sensitivity and specificity for clinical development and use. To this end, our group and others have advocated for the use of multiplexed protein models based on proteomic analysis of plasma samples to identify individuals at elevated risk for vascular disease. Our recent work in atherosclerosis disease has generated putative targets present in early atherosclerotic lesions with excellent sensitivity and specificity for detecting coronary artery disease in a separate patient cohort[8,9], generating strong initial targets for further validation in larger samples. In those experiments, tissue proteomes of vascular lesions (and normal tissue) were first analyzed to prioritize lesion markers, and then those markers were analyzed for their detectability and informativeness in plasma. While we have recently completed a proteomic analysis of descending thoracic aortic disease[10], we also observed in our previous work that the circulating levels of putative markers do not always match what is seen in tissue, and we

hypothesize that direct comparison of plasma proteomes of disease and control patients will be an efficient and accurate approach for identifying informative protein candidates for further biomarker development in descending TAAD.

Machine learning (ML) is a promising tool for automated classification of groups from proteomics data[11]. ML can take any collection of input data and estimate a mathematical function that predicts binary outcome, such as disease versus control, or even a continuous measure like age. ML has enabled discovery of biomarkers associated with cancer, COVID-19 disease severity[12–15], and subtypes of diseases[16]. ML model interpretation or feature selection methods can be used to reveal which proteins are relevant for prediction of disease versus control groups, and prior work suggests that clustering of similar proteins as a feature selection technique before applying ML methods enables accurate disease classification[17]. These successes in the application of ML to classify disease have led us to explore whether ML models given inputs of plasma peptides or proteins could distinguish descending TAAD from controls, or even distinguish subtypes of disease.

The goal of the current work was to provide preliminary insight into mass spectrometry detectable, circulating proteomic signatures of descending aortic disease with potential for eventual disease screening and prognosis. Toward this goal, we profiled the plasma proteomes of 202 individuals with or without descending thoracic aorta disease and applied a ML to identify protein features best able to discriminate between disease-free and any-disease status, as well as features that could discriminate, albeit less robustly, between cases of aneurysm alone and cases with type B dissections that occurred absent predisposing aneurysms.

Methods

Sample Collection and Study Design

Patients with descending thoracic aortic disease were selected retrospectively from a biorepository of aortic disease patients hosted by Dr. Milewicz and team at the University of Texas, Houston. All patients with a well-preserved plasma sample and a simple diagnosis of descending thoracic aortic aneurysm only descending type B dissection only (Type B) or a combination of descending aneurysm and dissection (BDTAA) were selected for proteomic analysis. Blood samples were collected prior to surgery and held in the patient's room or nurse's station until transported by the research nurse to our laboratory (in the same day). On receipt, each sample was logged into the computerized biorepository database and labeled with a unique bar-coded identity number and processed into plasma within two hours. The collection tubes were gently inverted 8 – 10 times and then centrifuged at 1650 RCF for 25 minutes at 22°C. The plasma layer of each tube was transferred to labeled 2ml cryovial tubes (0.5mL per tube) and frozen at -80°C until further use. One aliquot per patient was shipped on dry ice to the proteomics research team at Cedars-Sinai Medical Center for plasma proteomic sample preparation and analysis. Control plasma samples were selected from biobanked samples collected for the EISNER study as described in previous publications[18], with selected cohort members matched to achieve similar age and sex composition relative to disease patients. To control for higher prevalence of atherosclerosis and coronary artery disease (CAD) among patients with descending aortic disease[1], the control cohort included patients with computed tomography verified CAD as manifested by coronary artery calcification and elevated levels of known atherosclerosis risk factors and no visible aortic disease. All patients included in this study provided informed consent and their recruitment and participation was approved by the institutional review boards of their corresponding institutions (UT Houston, Cedars-Sinai Medical Center).

Sample Preparation for Liquid Chromatography Mass Spectrometry
Proteins from 5uL of plasma were processed for protein denaturation, reduction, alkylation, and tryptic digestion using the manufacturer protocols for the Protifi (Farmingdale, NY) S-Trap protein sample preparation workflow. Resulting peptides were quantified by BCA assay and 2uL of peptide suspension from each sample was pooled to make a master mix used for quality control monitoring purposes and for generation of peptide assay libraries for peptide and protein identification from individual DIA-MS samples (see below).

Mass Spectrometry Acquisition
Individual plasma samples. Mass spectrometry data were acquired on an Orbitrap Exploris 480 (ThermoFisher, Bremen, Germany) instrument with LC separation on an Ultimate 3000 HPLC system using a trap-elute set up on a 150 mm long, 0.3 mm inner diameter reversed phase column (Phenomenex, Luna Polar C18 3 um). A binary analytical gradient using 0.1 % formic acid in water (mobile phase A) and 0.1% formic acid in acetonitrile (mobile phase B) was delivered as follows at 9.5 uL/min: start at 1% B and hold for 2 minutes, ramp to 4%B in 30 seconds, ramp to 12% B over 20 minutes, ramp to 27% B in 24 minutes, ramp to 45% B over 16 minutes (60 minutes total). A separate cleaning equilibration method ran at 98% B for 8 minutes and equilibrated at 2% B for 2 minutes.

The peptides eluted from the analytical column into a Newomics M3 8-nozzle emitter and electrosprayed at 3 kV into a 300°C ion transfer tube temperature. The mass spectrometer was operated in data independent acquisition (DIA) mode acquiring an MS1 scan for 100msec on all ions between 400-1100 m/z and then completing a series of 25msec MS/MS fragment scans on 50 equally spaced 12 m/z width precursor isolation windows. Orbitrap resolution and normalized AGC target were 60,000 and 200% for MS1 and 15,000 and 400% for MS2. Collision energy for HCD fragmentation was set to 30%. The acquisition sequence included repeated sampling of pooled digest to monitor MS QC as well as evenly spaced samples of pooled plasma from across the 3 x 96-well digestion plates to monitor digestion QC.

Gas Phase Fractionation based Library Generation
A sample pool was used to generate a spectral library specific to this sample type and analytical platform. Gas phase fractionation limits the scope of the mass spectrometer to a narrow m/z range thus exhaustively fragmenting the corresponding peptide ions and maximizing the probability of their identification and incorporation in the generated library. Multiple injections probing different narrow m/z ranges are compiled to cover the entire range of interest (400 to 1000 m/z). Two complementary approaches were used to generate this library: data dependent acquisition (DDA) of peptides within 120 m/z wide mass ranges (400 to 520 m/z, 520 to 640 m/z … 880 - 1000 m/z) in duplicate; and data independent acquisition (DIA) using 1 m/z wide isolation windows covering 40 m/z at a time (400 to 440 m/z, 440 to 480 m/z … 960 to 1000 m/z). Other than the mass ranges and isolation window widths the method settings were matched to the data acquisition method.

Proteomic Dataset Generation
Library construction. Gas phase fractionated DIA and DDA runs were analyzed using the FragPipe platform[19]. For DIA mode, DIA-Umpire was first used to extract pseudospectra[20]. DIA pseudospectra and DDA spectra were searched separately using the FragPipe workflow to perform spectral matching, PSM probability scoring and then spectral library generation against a Uniprot Human FASTA predicted protein sequence database. Search settings were asl follows he mass errors were set to +/-8 ppm for precursor and fragment masses in DDA and +/- 1 m/z for the same masses in the DIA datasets. Carbamidomethylation of cysteine was set as static and methionine oxidation, phosphorylations of

serine, threonine, and tyrosine, N-terminal acetylation, pyroQ, pyroC, and pyroE, were selected as variable modifications. Peptides identified at 1% FDR were compiled into a spectral library with EasyPQP. Spectral libraries from the DDA and DIA runs were merged at the level of the final library tsv document. We assumed that a DIA-based peptide identification would provide the best representation of the fragments and their relative intensities for identification within a subsequent DIA run. Thus, only the unique peptides from the DDA library not seen in the DIA runs were appended to the DIA library. The final library contained 8,819 precursors and 407 proteins.

Individual subject peptide and protein quantification. Individual DIA runs were processed using DIA-NN[12] by searching against the sample-specific libraries (described above) using double pass mode and match between runs. Retention time-based normalization setting was used and maxLFQ calculated protein intensities, provided in the main DIANN output matrices, were used for further analysis.

Data Cleaning

Peptides and proteins with at least one missing value were removed from further analysis, reducing the number of peptides from 8,243 to 1,471 and proteins from 357 to 234. Thirty-nine proteins with multiple uniprot identifiers in their group were removed, resulting in 195 proteins measured across all samples. Only patients with complete demographic and sample collection data (i.e., age, sex, race, collection tube, and batch number) were carried forward for analysis and four patients were removed who had duplicate rows between both peptide and protein datasets, resulting in 202 unique patients. Peptide and protein quantities were $\log_2$ transformed across each sample and corrected for batch effects using pyComBat[21]. Three peptides with zero value quantities were removed, resulting in a final 1,468 peptides quantified across all samples.

Data Visualization

Of the 202 patients, the aortic disease group (150 patients) consisted of combined DTAA, Type B and BDTAA samples and the control group (52 patients) consisted of combined healthy and CAD samples. Reduction of $\log_2$ transformed quantities from 1,468 peptides and 195 proteins to a two dimensional space using UMAP (Uniform Manifold Approximation and Projection for Dimension)[22] and partial least squares-discriminant analysis (PLS-DA) allowed for visualization of peptide and protein clustering. All UMAP representations of data were unsupervised, meaning the algorithm is provided the feature but not the class information (e.g. disease or control), so groups are formed based on the input data alone. This contrasts with PLS-DA, which bases groupings on input and output data. UMAP was additionally used to visualize differences in age, sex, race, collection tube, batch and individual cohorts between aortic disease and control groups. Volcano plots were used to visualize differentially expressed proteins and peptides between disease and control groups using all samples. Using the anti-log of the batch-corrected protein quantities, $\log_2$ fold changes were calculated by subtracting the $\log_2$ mean quantity for each protein in the control group from the $\log_2$ mean quantity in the disease group.

Statistical Analysis

Data cleaning, analysis, and model training were performed in Python version 3.7.11 (SciKit-Learn[23], SciPy[24], seaborn[25], Matplotlib[26], Plotly[27], and Statsmodels[28]). Age was compared between groups using a Wilcoxon Rank Sum Test due to non-normal distribution (Shapiro-Wilks p-value <0.05). Fisher's Exact Test was used to compare categorical variables (sex and ethnicity) between groups. Significantly expressed peptides and proteins between disease and control and individual disease groups were determined using independent two-sample t-tests and Benjamini-Hochberg (BH) multiple hypothesis testing correction. Adjusted p-values <0.01 were considered statistically significant. To avoid leakage of test data into train data when performing feature selection[29], t-tests were performed only on the 80%

train set to select significantly different peptides between groups, and then the prediction was made on the 20% test data filtered for these features. Variable importance in projection (VIP) scores, which represent the weights/importance of each feature in the PLS-DA model, were calculated using SciKit-Learn PLSRegression model.

Machine Learning

The goal of ML classification was to use an initial proteomics training set example to predict whether a patient belonged to the disease or control category as well as distinguish between diseases. Data was split into 80% training and 20% final test sets. To avoid overfitting, the 80% training split was used to tune model hyperparameters via a random search with 5-fold cross validation optimized on F1-score, and the 20% test set was held-out until final evaluation using the best model from the random hyperparameter search. This was repeated with up to 200 random sets of hyperparameters, i.e., up to 1,000 models were trained for each ML method. The best hyperparameters were then used to refit each model with the entire 80% training set before assessing final performance with the 20% test set. The model output was stratified during data splitting to represent the proportion of classes in the whole dataset. Due to imbalanced classes, metrics that focus on the minority class, including precision, recall, F1-score and Matthew's correlation coefficient (MCC) scores, were chosen to represent generalization performance of the test set. MCC is zero when prediction is random and one when false positives and false negatives are zero (i.e., perfect classification). All peptides/proteins and those with significantly different means (using only the 80% training data, BH-adjusted p-value < 0.01) were used as inputs to train one of five supervised ML classification algorithms plus logistic regression: Gradient Boosting Decision Trees (GB)[30], Support Vector Classification (SVC)[31], Random Forest (RF)[32], Extra-Trees (ET)[33], Logistic Regression (LR)[34], and K-nearest neighbors (KNN)[35]. Yellowbrick[36] package in Python was used to generate classification reports for ML models. Precision-recall (PR) curves were shown for the models with the highest F1-scores on the training sets.

Permutation Importance for Prognostic Markers of Disease

To account for the presence of correlated proteins within our dataset, we grouped quantitatively similar proteins using hierarchical clustering analysis before performing model interpretation. This way, we could obtain a ranked list of important uncorrelated proteins, which would provide a better picture of uncorrelated pathways involved in disease pathology. Using only the 80% train set to avoid biasing the feature selection method with test data[29], we calculated the Spearman correlation coefficients between each protein pair, converted the correlation matrix to a condensed distance matrix, and applied Ward's linkage to cluster proteins based on distance. In the dendrogram, the number of vertical lines intersecting a horizontal line drawn at each linkage distance threshold represents the number of clusters at that distance. To identify the linkage distance threshold corresponding to optimal model performance, we trained six ML models with a single protein from each cluster at various distance thresholds. Hyperparameter tuning was performed using random search allowing up to 100 iterations of random hyperparameter combinations using the 80% training set and 5-fold cross validation optimized on F1-score. The optimal linkage distance threshold was selected at the distance after which model performance declined, corresponding to the fewest number of features a model can be trained on to still perform well. F1-score, PR AUC and accuracy scores were calculated for the test set prediction using the model with the best F1-score on the train set to ensure good model performance was sustained and features important for predictions were reliable.

Permutation feature importance, or a decrease in accuracy score when a single feature's value is randomly shuffled, for the selected proteins was calculated on the test set with each feature's value

being randomly shuffled 10 times. Proteins with the largest decrease in accuracy score were most important to the model's predictions.

Biological Pathway Analysis

Three different methods were used to identify the most important proteins that discriminate disease from control patients to submit for pathway analysis. The first method selected proteins with the top permutation importance from the highest F1-scoring model that was trained using only sentinel proteins from each correlated protein group. Second, we identified the same number of proteins with the highest VIP scores, or the most contributory in class separation, in the PLS-DA model. Last, we selected the same number of proteins with the lowest adjusted p-values. The unique proteins that were present in the compiled list from all three datasets were submitted for pathway analysis.

GO Biological Process term enrichment analysis was performed using the ClueGO (version 2.5.9)[37] application within Cytoscape (version 3.9.1)[38]. GO database release date was 5/25/2022. The default parameters were used, except: GO term fusion was turned on, the threshold for statistical significance was set to <0.0001, and the GO tree interval was set to 5-6. Enriched terms were then manually filtered to keep only non-redundant terms that connected all the proteins to the network.

Code and Data Availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange archive via the PRIDE partner repository with the data set identifier PXD041337 and 10.6019/PXD041337.
Username: reviewer_pxd041337@ebi.ac.uk
The password for the data will be made public upon manuscript acceptance.
The python code for data analysis in JupyterLab notebooks can be found at https://github.com/xomicsdatascience/Aneurysm-ML.

Results

Study and Data Overview

Figure 1A depicts the study workflow. Plasma samples were obtained from healthy individuals, patients with CAD but not aneurysm, and individuals diagnosed with one of three types of descending thoracic aortic disease (Type B DTAA, and BDTAA). Plasma proteomes were generated using DIA-MS and searched against a custom library of sample-specific peptides generated from pooled study plasma samples. Quality control samples were included throughout the data acquisition to account for both digestion and mass spectrometry performance (**Figure S1**).

There were no significant differences in age, sex or ethnicity between disease and control groups (**Table 1**). Majority of patients were older (median age 61 years in disease and 63.5 years in control), Caucasian and male. Differences in patient characteristics and collection techniques were also visualized in a two-dimensional space with unsupervised UMAP. Race, sex and age were randomly distributed between disease and control peptide clusters (**Figure S2A-D**). Control samples were collected exclusively in EDTA-K2 tubes and aortic disease samples were collected at random in either EDTA-K2, acid citrate dextrose or heparin tubes (**Figure S2E**). After batch correction, samples across all four sample processing batches were homogeneously distributed between disease and control peptide clusters (**Figure S2F**).

	Disease (N=150)	**Control (N=52)**	**p-value**
Age (yrs) median (IQR)	61 (52-70)	63.5 (58.5-68.75)	0.17
Sex count (%)	M: 90 (60%) F: 60 (40%)	M: 27 (52%) F: 25 (48%)	0.33
Ethnicity count (%)	Caucasian: 82 (54.7%) African American/Black: 44 (29.3%) Hispanic/Latino: 21 (14%) Asian: 1 (0.7%) Other: 2 (1.3%)	Caucasian: 27 (51.9%) African American/Black: 20 (38.5%) Hispanic/Latino: 4 (7.7%) Asian: 1 (1.9%) Other: 0 (0%)	0.41

Table 1. Baseline characteristics of disease and control patients. Continuous variables are reported as median (IQR) and categorical as counts (%).

Proteome analysis was conducted in a cohort of 202 subjects (75 Type B, 62 DTAA, 13 BDTAA, 52 controls; **Figure 1B**). After filtering for highest quality and most consistent protein identifications, a total of 1,468 peptides and 195 proteins were quantified across all 202 samples included in this study. Volcano plots depict the 853 peptides (**Figure 1C**) and 99 proteins (**Figure 1D**) that were significantly (BH-adjusted p-value < 0.01) expressed between pooled disease and control cohorts. Peptides (**Figure 1E, F**) and proteins (**Figure 1G, H**) dysregulated in aortic disease compared to control samples visually separated into two distinct groups using unsupervised UMAP and PLS-DA, except for one control sample that appeared to be more like the disease cluster (**Figure 1G**). Individual disease samples and CAD/healthy controls were distributed randomly among the disease and control clusters (**Figure S1G**).

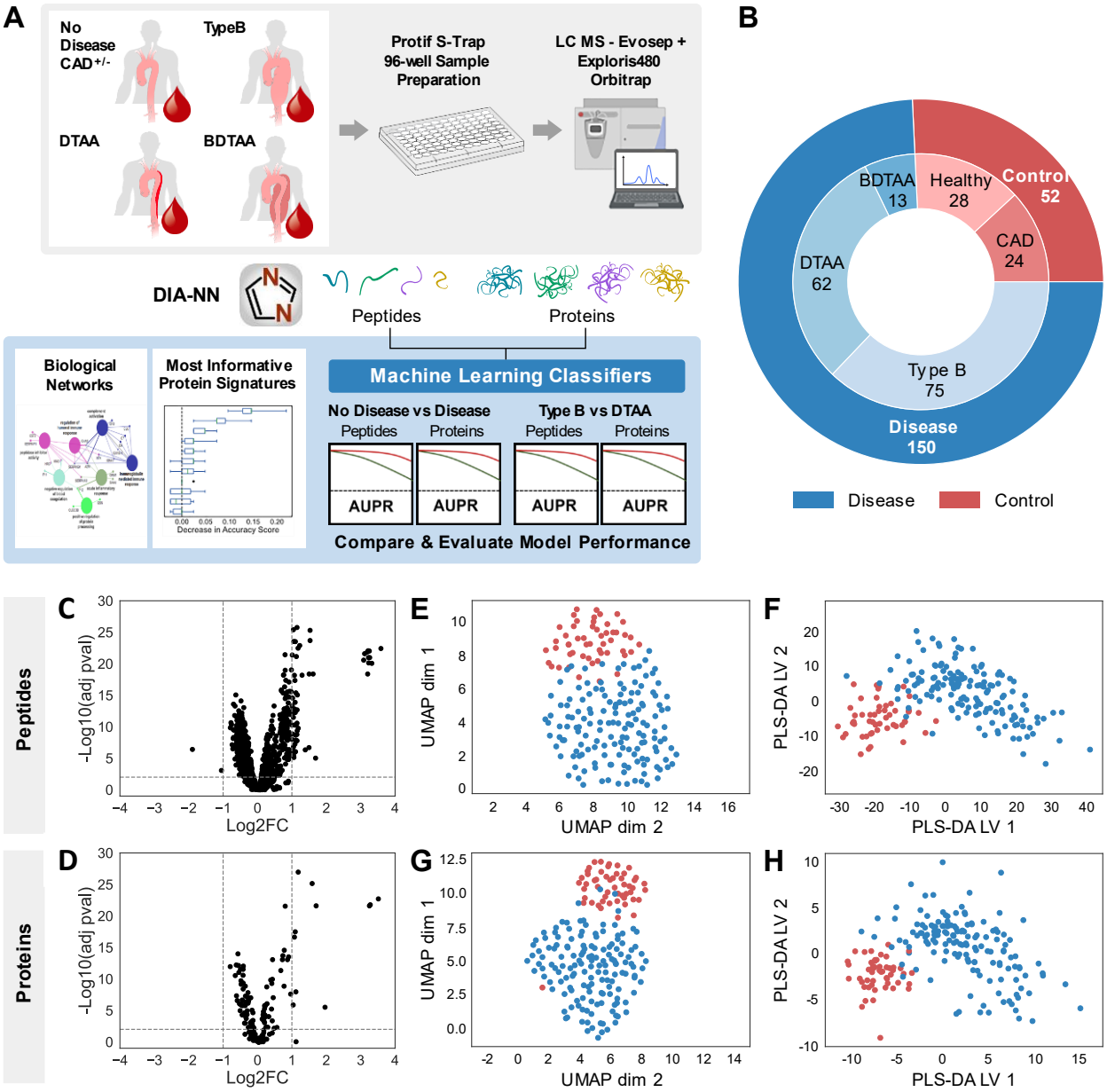

Figure 1. Study overview and plasma proteomics dysregulated in aortic disease compared to control samples. **(A)** Study overview cartoon. **(B)** Pie chart of patients in each group: 52 control patients (Healthy and CAD) and 150 aortic disease patients (BDTAA, DTAA, Type B). **(C)** Volcano plot of 1,468 peptides quantified across all samples; 853 peptides were differentially expressed (Benjamini-Hochberg adjusted p-value < 0.01) between disease and control using all samples. **(D)** Volcano plot of 195 proteins quantified across all samples; 99 proteins were differentially expressed between disease and control using all samples. **(E)** Unsupervised UMAP dimension reduction of all peptides showing separation between disease and control. **(F)** PLS-DA showing separation between disease and control. **(G)** Unsupervised UMAP of all proteins showing separation between disease and control. **(H)** PLS-DA showing separation between disease and control. In all UMAP, red dots are control, blue are cases of descending disease.

ML Models for Disease vs Control and Between-Disease Discrimination

Several experiments were first carried out to understand what models and dataset were best for this classification task. Various ML models were trained to classify disease versus control using four different input datasets: (1) all proteins, (2) all peptides, (3) significant proteins, or (4) significant peptides. To prevent leakage of data from the test set into the training set,[29] significant peptides or proteins were determined using only the 80% training data to do statistical testing. Input of proteins that were significantly different between all disease versus all control patients resulted in the highest precision-recall area under the curve (PR AUC 0.99) (**Figure 2A**). Significant peptides had the second-best performance (PR AUC 0.96), and all proteins had better performance (PR AUC 0.86) compared to all peptides (PR AUC 0.82). SVC was the highest F1-scoring model on significant peptides train data, GB for all peptides, and KNN for all/significant proteins. Test set precision, recall, and F1-scores for the optimized model from the training set were higher when using all proteins compared to all peptides on a per-class basis (**Figure 2B-C**).

A comparison of all tested models showed that any model does very well with this data; F1-score and MCC scores were overall high across all models using significant peptides and proteins, and the three non-tree-based models, SVC, KNN and LR, performed better than ET and RF (**Figure 2D-E**). For prediction of disease versus control using significant peptides with the best model SVC, the test set scores were: F1-score = 0.96, balanced accuracy = 0.98, and MCC = 0.94. Using the best model for significant proteins (KNN), test set scores were: F1-score was 0.96, balanced accuracy 0.98, and MCC 0.94. High performance metrics on the test set for all models signify the excellent ability to generalize well to unseen data.

There were no significantly different peptides or proteins between any of the three pairwise disease comparisons (Type B versus DTAA, BDTAA versus DTAA, and Type B versus BDTAA). All proteins and all peptides thus were used as inputs to the ML models that tried to classify these different diseases. Generalization performance of models trained using BDTAA samples was limited by this group's small sample size of 13 patients, with only three of those patients in the 20% test set. Thus, we focused on the Type B versus DTAA comparison, which had acceptable performance metrics on the test set. The use of all proteins led to a higher PR AUC (0.77) compared to peptides (**Figure 2F**) in distinguishing between diseases. The best models trained on all proteins and all peptides to predict Type B versus DTAA were RF and LR, respectively. Test MCC and F1-scores were overall lower for classification of this output compared to the disease versus control output and LR and RF had the highest F1-score and MCC on the test set when given all proteins as input, signifying these models' good generalizability to unseen data (**Figure 2G**).

Proteomic Marker Identification using Permutation Importance

Disease vs Control

A main goal of this study was to develop models that allow identification of markers that can be measured in the clinical setting to differentiate between those with and without thoracic aortic disease as well as between dissection and aneurysm aortic diseases.

When ML models are trained from correlated features, they may learn to rely on one arbitrary representative of the group of correlated features to make predictions. This is especially true of tree-based models. To avoid losing information about correlated features that are dropped by the model as unimportant during training, a second ML strategy was used before model interpretation to determine feature importance. Using the 80% train set, correlation between protein features was assessed using a

heatmap of Spearman rank-order correlation coefficients (**Figure 3A**), and then hierarchical clustering of proteins based on linkage distance was visualized (**Figure 3B**). A single protein was selected at random from each group of correlated proteins to be used for model training. This allowed tracing each protein selected by the model as 'important' back to the larger group of correlated proteins after we performed feature importance on the best model.

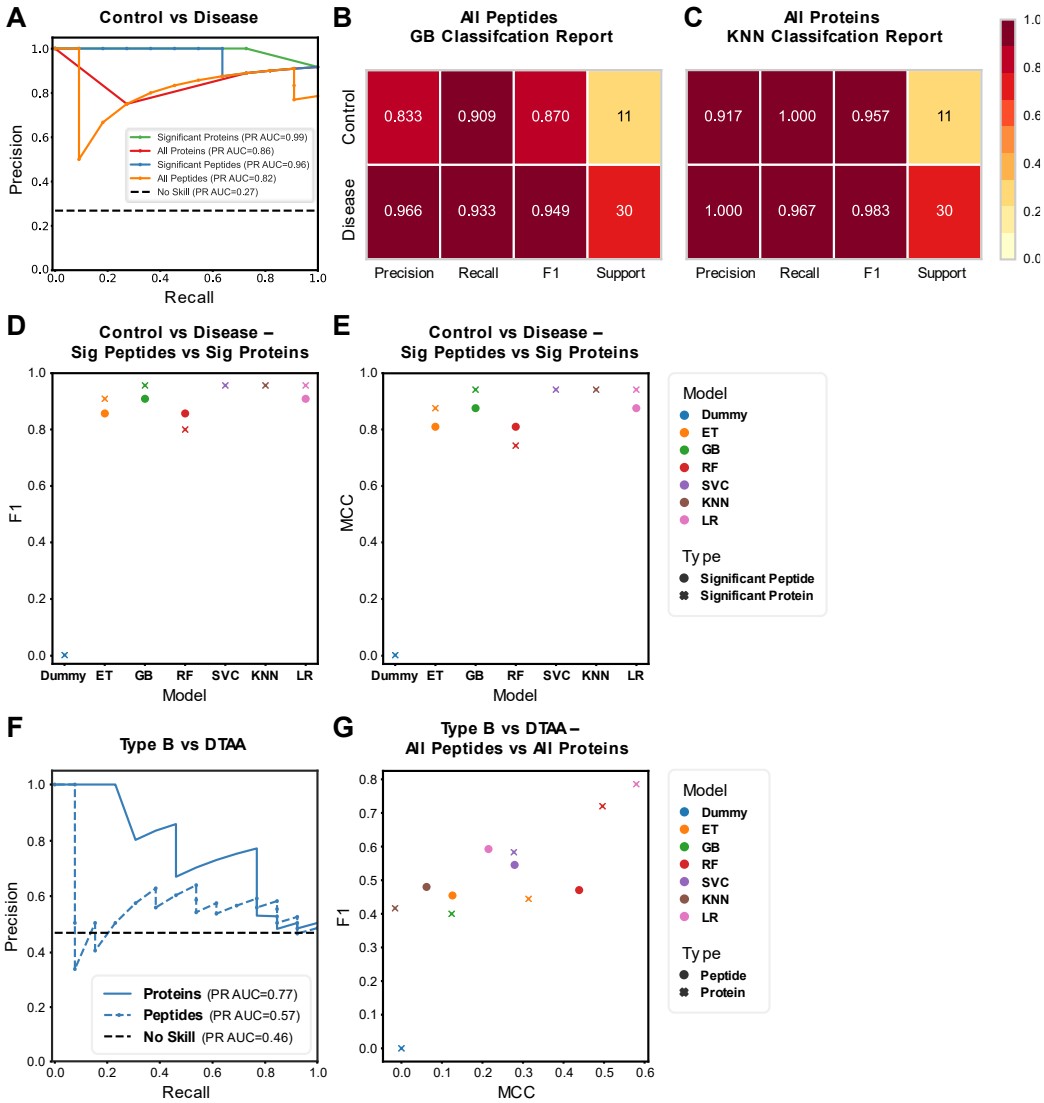

Figure 2. Machine learning (ML) approach for classification of all aortic disease patients versus all control patients and pairwise disease comparisons. **(A)** Precision-recall area under the curve (PR AUC) showing best model performances for each of four inputs (all peptides, all proteins, differentially expressed peptides, differentially expressed proteins) to predict disease vs control. Of the 1,468 peptides and 195 proteins, 792 peptides and 90 proteins were differentially expressed between disease versus control groups in the 80% training set. (B) Test set metrics for model with highest F1-score from training set (gradient boosting classifier [GB]) using all peptides input to predict disease vs control. **(C)** Test set metrics for model with highest F1-score from training set (k-nearest neighbors classifier [KNN]) using all proteins input to predict disease vs control. **(D)** Test set F1-scores across several optimized models predicting disease vs control using differentially expressed peptides (highest F1-scoring model from training set was SVC) or proteins (highest F1-scoring model from training set was KNN) as inputs. **(E)** Test set Matthew's correlation coefficient (MCC) scores across several optimized models predicting disease from control using differentially expressed peptides or proteins as inputs. **(F)** PR AUC curves for best models given all peptides vs all proteins to distinguish Type B vs DTAA. **(G)** F1 vs

Using the 80% training set of all proteins to predict disease versus control with various thresholds of linkage distance to filter features, several models performed well until reaching a linkage distance of 1.6, after which F1-scores declined (**Figure 3C**). At the cut-off of 1.6, there were seven correlated protein clusters from the original 195 proteins (**Figure 3B**). Using those seven proteins as input, SVC (C=100, gamma=0.01, probability=True) had the best F1-score of 0.89 on the train set out of the six models to predict disease vs control. The dummy model F1 remained at zero, indicating no predictive power. Performance of the SVC model was evaluated when given the 20% unseen test set with the seven selected proteins. SVC test metrics remained high (accuracy 0.88, F1-score 0.78, and PR AUC 0.90), which indicates the model is not significantly overfitting on training data. All models performed well when linkage distances were less than 1.6 (**Figure 3C**), so removing correlated features before training models was deemed unnecessary for prior assessments of ML performance in **Figure 2**.

We next sought to identify the most predictive proteins driving model performance in order to identify biomarker candidates and to better understand potential underlying biology. The permutation importance analysis (**Figure 3D**) shows that the sentinel protein P02042 (HBD) led to the largest increase in the model's prediction error after permuting this feature's value, followed by O75882 (attractin) and O00187 (mannan-binding lectin serine protease 2). HBD quantity was higher and attractin was lower in the disease group plasma (**Figure 3E**). Proteins with decrease in accuracy close to zero do not contribute to improved performance on unseen test data. Correlated proteins within each of the seven clusters and average decrease in accuracy score for each sentinel protein are listed in **Table S1**. Five proteins were present in the HBD cluster (i.e., HBD, hemoglobin subunit beta [HBB], hemoglobin subunit alpha [HBA], fibronectin [FN1], cytoplasmic actin 1 [ACTB]) and 47 were present in the attractin cluster.

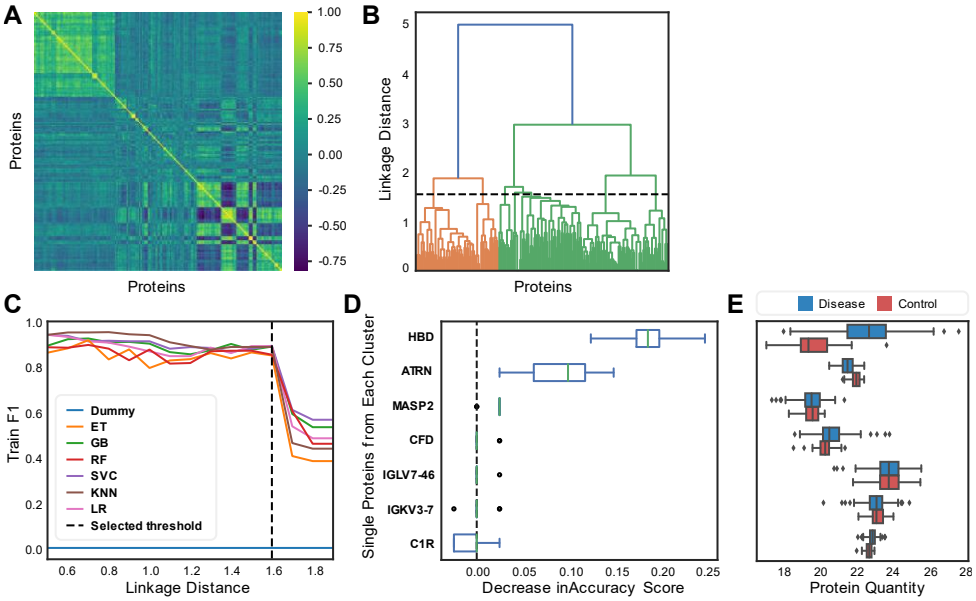

Figure 3. Machine learning (ML) approach for discovery of most predictive proteins between all aortic diseases compared to all control samples. **(A)** Heatmap of Spearman rank-order correlation coefficients for each pairing between 195 proteins (zero implies no correlation). **(B)** Hierarchical clustering of training data using Spearman rank-order correlations and Ward's linkage; a linkage distance of 1.6 yields 7 uncorrelated protein clusters. **(C)** Average F1-scores from 5-fold cross validation for each ML model with the training data across various linkage distance thresholds from 0.5 to 2. A single protein from each cluster was arbitrarily selected as input to the model at each threshold; a linkage distance of 1.6 was selected as the point where 7 single proteins kept from each cluster maintained good model performance. **(D)** Permutation importance plot calculated on test set showing the decrease in accuracy scores of 7 sentinel proteins. **(E)** Protein quantities for 7 sentinel proteins between disease and control groups.

To ensure capture of the most important proteins in discriminating disease versus control, these 52 proteins from the first to most important correlated groups from permutation importance were combined with the 52 proteins with the highest VIP scores in the PLS-DA model and the 52 proteins with the lowest BH-adjusted p-values. Thirty-nine unique proteins were present in the intersection of all three datasets and 64 proteins were present when the three datasets were combined and sorted for all unique proteins (**Figure 4A, Table S2**), which were then input to GO term enrichment analysis. Proteins involved in coagulation, protein-lipid complex remodeling, acute inflammatory response, cellular oxidant detoxification, complement activation and peptidase regulator activity were significantly enriched in the plasma of aortic disease patients compared to controls (**Figure 4B-C**). Log_2 fold change values for each of the 64 proteins input to ClueGO are listed in **Table S2.**

Type B vs DTAA

A heatmap of Spearman correlation coefficients between 195 proteins for the Type B and DTAA samples also revealed presence of correlated proteins (**Figure S3A).** Hierarchical clustering of proteins was performed and average training F1-scores across linkage distances from 0 to 2 were visualized for various ML models (**Figure S3B-C).** There was minimal variation in training F1-scores across number of features input to the models, however at a threshold of 0.8, GB peaked in F1-score. Forty-six protein clusters were present at this threshold and GB was the best F1-scoring model (train F1-score 0.63). **Figure S3D** shows good generalizability of the optimized GB model (learning rate=1, max depth=5, max features=log2, min samples leaf=20, min samples split=20, n estimators=10) tested with the 46 protein inputs, although modest performance metrics indicate potential overfitting on the train set and need for larger cohort sizes for future studies (accuracy 0.75, F1-score 0.70, and PR AUC 0.81). Permutation importance analysis (**Figure S3E**) shows the most important ten sentinel proteins from the 46 clusters and boxplots of the protein quantities for each of the 10 proteins are exhibited in **Figure S2F.** Correlated proteins within each of the 46 clusters are listed in **Table S3**.

There were 30 proteins within the top 10 clusters, which were compared to the proteins with the highest VIP scores in the PLS-DA model. No proteins were significantly different between Type B and DTAA. Five unique proteins were present in the intersection of the two datasets and 55 proteins were present when the two datasets were combined and sorted for all unique proteins (**Figure S4A, Table S4**), which were then input to GO term enrichment analysis. Proteins in the following pathways were significantly enriched in the plasma of Type B dissection compared to DTAA patients: peptidase inhibition, complement activation, blood coagulation, protein processing and acute inflammatory response (**Figure S4B-C).** Log_2 fold change values for each of the 55 proteins input to ClueGO are listed in **Table S4.**

Discussion

These data represent the most comprehensive analysis we are aware of describing the circulating proteome from patients with descending aortic disease to date. ML with protein-level data generated the most predictive models, but the performance gain was modest relative to peptide-level modeling. As the process of identifying informative features for eventual biomarker panel production is arguably more intuitive at the protein level and proteins appeared highly correlated, we segregated proteins into similar clusters, and used permutation-based importance ranking to identify correlated protein groups that were informative for separating patients between any disease and control or DTAA and Type B conditions. We found that we can train a model with only seven correlated proteins to achieve comparable performance to a model trained with all 90 differentially expressed proteins in classifying disease from control. Clusters important in distinguishing any descending disease from control patients

included free hemoglobin proteins, a large cluster of proteins involved in inflammation, apolipoprotein structure, and coagulation, as well as a cluster of complement proteins.

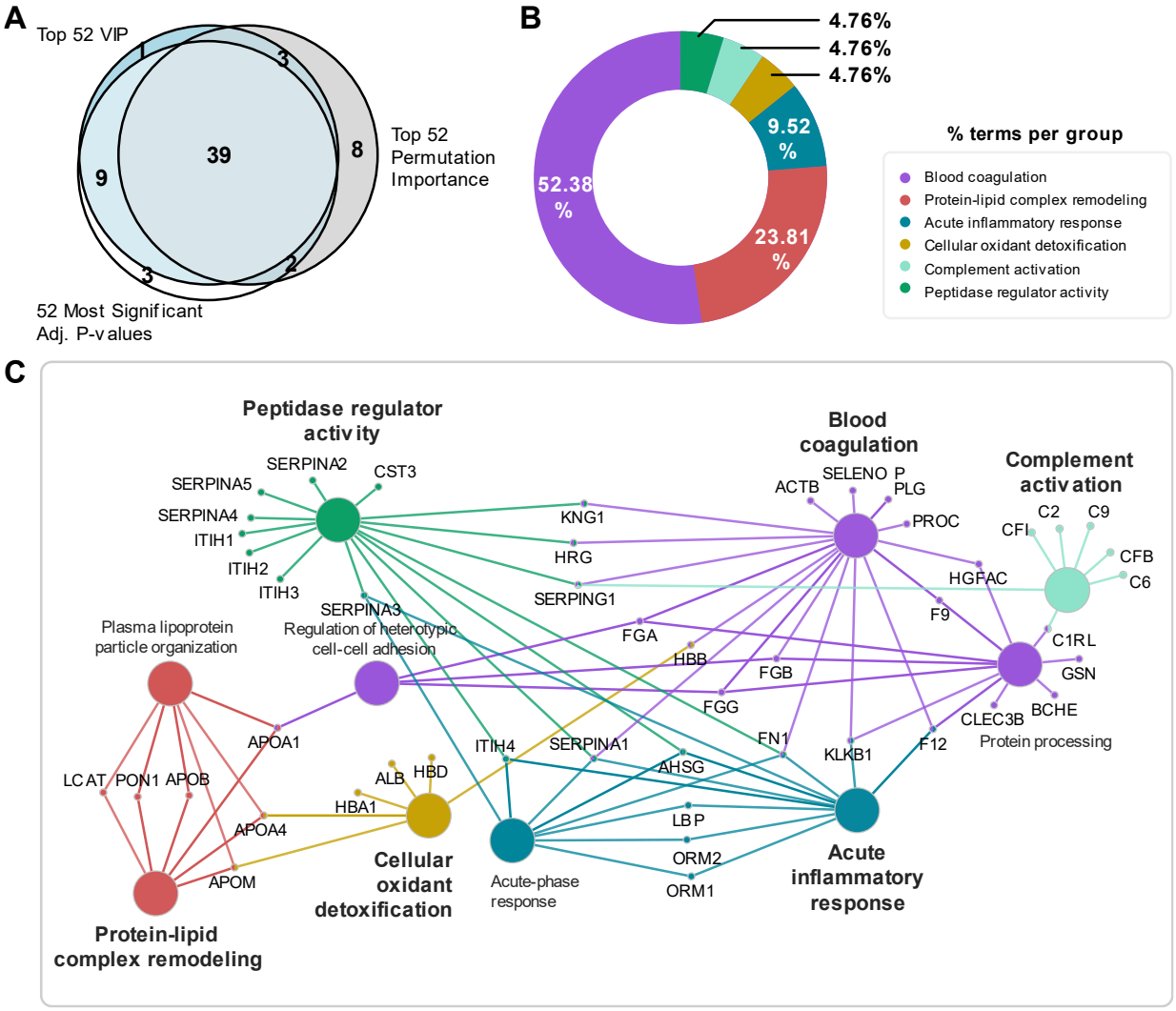

Figure 4. Biological term enrichment analysis of proteins altered between thoracic aortic disease vs control samples. **(A)** Venn diagram showing overlap among 52 proteins derived from top two proteins (HBD, ATRN) from permutation importance plot, 52 proteins with highest variable importance in projection (VIP) scores, and proteins with lowest 52 adjusted p-values. Sixty-four unique proteins were present when all three lists of proteins were combined. The 64 proteins were submitted to ClueGO for network analysis. **(B)** Pie chart of biological pathways enriched in aortic disease vs control. **(C)** Enriched biological pathways with co-expressed proteins dysregulated in aortic disease vs control.

For distinguishing DTAA from Type B dissection, the clusters were smaller in protein numbers and overall lower individual impacts on total model performance, but top clusters included proteins involved in coagulation, complement, and inflammation. Importantly, while traditional statistical analysis identified many differentially expressed proteins (DEPs) between control and any descending disease, we were unable to detect DEPs between Type B and DTAA patients, yet the ML approaches still differentiated patients between these two groups to some extent (AUPR=0.77) using the test set data not used during model training. ML is likely picking up patterns across many measured proteins,

compared to statistical tests that ask if one protein in aggregate has a different mean value between the groups. This may be useful for complex diseases that are heterogeneous across individuals; ML models can learn multiple signatures leading to disease. The main downside of using ML is that it requires many samples, typically hundreds, compared to statistics, which can be performed with as few as three replicates per group. Thus, modeling larger proteomic datasets using more sophisticated and modern approaches may be a potent approach for gaining new insight into the power of proteomic signatures for predictive biomarker development.

While ML has been deployed already for modeling proteomic data to discover putative biomarker signatures, few studies to date have examined how use of peptide level data impact overall model performance. Since peptides are the actual unit of measurement in these shotgun proteomic experiments, there is a risk that some fidelity in detecting meaningful biological events could be lost in the process of aggregating peptide level data into representative protein abundances[39,40]. Specifically, information on truncated isoforms or post-translationally modified sites can be lost, or at least diluted when potential evidence of different proteoforms is aggregated together into a single protein level estimate. In this study, we examined predictive performance of our plasma proteomic data at both the peptide and protein levels using a panel of different ML approaches. Overall, we observed that protein level data provided somewhat higher predictive performance, but the improvement was incremental in our case relative to peptide-level results. In the disease versus control comparison, models trained with both peptides and proteins were highly effective; therefore, it would be hard to show a clear benefit with either dataset. The incremental benefit we observed may prove to be of even stronger magnitude and more adequately powered in large scale studies or in other disease systems, and thus our data provide an example for future studies supporting exploration of peptide-level proteomics data for biomarker discovery applications.

The three most highly differential proteins in the comparison of disease vs non-disease were all subunits of hemoglobin, with >4-fold higher abundance in plasma from patients with aortic disease. It was recently reported that these proteins are also significantly elevated among individuals with bicuspid aortic valve associated ascending aneurysms, and that their levels in plasma correlate positively with aortic diameter [41]. Presence of free hemoglobin has also been observed in the media of human abdominal aortic aneurysms and is involved in monocyte recruitment to regions of medial injury[42]. Thus, while the observation of these proteins could reflect a technical bias caused by, for instance, RBC lysis during processing, there is also biological evidence that altered detection in plasma could be a disease-related event and candidates for further study as a potential biomarker.

We also observed several proteins whose differential abundance was similar between plasma and our previously published proteomic analysis of tissue samples comparing descending disease and control [10]. Among these, the monocyte marker CD14 and fibrinogens were uniformly upregulated whereas the protease inhibitor SERPINA5 (PAI-3), transthyrein (TTR), and Apoliprotein A4 (APOA4) were all prominently down regulated in both plasma and tissue from descending aortic disease patients, though APOA4 was even more substantially downregulated in dissection relative to aneurysm plasma and tissue. Elevated fibrinogen levels in our study are consistent with another recent report[43] demonstrating increased fibrinogens in patients with late stage descending aneurysm and type B dissection. CD14 is a well-known monocyte marker[44] and its elevation likely reflects local (tissue) and global (plasma) inflammation and monocyte recruitment[42,45]. The reduction of SERPIN5A/PAI-3 in tissue and plasma from patients with descending disease is interesting, as this protein has also been linked to protection against migration and malignancy in some cancers[46]. Specifically, SERPIN5A has been shown to bind to fibronectin and reduce integrin signaling, thus inhibiting cell migration[46]. Its reduction in descending

aortic tissue could reflect an important mechanistic role, given emerging evidence for the involvement of mechano- and matrix-sensing in aneurysm and dissection pathogenesis[47]. Further study into the potential for a mechanistic role of this protein in modulating vascular biology during aneurysm and/or dissection pathogenesis is warranted based on these preliminary findings. Taken together, many of the proteins selected by the ML models as highly informative for discriminating disease from non-disease are supported by solid corroborating biological evidence and for some, prior identification as putative biomarkers for thoracic aortic disease, thus providing evidence for the validity of this approach for identifying informative plasma biomarker candidates.

The performance of the ML approach was somewhat reduced for differentiating the plasma proteomes of descending aneurysm and descending dissection patients, maybe due in part to smaller sample sizes between these two disease groups. Furthermore, plasma samples were collected prior to surgery and thus at very late-stage disease. By this time, many aneurysm patients may have very similar overall pro-inflammatory plasma proteome signatures relative to aortic dissection patients. That being said, in our previous work we were able to differentiate proteomic signatures of late-stage aneurysm versus dissection tissues from these patients[10]. Interestingly, for proteins seen in both our previous tissue comparison and this plasma proteomic analysis, the direction of abundance differences was consistent between tissue and plasma, though not statistically significant. One interesting standouts in the list of informative proteins differentiating aneurysm and dissection was SAA4 (elevated in dissection). Overall levels of circulating Serum Amyloid A were recently identified as a potential biomarker for acute ascending and type B aortic dissection[48]. While the prior mentioned study did not differentiate between SAA subtypes (e.g., SAA1, SAA2, or SAA4), this work generally supports the biological relevance of SAA4 protein as potentially important for distinguishing aneurysm from dissection in descending thoracic aortic disease.

This study is a preliminary effort to address a pressing need for informative biomarkers for descending thoracic disease, and while powerful and biologically plausible new hypotheses have been generated, there are some weaknesses to mention. It is likely that small sample sizes impacted discriminative power and performance of the ML classifier, and future studies that expand the numbers of patients are needed. Samples were collected at very late-stage disease, just prior to surgical intervention. While this can be helpful in distinguishing disease states at their most extreme, the highest translational and clinical impact will come from biomarkers that can detect and distinguish disease at very early stages of development. Samples from the control group were also collected under a separate study protocol at a different hospital, which could lead to technical artifacts between disease and control groups. Thus, future work is needed to determine the robustness of the selected candidate markers in additional patients at later disease stage and, importantly, then determine which putative biomarkers may be informative at detecting early-stage disease and predicting risk for severe outcomes.

Conclusions:

The data presented in this preliminary report provide a framework and preliminary protein signature from which ongoing efforts will be built and support the power of ML for identifying biomarker candidates and building discriminative models to distinguish between biological states within the context of descending thoracic aortic disease.

Supporting Information:

Table S1. 7 representative cluster proteins and their correlated proteins (xlsx)

Table S2. 64 proteins between disease vs control from top two permutation importance proteins (xlsx)

Table S3. 46 representative cluster proteins and their correlated proteins (xlsx)

Table S4. 55 proteins between Type B vs DTAA from top ten permutation importance proteins (xlsx)

Figure S1. Quality Control analysis of digestion and mass spectrometry reference plasma pools

Figure S2. UMAPs colored by demographics, collection technique and cohorts (svg)

Figure S3. Discovery of predictive proteins distinguishing Type B from DTAA using machine learning (svg)

Figure S4. Biological term enrichment for proteins altered between Type B and DTAA (svg)

Supplemental Table Legends:

Table S1: 7 representative cluster proteins and all of the correlated proteins within each cluster; the 7 representative proteins were input to models classifying disease vs control

Table S2: 64 unique proteins between disease vs control identified from combination of top 52 VIP scores/lowest 52 BH-adjusted p-values/52 proteins from top two permutation importance proteins. $+\log_2$FC is up in disease, $-\log_2$FC is up in control

Table S3: 46 representative cluster proteins and all of the correlated proteins within each cluster; the 46 representative proteins were input to models classifying Type B dissection vs DTAA

Table S4: 55 unique proteins between Type B dissection vs DTAA identified from combination of top 30 VIP scores/30 proteins from top ten permutation importance proteins. $+\log_2$FC is up in Type B, $-\log_2$FC is up in DTAA

Author contributions

Conceptualization, A.M., J.G.M., S.P., A.A., D.M., D.G; Methodology, A.M., J.G.M., S.P., S.K., M.A., D.G.; Software, A.M., J.G.M.; Formal Analysis, A.M., J.G.M., S.P.; Investigation, A.M., J.G.M., S.P., M.A., S.K.; Resources, J.G.M., S.P., Data Curation, A.M., J.G.M., S.P.; Writing – Original Draft. A.M., J.G.M., S.P.; Writing – Reviewing & Editing, A.M., J.G.M., S.P., S.K., M.A., D.G., D.B., K-Y.C., P.K.S., D.M., A.A.; Visualization, A.M., J.G.M., S.P.; Supervision, J.G.M., S.P.; Project Administration, A.M., J.G.M., S.P., A.A., D.M., D.B.; Funding Acquisition, J.G.M., S.P., A.A., D.M.

Acknowledgements

This work was supported by the Cedars-Sinai Leon Fine award for Translational Research (*S.P, A.A)* as well as grants from the NIH NIGMS (*R35 GM142502 J.G.M.)*, the NIH NHLBI *(R00HL128787, S.P.)* and *(1P01HL110869-01A1, D.M.). The authors thank* Dasom Hwang for help with graphic design and the Cedars-Sinai Proteomics and Metabolomics Core facility for support with Mass Spectrometry data acquisition.

References

(1) Isselbacher, E. M. Thoracic and Abdominal Aortic Aneurysms. *Circulation* **2005**, *111* (6), 816–828. https://doi.org/10.1161/01.CIR.0000154569.08857.7A.

(2) Hagan, P. G.; Nienaber, C. A.; Isselbacher, E. M.; Bruckman, D.; Karavite, D. J.; Russman, P. L.; Evangelista, A.; Fattori, R.; Suzuki, T.; Oh, J. K.; Moore, A. G.; Malouf, J. F.; Pape, L. A.; Gaca, C.; Sechtem, U.; Lenferink, S.; Deutsch, H. J.; Diedrichs, H.; Marcos y Robles, J.; Llovet, A.; Gilon, D.; Das, S. K.; Armstrong, W. F.; Deeb, G. M.; Eagle, K. A. The International Registry of Acute Aortic Dissection (IRAD): New Insights into an Old Disease. *JAMA* **2000**, *283* (7), 897–903. https://doi.org/10.1001/jama.283.7.897.

(3) Bickerstaff, L. K.; Pairolero, P. C.; Hollier, L. H.; Melton, L. J.; Van Peenen, H. J.; Cherry, K. J.; Joyce, J. W.; Lie, J. T. Thoracic Aortic Aneurysms: A Population-Based Study. *Surgery* **1982**, *92* (6), 1103–1108.

(4) Olsson, C.; Thelin, S.; Ståhle, E.; Ekbom, A.; Granath, F. Thoracic Aortic Aneurysm and Dissection: Increasing Prevalence and Improved Outcomes Reported in a Nationwide Population-Based Study of More than 14,000 Cases from 1987 to 2002. *Circulation* **2006**, *114* (24), 2611–2618. https://doi.org/10.1161/CIRCULATIONAHA.106.630400.

(5) Tsai, T. T.; Fattori, R.; Trimarchi, S.; Isselbacher, E.; Myrmel, T.; Evangelista, A.; Hutchison, S.; Sechtem, U.; Cooper, J. V.; Smith, D. E.; Pape, L.; Froehlich, J.; Raghupathy, A.; Januzzi, J. L.; Eagle, K. A.; Nienaber, C. A.; International Registry of Acute Aortic Dissection. Long-Term Survival in Patients Presenting with Type B Acute Aortic Dissection: Insights from the International Registry of Acute Aortic Dissection. *Circulation* **2006**, *114* (21), 2226–2231. https://doi.org/10.1161/CIRCULATIONAHA.106.622340.

(6) Balmforth, D.; Harky, A.; Adams, B.; Yap, J.; Shipolini, A.; Roberts, N.; Uppal, R.; Bashir, M. Is There a Role for Biomarkers in Thoracic Aortic Aneurysm Disease? *Gen. Thorac. Cardiovasc. Surg.* **2019**, *67* (1), 12–19. https://doi.org/10.1007/s11748-017-0855-0.

(7) Shalhub, S.; Dua, A.; Brooks, J. Biomarkers in Descending Thoracic Aortic Dissection. *Semin. Vasc. Surg.* **2014**, *27* (3–4), 196–199. https://doi.org/10.1053/j.semvascsurg.2015.01.001.

(8) Herrington, D. M.; Mao, C.; Parker, S. J.; Fu, Z.; Yu, G.; Chen, L.; Venkatraman, V.; Fu, Y.; Wang, Y.; Howard, T. D.; Jun, G.; Zhao, C. F.; Liu, Y.; Saylor, G.; Spivia, W. R.; Athas, G. B.; Troxclair, D.; Hixson, J. E.; Vander Heide, R. S.; Wang, Y.; Van Eyk, J. E. Proteomic Architecture of Human Coronary and Aortic Atherosclerosis. *Circulation* **2018**, *137* (25), 2741–2756. https://doi.org/10.1161/CIRCULATIONAHA.118.034365.

(9) Parker, S. J.; Chen, L.; Spivia, W.; Saylor, G.; Mao, C.; Venkatraman, V.; Holewinski, R. J.; Mastali, M.; Pandey, R.; Athas, G.; Yu, G.; Fu, Q.; Troxlair, D.; Vander Heide, R.; Herrington, D.; Van Eyk, J. E.; Wang, Y. Identification of Putative Early Atherosclerosis Biomarkers by Unsupervised Deconvolution of Heterogeneous Vascular Proteomes. *J. Proteome Res.* **2020**, *19* (7), 2794–2806. https://doi.org/10.1021/acs.jproteome.0c00118.

(10) Saddic, L.; Orosco, A.; Guo, D.; Milewicz, D. M.; Troxlair, D.; Heide, R. V.; Herrington, D.; Wang, Y.; Azizzadeh, A.; Parker, S. J. Proteomic Analysis of Descending Thoracic Aorta Identifies Unique and Universal Signatures of Aneurysm and Dissection. *JVS-Vasc. Sci.* **2022**, *3*, 85–181. https://doi.org/10.1016/j.jvssci.2022.01.001.

(11) Swan, A. L.; Mobasheri, A.; Allaway, D.; Liddell, S.; Bacardit, J. Application of Machine Learning to Proteomics Data: Classification and Biomarker Identification in Postgenomics Biology. *OMICS J. Integr. Biol.* **2013**, *17* (12), 595–610. https://doi.org/10.1089/omi.2013.0017.

(12) Messner, C. B.; Demichev, V.; Wendisch, D.; Michalick, L.; White, M.; Freiwald, A.; Textoris-Taube, K.; Vernardis, S. I.; Egger, A.-S.; Kreidl, M.; Ludwig, D.; Kilian, C.; Agostini, F.; Zelezniak, A.; Thibeault, C.; Pfeiffer, M.; Hippenstiel, S.; Hocke, A.; von Kalle, C.; Campbell, A.; Hayward, C.; Porteous, D. J.; Marioni, R. E.; Langenberg, C.; Lilley, K. S.; Kuebler, W. M.; Mülleder, M.; Drosten,

C.; Suttorp, N.; Witzenrath, M.; Kurth, F.; Sander, L. E.; Ralser, M. Ultra-High-Throughput Clinical Proteomics Reveals Classifiers of COVID-19 Infection. *Cell Syst.* **2020**, *11* (1), 11-24.e4. https://doi.org/10.1016/j.cels.2020.05.012.

(13) Shen, B.; Yi, X.; Sun, Y.; Bi, X.; Du, J.; Zhang, C.; Quan, S.; Zhang, F.; Sun, R.; Qian, L.; Ge, W.; Liu, W.; Liang, S.; Chen, H.; Zhang, Y.; Li, J.; Xu, J.; He, Z.; Chen, B.; Wang, J.; Yan, H.; Zheng, Y.; Wang, D.; Zhu, J.; Kong, Z.; Kang, Z.; Liang, X.; Ding, X.; Ruan, G.; Xiang, N.; Cai, X.; Gao, H.; Li, L.; Li, S.; Xiao, Q.; Lu, T.; Zhu, Y.; Liu, H.; Chen, H.; Guo, T. Proteomic and Metabolomic Characterization of COVID-19 Patient Sera. *Cell* **2020**, *182* (1), 59-72.e15. https://doi.org/10.1016/j.cell.2020.05.032.

(14) Shu, T.; Ning, W.; Wu, D.; Xu, J.; Han, Q.; Huang, M.; Zou, X.; Yang, Q.; Yuan, Y.; Bie, Y.; Pan, S.; Mu, J.; Han, Y.; Yang, X.; Zhou, H.; Li, R.; Ren, Y.; Chen, X.; Yao, S.; Qiu, Y.; Zhang, D.-Y.; Xue, Y.; Shang, Y.; Zhou, X. Plasma Proteomics Identify Biomarkers and Pathogenesis of COVID-19. *Immunity* **2020**, *53* (5), 1108-1122.e5. https://doi.org/10.1016/j.immuni.2020.10.008.

(15) Overmyer, K. A.; Shishkova, E.; Miller, I. J.; Balnis, J.; Bernstein, M. N.; Peters-Clarke, T. M.; Meyer, J. G.; Quan, Q.; Muehlbauer, L. K.; Trujillo, E. A.; He, Y.; Chopra, A.; Chieng, H. C.; Tiwari, A.; Judson, M. A.; Paulson, B.; Brademan, D. R.; Zhu, Y.; Serrano, L. R.; Linke, V.; Drake, L. A.; Adam, A. P.; Schwartz, B. S.; Singer, H. A.; Swanson, S.; Mosher, D. F.; Stewart, R.; Coon, J. J.; Jaitovich, A. Large-Scale Multi-Omic Analysis of COVID-19 Severity. *Cell Syst.* **2021**, *12* (1), 23-40.e7. https://doi.org/10.1016/j.cels.2020.10.003.

(16) Bauer, Y.; de Bernard, S.; Hickey, P.; Ballard, K.; Cruz, J.; Cornelisse, P.; Chadha-Boreham, H.; Distler, O.; Rosenberg, D.; Doelberg, M.; Roux, S.; Nayler, O.; Lawrie, A. Identifying Early Pulmonary Arterial Hypertension Biomarkers in Systemic Sclerosis: Machine Learning on Proteomics from the DETECT Cohort. *Eur. Respir. J.* **2021**, *57* (6), 2002591. https://doi.org/10.1183/13993003.02591-2020.

(17) Shi, Z.; Wen, B.; Gao, Q.; Zhang, B. Feature Selection Methods for Protein Biomarker Discovery from Proteomics or Multiomics Data. *Mol. Cell. Proteomics* **2021**, *20*, 100083. https://doi.org/10.1016/j.mcpro.2021.100083.

(18) Rozanski, A.; Gransar, H.; Shaw, L. J.; Kim, J.; Miranda-Peats, L.; Wong, N. D.; Rana, J. S.; Orakzai, R.; Hayes, S. W.; Friedman, J. D.; Thomson, L. E. J.; Polk, D.; Min, J.; Budoff, M. J.; Berman, D. S. Impact of Coronary Artery Calcium Scanning on Coronary Risk Factors and Downstream Testing the EISNER (Early Identification of Subclinical Atherosclerosis by Noninvasive Imaging Research) Prospective Randomized Trial. *J. Am. Coll. Cardiol.* **2011**, *57* (15), 1622–1632. https://doi.org/10.1016/j.jacc.2011.01.019.

(19) Yu, F.; Haynes, S. E.; Nesvizhskii, A. I. IonQuant Enables Accurate and Sensitive Label-Free Quantification With FDR-Controlled Match-Between-Runs. *Mol. Cell. Proteomics MCP* **2021**, *20*, 100077. https://doi.org/10.1016/j.mcpro.2021.100077.

(20) Tsou, C.-C.; Avtonomov, D.; Larsen, B.; Tucholska, M.; Choi, H.; Gingras, A.-C.; Nesvizhskii, A. I. DIA-Umpire: Comprehensive Computational Framework for Data-Independent Acquisition Proteomics. *Nat. Methods* **2015**, *12* (3), 258–264. https://doi.org/10.1038/nmeth.3255.

(21) Behdenna, A.; Haziza, J.; Azencott, C.-A.; Nordor, A. PyComBat, a Python Tool for Batch Effects Correction in High-Throughput Molecular Data Using Empirical Bayes Methods. bioRxiv March 18, 2020, p 2020.03.17.995431. https://doi.org/10.1101/2020.03.17.995431.

(22) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv September 17, 2020. https://doi.org/10.48550/arXiv.1802.03426.

(23) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D. Scikit-Learn: Machine Learning in Python. *Mach. Learn. PYTHON* 6.

(24) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.;

Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, İ.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17* (3), 261–272. https://doi.org/10.1038/s41592-019-0686-2.

(25) Waskom, M. L. Seaborn: Statistical Data Visualization. *J. Open Source Softw.* **2021**, *6* (60), 3021. https://doi.org/10.21105/joss.03021.

(26) Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9* (3), 90–95. https://doi.org/10.1109/MCSE.2007.55.

(27) *Plotly: Low-Code Data App Development*. https://plotly.com/ (accessed 2022-08-03).

(28) Seabold, S.; Perktold, J. Statsmodels: Econometric and Statistical Modeling with Python. **2010**, 5.

(29) Desaire, H. How (Not) to Generate a Highly Predictive Biomarker Panel Using Machine Learning. *J. Proteome Res.* **2022**, *21* (9), 2071–2074. https://doi.org/10.1021/acs.jproteome.2c00117.

(30) *sklearn.ensemble.GradientBoostingClassifier — scikit-learn 1.1.1 documentation*. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html (accessed 2022-08-03).

(31) *sklearn.svm.SVC*. scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.svm.SVC.html (accessed 2022-08-03).

(32) *sklearn.ensemble.RandomForestClassifier*. scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (accessed 2022-08-03).

(33) *sklearn.ensemble.ExtraTreesClassifier*. scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html (accessed 2022-08-03).

(34) *sklearn.linear_model.LogisticRegression*. scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.linear_model.LogisticRegression.html (accessed 2022-08-03).

(35) *sklearn.neighbors.KNeighborsClassifier*. scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html (accessed 2022-08-03).

(36) Bengfort, B.; Bilbro, R.; Johnson, P.; Billet, P.; Roman, P.; Deziel, P.; McIntyre, K.; Gray, L.; Ojeda, A.; Schmierer, E.; Morris, A.; Morrison, M. Yellowbrick v1.3, 2021. https://doi.org/10.5281/zenodo.4525724.

(37) Bindea, G.; Mlecnik, B.; Hackl, H.; Charoentong, P.; Tosolini, M.; Kirilovsky, A.; Fridman, W.-H.; Pagès, F.; Trajanoski, Z.; Galon, J. ClueGO: A Cytoscape Plug-in to Decipher Functionally Grouped Gene Ontology and Pathway Annotation Networks. *Bioinformatics* **2009**, *25* (8), 1091–1093. https://doi.org/10.1093/bioinformatics/btp101.

(38) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003**, *13* (11), 2498–2504. https://doi.org/10.1101/gr.1239303.

(39) Plubell, D. L.; Käll, L.; Webb-Robertson, B.-J.; Bramer, L. M.; Ives, A.; Kelleher, N. L.; Smith, L. M.; Montine, T. J.; Wu, C. C.; MacCoss, M. J. Putting Humpty Dumpty Back Together Again: What Does Protein Quantification Mean in Bottom-Up Proteomics? *J. Proteome Res.* **2022**, *21* (4), 891–898. https://doi.org/10.1021/acs.jproteome.1c00894.

(40) Dermit, M.; Peters-Clarke, T. M.; Shishkova, E.; Meyer, J. G. Peptide Correlation Analysis (PeCorA) Reveals Differential Proteoform Regulation. *J. Proteome Res.* **2021**, *20* (4), 1972–1980. https://doi.org/10.1021/acs.jproteome.0c00602.

(41) Harrison, O. J.; Cagampang, F.; Ohri, S. K.; Torrens, C.; Salhiyyah, K.; Modi, A.; Moorjani, N.; Whetton, A. D.; Townsend, P. A. Candidate Plasma Biomarkers for Predicting Ascending Aortic Aneurysm in Bicuspid Aortic Valve Disease. *J. Cardiothorac. Surg.* **2018**, *13* (1), 76. https://doi.org/10.1186/s13019-018-0762-1.
(42) Rubio-Navarro, A.; Amaro Villalobos, J. M.; Lindholt, J. S.; Buendía, I.; Egido, J.; Blanco-Colio, L. M.; Samaniego, R.; Meilhac, O.; Michel, J. B.; Martín-Ventura, J. L.; Moreno, J. A. Hemoglobin Induces Monocyte Recruitment and CD163-Macrophage Polarization in Abdominal Aortic Aneurysm. *Int. J. Cardiol.* **2015**, *201*, 66–78. https://doi.org/10.1016/j.ijcard.2015.08.053.
(43) Li, S.; Yang, J.; Dong, J.; Guo, R.; Chang, S.; Zhu, H.; Li, Z.; Zhou, J.; Jing, Z. Neutrophil to Lymphocyte Ratio and Fibrinogen Values in Predicting Patients with Type B Aortic Dissection. *Sci. Rep.* **2021**, *11* (1), 11366. https://doi.org/10.1038/s41598-021-90811-5.
(44) Zamani, F.; Zare Shahneh, F.; Aghebati-Maleki, L.; Baradaran, B. Induction of CD14 Expression and Differentiation to Monocytes or Mature Macrophages in Promyelocytic Cell Lines: New Approach. *Adv. Pharm. Bull.* **2013**, *3* (2), 329–332. https://doi.org/10.5681/apb.2013.053.
(45) Blomkalns, A. L.; Gavrila, D.; Thomas, M.; Neltner, B. S.; Blanco, V. M.; Benjamin, S. B.; McCormick, M. L.; Stoll, L. L.; Denning, G. M.; Collins, S. P.; Qin, Z.; Daugherty, A.; Cassis, L. A.; Thompson, R. W.; Weiss, R. M.; Lindower, P. D.; Pinney, S. M.; Chatterjee, T.; Weintraub, N. L. CD14 Directs Adventitial Macrophage Precursor Recruitment: Role in Early Abdominal Aortic Aneurysm Formation. *J. Am. Heart Assoc.* **2013**, *2* (2), e000065. https://doi.org/10.1161/JAHA.112.000065.
(46) Jing, Y.; Jia, D.; Wong, C.-M.; Oi-Lin Ng, I.; Zhang, Z.; Liu, L.; Wang, Q.; Zhao, F.; Li, J.; Yao, M.; Wu, X.; He, X. SERPINA5 Inhibits Tumor Cell Migration by Modulating the Fibronectin-Integrin B1 Signaling Pathway in Hepatocellular Carcinoma. *Mol. Oncol.* **2014**, *8* (2), 366–377. https://doi.org/10.1016/j.molonc.2013.12.003.
(47) Humphrey, J. D.; Schwartz, M. A.; Tellides, G.; Milewicz, D. M. Role of Mechanotransduction in Vascular Biology: Focus on Thoracic Aortic Aneurysms and Dissections. *Circ. Res.* **2015**, *116* (8), 1448–1461. https://doi.org/10.1161/CIRCRESAHA.114.304936.
(48) He, Y.; Ma, C.; Xing, J.; Wang, S.; Ji, C.; Han, Y.; Zhang, J. Serum Amyloid a Protein as a Potential Biomarker in Predicting Acute Onset and Association with In-Hospital Death in Acute Aortic Dissection. *BMC Cardiovasc. Disord.* **2019**, *19* (1), 282. https://doi.org/10.1186/s12872-019-1267-0.