Differentiation between Descending Thoracic Aortic Diseases using Machine Learning and Plasma Proteomic Signatures

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[3], 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] Smidt 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

Background: Descending thoracic aortic aneurysms and dissections can go undetected until severe and catastrophic, and few clinical indices exist to screen for aneurysms or predict risk of dissection. **Methods:** This study generated a plasma proteomic dataset from 75 patients with descending type B dissection (Type B) and 62 patients with descending thoracic aortic aneurysm (DTAA). Standard statistical approaches were compared to supervised machine learning (ML) algorithms to distinguish Type B from DTAA cases. Quantitatively similar proteins were clustered based on linkage distance from hierarchical clustering and ML models were trained with uncorrelated protein lists across various linkage distances with hyperparameter optimization using 5-fold cross validation. Permutation importance (PI) was used for ranking the most important predictor proteins of ML classification between disease states and the proteins among the top 10 PI protein groups were submitted for pathway analysis. **Results:** Of the 1,549 peptides and 198 proteins used in this study, no peptides and only one protein, hemopexin (HPX), were significantly different at an adjusted p-value <0.01 between Type B and DTAA cases. The highest performing model on the training set (Support Vector Classifier) and its corresponding linkage distance (0.5) were used for evaluation of the test set, yielding a precision-recall area under the curve of 0.7 to classify between Type B from DTAA cases. The five proteins with the highest PI scores were immunoglobulin heavy variable 6-1 (IGHV6-1), lecithin-cholesterol acyltransferase (LCAT), coagulation factor 12 (F12), HPX, and immunoglobulin heavy variable 4-4 (IGHV4-4). All proteins from the top 10 most important correlated groups generated the following significantly enriched pathways in the plasma of Type B versus DTAA patients: complement activation, humoral immune response, and blood coagulation. **Conclusions:** We conclude that ML may be useful in differentiating the plasma proteome of highly similar disease states that would otherwise not be distinguishable using statistics, and, in such cases, ML may enable prioritizing important proteins for model prediction.

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 (DTAA) and Stanford type B dissections (Type B). Descending TAAD, also termed Type B dissection 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. 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. 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 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. These studies have all focused on biomarkers to distinguish aneurysm and/or dissection from normal and/or cases of acute coronary distress not caused by aneurysm or dissection. Biomarkers that can distinguish aneurysm from dissection may also be of clinical interest, as these molecules could aid in therapeutic decisions regarding timing of surgical intervention as aneurysmal tissue progresses toward increasing likelihood for dissection and degeneration. Our recent proteomic analysis of aneurysmal versus dissected descending thoracic aortic tissue found numerous proteins that were differentially expressed between the two groups[8], however it is unclear whether any of these tissue-derived proteins would be altered in the circulation and indicative of the two disease states.

Machine learning (ML) is a promising tool for automated classification of groups from proteomics data[9]. ML can take any collection of input data and estimate a mathematical function that predicts a categorical outcome, such as the presence or absence of a disease, or a continuous measure like age. Unlike statistical models, which allow for a quantitative measure of confidence for a relationship, ML can find patterns in unwieldy data with nonlinear interactions[10]. ML is also helpful when there are more input variables than number of subjects[10]. Thirdly, ML model interpretation methods can be used to reveal which proteins are relevant for differentiating between two similar diseases[11]. The application of ML to healthcare 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 model prediction, and prior work suggests that clustering of similar proteins as a feature selection technique before applying ML methods enables accurate disease classification[11]. These successes in the application of ML have led us to explore whether ML models given inputs of plasma proteins could distinguish between aneurysm and dissection.

The goal of the current work was to leverage ML to distinguish the plasma proteomes from two similar diseases not otherwise distinguishable using a standard statistical approach and provide preliminary insight into mass spectrometry detectable, circulating proteomic signatures capable of separating them. Toward this goal, we profiled the plasma proteomes of individuals with DTAA and Type B and applied ML strategies to identify protein features best able to discriminate between them.

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 an isolated diagnosis of DTAA or Type B dissection 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. All patients included in this study provided informed consent and their recruitment and participation was approved by the institutional review boards of UT Houston.

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[17]. For DIA mode, DIA-Umpire was first used to extract pseudospectra[18]. 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 as follows: 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 or zero value were removed from further analysis, reducing the number of peptides from 8,243 to 1,549 and proteins from 357 to 238 quantified across all samples. Forty proteins with multiple Uniprot identifiers in their group were removed, resulting in 198 proteins measured across all samples. Only patients with complete demographic and sample collection data (i.e., age, sex, race) were carried forward for analysis and four patients were removed who had duplicate rows between both peptide and protein datasets, resulting in 137 patients. Peptide and protein quantities were $\log_2$ transformed across each sample and corrected for batch effects using pyComBat[19].

Statistical Analysis

Data cleaning, analysis, and model training were performed in Python version 3.7.11 (SciKit-Learn[20], SciPy[21], seaborn[22], Matplotlib[23], Plotly[24], and Statsmodels[25]). Volcano plots were used to visualize the presence of any differentially expressed proteins and peptides between diseases. Log_2 fold changes (FC) 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. P-values were determined using independent two-sample t-tests with Benjamini-Hochberg (BH) multiple hypothesis testing correction. 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. BH adjusted p-values <0.01 were considered statistically significant. To avoid test data leak into train data when performing feature selection[26], 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.

Machine Learning and Feature Importance

To account for the presence of correlated proteins within our dataset, we grouped quantitatively similar proteins using hierarchical clustering analysis before performing model interpretation to generate a list of uncorrelated protein groups ranked by their level of importance when classifying between disease states. Using only the 80% train set to avoid biasing the feature selection method with test data[26], 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 a linkage distance threshold represents the number of clusters at that distance. To identify the linkage distance threshold corresponding to optimal model performance, we trained different ML models with a single, representative protein from each cluster at various distance thresholds. The six supervised ML classification algorithms used were Gradient Boosting Decision Trees (GB)[27], Support Vector Classification (SVC)[28], Random Forest (RF)[29], Extra-Trees (ET)[30], Logistic Regression (LR)[31], and K-nearest neighbors (KNN)[32]. 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, were chosen to represent generalization performance of the test set. The model with the highest F1-score on the training set and the number of features at the corresponding linkage distance were used for evaluation of the test set. F1-score, PR AUC and accuracy scores using this best model were reported for the test set.

Permutation importance (PI), or a decrease in accuracy score when a single feature's value is randomly shuffled, was calculated on the test set using the number of features determined at the optimal linkage distance. Each feature's value was randomly shuffled 10 times and a decrease in accuracy score was calculated each time. The mean of the 10 scores was calculated for each protein and the mean decrease in accuracy scores were ordered from high to low. Proteins with the largest mean decrease in accuracy score were most important to the model's predictions, and the top 10 proteins were visualized in a box and whisker plot.

Biological Pathway Analysis

All proteins among the top 10 PI protein groups at selected linkage distance were submitted for pathway analysis. GO Biological Process term enrichment analysis was performed using the ClueGO (version 2.5.9)[33] application within Cytoscape (version 3.9.1)[34]. 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 3-8. Enriched terms were then manually filtered to keep only non-redundant terms that connected all the proteins to the network.

Results

Figure 1A depicts the study workflow. Plasma samples were obtained from 137 individuals, of which 75 were diagnosed with isolated Type B and 62 were diagnosed with isolated DTAA. Plasma proteomes were generated using DIA-MS and searched against a custom library of sample-specific peptides generated from pooled study plasma samples. Six ML classification algorithms were then trained on the protein quantities from DIA-NN to predict Type B or DTAA disease and feature importance allowed for ranking of most important protein predictors of disease.

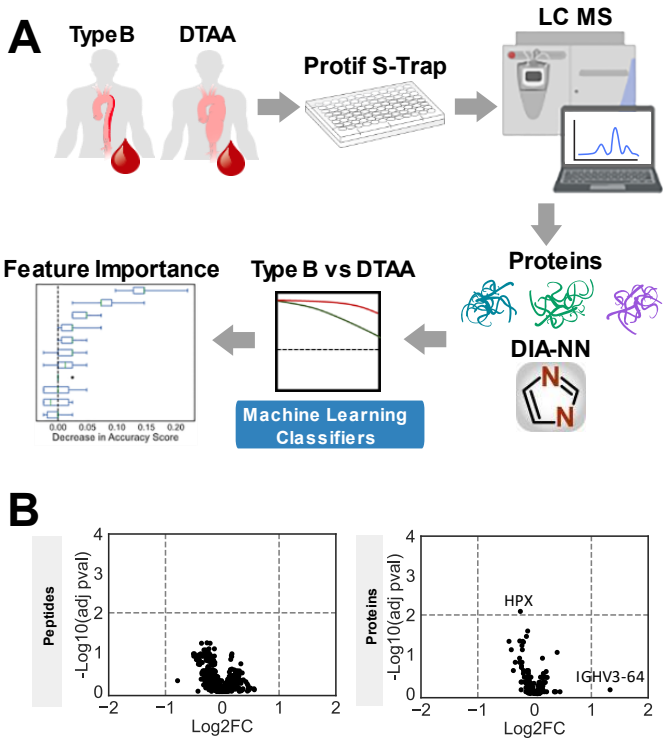

Figure 1. Study overview. **(A)** Proteomic dataset generation and analysis for classification between Type B and DTAA cases and identification of important protein predictors. **(B)** Left: Volcano plot of 1,549 peptides quantified across all samples; no peptides were differentially expressed between diseases. Right: Volcano plot of 198 proteins quantified across all samples; one protein was differentially expressed between diseases. A positive $\log_2$FC indicates protein mean is higher in Type B. Negative $\log_{10}$(adjusted p-value) > 2 corresponds to an adjusted p-value <0.01.

After filtering for the highest quality and most consistent protein identifications, a total of 198 proteins and 1,549 peptides were quantified across all 137 samples included in this study. Volcano plots of the negative $\log_{10}$ of the B-H adjusted p-value for all 198 proteins (left) and 1,549 peptides (right) between DTAA and Type B as a function of the $\log_2$FC of the mean quantities between each group are plotted in **Figure 1B**. A positive fold change indicates the mean protein quantity was higher in Type B relative to DTAA. There were no peptides and only one protein (hemopexin, HPX; p=0.008) that were significantly different between groups. One protein additionally had a $\log_2$ FC greater than one, immunoglobulin heavy variable-64 (IGHV3-64). Quality control samples were included throughout the data acquisition to account for both digestion and mass spectrometry performance (**Figure S1**).

Patient demographics were compared between groups. Age was significantly higher in the DTAA group while there were no significant differences in sex or ethnicity between Type B and DTAA (**Table 1**). Median age was 57 years in Type B and 66 years in DTAA (p=0.002). Sixty-one percent of patients with Type B and 58% of those with DTAA were male. The most prevalent ethnicity in each group was Caucasian, followed by African American/Black and Hispanic/Latino.

	Type B (N=75)	DTAA (N=62)	p-value
Age (yrs) median (IQR)	57 (51-67)	66 (58-74)	0.002
Sex count (%)	M: 46 (61%) F: 29 (39%)	M: 36 (58%) F: 26 (42%)	0.73
Ethnicity count (%)	Caucasian: 37 (49.3%) African American/Black: 24 (32%) Hispanic/Latino: 13 (17.3%) Asian: 1 (1.3%) Other: 0 (0%)	Caucasian: 41 (66.1%) African American/Black: 12 (19.4%) Hispanic/Latino: 7 (11.3%) Asian: 0 (0%) Other: 2 (3.2%)	0.07

Table 1. Baseline characteristics of patients with Type B or DTAA. Continuous variables are reported as median (IQR) and categorical as counts (%).

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 the correlated features within a group, a clustering strategy was used before ML model training and interpretation. Using the 80% train set, we visualized correlation between the 198 protein features using a heatmap of Spearman rank-order correlation coefficients. The heatmap indicates distinct groups of highly correlated proteins (**Figure 2A**). Hierarchical clustering of proteins was performed across linkage distances from zero to five; as linkage distance increases, there are fewer uncorrelated protein clusters **(Figure 2B)**. For example, at a linkage distance of four, all proteins are grouped into only two protein clusters. The number of protein clusters at each linkage distance is depicted in **Figure 2C**. 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. Average training F1-scores across linkage distances from zero to five were visualized for six ML models **(Figure 2D).** Training F1-scores generally declined across models as linkage distance increased and there were fewer features input to the models. The highest F1-scoring model using the training set at any linkage distance was SVC with a score of 0.67 at a linkage distance of 0.5. One hundred eleven protein clusters were present at this threshold. **Figure 2E** shows generalizability of the optimized SVC model on the test set (accuracy 0.74, F1-score 0.67, and PR AUC 0.69). Optimal hyperparameters for the SVC model were C=0.1, gamma=1, kernel=poly, probability=True, random_state=42 when tested with the 111 representative proteins from each cluster. PI scores were then calculated for these 111 proteins. **Figure 2F** shows box and whisker plots for the decrease in accuracy score across the 10 permutations for the top 10 most important sentinel proteins sorted by their mean decrease in accuracy score. The mean decrease in accuracy score and representative protein for each cluster are listed in **Table S1**.

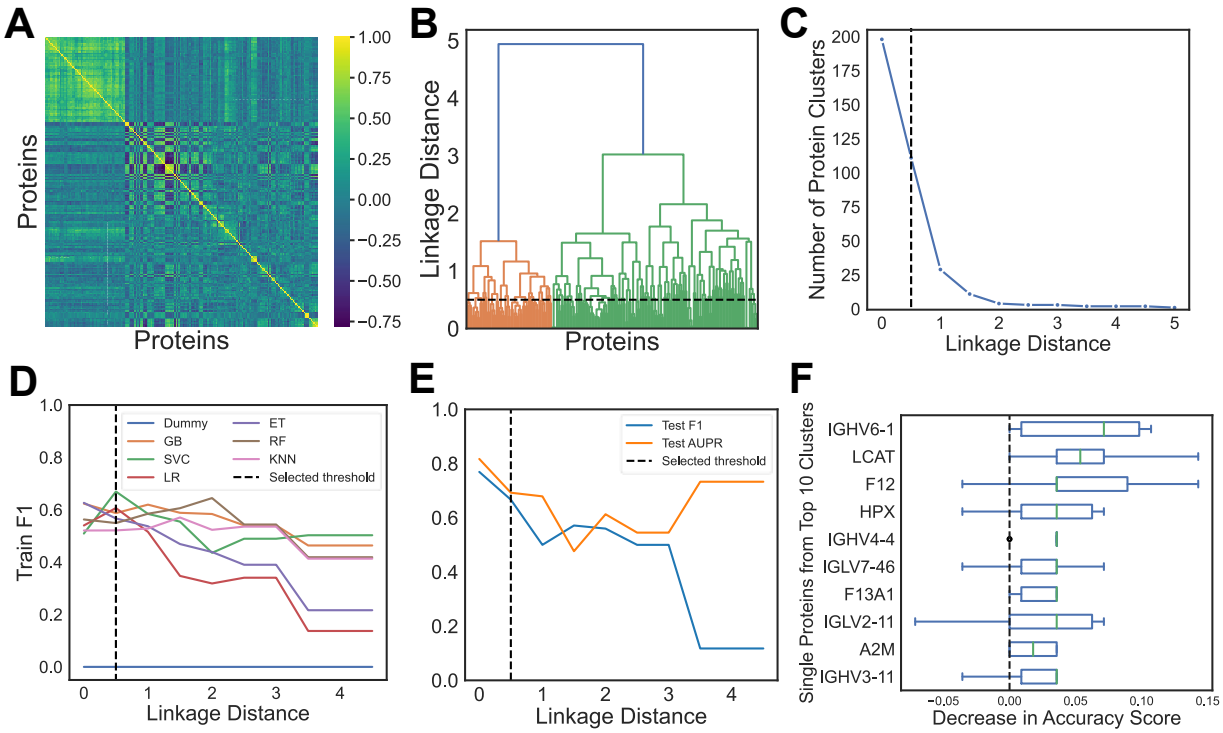

Figure 2. Machine learning (ML) approach for discovery of most predictive proteins between Type B dissection compared to DTAA samples. **(A)** Heatmap of Spearman rank-order correlation coefficients for each pairing between 198 proteins (zero implies no correlation). **(B)** Hierarchical clustering on Spearman rank-order correlations using Ward's linkage; a linkage distance of 0.5 yields 111 correlated protein clusters. **(C)** Number of uncorrelated protein clusters at each linkage distance. **(D)** Average F1-scores from 5-fold cross validation for each ML model with the training data across linkage distance thresholds from 0 to 5. A single protein from the clusters at each threshold was arbitrarily selected as input to the models; SVC at a linkage distance of 0.5 had the highest F1-score on the train set compared to all other models. At this threshold, there were 111 protein clusters. **E**) F1-score and PR AUC for test set across various thresholds for SVC showing good performance at 0.5 linkage distance threshold (SVC test F1-score 0.67, PR AUC 0.69). (**F**) Box and whisker plots of the distribution of PI scores for the top 10 sentinel proteins from the total 111 clusters. Each box has a line at the median and extends between the lower and upper quartiles of the PI distribution for that protein.

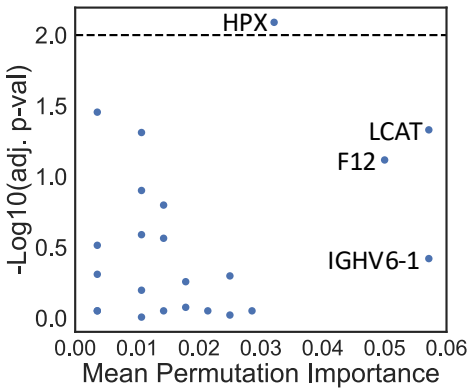

Figure 3. Comparison of -log10 (adjusted p-values) to mean PI for all proteins with a positive mean PI score.

Using these mean PI scores, we filtered for only the proteins with a positive score. The threshold to apply as a cut off to the PI score is objective[10], i.e., we chose values above zero, but a higher PI may have been selected to filter the number of proteins. There were 23 sentinel proteins with a positive PI score and a total of 38 proteins making up the 23 sentinel protein clusters. PI scores provide a ranked list of important proteins in prediction between the two diseases, versus statistics which selects a protein based on adjusted p-value cut-off.[10] To visualize the difference in level of information provided by each method, negative $\log_{10}$ of the adjusted p-values derived for each protein between the two groups was plotted as a function of the mean PI scores for these 23 proteins (**Figure 3**). While HPX is the only significantly different protein, there are proteins (LCAT, F12 and IGHV6-1) chosen by

the model as more 'important' than HPX for classification between diseases. There are also a number of proteins that have similar mean PI scores to HPX. Log_2FC, B-H adjusted p-values and mean PI scores for the top 10 protein clusters, mapping to 19 total proteins, depicted in **Figure 2F** are listed in **Table 2.** IGHV6-1 had the highest PI score, however it's quantity between groups was not significantly different (B-H adjusted p-value 0.38). A log_2FC of -0.17 indicates the mean quantity of IGHV6-1 was higher in the DTAA group compared to the Type B group. Mean quantities per group, log_2FC, negative log_{10} adjusted p-values and adjusted p-values for all 198 proteins are listed in **Table S2.**

Clustered Proteins	**Log_2FC**	**B-H adjusted p-value**	**Mean PI**
IGHV6-1	-0.17	0.38	0.057
LCAT HRG HGFAC PGLYRP2	-0.21	0.047	0.057
F12	-0.21	0.076	0.050
HPX	-0.25	0.0081	0.032
IGHV4-4 IGHV1-45 IGHV1-24 IGHV5-51 IGLV1-40	-0.055	0.89	0.029
IGLV7-46	-0.070	0.95	0.025
F13A1 F13B	-0.24	0.50	0.025
IGLV2-11	0.071	0.89	0.021
A2M FBLN1	-0.14	0.55	0.018
IGHV3-11	-0.10	0.84	0.018

Table 2. Log_2FC, adjusted p-values, and mean PI scores for 10 highest PI scoring clusters. Representative cluster protein is shown in bold for each cluster. Positive log_2FC indicates mean protein quantity is higher in Type B group. Negative log_2FC indicates mean protein quantity is higher in DTAA group.

These 19 proteins from the top 10 clusters shown in **Figure 2F** 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: complement activation, humoral immune response mediated by circulating immunoglobin, and blood coagulation/fibrin clot formation (**Figure 4).**

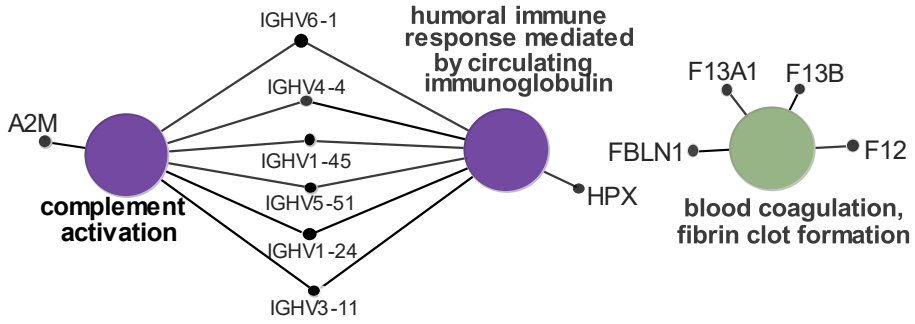

Figure 4. Biological term enrichment analysis of proteins altered between Type B dissection vs DTAA samples.

Discussion

These data represent the most comprehensive analysis we are aware of describing the circulating proteome from patients with descending aortic disease to date. 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 DTAA and Type B conditions. We were able to find only one differentially expressed protein between Type B and DTAA patients, yet the ML approaches still differentiated patients between these two groups to some extent (AUPR=0.7) using the test set data not used during model training. We found that we can reduce our protein data to a list of 111 uncorrelated proteins to train the highest performing model. Clusters important in distinguishing between the diseases included proteins involved in inflammation and coagulation.

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.

Many of the informative proteins selected by the ML model demonstrated similar trends for differential abundance in our previous proteomic analysis of tissue samples comparing Type B and DTAA [8]. Plasma HPX was both a ML model selected and significantly abundant protein between aneurysm and dissection cases in our study, and also demonstrated a trend toward increased abundance in aneurysm tissue relative to dissected tissue. Hemopexin is a heme scavenging protein considered to be generally protective against cardiovascular disease and atherosclerosis[35,36]. Similarly matched trends for abundance in both circulating plasma and tissue proteome of DTAA relative to Type B patients was observed for another heme scavenger, A2M, as well as proteins IGHV6-1, HRG, PGLYRP2 and F13A and B. Prominent involvement of immunoglobulins including IGHV6-1 is consistent with recent reports of a potentially pathogenic role for B cells and immunoglobulin deposition in abdominal aortic aneurysm (AAA)[37] , and suggests similar involvement in the descending thoracic aorta. Factor 13A and B are

fibrinolytic proteins with gene polymorphisms associated with AAA and hemolytic aneurysmal subarachnoid hemorrhage in the brain[38,39] . One other 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[40]. 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. Taken together, many of the proteins selected by the ML models as highly informative for discriminating diseases 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.

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. By this time, many aneurysm patients may have very similar overall pro-inflammatory plasma proteome signatures relative to aortic dissection patients. 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 and thus both predict adverse progression and provide theranostics to monitor effectiveness of pharmacological intervention. In addition, dissection absent a prior aneurysm may represent a very distinct pathogenic process for which late-stage aneurysm biomarkers cannot predict, and from which biomarkers of Type B dissection alone will not transfer to cases of dissection after significant aneurysm degeneration. 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.

Supplemental Figure Legend:

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

Supplemental Table Legends:

Table S1: 111 representative cluster proteins and all the correlated proteins within each cluster; the 111 representative proteins were used as inputs to the highest performing model classifying between disease. (xlsx)

Table S2: Mean protein quantities in each disease group, mean difference, $\log_2$(Type B/DTAA), negative $\log_{10}$ adjusted p-values and adjusted p-values for all 198 proteins. Positive $\log_2$(Type B/DTAA) indicates mean is higher in Type B group. Negative log10(adjusted p-value) > 2 corresponds to an adjusted p-value <0.01. (xlsx)

Declarations

Ethics approval and consent to participate

All patients included in this study provided informed consent and their recruitment and participation was approved by the institutional review boards of UT Houston.

Consent for publication

Not applicable.

Availability of data and materials

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.

Competing interests

The authors declare that they have no competing interests.

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.

Funding

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 the NIH NHLBI *(1P01HL110869, D.M.).*

Acknowledgements

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) 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.

(9) 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.

(10) Bzdok, D.; Altman, N.; Krzywinski, M. Statistics versus Machine Learning. *Nat. Methods* **2018**, *15* (4), 233–234. https://doi.org/10.1038/nmeth.4642.

(11) 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.

(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) 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.

(18) 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.

(19) 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.

(20) 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.

(21) 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.

(22) Waskom, M. L. Seaborn: Statistical Data Visualization. *J. Open Source Softw.* **2021**, *6* (60), 3021. https://doi.org/10.21105/joss.03021.
(23) Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9* (3), 90–95. https://doi.org/10.1109/MCSE.2007.55.
(24) *Plotly: Low-Code Data App Development*. https://plotly.com/ (accessed 2022-08-03).
(25) Seabold, S.; Perktold, J. Statsmodels: Econometric and Statistical Modeling with Python. **2010**, 5.
(26) 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.
(27) *sklearn.ensemble.GradientBoostingClassifier — scikit-learn 1.1.1 documentation*. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html (accessed 2022-08-03).
(28) *sklearn.svm.SVC*. scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.svm.SVC.html (accessed 2022-08-03).
(29) *sklearn.ensemble.RandomForestClassifier*. scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (accessed 2022-08-03).
(30) *sklearn.ensemble.ExtraTreesClassifier*. scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html (accessed 2022-08-03).
(31) *sklearn.linear_model.LogisticRegression*. scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.linear_model.LogisticRegression.html (accessed 2022-08-03).
(32) *sklearn.neighbors.KNeighborsClassifier*. scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html (accessed 2022-08-03).
(33) 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.
(34) 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.
(35) Mehta, N. U.; Reddy, S. T. Role of Hemoglobin/Heme Scavenger Protein Hemopexin in Atherosclerosis and Inflammatory Diseases: *Curr. Opin. Lipidol.* **2015**, *26* (5), 384–387. https://doi.org/10.1097/MOL.0000000000000208.
(36) Mehta, N. U.; Grijalva, V.; Hama, S.; Wagner, A.; Navab, M.; Fogelman, A. M.; Reddy, S. T. Apolipoprotein E $^{-/-}$ Mice Lacking Hemopexin Develop Increased Atherosclerosis via Mechanisms That Include Oxidative Stress and Altered Macrophage Function. *Arterioscler. Thromb. Vasc. Biol.* **2016**, *36* (6), 1152–1163. https://doi.org/10.1161/ATVBAHA.115.306991.
(37) Furusho, A.; Aoki, H.; Ohno-Urabe, S.; Nishihara, M.; Hirakata, S.; Nishida, N.; Ito, S.; Hayashi, M.; Imaizumi, T.; Hiromatsu, S.; Akashi, H.; Tanaka, H.; Fukumoto, Y.

Involvement of B Cells, Immunoglobulins, and Syk in the Pathogenesis of Abdominal Aortic Aneurysm. *J. Am. Heart Assoc.* **2018**, *7* (6), e007750. https://doi.org/10.1161/JAHA.117.007750.

(38) Macrae, F. L.; Evans, H. L.; Bridge, K. I.; Johnson, A.; Scott, D. J. A.; Ariëns, R. A. S. Common FXIII and Fibrinogen Polymorphisms in Abdominal Aortic Aneurysms. *PloS One* **2014**, *9* (11), e112407. https://doi.org/10.1371/journal.pone.0112407.

(39) Arati, S.; Chetan, G. K.; Sibin, M. K.; Bhat, D. I.; Vazhayil, V.; Narasingarao, K. V. L. Prognostic Significance of Factor XIIIA Promoter Methylation Status in Aneurysmal Subarachnoid Haemorrhage (aSAH). *BMC Cardiovasc. Disord.* **2019**, *19* (1), 170. https://doi.org/10.1186/s12872-019-1146-8.

(40) 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.