KTH ROYAL INSTITUTE
OF TECHNOLOGY

Degree Project in Biotechnology

First cycle 30 credits

Profiling the Blood Proteome in Autoimmune Disease Using Proximity Extension Assay

JULIA ASP

Stockholm, Sweden 2023

Author: Julia Asp, *juliaasp@kth.se*

2023-05-26

Examiner: Patrik Ståhl

Supervisor: Dr. Linn Fagerberg

Co-supervisor: Prof. Lars Klareskog

Karolinska Institutet, SciLifeLab

Abstract

Autoimmune diseases are complex, chronic, inflammatory conditions characterized by dysregulation of the immune system, resulting in inflammation and damage to various tissues, cells and organs. These diseases significantly impact individuals' quality of life and often contribute to increased mortality risk in the presence of comorbidities. However, due to the diverse array of symptoms associated with different autoimmune diseases, accurate diagnosis, prognosis, and treatment evaluation pose significant challenges. Thus, there is a pressing need for the discovery of novel biomarkers.

In this study, a comprehensive analysis of 944 plasma samples using the Olink® Explore platform was conducted, generating data on 1463 unique proteins. Based on the expression data, associated proteins were identified for six selected autoimmune diseases, namely multiple sclerosis, myositis, rheumatoid arthritis, systemic sclerosis, Sjögren's syndrome, and systemic lupus erythematosus, as well as some of their defined subgroups. These are prospective biomarkers and have the potential to aid in early diagnosis, therapeutic intervention, subgroup identification, disease differentiation, and disease prognosis. Notably, some of these proteins have not been previously associated with the specific diseases in the existing literature, especially not in plasma samples, thereby offering intriguing new perspectives for biomarker development. However, it is of great importance to conduct robust validation studies in independent cohorts to confirm the outcomes of this study.

In summary, our findings highlight the potential utility of these proteomic plasma biomarkers in improving the early detection, subgroup characterization, and disease differentiation of autoimmune diseases. The identification of these proteins will hopefully stimulate further investigation in the field of biomarker research and potential advancements in personalized medicine.

Keywords

Plasma proteomics, proximity extension assay, autoimmune disease, differential expression, machine learning, biomarkers.

Sammanfattning

Autoimmuna sjukdomar är en samling komplexa, kroniska, inflammatoriska sjukdomstillstånd som kännetecknas av dysreglering av immunsystemet, vilket resulterar i inflammation och skada av vävnader, celler och organ. Dessa sjukdomar har en betydande inverkan på individens livskvalitet och bidrar ofta till ökad dödsrisk där komorbiditeter föreligger. Emellertid medför den varierande symptombilden för olika autoimmuna sjukdomar betydande utmaningar för att uppnå noggrann diagnos, prognos och utvärdering av behandling. Det finns därför ett påtagligt behov av att upptäcka nya biomarkörer.

I denna studie utfördes en omfattande analys av 944 plasmaprover med hjälp av Olink® Explore-plattformen, vilket genererade data för 1463 unika proteiner. Baserat på uttrycksdata identifierades proteiner förknippade med de sex utvalda autoimmuna sjukdomarna multipel skleros, myosit, reumatoid artrit, systemisk skleros, Sjögrens sjukdom och systemisk lupus erythematosus samt några av deras definierade subgrupper. Dessa potentiella biomarkörer kommer eventuellt att underlätta tidig diagnos, sjukdomsdifferentiering och prognos. Flertalet av dessa proteiner har ännu aldrig kopplats till de här specifika sjukdomarna i litteraturen, särskilt inte från plasmaprover, vilket ger spännande nya perspektiv för biomarkörsutveckling. Det är dock av största vikt att genomföra robusta valideringsstudier i oberoende kohorter.

Sammanfattningsvis belyser våra resultat den potentiella brukbarheten hos dessa proteomiska plasmabiomarkörer för att förbättra tidig sjukdomsdetektering, karakterisering av subgrupper och sjukdomsdifferentiering att stimulera. Förhoppningsvis kan dessa resultat stimulera till vidare forskning inom området för biomarkörer och potentiella framsteg inom individbaserad medicin

Acknowledgements

First and foremost, I would like to express my sincerest gratitude to my supervisor Linn Fagerberg, my co-supervisor Lars Klareskog and the director of the Human Protein Atlas Mathias Uhlén for the opportunity to be a part of this project and learn from their incredible work and contributions to science. Their support, time, constructive criticism and insights have been invaluable and inspirational throughout the entire research process.

I am deeply thankful to Leonid Padyukov and María Bueno Álvez for the unwavering support, insightful feedback, discussions, shared resources, and patience throughout the entire project. Their dedication and expertise have been instrumental in shaping the direction and quality of this work.

I am also grateful to Josefin Kenrick, Emil Johansson, Richard Tjörnhammar and Emilio Skarwan for all the engaging discussions, support, collaborative efforts and passion for science that helped my project work and kept me motivated during the spring.

Last but not least, I would like to express my deepest gratitude to my friend Mario Reiser for the invaluable guidance he provided in the last stages of the project that immensely enriched the quality of this work.

Table of Contents

Introduction

This project is a collaborative effort between the Human Protein Atlas (HPA) and the Department of Rheumatology at the Karolinska Institute. In 2022, the HPA, launched the Human Disease Blood Atlas (HDBA), aiming to map out human plasma proteins associated with various diseases and facilitate biomarker discovery. The objective of this master's thesis project is to contribute to the HDBA by pursuing an exploratory analysis of the plasma proteome in patients with autoimmune disease, with the goal of identifying new potential biomarkers that could be contribute to early diagnosis, disease prediction, prognosis and evaluation of clinical treatments.

The definition of biomarkers varies, ranging from the European Medicines Agency's (EMA) description as *"A biological molecule found in blood, other body fluids, or tissues that can be used to follow body processes and disease in humans and animals"* to the broader definition by the U.S. Food and Drug Administration (FDA) as *"Characteristics that are objectively measured as indicators of health, disease, or a response to an exposure or intervention including therapeutic interventions."* (EMA, n.d.; FDA, 2022). An ideal biomarker should possess both diagnostic and prognostic value (Yang et al., 2022).

Unfortunately, there is a significant lack of clinical biomarkers for autoimmune diseases. Although autoantibodies can be useful in certain cases, they often fall short in terms of robustness and relevance. Therefore, the identification of new biomarkers is crucial for improving diagnosis, enabling early intervention, and effectively managing autoimmune diseases, ultimately enhancing patients' quality of life and reducing mortality risks. The discovery of biomarkers for autoimmune disorders is particularly significant for ensuring equitable patient care beyond the Western world, where resources may be limited and the current clinical techniques used for diagnosis are often expensive and inaccessible (Finckh et al., 2022).

Background

1.1 Autoimmune diseases

Autoimmune diseases are a complex group of disorders wherein the body's innate and adaptive immune system attacks its own healthy tissue (Wang et al., 2015). One of the most significant genetic associations in autoimmune disease lies within the major histocompatibility complex (MHC), which, in humans, generates the gene product known as human leukocyte antigens (HLA). These cell surface proteins aid in distinguishing between self and non-self (Nordquist & Jamil, 2022). Despite the strong association, the MHC has not demonstrated predictive power for clinicians, indicating that disease activation involves more complex mechanisms than a monogenetic mutation. Although the root genetic cause of most autoimmune disorders remains elusive, research suggests that disease symptoms frequently arise from inadequate clonal deletion, promoting proliferation of autoreactive T- and B-cells. Furthermore, studies have indicated that autoimmune disease initiation often requires environmental triggers, such as increased adipose tissue, reduced D-vitamin levels, smoking, infections with various microorganisms (e.g., the Epstein-Barr virus), toxins, insufficient nutrition, stress, microbiota, xenobiotics, and exposure to certain implants. Notably, the occurrence of autoimmune disease disproportionately affects women, even though the extent of this disparity varies depending on the disease and sometimes specific subgroups. The underlying cause of this uneven distribution among sexes within autoimmune diseases remains unknown (Wang et al., 2015).

The objective of this project is to analyze six selected autoimmune diseases, namely multiple sclerosis (MS) and five rheumatic disorders: myositis, rheumatoid arthritis (RA), scleroderma (Ssca), Sjögren's syndrome (SS) and systemic lupus erythematosus (SLE). Rheumatic disease is an umbrella term encompassing over 100 distinct autoimmune conditions that affect joints, muscles, tendons, ligaments and bones, with the most common symptom being joint pain (Sangha, 2000). These conditions often exhibit similar immunological, laboratory, and clinical manifestations, posing challenges in their classification and accurate diagnosis. Patients may also experience overlapping diseases, symptoms or even progression into another rheumatic disorder. For instance, Raynaud's phenomenon is a shared symptom among many rheumatic disorders as well as the presence of the autoantibodies to t-RNA synthetases in the sera of RA, Ssca and myositis patients. Antisynthetase-positive patients frequently present with interstitial lung disease (ILD) (Moutsopoulos, 2021).

Diagnosing autoimmune disease, particularly MS and rheumatic disorders, presents a difficulty due to high inter- and intraindividual heterogeneity as well as overlapping clinical manifestations. The absence of pathognomonic clinical and paraclinical features has prompted the development of diagnostic criteria. In the case of rheumatic disorders, these criteria are referred to as the European League Against Rheumatism (EULAR)/American College of Rheumatology (ACR) classification criteria. The EULAR/ACR criteria enable differentiation between healthy individuals and those with rheumatic diseases, also providing an indication of specific disease class. These criteria serve as valuable screening tools for early diagnosis; however, clinical tests are still necessary to confirm a diagnosis. Early diagnosis is crucial in all autoimmune diseases to prevent inflammatory episodes and mitigate permanent tissue damage (Singh et al., 2006).

1.1.1 Multiple Sclerosis

Multiple sclerosis is an incurable, chronic, inflammatory, demyelinating disease of the central nervous system (CNS) predominantly affecting women (75%) (Reich et al., 2018) and patients aged 20-40 years (Yamout et al., 2020). While the pathogenesis of MS shares certain traits with other non-CNS autoimmune diseases, the complete disease pathways remain unidentified. Researchers have observed the involvement of helper T-cells, cytotoxic T-cells and B-cells in MS lesions throughout the CNS. Another notable attribute of the disease is the activation of astrocytes and microglia in acute MS plaques, with the latter serving as a brain clean-up system for removing foreign or damaged substances. Immune response triggers in MS lead to chronic activation of microglia, resulting in white matter lesions (Reich et al., 2018).

The diagnosis of MS is established using the McDonald criteria, a universally accepted set of guidelines. These criteria describe clinical, imaging, and laboratory indications which collectively determine whether a patient is MS-positive or *probably* MS-positive and assign them to specific subgroups. These subgroups include clinically isolated syndrome (CIS), relapsing-remitting MS (RRMS), secondary-progressive MS (SPMS), primary-progressive MS (PPMS) and progressive-relapsing MS (PRMS). Each subgroup represents different disease progression and severities. Currently, there is no single laboratory test that can definitively diagnose MS. Magnetic resonance imaging (MRI) is the most effective method for observing lesions and disease progression, however lesions can coincide with other diseases and MRI accessibility is limited geographically and economically in many parts of the world (Thompson et al., 2018).

Early intervention with drugs that delay disease onset or progression has shown improved prognosis in MS (Simonsen et al., 2020). Individuals with MS have an 80% higher risk of mortality due to comorbidity compared to those without MS (Titcomb et al., 2022). The U.S. Food and Drug Administration (FDA) has approved immunomodulatory drugs for MS that target the immune system to counteract relapsing inflammatory episodes. These treatments dramatically improve survival rates and slow the progression of RRMS. One study even reported that 46% of patients showed *no evidence of disease activity* after one year, although only 8% of patients had sustained this condition after seven years. Rituximab has proven to be the most efficient therapy for MS which indicates and important contribution from B-cells in the disease mechanism. Currently, there is no established treatment for progressive MS (Baecher-Allan et al., 2018).

1.1.2 Myositis

Idiopathic inflammatory myopathies (IIM), also known as myositis, are a group of rare, heterogeneous, chronic, inflammatory disorders affecting the skeletal muscles. These conditions are characterized by muscle weakness and inflammation. In many cases, IIMs can involve multiple organs such as the joints, skin, lungs, heart and gastrointestinal tract, which can impact disease progression and potentially have lethal consequences (Lundberg et al., 2018). The onset of myositis is predominantly seen in females, occurring either in childhood or adulthood, with a mean age of onset at 57 years (Parker et al., 2022).

Diagnosis is determined with the 2017 *EULAR/ACR Classification Criteria for Adult and Juvenile IIMs and their Major Subgroups*. These criteria include the evaluation of various factors such as age of onset, muscle weakness, skin manifestations, dysphagia, paraclinical measurements of antisynthetase autoantibodies and serum levels of creatine kinase, lactate dehydrogenase or aspartate aminotransferase, muscle biopsy attributes and other clinical features. The most common subgroups of IIMs include dermatomyositis (DM), polymyositis (PM) and inclusion body myositis (IBM) (Lundberg et al., 2017). IBM is the only type of myositis that is more prevalent in men than in women.

Around 60% of IIM patients exhibit myositis-specific autoantibodies (MSAs), which are important for predicting organ involvement. Autoantibodies present in myositis and other diseases such as SLE, Ssca or Sjögren’s syndrome are instead referred to as myositis-associated autoantibodies (MAAs). Approximately 20-30% of IIM patients do not have any known autoantibodies and are classified as having seronegative IIM. Among the common MAAs is the antisynthetase autoantibody; patients displaying these antibodies are diagnosed with antisynthetase syndrome (ASyS). Current smoking is strongly associated with the development of ASyS (Lundberg et al., 2021), and in 90% of those patients, ILD is detected (Parker et al., 2022). ILD is also found in up to 78% of IIM patients. Cardiovascular disease, ILD and malignancies are the primary causes of death in individuals with IIMs, with ten-year survival rates ranging from 20% to 90%. Immune-mediated necrotizing myopathy (IMNM) is a rare and more aggressive type of myositis.

Early diagnosis and intervention are crucial for improving quality of life and mortality in myositis. However, no FDA-approved therapies currently exist due to the complexity of the disease and the lack of sufficient clinical evidence. Nonetheless, steroid hormones and immunosuppressants are commonly used in refractory cases of IIMS, although evidence for their efficacy is weak. Monoclonal antibodies, e.g., *Rituximab,* are being investigated as potential therapies for IIMs. (Lundberg et al., 2021).

1.1.3 Rheumatoid Arthritis

Rheumatoid arthritis is a chronic inflammatory disease affecting the joints. Its symptoms include joint swelling, tenderness, stiffness, and pain, leading to the destruction of synovial joints. Consequently, individuals with RA experience debilitation and increased risk of premature death (Aletaha et al., 2010). The disease progresses from an early stage characterized by early reactivity and limited adaptive immune responses to a systemic inflammation with elevated autoantibody levels and both innate and adaptive immune responses attacking tissues and causing permanent damage. An important distinction is that this pathological course primarily applies to seropositive RA, the most common form of the disease. RA can be divided into two major subgroups: seropositive and seronegative, based on the presence or absence of autoantibodies such as rheumatoid factor (RF) and anti-cyclic citrullinated peptide (anti-CCP or ACPA). Seropositive RA is defined by the presence of these autoantibodies and vice versa. Serological manifestations of seropositive RA can precede clinical manifestations by decades; however, the presence of autoantibodies does not always result in disease development (Deane & Holers, 2021). Seronegative RA differs from seropositive RA mainly in terms of prognosis (Aletaha et al., 2010). Most risk factors linked to seropositive RA, including smoking, heredity, toxins, and lifestyle choices, appear to have minimal to negligible effects on seronegative patients. Consequently, diagnosing and treating seronegative patients pose greater challenges (Pratt & Isaacs, 2014).

Comorbidities in RA are associated with higher mortality rates compared to the general population. For instance, individuals with RA face a 50% increased risk of cardiovascular mortality. While the disease is more common in women than in men across all age groups, the sex discrepancy is more pronounced among younger patients (Finckh et al., 2022). Disease onset typically occurs between the ages of 30 and 50 (Köhler et al., 2019). The most significant mortality risks in RA are disease progression into ILD and renal complications (Finckh et al., 2022).

Early introduction of treatment plays a crucial role in decreasing the accrual of joint damage, minimizing disability, and improving clinical outcomes (Aletaha et al., 2010). The implementation of a treat-to-target strategy and early intervention has been successful in Western populations over the past two decades, resulting in reduced disability, pain, disease progression and overall disease activity. Despite this, there is a pressing need for even earlier detection and initiation of antirheumatic therapy to achieve remission in RA and to improve outcomes outside of the Western population (Finckh et al., 2022). Acute inflammatory flare-ups are often managed with glucocorticoids, while disease-modifying antirheumatic drugs (DMARD), with methotrexate being the most common, are employed for long-term control of inflammation. The main objective is to alleviate pain, control inflammation and ultimately achieve remission (Köhler et al., 2019). In the case of seronegative RA, anti-tumor necrosis factor (TNF) therapy appears to have superior efficacy compared to seropositive RA (Pratt & Isaacs, 2014).

1.1.4 Scleroderma

Scleroderma, also referred to as systemic sclerosis, is a rare inflammatory disease of the connective tissue that leads to long-term fibrosis of the skin and internal organs. The clinical manifestations and prognosis of Ssca vary, ranging from milder forms characterized by Raynaud's phenomenon to aggressive cases with rapid disease progression and diffuse skin thickening due to progressive fibrosis of organs and tissues (Araújo et al., 2017). While Ssca is more prevalent in women, men tend to experience higher mortality rates and organ-related complications (Hughes et al., 2020). The majority of patients develop the disease between the ages of 40 and 50, although onset can occur earlier or later in life (Moinzadeh et al., 2020).

Diagnosis of Ssca is based on the 2013 EULAR/ACR classification criteria for systemic sclerosis. These criteria evaluate several indicators, including skin thickening, swollen fingers, Raynaud's syndrome, finger ulcers or scarring, enlarged capillaries, ILD and Ssca-specific autoantibodies for example the anti-centromere antibody (van den Hoogen et al., 2013). ILD is the leading cause of death in Ssca patients, with a prevalence of 30% and a 10-year mortality rate of up to 40% (Perelas et al., 2020). Treating Ssca poses challenges as it can affect multiple organ systems, requiring personalized medicine. Additionally, underlying diseases and therapeutic interventions can lead to complications. Therapeutic strategies include immunomodulatory treatments and vasodilators (Volkmann et al., 2022).

Mortality rates in Ssca vary widely, with patient survival after initial diagnosis ranging from 3 to 20 years. Poor survival rate is associated with more severe phenotypes displaying end-organ complications such as IDL and renal involvement (Moore & Steen, 2021). Scleroderma renal crisis affects 10-15% of patients and causes renal failure and severe hypertension. However, with intensive medical intervention, the incidence of death due to renal crisis has decreased in recent years (Chrabaszcz et al., 2020).

1.1.5 Sjögren's Syndrome

Sjögren's syndrome is a systemic autoimmune disease that manifests clinically as dry mouth and dry eyes, resulting from immune-mediated damage to the salivary and lacrimal glands. It is the second most prevalent systemic autoimmune disease, and in 15-30% of cases, SS overlaps with other autoimmune disorders (Hernández-Molina et al., 2015) leading to the classification of the disease into subgroups. Primary SS refers to cases where patients only exhibit Sjögren's-related symptoms, while secondary SS occurs when patients present both Sjögren's symptoms and symptoms of another rheumatic disorder (Shiboski et al., 2017). The disease predominantly affects women, with the mean age of onset being between 40 and 50 (Hernández-Molina et al., 2015). In most cases SS symptoms impact quality of life negatively and have a severe effect on oral health. In severe cases, particularly when primary SS progresses to secondary SS with additional complications such as ILD or renal involvement, the autoimmune disease can become life-threatening (Zhan et al., 2023).

The diagnosis of primary SS relies on the 2016 EULAR/ACR classification criteria, which include five objective items: inflammation of the labial salivary glands and duration, positive anti-Sjögren's-syndrome-related antigen A autoantibody (anti-SSA/Ro), ocular staining score, Schirmer's test, and unstimulated salivary flow rate (Shiboski et al., 2017). Currently, treatment options are limited to symptom relief as there is no official cure for the disease. Autoantibodies have been detected in up to 66% of patients before symptom onset, suggesting a potential role in disease development. Therefore, a molecular-based classification of the disease is strongly desired (Jonsson, 2022). The traditional therapeutic strategy for SS involves the use of antirheumatic drugs, i.e., glucocorticoids, which have a disease-modifying effect and help alleviate inflammation. Ongoing research focuses on novel therapies, including biotherapeutic approaches that target and block inflammation-related receptors or neutralize antibodies. Additionally, immune inhibitors, such as BAFF receptor blockers, are used to manage disease symptoms, even though they lack pharmacological effects (Zhan et al., 2023).

1.1.6 Systemic Lupus Erythematosus

SLE is one of the most complex systemic autoimmune diseases due to the considerable heterogeneity of the clinical manifestations of the disease. It predominantly affects women (90%) (Bernatsky et al., 2006) between the ages of 15 and 45. This inflammatory disease can target multiple organs, including the skin, heart, lungs, joints, and kidneys, leading to varying symptoms that may come and go over time (NIH, 2022). Symptoms can range from skin rashes and alopecia to arthritic and neuropsychiatric manifestations, likely resulting from complex immune dysregulation (Lazar & Kahlenberg, 2023). The protean nature of the clinical presentation often overlaps with other diseases, further complicating its identification, definition, and diagnosis. The EULAR/ACR classification criteria for SLE are used to establish a diagnosis, incorporating various weighted classification items such as fever, leukopenia, thrombocytopenia, psychosis, alopecia, oral ulcers, acute pericarditis, joint involvement, proteinuria and specific autoantibodies among others antiphospholipid antibodies, antinuclear antibody (ANA) and anti-Sm antibodies (Aringer et al., 2019).

SLE patients have a higher risk of mortality compared to the general population, with circulatory disease, renal disease, and malignancies, including hematological and lung cancers, being the main cause of death. Some studies indicate that male SLE patients have a higher mortality rate than female SLE patients (Bernatsky et al., 2006). On a positive note, the mortality rates associated with SLE significantly decreased, with the 15-year survival rate increasing from 50% in 1948 to 85-95% today (Dörner & Furie, 2019). Antimalarial therapy, particularly hydroxychloroquine (HCQ), is the most common treatment option for SLE. It aims to reduce risk of disease flares, improve life expectancy, decrease thrombosis risk, and minimize cutaneous and musculoskeletal manifestations. Glucocorticoids are also used to quickly control flare-ups, and antibody-based immunomodulatory treatments have recently gained FDA approval (Lazar & Kahlenberg, 2023).

1.2 The Plasma Proteome

The human plasma proteome is the intersection of proteomics, medicine and the diagnostic industry. Plasma, which constitutes 55% of blood, is defined as the liquid component of blood containing salts, enzymes, other proteins and water. The remaining 45% consists of leukocytes and erythrocytes. The plasma proteome is unique in its inclusion of proteins not only related to blood function but also subsets of proteins derived from all other tissues in the body. Analyzing the plasma proteome poses significant challenges due to the plethora of albumin and the heterogeneity in size and abundance, particularly of glycoproteins (Anderson & Anderson, 2002). However, advancements in techniques have made this task more feasible (Wik et al., 2021). The dynamic range of proteins in plasma is immense, with differences in abundance spanning up to a factor of 10^{10}. For instance, despite the substantial difference in abundance, albumin and interleukin 6 are commonly used as indicators of liver disease and inflammation or infection, respectively. Proteins found in plasma encompass secreted proteins from solid tissues, immunoglobulins, receptor ligands such as hormones and cytokines, temporary passenger proteins (e.g., lysosomal proteins), tissue leakage products resulting from damage or cell death, aberrant secretions from tumors or diseased tissue, and foreign proteins from infectious organisms, among others.

The exploration of the human plasma proteome has gained momentum in the field of biomarker discovery, driven by the concept of pathological protein leakage into the bloodstream. Extensive efforts have been dedicated to mapping the complete plasma proteome, aiming to identify all proteins above the limit of detection (LOD). Among the proteins of interest are tissue leakage proteins and interleukins, which hold pathological significance in various diseases and inflammatory episodes. In comparison to other bodily fluids such as cerebrospinal fluid (CSF), saliva, or urine, plasma samples offer advantages in terms of invasiveness during sampling, ease of separation, analysis and practicality. However, it is important to note that not all proteins from these other fluids leak or secrete into the bloodstream, at least not above the LOD.

Another challenge with plasma samples lies in the half-life of the proteins, from the point of first disease symptoms to their survival outside the body. Extended storage at -70 °C has been shown to effectively preserve the protein structures and activity. Factors such as drugs, medication, lifestyle choices and genetics significantly impact blood protein levels. The genetic noise alone can account for 12-95% of protein abundance variations, depending on the specific protein (Anderson & Anderson, 2002). Recent studies have confirmed the average variation per protein of approximately 61-62%. These studies integrate protein assays with genome-wide association studies (GWAS), with a specific focus on investigating and identifying protein quantitative trait loci (pQTLs). The identification of these pQTLs is expected to enhance our understanding of disease pathology and shed a light on proteins that are likely to exhibit higher prevalence in the plasma proteome (Kim et al., 2013).

1.3 Proximity Extension Assay

The Proximity Extension Assay (PEA) is a multiplex immunoassay, provided by Olink Proteomics AB, that enables detection of low abundance proteins in biological samples. The fundamental principles of this technology involve antibodies that are linked to DNA-encoded tags, which then locate and bind to target proteins. When two matched probes bind to the same protein, the oligonucleotides hybridize and create an amplicon aided by DNA polymerase. The target sequences are subsequently amplified, detected, and quantified through PCR, generating a relative quantification output per protein from different samples. In 2021, Olink released the Olink® Explore platform, with 1463 validated proteins, which combines the PEA technology with Next Generation Sequencing (NGS), specifically Illumina sequencing, to automate and enhance the capacity for high-throughput screenings.

When performing plasma protein analysis using PEA, sample sizes need to be in the range of a few microliters. Depending on the natural concentration of the target proteins in blood, the samples are diluted at four different dilution ratios (1:1, 1:10, 1:100, 1:1000). Subsequently, the samples are run in the four different dilution panels, each containing different proteins. This targeted approach enables a multiplex, high throughput process with high specificity, sensitivity, and minimal sample consumption (Wik et al., 2021). One technical challenge associated with PEA is the so-called *hooking effect,* which occurs when there is an excess of antigens relative to the antibody probes. This effect can falsely lower the intensity values detected, leading to the misconception of low protein abundance instead of very high abundance (Olink, 2016).

1.4 Linear Regression Model

Linear regression is a statistical modeling technique employed to analyze data, infer causality, and make quantitative predictions about future outcomes. The modeling can be achieved through simple linear regression or multiple linear regression, depending on the number of independent variables used to predict the dependent variable. The basic principle is to determine the coefficients for each independent variable that most accurately fit the data. These regression coefficients can be estimated with least squares, maximum likelihood, robust estimation, Bayesian approach or ridge regression, among many other methods. Linear regression is widely used in differential expression analysis due to its robust mathematical foundation, interpretability, and its role as a fundamental component in more advanced modeling techniques, in particular machine learning models (Su et al., 2012). An example of the incorporation of multiple linear regression models is the *limma* (Linear Models for Microarray and RNA-Seq Data) package in R.

1.4.1 Limma

BioConductor is an open-source R-based software, specifically developed for statistical genomics. Among the various software packages available in BioConductor, limma is a versatile tool designed for gene expression analysis of array-based experiments, including microarrays, protein arrays, and polymerase chain reaction (PCR). Additionally, it can be applied to RNA-seq data and other high-throughput omics datasets. To utilize the limma package, the data should be formatted as a matrix of expression values, where each row represents a feature (e.g., gene or protein) and each column corresponds to a sample. The statistical framework of limma is well-suited for large-scale differential expression studies and encompasses algorithms that facilitate information borrowing, quantitative weighting, variance modeling, and data pre-processing. Notably, it also incorporates robust statistical methods to handle datasets with small sample sizes.

Limma adopts a feature-wise linear modeling approach, enabling analysis of complex experimental designs and flexible hypothesis testing. Moreover, it incorporates global variability analysis across the entire dataset by estimating hyperparameters that capture correlations between features and samples, as well as variability in sample quality. Correlation between samples is assessed similarly to a random effects model. Empirical Bayes methods are employed to facilitate information borrowing between features, with the ability to incorporate mean-variance trends, particularly important for gene expression data at lower intensities or abundances. The inclusion of quantitative weights throughout the statistical analysis allows for the correction of different factors such as sex and age, enhancing the modeling for global characteristics of the data and improving statistical power and accuracy. These weights can be either preset based on external information or estimated from the data itself. By avoiding ad hoc decisions, the use of weights increases the power to detect differentially expressed genes, providing a more robust and model-based approach.

In limma, fitting the same linear model to each feature enables the borrowing of strength between features to moderate the residual variances. This approach leads to compromise between the feature-wise estimator and the global variability across all features, effectively increasing the effective degrees of freedom with which feature-wise variances are estimated. This is particularly advantageous in experiments with small sample sizes, as it enhances the reliability of statistical conclusions. Other statistical approaches in differential expression analysis often require splitting the data into smaller subsets, limiting the global analysis of the entire experiment. Furthermore, these techniques may lack the straightforwardness and reproducibility offered by limma. A distinctive feature of limma is its ability to handle variations in sample quality in a graduated manner through quantitative weights (Ritchie et al., 2015). Comparisons with other common non-linear, non-weighted approaches, such as DESeq2, have shown that limma can provide more precise results by identifying a greater number of differentially expressed genes (Tong, 2021). Overall, limma is highly regarded in the field of gene expression due to its statistical rigor, user-friendly interface, and its ability to handle high- throughput datasets. However, it should be noted that limma does not support multinomial regression modeling, which requires alternative techniques, such as machine learning.

1.5 Machine Learning

Machine learning (ML) is a field of computer science and artificial intelligence that focuses on enabling machines to learn without explicit programming. With the exponential growth of biological data in the field of omics, ML algorithms have gained significant traction in bioinformatics to extract knowledge from large and diverse datasets. ML techniques are primarily used for tasks such as feature selection, classification, clustering, and prediction of biological data. One of the key advantages of ML is its ability to uncover patterns and generate insights from heterogeneous datasets (Shastry & Sanjay, 2020).

ML can be broadly categorized into supervised and unsupervised learning. Supervised learning relies on prelabeled data and is typically used for tasks such as diagnostics and classification. Common supervised learning models include linear regression, naive Bayes, logistic regression, decision trees, ensemble methods e.g., random forest, and super vector machines (SVM). In supervised learning, models are trained on a dataset, and their weights are adjusted iteratively until an appropriate fit is achieved. However, challenges such as overfitting and underfitting can arise. Overfitting occurs when the model learns the patterns and noise of the specific dataset too well, making it less effective in generalizing to unseen data. On the other hand, underfitting happens when the model is too simplistic to capture important relationships in the data (IBM, n.d.). Several techniques can mitigate these challenges, including data scaling, feature selection, bootstrapping, cross-validation, hyperparameter tuning specific to the chosen model (Pudjihartono et al. 2022).

There are several methods available for evaluating the performance of a machine learning classification model. These metrics provide insights into the model's ability to make accurate predictions. Some commonly used evaluation metrics include the receiver operating characteristic (ROC), the area under the curve (AUC), accuracy, precision, recall, and F1 score. The ROC plots illustrate the relationship between the true positive rate and the false positive rate. The AUC score represents the area under this curve and serves as a measure of the model's overall performance. An AUC value, closer to 1, indicates better predictive ability. Accuracy is a metric that measures the proportion of the correct predictions made by the model out of all predictions. It provides an overall assessment of the model's correctness. Precision evaluates the model's ability to avoid false positives. It calculates the proportion of correctly predicted positive instances out of all instances predicted as positive. A high precision indicates a low rate of false positives. Recall, also known as sensitivity or true positive rate, measures the model's ability to detect all positive instances. It calculates the proportion of correctly predicted positive instances out of all actual positive instances. A high recall indicates low false negatives. The F1 score combines precision and recall into a single metric. It is the harmonic mean of precision and recall, providing a balanced assessment of both measures. A high F1 score indicates better overall performance in terms of both precision and recall. Additionally, the macro F1 score calculates the arithmetic mean of all F1 scores for each class. This metric can be useful when dealing with imbalanced datasets, where different classes have varying sample sizes (Gouette & Gaussier, 2005).

1.5.1 GLMNet

Glmnet is an R-based package for ML and stands for *Generalized Linear Models* (GLMs) *with the Elastic Net* (Net) *penalty*. This package provides fast algorithms for fitting GLMs with penalized maximum likelihood and regularization. In the presence of highly correlated predictors, lasso regression tends to select one predictor as important and discard the rest, while ridge regression shrinks the coefficients of correlated predictors towards each other. The elastic net penalty combines the two regression types allowing them to borrow strengths from each other and creating an embedded feature selection in the model (Friedman et al., 2010). The elastic net also permits tuning of the penalty term through the penalty parameters α and λ, which control the type (ridge ($\alpha = 0$) or lasso ($\alpha = 1$)) and amount of shrinkage, respectively. Furthermore, the tuning parameters α and λ can be automatically chosen using built-in k-fold cross-validation in the model (Engebretsen & Bohlin, 2019).

Glmnet is capable of handling large datasets efficiently, with short iteration and computation times. It includes models for regression, two-class logistic regression, and multinomial regression problems, enabling the fitting of generalized linear models to various response distributions, including Gaussian, binomial, multinomial, Poisson, etc. In other words, glmnet can generate both continuous and categorical output. The regularization provided by the λ-parameter helps prevent overfitting, making it well suited for large datasets (Friedman et al., 2010). Glmnet also offers a user-friendly interface with built-in functions for visualization, extracting feature importance lists with importance scores, classification confusion matrices and other performance measures (Hastie et al., 2023).

While logistic regression models are typically used for binomial classification, multinomial regression is employed for modeling categorical outcomes with more than two categories. In glmnet, multinomial regression generalizes the binomial logistic regression model to a multi-logit model (Friedman et al., 2010).

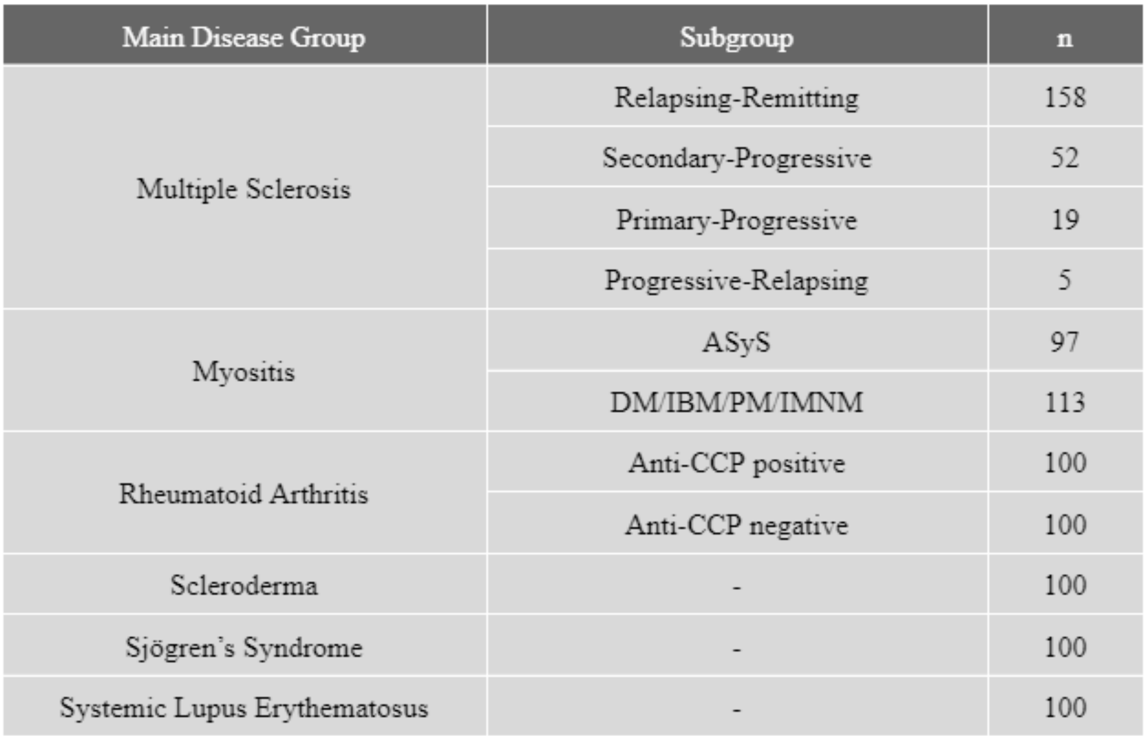
Methodology

2.1 Data Generation

2.1.1 Sampling

A total of 944 plasma samples were supplied by research groups at the Karolinska Institute, from the biobank at the Division of Rheumatology, Department of Medicine. The samples were collected from individuals at the date of first diagnosis over the course of a decade, and the diagnoses were assessed by experienced rheumatologists or neurologists (for MS). *Table 1* summarizes the distribution of samples among different diseases and subgroups. All RA samples were sourced from the EIRA study, while one third of the MS samples came from the EIMS study (EIRA Sweden, 2020; Andersen, 2012). Except for some EIMS samples, all other samples were collected at Karolinska University Hospital in EDTA tubes, separated from blood cells, and frozen within one hour or up to 6 days after initial sampling. A fraction of EIMS samples were collected at the local health centers and sent by regular mail to the biobank for further separation and storage. All samples have been stored at -80 °C. The majority of the samples came from individuals of northern European descent.

Table 1. *The distribution of samples per disease and subgroup.*



Main Disease Group	Subgroup	n
Multiple Sclerosis	Relapsing-Remitting	158
	Secondary-Progressive	52
	Primary-Progressive	19
	Progressive-Relapsing	5
Myositis	ASyS	97
	DM/IBM/PM/IMNM	113
Rheumatoid Arthritis	Anti-CCP positive	100
	Anti-CCP negative	100
Scleroderma	-	100
Sjögren's Syndrome	-	100
Systemic Lupus Erythematosus	-	100

2.1.2 Proximity Extension Assay

In 2022, the 944 samples were sent to the Olink Explore Lab at SciLifeLab in Uppsala, where the Olink® Explore platform was run with 1463 unique protein assays. Three assays (TNF, IL6 and CXCL8) were replicated in all four panels for control purposes. Olink provided a data file containing approximately 1.4 million log2-transformed intensity values, referred to as *normalized protein expression* (NPX) values, for all samples and proteins. These samples were a part of a larger study comprising a total of 10,000 samples that were normalized and bridged together. The 10,000 samples were run in four different batches over a period of 1,5 years. The autoimmune cohort was run in batch 3.

2.1.3 Healthy Cohort Data

A subset of the 10,000 samples came from the impaired glucose tolerance (IGT) cohort, a cross-sectional study of healthy, diabetic and prediabetic individuals. The healthy subset of the IGT cohort was called the normal glucose tolerance (NGT) group. These individuals were randomly recruited from the Gothenburg area, Sweden, through the census registry. All individuals in this cohort were aged between 50 and 64 (Wu et al., 2020). The healthy cohort was run in batch 2.

2.2 Data Preparation

All NPX values flagged with QC-warnings were removed. Among the three replicated assays, the assay with the highest mean NPX value was retained for each protein. Additionally, based on earlier investigations from the research group, 480 highly correlated and highly variable proteins were excluded due to suspicion that the correlation was driven by technical parameters rather than biological factors.

2.3 Differential Expression Analysis

Differential expression analysis was performed using two different methods in R (version 4.2.1). Initially, a normal distribution test, Shapiro-Wilk W-test (Shapiro & Wilk, 1965), was conducted, revealing a non-normal, skewed distribution for 90% of the data. Therefore, the non-parametric Mann-Whitney U-test (MWU) (Mann & Whitney, 1947) was employed. Subsequently, the linear regression model package limma was used, incorporating weights to account for sex and age and adjust for these factors. Both MWU and limma analyses were performed in a bimodal manner, comparing each disease to the healthy cohort and then to all other rheumatic diseases within the autoimmune cohort, as illustrated in *figure 1*. The p-values, from both methods, were adjusted using Benjamini-Hochberg correction (Benjamini & Hochberg, 1995). For the subgroup analysis of myositis and RA, MWU and limma were applied, designating one subgroup as the case and the other subgroup as control. In the case of MS subgroups, one subgroup was considered the case, while the remaining three were grouped into a control group, similar to the inter-disease comparisons.

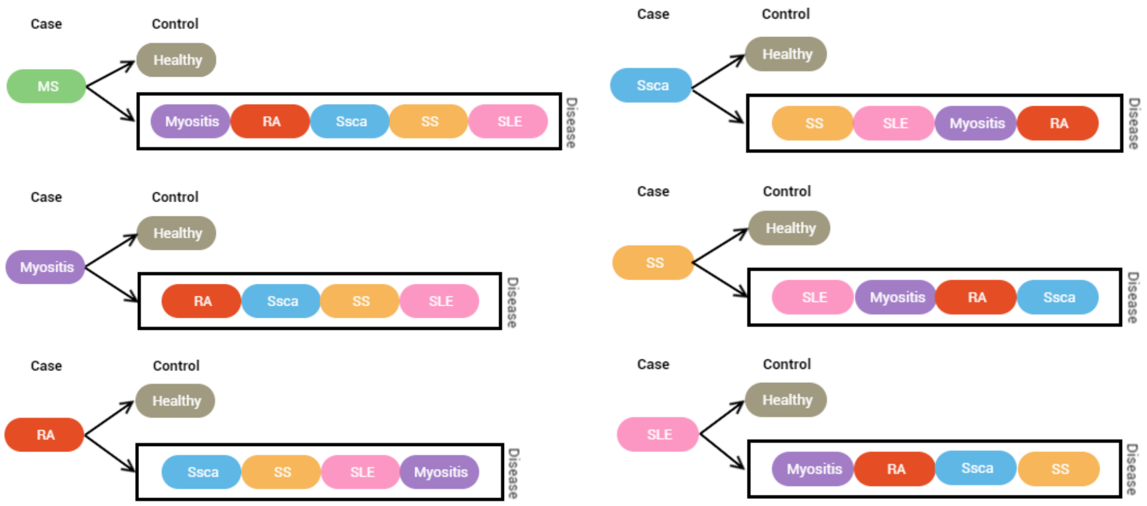


Figure 1. *A schematic of the case and control categorization for limma and Mann-Whitney U-test.*

2.4 Machine Learning

A machine learning model for disease classification was built using the glmnet engine in R and the tidymodels R package (version 1.0.0). The model was trained and evaluated on the autoimmune cohort data, with assays serving as features and the six disease groups as classes. The following parameters were adjusted for model optimization: seed setting, data splitting, feature scaling, feature selection, and class balancing. Model performance was evaluated by macro-F1 scores. The final model employed a 70/30 split of training and test data, with all features scaled using z-score normalization (Zhang et al., 2014). Feature selection and hyperparameter optimization was performed using built-in k-fold cross-validation with 5 folds, and ultimately, class balancing was not applied as it did not have any positive effect on the evaluation score.

2.5 Protein Selection

To identify potential plasma protein biomarkers with highest association to each disease, adjusted p-values from the MWU and limma differential expression analyses were combined with the importance scores from the machine learning model. Proteins with the lowest adjusted p-values and highest importance scores were considered to be the most relevant biomarker candidates in patients of a particular disease compared to all other autoimmune cohort patients and NGT cohort individuals. In the case of the subgroups, potential protein biomarkers were selected solely based on MWU and limma p-values.

Results

The initial phase of the project involved exploring various aspects of the data. *Figure 2,* displays the distribution of sex and age, revealing a higher representation of younger age groups among the MS and SLE cohorts. It is also evident from the figure that the majority of the samples were from female patients, which aligns with the natural occurrence of the diseases. Additionally, PCA and UMAP plots were utilized to assess other data characteristics for all samples and proteins, as depicted in *figure 2.* These plots indicate a certain degree of heterogeneity in terms of sex differences and a clear separation among disease groups. This first observation suggested the presence of detectable expression differences related to disease groups. the first indication that there should be detectable expression differences related to disease group. Among the disease clusters in the PCA and UMAP plots, the most distinct separation was observed for MS, myositis and RA, while the remaining three diseases exhibited more overlapping patterns with each other.

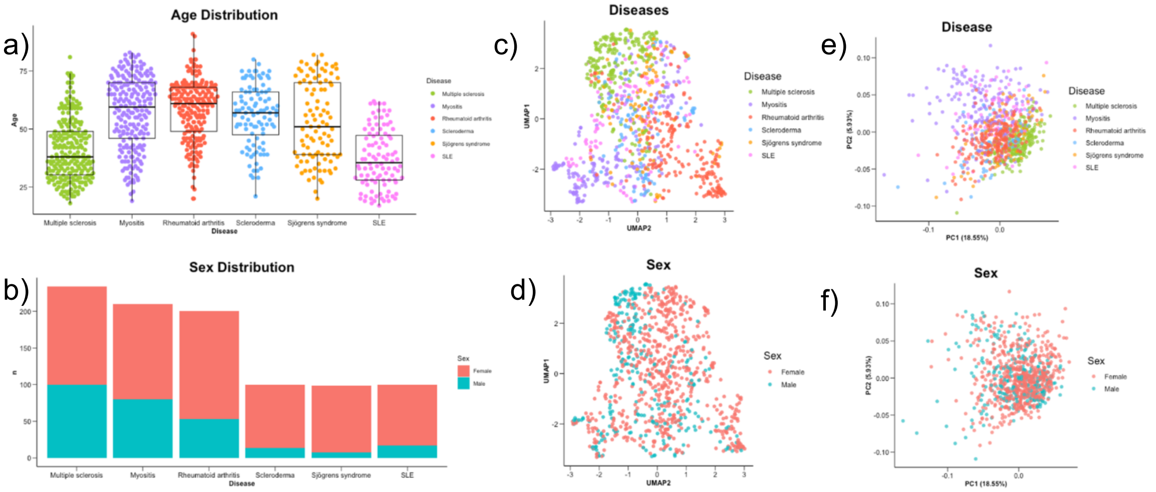


Figure 2. *a) Boxplots of age distribution per disease group. b) Bar plot of the sex distribution within each disease group. c) UMAP plot of all samples and all proteins colored by disease group. d) UMAP plot of all samples and all proteins colored by sex. e) PCA plot of all samples and all proteins colored by disease group. f) PCA plot of all samples and all proteins colored by sex.*

The analysis methods used in this project were differential expression analysis (limma), non-parametric testing (MWU), and machine learning (glmnet). Both MWU test and limma model successfully identified differentially expressed proteins. These two methods also exhibited a high correlation in terms of p-values in most cases, indicating consistent results. However, notable differences were observed when comparing the results obtained from different case and control combinations, likely attributable to the healthy data being a separate, distinct cohort and processed in a separate batch. These differences are illustrated with heatmaps in the subsequent sections.

The glmnet model gave high macro-F1 scores (0.85~0.9) after tuning the parameters. Nevertheless, it was discovered that the score did not consider the model's tendency to select insignificant proteins as important features. To address this, feature scaling was implemented before training the model. Furthermore, the choice of k-fold cross-validation had a notable impact on the macro-F1 score, with five folds being determined as optimal in terms of computation time and predictive power. The data was split into 70% for training and 30% for testing, which yielded the best results for the model.

Class balancing was experimented with by randomly sampling 100 samples from each disease group. This approach had a slight negative impact on the macro-F1 score and was therefore not included. Lastly, varying seed values during model training had mixed effects on the scores, suggesting that the randomization process had some influence on the results. Nevertheless, the macro-F1 scores consistently fell within the range of 0.83~0.91, indicating a relatively robust decision-making process by the model. It should be declared that further hyperparameter tuning was not explored, which could have potentially yielded more significant effects on the results.

Scaling features had the most pronounced effect on the selection of proteins deemed important by the model. This can be attributed to the separate quantification and, in some cases, dilution of assays, resulting in a wide range of NPX values. Consequently, the model may make topological inferences that do not align with the actual data. The final model achieved an F1-score of 0.8975622. In *figure 3*, the ROC curves per disease for the final model are displayed, along with the corresponding AUC scores. The macro-AUC score is 0.9983, and for MS and SLE, the AUC scores were 1, indicating excellent predictive ability of the model in these specific diseases.

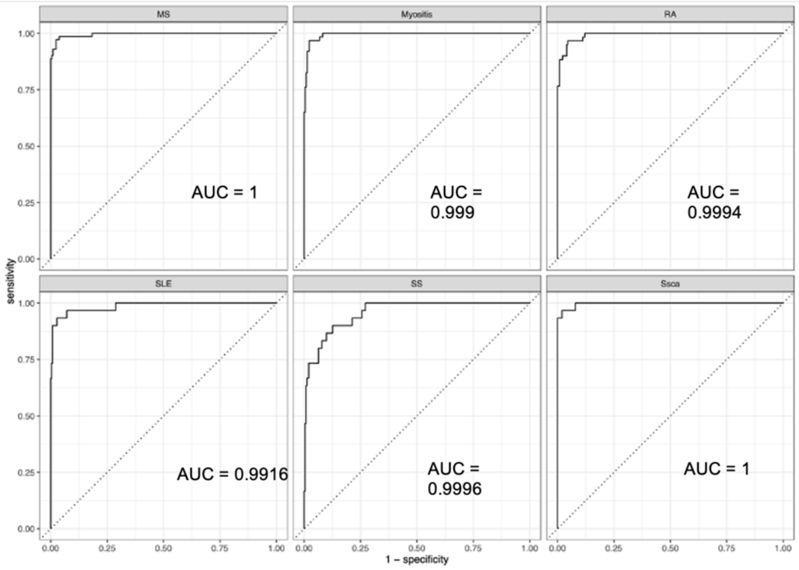


Figure 3. *Displays the ROC-curves and corresponding AUC scores per disease class from the final glmnet model when run on the test data. Order of plots from top left: MS, myositis, RA, SLE, SS, Ssca.*

The results for each disease and subgroup are described and illustrated as follows. Firstly, heatmaps are provided to visualize the performance of limma, MWU, and glmnet across different case-control combinations. Secondly, a volcano plot and boxplot are presented, depicting the results of the limma model when using the disease as the case and all other rheumatic diseases as controls. This particular case-control combination was considered the most reliable since all NPX values were obtained from the same cohort and batch. To ensure accuracy, the analysis of rheumatic disorders excluded the MS samples due to the clinical similarities of rheumatic diseases and the adherence to the same sampling pool. Thirdly, a lollipop plot and boxplot are included to showcase the most important classification features predicted by the glmnet model. These plots provide insights into the proteins that contribute significantly to the classification. Additionally, a table is added for each disease, summarizing the top ten proteins with the highest differential expression and importance scores.

3.1 Multiple Sclerosis

The MS cohort exhibited significant differences from the other autoimmune disease cohorts, and in most cases, showed a similar protein expression to the healthy cohort. The limma analysis, using MS as the case and the rheumatic disorders as controls, resulted in down-regulation of most proteins. Conversely, when limma was run with the healthy cohort as the control, the MS cohort showed predominantly up-regulated proteins. *Figure 4* illustrates the discrepancy between the results obtained from different case and control combinations. The results from limma and MWU analyses with rheumatic disorders as controls showed similar outcomes, while limma and MWU analyses with the healthy cohort as the control presented the opposite proteins as highly significant.

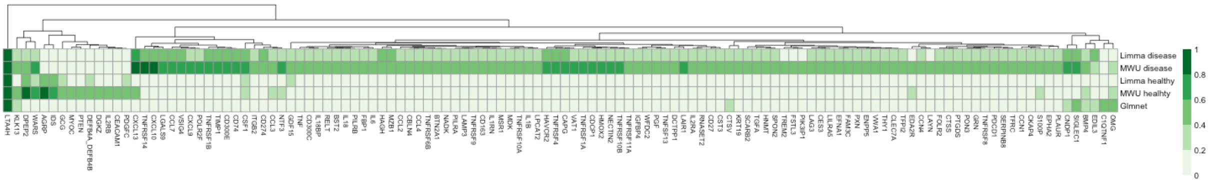


Figure 4. *A heatmap plot of the MS cohort showing some proteins on the x-axis and methods together with case-control combination on the y-axis. The color scale represents the significance scoring by the method, either adjusted p-value or importance score, with 1 representing the highest significance. Limma/MWU disease: other rheumatic diseases were used as control. Limma/MWU healthy: the healthy cohort was used as control.*

The results from the differential expression analysis and machine learning are depicted in *figure 5*. LTA4H emerged as the protein that stood out in both models, exhibiting significantly lower p-values and higher importance scores. The boxplots from the limma results in *figure 5b)* demonstrate that the most differentially expressed proteins were similar to the healthy cohort, indicating no significant differential expression. The finding aligns with the results observed in the heatmap shown in *figure 4.* The boxplots in *figure 5d)* reveal that the glmnet model successfully identified three proteins, namely OMG, EDIL3, and CTSV, which exhibited consistent up- or downregulation compared to all other cohorts. OMG is a cell adhesion molecule involved in the myelination process in the CNS (HPA, n.d.). EDIL3, associated with promoting adhesion in epithelial cells and inhibiting the formation of vascular-like structures, has been found to be upregulated in various types of cancers. CTSV, a protease that may play a role in corneal physiology, has no specific mention in literature regarding MS (HPA, n.d.).

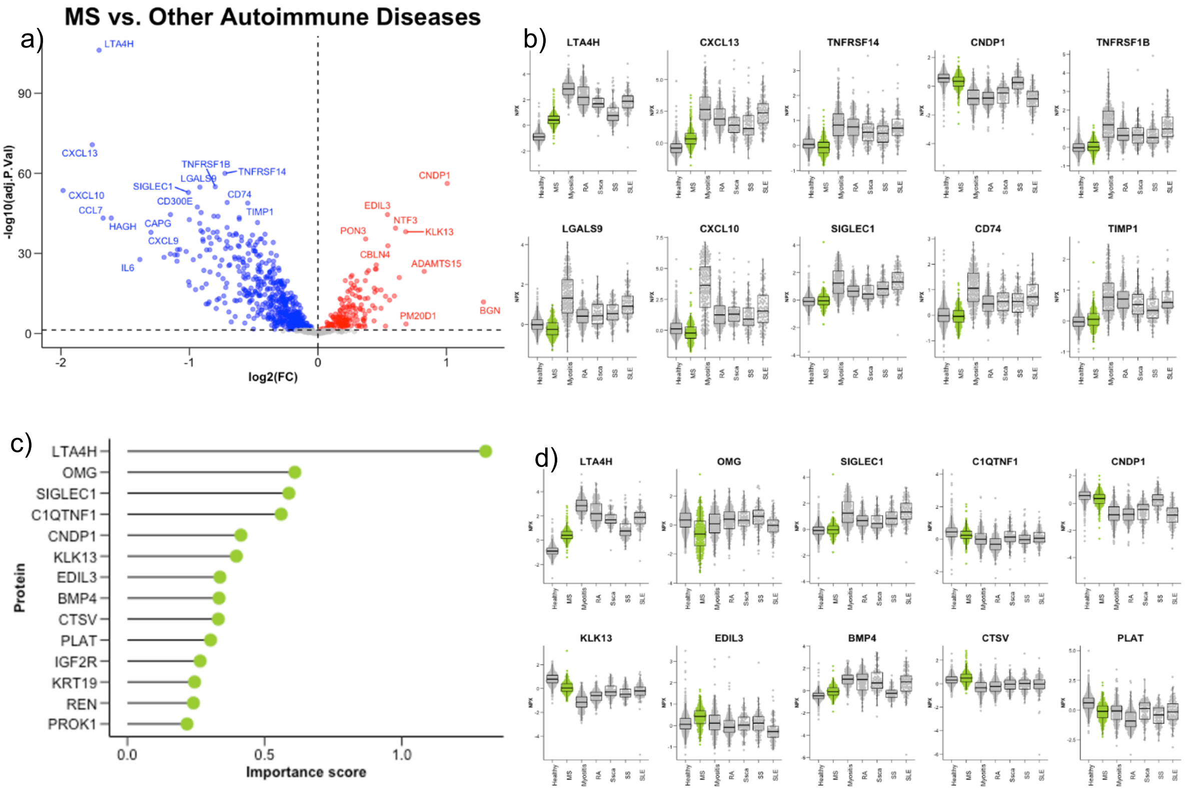


Figure 5. *a) Volcano plot of the resulting p-values and fold change when running limma for the MS cohort as case and the remaining diseases as control. b) Boxplots of the top ten most significantly differentially expressed proteins according to the limma analysis with NPX values on the y-axis and cohort names on the x-axis. c) Lollipop plot of the top, most important, features and their importance scores as assigned by the glmnet model. d) Boxplots of the top ten most important features assigned by the glmnet model. The MS cohort is shown in green.*

Table 2 presents the top ten proteins resulting from the limma and glmnet analyses, with two proteins, LTA4H and SIGLEC1, appearing in both analyses. These proteins were identified as downregulated. LTA4H has no mention in literature related to MS, while SIGLEC1 has been associated with MS, particularly the progressive type. Several studies have discussed SIGLEC1 as a biomarker for active neuroinflammation in the brain when examining monocytes, although it has not been found to be upregulated in blood (Ostendorf et al., 2021). Among the remaining proteins, most have been discussed in the literature in relation to MS. Notable mentions include CXCL13, proposed biomarker in CSF (DiSano et al., 2020); TNFRSF14, associated with MS risk when downregulated (Torre-Fuentes et al., 2020); TIMP1, elevated in patients with MS serum samples (Trentini et al., 2015); and CD74, upregulated in MS, and currently targeted by a monoclonal antibody in clinical trials (Haran et al., 2018).

Table 2. *Shows the top ten proteins picked by the differential expression (DE) model limma (with rheumatic disease as control group) and the machine learning (ML) model glmnet (using all disease groups). It also includes the full protein names and the number of proteins picked by both models.*

Rank	DE (limma)	Full protein name	ML (glmnet)	Full protein name
1	LTA4H	*Leukotriene A4 hydrolase*	LTA4H	*Leukotriene A4 hydrolase*
2	CXCL13	*C-X-C motif chemokine ligand 13*	OMG	*Oligodendrocyte myelin glycoprotein*
3	TNFRSF14	*TNF receptor superfamily member 14*	SIGLEC1	*Sialic acid binding Ig like lectin 1*
4	CNDP1	*Carnosine dipeptidase 1*	C1QTNF1	*C1q and TNF related 1*
5	TNFRSF1B	*TNF receptor superfamily member 1B*	CNDP1	*Carnosine dipeptidase 1*
6	LGALS9	*Galectin 9*	KLK13	*Kallikrein related peptidase 13*
7	CXCL10	*C-X-C motif chemokine ligand 10*	EDIL3	*EGF like repeats and discoidin domains 3*
8	SIGLEC1	*Sialic acid binding Ig like lectin 1*	BMP4	*Bone morphogenetic protein 4*
9	CD74	*CD74 molecule*	CTSV	*Cathepsin V*
10	TIMP1	*TIMP metallopeptidase inhibitor 1*	PLAT	*Plasminogen activator, tissue type*
Overlap: 2 proteins				

3.1.1 Multiple Sclerosis – Subgroups

When examining the subgroups of MS, namely RRMS, SPMS and PPMS, there were minimal differences in the expression levels of proteins among these groups, as depicted in the volcano plots shown in *figure 6.* The PRMS subgroup exhibited a slightly higher number of proteins with significant upregulation; however, it is important to note that this subgroup consisted of only five samples, rendering these findings statistically insignificant.

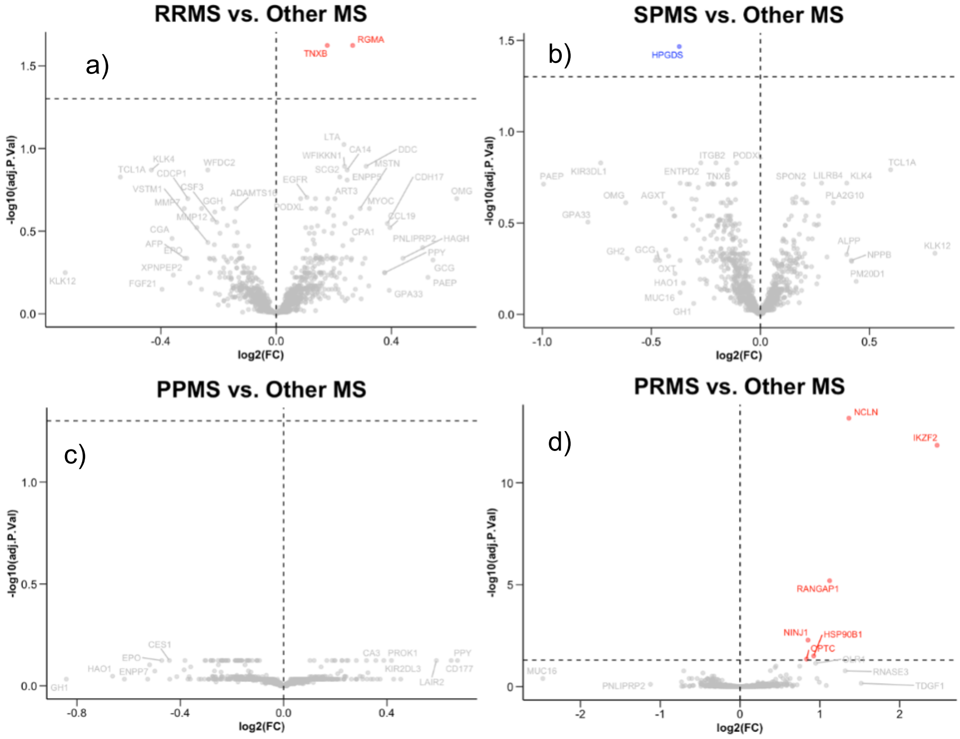


Figure 6. *Volcano plots from limma results when using a) relapsing-remitting MS as case and other MS as control, b) secondary-progressive MS as case and other MS as control, c) primary-progressive MS as case and other MS as control, d) progressive-remitting MS as case and other MS as control.*

Figure 7. presents the boxplots of all the proteins that were identified as significantly up- or down-regulated. Many of these proteins have been mentioned in literature in relation to MS. For instance, RGMA has been demonstrated to be involved in the pathogenesis of MS in mouse models, although no specific subgroup mention was found (Zhang et al., 2022). TXNB has also been associated with MS and SLE in GWAS studies (Tajuddin et al., 2016).

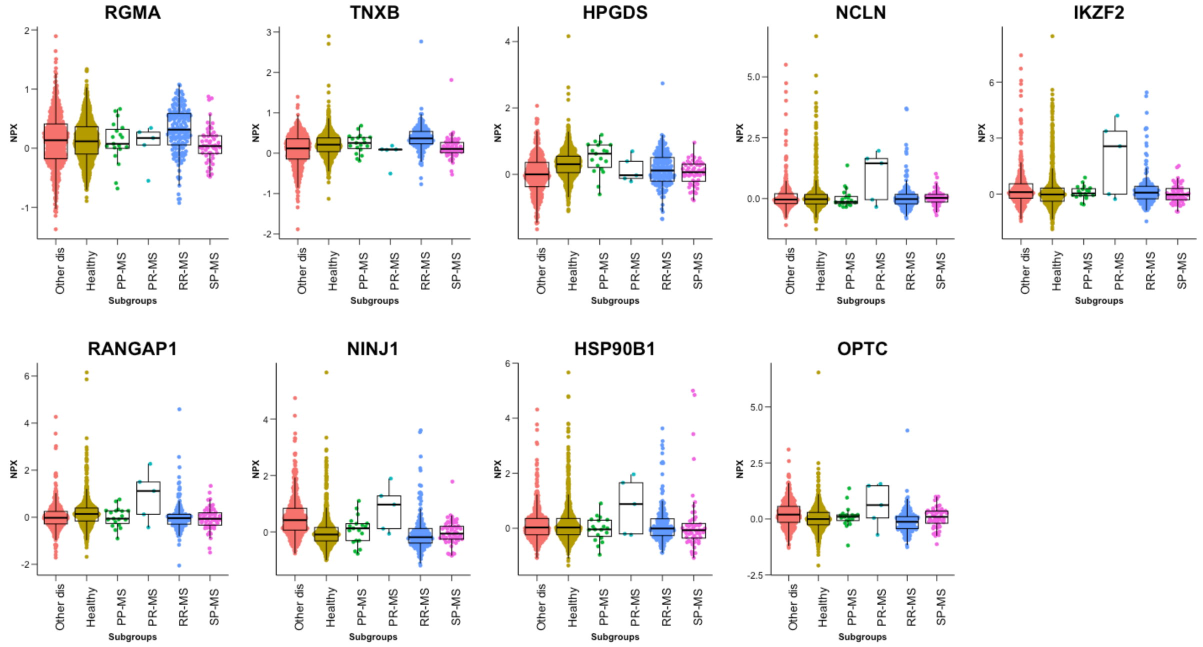


Figure 7. *Boxplots of the 9 differentially expressed protein from all four MS subgroup analyses, with NPX values on the y-axis and cohort or subgroup names on the x-axis.*

3.2 Myositis

In the myositis analysis, all three methods and both control combinations, exhibited relatively similar performance and ranking of the topmost important proteins. This is illustrated in the heatmap presented in *figure* 8 where there is a clustering occurring in two darker areas.

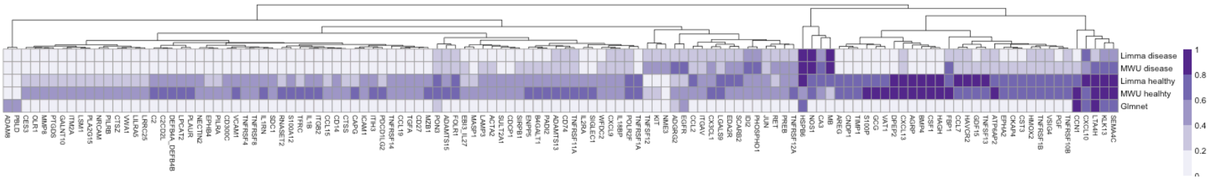


Figure 8. *A heatmap plot of the myositis cohort showing some proteins on the x-axis and methods together with case-control combination on the y-axis. The color scale represents the significance scoring by the method, either adjusted p-value or importance score, with 1 representing the highest significance. Limma/MWU disease: other rheumatic diseases were used as control. Limma/MWU healthy: the healthy cohort was used as control.*

The myositis analysis unveiled several highly differentially expressed proteins, as indicated by both their high p-values and large fold changes, as depicted in the volcano plot in *figure 9a)* Among the top ten proteins ranked by both models, HSPB6, CXCL10, KLK13, SEMA4C, and LTA4H were identified by both machine learning and differential expression analysis, as presented in *table 3.*

HSPB6, a small heat shock protein, plays a regulatory role in muscle function. Its upregulation has been associated with cardio protection and angiogenesis following induced damage (HPA, n.d.). Moreover, HSPB6 has been linked to reactive oxygen species (ROS) homeostasis (Capitanio et al., 2015), which are generated during mitochondrial oxidative metabolism and can also be amplified as a cellular response to cytokines, bacterial invasion or xenobiotics. (Ray et al., 2012). Overexpression of this protein has been observed in various myopathies including the IIMs. (Merino-Jiménez et al., 2019)

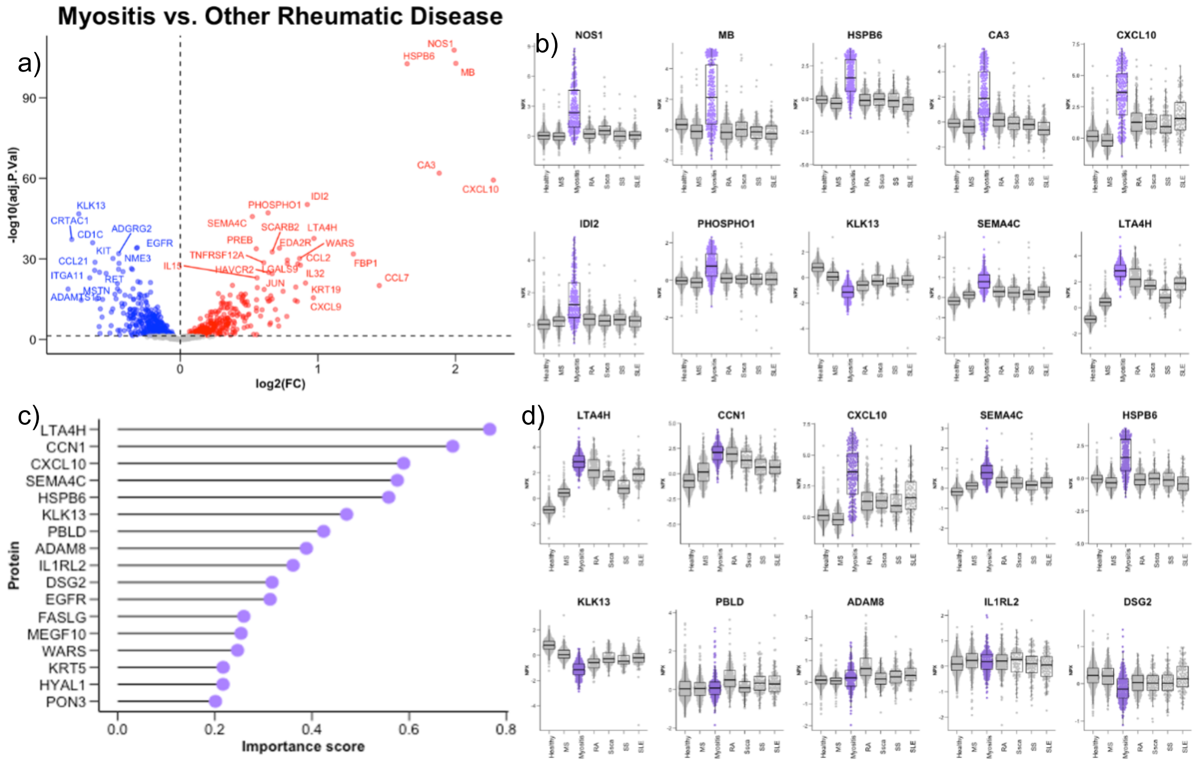


Figure 9. *a) Volcano plot of the resulting p-values and fold change when running limma for the myositis cohort as case and the remaining rheumatic diseases as control. b) Boxplots of the top ten most significantly differentially expressed proteins according to the limma analysis with NPX values on the y-axis and cohort names on the x-axis. c) Left: Lollipop plot of the top, most important, features and their importance scores as assigned by the glmnet model. d) Boxplots of the top ten most important features assigned by the glmnet model. The myositis cohort is shown in purple.*

CXCL10 is a pro-inflammatory cytokine, while KLK13 belongs to a subgroup of serine proteases that play various roles in the body and have been implicated in carcinogenesis. However, literature does not mention a direct relation between KLK13 and IIMs. SEMA4C is a cell surface protein involved in cell-cell signaling and muscle cell differentiation, whereas LTA4H catalyzes the conversion of LTA4 into a pro-inflammatory mediator leukotriene. LTA4H is recognized as a biomarker for chronic obstructive pulmonary disease (COPD) (HPA, n.d.) and has also been implicated in the pathogenesis of IIMs (Korotkova & Lundberg, 2014).

The boxplots presented in *figure 9* reveal several proteins with elongated plot profiles, indicating the presence of subgroups. However, these subgroups were not related to the pre-assigned myositis subgroups in this study. The limma analysis further revealed the significantly upregulated proteins NOS1, MB, CA3, IDI2 and PHOSPHO1 that were not identified by glmnet in the final model. NOS1 has been shown to be upregulated in myositis (Tews & Goebel, 1998), as has MB (Kagan, 1977).

Table 3. *Shows the top ten proteins picked by the differential expression (DE) model limma (with rheumatic disease as control group) and the machine learning (ML) model glmnet (using all disease groups). It also includes the full protein names and the number of proteins picked by both models.*

Rank	DE (limma)	Full protein name	ML (glmnet)	Full protein name
1	NOS1	*Nitric oxide synthase 1*	LTA4H	*Leukotriene A4 hydrolase*
2	MB	*Myoglobin*	CCN1	*Cellular communication network factor 1*
3	HSPB6	*Heat shock protein family B member 6*	CXCL10	*C-X-C motif chemokine ligand 10*
4	CA3	*Carbonic anhydrase 3*	SEMA4C	*Semaphorin 4C*
5	CXCL10	*C-X-C motif chemokine ligand 10*	HSPB6	*Heat shock protein family B member 6*
6	IDI2	*Isopentenyl-diphosphate delta isomerase 2*	KLK13	*Kallikrein related peptidase 13*
7	PHOSPHO1	*Phosphoethanolamine/ phosphocoline phosphatase 1*	PBLD	*Phenazine biosynthesis like protein domain containing*
8	KLK13	*Kallikrein related peptidase 13*	ADAM8	*ADAM metallopeptidase domain 8*
9	SEMA4C	*Semaphorin 4C*	IL1RL2	*Interleukin 1 receptor like 2*
10	LTA4H	*Leukotriene A4 hydrolase*	DSG2	*Desmoglein 2*
Overlap: 5 proteins				

3.2.1 Myositis – Subgroups

For the myositis subgroups a protein that really stood out was the SFTPD, a protein that contributes to the lung's defence against toxins, microorganisms and organic antigens (HPA, n.d.) As shown in the volcano plot of *figure 10*, this protein was highly upregulated in patients with ASyS who often do develop ILD.

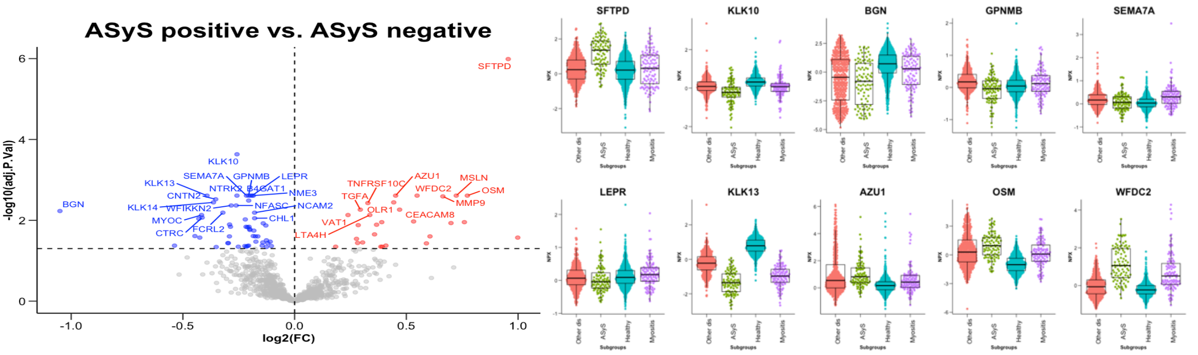


Figure 10. *Left: Volcano plot from the limma analysis of ASyS positive patients as case and ASyS negative patients as control. Right: Boxplots of the top ten protein with lowest p-value and highest fold change, with NPX values on the y-axis and cohort names on the x-axis. These plots show the healthy cohort, the myositis subgroups and the remaining disease cohorts as one called "Other dis".*

3.3 Rheumatoid Arthritis

The analysis of the RA cohort using different observational combinations yielded consistent results, as demonstrated in *figure 11*. The darker clustering area in the bottom right of the figure shows strong agreement between the methods.

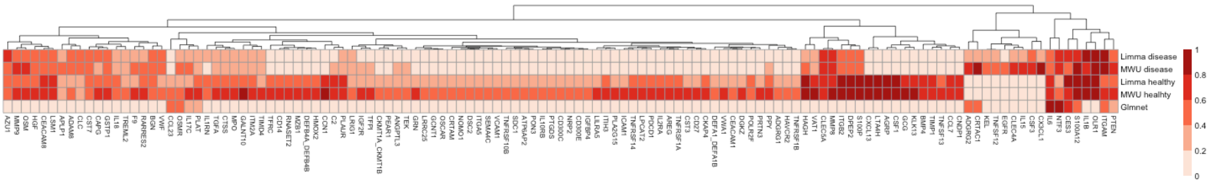


Figure 11. *A heatmap plot of the RA cohort showing some proteins on the x-axis and methods together with case-control combination on the y-axis. The color scale represents the significance scoring by the method, either adjusted p-value or importance score, with 1 representing the highest significance. Limma/MWU disease: other rheumatic diseases were used as control. Limma/MWU healthy: the healthy cohort was used as control.*

Furthermore, the results from the limma analysis indicated high significance and fold change, as evidenced by the distinct boxplots displayed in *figure 12b)*. Notably, four proteins, namely IL1B, OLR1, AZU1 and MMP8, suggested the presence of another type of subgrouping that is unrelated to the preassigned clinical subgroups of seropositive and seronegative RA.

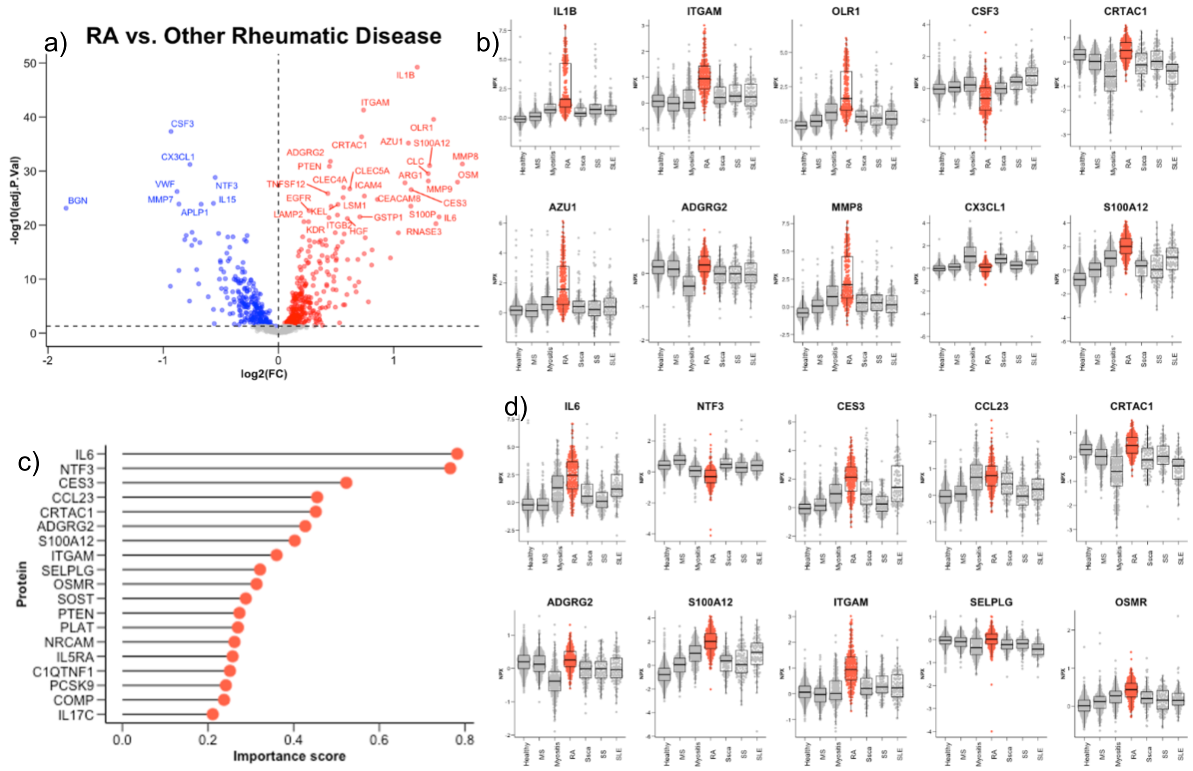


Figure 12. *a) Volcano plot of the resulting p-values when running limma for the RA cohort as case and the remaining rheumatic diseases as control. b) Boxplots of the top ten most significantly differentially expressed proteins according to the limma analysis with NPX values on the y-axis and cohort names on the x-axis. c) Lollipop plot of the top, most important, features and their importance scores as assigned by the glmnet model. d) Boxplots of the top ten most important features assigned by the glmnet model. The RA cohort is shown in red.*

Table 4 reveals an overlap of four proteins between the glmnet model and the limma model, namely ITGAM, CRTAC1, ADGRG2 and S100A12. In the next section, CRTAC1 is discussed as a protein that exhibits differential expression between seropositive and seronegative RA. However, it is worth noting that the CRTAC1 is also highly expressed in all RA patients compared to other disease groups. ITGAM, although not specifically mentioned in literature regarding RA, has been identified as a strong susceptibility gene for SLE. (Järvinen et al., 2010) ITGAM is an integrin involved in adhesive interactions with macrophages, granulocytes, and monocytes. ADGRG2 has no direct association with RA in the literature; it is an orphan receptor that appears to be involved in male fertility and has been linked to prostate cancer. S100A12, on the other hand, is a pro-inflammatory protein that plays a prominent role in regulating inflammatory response and immune system function. (HPA, n.d.) These proteins have been discussed in relation to RA for several decades (Foell et al., 2004). More recently, S100A12 has been proposed as a blood biomarker for early diagnosis of RA (Wang et al., 2022)

Table 4. *Shows the top ten proteins picked by the differential expression (DE) model limma (with rheumatic disease as control group) and the machine learning (ML) model glmnet (using all disease groups). It also includes the full protein names and the number of proteins picked by both models.*

Rank	DE (limma)	Full protein name	ML (glmnet)	Full protein name
1	IL1B	*Interleukin beta 1*	IL6	*Interleukin 6*
2	ITGAM	*Integrin subunit alpha M*	NTF3	*Neurotrophin 3*
3	OLR1	*Oxidized low density lipoprotein receptor 1*	CES3	*Carboxylesterase 3*
4	CSF3	*Colony stimulating factor 3*	CCL23	*C-C motif chemokine ligand 23*
5	CRTAC1	*Cartilage acidic protein 1*	CRTAC1	*Cartilace acidic protein 1*
6	AZU1	*Azurocidin 1*	ADGRG2	*Adhesion G protein-coupled receptor G2*
7	ADGRG2	*Adhesion G protein-coupled receptor G2*	S100A12	*S100 calcium binding protein A12*
8	MMP8	*Matrix metallopeptidase 8*	ITGAM	*Integrin subunit alpha M*
9	CXCL1	*C-X-C motif chemokine ligand 1*	SELPLG	*Selectin P ligand*
10	S100A12	*S100 calcium binding protein A12*	OSMR	*Oncostatin M receptor*
Overlap: 4 proteins				

3.3.1 Rheumatoid Arthritis – Subgroups

The RA subgroup analysis also revealed several significant proteins. However, the downregulated part of the volcano plot in *figure 13* is particularly interesting, as these two proteins are upregulated in seronegative patients. This finding is noteworthy because there are currently no established biomarkers for seronegative patients.

One of these proteins is NPPB, a cardiac hormone known for mediating cardio-renal homeostasis (HPA, n.d.) but with no reported association with rheumatoid arthritis or seronegative patients. On the other hand, recent research conducted in collaboration with the UK Biobank has highlighted the role of cartilage acidic protein 1 (CRTAC1). The study suggested that CRTAC is upregulated in the plasma of osteoarthritis patients and may serve as an indicator of future risk of joint replacement (Sturkarsdottir et al., 2022). Although osteoarthritis is a non-autoimmune form of arthritis, the similarity in CRTAC1 expression with the autoimmune seronegative type is intriguing.

Another significant protein is PDCD1, a programmed cell death protein that is significantly associated with rheumatoid arthritis and other autoimmune conditions. (Siwiec & Majdan, 2015). Lastly, SERPINA9, a protease inhibitor, has no specific mention in literature regarding seropositive RA (HPA, n.d.).

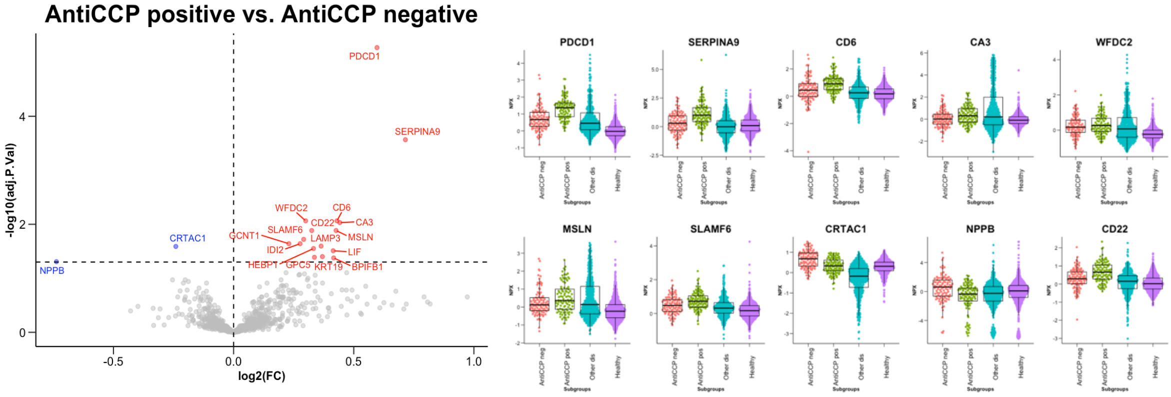


Figure 13. *Left: Volcano plot from the limma analysis of seropositive patients as case and seronegative patients as control. Right: Boxplots of the top ten protein with lowest p-value and highest fold change, with NPX values on the y-axis and cohort names on the x-axis. These plots show the healthy cohort, the RA subgroups and the remaining disease cohorts as one called "Other dis".*

3.4 Scleroderma

Figure 14 clearly demonstrates that all five modeling procedures consistently identified four proteins that are both significantly expressed and highly important for classification of scleroderma. This convergence of results highlights the robustness and relevance of these proteins as potential biomarkers for scleroderma.

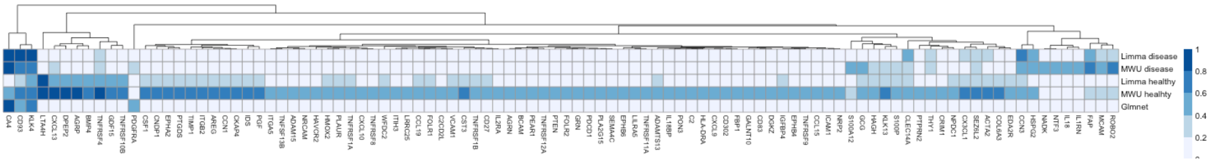


Figure 14. *A heatmap plot of the Ssca cohort showing some proteins on the x-axis and methods together with case-control combination on the y-axis. The color scale represents the significance scoring by the method, either adjusted p-value or importance score, with 1 representing the highest significance. Limma/MWU disease: other rheumatic diseases were used as control. Limma/MWU healthy: the healthy cohort was used as control.*

The distinct profiles observed in the boxplots of *figure 15* further support the promise of these top scleroderma markers for biomarker discovery. Not only do these proteins exhibit high p-values and fold change, but they also display highly distinct importance score profiles in machine learning, which is favorable for their potential use in classification.

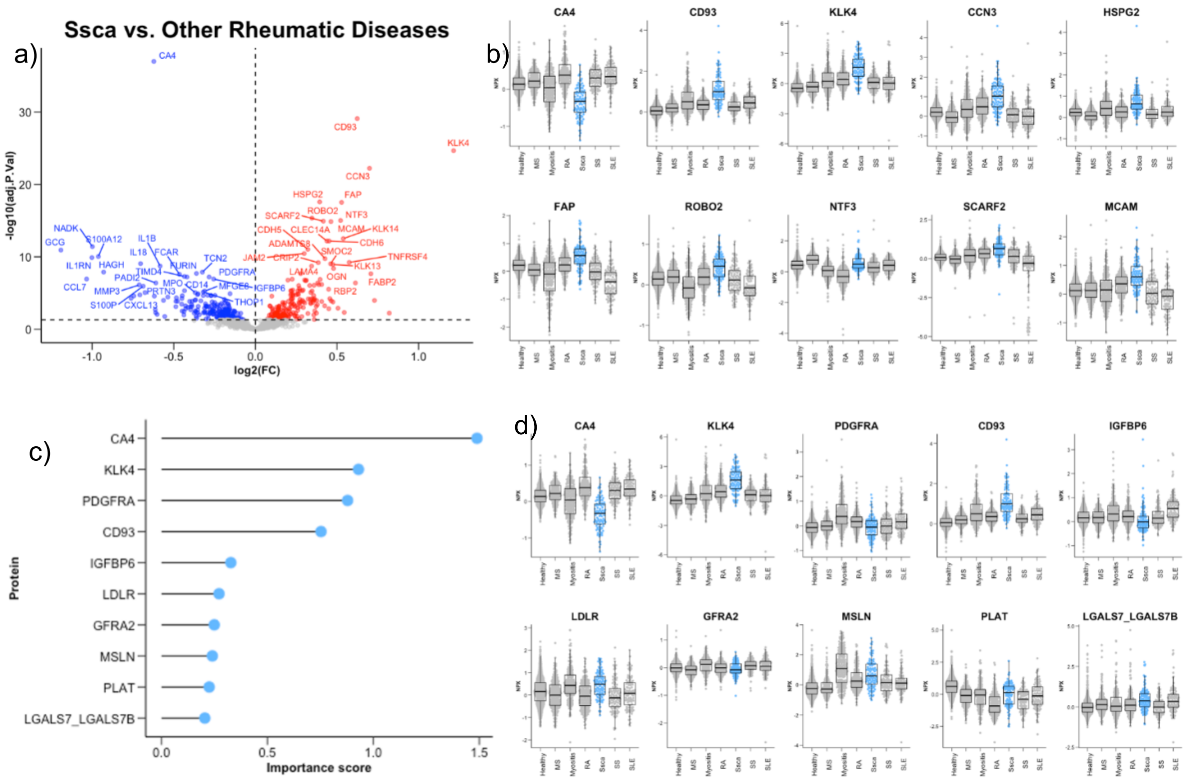


Figure 15. *a) Volcano plot of the resulting p-values when running limma for the Ssca cohort as case and the remaining rheumatic diseases as control. b) Boxplots of the top ten most significantly differentially expressed proteins according to the limma analysis with NPX values on the y-axis and cohort names on the x-axis.. c) Lollipop plot of the top, most important, features and their importance scores as assigned by the glmnet model. d) Boxplots of the top ten most important features assigned by the glmnet model. The Ssca cohort is shown in blue.*

The three proteins that are consistently identified by both limma and glmnet can be extracted from *table 5*: CA4, CD93 and KLK4. CA4 a carbonic anhydrase, plays a crucial role in the reversible hydration of carbon dioxide and is an FDA-approved drug target (HPA, n.d.). Previous studies have discussed the association of carbonic anhydrase II (CA2) with lung disease in systemic sclerosis patients, wherein autoantibodies against CA2 were detected (Alesandri et al., 2009). CD93, a receptor for mannose-binding lectin and pulmonary surfactant protein A, has shown elevated levels in the serum of systemic sclerosis patients (Yanaba et al., 2012). Although KLK4, a kallikrein-related peptidase, is not directly linked to scleroderma in the literature, it has been characterized as a cancer-associated gene. (HPA, n.d.)

Table 5. *Shows the top ten proteins picked by the differential expression (DE) model limma (with rheumatic disease as control group) and the machine learning (ML) model glmnet (using all disease groups). It also includes the full protein names and the number of proteins picked by both models.*

Rank	DE (limma)	Full protein name	ML (glmnet)	Full protein name
1	CA4	*Carbonic anhydrase 4*	CA4	*Carbonic anhydrase 4*
2	CD93	*CD93 molecule*	KLK4	*Kallikrein related peptidase 4*
3	KLK4	*Kallikrein related peptidase 4*	PDGFRA	*Platelet derived growth factor receptor alpha*
4	CCN3	*Cellular communication network factor 3*	CD93	*CD93 molecule*
5	HSPG2	*Heparan sulfate proteoglycan 2*	IGFBP6	*Insulin like growth factor binding protein 6*
6	FAP	*Fibroblast activation protein alpha*	LDLR	*Low density lipoprotein receptor*
7	ROBO2	*Roundabout guidance receptor 2*	GFRA2	*GDNF family receptor alpha 2*
8	NTF3	*Neutrophin 3*	MSLN	*Mesothelin*
9	SCARF2	*Scavenger receptor class F member 2*	PLAT	*Plasminogen activator, tissue type*
10	MCAM	*Melanoma cell adhesion molecule*	LGALS7/7B	Galectin 7/Galectin 7B
Overlap: 3 proteins				

3.5 Sjögren's Syndrome

The heatmap in *figure 16* illustrates the overlap between the MWU and limma analyses in the same case-control combinations. Similarly to the MS cohort, the disease and healthy cohorts as controls give quite contrary outcomes. Glmnet appears to have favored similarly to MWU and limma using the disease cohort as control.

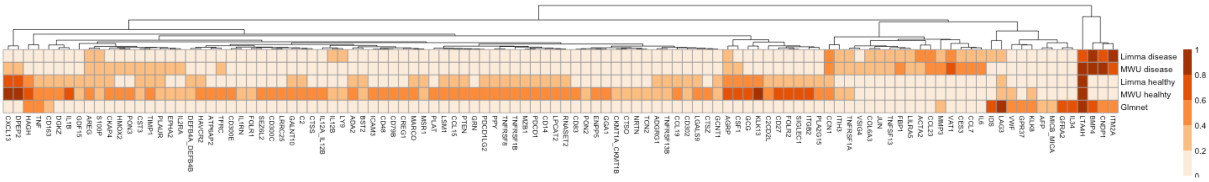


***Figure 16.** A heatmap plot of the SS cohort showing proteins on the x-axis and methods together with case-control combination on the y-axis. The color scale represents the significance scoring by the method, either adjusted p-value or importance score, with 1 representing the highest significance. Limma/MWU disease: other rheumatic diseases were used as control. Limma/MWU healthy: the healthy cohort was used as control.*

In *figure 17*, the linear regression and machine learning results are presented. The boxplots indicate that distinguishing SS from other diseases is challenging since the differences in expression are not prominently evident. Moreover, there are some extreme outliers in certain disease groups, which strongly influence the model's selection of important assays. An example of this is TNF in the SS cohort.

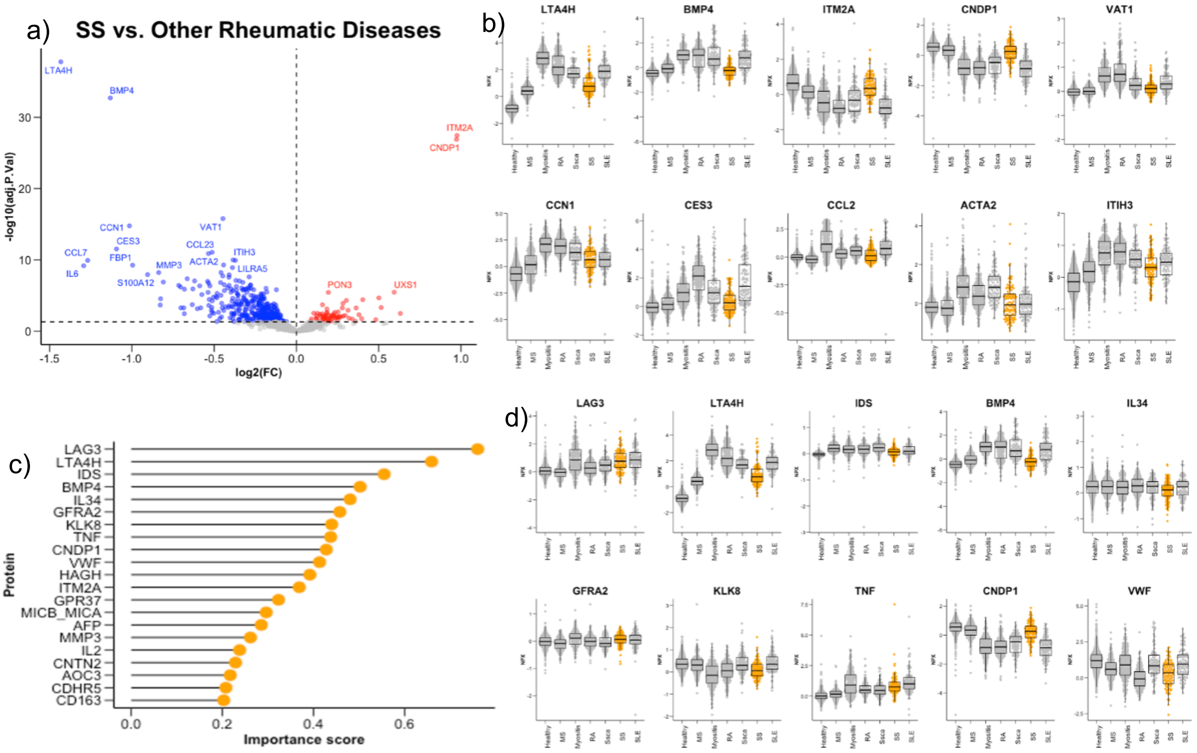


Figure 17. *a) Volcano plot of the resulting p-values when running limma for the SS cohort as case and the remaining rheumatic diseases as control. b) Boxplots of the top ten most significantly differentially expressed proteins according to the limma analysis with NPX values on the y-axis and cohort names on the x-axis. c) Lollipop plot of the top, most important, features and their importance scores as assigned by the glmnet model. d) Boxplots of the top ten most important features assigned by the glmnet model. The SS cohort is shown in orange.*

Table 6 clearly demonstrates that three of the most important classifying features remained the same as in the limma top ten selection. Both models identified LTA4H as a significant protein. However, it is difficult to ascertain from the boxplots in *figure 17* whether LTA4H is genuinely downregulated in SS or if the remaining diseases exhibit a strong upregulation, as it does not appear to be downregulated relative to the MS cohort and the healthy cohort. Similarly, the protein CNDP1 presents a similar challenge.

Among the notable mentions in the top ten results, BMP4, a growth factor known to be involved in SS and explored as a therapeutic target (Hu et al., 2020), stands out. Additionally, the von Willebrand factor, a plasma glycoprotein and a well-known marker for vascular damage, has been linked to SS and other rheumatic disorders when patients exhibit renal involvement and comorbidities (Yada et al., 2020). However, in this thesis project, it is unexpectedly downregulated, contrary to previous findings.

Table 6. *Shows the top ten proteins picked by the differential expression (DE) model limma (with rheumatic disease as control group) and the machine learning (ML) model glmnet (using all disease groups). It also includes the full protein names and the number of proteins picked by both models.*

Rank	DE (limma)	Full protein name	ML (glmnet)	Full protein name
1	LTA4H	*Leukotriene A4 hydrolase*	LAG3	*Lymphocyte activating 3*
2	BMP4	*Bone morphogenetic protein 4*	LTA4H	*Leukotriene A4 hydrolase*
3	ITM2A	*Integral membrane protein 2A*	IDS	*Iduronate 2-sulfatase*
4	CNDP1	*Carnosine dipeptidase 1*	BMP4	*Bone morphogenetic protein 4*
5	VAT1	*Vesicle mediated amine transport 1*	IL34	*Interleukin 34*
6	CCN1	*Cellular communication network factor 1*	GFRA2	*GNDF family receptor alpha 2*
7	CES3	*Carboxylesterase 3*	KLK8	*Kallikrein related peptidase 8*
8	CCL2	*C-C motif chemokine ligand 2*	TNF	*Tumor necrosis factor*
9	ACTA2	*Actin alpha 2, smooth muscle*	CNDP1	*Carnosine dipeptidase 1*
10	ITIH3	*Inter-alpha-trypsin inhibitor heavy chain 3*	VWF	*Von Willebrand factor*
Overlap: 3 proteins				

3.6 Systemic Lupus Erythematosus

The SLE cohort data exhibited the most inconsistency, both across analysis techniques and case-control variations, as depicted in the heatmap in *figure 18*. One peculiar observation is the strong overlap between MWU healthy and limma disease, as named in the figure. The inconsistency can likely be attributed to the presence of heavy outliers in this cohort.

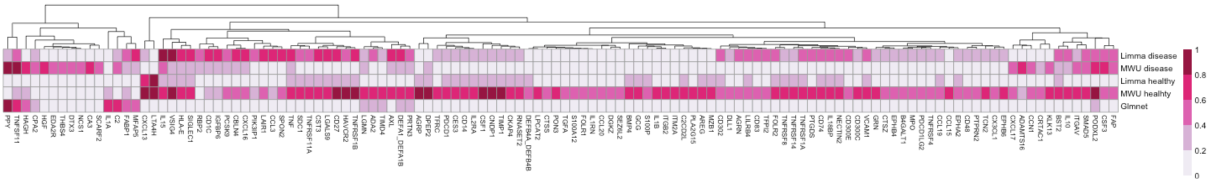


Figure 18. *A heatmap plot of the SLE cohort showing some proteins on the x-axis and methods together with case-control combination on the y-axis. The color scale represents the significance scoring by the method, either adjusted p-value or importance score, with 1 representing the highest significance. Limma/MWU disease: other rheumatic diseases were used as control. Limma/MWU healthy: the healthy cohort was used as control.*

Figure 19 presents the results from the limma and glmnet analyses, revealing a two-protein overlap between them, as indicated in *table 7*. Both models identified certain proteins with prominent outliers, including C2, TAFA5, and SLITRK2. Removing these outliers would likely improve the reliability of the findings. One of the overlapping proteins, MFAP5, has been indirectly associated with SLE through its involvement in inflammatory responses and the stimulation of cytokine secretion (Milwid et al., 2014).

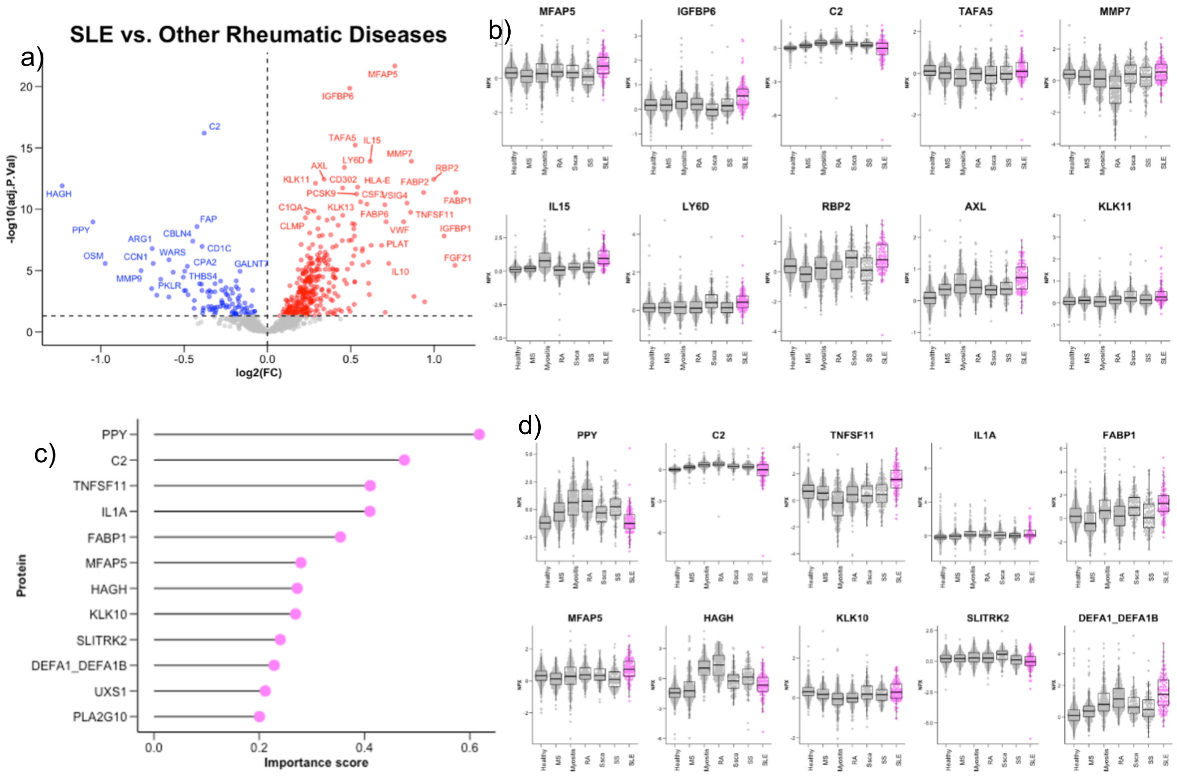


Figure 19. *a) Volcano plot of the resulting p-values when running limma for the SLE cohort as case and the remaining rheumatic diseases as control. b) Boxplots of the top ten most significantly differentially expressed proteins according to the limma analysis with NPX values on the y-axis and cohort names on the x-axis. c) Lollipop plot of the top, most important, features and their importance scores as assigned by the glmnet model. d) Boxplots of the top ten most important features assigned by the glmnet model. The SLE cohort is shown in pink.*

AXL, another selected protein, has been found to be elevated in the serum of SLE patients and is involved in regulating inflammatory cytokine release, among other functions (Orme et al., 2016). Inhibiting the AXL signaling pathways has been proposed as a potential treatment in SLE (Zhen et al., 2018). IL15, a cytokine that stimulates T-lymphocyte proliferation, has been shown to be elevated in sera of SLE patients. One study reported 38% elevation in SLE patients compared to the control group, where no elevations of protein was observed (Aringer et al., 2001). TNFSF11, is another cytokine that enhances the ability of dendritic cells to stimulate naive T-cell proliferation. Lastly, PPY, chosen by glmnet, is a pancreatic hormone and acts as a regulator of pancreatic and gastrointestinal functions (HPA, n.d.) and it has not been associated with SLE in previous literature.

Table 7. Shows the top ten proteins picked by the differential expression (DE) model limma (with rheumatic disease as control group) and the machine learning (ML) model glmnet (using all disease groups). It also includes the full protein names and the number of proteins picked by both models.

Rank	DE (limma)	Full protein name	ML (glmnet)	Full protein name
1	MFAP5	*Microfibril associated protein 5*	PPY	*Pancreatic polypeptide*
2	IGFBP6	*Insulin like growth factor binding protein 6*	C2	*Complement C2*
3	C2	*Complement C2*	TNFSF11	*TNF superfamily member 11*
4	TAFA5	*TAFA chemokine like family member 5*	IL1A	*Interleukin 1 alpha*
5	MMP7	*Matric metallopeptidase 7*	FABP1	*Fatty acid binding protein 1*
6	IL15	*Interleukin 15*	MFAP5	*Microfibril associated protein 5*
7	LY6D	*Lymphocyte antigen 6 family member D*	HAGH	*Hydroxyacylglutathione hydrolase*
8	RBP2	*Retinol binding protein 2*	KLK10	*Kallikrein related peptidase 10*
9	AXL	*AXL receptor tyrosine kinase*	SLITRK2	*SLIT and NTRK like family member 2*
10	KLK11	*Kallikrein related peptidase 11*	DEFA1/1B	*Defensin alpha 1/Dephensin alpha 1B*
Overlap: 2 proteins				

Discussion

The objective of this project was to identify potential biomarkers using two statistical methods: linear regression modeling for differential expression analysis and generalized linear models with elastic net penalty for disease classification. The limma model focuses on comparing quantitative expression differences using statistical inferences, assuming linear relationships between expression levels and experimental conditions. It employs a feature-wise, binomial approach with global variance analysis for information borrowing, and incorporates weights for increased statistical power and accuracy. On the other hand, the glmnet model can analyze all data and features simultaneously in a multinomial manner, making it well-suited for high-dimensional, topological data variations that are challenging to uncover with limma. By examining the data from different perspectives, limma and glmnet complement each other and provide more confidence in the validity of proteins detected by both methods as potential biomarkers.

The study successfully identified ten potential biomarkers per disease and model. In some cases, the top expressed proteins overlapped with the most important classification features. However, there are several other desirable attributes in a biomarker, such as the magnitude of difference in expression for one disease group compared to all others and the fold change. These attributes are crucial for robustness against batch and cohort effects. The greater the variance and spread of NPX values within the groups, the more challenging it becomes to distinguish one disease from another. With the exception of MS and Sjögren's syndrome, all other diseases had a minimum of two to three top proteins that satisfied these criteria. SS exhibited less prominent differentiation, likely due to biological factors. The other rheumatic disorders, with joint and organ involvement, likely display larger physiological differences that are detectable in plasma.

Most of the proteins identified in this project had previous mentions in the literature, ranging from biomarker studies and pathogenetic associations to being the subject of a clinical trials. While these proteins had been identified other bodily fluids, tissue samples or animal models, their discovery association with disease in this project highlights the power of the proximity extension assay technique in detecting low-abundance proteins and capturing early signs of disease symptoms.

Several challenges were encountered in this project, including the effects of poor control group selection, the lack of paraclinical validation for diagnoses, and skewed data. The analysis of the healthy control group often deviated from the inter-disease analysis in selecting significant proteins. This discrepancy was particularly evident in the analysis of the MS cohort, the limma identified proteins as significantly up-regulated when compared to the healthy cohort and significantly down-regulated compared to the autoimmune disease cohort. This created ambiguity as to whether the deviations in the healthy cohort were due to batch effects or if the MS cohort was biologically different from the rheumatic cohorts. In contrast, the glmnet model managed to identify three proteins with clear up- or down-regulation compared to all other cohorts. To address this issue in future studies, it is essential to incorporate a healthy control group from the same geographical region as the disease cases, sampled according to the same protocol, and processed in the same batch.

Subgroup analysis also presented challenges, stemming from uneven sample distribution and the absence of significant biological differences in some cases. Autoimmune diseases are diagnosed based on criteria that evolve over time. While RA and myositis subgroups were identified using serological data, the MS subgroups lacked confirmation from paraclinical markers due to the lack thereof. Furthermore, some of the identified proteins for both RA and myositis exhibited subgrouping behavior unrelated to the predefined clinical subgroups, suggesting avenues for further investigation.

The distribution of the sample data was skewed and contained numerous outliers, more prevalent in some disease groups than others, which impacted both the machine learning model and the linear regression model. For example, in SLE the selection of SLITRK2, C2 and TAFA5 proteins was based on one major outlier in each assay. This is most likely why there was a large discrepancy between MWU, limma and glmnet when it came to protein importance for the SLE cohort.

To improve future analyses, greater attention should be given to removing outliers, and exploring alternative scaling or normalization methods. Robust scaling, which considers data skewedness and is more resilient to outliers, could be a viable alternative. While z-score normalization improved the quality of the machine learning model outcome, it assumes a Gaussian distribution, which may explain instances where significant proteins were detected in some runs but received an importance score of zero upon changing parameters or conducting class balancing. Scaling should also positively impact the limma analysis, which performed reasonably well in its current form. The heterogeneity in data quality and magnitude of NPX values can vary exponentially between assays due to dilution differences and running in different panels, making it difficult for the regression models to handle despite their statistical robustness. This emphasizes the importance of careful preanalytical data wrangling when working with large biological datasets, a concept well-known in the field of "Big Data".

Furthermore, future studies should consider testing different hyperparameter tunings to examine their effects on the model. It is important to make the model more robust and the results more reproducible. Exploring the classification performance with a smaller set of features, such as selecting the top five to ten features per disease, can provide insights into the model's ability to classify disease accurately. Understanding the inner workings and decision-making processes of the model will be pivotal for future investigations.

Despite the potential for improvement in the preanalytical stages and the glmnet classification model, the overall workflow pipeline used in this project proved sufficient and will serve well in future analyses and research. The pipeline has been previously tested by the HPA in the cancer study, which was the first to launch on the HDBA, and it was also employed by two other master's students during the spring, who worked on different disease cohorts.

Future Perspectives

An updated version of the Olink® Explore platform, capable of running 3000 assays, has recently been released by Olink Proteomics. In an upcoming study, the HPA plans to assess the reproducibility of the results obtained from this study and other disease atlas projects by running new samples from a selection of the already analyzed disease cohorts on the new explore platform. If the results are reproducible, the subsequent step would involve initiating a project for quantitative analysis using plasma samples from both healthy individuals and those with diseases. Once the quantitative intervals are confirmed, phase I clinical trials can be initiated, progressing further into the development of potential biomarkers.

The results from this study will be made publicly available on the HDBA in the near future. These findings are anticipated not only to inspire new biomarker testing but hopefully offer valuable insights for pathologists and geneticists to further research and enhance our understanding of disease mechanisms. Ultimately, this work aims to contribute to the advancement of personalized medicine, leading to improved care and quality of life for individuals with autoimmune disease. Early diagnosis plays a key role in effectively managing these diseases, preventing them from becoming debilitating or fatal. The identification of reliable biomarkers is essential for the development of better therapeutic interventions. Furthermore, readily accessible biomarkers used for diagnosis and care indicators are vital in ensuring equitable healthcare conditions worldwide.

References

Alessandri, C., Bombardieri, M., Scrivo, R., Viganego, F., Conti, F., De Luca, N., Riccieri, V., Valesini, G. (2003). Anti-carbonic Anhydrase II Antibodies in Systemic Sclerosis: Association with Lung Involvement, Autoimmunity, 36:2, 85-89, DOI: 10.1080/0891693031000079239

Andersen, O. (2012). From the Gothenburg cohort to the Swedish multiple sclerosis registry. Acta Neurologica Scandinavica. Supplementum; 195: 13–19. https://doi.org/10.1111/ane.12023

Anderson, N. L., Anderson, N. G. (2002). The Human Plasma Proteome. *Mol Cell Proteomics*; 1(11): 845–867. https://doi.org/10.1074/mcp.R200007-MCP200

Aletaha, D., Neogi, T., Silman, A. J., Funovits, J., Felson, D. T., Bingham, C. O., 3rd, Birnbaum, N. S., Burmester, G. R., Bykerk, V. P., Cohen, M. D., Combe, B., Costenbader, K. H., Dougados, M., Emery, P., Ferraccioli, G., Hazes, J. M., Hobbs, K., Huizinga, T. W., Kavanaugh, A., Kay, J., Hawker, G. (2010). Rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. *Arthritis Rheum*; 62(9): 2569–2581. https://doi.org/10.1002/art.27584

Araújo, F.C., Camargo, C.Z. & Kayser, C. (2017). Validation of the ACR/EULAR classification criteria for systemic sclerosis in patients with early scleroderma. *Rheumatol Int*; 37: 1825–1833 https://doi.org/10.1007/s00296-017-3787-1

Aringer, M., Costenbader, K., Daikh, D., Brinks, R., Mosca, M., Ramsey-Goldman, R., Smolen, J.S., Wofsy, D., Boumpas, D.T., Kamen, D.L., Jayne, D., Cervera, R., Costedoat-Chalumeau, N., Diamond, B., Gladman, D.D., Hahn, B., Hiepe, F., Jacobsen, S., Khanna, D., Lerstrøm, K., Massarotti, E., McCune, J., Ruiz-Irastorza, G., Sanchez-Guerrero, J., Schneider, M., Urowitz, M., Bertsias, G., Hoyer, B.F., Leuchten, N., Tani, C., Tedeschi, S.K., Touma, Z., Schmajuk, G., Anic, B., Assan, F., Chan, T.M., Clarke, A.E., Crow, M.K., Czirják, L., Doria, A., Graninger, W., Halda-Kiss, B., Hasni, S., Izmirly, P.M., Jung, M., Kumánovics, G., Mariette, X., Padjen, I., Pego-Reigosa, J.M., Romero-Diaz, J., Rúa-Figueroa Fernández, Í., Seror, R., Stummvoll, G.H., Tanaka, Y., Tektonidou, M.G., Vasconcelos, C., Vital, E.M., Wallace, D.J., Yavuz, S., Meroni, P.L., Fritzler, M.J., Naden, R., Dörner, T. and Johnson, S.R. (2019). European League Against Rheumatism/American College of Rheumatology Classification Criteria for Systemic Lupus Erythematosus. *Arthritis Rheumatol*; 71: 1400–1412. https://doi.org/10.1002/art.40930

Aringer, M., Stummvoll, G. H., Steiner, G., Köller, M., Steinger, W., Höffler, E., Hiesberger, H., Smolen, J. S., Graninger, W. B. (2001). Serum interleukin-15 is elevated in systemic lupus erythematosus. *Rheumatology*; 40(8): 876–881. https://doi.org/10.1093/rheumatology/40.8.876

Baecher-Allen, C., Kaskow, B. J., Weiner, H. L. (2018). Multiple sclerosis: Mechanisms and immunotherapy. *Neuron*; 97(4): 742–768. https://doi.org/10.1016/j.neuron.2018.01.021

Bernatsky, S., Boivin, J. F., Joseph, L., Manzi, S., Ginzler, E., Gladman, D. D., Urowitz, M., Fortin, P. R., Petri, M., Barr, S., Gordon, C., Bae, S. C., Isenberg, D., Zoma, A., Aranow, C., Dooley, M. A., Nived, O., Sturfelt, G., Steinsson, K., Alarcón, G., Senécal, J. L., Zummer, M., Hanly, J., Ensworth, S., Pope, J., Edworthy, S., Rahman, A., Sibley, J., El-Gabalawy, H., McCarthy, T., St. Pierre, Y., Clarke, A., Ramsey-Goldman, R. (2006). Mortality in systemic lupus erythematosus. *Arthritis & Rheumatism*; 54: 2550–2557. https://doi.org/10.1002/art.21955

Benjamini, Y., Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J Royal Stat Society Series B (Methodological)*; 57(1): 289—300. https://www.jstor.org/stable/2346101

Capitanio, D., Vasso, M., De Palma, S., Fania, C., Torretta, E., Cammarata, F. P., Magnaghi, V., Procacci, P., Gelfi, C. (2015). Specific protein changes contribute to the differential muscle mass loss during ageing. *Proteomics;* 16(4): 645—656. https://doi.org/10.1002/pmic.201500395

Chrabaszcz, M., Małyszko, J., Sikora, M., Alda-Malicka, R., Stochmal, A., Matuszkiewicz-Rowińska, J., Rudnicka, L. (2020). Renal involvement in systemic sclerosis: An update. *Kidney Blood Press Res*; 45(4): 532–548. https://doi.org/10.1159/000507886

Deane, K.D. and Holers, V.M. (2021), Rheumatoid Arthritis Pathogenesis, Prediction, and Prevention: An Emerging Paradigm Shift. *Arthritis Rheumatol*; 73: 181–193. https://doi.org/10.1002/art.41417

DiSano, K. D., Gilli, F., Pachner, A. R. (2020). Intrathecally produced CXCL13: A predictive biomarker in multiple sclerosis. *Multiple Sclerosis J - Exp Transl Clin*; 6(4). doi:10.1177/2055217320981396

Dörner, T., Furie, R. (2019). Novel paradigms in systemic lupus erythematosus. *The Lancet*; 393(10188): 2344–2358. https://doi.org/10.1016/S0140-6736(19)30546-X

EIRA Sweden. (2020, Mar 09). *Welcome to EIRA.* EIRA Sweden. Retrieved on 26 of May, 2023 from: https://www.eirasweden.se/index1.htm

EMA. (n.d.). *Biomarker*. EMA. Retrieved on 22 of May, 2023 from: https://www.ema.europa.eu/en/glossary/biomarker

Engebretsen, S., Bohlin, J. (2019). Statistical predictions with glmnet. *Clin Epigenet*; 11: 123. https://doi.org/10.1186/s13148-019-0730-1

FDA. (2022, Sep 06). *Focus area: Biomarkers.* Retrieved on 22 of May, 2023 from: https://www.fda.gov/science-research/focus-areas-regulatory-science-report/focus-area-biomarkers#:~:text=Biomarkers%20are%20characteristics%20that%20are,or%20intervention%2C%20including%20therapeutic%20interventions

Finckh, A., Gilbert, B., Hodkinson, B., Bae, S. C., Thomas, R., Deane, K. D., Alpizar-Rodruigez, D., Lauper, K. (2022). Global epidemiology of rheumatoid arthritis. *Nat Rev Rheumatol*; 18: 591–602. https://doi.org/10.1038/s41584-022-00827-y

Friedman, J. H., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Soft*; 33(1): 1–22. https://doi.org/10.18637/jss.v033.i01

Foell, D., Wittowski, H., Hammerschmidt, I., Wulffraat, N., Schmeling, H., Frosch, M., Horneff, G., Kuis, W., Sorg, C., Roth, J. (2004). Monitoring neutrophil activatoin in juvenile arthritis by S100A12 serum concentrations. *Arthritis & Rheumatism*; 50(4): 1286—1295. https://doi.org/10.1002/art.20125

Goutte, C., Gaussier, E. (2005). A Probabilistic Interpretation of Precision, Recall and *F*-Score, with Implication for Evaluation. In: Losada, D.E., Fernández-Luna, J.M. (2005). *Advances Information Retrieval*. ECIR 2005. Lecture Notes in Computer Science vol 408. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-31865-1_25

Haran, M., Mirkin, V., Braester, A., Harpaz, N., Shevetz, O., Shtreiter, M., Greenberg, S., Mordich, O., Amram, O., Binsky-Ehrenreich, I., Marom, A., Shachar, I., Herishanu, Y., Ruchlemer, R., Berrebi, A., Valinsky, L., Shtalrid, M., & Shvidel, L. (2018). A phase I-II clinical trial of the anti-CD74 monoclonal antibody milatuzumab in frail patients with refractory chronic lymphocytic leukaemia: A patient based approach. *British J Haematology*; 182(1): 125–128. https://doi.org/10.1111/bjh.14726

Hastie, T., Quian, J., Tay, K. (2023, Mar 22). *An introduction to glmnet.* R-project. Retrieved on May 21, 2023 from: https://cloud.r-project.org/web/packages/glmnet/vignettes/glmnet.pdf

Hernández-Molina, G., Avila-Casado, C., Nuñez-Alvarez, C., Cárdenas-Velázquez, F., Hernández-Hernández, C., Calderillo, M. L., Marroquín, V., Recillas-Gispert, C., Romero-Díaz, J., Sánchez-Guerrero, J. (2015). Utility of the American–European Consensus Group and American College of Rheumatology Classification Criteria for Sjögren's syndrome in patients with systemic autoimmune diseases in the clinical setting. *Rheumatology*; 54(3): 441–448. https://doi.org/10.1093/rheumatology/keu352

HPA. (n.d.). *ADGRG2.* The Human Protein Atlas. Retrieved on Mar 15, 2023 from: https://www.proteinatlas.org/ENSG00000173698-ADGRG2%20S100A12%20The%20S100A12

HPA. (n.d.). *CA4.* The Human Protein Atlas. Retrieved on Mar 15, 2023 from: https://www.proteinatlas.org/ENSG00000167434-CA4

HPA. (n.d.). *CTSV.* The Human Protein Atlas. Retrieved on May 25, 2023 from: https://www.proteinatlas.org/ENSG00000136943-CTSV

HPA. (n.d.). *CXCL10.* The Human Protein Atlas. Retrieved on Mar 15, 2023 from: https://www.proteinatlas.org/ENSG00000169245-CXCL10

HPA. (n.d.). *EDIL3.* The Human Protein Atlas. Retrieved on May 25, 2023 from: https://www.proteinatlas.org/ENSG00000164176-EDIL3

HPA. (n.d.). *ITGAM.* The Human Protein Atlas. Retrieved on May 25, 2023 from: https://www.proteinatlas.org/ENSG00000169896-ITGAM

HPA. (n.d.). *KLK13.* The Human Protein Atlas. Retrieved on Mar 15, 2023 from: https://www.proteinatlas.org/ENSG00000167759-KLK13

HPA. (n.d.). *KLK14.* The Human Protein Atlas. Retrieved on Mar 15, 2023 from: https://www.proteinatlas.org/ENSG00000167749-KLK4

HPA. (n.d.). *NPPB.* The Human Protein Atlas. Retrieved on Mar 15, 2023 from: https://www.proteinatlas.org/ENSG00000120937-NPPB

HPA. (n.d.). *OMG.* The Human Protein Atlas. Retrieved on May 25, 2023 from: https://www.proteinatlas.org/ENSG00000126861-OMG

HPA. (n.d.). *PPY.* The Human Protein Atlas. Retrieved on May 25, 2023 from: https://www.proteinatlas.org/ENSG00000108849-PPY

HPA. (n.d.). *S100A12.* The Human Protein Atlas. Retrieved on Mar 15, 2023 from: https://www.proteinatlas.org/ENSG00000163221-S100A12

HPA. (n.d.). *SEMA4C.* The Human Protein Atlas. Retrieved on Mar 15, 2023 from: https://www.proteinatlas.org/ENSG00000168758-SEMA4C

HPA. (n.d.). *SFTPD.* The Human Protein Atlas. Retrieved on Mar 15, 2023 from: https://www.proteinatlas.org/ENSG00000133661-SFTPD

Hu, L., Xu, J., Wu, T. *et al.* (2020). Depletion of ID3 enhances mesenchymal stem cells therapy by targeting BMP4 in Sjögren's syndrome. *Cell Death Dis*; 11:172 https://doi.org/10.1038/s41419-020-2359-6

Hughes, M., Pauling, J. D., Armstrong-James, L., Denton, C. P., Galdas, P., Flurey, C. (2020). Gender-related differences in systemic sclerosis, *Autoimmunity Rev*; 19(4): 102494. https://doi.org/10.1016/j.autrev.2020.102494

IBM. (n.d.). *What is machine learning?* IBM. Retrieved on May 22, 2023 from: https://www.ibm.com/topics/machine-learning

Jonsson, R. (2022). Disease mechanisms in Sjögren's syndrome: What do we know? *Scandinavian J Immunol*; 95(3): e13145. https://doi.org/10.1111/sji.13145

Järvinen, T. M., Hellquist, A., Koskenmies, S., Einarsdottir, E., Panelius, J., Hasan, T., Julkunen, H., Padyukov, L., Kvarnström, M., Wahren-Herlenius, M., Nyberg, F., D'Amato, M., Kere, J., & Saarialho-Kere, U. (2010). Polymorphisms of the ITGAM gene confer higher risk of discoid cutaneous than of systemic lupus erythematosus. *PloS one*; 5(12): e14212. https://doi.org/10.1371/journal.pone.0014212

Kagan L. J. (1977). Myoglobinemia in inflammatory myopathies. *JAMA*; 237(14): 1448–1452. https://doi.org/10.1001/jama.237.14.1448

Kim, S., Swaminathan, S., Inlow, M., Risacher, S. L., Nho, K., Shen, L., Foroud, T. M., Petersen, R. C., Aisen, P. S., Soares, H., Toledo, J. B., Shaw, L. M., Trojanowski, J. Q., Weiner, M. W., McDonald, B. C., Farlow, M. R., Ghetti, B., Saykin, A. J., & Alzheimer's Disease Neuroimaging Initiative (ADNI). (2013). Influence of genetic variation on plasma protein levels in older adults using a multi-analyte panel. *PloS One;* 8(7): e70269. https://doi.org/10.1371/journal.pone.0070269

Korotkova, M., Lundberg, I. (2014). The skeletal muscle arachidonic acid cascade in health and inflammatory disease. *Nat Rev Rheumatol*; 10: 295–303. https://doi.org/10.1038/nrrheum.2014.2

Köhler, B. M., Günther, J., Kaudewitz, D., Lorenz, H. M. (2019). Current Therapeutic Options in the Treatment of Rheumatoid Arthritis. *J Clin Med*; 8(7): 938. https://doi.org/10.3390/jcm8070938

Lazar, S., Kahlenberg, J. M. (2023). Systemic lupus erythematosus: New diagnostic and therapeutic approaches. *Annu Rev Med*; 74: 339–352. https://doi.org/10.1146/annurev-med-043021-032611

Lundberg, I. E., Fujimoto, M., Vencovsky, J., Aggarwal, R., Holmqvist, M., Christopher-Stine, L., Mammen, A. L., & Miller, F. W. (2021). Idiopathic inflammatory myopathies. *Nat Rev Dis Primers*; *7*(1): 86. https://doi.org/10.1038/s41572-021-00321-x

Lundberg, I. E., Tjärnlund, A., Bottai, M., Werth, V. P., Pilkington, C., de Visser, M., Alfredsson, L., Amato, A. A., Barohn, R. J., Liang, M. H., Singh, J. A., Aggarwal, R., Arnardottir, S., Chinoy, H., Cooper, R. G., Dankó, K., Dimachkie, M. M., Feldman, B. M., Garcia-De La Torre, I., Gordon, P., the International Myositis Classification Criteria Project Consortium, the Euromyositis register, and the Juvenile Dermatomyositis Cohort Biomarker Study and Repository (UK and Ireland). (2017). European league against rheumatism/American college of rheumatology classification criteria for adult and juvenile idiopathic inflammatory myopathies and Their Major Subgroups. *Arthritis Rheumatol;* 69(12): 2271–2282. https://doi.org/10.1002/art.40320

Lundberg, I., de Visser, M., Werth, V. (2018). Classification of myositis. *Nat Rev Rheumatol*; 14: 269–278. https://doi.org/10.1038/nrrheum.2018.41

Mann, H. B., Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Ann Math Stat*; 18(1): 50—60. http://www.jstor.org/stable/2236101

Merino-Jiménez, C., García-Cruz, C., Aragón, J., Siqueiros-Márquez, L., Montañez, C. (2019). *Heat Shock Proteins Involved in Neuromuscular Pathologies.* In: Asea, A., Kaur, P. (eds) Heat Shock Proteins in Signaling Pathways. Heat Shock Proteins, vol 17. Springer, Cham. https://doi.org/10.1007/978-3-030-03952-3_21

Jack M Milwid, Jessica S Elman, Matthew Li, Keyue Shen, Arjun Manrai, Aaron Gabow, Joshua Yarmush, Yunxin Jiao, Anne Fletcher, Jungwoo Lee, Michael J Cima, Martin L Yarmush, Biju Parekkadan (2014). Enriched Protein Screening of Human Bone Marrow Mesenchymal Stromal Cell Secretions Reveals MFAP5 and PENK as Novel IL-10 Modulators. *Mol Therapy*; 22(5): 999—1007. https://doi.org/10.1038/mt.2014.17

Moinzadeh, P., Kuhr., K, Siegert, E., Ladner, U. M., Riemekasten, G., Günther, C., Kötter, I., Henes, J., Blank, N., Zeidler, G., Pfeiffer, C., Juche, A., Jandova, I., Ehrchen, J., Schmalzing, M., Susok, L., Schmeiser, T., Sunderkoetter, C., Distler, J. H. W., Worm, M., Kreuter, A., Krieg, T., Hunzelmann, N., Registry of the German Network for Systemic Scleroderma. (2020). Older age onset of systemic sclerosis – accelerated disease progression in all disease subsets, *Rheumatol*; 59(11): 3380–3389. https://doi.org/10.1093/rheumatology/keaa127

Moore, D. F., & Steen, V. D. (2021). Overall mortality. *J Scleroderma Relat Disord*; 6(1): 3–10. https://doi.org/10.1177/2397198320924873

Moutsopoulos, H. M. (2021). Autoimmune rheumatic diseases: One or many diseases? *J Transl Autoimmun;* 4: 100129. https://doi.org/10.1016/j.jtauto.2021.100129

National Institutes of Health. (2022, Okt). *Systemic Lupus Erythematosus (Lupus)*. NIH. Retrieved on May 3, 2023 from: https://www.niams.nih.gov/health-topics/lupus

Nordquist, H., & Jamil, R. T. (2022). *Biochemistry, HLA Antigens*. StatPearls Publishing. PMID: 31536268

Olink. (2016, Mar 29). *What is the high dose hook effect?*. Olink. Retrieved on March 27, 2023 from:https://olink.com/faq/what-is-the-high-dose-hook-effect/#:~:text=A%20high%20dose%20hook%20effect,lead%20to%20misinterpretation%20of%20results

Jacob J. Orme, Yong Du, Kamala Vanarsa, Jessica Mayeux, Li Li, Azza Mutwally, Cristina Arriens, Soyoun Min, Jack Hutcheson, Laurie S. Davis, Benjamin F. Chong, Anne B. Satterthwaite, Tianfu Wu, Chandra Mohan. (2016). Heightened cleavage of Axl receptor tyrosine kinase by ADAM metalloproteases may contribute to disease pathogenesis in SLE. *Clin Immunol*; 169: 58—68. https://doi.org/10.1016/j.clim.2016.05.011

Ostendorf, L., Dittert, P., Biesen, R., Duchow, A., Stiglbauer, V., Ruprecht, K., Bellmann-Strobl, J., Seelow, D., Stenzel1, W., Niesner, R. A., Hauser, A. E., Friedemann, P., Radbruch, H. (2021). SIGLEC1 (CD169): a marker of active neuroinflammation in the brain but not in the blood of multiple sclerosis patients. *Sci Rep.* 11: 10299. https://doi.org/10.1038/s41598-021-89786-0

Parker, M. J. S., Lilleker, J. B., Chinoy, H. (2022). Adult idiopathic inflammatory myopathies. *Medicine*; 50(1): 70–75. https://doi.org/10.1016/j.mpmed.2021.10.011

Perelas, A., Silver, R. M., Arrossi, A. V., Highland, K. B. (2020). Systemic sclerosis-associated interstitial lung disease. *The Lancet Respiratory Medicine*; 8(3): 304–320. https://doi.org/10.1016/S2213-2600(19)30480-1

Pratt, A. G., Isaacs, J. D. (2014). Seronegative rheumatoid arthritis: Pathogenetic and therapeutic aspects. *Best Practice Res Rheumatol*; 28(4): 651–659. https://doi.org/10.1016/j.berh.2014.10.016

Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., O'Sullivan, J. M. (2022). A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. *Front Bioinform;* 2: 927312. DOI: 10.3389/fbinf.2022.927312

Ray, P. D., Huang, B. W., & Tsuji, Y. (2012). Reactive oxygen species (ROS) homeostasis and redox regulation in cellular signaling. *Cell Signal*; 24(5): 981–990. https://doi.org/10.1016/j.cellsig.2012.01.008

Reich, D. S., Lucchinetti, C. F., Calabresi, P. A. (2018). Multiple Sclerosis. *N Engl J Med;* 378(2): 169–180. https://doi.org/10.1056/NEJMra1401483

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*; 43(7): 47. https://doi.org/10.1093/nar/gkv007

Royston, P. (1992). Approximating the Shapiro-Wilk W-test for non-normality. *Stat Comput*; 2: 117—119. https://doi.org/10.1007/BF01891203

Sangha, O. (2000). Epidemiology of rheumatic diseases. *Rheumatology;* 39(2): 3–12. https://doi.org/10.1093/rheumatology/39.suppl_2.3

Shapiro, S. S., Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*; 52: 591—611. https://doi.org/10.2307/2333709

Shastry, K.A., Sanjay, H.A. (2020). *Machine Learning for Bioinformatics.* In: Srinivasa, K., Siddesh, G., Manisekhar, S. (eds). Statistical Modelling and Machine Learning Principles for Bioinformatics Techniques, Tools, and Applications. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-15-2445-5_3

Shiboski, C. H., Shiboski, S. C., Seror, R., Criswell, L. A., Labetoulle, M., Lietman, T. M., Rasmussen, A., Scofield, H., Vitali, C., Bowman, S. J., Mariette, X., & International Sjögren's Syndrome Criteria Working Group. (2017). 2016 American College of Rheumatology/European League Against

Rheumatism Classification Criteria for Primary Sjögren's Syndrome: A Consensus and Data-Driven Methodology Involving Three International Patient Cohorts. *Arthritis Rheumatol*; 69(1): 35–45. https://doi.org/10.1002/art.39859

Simonsen, C. S., Flemmen, H. Ø., Lauritzen, T., Berg-Hansen, P., Moen, S. M., & Celius, E. G. (2020). The diagnostic value of IgG index versus oligoclonal bands in cerebrospinal fluid of patients with multiple sclerosis. *Mult Scler J Exp Transl Clin;* 6(1): 2055217319901291. https://doi.org/10.1177/2055217319901291

Singh, J. A., Solomon, D. H., Dougados, M., Felson, D., Hawker, G., Katz, P., Paulus, H., Wallace, C., & Classification and Response Criteria Subcommittee of the Committee on Quality Measures, American College of Rheumatology. (2006). Development of classification and response criteria for rheumatic diseases. *Arthritis and rheumatism*; 55(3): 348–352. https://doi.org/10.1002/art.22003

Siwiec, A., Majdan, M. (2015). The role of the PD-1 protein in pathogenesis of autoimmune diseases, with particular consideration of rheumatoid arthritis and systemic lupus erythematosus. *Postepy Higieny i Medycyny Doswiadczalnej (Online)*. 69: 534—542. DOI: 10.5604/17322693.1150784

Styrkarsdottir, U., Lund, S. H., Thorlefsson, G., Saevarsdottir, S., Gudbjartsson, D. F. THorsteinsdottis, U., Stefansson, K. (2022). *Arthitis Rheumatol*; 75(4): 544—552. https://doi.org/10.1002/art.42376

Su, X., Yan, X. and Tsai, C.-L. (2012). Linear regression. *WIREs Comp Stat*, 4: 275–294. https://doi.org/10.1002/wics.1198

Tajuddin, S. M., Schick, U. M., Eicher, J. D., Chami, N., Giri, A., Brody, J. A., Hill, W. D., Kacprowski, T., Li, J., Lyytikäinen, L. P., Manichaikul, A., Mihailov, E., O'Donoghue, M. L., Pankratz, N., Pazoki, R., Polfus, L. M., et al. (2016). Large-Scale Exome-wide Association Analysis Identifies Loci for White Blood Cell Traits and Pleiotropy with Immune-Mediated Diseases, *Am J Hum Gen*; 99(1): 22—39. https://doi.org/10.1016/j.ajhg.2016.05.003.

Tews, D. S., & Goebel, H. H. (1998). Cell death and oxidative damage in inflammatory myopathies. *Clinl Immunol Immunopathol*; 87(3): 240–247. https://doi.org/10.1006/clin.1998.4527

Thompson, A. J., Banwell, B. L., Barkhof, F., Carroll, W. M., Coetzee, T., Comi, G., Correale, J., Fazekas, F., Filippi, M., Freedman, M. S., Fujihara, K., Galetta, S. L., Hartung, H. P., Kappos, L., Lublin, F. D., Marrie, R. A., Miller, A. E., Miller, D. H., Montalban, X., Mowry, E. M., Sorensen, P. S., Tintoré, M., Traboulsee, A. L., Trojano, M., Uitdehaag, B. M. J., Vukusic, S., Waubant, E., Weinshenker, B. G., Reingold, S. C., Cohen, J. A. (2018). Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *The Lancet Neurology*; 17(2): 162–173. https://doi.org/10.1016/S1474-4422(17)30470-2

Titcomb, T. J., Bao, W., Du, Y., Liu, B., Snetselaar, L. G., & Wahls, T. L. (2022). Association of multiple sclerosis with risk of mortality among a nationally representative sample of adults in the United States. *Mult Scler J Exp Transl Clin*; 8(2): 20552173221104009. https://doi.org/10.1177/20552173221104009

Tong, Y. (2021). The comparison of limma and DESeq2 in gene analysis. *E3S Web Conf*; 271: 03058. https://doi.org/10.1051/e3sconf/202127103058

Torres-Fuentes, L., Matías-Guiu, J. A., Pytel, V., Montero-Escribano, P., Maietta, P., Álvarez, S., Gómez-Pinedo, U., Matías-Guiu, J. (2020). Variants of genes encoding TNF receptors and ligans and proteins regulating TNF activation in familial multiple sclerosis. *CNS Neruoscience Ther*; 26(11): 1178—1184. https://doi.org/10.1111/cns.13456

Trentini, A., Manfrinato, M. C., Castellazzi, M., Tamborino, C., Roversi, G., Volta, C. A., Baldi, E., Tola, M. R., Granieri, E., Dallocchio, F., Bellini, T., Fainardi, E., & Emilia-Romagna network for Multiple Sclerosis (ERMES) study group (2015). TIMP-1 resistant matrix metalloproteinase-9 is the predominant serum active isoform associated with MRI activity in patients with multiple sclerosis. *Multiple sclerosis (Houndmills, Basingstoke, England)*; 21(9): 1121–1130. https://doi.org/10.1177/1352458514560925

van den Hoogen, F., Khanna, D., Fransen, J., Johnson, S. R., Baron, M., Tyndall, A., Matucci-Cerinic, M., Naden, R. P., Medsger, T. A., Jr, Carreira, P. E., Riemekasten, G., Clements, P. J., Denton, C. P., Distler, O., Allanore, Y., Furst, D. E., Gabrielli, A., Mayes, M. D., van Laar, J. M., Seibold, J. R., Simms, R., Pope, J. E. (2013). Classification criteria for systemic sclerosis: an American College of Rheumatology/European League against Rheumatism collaborative initiative. *Arthritis and rheumatism*; 65(11): 2737–2747. https://doi.org/10.1002/art.38098

Volkmann, E. R., Andréasson, K., Smith, V. (2022). Systemic sclerosis. *The Lancet*; 401(10373): 304–318. https://doi.org/10.1016/S0140-6736(22)01692-0

Wang, L., Wang, F. S., Gershwin, M. E. (2015). Human autoimmune diseases: a comprehensive update. *J Intern Med;* 278: 369–395. https://doi.org/10.1111/joim.12395

Wang, Z. W., Zhang, L. J., Zhuang, Y., Lv, Z. F., Tan, Z. M. (2022). CKS2 and S100A12: Two novel diagnostic biomarkers for rheumatoid arthritis. https://doi.org/10.1155/2022/2431976

Wik, L., Nordberg, N., Broberg, J., Björkesten, J., Assarsson, E., Henriksson, S., Grundberg, I., Pettersson, E., Westerberg, C., Liljeroth, E., Falck, A., & Lundberg, M. (2021). Proximity extension assay in combination with next-generation sequencing for high-throughput proteome-wide analysis. *Mol Cell Proteomics: MCP*; 20: 100168. https://doi.org/10.1016/j.mcpro.2021.100168

Wu, H., Tremaroli, V., Schmidt, C., Lundqvist, A., Olsson, L. M., Krämer, M., Gummesson, A., Perkins, R., Bergström, G., Bäckhed, F. (2020). The gut microbiota in prediabetes and diabetes: A population-based cross-sectional study. *Cell Metab*; 32(3): 379–390. https://doi.org/10.1016/j.cmet.2020.06.011

Yada, N., Toshimoto, K., Kawashima, H., Yonemia, R., Nishimura, N., Tai, Y., Tsushima, E., Miyamoto, M., Ono, S., Matsumoto, M., Fujimoto, T., Nishio, K. (2020). Plasma level of von Willebran factor prpeptide at diagnosis: A marker of subsequent renal dysfunction in autoimmune rheumatic diseases. *Clin App Thrombosis/Hemostasis*; 16: 1—9. https://doi.org/10.1177/1076029620938874

Yamout, B., Sahraian, M., Bohlega, S., Al-Jumah, M., Goueider, R., Dahdaleh, M., Inshasi, J., Hashem, S., Alsharoqi, I., Khoury, S., Alkhawajah, M., Koussa, S., Al Khaburi, J., Almahdawi, A., Alsaadi, T., Slassi, E., Daodi, S., Zakaria, M., Alroughani R. (2020). Consensus recommendations for the diagnosis and treatment of multiple sclerosis: 2019 revisions to the MENACTRIMS guidelines. *Multiple Sclerosis and Related Disorders*; 37: 101459 https://doi.org/10.1016/j.msard.2019.101459

Yanaba , K., et al. (2012). Augmented production of soluble CD93 in patients with systemic sclerosis and clinical association with severity of skin sclerosis, *Brit J Dermatol*; 167(3): 542–547. https://doi.org/10.1111/j.1365-2133.2012.11020.x

Yang, J., Hamade, M., Wu, Q., Wang, Q., Axtell, R., Giri, S., Mao-Draayer, Y. (2022). Current and Future Biomarkers in Multiple Sclerosis. *Int J Mol Sci*; 23(11): 5877. https://doi.org/10.3390/ijms23115877

Zhan, Q., Zhang, J., Lin, Y., Chen, W., Fan, X., & Zhang, D. (2023). Pathogenesis and treatment of Sjogren's syndrome: Review and update. *Front Immunol*; 14: 1127417. https://doi.org/10.3389/fimmu.2023.1127417

Zhang, Z., Cheng, Y. & Liu, N.C. (2014). Comparison of the effect of mean-based method and *z*-score for field normalization of citations at the level of Web of Science subject categories. *Scientometrics*; 101:1679–1693. https://doi.org/10.1007/s11192-014-1294-7

Zhang, L., Tang, S., Ma, Y., Liu, J., Monnier, P., Li, H., Zhang, R., Yu, G., Zhang, M., Li, Y., Feng, J., Qin, X. (2022). RGMa Participates in the Blood–Brain Barrier Dysfunction Through BMP/BMPR/YAP Signaling in Multiple Sclerosis. *Front Immunol*; 13:861486. DOI: 10.3389/fimmu.2022.861486

Zhen, Y., Lee, I. J., Finkelman, F. D., Shao, W. H. (2018). Targeted inhibition of Axl receptor tyrosine kinase ameliorates anti-GBM-induced lupus-like nephritis. *J Autoim*; 93: 37—44. https://doi.org/10.1016/j.jaut.2018.06.001