biomedicines

MDPI

Article

Application of SWATH Mass Spectrometry and Machine Learning in the Diagnosis of Inflammatory Bowel Disease Based on the Stool Proteome

Elmira Shajari [1,2,3], David Gagné [1,2,3,4], Mandy Malick [1,2,3], Patricia Roy [1,2,3], Jean-François Noël [4], Hugo Gagnon [4], Marie A. Brunet [2,5], Maxime Delisle [2,6], François-Michel Boisvert [2,3] and Jean-François Beaulieu [1,2,3,*]

1 Laboratory of Intestinal Physiopathology, Faculty of Medicine and Health Sciences, Université de Sherbrooke, Sherbrooke, QC J1H 5N4, Canada
2 Centre de Recherche du Centre Hospitalier Universitaire de Sherbrooke, Sherbrooke, QC J1H 5N4, Canada
3 Department of Immunology and Cell Biology, Faculty of Medicine and Health Sciences, Université de Sherbrooke, Sherbrooke, QC J1H 5N4, Canada
4 Allumiqs, 975 Rue Léon-Trépanier, Sherbrooke, QC J1G 5J6, Canada
5 Department of Pediatrics, Faculty of Medicine and Health Sciences, Université de Sherbrooke, Sherbrooke, QC J1H 5N4, Canada
6 Department of Medicine, Faculty of Medicine and Health Sciences, Université de Sherbrooke, Sherbrooke, QC J1H 5N4, Canada
* Correspondence: jean-francois.beaulieu@usherbrooke.ca

Abstract: Inflammatory bowel disease (IBD) flare-ups exhibit symptoms that are similar to other diseases and conditions, making diagnosis and treatment complicated. Currently, the gold standard for diagnosing and monitoring IBD is colonoscopy and biopsy, which are invasive and uncomfortable procedures, and the fecal calprotectin test, which is not sufficiently accurate. Therefore, it is necessary to develop an alternative method. In this study, our aim was to provide proof of concept for the application of Sequential Window Acquisition of All Theoretical Mass Spectra-Mass spectrometry (SWATH-MS) and machine learning to develop a non-invasive and accurate predictive model using the stool proteome to distinguish between active IBD patients and symptomatic non-IBD patients. Proteome profiles of 123 samples were obtained and data processing procedures were optimized to select an appropriate pipeline. The differentially abundant analysis identified 48 proteins. Utilizing correlation-based feature selection (Cfs), 7 proteins were selected for proceeding steps. To identify the most appropriate predictive machine learning model, five of the most popular methods, including support vector machines (SVMs), random forests, logistic regression, naive Bayes, and k-nearest neighbors (KNN), were assessed. The generated model was validated by implementing the algorithm on 45 prospective unseen datasets; the results showed a sensitivity of 96% and a specificity of 76%, indicating its performance. In conclusion, this study illustrates the effectiveness of utilizing the stool proteome obtained through SWATH-MS in accurately diagnosing active IBD via a machine learning model.

Keywords: inflammatory bowel disease; IBD biomarkers; SWATH; DIA mass spectrometry; quantitative proteomics; machine learning; bioinformatics analysis; SVM; data mining

check for updates

Citation: Shajari, E.; Gagné, D.; Malick, M.; Roy, P.; Noël, J.-F.; Gagnon, H.; Brunet, M.A.; Delisle, M.; Boisvert, F.-M.; Beaulieu, J.-F. Application of SWATH Mass Spectrometry and Machine Learning in the Diagnosis of Inflammatory Bowel Disease Based on the Stool Proteome. *Biomedicines* **2024**, *12*, 333. https://doi.org/10.3390/biomedicines12020333

Academic Editor: Maria-Ioanna (Marianna) Christodoulou

Received: 7 December 2023
Revised: 17 January 2024
Accepted: 25 January 2024
Published: 1 February 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Inflammatory bowel disease (IBD) is a chronic disorder of the gastrointestinal tract that affects millions of people worldwide. It is characterized by inflammation of the intestinal mucosa, leading to symptoms such as abdominal pain, diarrhea, rectal bleeding, and weight loss [1]. During flare-ups, patients require drug treatment, such as steroids, immunosuppressants, and biological therapies, to reduce inflammation and promote healing [2]. Several other diseases and conditions can present symptoms similar to those of IBD, including celiac disease, irritable bowel syndrome (IBS), and infectious colitis [3]. However,

each of these diseases requires different treatments. Consequently, the rapid and accurate diagnosis of IBD flare-ups is essential to ensure appropriate treatment and management of this condition. This is especially true as IBD is associated with both unique and severe complications, sometimes requiring hospitalization and intestinal resection.

Currently, the gold standard for diagnosing and monitoring IBD is colonoscopy and biopsy, invasive procedures that can be uncomfortable and present risks of complications [4]. Moreover, IBD is a lifelong disease, and repeated colonoscopies are necessary for disease follow-up, representing a significant burden for patients. It is therefore necessary to develop non-invasive methods for IBD diagnosis and follow-up [5]. Stool biomarkers have emerged as a promising non-invasive approach for IBD diagnosis and monitoring because they are in direct contact with the affected area of inflammation and pathology in IBD and can be utilized repeatedly as required. Among stool biomarkers, protein biomarkers have several advantages over other molecules since they are more stable in stool samples and can provide information on the activity and severity of the disease. Calprotectin is a calcium-binding protein that is released by inflammatory cells and is highly elevated in the feces of patients with IBD [6]. Calprotectin is a common clinically used fecal biomarker to monitor disease activity and the response to treatment and to distinguish between IBD and other gastrointestinal conditions that may have similar symptoms. However, it is not always accurate, and false-positive or false-negative results can occur. Especially when the calprotectin value falls within the range of 100 to 300 μg/g, it can be challenging to predict the transition from the remission phase to the flare-up phase of IBD [7]. Given this, it is reasonable to expect that combining multiple biomarkers could enhance accuracy and sensitivity in diagnostic or research applications [8].

Recently, there have been promising developments in technology and platforms that can identify and measure a large number of targets simultaneously, such as mass spectrometry-based approaches. Mass spectrometry holds great potential for clinical proteomics, which is used for a comprehensive study of proteins in clinical samples with the aim of discovering the most relevant disease markers [9]. Data-independent acquisition (DIA) mass spectrometry enables comprehensive quantification of all detectable proteins in a sample and allows retrospective data analysis. It also has several advantages over data-dependent acquisition (DDA) for proteomic profiling, such as higher reproducibility, a lower missing value rate, and better quantification accuracy [10]. In comparison to various DIA methods [11], Sequential Window Acquisition of All Theoretical Mass Spectra (SWATH) typically provides a combination of deep proteome coverage capabilities with quantitative consistency and accuracy [11–13].

Overall, only a few published studies have used mass spectrometry (MS) analysis on human stool samples to identify protein profiles for specific pathologies, including IBD. For example, a pilot study was conducted on a cohort of 10 to discriminate between active and remission phases. However, they did not use a validation group and identified 30 differentially expressed proteins in two groups of five patients [14]. Another study was performed on a cohort of IBD patients, which utilized a spectrum analysis instead of quantitative data. Their validation cohort yielded low specificity (55%), and the standard operating procedure (SOP) for sample collection and storage in this study required dispatch to the laboratory within 2 h and freezing at −80 °C, which may not be compatible with the general constraints of a standard clinical setup [15]. Recently, Vitali et al. identified three single fecal biomarkers using 2-DIGE and MALDI-TOF/TOF MS on stool samples [16]. Among them, only RhoGDI2 showed better performance than calprotectin to discriminate control from IBD patients. However, this marker, like calprotectin, was not able to identify patients in the middle zone, encompassing those in remission and with moderate activity.

Nevertheless, these studies demonstrated the feasibility of using mass spectrometry on stool samples to identify specific biomarkers that can contribute to the diagnosis of IBD.

Alternatively, analyzing such a large DIA dataset, especially from complex samples such as stool, is challenging and necessitates advanced bioinformatics to identify reliable patterns. In this regard, machine learning (ML) and using advanced feature selection

methods have emerged as promising tools. Our hypothesis was that conducting a proteomic analysis on clinical laboratory samples that are intended for the fecal calprotectin test would enable the development of a highly sensitive and specific non-invasive stool test based on mass spectrometry. To investigate this hypothesis, we combined and applied our expertise in basic research, clinical practice, and bioinformatics to develop a precise machine learning model for the accurate diagnosis of active IBD patients from symptomatic non-IBD patients.

This study represents a significant advancement in the field by demonstrating the effectiveness of SWATH-DIA proteomic profiling in diagnosing active IBD patients from non-IBD controls. The novel integration of this proteomic approach with machine learning techniques to create a predictive model enhances the diagnostic accuracy. The model's practicality was confirmed through successful validation of a separate set of samples, achieving 96% sensitivity with a 0.96 AUC. Furthermore, the robustness of the model is evident in its ability to process data from multiple batches with different collection times, showcasing its real-world applicability. Importantly, the stool samples were obtained under clinically compatible SOP conditions, emphasizing the study's relevance to clinical practice.

2. Materials and Methods

2.1. Sample Collection and Research Ethics

A total of 123 samples was obtained from the Clinical Hematology Lab of the CIUSSS de l'Estrie-CHUS in the context of the fecal calprotectin (f-cal) testing program. The research protocol for accessing stool samples from patients that have been tested for f-cal includes a reverse consent procedure for using residual stool samples and accessing the related clinical data on the Ariane network for diagnosis. This protocol has been approved by the Research Ethics Committee of the CIUSSS de l'Estrie-CHUS (Protocol number 1991-17, 90-18; last date of approval 27 August 2023). Patients under 18 years were excluded from the study. When prescribed an f-cal test by their doctor, patients were instructed to collect a stool sample at home and bring it to the hospital within 24 h (according to the CHUS protocol, 2 h max at RT, within 24 h, but in the fridge (4 °C)). In the Hematology lab, a special device was used to collect a fixed amount of stool (~50 mg) and perform the extraction to be tested for calprotectin using ELISA. The remaining stool samples were stored frozen at −80 °C and waited for confirmation of the patient's lack of objection from the Archive Division before being stored in the lab and included in the study.

Furthermore, in our study, we excluded samples with ambiguous diagnoses, retaining only those with clear-cut diagnoses made using imaging, colonoscopy, fecal calprotectin tests, and histological data by the attending physician. The control group in our study consisted of individuals who consulted a doctor for symptoms mimicking IBD. However, subsequent tests confirmed the absence of IBD in these patients. The control group predominantly consisted of individuals with irritable bowel syndrome (IBS), and some had infectious colitis. Hence, we refer to them as symptomatic non-IBD controls.

2.2. Sample Preparation

Sample preparation was implemented as previously described [17]. Briefly, 100 mg of frozen stool specimens was solubilized in 1 mL of lysis buffer (25 mM Tris, 1% SDS, pH 7.5) and centrifuged. Then the aqueous phase between the pellet and the floating residual was recovered and stored at −80 °C until preparation for LC-MS/MS analysis. The concentration of solubilized proteins in the individual samples was measured using a BCA test. For reduction, the samples were treated with 10 mM dithiothreitol (DTT) and, for alkylation, the samples were exposed to 15 mM iodoacetamide. Subsequently, the quenching step was implemented using 10 mM DTT. The proteins were precipitated with cold acetone and methanol and digested with Trypsin/Lys-C. The cleaning and recovery of the peptides were performed with a reverse-phase Strata-X polymeric SPE sorbent column (Phenomenex, Torrance, CA, USA) according to the manufacturer's instructions. The recovered peptides were dried under nitrogen flow at 37 °C for 45 min and stored at 4 °C

until being resuspended in 20 µL of mobile phase solvent A (0.2% v/v formic acid and 3% DMSO v/v in water) before LC-MS/MS analysis.

2.3. SWATH-MS Data Acquisition

The acquisition of LC-MS/MS data was conducted at the proteomics facility located at Allumiqs Solutions in Sherbrooke, Quebec, Canada. Samples were analyzed using an Eksigent µUHPLC (Eksigent, Redwood City, CA, USA) coupled to an ABSciex TripleTOF 6600 mass spectrometer equipped with an electrospray interface with a 25 µm i.d. capillary. Data-Independent Acquisition (DIA) Sequential Window Acquisition of All Theoretical Mass Spectra (SWATH) acquisition mode was used to acquire raw data from the individual samples. The source voltage was set to 5.5 kV and maintained at 325 °C, the curtain gas was set at 35 psi, gas one was set at 27 psi, and gas two was set at 10 psi. Separation was performed on a reverse-phase Kinetex XB column with a 0.3 mm i.d., 2.6 µm particles, 150 mm (Phenomenex), which was maintained at 60 °C. Samples were injected by loop overfilling into a 5 µL loop. For the 60 min LC gradient, the mobile phase consisted of the following: solvent A (0.2% v/v formic acid and 3% DMSO v/v in water) and solvent B (0.2% v/v formic acid and 3% DMSO in EtOH) at a flow rate of 3 µL/min. DDA analyses were conducted with a 60 min LC gradient, while SWATH analyses utilized a 30 min LC gradient under the following conditions: 0 to 4 min, maintaining a constant 98%/2% solvent A/B mixture; 4 to 16 min, transitioning to a 75%/25% mixture; 16 to 21 min, transitioning to a 55%/45% mixture; 21 to 25 min, transitioning to 100% solvent B, which continued until 27 min; and 27 to 30 min for column re-equilibration. The decision to reduce the LC gradient length to 30 min for SWATH was driven by logistical considerations. To ensure optimal SWATH data quality, various combinations of parameters were assessed using variable acquisition windows for an MS scanning range from 350 to 1250 m/z. Parameters evaluated encompassed the number, width, and distribution of the SWATH windows, as well as ion accumulation times. Optimization of SWATH windows was executed using the SWATH Variable Window Calculator (Sciex), scaling window sizes across the m/z range based on the m/z intensity distribution. The selected optimized SWATH method was determined by identifying the combination that provided a minimum of 6 MS2 data points per peak while maximizing quantifiable proteins and peptides.

2.4. Spectral Library Generation

To generate an ion library, extracted proteins from a representative pool of samples (3 IBD and 3 symptomatic non-IBD patients) were separated on a 4–20% polyacrylamide gel and then reduced, alkylated, and digested in the gel. Peptides were extracted from the gel using successive rounds of dehydration and sonication and purified using reverse-phase SPE. Data-Dependent Acquisition (DDA) mode was used to acquire raw data from 12 gel fractions of a pooled sample. The spectral library was created following the procedure outlined in a previous study [17]. Briefly, the raw data (.wiff) files obtained in DDA and DIA mode were converted into mzML format with MSConvert (GUI) from ProteoWizard (v3.0.22074) [18]. Subsequently, we utilized FragPipe software (https://fragpipe.nesvilab.org/, accessed on 10 March 2022) to search the MS/MS spectra against the human proteome reviewed database (UP000005640; including isoforms and contaminants; accessible at www.uniprot.org (accessed on 15 March 2022), containing 20,411 reviewed proteins) via the MSFragger search engine [19]. This search was conducted with default open search parameters, specifying a peptide length between 6 and 42, using strict trypsin as the enzyme with a maximum of 1 missed cleavage allowed, setting the maximum fragment charge to 4, and designating methionine oxidation as a variable modification and carbamidomethylation as a fixed modification. The mass tolerance for precursor ions was set to ±20 ppm and for fragment ions at 20 ppm. The false discovery rate (FDR) for both peptide and protein identifications was set at 5%. The DDA and DIA-based libraries were merged and carefully filtered to remove duplicated precursors and we counted a total of 2000 proteins. This integration increased the human proteome coverage of the library.

2.5. Label Free Quantification Analysis

All DIA-converted data in mzML format were processed using DIA-NN software (version 1.8.1) with the following parameters: a fragment ion m/z range of 200 to 1800, a precursor m/z range of 300 to 1800, a precursor false-discovery rate (FDR) threshold of 1%, automatic settings for mass accuracy at both the MS2 and MS1 levels, and the scan window. Protein inference was set to 'Genes', and the quantification strategy was 'robust LC (high accuracy)'. Cross-run normalization was disabled, while match between runs (MBR) was enabled.

2.6. Statistical and Modeling Analysis

The statistical analysis was conducted with R software (version 4.2.2) and the basement of RStudio included packages ggplot2 for visualization, limma [20] for normalization, sva [21] for batch effect correction, and impute for imputation [22]. Differentially expressed proteins were identified using ProStar software (version 1.30.5) [23]. Machine learning and the feature selection analysis were mainly performed using freely available WEKA software (https://www.cs.waikato.ac.nz/ml/weka/, version 3.8.6, accessed on 15 January 2023) [24] and using R packages Caret (Classification And REgression Training) [25], caretEnsemble [26] and Boruta [27].

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE [28] partner repository with the dataset identifier PXD047585.

3. Results

3.1. Patient Demographics

A total of 123 stool samples were collected, including 70 active IBD patients and 53 gastrointestinal symptomatic non-IBD patients. The age distribution of the samples was 48.3 ± 19.8 (mean $\pm$ SD) and ranged from 18 to 90 years, and the sex distributions in each group lay approximately in an equal range (53% F vs. 47% M) with no statistical difference.

3.2. MS Analysis and Generating the Spectral Library

The samples were analyzed using SWATH-MS in four distinct batches with four replicated samples in batches for the batch effect diagnosis. Initially, we used batches 1–3 including 78 samples for the retrospective analysis and model training, while keeping aside batch 4 with 45 samples as a prospective blind testing group. For accurate peptide identification, we utilized the combined library (DDA and DIA) in conjunction with MBR (match between runs) within the DIA-NN software. The DIA-NN software employs collections of deep neural networks to enhance the ability to match DIA fragmentation patterns with spectral libraries, thereby improving sensitivity [29]. Moreover, enabling the match between runs (MBR) parameter led to an increase in the average number of identified entities and significantly improved data completeness by reducing the occurrence of missing values (https://github.com/vdemichev/DiaNN, accessed on 15 May 2022). An estimated 1250 proteins and 9000 peptides were identified and quantified.

3.3. Data Preprocessing

To obtain a precise differential expression protein (DEP), it is necessary to conduct an accurate data analysis of quantitative proteomic studies. This involves various key steps in data processing, including normalization, batch effect correction, imputation of missing values, and appropriate statistical analysis [30–33]. Since there is currently no established standard procedure for data processing in quantitative proteomics, to ensure an accurate biomarker analysis, we optimized each analytical step and identified an appropriate pipeline, as summarized in Figure 1. To begin the analysis, we first eliminated contaminants and proteins that had less than 70% valid values in each batch. After completing this step, we were left with a total of 250 proteins for further analysis. Afterward, a logarithm transformation ($\log_2$) was applied to the intensity values as common practice for normalizing skewed data and approximating a normal distribution. To evaluate the data

structure initially, we used a box plot to observe differences in variances and means [31,33] (see Figure 2a). From Figure 2a, it is clear that there is significant variation among the samples and batches, which indicates the batch effect. In order to eliminate the unwanted non-biological variability caused by differences in procedures of sample collection, storage, preparation, and spectral data acquisition, a normalization process was necessary [33].

3.3.1. Normalization and Batch Effect Correction

Normalization methods must be selected carefully. Several studies have been carried out on this topic. Dubois et al. systematically evaluated various commonly used normalization methods on a large MS-based proteomic dataset [34]. The results indicated that there was superior performance for certain methods, including sample quantile normalization and median centering. Due to the approximate similarity of sample proteomes, we employed quantile normalization, which is designed to align the distributions of different samples by matching their quantiles [35]. Zhao et al. suggest utilizing a "class-specific" approach for quantile normalization in their study. However, as we intended to apply the final pipeline to an unseen dataset with a blind group label, we opted for overall normalization (regardless of classification) instead [36]. Figure 2b shows the intensity distributions after quantile normalization, displaying high similarity, which is desirable in experiments in which most features are expected to remain constant. Although normalization improves comparability among samples, it primarily focuses on aligning their overall patterns. Consequently, even after normalization, batch effects that specifically impact particular proteins or protein groups can remain a significant source of variance. To explore if data were affected by batches, a principal component analysis (PCA) was performed using batch labels. The results depicted in Figure 2c highlight the considerable influence of the batch on the sample distribution and clustering of samples. Moreover, the replicated samples were generally not closely grouped, except for replicate D, which could randomly position. This clustering can be caused by external experimental factors such as technical and temporal variability. In addition, we attempted to apply median-centering normalization as an alternative to quantile normalization to assess its impact on the batch effect. However, the PCA results did not demonstrate any noticeable improvement with this approach (Supplementary Figure S1). To remove the batch effect, we used the ComBat method, which is a popular and widely used method for gene expression data but is also applicable to proteomics data [37]. ComBat offers an enhanced variant of the mean shift that makes use of a Bayesian framework. The application of the ComBat algorithm to normalized data yielded a substantial improvement in correcting batch effects, as seen in Figure 2d,e. This improvement was evident in the closest representation of the replicated samples of each batch, as observed in the PCA analysis shown in Figure 2f. Furthermore, even though we were aware that it is preferable to perform batch correction after normalization [38], we wanted to explore if the order in which normalization and batch correction are implemented had any impact on the outcome. However, their PCA comparison indicated there were no significant differences observed between the two approaches (Supplementary Figure S2).

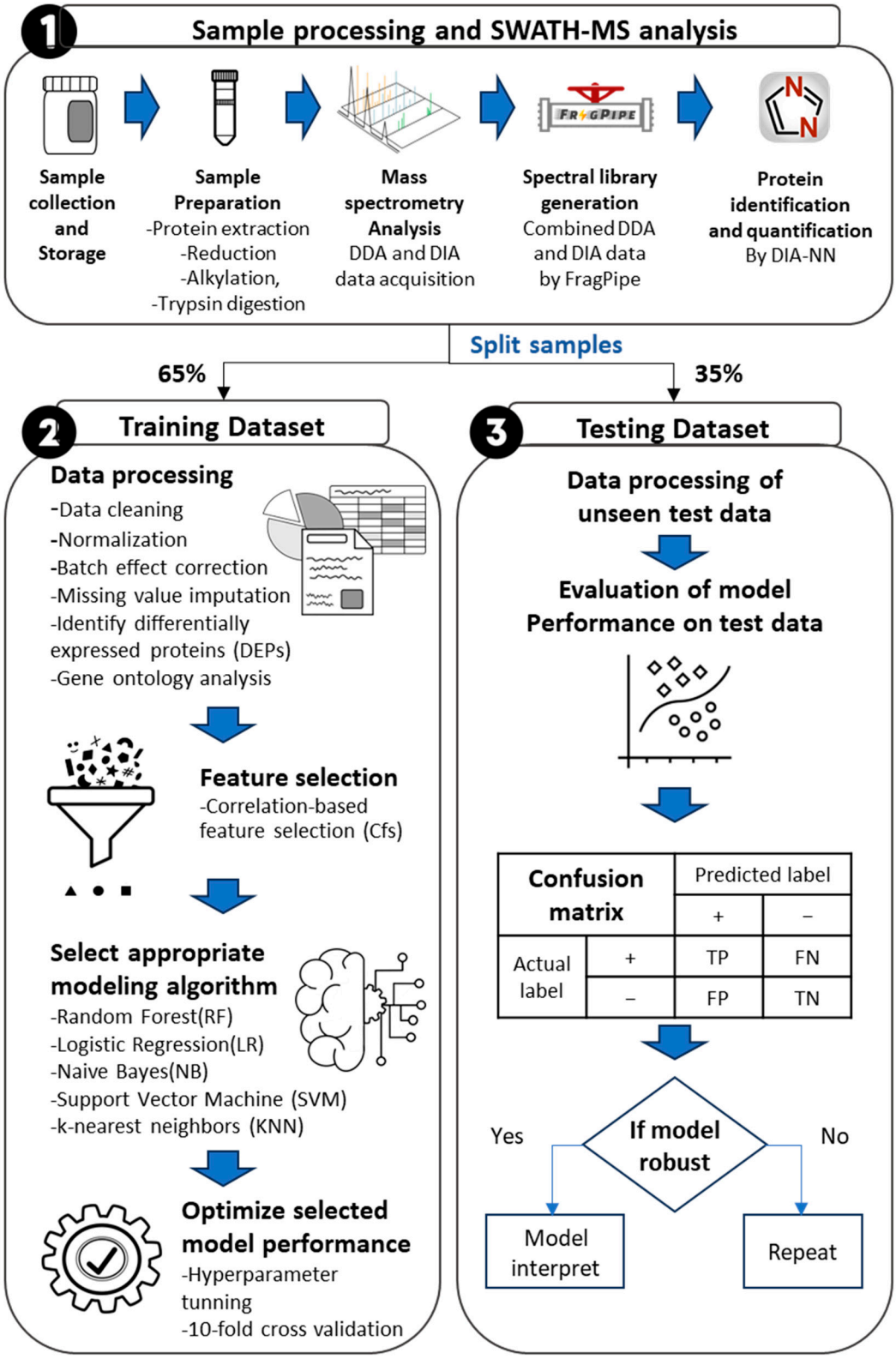


Figure 1. General workflow. This schematic representation outlines the experimental procedure, which consists of three main steps: (**1**) sample processing and SWATH-MS analysis—This step involves obtaining proteome data from stool samples. (**2**) Data processing, training, and optimizing the machine learning model—in this phase, a machine learning model is trained and optimized using 78 training samples. (**3**) Evaluation of model performance—the final step involves evaluating the model's performance using 45 prospective samples (testing set).

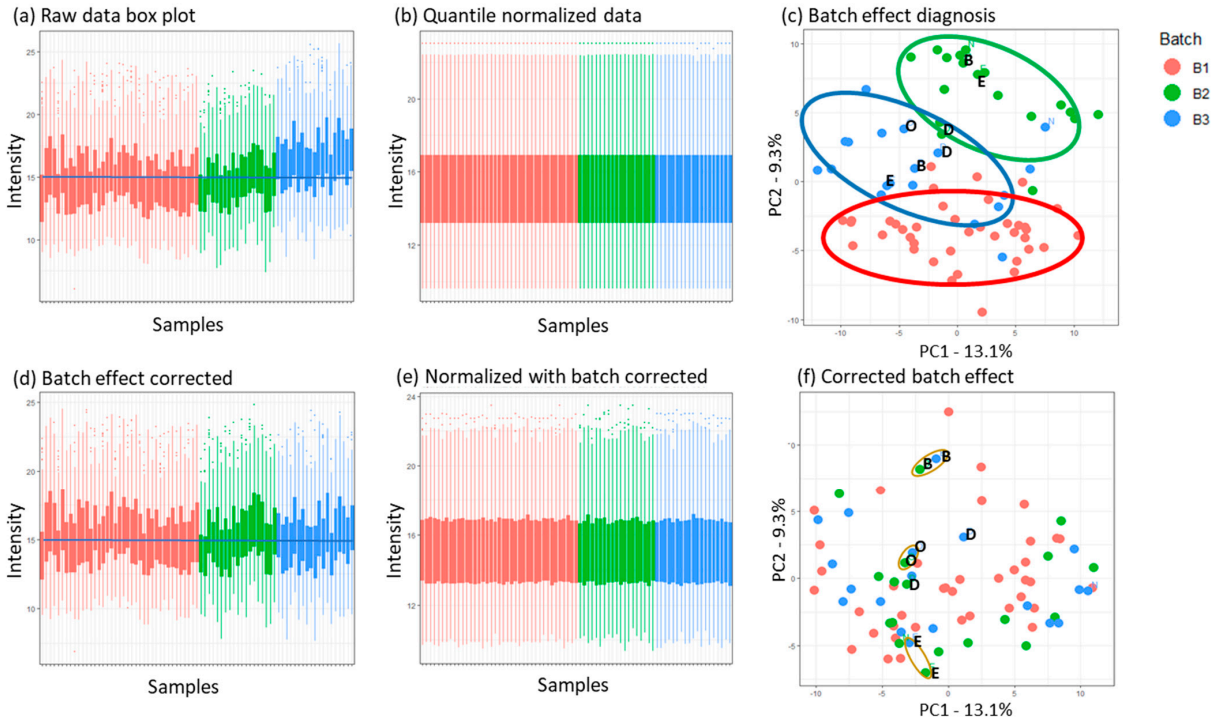


Figure 2. Identification and correction of the batch effect. (**a**) Box plot illustrating the protein distribution in the unprocessed log-transformed data format across three sample batches. (**b**) Box plot displaying the quantile normalized data. (**c**) PCA analysis of the normalized data, revealing clear clustering due to batch effects. (**d**) Box plot representing the influence of the ComBat batch effect correction on the initial data. (**e**) Box plot of data after batch correction of normalized data. (**f**) PCA analysis indicating the successful elimination of batch effects by the close representation of the replicated samples in different batches. The letters in panels (**c**,**f**) illustrate the replicated samples in different batches, which are expected to be seen in close proximity to each other. This expectation is fulfilled after batch correction.

3.3.2. Missing Value Imputation

Missing values (MVs) are commonly encountered in quantitative proteomics datasets, primarily due to the limitations of protein detection and random fluctuations that occur during the process of data acquisition. The presence of MVs necessitates the consideration of their removal or imputation [33]. To determine the most appropriate approach for handling missing values, it is crucial to identify the origins and types of these missing values [39]. In general, MVs can be categorized into three types: missing values not at random (MNAR), missing values at random (MAR), and missing values completely at random (MCAR) [40]. The analysis of the data from each batch and condition, categorized as symptomatic non-IBD or active IBD, revealed the absence of intentionally missing values. In other words, we did not have proteins that were exclusively present under just one condition. Additionally, comparing replicated samples confirmed the random nature of the MVs. Various studies have assessed different imputation methods to handle missing values [40,41]. However, these studies have yielded varying results in terms of the best method due to the differences in the datasets used. Nevertheless, random forest (RF) [42] and k-nearest neighbors (KNN) [43] are commonly recommended for addressing random missingness [31].

Notably, Wang et al. have introduced the NAguideR toolkit [44], which incorporates 23 commonly used imputation methods and provides evaluation criteria to assist researchers in selecting the most suitable method for their dataset. When we applied this toolkit to our dataset, RF and KNN ranked first and second as the most appropriate imputation methods. After evaluating both methods, we found that neither of them was significantly

superior to the other. However, we ultimately decided to proceed with the KNN method, considering N = 5 for the processed data, which means that the KNN method utilizes a machine learning algorithm to estimate missing values based on the values of their five closest neighbors in the feature space.

3.3.3. Identifying Differentially Expressed Proteins (DEPs)

Following data cleaning and preprocessing, protein abundance data were prepared for further statistical analysis and downstream investigation. Our goal was to identify the subset of proteins that demonstrated significant changes between the two conditions among the pool of 250 proteins. The *t*-test and limma [20] are two widely used hypothesis testing methods. In our analysis, we utilized ProStar software [23], which incorporates both of these methods and provides options for both the *t*-tests (Student's and Welch's) and limma. Considering our assumption of varying data variation and different sample sizes in the two study groups, we chose to employ Welch's *t*-test [45]. We applied two criteria via ProStar to identify differentially expressed proteins, a fold change (FC) ratio of at least 1.6 (i.e., $|\mathrm{Log_2(FC)}| \geq 0.70$) and a *p*-value less than 0.05 (i.e., $\mathrm{Log_{10}}(p\text{-value}) \geq 1.3$), resulting in the filtration of 201 proteins [46]. The subsequent *p*-value calibration plot assessed the *p*-value distribution and allowed an FDR estimation adjustment using various statistical methods, such as st.boot, st.spline, langaas, Benjamini–Hochberg, etc. [47]. This calibration plot ensures an evaluation of how well observed *p*-values align with the expected behavior under specific assumptions about the proportion of differentially and non-differentially abundant proteins. In this analysis, the st.boot (Bootstrap) method demonstrated superior performance, yielding a pi0 value of 0.05, indicative of effective control over the false discovery rate (FDR) (below 1%). Using these criteria for the training group, we identified 48 DEPs, as shown in the volcano plot (Figure 3). Detailed data related to the *p*-values and fold changes of these 48 proteins are provided in Supplementary Table S1. Compared to symptomatic non-IBD patients, there are 32 proteins presented as upregulated and 16 proteins shown as downregulated.

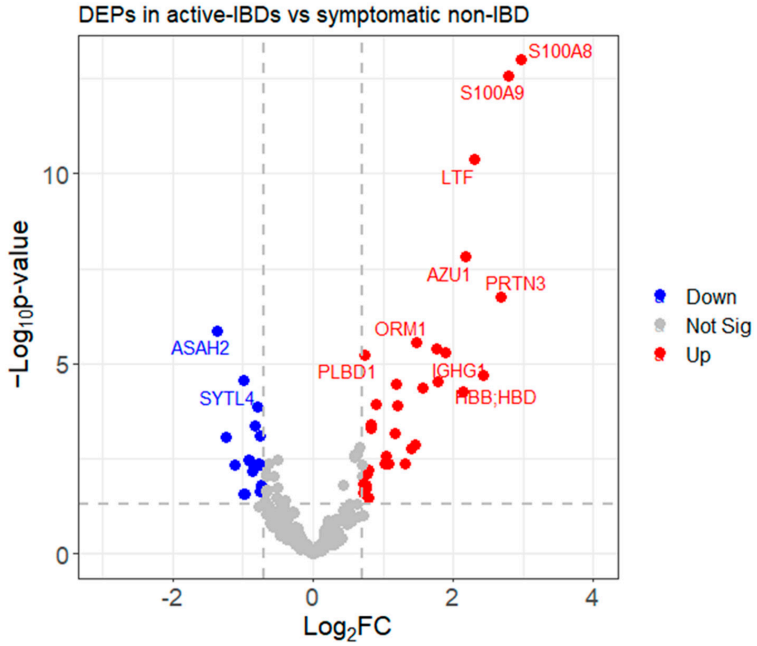


Figure 3. The volcano plot summarizes differentially expressed proteins (DEPs) detected in samples obtained from active IBD patients and symptomatic non-IBD patients. Among the approximately 300 proteins detected as being consistently present in samples obtained from IBD patients, 48 were initially identified either as reduced (blue dots) or increased (red dots). Group comparisons between active IBD patients and symptomatic non-IBD patients were calculated using Welch's *t*-test with a $|\mathrm{Log_2(FC)}| \geq 0.70$ and *p*-value ≥ 1.3 in ProStar.

3.4. Functional Enrichment Analysis

KEGG and gene ontology enrichment analyses were performed with ShinyGO (http://bioinformatics.sdstate.edu/go/, accessed on 1 July 2023) in order to gain insights into the functional roles of dysregulated proteins. Gene ontology analysis revealed that these 48 proteins are mainly involved in biological processes related to inflammatory and immune responses, particularly neutrophil- and myeloid cell-related processes, as we expected (Figure 4a), which confirms the upregulation of inflammatory genes in patients with active IBD compared to symptomatic non-IBD patients. Moreover, KEGG pathway enrichment analysis results revealed seven pathways that are significantly affected by DEPs, including the renin–angiotensin system, protein digestion and absorption, mineral absorption, the complement and coagulation cascade, the IL-17 signaling pathway, pancreatic secretion, and Chagas disease. The "renin–angiotensin system (RAS)" pathway is highly enriched and likely plays a significant role in IBD, as illustrated in Figure 4b. Some studies have reported altered levels and activities of RAS components in the inflamed mucosa. These studies suggest that RAS inhibition can have anti-inflammatory effects on IBD. That is why pharmacologically inhibiting the classic RAS pathway using ACE inhibitors and angiotensin II receptor blockers (ARBs) has been a well-established strategy to treat hypertension [48]. The next highest fold enriched pathway includes protein digestion and absorption, reflecting alterations in digestive functions and nutrient absorption in IBD patients [49]. Moreover, the role of the IL-17 pathway in the pathogenesis of IBD and its involvement in inflammatory cytokine production, neutrophil recruitment, and tissue remodeling has been demonstrated [50].

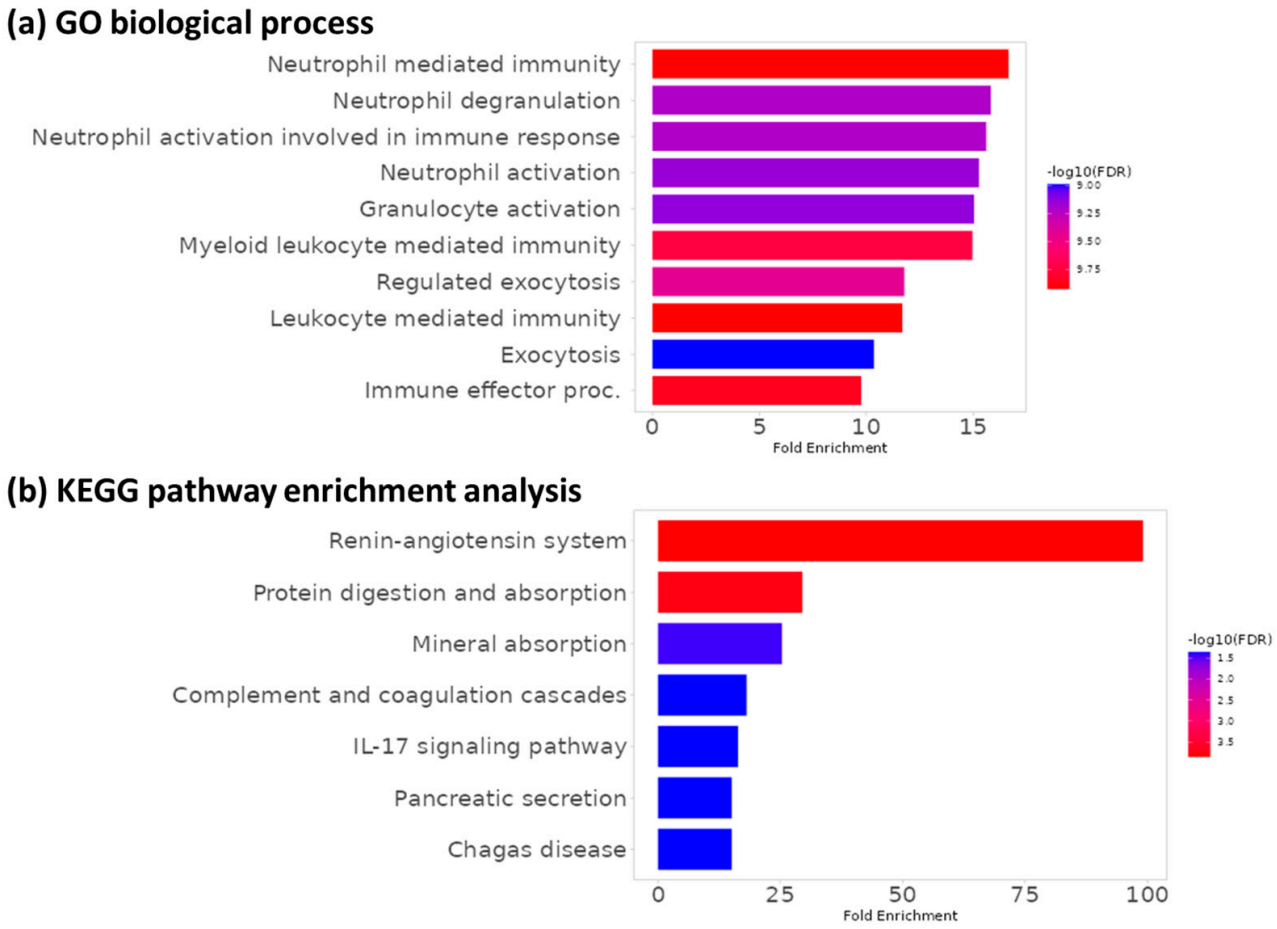


Figure 4. Functional analysis of differentially expressed proteins in the two groups. (**a**) The significant gene ontology analysis of 48 DEPs proteins (p-value < 0.05) (**b**) KEGG pathway enrichment analysis of 48 DEPs proteins in active IBD patients and asymptomatic non-IBD patients.

3.5. Feature Selection

To construct a predictive model, it is essential to select relevant features while removing redundant and irrelevant ones through feature selection. This process reduces data dimensionality, improves model performance, and reduces overfitting. In this study, "features" refer to the "proteins", and we aimed to reach a reasonable number of proteins as biomarkers. Two common classic feature selection models are the filter and wrapper methods. The main difference between them is that a filter model selects features based on intrinsic data properties, while a wrapper model involves a learning algorithm in determining feature quality [51]. To identify the most relevant features among the 48 DEPs, we assessed five well-known feature selection methods, including correlation-based feature selection (Cfs), Boruta, information gain, gain ratio, and the wrapper method in WEKA software. Among these methods, the Cfs method demonstrated superior prediction performance compared to the others. Cfs is a filter-based feature selection method that chooses features based on their maximum correlation with the class variable and minimum intercorrelation [52]. As feature reduction offers several benefits, including speeding up algorithm processing time, improving data quality, enhancing algorithm predictive power, and making results more understandable, we aimed to investigate whether we could reduce these 16 proteins without compromising classification performance [33]. To refine our selection, we excluded proteins with less attribute weight, resulting in the elimination of five that had minimal impact on classification performance. Seeking further optimization, we assessed protein–protein correlations among the remaining 11 proteins and removed the ones with a high intercorrelation and lower weight attribute. This iterative process led to a reduction in the number of proteins to seven. Figure 5 illustrates the correlation heatmap among proteins, with the selected ones highlighted. This visualization demonstrates that the selected proteins are primarily chosen from distinct clusters, confirming their low intercorrelation.

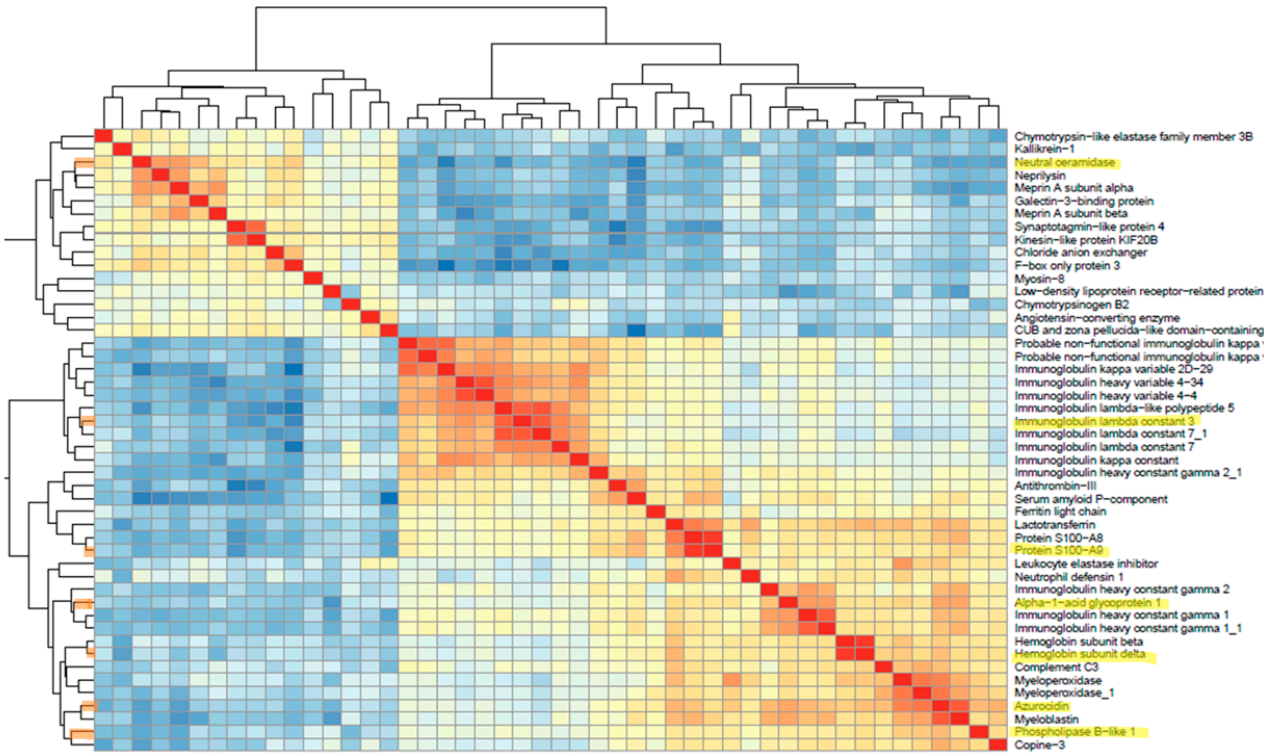


Figure 5. Protein correlation heatmap. This heatmap reveals two prominent clusters: one comprising upregulated proteins and the other containing downregulated proteins. It visually represents correlation strength, with stronger correlations depicted in red and weaker ones in blue. Notably, the seven selected proteins for our model are highlighted in the map, each of which is mainly associated with distinct clusters exhibiting low intercorrelations.

These seven proteins included six upregulated proteins and one downregulated protein, as displayed in Table 1. Figure 6 illustrates the enhancement in unsupervised group classification via PCA across three datasets: the original dataset comprising 250 proteins, the dataset following the DEP analysis of 48 proteins, and the dataset featuring the seven proteins selected through the feature selection process. Based on these plots, it is visually evident how effectively the two groups separate as we reduce the number of proteins, and the cumulative proportion of variance explained by the first two principal components significantly increases from 22% to 61%.

Table 1. Characteristics of selected proteins. The list of the final seven selected proteins for a prediction model, including their fold changes and *p*-values, across the two groups. The attribute weights show different levels of importance to different features (proteins) during the model training and classification process.

Protein Name	Gene	Fold Change	*p*-Value	Attribute Weight
Protein S100-A9	S100A9	6.9	0.0000	−3.9813
Azurocidin	AZU1	4.5	0.0000	−2.7925
Immunoglobulin lambda constant 3	IGLC3	2.0	0.0044	−2.4284
Hemoglobin subunit delta	HBB	5.4	0.0000	−2.2529
Phospholipase B-like 1	PLBD1	1.7	0.0000	−1.6708
Alpha-1-acid glycoprotein 1	ORM1	2.8	0.0000	−0.9056
Neutral ceramidase	ASAH2	−2.6	0.0000	1.1675

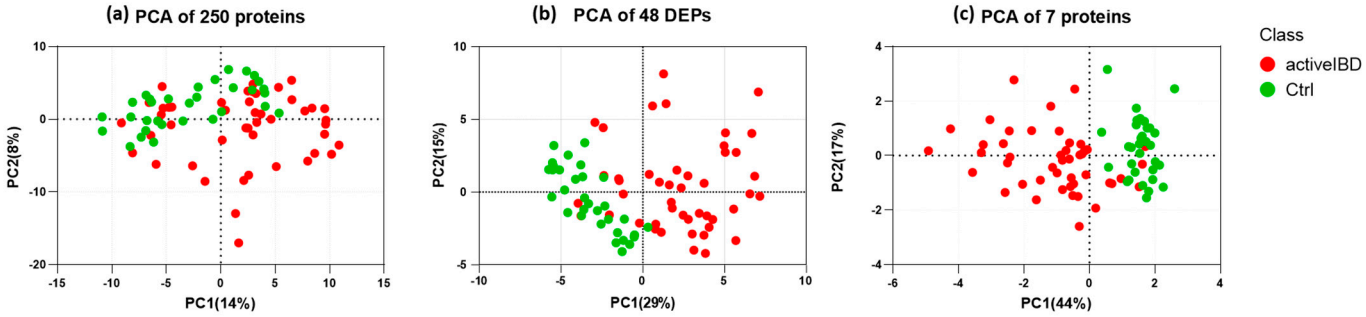


Figure 6. Unsupervised group classification via principal component analysis (PCA) across three datasets: (**a**) the original dataset comprising 250 proteins, (**b**) the dataset following the DEP analysis of 48 proteins, and (**c**) the dataset featuring the seven proteins selected through the feature selection process.

3.6. Selecting the Appropriate Machine Learning Algorithm

Machine learning (ML) is a powerful tool in bioinformatic analysis. Supervised machine learning refers to using quantitative proteome data with known clinical conditions to train a model for the prediction of prospective samples [53]. To identify the most appropriate predictive classifier according to the nature of the data, the five most popular machine learning methods, including support vector machines (SVMs), random forests (RFs), logistic regression (LR), k-nearest neighbors (KNN) and naive Bayes (NB), were evaluated. Table 2 displays the performance metrics of five classifiers for predicting active IBD patients from symptomatic non-IBD patients in terms of accuracy, precision, recall, F-score, area under the ROC curve (AU-ROC), and area under the precision and recall curve (AU-PRC). Detailed information on each of these parameters is presented in Table 3. These metrics were obtained using WEKA [24]. The results indicate that the SVM classifier outperforms the others in all criteria.

Table 2. Performance metrics of classifiers. This table compares the performance metrics of each classifier based on the prediction results obtained from the training and validation of 78 samples using 10-fold cross-validation. The SVM model outperforms the other classifiers based on the first four metrics. However, when considering the area under the curve (AUC), the RF classifier outperforms the others. Further analysis confirms that the SVM model works better for these data and provides more accurate predictions for blind data.

Classifier	Accuracy	Precision	Recall	F-Score	AU-ROC	AU-PRC
SVM	95%	0.97	0.93	0.96	0.95	0.96
NB	90%	0.94	0.90	0.92	0.93	0.94
LR	88%	0.89	0.91	0.90	0.92	0.92
KNN	88%	0.91	0.89	0.90	0.90	0.89
RF	87%	0.89	0.89	0.89	0.94	0.93

Table 3. General definitions of different performance metrics for model selection.

Measure	Evaluation Focus
Accuracy	• The overall effectiveness of a classifier
Precision	• The proportion of positive instances among all instances classified as positive
Recall (Sensitivity)	• The proportion of positive instances correctly classified as positive out of all positive instances in the data
F-score	• The harmonic mean of precision and recall and provides a combined measure of both
ROC Area	• The area under the receiver operating characteristic (ROC) curve, which is a graphical representation of the trade-off between the true positive rate and the false positive rate for different threshold values of the classification model
PRC Area	• The area under the precision–recall curve (PRC), which is a graphical representation of the trade-off between precision and recall for different threshold values of the classification model

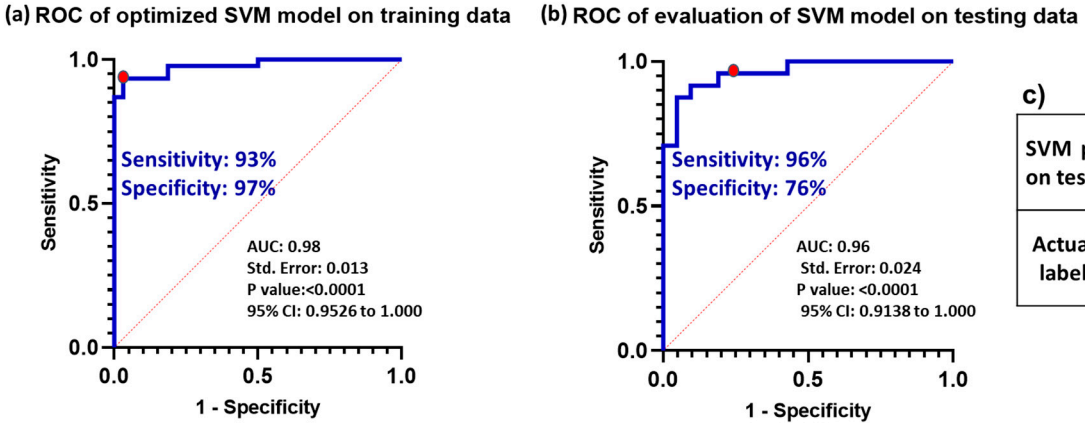
3.7. Optimizing the Selected Model Performance

In the context of machine learning, there is a risk of a model becoming overly proficient at learning from the training data, a phenomenon known as overfitting. This entails not only capturing the inherent patterns but also incorporating noise or random variations present in the data. Consequently, an overfit model performs exceptionally well on the training data but faces challenges when attempting to apply its knowledge to new and unfamiliar data. Two strategies that help to avoid overfitting are cross-validation and hyperparameter tuning [54]. In this study we used 10-fold cross-validation for the training data where the data divided into ten subsets, with nine parts of the data used for training and one part for validation in each fold. The experiment was then repeated 10 times, with each of these subsets serving as the validation group. The final result indicated the average across the 10 folds, which provides a more realistic assessment of the model's performance. Moreover, most machine learning algorithms have parameters that can be adjusted, referred to as hyperparameters. These hyperparameters are critical for building robust and accurate models, as they help find the balance between bias and variance, thereby preventing the model from overfitting. Two common effective techniques for hyperparameter tuning are grid search and random search [55]. Rafael et al. have demonstrated equal predictive performance for grid and random search techniques in SVM hyperparameter tuning [56]. In this analysis, we employed the "tuneLength = 10" function for each classifier as a grid search method. This means that the system will perform hyperparameter tuning by randomly selecting 10 different combinations of hyperparameters for each classifier and evaluating their performance using cross-validation to find the best hyperparameter

settings. The analysis indicates that, for an SVM classifier with a polynomial kernel of degree 1, a scale parameter of 0.001, and a cost parameter of eight (C = 8), it outperforms other configurations and achieves an accuracy of 0.95, a sensitivity of 0.93, and a specificity of 0.97 in classifying the training dataset.

3.8. Model Validation with Prospective Data

To validate the optimized model, it was applied to 40 blind samples from batch 4. The prediction results indicated 96% sensitivity and 76% specificity, as shown in the confusion matrix in Figure 7. Figure 7 also illustrates that the area under the ROC curve was equal to 0.96. The results highlight the high accuracy and performance of the generated model in accurately classifying the blind data.



c)

SVM prediction on test dataset		Predicted label	
		aIBD	Ctrl
Actual label	aIBD	23 (TP)	1 (FN)
	Ctrl	5 (FP)	16 (TN)

Figure 7. (**a**) The ROC curve analysis involved applying an optimized SVM model to the training dataset to determine the threshold, resulting in 93% sensitivity and 97% specificity as indicated by the red dot. (**b**) In the ROC curve analysis of the model applied to the testing dataset, we obtained an AUC of 0.96. Using a previously selected threshold, the model achieved 96% sensitivity and 76% specificity (red dot). (**c**) The confusion matrix displays the counts of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). The results highlight the high accuracy and exceptional performance of the generated model in accurately classifying the blind data.

4. Discussion

This study demonstrated the potential use of SWATH-DIA proteomic profiling of stool samples as a tool for diagnosing active-IBD patients from symptomatic non-IBD patients. This was achieved by employing machine learning techniques to develop a robust predictive model. To accomplish this, we designed an experiment with three main steps of (1) data acquisition and processing, (2) training and optimizing a machine learning model based on 78 retrospective samples, and (3) validating the model's performance on 45 prospective samples. Achieving 96% sensitivity with a 0.96 AUC using a blind dataset confirmed the model's robustness, also indicating our ability to successfully and effectively process the data obtained from four separate batches with different collection times. The processing steps included the successful removal of batch effects and employed effective methods for normalization and missing value imputation.

We have corrected the batch effect using the ComBat method. ComBat starts by adjusting each batch of data separately to have similar means and variances, and then calculates the differences between the batches and uses this information to "harmonize" the data. ComBat adjusts the data for each sample in a way that minimizes the batch-related differences while preserving the true biological differences [37].

To impute missing values, it is crucial to understand the nature of the data and determine the reasons for their absence, which will guide the selection of an appropriate imputation method. Upon comparing the replicated samples, we observed that the missing values were missing at random. "Zero", "mean", and "minimum value" are the straightfor-

ward imputation methods commonly used, but they may not always be suitable, especially when the missing values occur randomly and are not due to limits of detection or actual missing data. In such cases, imputing them with these methods could introduce bias into the analysis. Therefore, we chose to employ the k-nearest neighbors (KNN) imputation method with a setting of five neighbors. This implies that it leverages information from the five most similar samples in the dataset to estimate the missing values.

Among the differentially expressed proteins (DEPs), the highest-scoring proteins in the volcano plot were S100A8 and S100A9, which are well-known neutrophil-derived proteins predominantly found as the S100A8/S100A9 complex, also known as calprotectin. This finding further confirms the correctness of the analysis pathway.

Utilizing all 48 differentially expressed proteins as biomarker signatures for classification may not be practical. Therefore, we needed to reduce the number of biomarkers without compromising prediction accuracy. However, selecting only the best proteins and combining them based on previous studies does not guarantee an improvement in overall classification performance. Furthermore, in machine learning, a specific coefficient is assigned to each biomarker, known as a weight, based on its importance and effect on classification to achieve an optimal result. For instance, Mooiweer et al. found that the combination of fecal hemoglobin and calprotectin did not enhance their predictive accuracy compared to using fecal Hb and FC individually [57]. Similarly, Schröder et al. found that the combination of calprotectin, lactoferrin, and neutrophile elastase did not increase predictive accuracy when compared with calprotectin alone [58]. In this regard, using correlation-based feature selection in this study helped us to only keep the seven most relevant proteins with maximum correlations with the class variable and minimum intercorrelation. For instance, retaining both the S100A9 and S100A8 proteins does not provide significant additional informative value because both of them are subunits of calprotectin and exhibit a high correlation with each other. Moreover, the correlation heatmap in Figure 5 indicates that S100A9 and S100A8 also share a high correlation with lactoferrin, and there is also a noticeable correlation between azurocidin, myeloblastin, and myeloperoxidase. Although all of them were identified previously as potential IBD markers, keeping one of them would give us almost similar results.

The seven selected proteins include the upregulated proteins S100A9, azurocidin (AZU1), immunoglobulin lambda constant 3, hemoglobin subunit delta, phospholipase B-like 1 (PLBD1), and alpha-1-acid glycoprotein 1 (alpha 1-AGP), and the downregulated protein neutral ceramidase (ASAH2). Two of these proteins, S100A9 and AZU1, are associated with neutrophils and play a key role in the host's defense against bacterial infections. S100A9 is, in fact, a subunit of calprotectin, accounting for approximately 60% of the total soluble proteins in the cytosol fraction of neutrophils, while AZU1 is found in the azurophilic granules of neutrophils, alongside other proteins [59]. Hemoglobin delta is linked to occult intestinal bleeding in IBD patients, and previous research has highlighted a correlation between fecal hemoglobin and calprotectin [57]. Immunoglobulin lambda is a light chain of hemoglobin and can be indicative of an active immune system in IBD patients. The increase in free light chains (FLCs), including kappa and lambda immunoglobulins, in plasma has previously been shown in diabetes and immune system abnormalities, as well as autoimmune-based inflammatory diseases [60,61]. However, the dysregulation of lambda light chains in stool and its relevance to IBD have not been studied in detail. PLBD1 is a phospholipase that can generate lipid mediators of inflammation and was first identified in neutrophils [62]. However, to the best of our knowledge, its relationship with IBD has not been specifically investigated. Alpha 1-AGP is one of the major acute phase proteins in humans, and its serum concentration increases in response to systemic tissue injury, inflammation, or infection [63]. Takashi et al. demonstrated a significant increase in fecal alpha 1-AGP in active IBD patients compared to non-active patients, suggesting alpha 1-AGP as a potential biomarker for evaluating IBD activity [64]. ASAH2 is involved in breaking down ceramides to sphingosines. Its downregulation in IBD causes ceramide accumulation in microdomains of cholesterol- and sphingolipid-enriched membranes,

resulting in an impairment of the barrier function of the gut [65,66]. The loss of ASAH2 causes elevated levels of sphingosine-1-phosphate and systemic inflammation in ASAH2 knockout mice [67]. These proteins collectively offer insights into the complex molecular mechanisms and potential biomarkers associated with IBD.

The superiority of SVM over other models can be attributed to various factors, including the characteristics of the data, the nature of the classes, the distribution of the features, and the inherent strengths and weaknesses of each algorithm [68]. Some advantages of SVM over other classifiers include being less prone to overfitting due to its optimization process and regularization (controlled by the parameter C and gamma), and greater robustness to outliers and noisy data [69,70].

SVM serves as a robust technique for constructing a classifier [71]. Its primary objective is to establish a decision boundary between two classes, facilitating the classification of data points based on their features. This decision boundary, referred to as a hyperplane, is positioned in a manner that maximizes its distance from the nearest data points of each class, which are known as support vectors [72]. Vapnik initially introduced the SVM algorithm in 1963 to create linear classifiers [73]. Additionally, SVMs can employ kernel methods to model complex, non-linear patterns in higher dimensions. The choice of a suitable kernel function, among other considerations, can significantly impact the performance of an SVM model. However, there is no definitive method to determine the optimal kernel for a specific pattern recognition problem. It often involves a trial-and-error approach, beginning with a basic SVM and experimenting with various standard kernel functions [72]. In this study, the selection of the optimal kernel function is part of the hyperparameter tuning process. Depending on the nature of the data, one kernel (with a degree of one) outperforms the others. This configuration is commonly referred to as "Linear SVM" or "SVM with a Linear Kernel" [74]. This setup assumes that the data is linearly separable, which could be considered an advantage in simplifying the model complexity.

Let us take a closer look at the cost and gamma hyperparameters to gain insights into their impacts on the model. The scale parameter (γ or gamma) controls how tightly the SVM model fits the training data. The usual range for the gamma parameter is typically between 0.01 and 10. Opting for smaller values, such as our chosen value of 0.001, implies a more extensive decision boundary. In contrast, larger values like one or 10 result in narrower decision boundaries, which, if not carefully considered, can potentially trigger overfitting. On the other hand, the cost parameter (C) in SVM controls the trade-off between training error and testing error. The usual range for the cost parameter typically lies between 0.1 and 1000. A smaller C allows for a larger margin and tolerates some misclassification of training points. In our dataset, C = 8 exhibited better performance than the other values. This value strikes a balance between being not too large, which could lead to overfitting, and not too small, which could risk underfitting.

One limitation of this study is that it involved Canadian IBD patients aged 18 and above. Therefore, applying the machine learning algorithm to populations from different regions and ages should be approached with caution. Additionally, while we were able to correct the batch effect, it is essential to note that all samples were analyzed using a single mass spectrometer. To ensure the broader applicability of this method in different clinical laboratories, it would be advantageous to analyze data from various spectrometers.

The primary objective of this study was to provide the proof of concept that a SWATH-based MS analysis can be advantageously used as an additional tool for assisting the gastroenterologist through a protein signature. This, in turn, can significantly enhance the effectiveness of IBD therapy and overall disease management. Moreover, this approach offers substantial advantages in terms of expediting and improving the precision of IBD diagnoses, thereby preventing the deterioration of the patient's condition due to delayed colonoscopy or inaccurate diagnosis. It also ensures the optimal prescription of drugs from the outset, maximizing treatment efficacy. Additionally, by reducing the necessity for unnecessary colonoscopies, it not only carries financial benefits but also minimizes patient

discomfort and anxiety, saves time, enhances convenience, and streamlines the diagnosis and monitoring processes.

In conclusion, this study presents a proof of concept for the application of SWATH for precise IBD diagnosis using stool proteomics and showcases the effectiveness of the data processing and machine learning approaches. Additionally, it highlights the potential of this method for classifying Crohn's disease (CD) vs. ulcerative colitis (UC) and distinguishing active IBD from remission. The creation of a non-invasive, precise, and sensitive method for diagnosing and monitoring IBD could have a substantial positive impact on the quality of life of IBD patients and lessen the burden of unnecessary or repeated invasive procedures.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/biomedicines12020333/s1, Figure S1: The comparison examines the effects of two different normalization methods on batch effect correction; Figure S2: The comparison examines the effect of normalization before and after batch effect correction; Table S1: List of 48 differentially expressed proteins with their corresponding p-values and fold-change values.

Author Contributions: Conceptualization, E.S., D.G. and J.-F.B.; methodology, E.S., D.G., P.R. and J.-F.N.; collection of samples, E.S., D.G., M.M., P.R. and M.D.; data curation and formal analysis, E.S. and D.G.; statistical and machine learning analysis, E.S. and M.A.B.; funding acquisition, E.S., D.G., M.D., M.A.B., F.-M.B. and J.-F.B.; resources, H.G. and J.-F.B.; writing—original draft preparation, E.S.; writing—review and editing, E.S., D.G., M.M., P.R., J.-F.N., H.G., M.A.B., M.D., F.-M.B. and J.-F.B.; supervision, M.A.B., M.D., F.-M.B. and J.-F.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by grants from Crohn's and Colitis Canada (grant # 1031647) to M.D., M.A.B., F.M.B. and J.F.B., and the Natural Sciences and Engineering Research Council of Canada through a cooperative ENGAGE grant to J.F.B. with Allumiqs Solutions. E.S. was the recipient of a doctoral studentship from the Faculty of Medicine and Health Science of the Université de Sherbrooke. D.G. was the recipient of an MITACS postdoctoral fellowship obtained in collaboration with Allumiqs Solutions. J.F.B. was the recipient of the Canada Research Chair in Intestinal Physiopathology.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board (or Ethics Committee) of the Centre Hospitalier Universitaire de Sherbrooke (protocol code 1991-17, 90-18, last date of approval 27 August 2023).

Data Availability Statement: The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository (http://www.ebi.ac.uk/pride, accessed 6 December 2023) with the dataset identifier PXD047585.

Acknowledgments: The authors thank the personnel of the Hematology Lab of the Centre Hospitalier Universitaire de Sherbrooke (CHUS) for their cooperation in the daily collection of samples; the patients of the CHUS for their consent to participate to the project; and Elizabeth Herring for reviewing the English of the manuscript.

Conflicts of Interest: H.G., J.F.N. and D.G. are the CSO & Director and employees, respectively, of Allumiqs. E.S., D.G., M.M., H.G., J.F.N. and J.F.B. are the inventors of the intellectual property owned by TransferTech Sherbrooke, a valorization society for the Université de Sherbrooke, and the subject of a provisional patent. The other authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of the data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Baumgart, D.C.; Carding, S.R. Inflammatory bowel disease: Cause and immunobiology. *Lancet* **2007**, *369*, 1627–1640. [CrossRef]
2. Pithadia, A.B.; Jain, S. Treatment of inflammatory bowel disease (IBD). *Pharmacol. Rep.* **2011**, *63*, 629–642. [CrossRef] [PubMed]
3. Langshaw, A.; Rosen, J.; Pensabene, L.; Borrelli, O.; Salvatore, S.; Thapar, N.; Concolino, D.; Saps, M. Overlap between functional abdominal pain disorders and organic diseases in children. *Rev. Gastroenterol. México* **2018**, *83*, 268–274. [CrossRef] [PubMed]
4. Fisher, D.A.; Maple, J.T.; Ben-Menachem, T.; Cash, B.D.; Decker, G.A.; Early, D.S.; Evans, J.A.; Fanelli, R.D.; Fukami, N.; Hwang, J.H. Complications of colonoscopy. *Gastrointest. Endosc.* **2011**, *74*, 745–752. [CrossRef]

5. Noiseux, I.; Veilleux, S.; Bitton, A.; Kohen, R.; Vachon, L.; White Guay, B.; Rioux, J.D. Inflammatory bowel disease patient perceptions of diagnostic and monitoring tests and procedures. *BMC Gastroenterol.* **2019**, *19*, 30. [CrossRef] [PubMed]
6. Lopez, R.N.; Leach, S.T.; Lemberg, D.A.; Duvoisin, G.; Gearry, R.B.; Day, A.S. Fecal biomarkers in inflammatory bowel disease. *J. Gastroenterol. Hepatol.* **2017**, *32*, 577–582. [CrossRef]
7. Laserna-Mendieta, E.J.; Lucendo, A.J. Faecal calprotectin in inflammatory bowel diseases: A review focused on meta-analyses and routine usage limitations. *Clin. Chem. Lab. Med. (CCLM)* **2019**, *57*, 1295–1307. [CrossRef] [PubMed]
8. Rokkas, T.; Portincasa, P.; Koutroubakis, I.E. Fecal calprotectin in assessing inflammatory bowel disease endoscopic activity: A diagnostic accuracy meta-analysis. *J. Gastrointest. Liver Dis.* **2018**, *27*, 299–306. [CrossRef] [PubMed]
9. Pham, T.V.; Piersma, S.R.; Oudgenoeg, G.; Jimenez, C.R. Label-free mass spectrometry-based proteomics for biomarker discovery and validation. *Expert Rev. Mol. Diagn.* **2012**, *12*, 343–359. [CrossRef]
10. Sajic, T.; Liu, Y.; Aebersold, R. Using data-independent, high-resolution mass spectrometry in protein biomarker research: Perspectives and clinical applications. *PROTEOMICS–Clin. Appl.* **2015**, *9*, 307–321. [CrossRef]
11. Ludwig, C.; Gillet, L.; Rosenberger, G.; Amon, S.; Collins, B.C.; Aebersold, R. Data-independent acquisition-based SWATH-MS for quantitative proteomics: A tutorial. *Mol. Syst. Biol.* **2018**, *14*, e8126. [CrossRef] [PubMed]
12. Anjo, S.I.; Santa, C.; Manadas, B. SWATH-MS as a tool for biomarker discovery: From basic research to clinical applications. *Proteomics* **2017**, *17*, 1600278. [CrossRef]
13. Sidoli, S.; Lin, S.; Xiong, L.; Bhanu, N.V.; Karch, K.R.; Johansen, E.; Hunter, C.; Mollah, S.; Garcia, B.A. Sequential Window Acquisition of all Theoretical Mass Spectra (SWATH) Analysis for Characterization and Quantification of Histone Post-translational Modifications*[S]. *Mol. Cell. Proteom.* **2015**, *14*, 2420–2428. [CrossRef]
14. Fabian, O.; Bajer, L.; Drastich, P.; Harant, K.; Sticova, E.; Daskova, N.; Modos, I.; Tichanek, F.; Cahova, M. A Current State of Proteomics in Adult and Pediatric Inflammatory Bowel Diseases: A Systematic Search and Review. *Int. J. Mol. Sci.* **2023**, *24*, 9386. [CrossRef]
15. Basso, D.; Padoan, A.; D'Incà, R.; Arrigoni, G.; Scapellato, M.L.; Contran, N.; Franchin, C.; Lorenzon, G.; Mescoli, C.; Moz, S. Peptidomic and proteomic analysis of stool for diagnosing IBD and deciphering disease pathogenesis. *Clin. Chem. Lab. Med. (CCLM)* **2020**, *58*, 968–979. [CrossRef]
16. Vitali, R.; Palone, F.; Armuzzi, A.; Fulci, V.; Negroni, A.; Carissimi, C.; Cucchiara, S.; Stronati, L. Proteomic analysis identifies three reliable biomarkers of intestinal inflammation in the stools of patients with Inflammatory Bowel Disease. *J. Crohn's Colitis* **2023**, *17*, 92–102. [CrossRef]
17. Gagné, D.; Shajari, E.; Thibault, M.-P.; Noël, J.-F.; Boisvert, F.-M.; Babakissa, C.; Levy, E.; Gagnon, H.; Brunet, M.A.; Grynspan, D. Proteomics Profiling of Stool Samples from Preterm Neonates with SWATH/DIA Mass Spectrometry for Predicting Necrotizing Enterocolitis. *Int. J. Mol. Sci.* **2022**, *23*, 11601. [CrossRef]
18. Adusumilli, R.; Mallick, P. Data conversion with ProteoWizard msConvert. *Proteom. Methods Protoc.* **2017**, *1550*, 339–368.
19. Kong, A.T.; Leprevost, F.V.; Avtonomov, D.M.; Mellacheruvu, D.; Nesvizhskii, A.I. MSFragger: Ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nat. Methods* **2017**, *14*, 513–520. [CrossRef]
20. Ritchie, M.E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C.W.; Shi, W.; Smyth, G.K. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **2015**, *43*, e47. [CrossRef]
21. Leek, J.T.; Johnson, W.E.; Parker, H.S.; Jaffe, A.E.; Storey, J.D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **2012**, *28*, 882–883. [CrossRef]
22. Hastie, T.; Tibshirani, R.; Narasimhan, B.; Chu, G. Impute: Imputation for Microarray Data, R Package Version 1.76.0 2023. Available online: https://bioconductor.org/packages/impute (accessed on 1 April 2023).
23. Wieczorek, S.; Combes, F.; Lazar, C.; Giai Gianetto, Q.; Gatto, L.; Dorffer, A.; Hesse, A.-M.; Coute, Y.; Ferro, M.; Bruley, C. DAPAR & ProStaR: Software to perform statistical analyses in quantitative discovery proteomics. *Bioinformatics* **2017**, *33*, 135–136.
24. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. The WEKA data mining software: An update. *ACM SIGKDD Explor. Newsl.* **2009**, *11*, 10–18. [CrossRef]
25. Kuhn, M. A Short Introduction to the caret Package. *R Found Stat. Comput.* **2015**, *1*, 1–10.
26. Deane-Mayer, Z.A.; Knowles, J.E.; Deane-Mayer, M.Z.A. Package 'caretEnsemble'. 2016. Available online: https://mirrors.nic.cz/R/web/packages/caretEnsemble/caretEnsemble.pdf (accessed on 1 May 2023).
27. Kursa, M.B.; Rudnicki, W.R. Feature selection with the Boruta package. *J. Stat. Softw.* **2010**, *36*, 1–13. [CrossRef]
28. Perez-Riverol, Y.; Bai, J.; Bandla, C.; Garcia-Seisdedos, D.; Hewapathirana, S.; Kamatchinathan, S.; Kundu, D.J.; Prakash, A.; Frericks-Zipper, A.; Eisenacher, M.; et al. The PRIDE database resources in 2022: A hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res.* **2022**, *50*, D543–D552. [CrossRef]
29. Demichev, V.; Messner, C.B.; Vernardis, S.I.; Lilley, K.S.; Ralser, M. DIA-NN: Neural networks and interference correction enable deep proteome coverage in high throughput. *Nat. Methods* **2020**, *17*, 41–44. [CrossRef]
30. Bai, M.; Deng, J.; Dai, C.; Pfeuffer, J.; Sachsenberg, T.; Perez-Riverol, Y. LFQ-Based Peptide and Protein Intensity Differential Expression Analysis. *J. Proteome Res.* **2023**, *22*, 2114–2123. [CrossRef] [PubMed]
31. Chen, C.; Hou, J.; Tanner, J.J.; Cheng, J. Bioinformatics methods for mass spectrometry-based proteomics data analysis. *Int. J. Mol. Sci.* **2020**, *21*, 2873. [CrossRef] [PubMed]
32. Lin, M.-H.; Wu, P.-S.; Wong, T.-H.; Lin, I.-Y.; Lin, J.; Cox, J.; Yu, S.-H. Benchmarking differential expression, imputation and quantification methods for proteomics data. *Brief. Bioinform.* **2022**, *23*, bbac138. [CrossRef] [PubMed]

33. Spratt, H.M.; Ju, H. Statistical Approaches to Candidate Biomarker Panel Selection. *Adv. Exp. Med. Biol.* **2016**, *919*, 463–492. [CrossRef] [PubMed]
34. Dubois, E.; Galindo, A.N.; Dayon, L.; Cominetti, O. Comparison of normalization methods in clinical research applications of mass spectrometry-based proteomics. In Proceedings of the 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), Vina del Mar, Chile, 27–29 October 2020; IEEE Publisher: Vina del Mar, Chile, 2020; pp. 1–10. [CrossRef]
35. Callister, S.J.; Barry, R.C.; Adkins, J.N.; Johnson, E.T.; Qian, W.-j.; Webb-Robertson, B.-J.M.; Smith, R.D.; Lipton, M.S. Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *J. Proteome Res.* **2006**, *5*, 277–286. [CrossRef]
36. Zhao, Y.; Wong, L.; Goh, W.W.B. How to do quantile normalization correctly for gene expression data analyses. *Sci. Rep.* **2020**, *10*, 1–11. [CrossRef]
37. Johnson, W.E.; Li, C.; Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **2007**, *8*, 118–127. [CrossRef]
38. Čuklina, J.; Lee, C.H.; Williams, E.G.; Sajic, T.; Collins, B.C.; Rodríguez Martínez, M.; Sharma, V.S.; Wendt, F.; Goetze, S.; Keele, G.R. Diagnostics and correction of batch effects in large-scale proteomic studies: A tutorial. *Mol. Syst. Biol.* **2021**, *17*, e10240. [CrossRef]
39. Kong, W.; Hui, H.W.H.; Peng, H.; Goh, W.W.B. Dealing with missing values in proteomics data. *Proteomics* **2022**, *22*, 2200092. [CrossRef]
40. Wei, R.; Wang, J.; Su, M.; Jia, E.; Chen, S.; Chen, T.; Ni, Y. Missing value imputation approach for mass spectrometry-based metabolomics data. *Sci. Rep.* **2018**, *8*, 1–10. [CrossRef]
41. Hasan, M.K.; Alam, M.A.; Roy, S.; Dutta, A.; Jawad, M.T.; Das, S. Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021). *Inform. Med. Unlocked* **2021**, *27*, 100799. [CrossRef]
42. Stekhoven, D.J.; Bühlmann, P. MissForest—Non-parametric missing value imputation for mixed-type data. *Bioinformatics* **2012**, *28*, 112–118. [CrossRef]
43. Troyanskaya, O.; Cantor, M.; Sherlock, G.; Brown, P.; Hastie, T.; Tibshirani, R.; Botstein, D.; Altman, R.B. Missing value estimation methods for DNA microarrays. *Bioinformatics* **2001**, *17*, 520–525. [CrossRef] [PubMed]
44. Wang, S.; Li, W.; Hu, L.; Cheng, J.; Yang, H.; Liu, Y. NAguideR: Performing and prioritizing missing value imputations for consistent bottom-up proteomic analyses. *Nucleic Acids Res.* **2020**, *48*, e83. [CrossRef]
45. West, R.M. Best practice in statistics: Use the Welch *t*-test when testing the difference between two groups. *Ann. Clin. Biochem.* **2021**, *58*, 267–269. [CrossRef]
46. Wieczorek, S.; Combes, F.; Borges, H.; Burger, T. Protein-level statistical analysis of quantitative label-free proteomics data with ProStaR. *Proteom. Biomark. Discov. Methods Protoc.* **2019**, *1959*, 225–246.
47. Giai Gianetto, Q.; Combes, F.; Ramus, C.; Bruley, C.; Couté, Y.; Burger, T. Calibration plot for proteomics: A graphical tool to visually check the assumptions underlying FDR control in quantitative experiments. *Proteomics* **2016**, *16*, 29–32. [CrossRef]
48. Lo, S.W.; Segal, J.P.; Lubel, J.S.; Garg, M. What do we know about the renin angiotensin system and inflammatory bowel disease? *Expert Opin. Ther. Targets* **2022**, *26*, 897–909. [CrossRef]
49. Peuhkuri, K.; Vapaatalo, H.; Korpela, R. Even low-grade inflammation impacts on small intestinal function. *World J. Gastroenterol. WJG* **2010**, *16*, 1057. [CrossRef] [PubMed]
50. Geremia, A.; Jewell, D.P. The IL-23/IL-17 pathway in inflammatory bowel disease. *Expert Rev. Gastroenterol. Hepatol.* **2012**, *6*, 223–237. [CrossRef]
51. Liu, H.; Motoda, H. *Computational Methods of Feature Selection*; CRC Press: Boca Raton, FL, USA, 2007; 440p. [CrossRef]
52. Hall, M.A. Correlation-Based Feature Selection for Machine Learning. Ph.D. Thesis, The University of Waikato, Hamilton, New Zaeland, 1999. Available online: https://hdl.handle.net/10289/15043 (accessed on 12 June 2023).
53. Greener, J.G.; Kandathil, S.M.; Moffat, L.; Jones, D.T. A guide to machine learning for biologists. *Nat. Rev. Mol. Cell Biol.* **2022**, *23*, 40–55. [CrossRef]
54. Ying, X. An Overview of Overfitting and Its Solutions. *Proc. J. Phys. Conf. Ser.* **2019**, *1168*, 022022. [CrossRef]
55. Bischl, B.; Binder, M.; Lang, M.; Pielok, T.; Richter, J.; Coors, S.; Thomas, J.; Ullmann, T.; Becker, M.; Boulesteix, A.L. Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2023**, *13*, e1484. [CrossRef]
56. Mantovani, R.G.; Rossi, A.L.; Vanschoren, J.; Bischl, B.; De Carvalho, A.C. Effectiveness of random search in SVM hyper-parameter tuning. In Proceedings of the 2015 International Joint Conference on Neural Networks (IJCNN), Killarney, Ireland, 12–17 July 2015; IEEE Publisher: Killarney, Ireland; pp. 1–8. [CrossRef]
57. Mooiweer, E.; Fidder, H.H.; Siersema, P.D.; Laheij, R.J.; Oldenburg, B. Fecal hemoglobin and calprotectin are equally effective in identifying patients with inflammatory bowel disease with active endoscopic inflammation. *Inflamm. Bowel Dis.* **2014**, *20*, 307–314. [CrossRef]
58. Schröder, O.; Naumann, M.; Shastri, Y.; Povse, N.; Stein, J. Prospective evaluation of faecal neutrophil-derived proteins in identifying intestinal inflammation: Combination of parameters does not improve diagnostic accuracy of calprotectin. *Aliment. Pharmacol. Ther.* **2007**, *26*, 1035–1042. [CrossRef]

59. dos Santos Ramos, A.; Viana, G.C.S.; de Macedo Brigido, M.; Almeida, J.F. Neutrophil extracellular traps in inflammatory bowel diseases: Implications in pathogenesis and therapeutic targets. *Pharmacol. Res.* **2021**, *171*, 105779. [CrossRef]
60. Matsumori, A.; Shimada, T.; Shimada, M.; Drayson, M.T. Immunoglobulin free light chains: An inflammatory biomarker of diabetes. *Inflamm. Res.* **2020**, *69*, 715–718. [CrossRef]
61. Napodano, C.; Pocino, K.; Rigante, D.; Stefanile, A.; Gulli, F.; Marino, M.; Basile, V.; Rapaccini, G.L.; Basile, U. Free light chains and autoimmunity. *Autoimmun. Rev.* **2019**, *18*, 484–492. [CrossRef]
62. Xu, S.; Zhao, L.; Larsson, A.; Venge, P. The identification of a phospholipase B precursor in human neutrophils. *FEBS J.* **2009**, *276*, 175–186. [CrossRef]
63. Fournier, T.; Medjoubi-N, N.; Porquet, D. Alpha-1-acid glycoprotein. *Biochim. Et Biophys. Acta (BBA)-Protein Struct. Mol. Enzymol.* **2000**, *1482*, 157–171. [CrossRef]
64. Watanabe, T.; Aoyagi, K.; Nimura, S.; Eguchi, K.; Tomioka, Y.; Sakisaka, S. New fecal biomarker, α1-acid glycoprotein, for evaluation of inflammatory bowel disease: Comparison with calprotectin and lactoferrin. *Fukuoka Univ. Med. J.* **2013**, *40*, 155–162.
65. Bock, J.; Liebisch, G.; Schweimer, J.; Schmitz, G.; Rogler, G. Exogenous sphingomyelinase causes impaired intestinal epithelial barrier function. *World J. Gastroenterol. WJG* **2007**, *13*, 5217. [CrossRef]
66. Parveen, F.; Bender, D.; Law, S.-H.; Mishra, V.K.; Chen, C.-C.; Ke, L.-Y. Role of ceramidases in sphingolipid metabolism and human diseases. *Cells* **2019**, *8*, 1573. [CrossRef]
67. Snider, A.J.; Wu, B.X.; Jenkins, R.W.; Sticca, J.A.; Kawamori, T.; Hannun, Y.A.; Obeid, L.M. Loss of neutral ceramidase increases inflammation in a mouse model of inflammatory bowel disease. *Prostaglandins Other Lipid Mediat.* **2012**, *99*, 124–130. [CrossRef] [PubMed]
68. Karamizadeh, S.; Abdullah, S.M.; Halimi, M.; Shayan, J.; javad Rajabi, M. Advantage and drawback of support vector machine functionality. In Proceedings of the 2014 International Conference on Computer, Communications, and Control Technology (I4CT), Langkawi, Malaysia, 2–4 September 2014; IEEE Publisher: Langkawi, Malaysia; pp. 63–65. [CrossRef]
69. Burbidge, R.; Buxton, B. An introduction to support vector machines for data mining. *Keynote Pap. Young OR12* **2001**, 3–15. Available online: https://api.semanticscholar.org/CorpusID:8133449 (accessed on 15 June 2023).
70. Singh, A.; Thakur, N.; Sharma, A. A review of supervised machine learning algorithms. In Proceedings of the 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 16–18 March 2016; IEEE Publisher: New Delhi, India; pp. 1310–1315.
71. Hsu, C.-W.; Chang, C.-C.; Lin, C.-J. *A Practical Guide to Support Vector Classification*; National Taiwan University: Taipei, Taiwan, 2003; pp. 1396–1400. Available online: https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf (accessed on 2 July 2023).
72. Huang, S.; Cai, N.; Pacheco, P.P.; Narrandes, S.; Wang, Y.; Xu, W. Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genom. Proteom.* **2018**, *15*, 41–51.
73. Vapnik, V.N. Pattern recognition using generalized portrait method. *Autom. Remote Control* **1963**, *24*, 774–780.
74. Goel, A.; Srivastava, S.K. Role of kernel parameters in performance evaluation of SVM. In Proceedings of the 2016 Second International Conference on Computational Intelligence & Communication Technology (CICT), Ghaziabad, India, 12–13 February 2016; IEEE Publisher: Ghaziabad, India; pp. 166–169. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.