Improving Differentiation of Crohn's Disease and Ulcerative Colitis Proteomes through Protein-Wide Association Study Feature Selection in Machine Learning

Running Title: Protein-Wide Association and Machine Learning in IBD

Mark G. Gorelik[1,2], Aaron J. Gorelik[3], Skye R.S. Fishbein[1,2], Tara Fehlmann[4], Parakkal Deepak[5#], Ryan Bogdan[3], Gautam Dantas[1,2,6,7,8 #], Umang Jain[1#]* and SPARC IBD Investigators[9]

[1]Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO, USA

[2]The Edison Family Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, MO, USA

[3]Department of Psychological & Brain Sciences, Washington University in St. Louis, St. Louis, MO, USA

[4] Crohn's and Colitis Foundation, New York, New York, USA

[5]Division of Gastroenterology, John T. Milliken Department of Medicine, Washington University School of Medicine, St. Louis, MO, USA

[6]Department of Molecular Microbiology, Washington University School of Medicine, St. Louis, MO, USA

[7]Department of Biomedical Engineering, Washington University in St Louis, St. Louis, MO, USA

[8]Department of Pediatrics, Washington University School of Medicine, St. Louis, MO, USA

[9]See below for SPARC-IBD Investigator Affiliations

#-co-senior authors

SPARC-IBD Investigators

Site	Principal Investigator
Baylor College of Medicine	Richa Shukla, MD
Baylor University Medical Center	Themistocles Dassopoulos, MD
Brigham & Women's Hospital	Scott B. Snapper, MD, PhD Joshua R. Korzenik, MD
Indiana University	Matthew Bohm, MD
Mayo Clinic	Laura Raffals, MD
Medical College of Wisconsin	Poonam Beniwal-Patel, MD
NYU Langone Medical Center	David Hudesman, MD
Scripps Healthcare	Mazer Ally, MD Gauree Konijeti, MD Rebecca Matro, MD
University Gastroenterology	Sheldon Lidofsky, MD
University of Alabama	Kirk Russ, MD
University of Cincinnati Medical Center	Loren Brook, MD
University of Chicago	Joel Pekow, MD
University of Maryland	Raymond Cross, MD
University of Michigan	Shrinivas Bishu, MD
University of Pennsylvania	Meenakshi Bewtra, MD, MPH, PhD James D Lewis, MD, MSCE
University of Pittsburgh	Richard Duerr, MD
University of Wisconsin	Sumona Saha MD, MS Freddy Caldera DO, MS
Vanderbilt University Medical Center	Elizabeth Scoville, MD, MSCI
Washington University School of Medicine	Parakkal Deepak, MBBS, MS

Funding information: This work is supported by SPARC IBD PLEXUS Grant from Crohn's & Colitis Foundation and NIH Washington University-DDRCC Grant Number P30 DK052574 to U.J. M.G.G and S.R.F are supported by awards from the Pediatric Gastroenterology Research Training Program (T32 DK077653). A.J.G is supported by NSF (DGE-213989). PD is supported by a Junior Faculty Development Award from the American College of Gastroenterology and IBD Plexus of the Crohn's & Colitis Foundation.

Presentation at a meeting: A portion of this study was presented at the 2024 Digestive and Diseases Week (May 18-21) in Washington D.C.

Acknowledgements: The authors would like to thank Sarah E. Paul for her initial analytical support and Kevin S. Blake of the LGM Scientific Editing Service at the Department of Pathology and Immunology, Washington University School of Medicine for scientific editing support. The authors also thank the staff at The Edison Family Center for Genome Sciences & Systems Biology at the Washington University School of Medicine in St Louis, including E. Martin and B. Koebbe for computational support, and B. Dee, K. Matheny, J. Theodore and K. Page for administrative support. Finally, we would like to thank the members of the Dantas, Bogdan, and Jain labs for helpful general discussions and comments on the manuscript. The results published here are in whole based on data from the Study of a Prospective Adult Research Cohort with IBD (SPARC IBD). SPARC IBD is a component of the Crohn's & Colitis Foundation's IBD Plexus data exchange platform. SPARC IBD enrolls patients with an established or new diagnosis of IBD from sites throughout the United States and links data collected from the electronic health record and study specific case report forms. Patients also provide blood, stool and biopsy samples at selected times during follow-up. The design and implementation of the SPARC IBD cohort has been previously described[1]. The SPARC IBD data

are available upon approved application to Crohn's & Colitis Foundation IBD Plexus (https://www.crohnscolitisfoundation.org/ibd-plexus).

Conflict of Interest. PD: has received research support under a sponsored research agreement unrelated to the data in the paper and/or consulting from AbbVie, Arena Pharmaceuticals, Boehringer Ingelheim, Bristol Myers Squibb, Janssen, Pfizer, Prometheus Biosciences, Takeda Pharmaceuticals, Roche Genentech, Scipher Medicine, Fresenius Kabi, Teva Pharmaceuticals, Landos Pharmaceuticals, Iterative scopes and CorEvitas, LLC. U.J. has received research support from Boehringer Ingelheim.

Ethics Approval. The study protocol was approved by the Institutional Review Board (IRB) at the University of Pennsylvania which is the single IRB for the SPARC IBD study.

Author contributions: M.G.G., P.D., R.B., G.D. and U.J designed the study, T.F., P.D., and U.J. collected and acquired data, M.G.G., A.J.G., S. R.S.F., and T.F.analysed data, M.G.G., and U.J wrote the manuscript. All authors approved the manuscript.

Abbreviations: Inflammatory bowel diseases (IBD) , Protein-Wide Association Study (PWAS), Crohn's disease (CD), Ulcerative colitis (UC)

Abstract:

Background and Aims: Diagnostic differentiation between Crohn's disease (CD) and ulcerative colitis (UC) is crucial for timely and suitable therapeutic measures. The current gold standard for differentiating between CD and UC involves endoscopy and histology, which are invasive and costly. We aimed to identify blood plasma proteomic signatures using a Protein-Wide Association Study (PWAS) approach to differentiate CD from UC and evaluate the efficacy of these signatures as features in machine learning (ML) classifiers.

Methods: Among participants (n=1,106; n_{CD}=636; n_{UC}=470) of the Study of a Prospective Adult Research Cohort with IBD (SPARC), plasma protein (n=2,920) levels were estimated using Olink proteomics. A PWAS with Bonferroni correction for multiple testing was used to identify proteins associated with disease states after controlling for age, sex, and disease severity. ML classifiers examined the diagnostic utility of these models. Feature importance was determined via SHapley Additive exPlanations (SHAP) analysis.

Results: Thirteen proteins which were significantly differentially abundant in CD vs UC (all |β|s > 0.22, all adjusted p values < 8.42E-06). Random forest models of proteins differentiated between CD and UC with models trained only on PWAS identified proteins (Average ROC-AUC 0.73) outperforming models trained of the full proteome (Average ROC-AUC 0.62). SHAP analysis revealed that Granzyme B, insulin-like peptide 5 (INSL5), and interleukin-12 subunit beta (IL-12B) were the most important features.

Conclusions: Our findings demonstrate that PWAS-based feature selection approaches are a powerful method to identify features in complex, noisy datasets. Importantly, we have identified novel peptide based biomarkers such as INSL5, that can be potentially used to complement existing strategies to differentiate between CD and UC.

Keywords: Machine Learning, PWAS, IBD

INTRODUCTION

Inflammatory bowel diseases (IBD) are chronic relapsing and remitting inflammatory disorders of the gastrointestinal tract. They affect more than 6 million people worldwide, and in the United States alone more than 70,000 cases of IBD are diagnosed each year[2,3]. Patients with IBD experience markedly decreased quality of life, high disease- and treatment-related morbidity, and often endure complications requiring hospitalizations and surgeries[4–8]. IBD is generally subtyped as either Crohn's disease (CD) or ulcerative colitis (UC), with each differing in the areas of manifestation and the resulting sequela[9–11]. Specifically, CD can affect any region of the gastrointestinal (GI) tract and generally presents with transmural inflammation, while UC is restricted to the colon and is characterized by mucosal ulceration[9–11]. While CD and UC present distinct clinical complications, CD's ability to affect any region of the GI tract, including regions affected by UC, makes discriminating between them challenging[12–14].

As each disease requires distinct therapeutic strategies, being able to accurately and efficiently differentiate CD from UC has significant consequences for clinical care. For example, surgery is not a definitive cure for CD and can result in further complications[15–19]. Current practices rely on endoscopy to discriminate CD from UC; however, endoscopy is invasive, expensive, and carries significant risk to the patient[20]. To complement endoscopic procedures, blood and fecal markers are often used; however, none of these tests have proven sufficient to enable the differentiation of CD and UC[21–27]. For instance, serum antibodies against *Saccharomyces cerevisiae* (ASCA) and bacterial antigens have limited accuracy and suffer from low sensitivity, rendering these tests relatively nonspecific to subtype IBD[21–24]. Other markers such as fecal calprotectin and Lipocalin-2 can identify inflammatory status but do not enable differentiation between CD and UC[25–27]. Given the rising prevalence of IBD worldwide, its high morbidity, and its substantial negative impact on quality of life, there is an urgent need for

diagnostic tools that enable early differentiation of CD from UC, and are easier to use, non-invasive, and less costly than those currently available[28].

Advanced proteomics technologies offer novel avenues for comprehending pathophysiological mechanisms and pinpointing potential clinical biomarkers in complex diseases. Recent breakthroughs, exemplified by the Olink platform, have revealed novel protein biomarkers for multiple diseases in blood and plasma [29,30,31], ensuring heightened sensitivity, precision, and specificity, while also requiring minimal sample volumes. The output data of the Olink platform can be then applied as features for machine learning (ML)-based classification analysis[32]. Unfortunately, even with these technological advancements, omics data is still often affected by the "curse of dimensionality"[33], where the number of features captured far exceeds the number of samples which can result in models fitting to spurious patterns[33]. ML models trained on high dimensionality data may fail to generalize to real world data unless the sample size is sufficiently large enough (normally at least 5 samples per feature[34]) to separate signal versus noise[33]. However, generating such large omics datasets can be both costly and time consuming. To mitigate the "curse of dimensionality" without the costly and time-consuming process of generating large omics datasets, feature selection methods are often used to identify informative features in high dimensionality datasets before model training[33,35]. In particular, GWAS (Genome-Wide Association Study) and PheWAS (Phenome-Wide Association Study) approaches have proven to be extremely effective for feature selection[35]. Here we leverage a PWAS (Protein-Wide Association Study)-based approach to identify informative features in a high dimensionality Olink proteomics dataset from IBD patients. This approach identified 13 proteins which distinguish CD from UC using plasma samples.

MATERIALS & METHODS

Participants and Sample Collection

The Study of Prospective Adult Research Cohort of IBD (SPARC IBD) is an ongoing longitudinal cohort study of patients with IBD recruited from 17 academic medical centers across the United States[1]. Plasma samples used in this study were obtained from N = 1106 individuals (n_{CD} = 636; n_{UC} =470).

Demographic, disease-related, and patient-reported data were collected during the following visits: 1) during routine GI office visits (2016-2021), 2) quarterly by sending surveys to patients, and 3) before a scheduled colonoscopy. All collections generated highly structured electronic case report forms (eCRF). Bio-samples of each respective patient's blood and stool were collected at enrolment and at the time of each patient's colonoscopy. Further, blood samples were collected if a patient or provider reported key medication changes. Initially collected samples were used in this study. Clinical data is transferred from sites on a periodic basis and stored in IBD Plexus, Crohn's & Colitis Foundation's exchange platform (see[1] for details).

Olink Proteomics, normalization, and filtering

Plasma was purified from blood and stored in EDTA. Proteins within plasma were estimated using Olink Explore 384 panels (i.e., Cardiometabolic, Cardiometabolic II, Inflammation, Inflammation II, Neurology, Neurology II, Oncology, Oncology II panels; Olink Proteomics) Protein levels were estimated as Olink's arbitrary units, Normalized Protein eXpression (NPX) values on a $\log_2$ scale. NPX values which did not pass the following quality control metrics were filtered out: 1) at least 500 counts per specific combination of sample and assay, 2) the deviation from the median value of the incubation- and amplification controls for each individual sample did not exceed +/-0.3 NPX for either of the internal controls, and 3) the deviation of the median of the negative controls must be ≤5 standard deviations from the set predefined manufacturer value. Samples across plates were normalized via the intensity

normalization method. The following Explore 384 assays did not meet Olink's batch release quality control criteria and are therefore not included in this study: KNG1 (Inflammation II), TNFSF9 (Inflammation II), TOM1L2 (Neurology II), SMAD1 (Oncology), and ARHGAP25 (Oncology).

Statistical Analyses

Protein Wide Association Study (PWAS)

A Protein Wide Association Study (PWAS) was performed on all proteins passing quality control described above (n=2,920) using the glmer function in the lme4 package[36] as previously described in phenome-wide association studies[37,38]. . CD/UC disease status was regressed on each individual protein in a mixed effects logistic regression with age and sex as fixed effects covariates and disease activity (Simple Crohn's Disease Activity Index for CD and 6-point Mayo Score for UC[1]) was treated as a random effect. To adjust for multiple testing, a Bonferroni-corrected proteome wide significance threshold was used (0.05/2, 920 = 0.0000171 alpha level).

Principle Component Analysis (PCA)

Principal component analysis (PCA) was initially performed incorporating the measured values of all proteins and just proteins identified via the PWAS analysis. Analysis of Similarities (ANOSIM) was performed using the ANOSIM function in the vegan package[39].

Machine Learning Methods

Using Scikit learn based implementations of random forests we tested the following feature sets: All proteomics features and patient features (Age, Sex, Disease Severity), proteomics features which passed the Bonferroni cutoff and patient features, and just proteomics features. Of the samples, 20% were reserved for a holdout validation dataset which was also used for SHAP value analysis[40]. The remaining 80% of the data was split into a train/test split (70/30) and cross validated 30 times.

The following packages and versions were used for analysis in R v4.3.2: ibdplexus (0.1.0), tidyverse (1.3.1), stringr (1.5.1), readxl (1.4.3), OlinkAnalyze (3.7.0), data.table (1.15.0), lmerTest (3.1), lme4 (1.1)[36], readxl (1.4.3), dplyr (1.1.4), ggplot2 (3.4.3), reshape2 (1.4.4), ggrepel (0.9.5), forcats (1.0.0), ggsci (3.0.0), RColorBrewer (1.1.3), optimx (2023-10.21), minqa (1.2.6), dfoptim (2023.1.0), survey (4.2.1), scales (1.3.0), ggnewscale (0.4.10), ggpubr (0.6.0), gplots (3.1.31), psych (2.4.1), MuMIn (1.47.5), vegan (2-6-6.1), NatParksPalettes (0.2.0), and ggfortify (0.4.16).

The following packages and versions were used for analysis in Python v3.10.9: pandas[41](1.5.3), sklearn[42,43] (1.3.2), numpy[44] (1.23.5), and shap[40](0.43.0).

Formulas

TP: True Positive

TN: False Negative

FP: False Positive

FN: False Negative

Accuracy = (TP+FN)/(TP+FP+TN+FN)

Sensitivity=TP/(TP+FN)

Specificity=TN/(TN+FP)

RESULTS

Subject Characteristics

Details of the subjects' characteristics are shown in Table 1. A total of 1,106 individuals (CD n = 636; UC n=470) from 17 medical centers were included. CD patients were on average, 42.26 years old and 62% of the CD patients were female. UC patients were on average 43.98 years old and 50.95% of the total UC patients were female.

Specific proteins differentiate the proteomic profiles of Crohn's disease and Ulcerative colitis

The studied plasma proteome dataset includes measures of 2,920 protein levels across 1,106 patients from the Crohn's and Colitis Foundation dataset[1] **(Table 1, Fig 1)**. We initially ran a principal component analysis (PCA) (**Fig 2A**) and an analysis of similarities (ANOSIM) which revealed that the global proteomes of CD and UC do not differ (p=0.21, R=0.004). We then conducted a Protein Wide Association Study (PWAS) analysis adapted from previous study [38,45] to filter for proteins measured at significantly different levels between CD and UC. Age and sex were included as fixed effects with disease severity as random effect. The PWAS-based approach identified thirteen proteins that were significant after Bonferroni correction **(Table 2, Fig 2B, Table S1)**. Five of these proteins, INSL5, IL12B, IL12AB, HRG, and LY96 were more abundant in CD relative to UC; in contrast, eight proteins, FGF19, EPCAM, NOS2, GPA33, GUC2A, GRAB, FGFR4, MMP10 were more abundant in UC relative to CD (**Fig S1**). Performing an ANOSIM test and PCA analysis (**Fig 2C**) on the proteins identified via PWAS after multiple test correction revealed that there was significant a difference between the CD and UC cohorts (p=0.001, R=0.1247).

Grouping Specific proteins improve prediction of Crohn's disease and Ulcerative Colitis

We aimed to determine whether the PWAS identified proteomic features could lead to improved differentiation of CD and UC via ML classification. To do this we trained random forest models on feature sets composed of different combinations of proteomic features and patient

features (age, sex, disease severity). We generated three features sets which contained the following: 1) the entire proteome and patient features ("Full Feature Set"), 2) subset of proteomic features which only included the thirteen significant proteins (e.g., INSL5) identified by the PWAS and patient features (referred to as "PWAS and Patient Features"), and 3) just the thirteen significant proteins identified by the PWAS without patient features ("PWAS Features"). Models trained on the "Full Feature Set" had significantly higher specificity, but significantly lower accuracy, sensitivity, and ROC-AUC scores compared to the other two feature sets **(Fig 3A-D, Table 3**). Models trained on "PWAS and Patient Features" and just "PWAS Features" were significantly more accurate and sensitive than models trained on the "Full Feature Set," they did not differ significantly in performance from each other (**Fig 3A-D**). Next, we interrogated the ML models using SHapley Additive exPlanations (SHAP) analysis, which can infer feature importance, to determine if patient features were important for model performance. Interestingly, SHAP feature importance analysis suggested patient features were not informative, where Age, Disease Severity, and Sex were the ranked as the three least important features in the "PWAS and Patient Features" model suggesting that the patient features contributed the least to model performance **(Fig 2B, 4AB, S1)**. Notably, interrogating the ML models revealed that the three most important features including granzyme B, Insulin-like peptide 5 (INSL5), and Interleukin 12B (IL12B) were conserved between models. **(Fig 4AB)**. This demonstrates that PWAS-based approaches act as a filter for identifying more informative features which in turn improve prediction performance.

DISCUSSION

Accurately classifying subtypes of IBD poses a significant clinical challenge. Identification of noninvasive biomarkers that can increase the accuracy of diagnosing and subtyping IBD is a major unmet need. There has been a growing interest in using proteomics to identify new biomarkers for the differentiation of IBD; however, these studies have been limited by the number of proteins measured[46,47]. Here, we used a highly sensitive proximity extension assay and measured 2920 proteins in the plasma of IBD patients. Protein wide association analysis with age and sex as fixed effects identified 13 proteins that are significantly different between CD and UC. Further, using used multiple feature sets in random forest models, we discovered that PWAS identified proteins could distinguish between CD and UC with high accuracy and sensitivity. Taken together, we have identified a novel set of proteins in blood that can potentially complement other existing biomarkers to accurately subtype IBD.

Machine learning algorithms are increasingly being utilized to analyze medical data to diagnose diseases, predict their severity, and monitor their progression. Recent work on diagnosing IBD using ML approaches has also been successful, achieving high levels of performance[12,48–50]. For instance, supervised learning models on RNA sequencing data enabled CD and UC differentiation[12]. Similarly, deep learning networks have been used on endoscopic images to accurately predict the severity of the disease in IBD[48,49]. Although proteomic datasets have been generated in IBD, the application of ML techniques to analyze such datasets has been limited[47]. Furthermore, previous IBD-focused proteomics datasets have measured smaller panels of proteins[47,51–53]. We used a combined PWAS-based feature selection and ML models on a large dataset of proteins to identify novel signatures that could accurately subtype IBD. Importantly, in contrast to previous studies that have primarily focused on inflammatory markers, we analyzed proteins that are involved in a diverse array of processes including, hormonal

regulation, inflammation, cancer, and brain gut axis. Our findings suggest that PWAS based ML approaches could improve subtyping of IBD patients.

Several proteins in our cohort have been validated by other IBD studies focused on differentiating CD from UC. A study by Bourgonje et al. also employed a proximity extension assay (Olink) and measured 92 proteins identifying FGF19, IL12B, and MMP10 to be differentially abundant between CD and UC[53]. In another study by Di Narzo et al., the authors used a SOMAmer-based capture array to measure protein levels in plasma (n=244) and discovered that Granzyme B, FGF19, and MMP10 were downregulated in CD relative to UC, mirroring our results[52]. Importantly, our findings combined with others have identified FGF19 and MMP10 as consistent plasma-based biomarkers which can be used differentiate CD from UC[52,53].

Among the 13 differentially abundant proteins significant in the PWAS after multiple correction, Granzyme B, IL12B, and INSL5 were the most informative for model prediction (**Fig 4A-B**). INSL5, to date, has not been measured in similar proteomics studies focused on IBD[52,53]. Notably, depletion of *INSL5* transcripts in mucosal tissue has been associated with IBD[54], and our study further implicates the INSL5 peptide as differentially abundant between CD and UC. INSL5 is a peptide hormone that is expressed in the colonic epithelium [54–56]. Because INSL5 is a microbially regulated molecule, it is possible that UC, but not CD, specific microbes alter its production[57]. Indeed, both bacterial and fungal microbiota are known to be different between UC and CD[58–60]. Another possible reason for the decreased abundance of INSL5 in UC relative to CD is the loss of colonic epithelial cells due to ulceration, a prominent feature of UC[61]. Future studies are needed to elucidate how INSL5 is regulated and the mechanisms by which INSL5 modulates the severity of the disease[54,62]. In addition to INSL5, Granzyme B was also a powerful predictive feature and was elevated in UC. Granzyme B is a serine protease released by lymphocytes which can trigger apoptosis[63,64]. Similar to our findings, Di Narzo et al. identified

elevated Granzyme B protein levels in CD compared to UC[52]. Further, levels of Granzyme B have been reported to predict treatment responses in IBD as its levels are significantly lower in responders compared to non-responder populations[65]. While these initial findings are promising, it is unclear how the levels of these targets fluctuate throughout disease specific treatments and subtype. Future studies utilizing longitudinal samples are needed to ascertain its association with IBD subtypes.

Our study has several strengths: (1) We utilized a relatively large sample size with patients from 17 different medical centers; (2) Because the SPARC cohort follows standard guidelines, it allows investigators to maintain consistency in both data and bio-sample collection; (3) Our data analysis controlled for multiple parameters including age and sex; (4) We assessed over 2,900 proteins using the Olink platform, enabling us to capture differences across a wide range. Limitations include: (1) absence of a healthy cohort, (2) a single time point of blood collection, (3) and need for validation in non-North American cohorts. Future studies including a healthy control group and longitudinal data would enable exploration of the complex nature of IBD, focusing on the complex spatial-temporal dynamics of IBD location and flare up. This would add important context for leveraging proteins such as INSL5 whether alone or in combination with other markers to differentiate between CD and UC.

Overall, the results of this study provide evidence that applying a PWAS-based approach to filter for potentially relevant proteins improves ML model predication for differentiation between CD and UC. Importantly, the informative biomarkers identified in our study have not been previously examined in the context of differentiating CD from UC. We speculate that this approach may identify new targets for biomarker research and improve mechanistic understanding of disease states.

References:

1. Raffals LE., Saha S., Bewtra M., Norris C., Dobes A., Heller C., et al. The Development and Initial Findings of A Study of a Prospective Adult Research Cohort with Inflammatory Bowel Disease (SPARC IBD). *Inflammatory Bowel Diseases* 2022;**28**(2):192–9. Doi: 10.1093/ibd/izab071.
2. Kaplan GG. The global burden of IBD: from 2015 to 2025. *Nat Rev Gastroenterol Hepatol* 2015;**12**(12):720–7. Doi: 10.1038/nrgastro.2015.150.
3. Duryee MJ., Ahmad R., Eichele DD., Hunter CD., Mitra A., Talmon GA., et al. Identification of Immunoglobulin G Autoantibody Against Malondialdehyde-Acetaldehyde Adducts as a Novel Serological Biomarker for Ulcerative Colitis. *Clin Transl Gastroenterol* 2022;**13**(4):e00469. Doi: 10.14309/ctg.0000000000000469.
4. Soriano CR., Powell CR., Chiorean MV., Simianu VV. Role of hospitalization for inflammatory bowel disease in the post-biologic era. *World J Clin Cases* 2021;**9**(26):7632–42. Doi: 10.12998/wjcc.v9.i26.7632.
5. The global, regional, and national burden of inflammatory bowel disease in 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet Gastroenterol Hepatol* 2019;**5**(1):17–30. Doi: 10.1016/S2468-1253(19)30333-4.
6. Xavier RJ., Podolsky DK. Unravelling the pathogenesis of inflammatory bowel disease. *Nature* 2007;**448**(7152):427–34. Doi: 10.1038/nature06005.
7. Mitropoulou M-A., Fradelos EC., Lee KY., Malli F., Tsaras K., Christodoulou NG., et al. Quality of Life in Patients With Inflammatory Bowel Disease: Importance of Psychological Symptoms. *Cureus* n.d.;**14**(8):e28502. Doi: 10.7759/cureus.28502.
8. Bernstein CN., Nabalamba A. Hospitalization, Surgery, and Readmission Rates of IBD in Canada: A Population-Based Study. *Official Journal of the American College of Gastroenterology | ACG* 2006;**101**(1):110.
9. Panaccione R. Mechanisms of Inflammatory Bowel Disease. *Gastroenterol Hepatol (N Y)* 2013;**9**(8):529–32.
10. Graham DB., Xavier RJ. Pathway Paradigms Revealed from the Genetics of Inflammatory Bowel Disease. *Nature* 2020;**578**(7796):527–39. Doi: 10.1038/s41586-020-2025-2.
11. Guan Q. A Comprehensive Review and Update on the Pathogenesis of Inflammatory Bowel Disease. *J Immunol Res* 2019;**2019**:7247238. Doi: 10.1155/2019/7247238.
12. Park S-K., Kim S., Lee G-Y., Kim S-Y., Kim W., Lee C-W., et al. Development of a Machine Learning Model to Distinguish between Ulcerative Colitis and Crohn's Disease Using RNA Sequencing Data. *Diagnostics (Basel)* 2021;**11**(12):2365. Doi: 10.3390/diagnostics11122365.
13. von Stein P., Lofberg R., Kuznetsov NV., Gielen AW., Persson J., Sundberg R., et al. Multigene Analysis Can Discriminate Between Ulcerative Colitis, Crohn's Disease, and Irritable Bowel Syndrome. *Gastroenterology* 2008;**134**(7):1869–81. Doi: 10.1053/j.gastro.2008.02.083.
14. Tontini GE., Vecchi M., Pastorelli L., Neurath MF., Neumann H. Differential diagnosis in inflammatory bowel disease colitis: State of the art and future perspectives. *World J Gastroenterol* 2015;**21**(1):21–46. Doi: 10.3748/wjg.v21.i1.21.
15. Kanazawa A., Yamana T., Okamoto K., Sahara R. Risk Factors for Postoperative Intra-abdominal Septic Complications after Bowel Resection in Patients with Crohn's Disease. *Diseases of the Colon & Rectum* 2012;**55**(9):957. Doi: 10.1097/DCR.0b013e3182617716.
16. Lewis RT., Maron DJ. Efficacy and Complications of Surgery for Crohn's Disease. *Gastroenterol Hepatol (N Y)* 2010;**6**(9):587–96.

17. Guo K., Ren J., Li G., Hu Q., Wu X., Wang Z., et al. Risk factors of surgical site infections in patients with Crohn's disease complicated with gastrointestinal fistula. *Int J Colorectal Dis* 2017;**32**(5):635–43. Doi: 10.1007/s00384-017-2751-6.
18. Post S., Betzler M., von Ditfurth B., Schürmann G., Küppers P., Herfarth C. Risks of intestinal anastomoses in Crohn's disease. *Ann Surg* 1991;**213**(1):37–42.
19. Singh S., Nguyen GC. Management of Crohn's Disease After Surgical Resection. *Gastroenterology Clinics of North America* 2017;**46**(3):563–75. Doi: 10.1016/j.gtc.2017.05.011.
20. Kavic SM., Basson MD. Complications of endoscopy. *The American Journal of Surgery* 2001;**181**(4):319–32. Doi: 10.1016/S0002-9610(01)00589-X.
21. Prideaux L., De Cruz P., Ng SC., Kamm MA. Serological Antibodies in Inflammatory Bowel Disease: A Systematic Review. *Inflammatory Bowel Diseases* 2012;**18**(7):1340–55. Doi: 10.1002/ibd.21903.
22. WALKER LJ., ALDHOUS MC., DRUMMOND HE., SMITH BRK., NIMMO ER., ARNOTT IDR., et al. Anti-Saccharomyces cerevisiae antibodies (ASCA) in Crohn's disease are associated with disease severity but not NOD2/CARD15 mutations. *Clin Exp Immunol* 2004;**135**(3):490–6. Doi: 10.1111/j.1365-2249.2003.02392.x.
23. Zhou G., Song Y., Yang W., Guo Y., Fang L., Chen Y., et al. ASCA, ANCA, ALCA and Many More: Are They Useful in the Diagnosis of Inflammatory Bowel Disease? *Digestive Diseases* 2016;**34**(1–2):90–7. Doi: 10.1159/000442934.
24. Reese GE., Constantinides VA., Simillis C., Darzi AW., Orchard TR., Fazio VW., et al. Diagnostic Precision of Anti-Saccharomyces cerevisiae Antibodies and Perinuclear Antineutrophil Cytoplasmic Antibodies in Inflammatory Bowel Disease. *Official Journal of the American College of Gastroenterology | ACG* 2006;**101**(10):2410.
25. Oikonomou KA., Kapsoritakis AN., Theodoridou C., Karangelis D., Germenis A., Stefanidis I., et al. Neutrophil gelatinase-associated lipocalin (NGAL) in inflammatory bowel disease: association with pathophysiology of inflammation, established markers, and disease activity. *J Gastroenterol* 2012;**47**(5):519–30. Doi: 10.1007/s00535-011-0516-5.
26. Barnes EL., Burakoff R. New Biomarkers for Diagnosing Inflammatory Bowel Disease and Assessing Treatment Outcomes. *Inflamm Bowel Dis* 2016;**22**(12):2956–65. Doi: 10.1097/MIB.0000000000000903.
27. Jukic A., Bakiri L., Wagner EF., Tilg H., Adolph TE. Calprotectin: from biomarker to biological function. *Gut* 2021;**70**(10):1978–88. Doi: 10.1136/gutjnl-2021-324855.
28. Barnes EL., Liew C-C., Chao S., Burakoff R. Use of blood based biomarkers in the evaluation of Crohn's disease and ulcerative colitis. *World J Gastrointest Endosc* 2015;**7**(17):1233–7. Doi: 10.4253/wjge.v7.i17.1233.
29. Wang T., Yang S., Long Y., Li Y., Wang T., Hou Z. Olink proteomics analysis uncovers the landscape of inflammation-related proteins in patients with acute compartment syndrome. *Front Immunol* 2023;**14**:1293826. Doi: 10.3389/fimmu.2023.1293826.
30. Kong T., Qu Y., Zhao T., Niu Z., Lv X., Wang Y., et al. Identification of novel protein biomarkers from the blood and urine for the early diagnosis of bladder cancer via proximity extension analysis. *J Transl Med* 2024;**22**:314. Doi: 10.1186/s12967-024-04951-z.
31. Gong Q., Fu M., Wang J., Zhao S., Wang H. Potential Immune-Inflammatory Proteome Biomarkers for Guiding the Treatment of Patients with Primary Acute Angle-Closure Glaucoma Caused by COVID-19. *J Proteome Res* 2024. Doi: 10.1021/acs.jproteome.4c00325.
32. Diaz-Canestro C., Chen J., Liu Y., Han H., Wang Y., Honoré E., et al. A machine-learning algorithm integrating baseline serum proteomic signatures predicts exercise responsiveness in overweight males with prediabetes. *Cell Rep Med* 2023;**4**(2):100944. Doi: 10.1016/j.xcrm.2023.100944.

33. Berisha V., Krantsevich C., Hahn PR., Hahn S., Dasarathy G., Turaga P., et al. Digital medicine and the curse of dimensionality. *NPJ Digit Med* 2021;**4**:153. Doi: 10.1038/s41746-021-00521-5.
34. *Pattern Recognition*. 2008.
35. Pudjihartono N., Fadason T., Kempa-Liehr AW., O'Sullivan JM. A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. *Front Bioinform* 2022;**2**:927312. Doi: 10.3389/fbinf.2022.927312.
36. Bates D., Mächler M., Bolker B., Walker S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 2015;**67**:1–48. Doi: 10.18637/jss.v067.i01.
37. Bastarache L., Denny JC., Roden DM. Phenome-Wide Association Studies. *JAMA* 2022;**327**(1):75–6. Doi: 10.1001/jama.2021.20356.
38. Gorelik AJ., Paul SE., Karcher NR., Johnson EC., Nagella I., Blaydon L., et al. A Phenome-Wide Association Study (PheWAS) of Late Onset Alzheimer Disease Genetic Risk in Children of European Ancestry at Middle Childhood: Results from the ABCD Study. *Behav Genet* 2023;**53**(3):249–64. Doi: 10.1007/s10519-023-10140-3.
30. Oksanen J., Simpson GL., Kindt R., Legendre P., Minchin PR., O'Hara RB., et al. Vegan□: community ecology package. *Http://Vegan.r-Forge.r-ProjectOrg/* 2010.
40. Lundberg SM., Lee S-I. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc.; 2017.
41. The pandas development team. pandas-dev/pandas: Pandas 2024. Doi: 10.5281/zenodo.10957263.
42. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 2011;**12**(85):2825–30.
43. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., et al. Scikit-learn: Machine Learning in Python 2018. Doi: 10.48550/arXiv.1201.0490.
44. Harris CR., Millman KJ., van der Walt SJ., Gommers R., Virtanen P., Cournapeau D., et al. Array programming with NumPy. *Nature* 2020;**585**(7825):357–62. Doi: 10.1038/s41586-020-2649-2.
45. Brandes N., Linial N., Linial M. PWAS: proteome-wide association study-linking genes and phenotypes by functional variation in proteins. *Genome Biol* 2020;**21**(1):173. Doi: 10.1186/s13059-020-02089-x.
46. Zhao JH., Stacey D., Eriksson N., Macdonald-Dunlop E., Hedman ÅK., Kalnapenkis A., et al. Genetics of circulating inflammatory proteins identifies drivers of immune-mediated disease risk and therapeutic targets. *Nat Immunol* 2023;**24**(9):1540–51. Doi: 10.1038/s41590-023-01588-w.
47. Kalla R., Adams AT., Bergemalm D., Vatn S., Kennedy NA., Ricanek P., et al. Serum proteomic profiling at diagnosis predicts clinical course, and need for intensification of treatment in inflammatory bowel disease. *Journal of Crohn's and Colitis* 2021;**15**(5):699–708. Doi: 10.1093/ecco-jcc/jjaa230.
48. Chierici M., Puica N., Pozzi M., Capistrano A., Donzella MD., Colangelo A., et al. Automatically detecting Crohn's disease and Ulcerative Colitis from endoscopic imaging. *BMC Medical Informatics and Decision Making* 2022;**22**(6):300. Doi: 10.1186/s12911-022-02043-w.
49. Takenaka K., Ohtsuka K., Fujii T., Negi M., Suzuki K., Shimizu H., et al. Development and Validation of a Deep Neural Network for Accurate Evaluation of Endoscopic Images From Patients With Ulcerative Colitis. *Gastroenterology* 2020;**158**(8):2150–7. Doi: 10.1053/j.gastro.2020.02.012.
50. Seeley EH., Washington MK., Caprioli RM., M'Koma AE. Proteomic Patterns of Colonic Mucosal Tissues Delineate Crohn's Colitis and Ulcerative Colitis. *Proteomics Clin Appl* 2013;**7**(0):10.1002/prca.201200107. Doi: 10.1002/prca.201200107.

51. Gisbert JP., Chaparro M. Clinical Usefulness of Proteomics in Inflammatory Bowel Disease: A Comprehensive Review. *J Crohns Colitis* 2019;**13**(3):374–84. Doi: 10.1093/ecco-jcc/jjy158.
52. Di Narzo AF., Brodmerkel C., Telesco SE., Argmann C., Peters LA., Li K., et al. High-Throughput Identification of the Plasma Proteomic Signature of Inflammatory Bowel Disease. *Journal of Crohn's and Colitis* 2019;**13**(4):462–71. Doi: 10.1093/ecco-jcc/jjy190.
53. Bourgonje AR., Hu S., Spekhorst LM., Zhernakova DV., Vich Vila A., Li Y., et al. The Effect of Phenotype and Genotype on the Plasma Proteome in Patients with Inflammatory Bowel Disease. *Journal of Crohn's and Colitis* 2022;**16**(3):414–29. Doi: 10.1093/ecco-jcc/jjab157.
54. Skok DJ., Hauptman N., Jerala M., Zidar N. Expression of Cytokine-Coding Genes BMP8B, LEFTY1 and INSL5 Could Distinguish between Ulcerative Colitis and Crohn's Disease. *Genes (Basel)* 2021;**12**(10):1477. Doi: 10.3390/genes12101477.
55. Thanasupawat T., Hammje K., Adham I., Ghia J-E., Del Bigio MR., Krcek J., et al. INSL5 is a novel marker for human enteroendocrine cells of the large intestine and neuroendocrine tumours. *Oncology Reports* 2013;**29**(1):149–54. Doi: 10.3892/or.2012.2119.
56. Liu C., Kuei C., Sutton S., Chen J., Bonaventure P., Wu J., et al. INSL5 is a high affinity specific agonist for GPCR142 (GPR100). *J Biol Chem* 2005;**280**(1):292–300. Doi: 10.1074/jbc.M409916200.
57. Lee YS., De Vadder F., Tremaroli V., Wichmann A., Mithieux G., Bäckhed F. Insulin-like peptide 5 is a microbially regulated peptide that promotes hepatic glucose production. *Mol Metab* 2016;**5**(4):263–70. Doi: 10.1016/j.molmet.2016.01.007.
58. Jain U., Ver Heul AM., Xiong S., Gregory MH., Demers EG., Kern JT., et al. Debaryomyces is enriched in Crohn's disease intestinal tissue and impairs healing in mice. *Science* 2021;**371**(6534):1154–9. Doi: 10.1126/science.abd0919.
59. Schirmer M., Franzosa EA., Lloyd-Price J., McIver LJ., Schwager R., Poon TW., et al. Dynamics of metatranscription in the inflammatory bowel disease gut microbiome. *Nat Microbiol* 2018;**3**(3):337–46. Doi: 10.1038/s41564-017-0089-z.
60. Franzosa EA., Sirota-Madi A., Avila-Pacheco J., Fornelos N., Haiser HJ., Reinker S., et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nature Microbiology* 2019;**4**(2):293–305. Doi: 10.1038/s41564-018-0306-4.
61. A K., P G. Ulcerative colitis: understanding its cellular pathology could provide insights into novel therapies. *Journal of Inflammation (London, England)* 2020;**17**. Doi: 10.1186/s12950-020-00246-4.
62. Pustovit RV., Zhang X., Liew JJ., Praveen P., Liu M., Koo A., et al. A Novel Antagonist Peptide Reveals a Physiological Role of Insulin-Like Peptide 5 in Control of Colorectal Function. *ACS Pharmacol Transl Sci* 2021;**4**(5):1665–74. Doi: 10.1021/acsptsci.1c00171.
63. Trapani JA., Sutton VR. Granzyme B: pro-apoptotic, antiviral and antitumor functions. *Curr Opin Immunol* 2003;**15**(5):533–43. Doi: 10.1016/s0952-7915(03)00107-9.
64. Kim TJ., Koo JS., Kim SJ., Hong SN., Kim YS., Yang S-K., et al. Role of IL-1ra and Granzyme B as biomarkers in active Crohn's disease patients. *Biomarkers* 2018;**23**(2):161–6. Doi: 10.1080/1354750X.2017.1387933.
65. Heidari P., Haj-Mirzaian A., Prabhu S., Ataeinia B., Esfahani SA., Mahmood U. Granzyme B PET Imaging for Assessment of Disease Activity in Inflammatory Bowel Disease. *Journal of Nuclear Medicine* 2024. Doi: 10.2967/jnumed.123.267344.

548

549

550

551

552

553

554

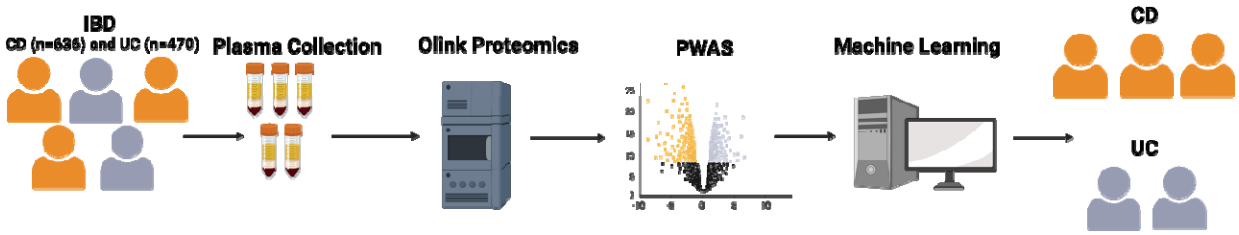
555

556

557

558

559 **Figures**



560

561 **Figure 1. Sample processing and analysis pipeline**. Blood plasma samples were collected

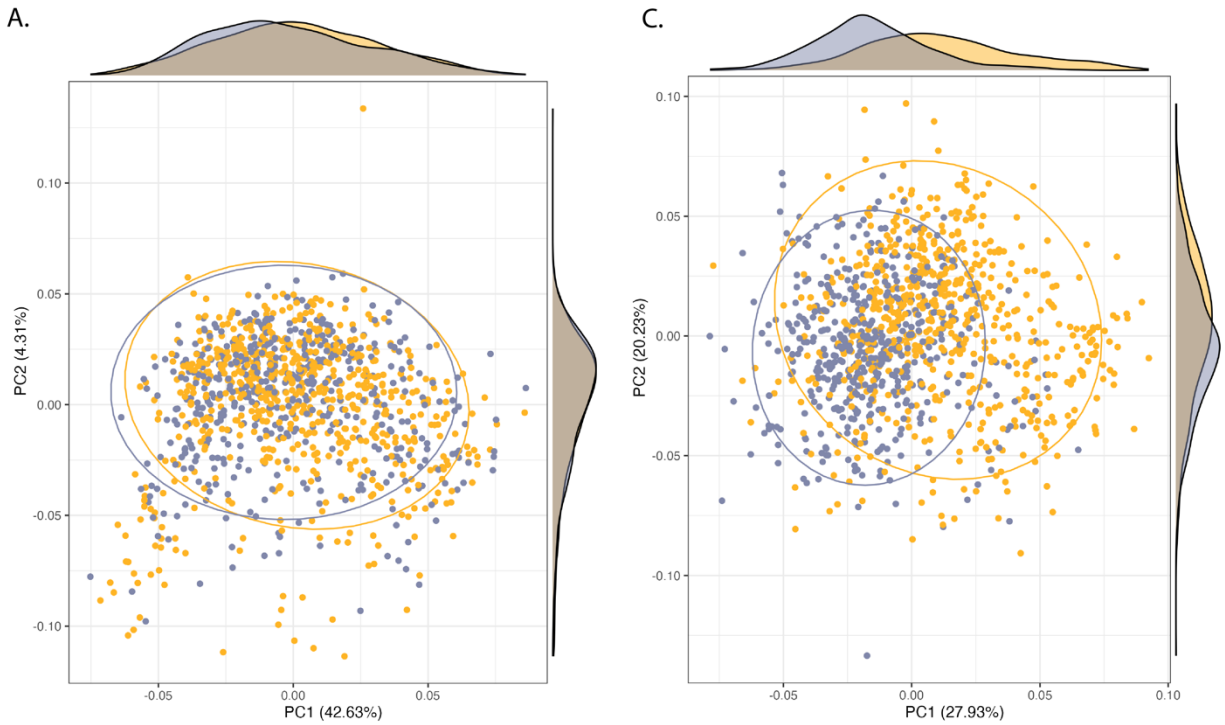
562 and processed as described in the methods and materials. Differentially abundant proteins were

563 identified in the PWAS analysis. Protein abundance was used as features for the machine

564 learning models to classify CD from UC.

565

566



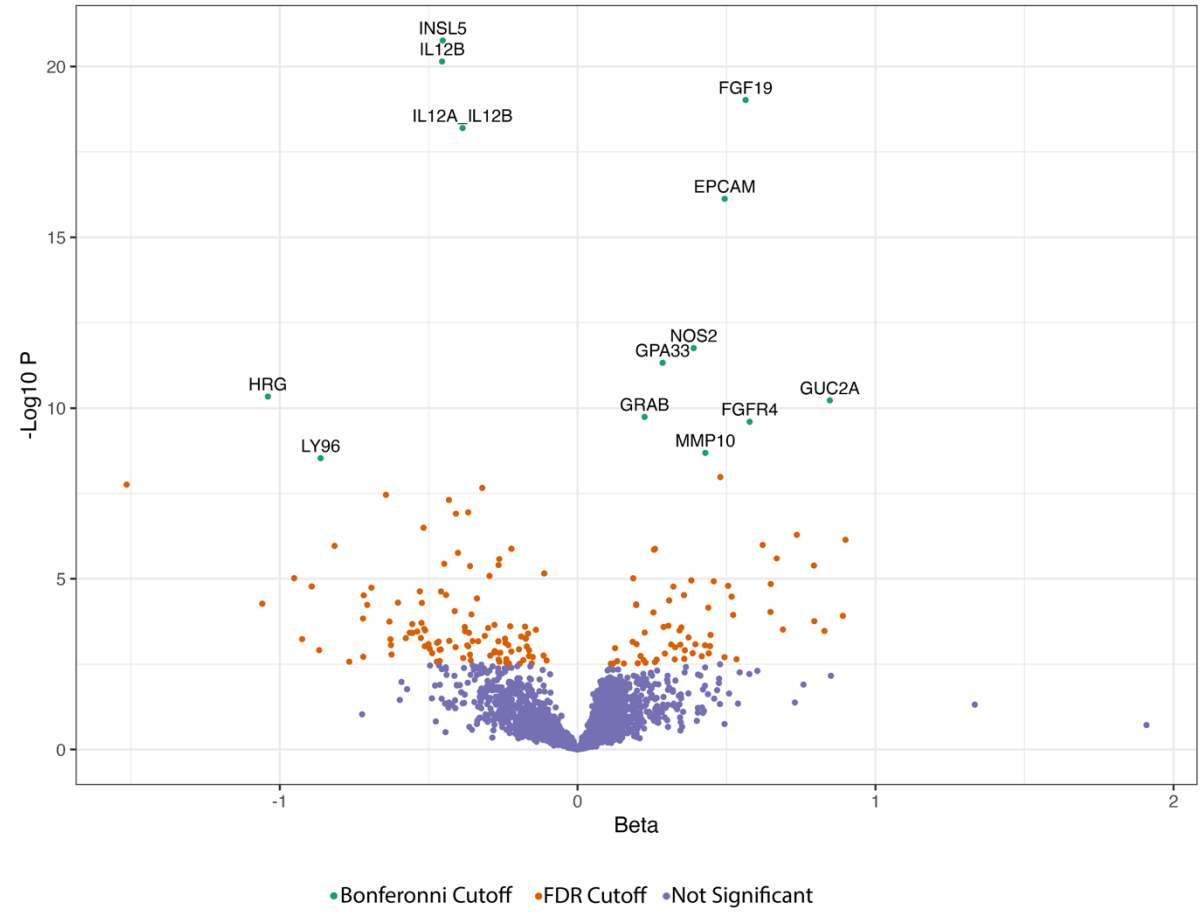
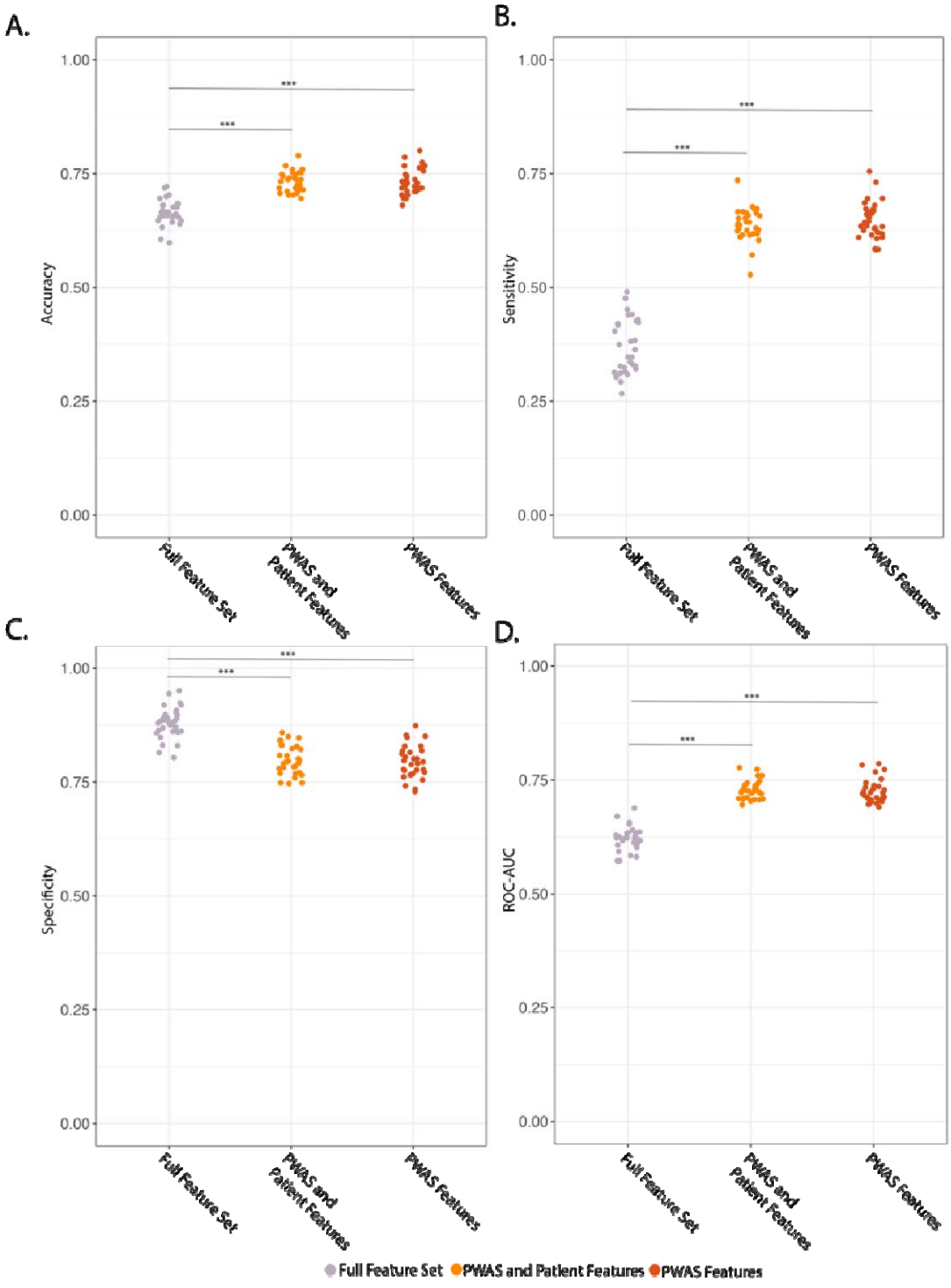


Figure 2. **PWAS analysis enables separation of the proteomic profiles of Ulcerative colitis and Crohn's disease**. A) Principal Component Analysis (PCA) of the global proteomics profiles of Crohn's disease and Ulcerative colitis. B) Volcano plot where the x axis is the calculated beta and the y axis is the negative $\log_{10}$ of the unadjusted p-value; green and labeled points had a Bonferroni adjusted p-value of less than 0.0000171 (used in Fig 1B), orange points had an FDR adjusted p value of less than .05, and purple points represent proteins with a p-value > .05. Negative beta values are associated with Crohn's disease and positive beta values are associated with Ulcerative colitis. C) PCA of the proteomics profile identified by the PWAS analysis. Ellipses represent 95% confidence bounds around group centroids.



579

580

581

Figure 3. Specific proteins improve machine learning based differentiation of CD and UC. A) Effect of feature set on model accuracy. B) Effect of feature set on model sensitivity. C) Effect of feature set on model specificity. D) Effect of feature set on model ROC-AUC. ***P<.001, ANOVA with Tukey's post hoc test.

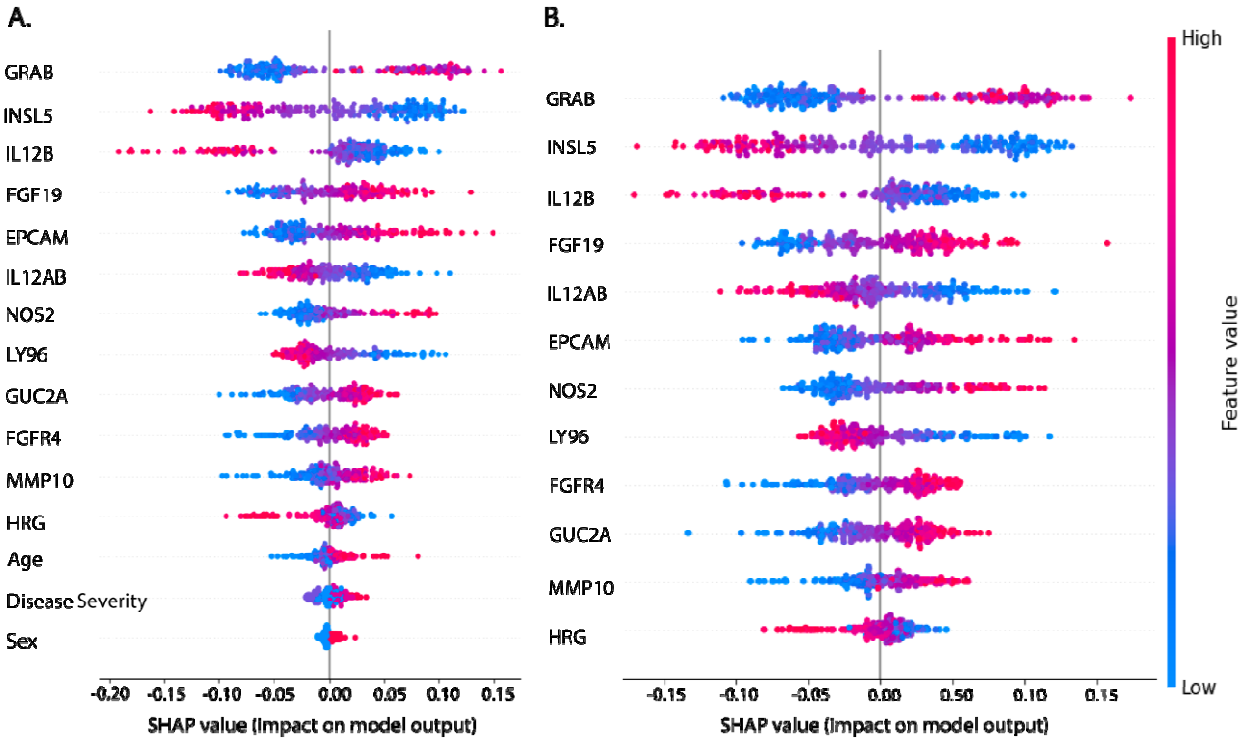
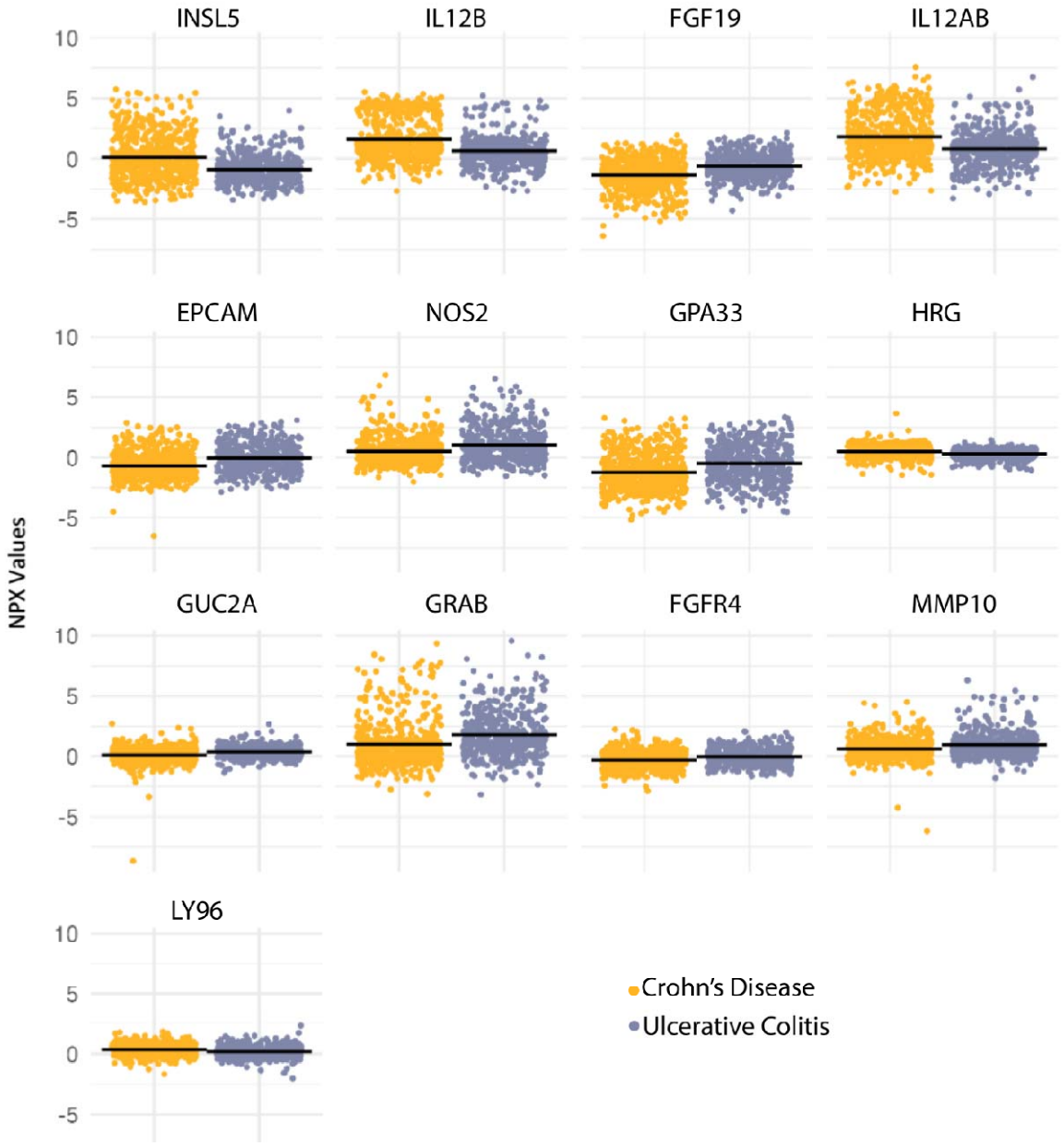


Figure 4. **Clinical Features do not improve model performance**. A) SHAP beeswarm plot of the validation dataset indicating feature importance in random forest models trained on patient associated features (Age, Sex, Disease Severity) and the thirteen proteins which are significantly associated with Crohn's disease and Ulcerative colitis. B) SHAP beeswarm plot of the validation dataset indicating feature importance in random forest models trained on just the thirteen proteins which are significantly associated with Crohn's disease and ulcerative colitis. Features are sorted in order of predicted importance in a descending manner.



609

610

611 **Supplemental figure 1.** The average NPX values for proteins which were significant after

612 Bonferroni correction (Fig 1B).

613

614

615

Table 1. Cohort Breakdown

		Crohn's Disease	***Ulcerative colitis***	***p value***
Number		636	470	
Age in years, mean (SD)		42.26 (14.25)	43.979 (14.6)	.414
Sex (%)	Female	397 (62)	240 (50.95)	.0002
Disease	Remission	228 (35.8)	205 (43.5)	<.0001
Severity (%)	Mild	156 (24.5)	154 (32.7)	
	Moderate	237 (37.2)	88 (18.6)	
	Severe	15 (2.35)	23 (4.88)	

632 **Table 2.** PWAS Results

Protein Name	Average Level in Crohn's Disease	Average Level in Ulcerative Colitis	PWAS Standardized Beta	PWAS p-value	PWAS FDR Adjusted p-value	PWAS Bonferroni Adjusted p-value
INSL5	0.143	-0.93	-0.453	1.73E-21	5.05E-18	5.05E-18
IL12B	1.772	0.683	-0.455	7.15E-21	1.04E-17	2.09E-17
FGF19	-1.398	-0.648	0.564	9.66E-20	9.40E-17	2.82E-16
IL12A-IL12B	1.982	0.839	-0.387	6.39E-19	4.66E-16	1.87E-15
EPCAM	-0.735	-0.087	0.494	7.49E-17	4.37E-14	2.19E-13
NOS2	0.550	1.131	0.390	1.76E-12	8.59E-10	5.15E-09
GPA33	-1.307	-0.515	0.286	4.72E-12	1.97E-09	1.38E-08
HRG	0.492	0.322	-1.040	4.58E-11	1.67E-08	1.34E-07
GUC2A	0.099	0.349	0.847	5.94E-11	1.93E-08	1.73E-07
GRAB	1.016	1.838	0.225	1.81E-10	5.30E-08	5.30E-07
FGFR4	-0.333	-0.054	0.578	2.50E-10	6.65E-08	7.31E-07

MMP10	0.612	0.991	0.429	2.03E-09	4.94E-07	5.93E-06
LY96	0.383	0.242	-0.863	2.89E-09	6.48E-07	8.42E-06

Table 3. Average Machine Learning Model Results

Metric	Full Feature Set	PWAS and Patient Features	Patient Features
Accuracy	66.29%	73.48%	72.6%
Sensitivity	0.36	0.64	0.63
Specificity	0.88	0.80	0.79
ROC-AUC	0.62	0.73	0.73

654