

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

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

4 Mark G. Gorelik^{1,2}, Aaron J. Gorelik³, Skye R.S. Fishbein^{1,2}, Tara Fehlmann⁴, Parakkal
5 Deepak^{5#}, Ryan Bogdan³, Gautam Dantas^{1,2,6,7,8 #}, Umang Jain^{1#*} and SPARC IBD
6 Investigators⁹

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

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

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

13 ⁴ Crohn's and Colitis Foundation, New York, New York, USA

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

16 ⁶Department of Molecular Microbiology, Washington University School of Medicine, St. Louis,
17 MO, USA

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

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

20 ⁹See below for SPARC-IBD Investigator Affiliations

21 #-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

22

23

24

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

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

33

34 **Acknowledgements:** The authors would like to thank Sarah E. Paul for her initial analytical
35 support and Kevin S. Blake of the LGM Scientific Editing Service at the Department of
36 Pathology and Immunology, Washington University School of Medicine for scientific editing
37 support. The authors also thank the staff at The Edison Family Center for Genome Sciences &
38 Systems Biology at the Washington University School of Medicine in St Louis, including E.
39 Martin and B. Koebbe for computational support, and B. Dee, K. Matheny, J. Theodore and K.
40 Page for administrative support. Finally, we would like to thank the members of the Dantas,
41 Bogdan, and Jain labs for helpful general discussions and comments on the manuscript. The
42 results published here are in whole based on data from the Study of a Prospective Adult
43 Research Cohort with IBD (SPARC IBD). SPARC IBD is a component of the Crohn's & Colitis
44 Foundation's IBD Plexus data exchange platform. SPARC IBD enrolls patients with an
45 established or new diagnosis of IBD from sites throughout the United States and links data
46 collected from the electronic health record and study specific case report forms. Patients also
47 provide blood, stool and biopsy samples at selected times during follow-up. The design and
48 implementation of the SPARC IBD cohort has been previously described¹. The SPARC IBD data

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

51

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

58

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

61

62 **Author contributions:** M.G.G., P.D., R.B., G.D. and U.J designed the study, T.F., P.D., and U.J.
63 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
64 wrote the manuscript. All authors approved the manuscript.

65

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

68

69

70

71

72

73 **Abstract:**

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

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

86 **Results:** Thirteen proteins which were significantly differentially abundant in CD vs UC (all $|\beta|s$
87 > 0.22 , all adjusted p values $< 8.42E-06$). Random forest models of proteins differentiated
88 between CD and UC with models trained only on PWAS identified proteins (Average ROC-AUC
89 0.73) outperforming models trained of the full proteome (Average ROC-AUC 0.62). SHAP
90 analysis revealed that Granzyme B, insulin-like peptide 5 (INSL5), and interleukin-12 subunit
91 beta (IL-12B) were the most important features.

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

96

97 **Keywords:** Machine Learning, PWAS, IBD

98

99

100 INTRODUCTION

101 Inflammatory bowel diseases (IBD) are chronic relapsing and remitting inflammatory
102 disorders of the gastrointestinal tract. They affect more than 6 million people worldwide, and in
103 the United States alone more than 70,000 cases of IBD are diagnosed each year^{2,3}. Patients
104 with IBD experience markedly decreased quality of life, high disease- and treatment-related
105 morbidity, and often endure complications requiring hospitalizations and surgeries⁴⁻⁸. IBD is
106 generally subtyped as either Crohn's disease (CD) or ulcerative colitis (UC), with each differing
107 in the areas of manifestation and the resulting sequela⁹⁻¹¹. Specifically, CD can affect any region
108 of the gastrointestinal (GI) tract and generally presents with transmural inflammation, while UC
109 is restricted to the colon and is characterized by mucosal ulceration⁹⁻¹¹. While CD and UC
110 present distinct clinical complications, CD's ability to affect any region of the GI tract, including
111 regions affected by UC, makes discriminating between them challenging¹²⁻¹⁴.

112 As each disease requires distinct therapeutic strategies, being able to accurately and
113 efficiently differentiate CD from UC has significant consequences for clinical care. For example,
114 surgery is not a definitive cure for CD and can result in further complications¹⁵⁻¹⁹. Current
115 practices rely on endoscopy to discriminate CD from UC; however, endoscopy is invasive,
116 expensive, and carries significant risk to the patient²⁰. To complement endoscopic procedures,
117 blood and fecal markers are often used; however, none of these tests have proven sufficient to
118 enable the differentiation of CD and UC²¹⁻²⁷. For instance, serum antibodies against
119 *Saccharomyces cerevisiae* (ASCA) and bacterial antigens have limited accuracy and suffer from
120 low sensitivity, rendering these tests relatively nonspecific to subtype IBD²¹⁻²⁴. Other markers
121 such as fecal calprotectin and Lipocalin-2 can identify inflammatory status but do not enable
122 differentiation between CD and UC²⁵⁻²⁷. Given the rising prevalence of IBD worldwide, its high
123 morbidity, and its substantial negative impact on quality of life, there is an urgent need for

124 diagnostic tools that enable early differentiation of CD from UC, and are easier to use, non-
125 invasive, and less costly than those currently available²⁸.

126 Advanced proteomics technologies offer novel avenues for comprehending
127 pathophysiological mechanisms and pinpointing potential clinical biomarkers in complex
128 diseases. Recent breakthroughs, exemplified by the Olink platform, have revealed novel protein
129 biomarkers for multiple diseases in blood and plasma^{29,30,31}, ensuring heightened sensitivity,
130 precision, and specificity, while also requiring minimal sample volumes. The output data of the
131 Olink platform can be then applied as features for machine learning (ML)-based classification
132 analysis³². Unfortunately, even with these technological advancements, omics data is still often
133 affected by the “curse of dimensionality”³³, where the number of features captured far exceeds
134 the number of samples which can result in models fitting to spurious patterns³³. ML models
135 trained on high dimensionality data may fail to generalize to real world data unless the sample
136 size is sufficiently large enough (normally at least 5 samples per feature³⁴) to separate signal
137 versus noise³³. However, generating such large omics datasets can be both costly and time
138 consuming. To mitigate the “curse of dimensionality” without the costly and time-consuming
139 process of generating large omics datasets, feature selection methods are often used to identify
140 informative features in high dimensionality datasets before model training^{33,35}. In particular,
141 GWAS (Genome-Wide Association Study) and PheWAS (Phenome-Wide Association Study)
142 approaches have proven to be extremely effective for feature selection³⁵. Here we leverage a
143 PWAS (Protein-Wide Association Study)-based approach to identify informative features in a
144 high dimensionality Olink proteomics dataset from IBD patients. This approach identified 13
145 proteins which distinguish CD from UC using plasma samples.

146

147

148

149 **MATERIALS & METHODS**

150 **Participants and Sample Collection**

151 The Study of Prospective Adult Research Cohort of IBD (SPARC IBD) is an ongoing
152 longitudinal cohort study of patients with IBD recruited from 17 academic medical centers
153 across the United States¹. Plasma samples used in this study were obtained from N = 1106
154 individuals ($n_{CD} = 636$; $n_{UC} = 470$).

155 Demographic, disease-related, and patient-reported data were collected during the
156 following visits: 1) during routine GI office visits (2016-2021), 2) quarterly by sending surveys to
157 patients, and 3) before a scheduled colonoscopy. All collections generated highly structured
158 electronic case report forms (eCRF). Bio-samples of each respective patient's blood and stool
159 were collected at enrolment and at the time of each patient's colonoscopy. Further, blood
160 samples were collected if a patient or provider reported key medication changes. Initially
161 collected samples were used in this study. Clinical data is transferred from sites on a periodic
162 basis and stored in IBD Plexus, Crohn's & Colitis Foundation's exchange platform (see¹ for
163 details).

164 **Olink Proteomics, normalization, and filtering**

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

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

179 **Statistical Analyses**

180 Protein Wide Association Study (PWAS)

181 A Protein Wide Association Study (PWAS) was performed on all proteins passing quality control
182 described above (n=2,920) using the glmer function in the lme4 package³⁶ as previously
183 described in phenome-wide association studies^{37,38}. . CD/UC disease status was regressed on
184 each individual protein in a mixed effects logistic regression with age and sex as fixed effects
185 covariates and disease activity (Simple Crohn's Disease Activity Index for CD and 6-point Mayo
186 Score for UC¹) was treated as a random effect. To adjust for multiple testing, a Bonferroni-
187 corrected proteome wide significance threshold was used ($0.05/2,920 = 0.0000171$ alpha
188 level).

189 Principle Component Analysis (PCA)

190 Principal component analysis (PCA) was initially performed incorporating the measured
191 values of all proteins and just proteins identified via the PWAS analysis. Analysis of Similarities
192 (ANOSIM) was performed using the ANOSIM function in the vegan package³⁹.

193 Machine Learning Methods

194 Using Scikit learn based implementations of random forests we tested the following
195 feature sets: All proteomics features and patient features (Age, Sex, Disease Severity),
196 proteomics features which passed the Bonferroni cutoff and patient features, and just
197 proteomics features. Of the samples, 20% were reserved for a holdout validation dataset which
198 was also used for SHAP value analysis⁴⁰. The remaining 80% of the data was split into a
199 train/test split (70/30) and cross validated 30 times.

200 The following packages and versions were used for analysis in R v4.3.2: ibdplexus (0.1.0),
201 tidyverse (1.3.1), stringr (1.5.1), readxl (1.4.3), OlinkAnalyze (3.7.0), data.table (1.15.0),
202 lmerTest (3.1), lme4 (1.1)³⁶, readxl (1.4.3), dplyr (1.1.4), ggplot2 (3.4.3), reshape2 (1.4.4),
203 ggrepel (0.9.5), forcats (1.0.0), ggsci (3.0.0), RColorBrewer (1.1.3), optimx (2023-10.21), minqa
204 (1.2.6), dfoptim (2023.1.0), survey (4.2.1), scales (1.3.0), ggnewscale (0.4.10), ggpubr (0.6.0),
205 gplots (3.1.31), psych (2.4.1), MuMIn (1.47.5), vegan (2-6-6.1), NatParksPalettes (0.2.0), and
206 ggfortify (0.4.16).

207 The following packages and versions were used for analysis in Python v3.10.9: pandas⁴¹(1.5.3),
208 sklearn^{42,43} (1.3.2), numpy⁴⁴ (1.23.5), and shap⁴⁰(0.43.0).

209

210 Formulas

211 TP: True Positive

212 TN: False Negative

213 FP: False Positive

214 FN: False Negative

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

216 Sensitivity= $TP/(TP+FN)$

217 Specificity= $TN/(TN+FP)$

218

219

220

221

222

223

224

225

226 RESULTS

227 Subject Characteristics

228 Details of the subjects' characteristics are shown in Table 1. A total of 1,106 individuals (CD n =
229 636; UC n=470) from 17 medical centers were included. CD patients were on average, 42.26
230 years old and 62% of the CD patients were female. UC patients were on average 43.98 years
231 old and 50.95% of the total UC patients were female.

232 Specific proteins differentiate the proteomic profiles of Crohn's disease and Ulcerative 233 colitis

234 The studied plasma proteome dataset includes measures of 2,920 protein levels across
235 1,106 patients from the Crohn's and Colitis Foundation dataset¹ (**Table 1, Fig 1**). We initially ran
236 a principal component analysis (PCA) (**Fig 2A**) and an analysis of similarities (ANOSIM) which
237 revealed that the global proteomes of CD and UC do not differ ($p=0.21$, $R=0.004$). We then
238 conducted a Protein Wide Association Study (PWAS) analysis adapted from previous study^{38,45}
239 to filter for proteins measured at significantly different levels between CD and UC. Age and sex
240 were included as fixed effects with disease severity as random effect. The PWAS-based
241 approach identified thirteen proteins that were significant after Bonferroni correction (**Table 2,**
242 **Fig 2B, Table S1**). Five of these proteins, INSL5, IL12B, IL12AB, HRG, and LY96 were more
243 abundant in CD relative to UC; in contrast, eight proteins, FGF19, EPCAM, NOS2, GPA33,
244 GUC2A, GRAB, FGFR4, MMP10 were more abundant in UC relative to CD (**Fig S1**).
245 Performing an ANOSIM test and PCA analysis (**Fig 2C**) on the proteins identified via PWAS
246 after multiple test correction revealed that there was significant a difference between the CD
247 and UC cohorts ($p=0.001$, $R=0.1247$).

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

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

252 features (age, sex, disease severity). We generated three features sets which contained the
253 following: 1) the entire proteome and patient features (“Full Feature Set”), 2) subset of
254 proteomic features which only included the thirteen significant proteins (e.g., INSL5) identified
255 by the PWAS and patient features (referred to as “PWAS and Patient Features”), and 3) just the
256 thirteen significant proteins identified by the PWAS without patient features (“PWAS Features”).
257 Models trained on the “Full Feature Set” had significantly higher specificity, but significantly
258 lower accuracy, sensitivity, and ROC-AUC scores compared to the other two feature sets (**Fig**
259 **3A-D, Table 3**). Models trained on “PWAS and Patient Features” and just “PWAS Features”
260 were significantly more accurate and sensitive than models trained on the “Full Feature Set,”
261 they did not differ significantly in performance from each other (**Fig 3A-D**). Next, we interrogated
262 the ML models using SHapley Additive exPlanations (SHAP) analysis, which can infer feature
263 importance, to determine if patient features were important for model performance. Interestingly,
264 SHAP feature importance analysis suggested patient features were not informative, where Age,
265 Disease Severity, and Sex were the ranked as the three least important features in the “PWAS
266 and Patient Features” model suggesting that the patient features contributed the least to model
267 performance (**Fig 2B, 4AB, S1**). Notably, interrogating the ML models revealed that the three
268 most important features including granzyme B, Insulin-like peptide 5 (INSL5), and Interleukin
269 12B (IL12B) were conserved between models. (**Fig 4AB**). This demonstrates that PWAS-based
270 approaches act as a filter for identifying more informative features which in turn improve
271 prediction performance.

272

273

274

275

276

277

278 **DISCUSSION**

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

290 Machine learning algorithms are increasingly being utilized to analyze medical data to
291 diagnose diseases, predict their severity, and monitor their progression. Recent work on
292 diagnosing IBD using ML approaches has also been successful, achieving high levels of
293 performance^{12,48–50}. For instance, supervised learning models on RNA sequencing data enabled
294 CD and UC differentiation¹². Similarly, deep learning networks have been used on endoscopic
295 images to accurately predict the severity of the disease in IBD^{48,49}. Although proteomic datasets
296 have been generated in IBD, the application of ML techniques to analyze such datasets has
297 been limited⁴⁷. Furthermore, previous IBD-focused proteomics datasets have measured smaller
298 panels of proteins^{47,51–53}. We used a combined PWAS-based feature selection and ML models
299 on a large dataset of proteins to identify novel signatures that could accurately subtype IBD.
300 Importantly, in contrast to previous studies that have primarily focused on inflammatory markers,
301 we analyzed proteins that are involved in a diverse array of processes including, hormonal

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

304 Several proteins in our cohort have been validated by other IBD studies focused on
305 differentiating CD from UC. A study by Bourgonje et al. also employed a proximity extension
306 assay (Olink) and measured 92 proteins identifying FGF19, IL12B, and MMP10 to be
307 differentially abundant between CD and UC⁵³. In another study by Di Narzo et al., the authors
308 used a SOMAmer-based capture array to measure protein levels in plasma (n=244) and
309 discovered that Granzyme B, FGF19, and MMP10 were downregulated in CD relative to UC,
310 mirroring our results⁵². Importantly, our findings combined with others have identified FGF19 and
311 MMP10 as consistent plasma-based biomarkers which can be used differentiate CD from
312 UC^{52,53}.

313 Among the 13 differentially abundant proteins significant in the PWAS after multiple
314 correction, Granzyme B, IL12B, and INSL5 were the most informative for model prediction (**Fig**
315 **4A-B**). INSL5, to date, has not been measured in similar proteomics studies focused on IBD^{52,53}.
316 Notably, depletion of *INSL5* transcripts in mucosal tissue has been associated with IBD⁵⁴, and
317 our study further implicates the INSL5 peptide as differentially abundant between CD and UC.
318 INSL5 is a peptide hormone that is expressed in the colonic epithelium⁵⁴⁻⁵⁶. Because INSL5 is
319 a microbially regulated molecule, it is possible that UC, but not CD, specific microbes alter its
320 production⁵⁷. Indeed, both bacterial and fungal microbiota are known to be different between UC
321 and CD⁵⁸⁻⁶⁰. Another possible reason for the decreased abundance of INSL5 in UC relative to
322 CD is the loss of colonic epithelial cells due to ulceration, a prominent feature of UC⁶¹. Future
323 studies are needed to elucidate how INSL5 is regulated and the mechanisms by which INSL5
324 modulates the severity of the disease^{54,62}. In addition to INSL5, Granzyme B was also a
325 powerful predictive feature and was elevated in UC. Granzyme B is a serine protease released
326 by lymphocytes which can trigger apoptosis^{63,64}. Similar to our findings, Di Narzo et al. identified

327 elevated Granzyme B protein levels in CD compared to UC⁵². Further, levels of Granzyme B
328 have been reported to predict treatment responses in IBD as its levels are significantly lower in
329 responders compared to non-responder populations⁶⁵. While these initial findings are promising,
330 it is unclear how the levels of these targets fluctuate throughout disease specific treatments and
331 subtype. Future studies utilizing longitudinal samples are needed to ascertain its association
332 with IBD subtypes.

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

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

350

351

352 **References:**

- 353 1. Raffals LE., Saha S., Bewtra M., Norris C., Dobes A., Heller C., et al. The Development and
354 Initial Findings of A Study of a Prospective Adult Research Cohort with Inflammatory Bowel
355 Disease (SPARC IBD). *Inflammatory Bowel Diseases* 2022;**28**(2):192–9. Doi:
356 10.1093/ibd/izab071.
- 357 2. Kaplan GG. The global burden of IBD: from 2015 to 2025. *Nat Rev Gastroenterol Hepatol*
358 2015;**12**(12):720–7. Doi: 10.1038/nrgastro.2015.150.
- 359 3. Duryee MJ., Ahmad R., Eichele DD., Hunter CD., Mitra A., Talmon GA., et al. Identification
360 of Immunoglobulin G Autoantibody Against Malondialdehyde-Acetaldehyde Adducts as a
361 Novel Serological Biomarker for Ulcerative Colitis. *Clin Transl Gastroenterol*
362 2022;**13**(4):e00469. Doi: 10.14309/ctg.0000000000000469.
- 363 4. Soriano CR., Powell CR., Chiorean MV., Simianu VV. Role of hospitalization for
364 inflammatory bowel disease in the post-biologic era. *World J Clin Cases* 2021;**9**(26):7632–
365 42. Doi: 10.12998/wjcc.v9.i26.7632.
- 366 5. The global, regional, and national burden of inflammatory bowel disease in 195 countries
367 and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study
368 2017. *Lancet Gastroenterol Hepatol* 2019;**5**(1):17–30. Doi: 10.1016/S2468-1253(19)30333-
369 4.
- 370 6. Xavier RJ., Podolsky DK. Unravelling the pathogenesis of inflammatory bowel disease.
371 *Nature* 2007;**448**(7152):427–34. Doi: 10.1038/nature06005.
- 372 7. Mitropoulou M-A., Fradelos EC., Lee KY., Malli F., Tsaras K., Christodoulou NG., et al.
373 Quality of Life in Patients With Inflammatory Bowel Disease: Importance of Psychological
374 Symptoms. *Cureus* n.d.;**14**(8):e28502. Doi: 10.7759/cureus.28502.
- 375 8. Bernstein CN., Nabalamba A. Hospitalization, Surgery, and Readmission Rates of IBD in
376 Canada: A Population-Based Study. *Official Journal of the American College of*
377 *Gastroenterology | ACG* 2006;**101**(1):110.
- 378 9. Panaccione R. Mechanisms of Inflammatory Bowel Disease. *Gastroenterol Hepatol (N Y)*
379 2013;**9**(8):529–32.
- 380 10. Graham DB., Xavier RJ. Pathway Paradigms Revealed from the Genetics of Inflammatory
381 Bowel Disease. *Nature* 2020;**578**(7796):527–39. Doi: 10.1038/s41586-020-2025-2.
- 382 11. Guan Q. A Comprehensive Review and Update on the Pathogenesis of Inflammatory Bowel
383 Disease. *J Immunol Res* 2019;**2019**:7247238. Doi: 10.1155/2019/7247238.
- 384 12. Park S-K., Kim S., Lee G-Y., Kim S-Y., Kim W., Lee C-W., et al. Development of a Machine
385 Learning Model to Distinguish between Ulcerative Colitis and Crohn’s Disease Using RNA
386 Sequencing Data. *Diagnostics (Basel)* 2021;**11**(12):2365. Doi:
387 10.3390/diagnostics11122365.
- 388 13. von Stein P., Lofberg R., Kuznetsov NV., Gielen AW., Persson J., Sundberg R., et al.
389 Multigene Analysis Can Discriminate Between Ulcerative Colitis, Crohn’s Disease, and
390 Irritable Bowel Syndrome. *Gastroenterology* 2008;**134**(7):1869–81. Doi:
391 10.1053/j.gastro.2008.02.083.
- 392 14. Tontini GE., Vecchi M., Pastorelli L., Neurath MF., Neumann H. Differential diagnosis in
393 inflammatory bowel disease colitis: State of the art and future perspectives. *World J*
394 *Gastroenterol* 2015;**21**(1):21–46. Doi: 10.3748/wjg.v21.i1.21.
- 395 15. Kanazawa A., Yamana T., Okamoto K., Sahara R. Risk Factors for Postoperative Intra-
396 abdominal Septic Complications after Bowel Resection in Patients with Crohn’s Disease.
397 *Diseases of the Colon & Rectum* 2012;**55**(9):957. Doi: 10.1097/DCR.0b013e3182617716.
- 398 16. Lewis RT., Maron DJ. Efficacy and Complications of Surgery for Crohn’s Disease.
399 *Gastroenterol Hepatol (N Y)* 2010;**6**(9):587–96.

- 400 17. Guo K., Ren J., Li G., Hu Q., Wu X., Wang Z., et al. Risk factors of surgical site infections in
401 patients with Crohn's disease complicated with gastrointestinal fistula. *Int J Colorectal Dis*
402 2017;**32**(5):635–43. Doi: 10.1007/s00384-017-2751-6.
- 403 18. Post S., Betzler M., von Ditfurth B., Schürmann G., Küppers P., Herfarth C. Risks of
404 intestinal anastomoses in Crohn's disease. *Ann Surg* 1991;**213**(1):37–42.
- 405 19. Singh S., Nguyen GC. Management of Crohn's Disease After Surgical Resection.
406 *Gastroenterology Clinics of North America* 2017;**46**(3):563–75. Doi:
407 10.1016/j.gtc.2017.05.011.
- 408 20. Kavic SM., Basson MD. Complications of endoscopy. *The American Journal of Surgery*
409 2001;**181**(4):319–32. Doi: 10.1016/S0002-9610(01)00589-X.
- 410 21. Prideaux L., De Cruz P., Ng SC., Kamm MA. Serological Antibodies in Inflammatory Bowel
411 Disease: A Systematic Review. *Inflammatory Bowel Diseases* 2012;**18**(7):1340–55. Doi:
412 10.1002/ibd.21903.
- 413 22. WALKER LJ., ALDHOUS MC., DRUMMOND HE., SMITH BRK., NIMMO ER., ARNOTT
414 IDR., et al. Anti-Saccharomyces cerevisiae antibodies (ASCA) in Crohn's disease are
415 associated with disease severity but not NOD2/CARD15 mutations. *Clin Exp Immunol*
416 2004;**135**(3):490–6. Doi: 10.1111/j.1365-2249.2003.02392.x.
- 417 23. Zhou G., Song Y., Yang W., Guo Y., Fang L., Chen Y., et al. ASCA, ANCA, ALCA and Many
418 More: Are They Useful in the Diagnosis of Inflammatory Bowel Disease? *Digestive Diseases*
419 2016;**34**(1–2):90–7. Doi: 10.1159/000442934.
- 420 24. Reese GE., Constantinides VA., Simillis C., Darzi AW., Orchard TR., Fazio VW., et al.
421 Diagnostic Precision of Anti-Saccharomyces cerevisiae Antibodies and Perinuclear
422 Antineutrophil Cytoplasmic Antibodies in Inflammatory Bowel Disease. *Official Journal of the*
423 *American College of Gastroenterology | ACG* 2006;**101**(10):2410.
- 424 25. Oikonomou KA., Kapsoritakis AN., Theodoridou C., Karangelis D., Germentis A., Stefanidis
425 I., et al. Neutrophil gelatinase-associated lipocalin (NGAL) in inflammatory bowel disease:
426 association with pathophysiology of inflammation, established markers, and disease activity.
427 *J Gastroenterol* 2012;**47**(5):519–30. Doi: 10.1007/s00535-011-0516-5.
- 428 26. Barnes EL., Burakoff R. New Biomarkers for Diagnosing Inflammatory Bowel Disease and
429 Assessing Treatment Outcomes. *Inflamm Bowel Dis* 2016;**22**(12):2956–65. Doi:
430 10.1097/MIB.0000000000000903.
- 431 27. Jukic A., Bakiri L., Wagner EF., Tilg H., Adolph TE. Calprotectin: from biomarker to biological
432 function. *Gut* 2021;**70**(10):1978–88. Doi: 10.1136/gutjnl-2021-324855.
- 433 28. Barnes EL., Liew C-C., Chao S., Burakoff R. Use of blood based biomarkers in the
434 evaluation of Crohn's disease and ulcerative colitis. *World J Gastrointest Endosc*
435 2015;**7**(17):1233–7. Doi: 10.4253/wjge.v7.i17.1233.
- 436 29. Wang T., Yang S., Long Y., Li Y., Wang T., Hou Z. Olink proteomics analysis uncovers the
437 landscape of inflammation-related proteins in patients with acute compartment syndrome.
438 *Front Immunol* 2023;**14**:1293826. Doi: 10.3389/fimmu.2023.1293826.
- 439 30. Kong T., Qu Y., Zhao T., Niu Z., Lv X., Wang Y., et al. Identification of novel protein
440 biomarkers from the blood and urine for the early diagnosis of bladder cancer via proximity
441 extension analysis. *J Transl Med* 2024;**22**:314. Doi: 10.1186/s12967-024-04951-z.
- 442 31. Gong Q., Fu M., Wang J., Zhao S., Wang H. Potential Immune-Inflammatory Proteome
443 Biomarkers for Guiding the Treatment of Patients with Primary Acute Angle-Closure
444 Glaucoma Caused by COVID-19. *J Proteome Res* 2024. Doi:
445 10.1021/acs.jproteome.4c00325.
- 446 32. Diaz-Canestro C., Chen J., Liu Y., Han H., Wang Y., Honoré E., et al. A machine-learning
447 algorithm integrating baseline serum proteomic signatures predicts exercise responsiveness
448 in overweight males with prediabetes. *Cell Rep Med* 2023;**4**(2):100944. Doi:
449 10.1016/j.xcrm.2023.100944.

- 450 33. Berisha V., Krantsevich C., Hahn PR., Hahn S., Dasarathy G., Turaga P., et al. Digital
451 medicine and the curse of dimensionality. *NPJ Digit Med* 2021;**4**:153. Doi: 10.1038/s41746-
452 021-00521-5.
- 453 34. *Pattern Recognition*. 2008.
- 454 35. Pudjihartono N., Fadason T., Kempa-Liehr AW., O'Sullivan JM. A Review of Feature
455 Selection Methods for Machine Learning-Based Disease Risk Prediction. *Front Bioinform*
456 2022;**2**:927312. Doi: 10.3389/fbinf.2022.927312.
- 457 36. Bates D., Mächler M., Bolker B., Walker S. Fitting Linear Mixed-Effects Models Using lme4.
458 *Journal of Statistical Software* 2015;**67**:1–48. Doi: 10.18637/jss.v067.i01.
- 459 37. Bastarache L., Denny JC., Roden DM. Phenome-Wide Association Studies. *JAMA*
460 2022;**327**(1):75–6. Doi: 10.1001/jama.2021.20356.
- 461 38. Gorelik AJ., Paul SE., Karcher NR., Johnson EC., Nagella I., Blaydon L., et al. A Phenome-
462 Wide Association Study (PheWAS) of Late Onset Alzheimer Disease Genetic Risk in
463 Children of European Ancestry at Middle Childhood: Results from the ABCD Study. *Behav*
464 *Genet* 2023;**53**(3):249–64. Doi: 10.1007/s10519-023-10140-3.
- 465 30. Oksanen J., Simpson GL., Kindt R., Legendre P., Minchin PR., O'Hara RB., et al. Vegan
466 community ecology package. [Http://Vegan.r-Forge.r-ProjectOrg/](http://Vegan.r-Forge.r-ProjectOrg/) 2010.
- 467 40. Lundberg SM., Lee S-I. A Unified Approach to Interpreting Model Predictions. *Advances in*
468 *Neural Information Processing Systems*, vol. 30. Curran Associates, Inc.; 2017.
- 469 41. The pandas development team. pandas-dev/pandas: Pandas 2024. Doi:
470 10.5281/zenodo.10957263.
- 471 42. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., et al. Scikit-learn:
472 Machine Learning in Python. *Journal of Machine Learning Research* 2011;**12**(85):2825–30.
- 473 43. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., et al. Scikit-learn:
474 Machine Learning in Python 2018. Doi: 10.48550/arXiv.1201.0490.
- 475 44. Harris CR., Millman KJ., van der Walt SJ., Gommers R., Virtanen P., Cournapeau D., et al.
476 Array programming with NumPy. *Nature* 2020;**585**(7825):357–62. Doi: 10.1038/s41586-020-
477 2649-2.
- 478 45. Brandes N., Linal N., Linal M. PWAS: proteome-wide association study-linking genes and
479 phenotypes by functional variation in proteins. *Genome Biol* 2020;**21**(1):173. Doi:
480 10.1186/s13059-020-02089-x.
- 481 46. Zhao JH., Stacey D., Eriksson N., Macdonald-Dunlop E., Hedman ÅK., Kalnapekis A., et
482 al. Genetics of circulating inflammatory proteins identifies drivers of immune-mediated
483 disease risk and therapeutic targets. *Nat Immunol* 2023;**24**(9):1540–51. Doi:
484 10.1038/s41590-023-01588-w.
- 485 47. Kalla R., Adams AT., Bergemalm D., Vatn S., Kennedy NA., Ricanek P., et al. Serum
486 proteomic profiling at diagnosis predicts clinical course, and need for intensification of
487 treatment in inflammatory bowel disease. *Journal of Crohn's and Colitis* 2021;**15**(5):699–
488 708. Doi: 10.1093/ecco-jcc/jjaa230.
- 489 48. Chierici M., Puica N., Pozzi M., Capistrano A., Donzella MD., Colangelo A., et al.
490 Automatically detecting Crohn's disease and Ulcerative Colitis from endoscopic imaging.
491 *BMC Medical Informatics and Decision Making* 2022;**22**(6):300. Doi: 10.1186/s12911-022-
492 02043-w.
- 493 49. Takenaka K., Ohtsuka K., Fujii T., Negi M., Suzuki K., Shimizu H., et al. Development and
494 Validation of a Deep Neural Network for Accurate Evaluation of Endoscopic Images From
495 Patients With Ulcerative Colitis. *Gastroenterology* 2020;**158**(8):2150–7. Doi:
496 10.1053/j.gastro.2020.02.012.
- 497 50. Seeley EH., Washington MK., Caprioli RM., M'Koma AE. Proteomic Patterns of Colonic
498 Mucosal Tissues Delineate Crohn's Colitis and Ulcerative Colitis. *Proteomics Clin Appl*
499 2013;**7**(0):10.1002/prca.201200107. Doi: 10.1002/prca.201200107.

- 500 51. Gisbert JP., Chaparro M. Clinical Usefulness of Proteomics in Inflammatory Bowel Disease:
501 A Comprehensive Review. *J Crohns Colitis* 2019;**13**(3):374–84. Doi: 10.1093/ecco-
502 jcc/jjy158.
- 503 52. Di Narzo AF., Brodmerkel C., Telesco SE., Argmann C., Peters LA., Li K., et al. High-
504 Throughput Identification of the Plasma Proteomic Signature of Inflammatory Bowel
505 Disease. *Journal of Crohn's and Colitis* 2019;**13**(4):462–71. Doi: 10.1093/ecco-jcc/jjy190.
- 506 53. Bourgonje AR., Hu S., Spekhorst LM., Zhernakova DV., Vich Vila A., Li Y., et al. The Effect
507 of Phenotype and Genotype on the Plasma Proteome in Patients with Inflammatory Bowel
508 Disease. *Journal of Crohn's and Colitis* 2022;**16**(3):414–29. Doi: 10.1093/ecco-jcc/jjab157.
- 509 54. Skok DJ., Hauptman N., Jerala M., Zidar N. Expression of Cytokine-Coding Genes BMP8B,
510 LEFTY1 and INSL5 Could Distinguish between Ulcerative Colitis and Crohn's Disease.
511 *Genes (Basel)* 2021;**12**(10):1477. Doi: 10.3390/genes12101477.
- 512 55. Thanasupawat T., Hammje K., Adham I., Ghia J-E., Del Bigio MR., Krcek J., et al. INSL5 is a
513 novel marker for human enteroendocrine cells of the large intestine and neuroendocrine
514 tumours. *Oncology Reports* 2013;**29**(1):149–54. Doi: 10.3892/or.2012.2119.
- 515 56. Liu C., Kuei C., Sutton S., Chen J., Bonaventure P., Wu J., et al. INSL5 is a high affinity
516 specific agonist for GPCR142 (GPR100). *J Biol Chem* 2005;**280**(1):292–300. Doi:
517 10.1074/jbc.M409916200.
- 518 57. Lee YS., De Vadder F., Tremaroli V., Wichmann A., Mithieux G., Bäckhed F. Insulin-like
519 peptide 5 is a microbially regulated peptide that promotes hepatic glucose production. *Mol*
520 *Metab* 2016;**5**(4):263–70. Doi: 10.1016/j.molmet.2016.01.007.
- 521 58. Jain U., Ver Heul AM., Xiong S., Gregory MH., Demers EG., Kern JT., et al. *Debaryomyces*
522 is enriched in Crohn's disease intestinal tissue and impairs healing in mice. *Science*
523 2021;**371**(6534):1154–9. Doi: 10.1126/science.abd0919.
- 524 59. Schirmer M., Franzosa EA., Lloyd-Price J., McIver LJ., Schwager R., Poon TW., et al.
525 Dynamics of metatranscription in the inflammatory bowel disease gut microbiome. *Nat*
526 *Microbiol* 2018;**3**(3):337–46. Doi: 10.1038/s41564-017-0089-z.
- 527 60. Franzosa EA., Sirota-Madi A., Avila-Pacheco J., Fornelos N., Haiser HJ., Reinker S., et al.
528 Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nature*
529 *Microbiology* 2019;**4**(2):293–305. Doi: 10.1038/s41564-018-0306-4.
- 530 61. A K., P G. Ulcerative colitis: understanding its cellular pathology could provide insights into
531 novel therapies. *Journal of Inflammation (London, England)* 2020;**17**. Doi: 10.1186/s12950-
532 020-00246-4.
- 533 62. Pustovit RV., Zhang X., Liew JJ., Praveen P., Liu M., Koo A., et al. A Novel Antagonist
534 Peptide Reveals a Physiological Role of Insulin-Like Peptide 5 in Control of Colorectal
535 Function. *ACS Pharmacol Transl Sci* 2021;**4**(5):1665–74. Doi: 10.1021/acspsci.1c00171.
- 536 63. Trapani JA., Sutton VR. Granzyme B: pro-apoptotic, antiviral and antitumor functions. *Curr*
537 *Opin Immunol* 2003;**15**(5):533–43. Doi: 10.1016/s0952-7915(03)00107-9.
- 538 64. Kim TJ., Koo JS., Kim SJ., Hong SN., Kim YS., Yang S-K., et al. Role of IL-1ra and
539 Granzyme B as biomarkers in active Crohn's disease patients. *Biomarkers* 2018;**23**(2):161–
540 6. Doi: 10.1080/1354750X.2017.1387933.
- 541 65. Heidari P., Haj-Mirzaian A., Prabhu S., Ataeinia B., Esfahani SA., Mahmood U. Granzyme B
542 PET Imaging for Assessment of Disease Activity in Inflammatory Bowel Disease. *Journal of*
543 *Nuclear Medicine* 2024. Doi: 10.2967/jnumed.123.267344.

545

546

547

548

549

550

551

552

553

554

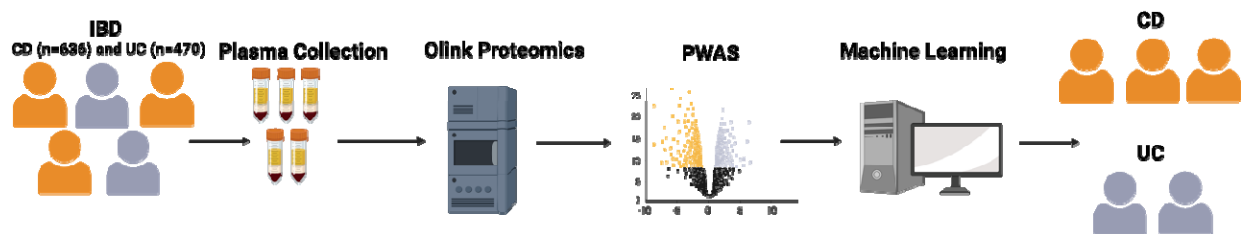
555

556

557

558

559 **Figures**



560

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

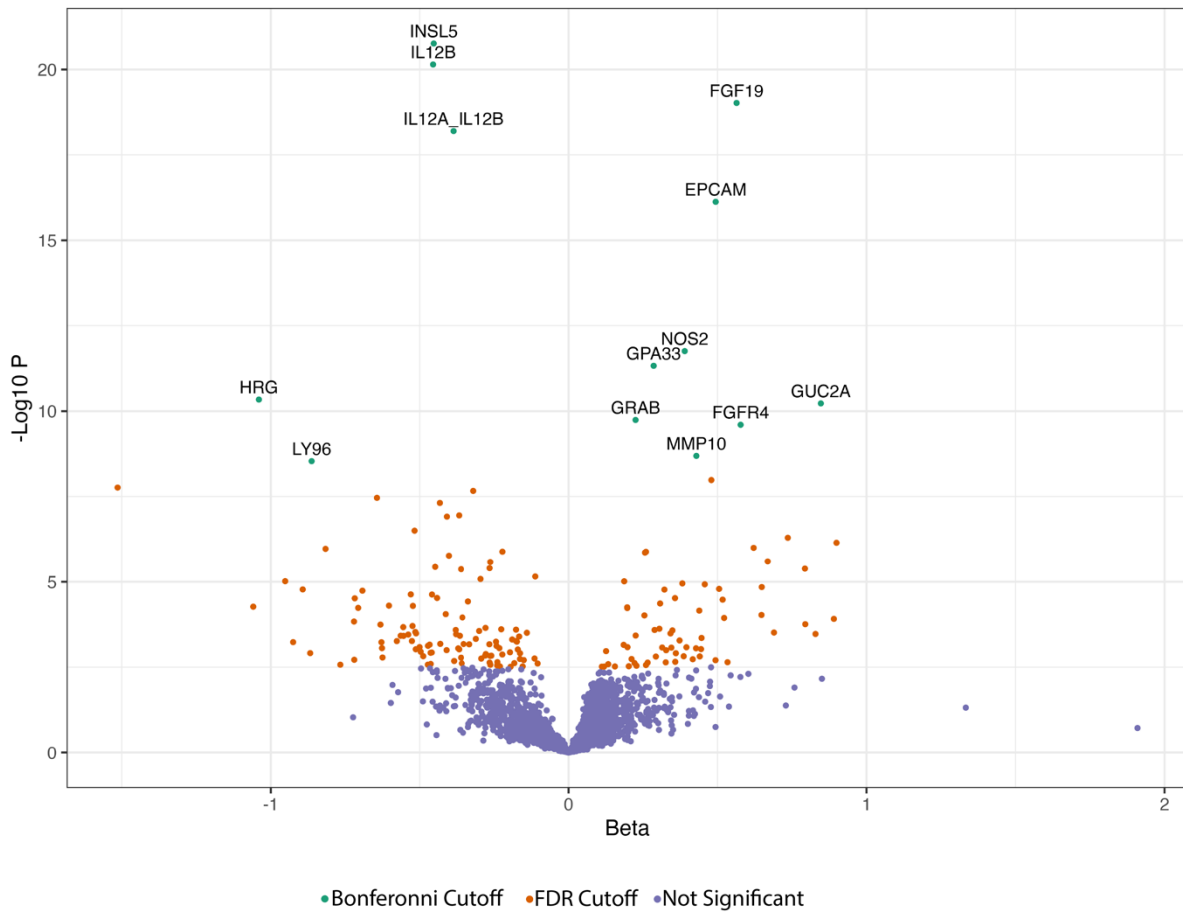
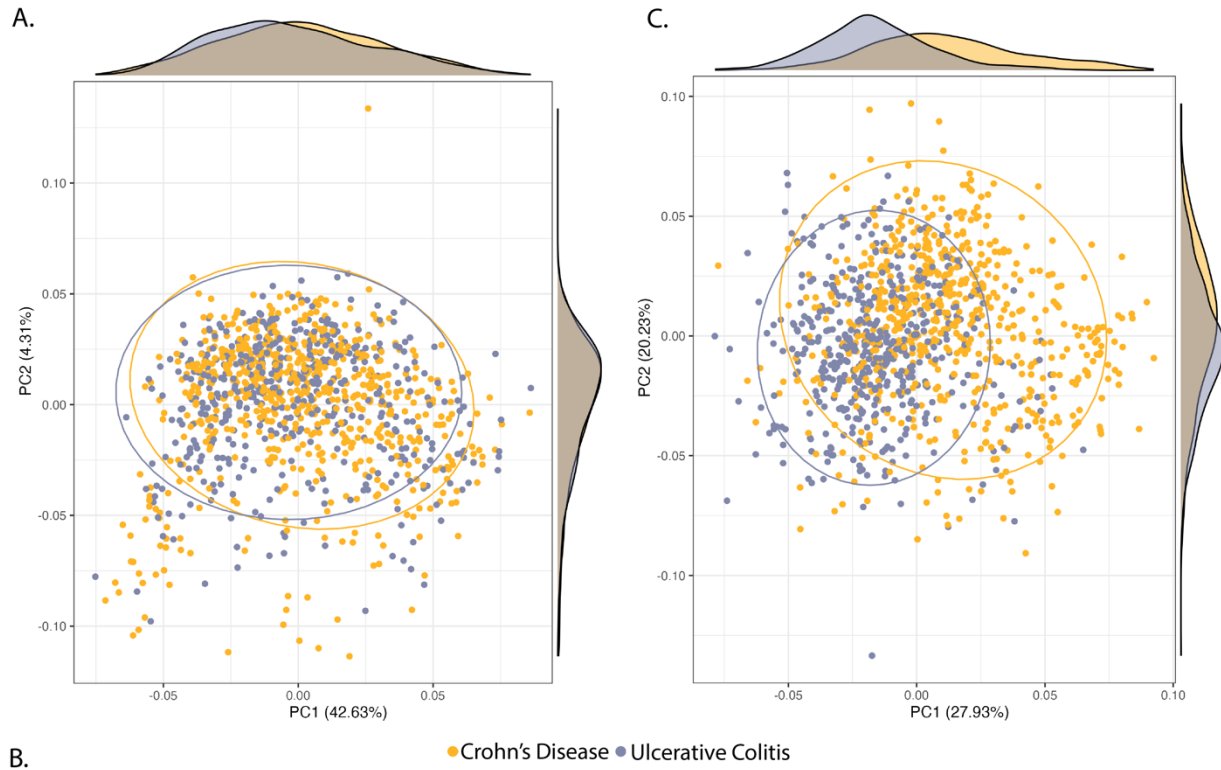
562 and processed as described in the methods and materials. Differentially abundant proteins were

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

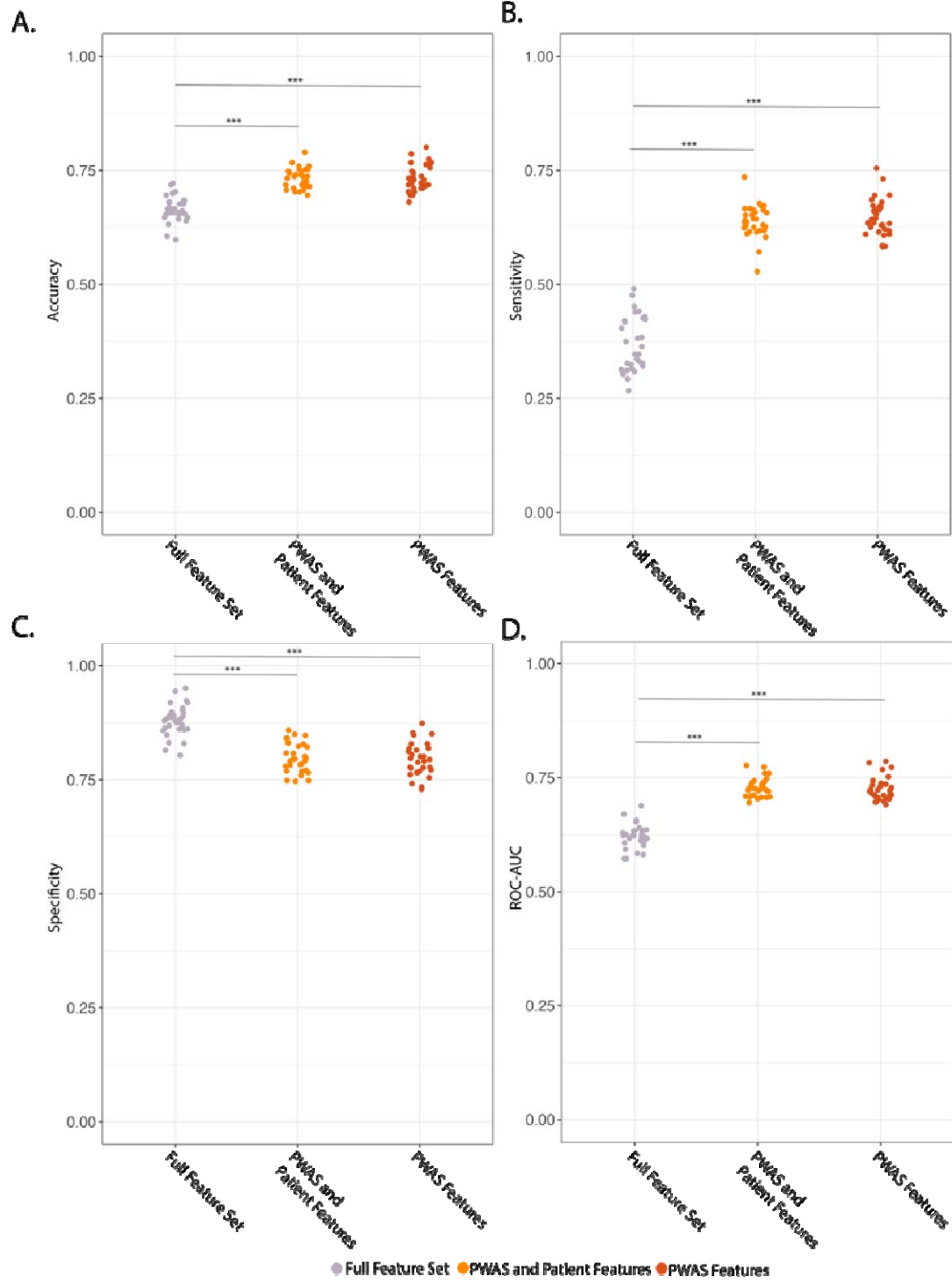
564 learning models to classify CD from UC.

565

566



568 **Figure 2. PWAS analysis enables separation of the proteomic profiles of Ulcerative colitis**
569 **and Crohn's disease.** A) Principal Component Analysis (PCA) of the global proteomics profiles
570 of Crohn's disease and Ulcerative colitis. B) Volcano plot where the x axis is the calculated beta
571 and the y axis is the negative \log_{10} of the unadjusted p-value; green and labeled points had a
572 Bonferroni adjusted p-value of less than 0.0000171 (used in Fig 1B), orange points had an FDR
573 adjusted p value of less than .05, and purple points represent proteins with a p-value > .05.
574 Negative beta values are associated with Crohn's disease and positive beta values are
575 associated with Ulcerative colitis. C) PCA of the proteomics profile identified by the PWAS
576 analysis. Ellipses represent 95% confidence bounds around group centroids.
577
578



579

580

581

582 **Figure 3. Specific proteins improve machine learning based differentiation of CD and UC.**

583 A) Effect of feature set on model accuracy. B) Effect of feature set on model sensitivity. C) Effect
584 of feature set on model specificity. D) Effect of feature set on model ROC-AUC. ***P<.001,
585 ANOVA with Tukey's post hoc test.

586

587

588

589

590

591

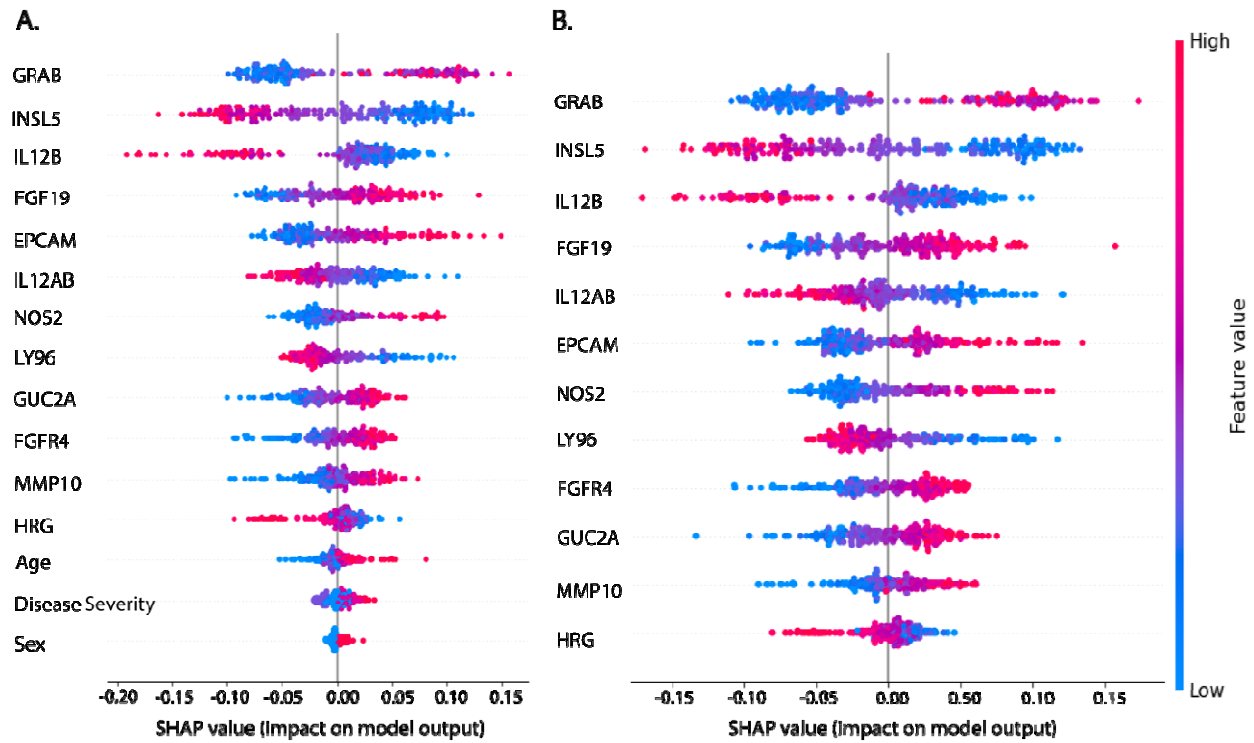
592

593

594

595

596



597

598

599

600 **Figure 4. Clinical Features do not improve model performance.** A) SHAP beeswarm plot of

601 the validation dataset indicating feature importance in random forest models trained on patient

602 associated features (Age, Sex, Disease Severity) and the thirteen proteins which are

603 significantly associated with Crohn's disease and Ulcerative colitis. B) SHAP beeswarm plot of

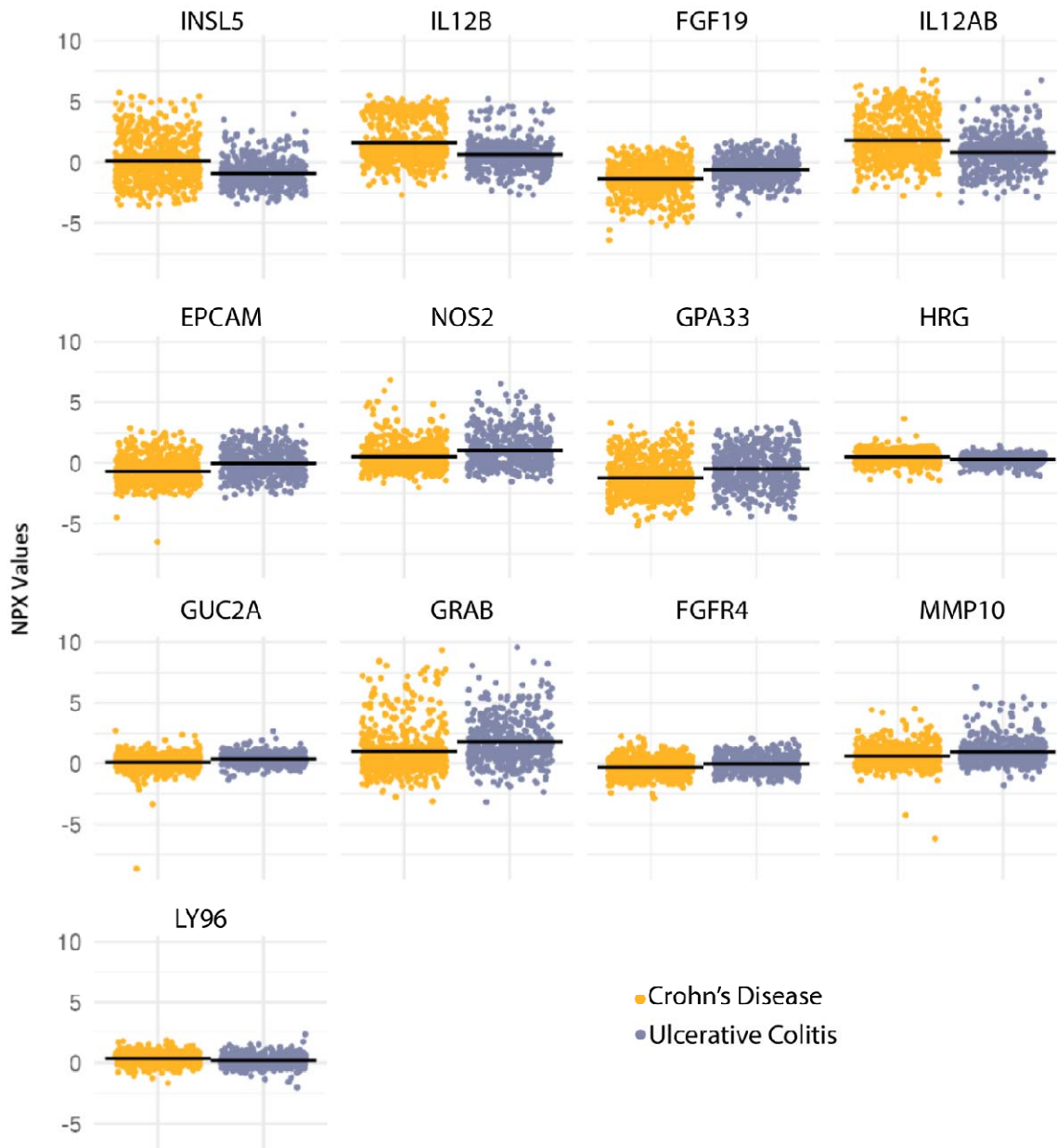
604 the validation dataset indicating feature importance in random forest models trained on just the

605 thirteen proteins which are significantly associated with Crohn's disease and ulcerative colitis.

606 Features are sorted in order of predicted importance in a descending manner.

607

608



609

610

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

612 Bonferroni correction (Fig 1B).

613

614

615

616 **Table 1.** Cohort Breakdown

	<i>Crohn's</i>	<i>Ulcerative colitis</i>	<i>p value</i>
	<i>Disease</i>		
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)	

617

618

619

620

621

622

623

624

625

626

627

628

629

630

631

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

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

633

634

635

636

637

638

639

640

641

642

643

644

645

646

647

648

649

650

651

652

653 **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