AI-Powered Analysis to Identify Potential Patient-Specific Therapy for Crohn's Disease Using Single-Cell RNA Sequencing

Rohan Vellamcheti

Abstract:
Inflammatory bowel disease (IBD), affecting over 3 million Americans, describes chronic inflammatory disorders of the digestive tract, with Crohn's disease as a major subtype affecting the small intestine and colon. Patients often cycle through multiple therapies including corticosteroids, 5-aminosalicylates, immunomodulators, anti-TNF agents, anti-integrin antibodies, and JAK inhibitors before achieving remission. This trial-and-error approach prolongs disease activity, reduces quality of life, and burdens patients and the economy. The current treatment paradigm relies on population-level responses rather than patient-specific biology, therefore precision medicine approaches are needed.This project developed an AI-powered pipeline (automated sequential steps) identifying treatments by combining single-cell RNA sequencing, which measures gene activity in individual cells, with machine learning. Using data from 347,017 cells across multiple Crohn's disease patients, gene set enrichment analysis, which identifies abnormally active cellular processes, was performed across 2,186 pathways from KEGG, Hallmark, and Reactome databases. Singular value decomposition, which mathematically groups related pathways by similarity, compressed them into 20 summary patterns. Random Forest classifiers were trained for both tissues (small intestine and colon) to score treatment compatibility.The pipeline distinguished effective from ineffective treatment matches across 480 pathway-patient pairs. The small intestine model achieved 73.8% accuracy and area under the curve (AUC) of 0.811, while the colon model achieved 75.7% accuracy and 0.767, where 0.5 is random chance. All 480 recommendations received high confidence scores above 0.7.This pipeline offers a data-driven path toward treatments matched to each patient's unique biology, potentially transforming Crohn's disease treatment.

Introduction

Inflammatory Bowel Disease (IBD), with its major diseases as Crohn's Disease (CD) and ulcerative colitis (UC), is a chronic inflammatory disorder of the gastrointestinal tract(1,2). Despite extensive research, the exact cause of IBD remains unknown and is believed to involve a complex interaction between genetics, the immune system, environmental factors, and the gut microbiome. In recent years, the initial occurrence of IBD has increased among children and younger adolescents, a period of time that is critical for growth and development of the human body(3–5). Constant inflammation during this time in life can disrupt nutrient absorption and redirect nutrients toward the immune system to fight the disease rather than growth, potentially resulting in not reaching growth potential or prolonged growth.

With advances in therapies, current treatments for IBD still remain largely based on trial and error rather than on a clear understanding of underlying mechanisms. Available therapies include 5-aminosalicylates (mesalamine, sulfasalazine), corticosteroids (prednisone, budesonide), immunomodulators (azathioprine, 6-mercaptopurine, methotrexate), anti-TNF agents (infliximab, adalimumab, certolizumab), anti-integrin antibodies (vedolizumab, natalizumab), anti-IL-12/23 antibodies (ustekinumab), and JAK inhibitors (tofacitinib,

upadacitinib)(6,7). IBD patients often undergo multiple different therapies and combinations of therapies before reaching remission, as treatments are selected based on population level outcomes rather than patient specific responses creating a large delay for patients to reach remission missing out on potential growth and development(8,9). This delay prolongs disease activity and increases the risk of complications such as fibrosis which is a major complication of CD marked by excess deposition of extracellular matrix, leading to stricturing and functional impairment of the organs affected by fibrosis(10–12).

Single-cell RNA sequencing (scRNA-seq) technology now enables profiling of individual cells within diseased tissue, revealing how different cell types respond uniquely to IBD(13,14). This cellular-level resolution creates opportunities for precision medicine approaches that could match patients to therapies based on their specific molecular profiles rather than population averages.However, translating single-cell transcriptomic signatures into actionable therapeutic recommendations remains a significant challenge, requiring integration of drug target databases with patient-specific cellular profiling and machine learning methods to identify optimal treatment strategies.

Prior Work:

ScRNA-seq in IBD
Multiple studies have used scRNA-seq to characterize cellular heterogeneity in IBD. A study in 2023 in Gastroenterology analyzed 409,001 cells from full-thickness CD resections, revealing previously unrecognized fibroblast (a cell in connective tissue that produces fibers) heterogeneity with CXCL14+ and MMP/WNT5A+ fibroblasts playing central signaling roles in strictures(abnormal narrowings or constriction of a body passage)(15). Importantly, this work validated cadherin-11 (CDH11) as a potential therapeutic target using both in laboratory cell culture experiments and animal models, demonstrating its profibrotic function.

A more recent study done in 2025 combined bulk sequencing with scRNA-seq to comprehensively characterize resected CD-related fibrosis, identifying dysregulated pathways and promising biomarkers including GREM1 (16)(exclusively upregulated in stricture fibroblasts), representing potential therapeutic targets. A longitudinal study examined anti-TNF treatment response in approximately 1 million cells, identifying myeloid and epithelial cell predictors of drug response for CD and UC (Ulcerative Colitis)(17).

Recent advances in AI-powered genomic analysis, such as GenePT, have demonstrated the potential of foundation models built from large language models like ChatGPT for interpreting gene expression patterns and generating embeddings for downstream tasks(18). GenePT creates contextualized gene representations by leveraging the natural language understanding capabilities of transformer architectures, showing promise for tasks like cell type annotation and gene function prediction.

Drug discovery in IBD
Despite advances in understanding IBD at a cellular level, most studies have not directly implemented single cell transcriptomic data with drug target databases to pinpoint therapeutic candidates. Computational methods for IBD drug discovery have focused primarily on target

identification through genomic studies, virtual screening, and molecular docking, with the drug discovery process beginning with target discovery followed by biological confirmation in cellular and animal models(19). In more recent work people have started to emphasize the integration of AI with multi-omics data to uncover molecular targets and personalize trial strategies, with multi-omics analysis helping researchers identify potential targets by understanding pathogenesis and simultaneous changes in the microbiome and biological processes(20,21).

However, no studies have linked systematically scRNA-seq cell population signatures directly to large-scale drug databases such as ChEMBL for therapeutic discovery in IBD. My project addresses this gap matching cellular gene expression patterns from scRNA-seq data to drugs predicted to target relevant inflammatory and fibrotic pathways, enabling a precision based approach to the selection of therapies.

Pipeline Overview

While these studies have advanced our understanding of CD cellular heterogeneity, a critical gap remains: no systematic framework exists to translate single-cell signatures into patient-specific therapeutic recommendations. Existing computational approaches focus on target identification but lack integration with patient-level transcriptomic profiling for precision medicine. The goal of this project is to find therapeutic options that directly target the mechanisms of action driving IBD pathology. By using single-cell RNA sequencing(scRNA-seq) to characterize cell specific gene expression patterns in affected tissues, specifically the small intestine and colon, this project aims to understand mechanisms responsible for inflammation and tissue damage.The approach of my project consists of essentially 3 phases:

1) Population-level drug-pathway analysis: Identifying which biological pathways are dysregulated across specific cell types using gene set enrichment analysis (GSEA) on differential expression patterns across multiple cell populations in CD-affected tissues (specifically the small intestine and colon).
2) Machine learning based patient-level pathology classification: Compressing pathway enrichment data into compact 20-dimensional representations using singular value decomposition (SVD), capturing co-activity patterns across cell types for each patient.
3) Random Forest(RF) based drug prioritization: Integrating SVD pathway embeddings with patient-specific enrichment features through a RF classifier to generate personalized therapeutic recommendations based on each pathway for each patient.

The following sections describe the computational methods applied to 18 Crohn's disease patients, present the results from both small intestine and colon tissues, and discuss the therapeutic implications for precision medicine in IBD management.

Methods:

Data Acquisition Single-cell RNA sequencing data was obtained from the Single Cell Portal (SCP2959)(22). The dataset, "Cleaned_raw_annotated_object_LK.v2.h5ad", consisted of pre-processed and annotated scRNA-seq data from patients with inflammatory bowel disease and control samples, containing 347,017 cells across 33,551 genes from various gastrointestinal

tissues. All computational analysis was performed using Python 3.13 in Visual Studio Code with version-controlled code.

Data Filtering and Quality Control To focus the analysis on CD pathology, the dataset was filtered to include only cells from patients diagnosed with Crohn's disease (MONDO_0005011) and healthy controls. Further tissue-specific filtering was applied to select cells from two primary sites of CD pathology, the small intestine and colon, which were analyzed separately throughout this study to capture tissue-specific disease mechanisms as analyzed in the results. These analyses were performed separately for each tissue to identify tissue-specific drug targets and cellular responses. Cell type annotations from the original dataset (annotation2v2) were used to identify distinct cell populations, which were visualized using UMAP dimensionality reduction to confirm cell population structure.

Population-Level Gene Set Enrichment Analysis (Phase 1) For each annotated cell type in both tissues, differentially expressed genes were identified between Crohn's disease and healthy samples using the Wilcoxon rank-sum test. Gene Set Enrichment Analysis (GSEA) was then performed using 2,186 biological pathways from three databases: KEGG (320 pathways), MSigDB Hallmark (50 pathways), and Reactome (1,816 pathways). GSEA ranked genes by their differential expression scores and assessed whether specific pathway gene sets appeared more frequently at the top or bottom of the ranked list than expected by chance. For each cell type and pathway combination, enrichment metrics were calculated including normalized enrichment score (NES), false discovery rate (FDR), and p-value. Pathways with FDR less than 0.25 and absolute NES greater than 1.5 were retained as significantly enriched, resulting in 4,137 significant enrichments for colon and comparable numbers for small intestine.

Pathway Embedding via Singular Value Decomposition (Phase 2) To reduce the high-dimensional pathway enrichment space into interpretable patterns, pathway embeddings were created using singular value decomposition (SVD). For each significantly enriched pathway, a pathway-by-cell-type matrix was constructed where each entry represented the normalized enrichment score (NES) for that pathway in that cell type. SVD was applied to decompose this matrix into orthogonal components, retaining the top 20 singular vectors to create 20-dimensional pathway embeddings. These embeddings captured co-activity relationships across cell populations, representing how pathways vary together across different cell types. The resulting low-dimensional representations preserved 211 pathways for colon and comparable numbers for small intestine, with each pathway represented as a 20-dimensional vector suitable for machine learning.

Feature Engineering for Drug-Pathway Matching (Phase 3) A comprehensive feature matrix was constructed for training machine learning models to predict drug-pathway compatibility. For each patient-pathway combination, features were engineered by combining the 20-dimensional SVD pathway embeddings with six additional enrichment-derived features: pathway activity score (weighted by cell type proportions in each patient), population-level enrichment signature (mean NES across cell types), cosine similarity of the pathway embedding, mean NES, mean FDR, and relative activity score. This resulted in a 26-dimensional feature vector for each patient-pathway pair. Patient-specific cell type proportions were calculated by counting cells of each type per patient and normalizing to create proportion vectors. Labels for supervised

learning were generated based on whether each patient had Crohn's disease (positive class, labeled 1) or was healthy (negative class, labeled 0), creating a binary classification problem for predicting effective drug-pathway matches.

Random Forest Classifier Training and Optimization Separate Random Forest classifiers were trained for small intestine and colon tissues. Each Random Forest consisted of 500 decision trees with the following hyperparameters optimized to prevent overfitting while maintaining strong predictive performance: maximum depth unconstrained (None), minimum samples required to split an internal node set to 5, minimum samples required at leaf nodes set to 2, and square root of total features considered at each split. Class weights were balanced to account for the imbalance between Crohn's disease and healthy samples (approximately 75% Crohn's, 25% healthy). Features were standardized using z-score normalization (zero mean, unit variance) before training. The dataset was split 80-20 into training and test sets using stratified sampling to maintain class proportions. Models were trained using all available CPU cores for parallel tree construction.

Model Evaluation and Cross-Validation Model performance was evaluated using multiple metrics on the held-out test set: accuracy (percentage of correct predictions), area under the receiver operating characteristic curve (ROC AUC), and area under the precision-recall curve (PR AUC). To assess model robustness and generalization, 5-fold stratified cross-validation was performed, where the data was split into 5 subsets and the model was trained 5 times, each time using 4 subsets for training and 1 for validation. Cross-validation AUC scores were averaged across folds to provide a more stable estimate of model performance. Feature importance scores were extracted from the trained Random Forest models to identify which pathway embedding dimensions and enrichment metrics most strongly influenced predictions. Prediction variance was calculated as the standard deviation of predicted probability scores across all patient-pathway combinations to ensure the model generated diverse, meaningful predictions rather than collapsed uniform outputs.

Patient-Specific Drug Recommendation Generation For each Crohn's disease patient in the dataset, the trained Random Forest model generated prediction scores (0-1 scale) for all pathway combinations, representing the predicted probability that each pathway would be an effective therapeutic target for that patient. These scores combined the patient's specific cellular transcriptomic profile (captured through cell type proportions and pathway activity) with population-level enrichment patterns learned during training. For each patient, pathways were ranked by their prediction scores, and the top 20 highest-scoring pathways were selected as personalized therapeutic recommendations. Drugs targeting these pathways were identified through the enrichment analysis performed in Phase 1, linking pathway recommendations to specific therapeutic compounds. Recommendations were analyzed across all patients to distinguish patient-specific candidates (high scores in individual patients only) from universal candidates (consistently high scores across multiple patients), enabling identification of both personalized and broadly applicable therapeutic strategies.

Visualization and Statistical Analysis Results were visualized through multiple complementary approaches: confusion matrices showing true positive, true negative, false positive, and false negative rates for model predictions; ROC curves and precision-recall curves

displaying model discrimination ability across probability thresholds; feature importance bar plots ranking the contribution of each feature to model predictions; heatmaps showing drug recommendation scores across patients and cell types; cross-validation box plots displaying model stability across folds; and patient-specific bar charts ranking top therapeutic candidates. All visualizations were generated at 300 DPI resolution for publication quality. Statistical significance of enrichment was assessed using false discovery rate correction for multiple hypothesis testing, with FDR threshold of 0.25 considered significant.

Full code can be found at the following github repository:
https://github.com/rohanv-git/aicrohns-py-repo

Model Development and Iteration The pipeline went through several major changes during development. The pathway representation and drug-pathway matching steps were originally built around deep learning, using a feedforward neural network and a Residual Neural Network respectively. The ResNet was chosen initially because it was designed to solve the vanishing gradient problem in deep networks, allowing many layers to be stacked without losing signal, which made it appealing for capturing complex relationships between pathway features. However, both models failed in practice for the same reason: deep learning requires far more labeled training data than this patient cohort could provide, and both collapsed to predicting near-average values for every input. ResNets in particular are also better suited to image recognition tasks where their skip connections help preserve spatial features across many layers, making them a poor fit for tabular biological data like pathway enrichment scores. The feedforward network was replaced with SVD because it identifies the dominant structure in the enrichment data mathematically without needing labeled examples and captured 95% of the variance in fewer than 13 components. The ResNet was replaced with a Random Forest because it performs well on smaller datasets, resists overfitting, and produces interpretable feature importance scores. The pathway database also expanded from drug2cell scoring against ChEMBL alone to gene set enrichment analysis across 2,186 curated pathways from KEGG, Hallmark, and Reactome, connecting gene expression changes to well-characterized biological processes. A feature construction bug causing collapsed predictions was identified and corrected, and hyperparameters were tuned to bring the train-test accuracy gap to an acceptable range. Each change reflected a shift toward methods better matched to the actual scale and structure of the data.

Results

P1 Data Overview and Quality Control

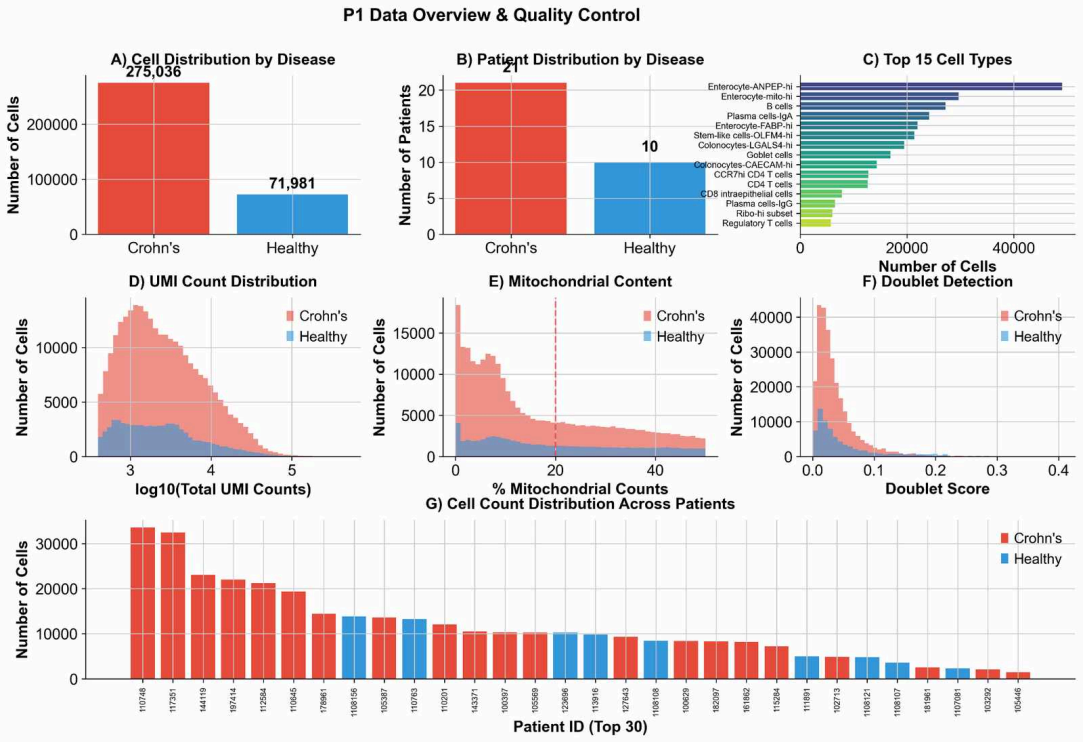


Figure 1. (A) Total cell counts by disease status showing 275,036 Crohn's disease cells and 71,981 healthy controls. (B) Patient distribution showing 21 CD donors and 10 healthy donors. (C) Top 15 most abundant cell types across the full dataset dominated by enterocyte populations with Enterocyte-ANPEP-hi as the largest group at over 40,000 cells. (D) UMI count distribution for CD and healthy cells showing the expected bimodal peak around 1,000 counts confirming adequate sequencing depth. (E) Mitochondrial content distribution with 20% threshold indicated by red dashed line confirming low proportions of damaged cells in both groups. (F) Doublet score distribution concentrated near zero for both groups confirming minimal cell multiplets. (G) Cell count distribution across the top 30 patients by donor ID with CD patients in red and healthy controls in blue.

The dataset contained 347,017 total cells from 31 patients across two primary tissues analyzed separately in this study: the small intestine and colon. Of these, 275,036 cells came from Crohn's disease patients and 71,981 came from healthy controls, spread across 21 CD donors and 10 healthy donors (Figure 1A, B). The 15 most common cell types were dominated by enterocyte populations, with Enterocyte-ANPEP-hi being the largest group at over 40,000 cells, followed by Enterocyte-mito-hi and B cells (Figure 1C). UMI count distributions showed the expected bimodal curves around 1,000 counts for both disease groups, confirming that cells were sequenced at adequate depth (Figure 1D). Most cells fell below the 20% mitochondrial content threshold shown by the red dashed line, which means the dataset had low numbers of damaged or dying cells that could interfere with analysis (Figure 1E). Doublet scores, which detect when two cells are accidentally captured together, were concentrated near zero for both groups, confirming that cell multiplets were rare in this dataset (Figure 1F). Cell counts across the top 30 patients ranged from about 1,000 to 32,000 cells per patient, with healthy donors spread throughout the range rather than clustered separately (Figure 1G).

P2 Data Filtering Summary — Small Intestine and Colon

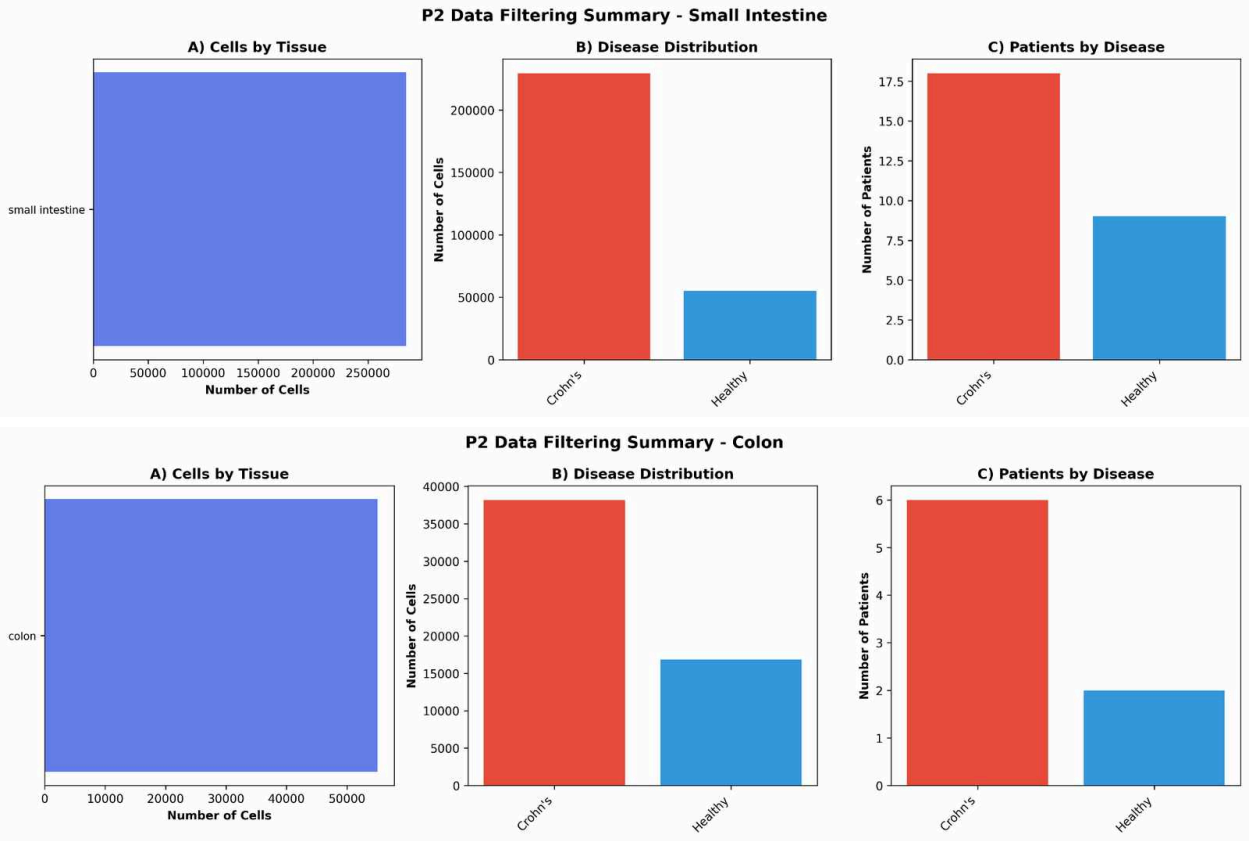


Figure 2. Small intestine: (A) Total cells retained after filtering showing 284,542 cells. (B) Disease distribution showing approximately 225,000 CD cells and 57,000 healthy cells. (C) Patient counts showing 18 CD patients and 9 healthy controls. Colon: same panels showing 55,065 total cells, 6 CD patients, and 2 healthy controls.

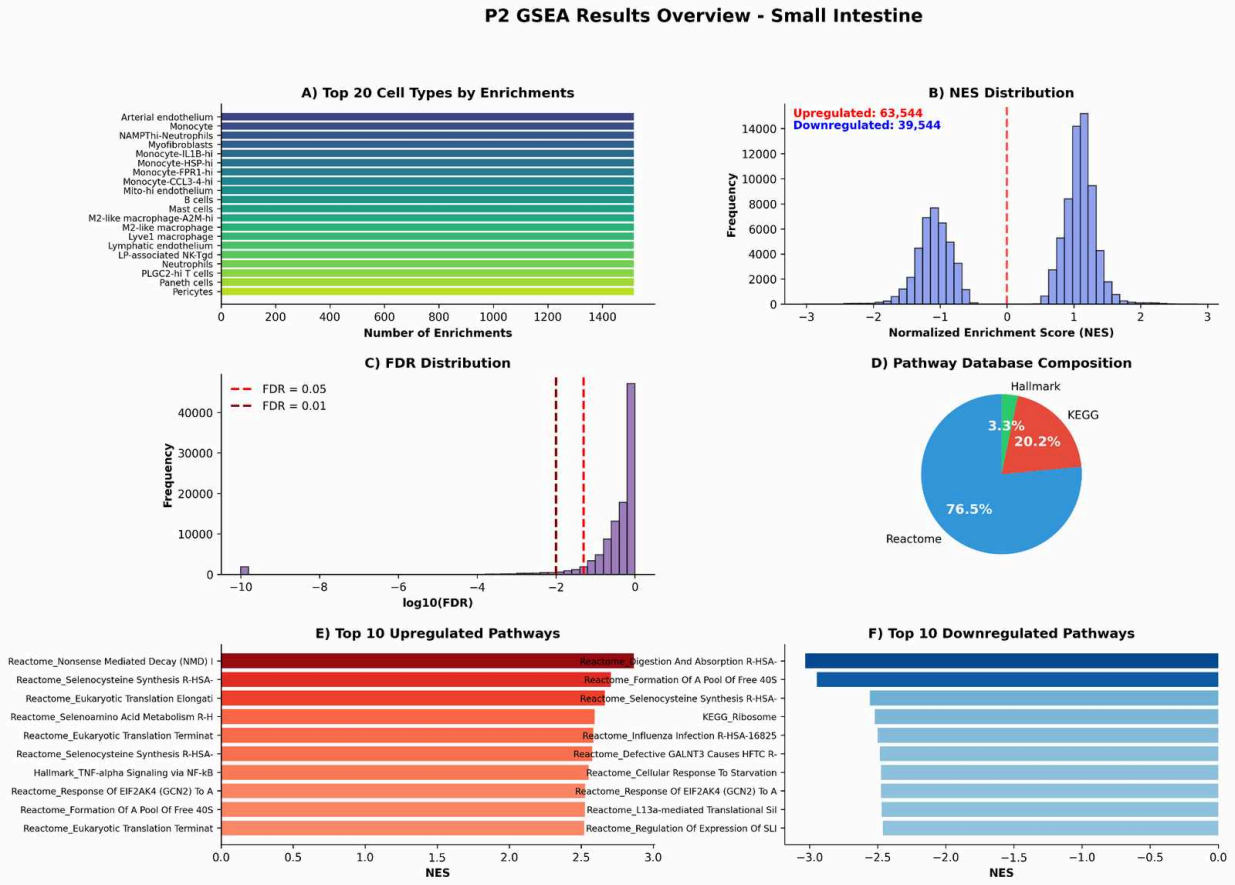
After filtering the full dataset down to just small intestine tissue, 284,542 cells remained from 18 CD patients and 9 healthy controls across 68 annotated cell types (Figure 2, small intestine). The ratio of CD to healthy cells was approximately 4 to 1, reflecting the larger number of CD patients in the study. All 68 cell types were kept for the next steps of analysis.

Colon filtering produced 55,065 cells from 6 CD patients and 2 healthy controls across 61 cell types (Figure 2, colon). The colon dataset was smaller than the small intestine dataset in both total cells and patient count.

Gene ranking is the process of scoring each gene by how differently it is expressed between CD and healthy samples, producing a ranked list from most upregulated to most downregulated for each cell type. This ranked list is then used as input for GSEA to determine which biological pathways are enriched at the top or bottom of the ranking. Gene ranking was performed for all 68 small intestine cell types, with approximately 33,500 genes ranked per cell type based on how differently they were expressed between CD and healthy samples.

Using the same gene ranking approach described above, the 61 colon cell types similarly produced about 33,500 ranked genes per cell type.The colon ranking score distribution was slightly tighter than the small intestine, peaking around a score of 12 to 13. The most abundant colon cell types were Enterocyte-mito-hi and Colonocytes-CEACAM-hi, consistent with what was seen in the colon cell type distribution.

P2 GSEA Results Overview — Small Intestine and Colon



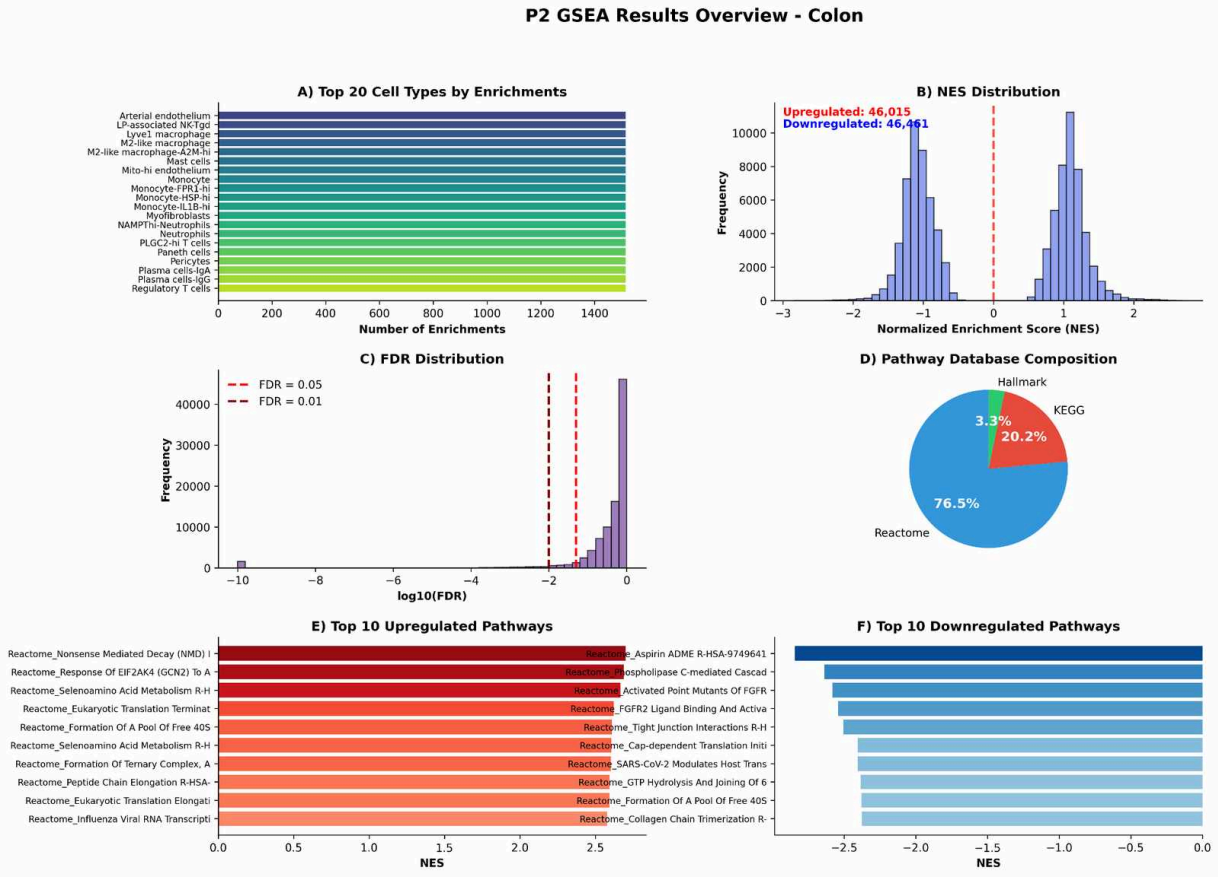


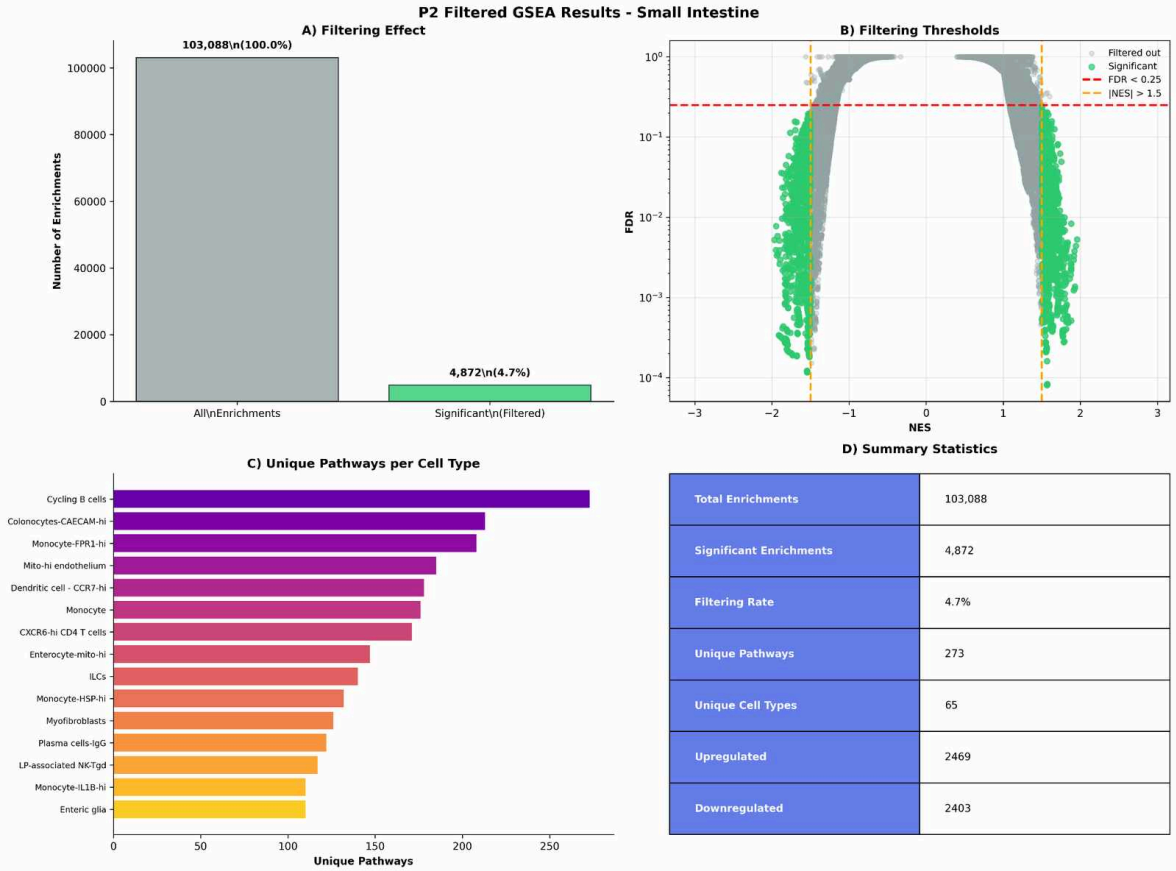
Figure 4. (A) Top 20 cell types by number of enrichments. (B) NES distribution showing 63,544 upregulated and 39,544 downregulated enrichments for small intestine and 46,015 upregulated and 46,461 downregulated for colon. (C) FDR distribution with significance thresholds indicated. (D) Pathway database composition showing Reactome at 76.5%, KEGG at 20.2%, and Hallmark at 3.3%. (E) Top 10 upregulated pathways by NES including Nonsense Mediated Decay and Selenocysteine Synthesis for small intestine. (F) Top 10 downregulated pathways by NES including Digestion and Absorption for small intestine and Reactome Tight Junction Interactions and Collagen Chain Trimerization for colon.

Gene set enrichment analysis across 2,186 pathways identified 103,088 total enrichment tests across all 68 small intestine cell types. Of these, 63,544 were upregulated and 39,544 were downregulated in CD compared to healthy tissue, meaning more pathways were turned up than turned down overall in CD small intestine (Figure 4, small intestine). The top upregulated pathways included Nonsense Mediated Decay, which involves genes such as UPF1, SMG1, and EIF4A3 that regulate degradation of faulty messenger RNA, and Selenocysteine Synthesis, involving SEPHS2 and SEPSECS. The top downregulated pathways included Digestion and Absorption, involving SLC transporters and digestive enzymes such as TREH and SI, suggesting that normal digestive functions were suppressed in CD small intestine while cellular stress responses were activated.

The colon GSEA identified 92,476 total enrichments across 60 cell types, with 46,015 upregulated and 46,461 downregulated, showing a more balanced pattern than the small

intestine (Figure 4, colon). The top downregulated colon pathways included Reactome Tight Junction Interactions, involving genes such as CLDN2, OCLN, and TJP1 that maintain the intestinal barrier, and Collagen Chain Trimerization, involving COL1A1, COL3A1, and COL6A1, both of which are relevant to the barrier dysfunction and fibrosis that are known complications of CD. This difference in the balance of up- and downregulated pathways between the two tissues points to the small intestine and colon having distinct molecular disease patterns in CD.

P2 Filtered GSEA Results — Small Intestine and Colon



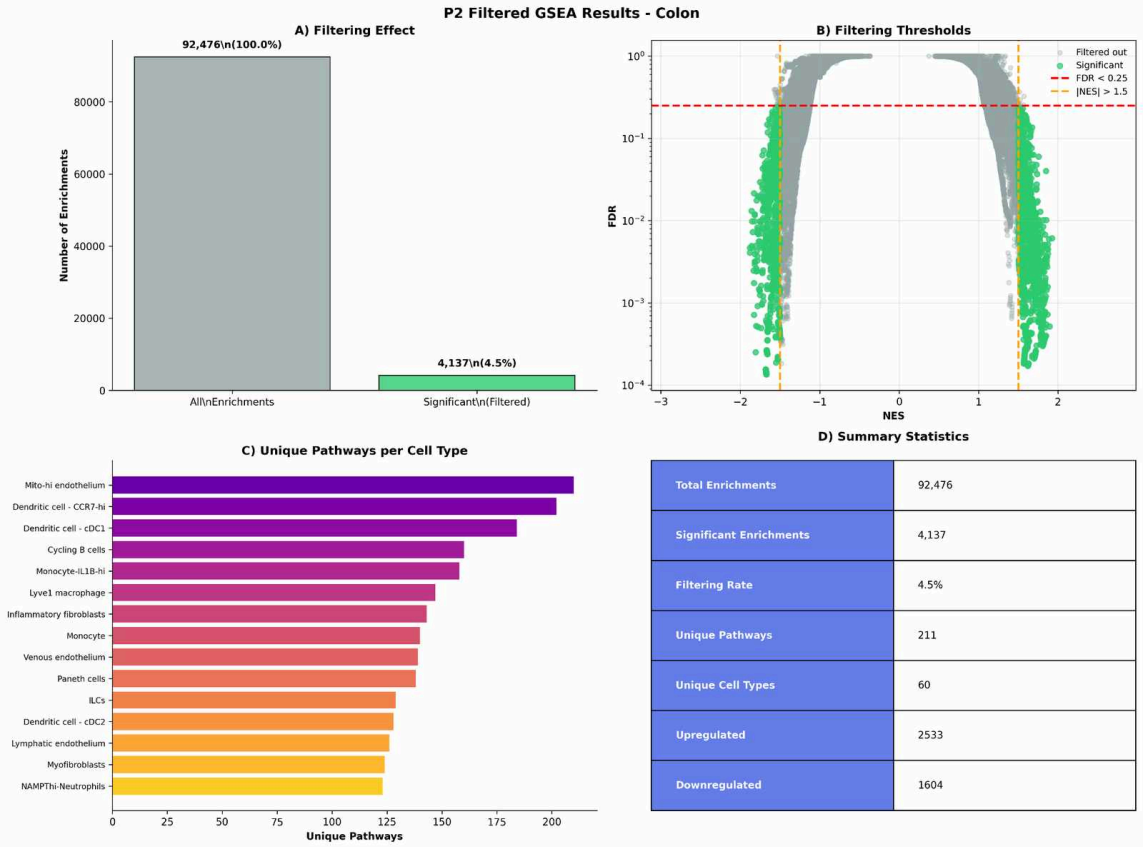


Figure 5. (A) Filtering effect showing reduction from 103,088 to 4,872 significant enrichments (4.7%) for small intestine and from 92,476 to 4,137 (4.5%) for colon. (B) Scatter plot of FDR versus NES confirming retained enrichments meet both significance thresholds of FDR less than 0.25 and absolute NES greater than 1.5. (C) Unique significant pathways per cell type with Cycling B cells showing the highest count at over 250 for small intestine and Mito-hi endothelium and Dendritic cell-CCR7-hi leading for colon. (D) Summary statistics showing total enrichments, significant enrichments, unique pathways, and cell type coverage for each tissue.

Applying the significance cutoffs of FDR less than 0.25 and absolute NES greater than 1.5 reduced the 103,088 small intestine enrichments down to 4,872 statistically significant ones, which is 4.7% of the total (Figure 5, small intestine). The scatter plot of filtering thresholds confirmed that all kept enrichments had the most extreme NES values and the lowest FDR values, meaning the filter kept only the most reliable and biologically meaningful signals. Among cell types, Cycling B cells had the highest number of unique significant pathways at over 250, including pathways related to cell cycle regulation such as Reactome Cell Cycle Checkpoints and Reactome DNA Repair, consistent with the high proliferative activity of cycling immune cells during active inflammation.

Colon filtering kept 4,137 of 92,476 total enrichments, a retention rate of 4.5%, resulting in 211 unique significant pathways across 60 cell types (Figure 5, colon). Mito-hi endothelium and Dendritic cell-CCR7-hi had the highest unique pathway counts at about 200 each. The colon had fewer unique significant pathways than the small intestine (211 vs 273), which is consistent with the smaller number of patients and cells in the colon dataset.

P3 Input Data from P2 — Small Intestine

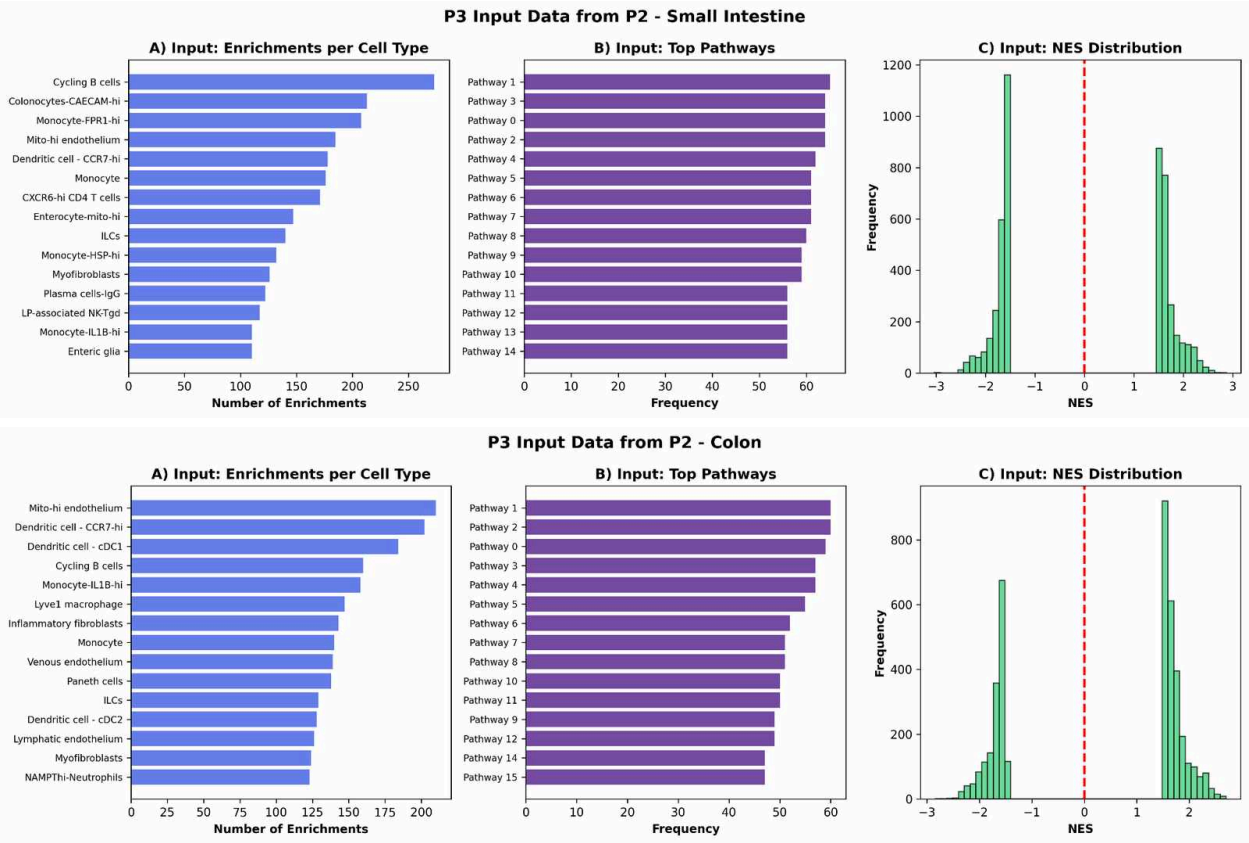
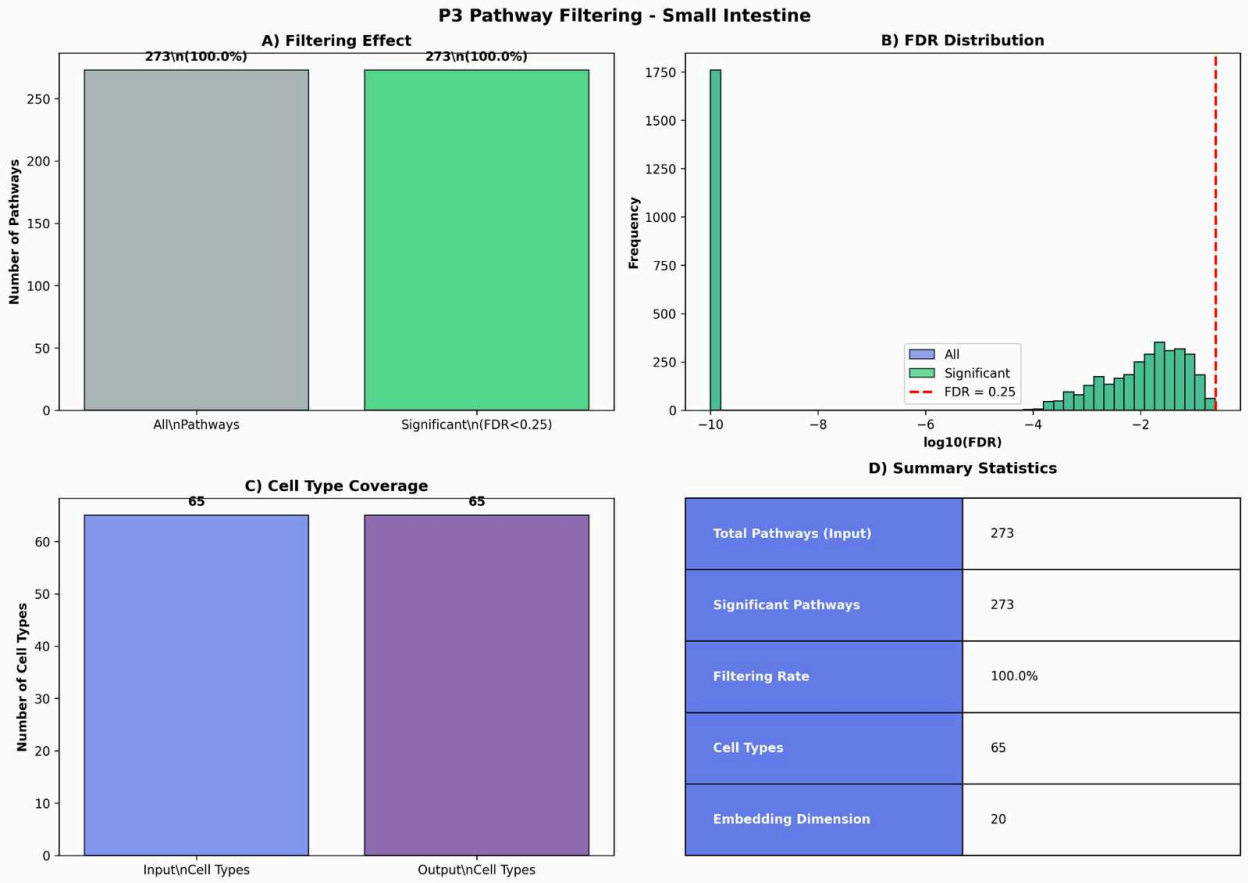


Figure 6. (A) Number of significant enrichments per cell type passed into the embedding stage. (B) Top pathways by frequency across cell types. (C) NES distribution showing a clear bimodal pattern with peaks around -2 and 1.5 for both tissues confirming strong biological signals were passed forward.

The third phase of the pipeline received 4,872 significant enrichments as input, covering 273 unique pathways across 65 cell types for the small intestine. These pathways included significantly enriched biological processes such as Reactome Cell Cycle Checkpoints, Reactome DNA Repair, Nonsense Mediated Decay, and Selenocysteine Synthesis among the upregulated, and Digestion and Absorption among the downregulated. The NES distribution of the input data showed a clear bimodal pattern with one peak around -2 and another around 1.5.

The colon P3 input consisted of 4,137 enrichments across 211 pathways and 60 cell types, with Mito-hi endothelium and Dendritic cell-CCR7-hi contributing the most enrichments. These pathways included significantly enriched processes such as Reactome Tight Junction Interactions and Collagen Chain Trimerization among the downregulated, and mitochondrial and cell cycle related pathways among the upregulated. The NES distribution showed the same bimodal distribution as the small intestine, confirming that the colon input data captured strong and reliable biological signals.

P3 Pathway Filtering — Small Intestine and Colon



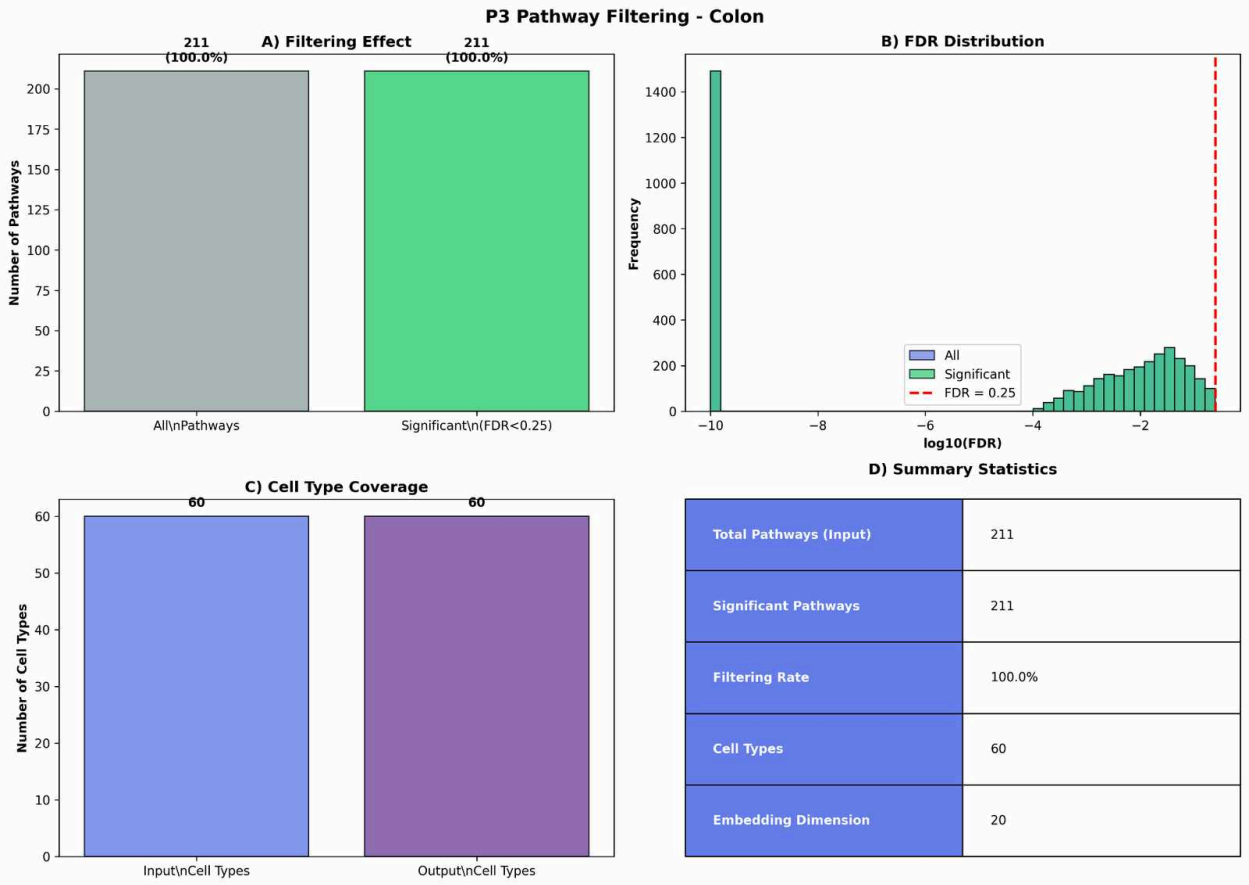
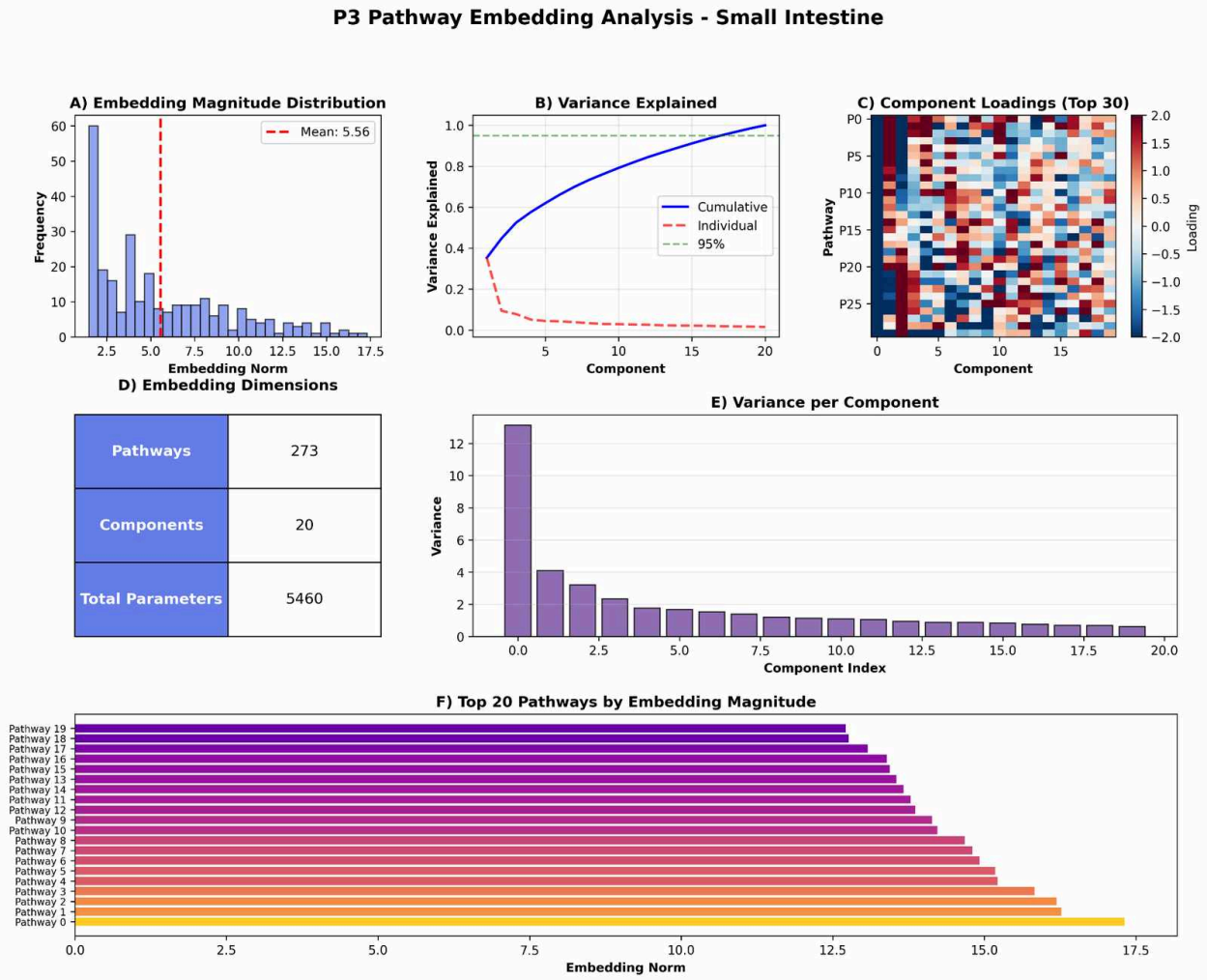


Figure 7. (A) Filtering effect showing all 273 small intestine and all 211 colon pathways passed at 100% retention. (B) FDR distribution showing the vast majority of pathway associations had extremely low FDR values well below the 0.25 threshold. (C) Summary statistics including total pathways, retention rate, cell types retained, and embedding dimension of 20.

All 273 small intestine pathways passed the FDR threshold, resulting in a 100% retention rate. The FDR distribution showed that the vast majority of pathway associations had extremely low FDR values, well below the 0.25 threshold. All 65 cell types were kept, and the embedding dimension was set to 20, meaning each pathway was represented as a vector of 20 numbers where each number captures a distinct axis of co-variation across cell types, giving a total of 273 pathways times 20 components, which equals 5,460 total parameters in the embedding space.

All 211 colon pathways also passed filtering at 100% retention, with all 60 cell types carried into the embedding step. The colon total parameter count was 211 pathways times 20 embedding dimensions, equaling 4,220 total parameters in the embedding space.

P3 Pathway Embedding Analysis — Small Intestine and Colon


P3 Pathway Embedding Analysis - Small Intestine
A) Embedding Magnitude Distribution
Mean: 5.56
Frequency
Embedding Norm
B) Variance Explained
Cumulative
Individual
95%
Variance Explained
Component
C) Component Loadings (Top 30)
Pathway
Loading
Component
D) Embedding Dimensions
Pathways
273
Components
20
Total Parameters
5460
E) Variance per Component
Variance
Component Index
F) Top 20 Pathways by Embedding Magnitude
Embedding Norm

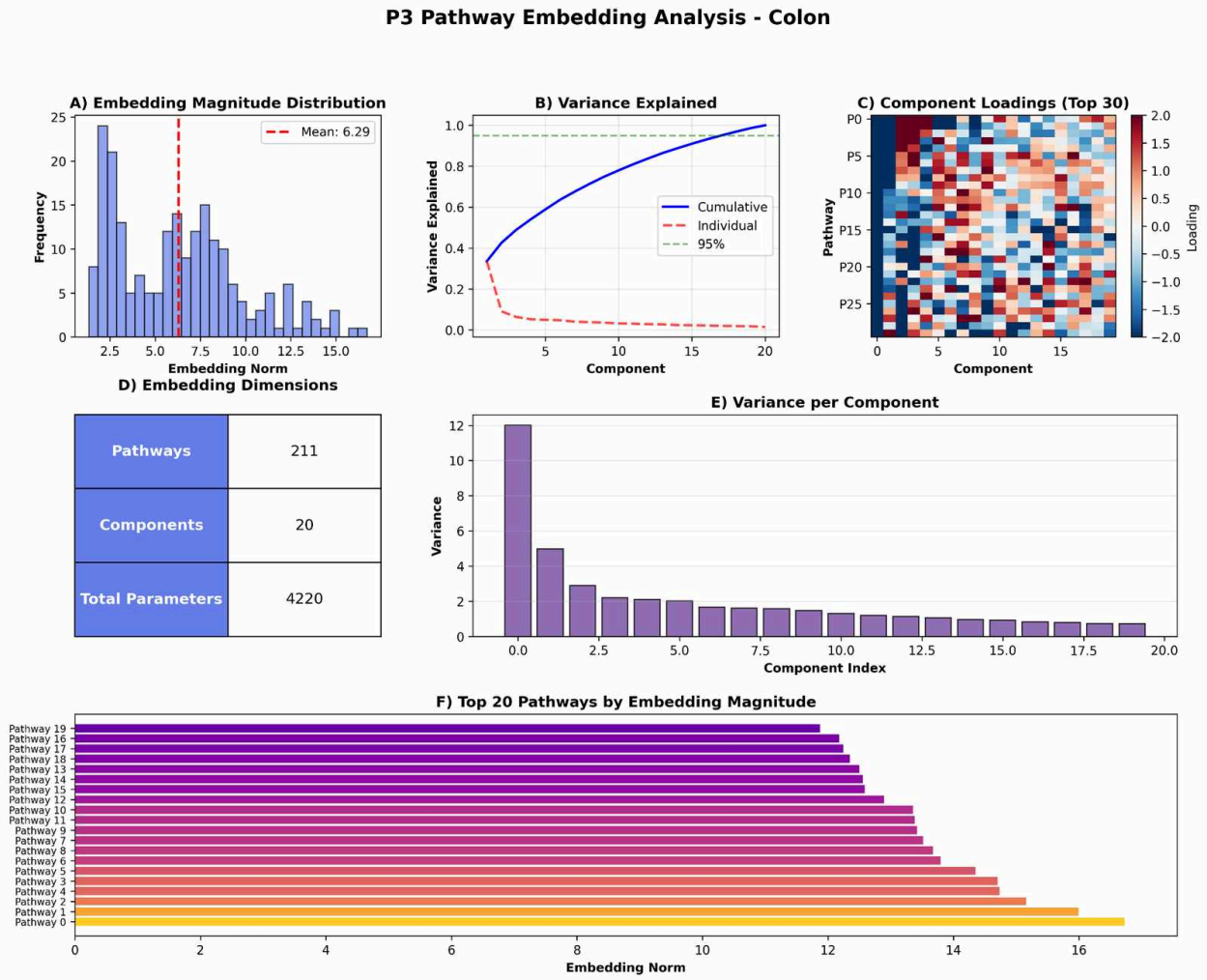


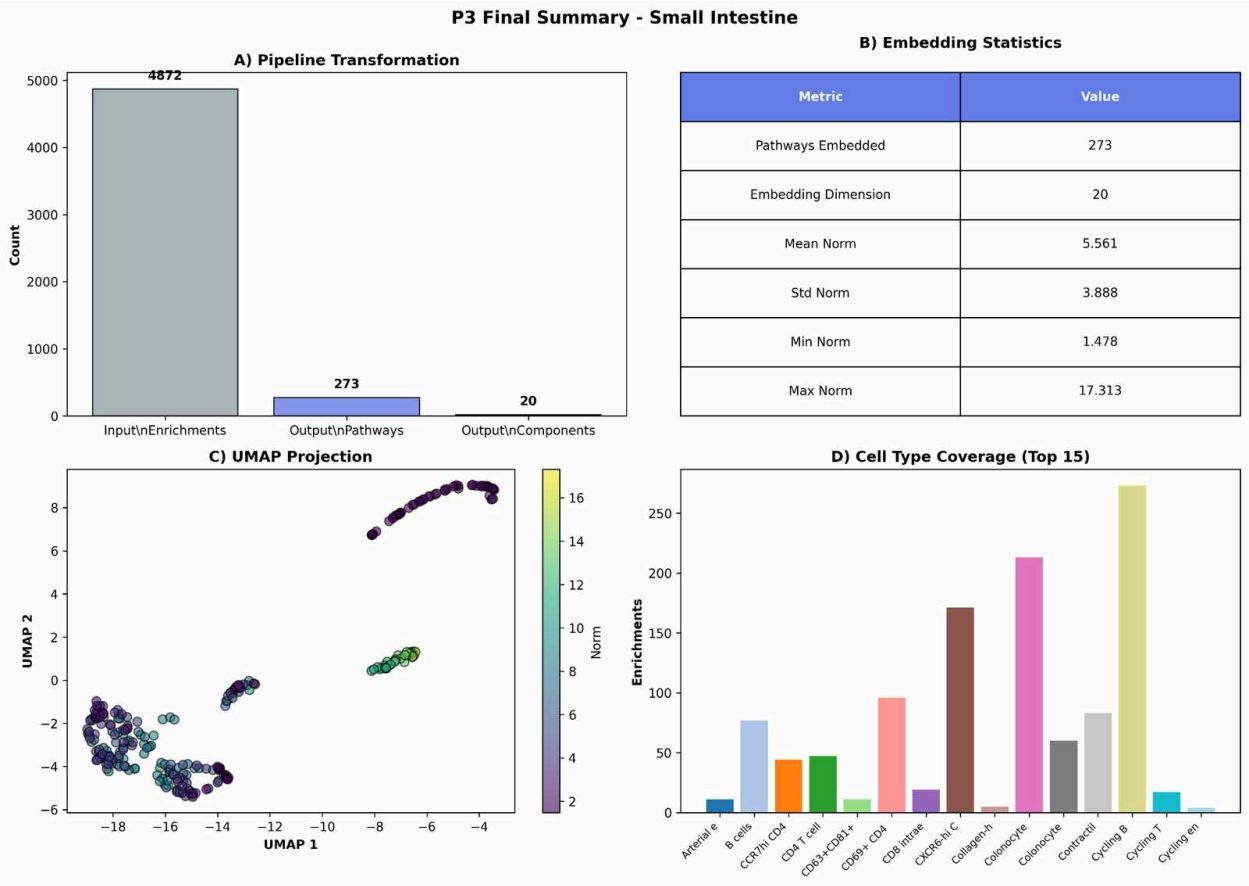
Figure 8. (A) Embedding magnitude distribution with mean norm indicated. (B) Variance explained curve showing cumulative variance reaching 95% by approximately the 12th to 13th component for small intestine and a comparable pattern for colon. (C) Component loadings heatmap showing a mix of positive and negative values across pathways and components indicating different components captured different biological patterns. (D) Variance per component bar chart showing a sharp drop after the first component consistent with one dominant biological signal driving most variation.

SVD applied to the small intestine data produced 20-dimensional pathway embeddings. The variance explained curve showed that the first component alone captured about 35% of the variance, and cumulative variance reached 95% by around the 12th or 13th component, meaning that 20 dimensions captured nearly all of the important structure in the data with very little information lost. The heatmap of component loadings showed a mix of positive and negative values across pathways and components, meaning different components picked up on different biological patterns. The variance per component bar chart showed a sharp drop after the first component, which is the expected shape when one dominant biological signal (the general CD inflammatory response) drives most of the variation. SVD was chosen as a computationally tractable starting approach for this dataset; more specialized single-cell

embedding methods may capture additional biological structure in future work. To validate that the SVD embeddings captured meaningful biological structure, pathways with similar biological functions were confirmed to cluster near each other in the PCA plot of the embedding space, suggesting the embeddings preserved functional relationships between pathways.

The colon SVD embedding produced a mean norm of 6.294 and standard deviation of 3.574, slightly higher than the small intestine. The variance explained curve followed the same pattern, with the first component explaining about 12 units of variance compared to 13 in the small intestine, reflecting the smaller pathway set. The loadings heatmap and variance per component plot both showed the same structure as the small intestine, confirming that the SVD approach worked consistently across both tissues. As with the small intestine, SVD served as a computationally tractable starting point, with more specialized single-cell embedding methods representing a direction for future improvement. PCA validation of the colon embeddings similarly showed functionally related pathways clustering near each other, confirming that the embeddings preserved meaningful biological structure.

P3 Final Summary — Small Intestine and Colon



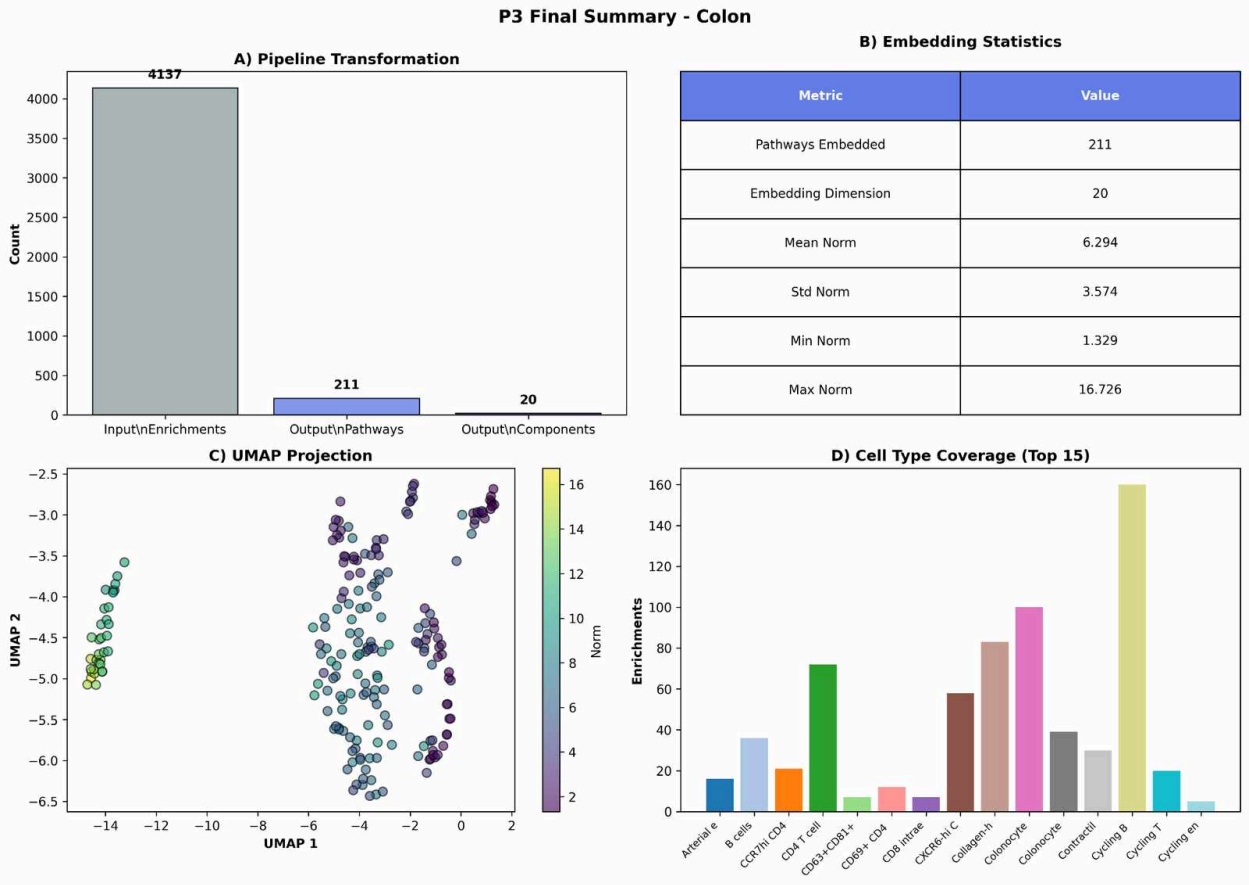


Figure 9. (A) Pipeline transformation showing reduction from 4,872 input enrichments to 273 pathway embeddings of 20 dimensions for small intestine and from 4,137 to 211 embeddings for colon. (B) Embedding statistics table showing mean norm, standard deviation, minimum, and maximum. (C) UMAP projection calculated on the original pathway enrichment matrix showing higher-norm pathways grouping separately from lower-norm pathways for small intestine and more fragmented groupings for colon consistent with the smaller pathway set. (D) Cell type enrichment contribution bar chart with Cycling B cells contributing the most enrichments for both tissues.

The P3 pipeline reduced 4,872 input enrichments down to 273 pathway embeddings of 20 dimensions each, a major reduction in complexity while keeping the key biological information. The UMAP projection was calculated on the original pathway enrichment matrix rather than on the SVD components, and showed that higher-norm pathways grouped separately from lower-norm pathways, suggesting that pathways with stronger co-activity signals across cell types form a distinct subset. Cycling B cells contributed the most enrichments to the embedding, consistent with their high pathway count in the filtered GSEA results.

The colon pipeline reduced 4,137 enrichments down to 211 embeddings of 20 dimensions, with embedding statistics showing a mean norm of 6.294, minimum of 1.329, and maximum of 16.726. The colon UMAP, also calculated on the original enrichment matrix, showed more fragmented groupings compared to the small intestine, likely because the smaller pathway set of

211 created less overlap between pathway neighborhoods. Cycling B cells again contributed the most enrichments to the embedding.

Cross-Tissue Drug-Pathway Analysis Summary

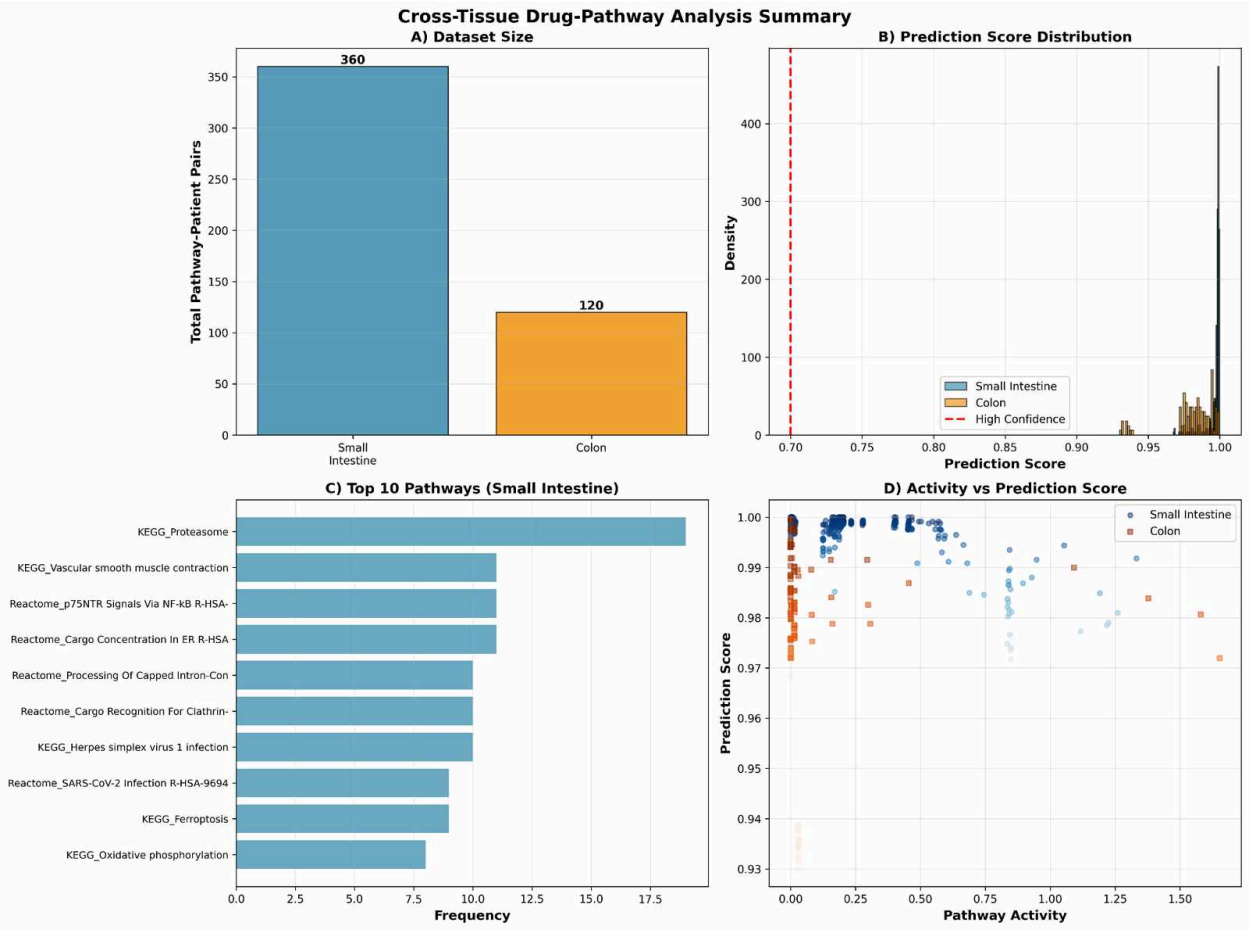


Figure 10. (A) Total pathway-patient pairs generated showing 360 for small intestine and 120 for colon. (B) Prediction score distribution for all recommendations across both tissues with the high confidence threshold of 0.7 indicated showing the vast majority of scores concentrated near 1.0. (C) Top 10 most frequently recommended small intestine pathways with KEGG Proteasome appearing in nearly every patient. (D) Scatter plot of pathway activity versus prediction score for small intestine and colon confirming high confidence predictions spanned a wide range of pathway activity values.

Across both tissues, 360 pathway-patient pairs were generated for the small intestine and 120 for the colon, with all 480 total recommendations receiving high confidence scores. The mean recommendation score was 0.997 for small intestine and 0.981 for colon. The most frequently recommended small intestine pathway was KEGG Proteasome, which appeared in recommendations for 19 of the 18 patient slots, meaning it was a top-20 pathway for almost every patient. The activity versus prediction score scatter plot showed that high-confidence predictions spanned a wide range of pathway activity values, confirming that the model used multiple features to arrive at its predictions rather than relying on pathway activity alone.

Patient x Pathway Recommendation Heatmaps

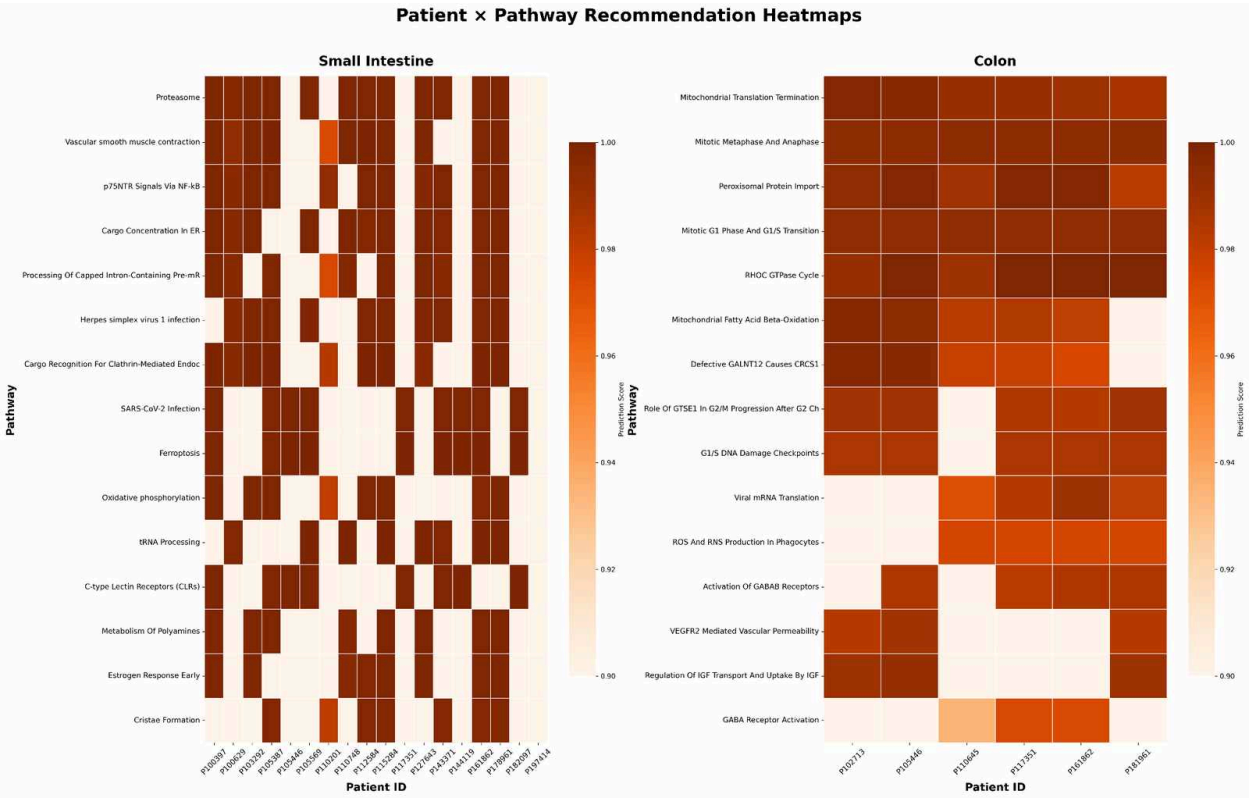


Figure 11. Prediction scores for the top 15 most frequently recommended pathways across all patients for small intestine (left) and colon (right). Darker orange cells indicate higher prediction scores approaching 1.0 while cream-colored cells indicate low scores. Universal therapeutic candidates visible as consistently dark rows include KEGG Proteasome for small intestine and Reactome Mitochondrial Translation Termination for colon. Patient-specific candidates visible as dark cells in only one or two columns include Reactome Metabolism of Polyamines and Hallmark Estrogen Response Early for small intestine and Reactome GABA Receptor Activation and Reactome VEGFR2 Mediated Vascular Permeability for colon.

The heatmaps showing prediction scores across all patients and pathways for both tissues revealed two types of therapeutic candidates. Universal candidates were pathways that received high scores across most or all patients, while patient-specific candidates were pathways that received high scores only for certain patients. For the small intestine, KEGG Proteasome was nearly universal across all 18 patients, while pathways like Reactome Metabolism of Polyamines and Hallmark Estrogen Response Early were high only in certain patients. For the colon, Reactome Mitochondrial Translation Termination and Mitotic Metaphase and Anaphase were nearly universal across all 6 colon patients, while pathways like Reactome VEGFR2 Mediated Vascular Permeability and GABA Receptor Activation were specific to only 1 to 3 patients. The presence of cream-colored (low score) cells throughout both heatmaps confirms that the model was not giving high scores to everything but was making real distinctions between good and poor pathway matches for each patient.

Patient 100629 Small Intestine Analysis

Patient 100629 received a top recommendation score of 0.999 with an up-regulated pathway activity of 0.35. The highest-ranked therapeutic pathways were Reactome Formation of ATP by Chemiosmotic Coupling (0.999), Reactome SARS-CoV-2-host Interactions (0.999), Reactome tRNA Processing (0.998), KEGG Proteasome (0.997), and Reactome Processing of Capped Intron-Containing Pre-mRNA (0.997). The pattern of multiple RNA processing and protein degradation pathways appearing at the top suggests that this patient's CD may be driven in part by cellular stress related to protein production and clearance, which is a mechanism relevant to several existing IBD therapies including immunomodulators.

Patient 115284 Small Intestine Analysis

Patient 115284 had a top recommendation score of 1.000 and a mean pathway activity of 0.1855. The top pathways were Reactome tRNA Processing (1.000), Reactome Cargo Recognition for Clathrin-Mediated Endocytosis (1.000), and Reactome Metabolism of Polyamines (1.000). The high score for polyamine metabolism sets this patient apart from Patient 100629, as polyamine metabolism is linked to cell growth and mucosal repair. This patient also showed high scores for Reactome p75NTR Signals Via NF-kB (0.999) and Reactome Cargo Concentration in ER (0.999), pointing to endoplasmic reticulum stress as an additional therapeutic target. The different top pathway profile compared to Patient 100629 shows that the model successfully identified different disease patterns in two patients from the same tissue.

Patient 117351 Small Intestine Analysis

Patient 117351 had a top recommendation score of 1.000 and a notably higher pathway activity score, with the top pathway, Reactome C-type Lectin Receptors, showing an up-regulated NES of 1.52. The top recommendations included KEGG Ferroptosis (1.000), KEGG Salmonella infection (0.999), KEGG Proteasome (0.999), and Reactome C-type Lectin Receptors (0.999). The appearance of ferroptosis, which is a type of cell death driven by iron-dependent oxidative damage, alongside innate immune receptor pathways suggests that this patient's disease is centered more on oxidative stress and innate immune activation than on the translational stress pattern seen in the other two small intestine patients. This is a meaningful difference that a population-level analysis would not be able to detect.

Patient 102713 Colon Analysis

Patient 102713's top colon recommendations included Reactome Mitochondrial Fatty Acid Beta-Oxidation (0.997), Reactome Defective GALNT12 Causes CRCS1 (0.997), Reactome Mitotic Metaphase and Anaphase (0.995), Reactome Peroxisomal Protein Import (0.994), and Reactome Mitotic G1 Phase and G1/S Transition (0.994). The high scores for cell cycle regulation pathways alongside mitochondrial metabolism pathways suggest that this patient may have significant dysregulation in how the intestinal lining repairs and renews itself, which is consistent with the known difficulty of mucosal healing in stricturing CD. The appearance of GALNT12, a pathway connected to colorectal cancer susceptibility, highlights the model's ability to identify disease-relevant targets beyond the standard inflammation-focused pathways.

Patient 117351 Colon Analysis

Patient 117351 in the colon showed a top recommendation for Reactome RHOC GTPase Cycle with a high activity score of 0.92, followed by Reactome Peroxisomal Protein Import (0.998), Reactome Mitotic Metaphase and Anaphase (0.995), Reactome Mitotic G1 Phase and G1/S Transition (0.994), and Reactome Mitochondrial Translation Termination (0.992). The RhoC GTPase pathway controls how cells organize their internal structure, and its high score alongside mitotic pathways points to disruption of epithelial cell organization in this patient's colon. Comparing this patient's small intestine profile (translational stress and innate immunity) with the colon profile (cell cycle and RhoC signaling) shows clearly that even within the same patient, the disease mechanism varies between tissues, which has direct implications for how therapy might be targeted.

Patient 161862 Colon Analysis

Patient 161862 received top recommendations for Reactome Peroxisomal Protein Import (0.998), Reactome Mitotic Metaphase and Anaphase (0.995), Reactome Mitotic G1 Phase and G1/S Transition (0.994), Reactome Viral mRNA Translation (0.990), and Reactome Mitochondrial Translation Termination (0.990). This patient shared the mitotic pathway signature with the other colon patients but also showed high scores for Reactome Activation of GABAB Receivers and Reactome G1/S DNA Damage Checkpoints. The appearance of GABA receptor pathways is notable because the enteric nervous system, which uses GABA signaling, is increasingly recognized as playing a role in IBD, and this recommendation highlights how the model can identify targets that go beyond conventional inflammatory mechanisms.

Discussion

Data Quality and Study Design

The quality control results confirmed that the dataset was clean and reliable for analysis. Low mitochondrial content, minimal doublets, and adequate sequencing depth across all patients meant that the downstream enrichment and machine learning results were unlikely to be driven by poor-quality cells. The larger number of CD cells compared to healthy cells (275,036 vs 71,981) reflects a study design focused on characterizing CD pathology, and this imbalance was corrected for during model training using balanced class weights so that the classifier did not simply learn to predict the majority class.

Tissue-Specific Enrichment Patterns and Their Biological Meaning

One of the most informative findings from the GSEA phase was the difference in enrichment patterns between the small intestine and colon. The small intestine showed more upregulated pathways than downregulated (63,544 vs 39,544), while the colon was nearly balanced (46,015 vs 46,461). This suggests that CD affects the small intestine mainly by turning up inflammatory and stress response programs, while in the colon the disease both activates some pathways and suppresses others more evenly. The suppression of tight junction and collagen organization pathways in the colon is particularly important because these pathways control how well the gut wall holds together and resists damage. Their downregulation is consistent with the barrier

dysfunction and fibrosis that are hallmarks of CD, and is supported by prior studies of CD stricturing(15,16).

The presence of Hallmark TNF-alpha Signaling via NF-kB among the most upregulated small intestine pathways is a direct confirmation of a known therapeutic target, because anti-TNF drugs such as infliximab and adalimumab are among the most widely used CD treatments(6). Identifying this pathway as highly enriched in the small intestine analysis provides biological validation that the GSEA results are capturing real disease-relevant signals rather than statistical noise.

Why SVD-Based Pathway Embeddings Work

The SVD results showed that most of the structure in the pathway enrichment data could be captured in far fewer than 20 dimensions, with 95% of variance explained by around 12 to 13 components. This rapid variance decay suggests that CD pathway activity across cell types is organized around a small number of major biological themes rather than being completely chaotic and unpredictable. The first and largest component most likely represents the general CD inflammatory response that is shared across many cell types, while later components capture more subtle cell-type-specific or pathway-specific patterns.

The fact that pathway embeddings spread out continuously in PCA space without forming discrete clusters is biologically meaningful. It suggests that the pathways involved in CD exist along a spectrum of related functions and are co-regulated in complex ways, rather than falling neatly into a few independent modules. This is consistent with the known biology of IBD, where immune activation, metabolic changes, barrier dysfunction, and fibrosis all interact with each other. The approach of using matrix decomposition to capture these co-activity relationships is similar in spirit to how GenePT uses language model embeddings to capture functional relationships between genes(18), though the SVD method used here works directly on the enrichment data and is therefore more directly interpretable for this specific application.

Understanding Model Performance

The Random Forest models achieved cross-validation AUC values of 0.811 and 0.767 for the small intestine and colon respectively. To put these numbers in context, an AUC of 0.5 means the model is no better than random guessing, while 1.0 means perfect prediction. Values in the range of 0.75 to 0.85 are generally considered good performance for biological prediction tasks, especially when working with limited patient numbers. The consistency of scores across the five cross-validation folds (standard deviation of 0.011 for small intestine and 0.031 for colon) shows that the models learned patterns that generalize beyond any single subset of the training data.

The difference in performance between the colon (0.767) and small intestine (0.811) models is likely caused by the much smaller colon training dataset, which had only 1,688 samples from 8 patients compared to 7,371 samples from 27 patients for the small intestine. With fewer training examples, the colon model had less variety to learn from, which explains both the lower AUC and the larger variation between folds.

What the Recommendations Tell Us About CD

The patient-pathway heatmaps and individual patient analyses identified two distinct types of therapeutic targets. Universal targets were pathways that received high recommendation scores across most or all patients in a tissue, suggesting they represent core disease mechanisms that are relevant regardless of individual differences. For the small intestine, KEGG Proteasome was the most universal target, appearing in top recommendations for nearly every patient. The proteasome is the cellular machinery responsible for breaking down damaged or unneeded proteins, and problems with this system have been linked to IBD because impaired protein degradation can allow inflammatory signaling molecules to build up inside cells(10). For the colon, the near-universal targets were Reactome Mitochondrial Translation Termination and mitotic cell cycle pathways, suggesting that energy production in mitochondria and cell division in the intestinal lining are core disease-relevant processes in colon CD.

Patient-specific targets were more interesting from a precision medicine perspective because they pointed to different disease mechanisms in different individuals. Patient 117351 in the small intestine stood out from the other patients by showing high scores for ferroptosis and C-type lectin receptor pathways, pointing to oxidative stress and innate immune activation as the main drivers of disease in this patient. Patients 100629 and 115284, by contrast, had top pathways related to RNA processing and protein transport, suggesting a cellular stress pattern centered on protein production and clearance. These differences would not be visible in a standard population-level analysis, which would average across all patients and potentially miss these individual-level signals.

The cross-tissue comparison of Patient 117351, who appeared in both the small intestine and colon analyses, was one of the most informative results in the study. This patient's small intestine profile was dominated by translational stress and innate immune pathways, while the colon profile was dominated by cell cycle regulation and RhoC GTPase signaling. This finding directly demonstrates that even in a single patient, CD affects different parts of the gut through different molecular mechanisms. In a clinical setting, this would suggest that the same patient might respond differently to the same therapy depending on which part of the gut is being treated, which is an insight that could only come from a tissue-specific, patient-specific analysis like this one.

Limitations

There are several important limitations to acknowledge. The classification models were trained to distinguish CD tissue from healthy tissue, not to predict whether a specific drug would work for a specific patient. The drug recommendations come from the fact that highly enriched pathways in CD are also pathways targeted by certain drugs, but there is no direct evidence in this dataset that those drugs would actually cause remission. Experimental testing of the top predicted pathway-drug pairs in laboratory models would be a necessary step before drawing any clinical conclusions(16).

The colon cohort of only 6 CD patients and 2 healthy controls is very small, and the colon model results should be interpreted with extra caution compared to the small intestine results. The dataset also comes from a specific patient population, so it is not certain that the findings would

apply equally to all CD patients in different geographic or demographic groups. Finally, because the data was collected at a single point in time rather than tracking patients over months or years, the analysis cannot capture how pathway activity changes with disease progression or treatment.

Conclusion

In conclusion, this pipeline identified potential therapeutic options at both the population and individual patient level for Crohn's disease. Universal candidates including KEGG Proteasome and Reactome Mitochondrial Translation pathways represent targets broadly relevant across patients, while patient-specific candidates including ferroptosis, polyamine metabolism, RhoC GTPase signaling, and GABA receptor pathways point to distinct disease mechanisms that population-level approaches would miss.

This project is not a finished clinical tool, and its predictions should be treated as hypotheses rather than conclusions. However, even with a limited patient cohort, the pipeline produced biologically plausible results and captured real patient-level heterogeneity, which suggests the core approach has genuine potential. With more data, larger cohorts, and experimental validation, a pipeline like this could grow into something that meaningfully aids doctors in making personalized treatment decisions. For CD patients who currently cycle through therapies through trial and error, particularly younger patients where prolonged disease activity affects growth and development, this kind of precision medicine framework represents a promising direction worth pursuing further.

Bibliography

1. Lewis JD, Parlett LE, Jonsson Funk ML, Brensinger C, Pate V, Wu Q, et al. Incidence, Prevalence, and Racial and Ethnic Distribution of Inflammatory Bowel Disease in the United States. Gastroenterology. 2023 Nov;165(5):1197-1205.e2. doi:10.1053/j.gastro.2023.07.003

2. Groundbreaking Study Estimates Nearly 1 in 100 Americans Has IBD [Internet]. Crohn’s & Colitis Foundation. Available from: https://www.crohnscolitisfoundation.org/groundbreaking-study-led-the-crohns-colitis-foundation-estimates-nearly-1-100-americans-has

3. Kontola K, Oksanen P, Huhtala H, Jussila A. Increasing Incidence of Inflammatory Bowel Disease, with Greatest Change Among the Elderly: A Nationwide Study in Finland, 2000–2020. J Crohns Colitis. 2023 May 3;17(5):706–11. doi:10.1093/ecco-jcc/jjac177

4. Rosen MJ, Dhawan A, Saeed SA. Inflammatory Bowel Disease in Children and Adolescents. JAMA Pediatr. 2015 Nov 1;169(11):1053. doi:10.1001/jamapediatrics.2015.1982

5. Benchimol EI, Fortinsky KJ, Gozdyra P, Van Den Heuvel M, Van Limbergen J, Griffiths AM. Epidemiology of pediatric inflammatory bowel disease: A systematic review of international trends: Inflamm Bowel Dis. 2011 Jan;17(1):423–39. doi:10.1002/ibd.21349

6. Torres J, Mehandru S, Colombel JF, Peyrin-Biroulet L. Crohn's disease. The Lancet. 2017 Apr;389(10080):1741–55. doi:10.1016/S0140-6736(16)31711-1

7. Sabino J, Verstockt B, Vermeire S, Ferrante M. New biologics and small molecules in inflammatory bowel disease: an update. Ther Adv Gastroenterol. 2019 Jan;12:1756284819853208. doi:10.1177/1756284819853208

8. Colombel JF, Sandborn WJ, Reinisch W, Mantzaris GJ, Kornbluth A, Rachmilewitz D, et al. Infliximab, Azathioprine, or Combination Therapy for Crohn's Disease. N Engl J Med. 2010 Apr 15;362(15):1383–95. doi:10.1056/NEJMoa0904492

9. Colombel JF, Panaccione R, Bossuyt P, Lukas M, Baert F, Vaňásek T, et al. Effect of tight control management on Crohn's disease (CALM): a multicentre, randomised, controlled phase 3 trial. The Lancet. 2017 Dec;390(10114):2779–89. doi:10.1016/S0140-6736(17)32641-7

10. Roda G, Chien Ng S, Kotze PG, Argollo M, Panaccione R, Spinelli A, et al. Crohn's disease. Nat Rev Dis Primer. 2020 Apr 2;6(1):22. doi:10.1038/s41572-020-0156-2

11. Bertin L, Crepaldi M, Zanconato M, Lorenzon G, Maniero D, De Barba C, et al. Refractory Crohn's Disease: Perspectives, Unmet Needs and Innovations. Clin Exp Gastroenterol. 2024 Oct;Volume 17:261–315. doi:10.2147/CEG.S434014

12. Bernstein CN, Hitchon CA, Walld R, Bolton JM, Sareen J, Walker JR, et al. Increased Burden of Psychiatric Disorders in Inflammatory Bowel Disease. Inflamm Bowel Dis. 2019 Jan 10;25(2):360–8. doi:10.1093/ibd/izy235

13. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med. 2018 Aug;50(8):1–14. doi:10.1038/s12276-018-0071-8

14. Zhang X, Li T, Liu F, Chen Y, Yao J, Li Z, et al. Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems. Mol Cell. 2019 Jan;73(1):130-142.e5. doi:10.1016/j.molcel.2018.10.020

15. Mukherjee PK, Nguyen QT, Li J, Zhao S, Christensen SM, West GA, et al. Stricturing Crohn's Disease Single-Cell RNA Sequencing Reveals Fibroblast Heterogeneity and Intercellular Interactions. Gastroenterology. 2023 Nov;165(5):1180–96. doi:10.1053/j.gastro.2023.07.014

16. Kong L, Subramanian S, Segerstolpe Å, Tran V, Shih AR, Carter GT, et al. Single-cell and spatial transcriptomics of stricturing Crohn's disease highlights a fibrosis-associated network. Nat Genet. 2025 Jul;57(7):1742–53. doi:10.1038/s41588-025-02225-y

17. Thomas T, Friedrich M, Rich-Griffin C, Pohin M, Agarwal D, Pakpoor J, et al. A longitudinal single-cell atlas of anti-tumour necrosis factor treatment in inflammatory bowel disease. Nat Immunol. 2024 Nov;25(11):2152–65. doi:10.1038/s41590-024-01994-8

18. Chen Y, Zou J. GenePT: A Simple But Effective Foundation Model for Genes and Cells Built From ChatGPT [Internet]. Bioinformatics; 2023 [cited 2026 Apr 19]. Available from: http://biorxiv.org/lookup/doi/10.1101/2023.10.16.562533 doi:10.1101/2023.10.16.562533

19. Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011 Jan;12(1):56–68. doi:10.1038/nrg2918

20. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, et al. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol. 2018 Jun;14(6):e8124. doi:10.15252/msb.20178124

21. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017 Dec;18(1):83. doi:10.1186/s13059-017-1215-1

22. Human CD Fibrosis Study Using Single-Cell and Spatial Data [Internet]. Single Cell Portal, Broad Institute. Available from: https://singlecell.broadinstitute.org/single_cell/study/SCP2959/human-cd-fibrosis-study-using-single-cell-and-spatial-data