

Using machine learning and molecular docking technology to Elucidate the Molecular Pathway of Perfluorooctanoic Acid-Induced Ulcerative Colitis

✉ ✉

Shanxi University of Chinese Medicine

✉ ✉

shz19650911@163.com

XXXXXXXX

Research Article

Keywords: Perfluorooctanoic acid, Ulcerative colitis, Bioinformatics, Machine learning, Molecular docking

Posted Date: October 28th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11489512/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Using machine learning and molecular docking technology to Elucidate the Molecular Pathway of Perfluorooctanoic Acid- Induced Ulcerative Colitis

Authors: Dingyan Guo¹, Huazhong Sun^{2*}

Affiliations:

¹ The First Clinical College, Shanxi University of Chinese Medicine, Taiyuan, 030000, China.

² Shanxi Provincial People's Hospital, Taiyuan, 030000, China.

Corresponding Author:

Huazhong Sun (孙化中)

Email: SHZ19650911@163.com

Tel: 13703589069

Abstract:

Objective: This study aims to elucidate the molecular mechanisms through which perfluorooctanoic acid (PFOA) triggers ulcerative colitis (UC). **Methods:** A differential expression analysis was conducted utilizing multiple datasets to pinpoint genes associated with UC. A multi-faceted computational analysis—leveraging network toxicology revealed 24 genes as potential molecular targets linking PFOA exposure to ulcerative colitis pathogenesis. Further machine learning discerned six core genes (MEP1B, NAAA, PANK3, SCD, SLC23A1, SLC6A14) as pivotal regulators. Notably, SLC6A14 and SCD exhibited significant upregulation, whereas MEP1B, NAAA, PANK3, and SLC23A1 showed marked downregulation ($P < 0.05$). Subsequently, molecular docking analysis revealed that PFOA exhibited high-affinity binding with six machine learning-derived core genes, suggesting its potential contribution to UC pathogenesis through these targets and their associated signaling pathways. These observations provide critical insights for future investigations into the mechanisms underlying PFOA-induced UC.

Keywords: Perfluorooctanoic acid, Ulcerative colitis, Bioinformatics, Machine learning, Molecular docking

Funding: This work was supported by the Natural Science Foundation of Shanxi Province (Grant No. 20210302123355) and the Research Project of Shanxi Provincial Administration of Traditional Chinese Medicine (Grant No. 2023ZYA022).

1. Introduction

Ulcerative colitis is a chronic inflammatory disease that primarily affects the colon, characterized by a complex etiology and pathogenesis. According to contemporary medical literature, it is associated with factors such as immune response, genetics, and environmental factors, infectious agents, and gut microbiota dysbiosis. The disease often follows a prolonged course with a tendency for recurrence, manifesting clinically through symptoms such as chronic diarrhea, presence of mucus or blood in the stool, frequently accompanied by abdominal discomfort. A subset of patients may exhibit atypical presentations, including isolated constipation or an absence of overt bloody stools [1]. Recent epidemiological data indicate a rapidly increasing incidence of uc in China [2, 3].

An epidemiological study investigated the health effects of PFOA exposure among workers, where the PFOA originated from emissions of a DuPont manufacturing facility. This study evaluated six autoimmune disorders (including ulcerative colitis) within the PFOA-exposed cohort. The sole significant finding was an augmented risk of ulcerative colitis that correlated with the level of PFOA exposure [4]. Research exploring the association between PFAS exposure and inflammatory bowel disease remains limited and primarily focuses on PFOA. In a retrospective analysis conducted in 2013 by Steenland et al., which assessed a community cohort of 32,000 individuals, a marked link between serum PFOA levels and UC was established, although no such relationship was observed with Crohn's disease [5]. A subsequent cohort study in 2015 by Steenland et al., involving over 3,000 workers, similarly identified a positive correlation between PFOA exposure and UC [6].

PFOA is emblematic of synthetic perfluoroalkyl substances (PFAS). Reports from various global regions corroborate the ubiquitous contamination of drinking water sources with PFAS accessible to the general populace [7, 8, 9]. An increasing body of research documents associations between PFAS and detrimental health impacts [10, 11]. Due to its thermal stability and chemical inertness, PFOA exhibits exceptional resistance to degradation [12, 13]. This persistence in the environment allows for long-distance migration via atmospheric transport, leading to contamination of distant water bodies and soil. Once introduced into the human body, PFOA can accumulate (predominantly within the blood, liver, and kidneys), proves difficult to eliminate through metabolic processes, and has the potential to biomagnify within the food chain. Research has demonstrated that the prolonged accumulation of perfluorooctanoic acid (PFOA) within the human system correlates with a myriad of health complications, including reproductive toxicity [14], nonalcoholic fatty liver disease [15], thyroid disorders [16], and malignancies such as testicular and renal cancers [17].

PFOA is proposed to promote ulcerative colitis (UC) through multiple mechanisms, including inducing a shift in tissue macrophages toward the anti-inflammatory M2 phenotype and stimulating a TH2-like immune reaction to particular antigens [18]. It could potentially disrupt the balance between TH1 and TH2 immune responses, which results in an elevation of TH2-like cytokines associated with hypersensitivity. Of particular note is interleukin-13 (IL-13), a pivotal cytokine in UC that is instrumental in driving and orchestrating mucosal inflammation, although the precise effects of IL-13 in this context remain to be fully elucidated [19]. The microbial ecosystem is essential for understanding the interaction between PFOA and UC. PFOA exposure is associated with systemic neutropenia and an increase in lipopolysaccharide (LPS)-associated cytokines released by macrophages, which may heighten the host's vulnerability to infections.[20]. Moreover, PFOA impacts non-specific immune responses, including the integrity of the intestinal mucosal barrier and humoral immunity [21]. Additionally, 16S rRNA gene sequencing has revealed that PFOA exposure leads to diminished gut microbiota diversity, an upsurge in the abundance of potential pathogens (e.g., Proteobacteria), and a marked decrease in beneficial bacteria such as Lactobacillus, indicating dysbiosis within the gut microbiota [22].

Despite advancements in understanding the relationship between PFOA and UC, existing research is fraught with significant limitations. Primarily, studies often focus predominantly on epidemiological aspects while neglecting comprehensive linear pathway analyses and failing to construct systemic "exposure-response-effect" networks that could unravel intricate associations. Furthermore, current investigations have typically overlooked the dynamic interactions between genomic damage and the remodeling of the immune microenvironment, complicating efforts to clarify pathogenic mechanisms. To bridge this research gap, the field of network toxicology presents a novel analytical framework. This approach integrates multi-omics datasets, constructs dynamic network models, and utilizes computational structural analysis, thereby facilitating a systematic exploration of how the molecular architectures of chemicals, such as PFOA, affect their biological activities (or chemical properties). Consequently, this methodology serves as a crucial tool for deciphering complex pathogenic processes.

In line with this framework, the current study employed a triangulated systems toxicology strategy: Initially, focusing on PFOA-induced UC, we reconstructed the pathogenesis network through the integration of multi-omics data. Subsequently, guided by topological and functional enrichment analyses, we prioritized key hub molecules within the network for screening via machine learning techniques. Lastly, through the employment of molecular docking technology, we validated the binding affinity and thermodynamic properties of

interactions between these key molecules and potential targets. Ultimately, the goal of this study is to clarify the multiscale pathogenic network associated with PFOA-induced UC and to identify druggable nodes that may facilitate chemical prevention and targeted treatment of UC, thus addressing a critical and unmet medical requirement within Vulnerable groups.

2. Methods

2.1 Screening of Disease-Relevant Targets

Four transcriptomic datasets related to ulcerative colitis (UC) (GSE87473, GSE38713, GSE107499, GSE224758) were sourced from GEO database. We designated GSE87473 and GSE38713 as the experimental cohort and combined GSE107499 with GSE224758 to form the validation cohort. To mitigate batch effects, two normalization strategies were employed: Surrogate Variable Analysis (SVA) was performed using the SVA package to model and adjust for unknown confounders, and ComBat correction was applied to remove residual batch variations via a parametric empirical Bayes framework. The application of Principal Component Analysis (PCA) on the adjusted dataset demonstrated a marked enhancement in the clustering of samples from distinct batches within the reduced-dimensional space, thereby validating the effectiveness of the data correction process. The comprehensive analytical procedure is depicted in Figure 1.

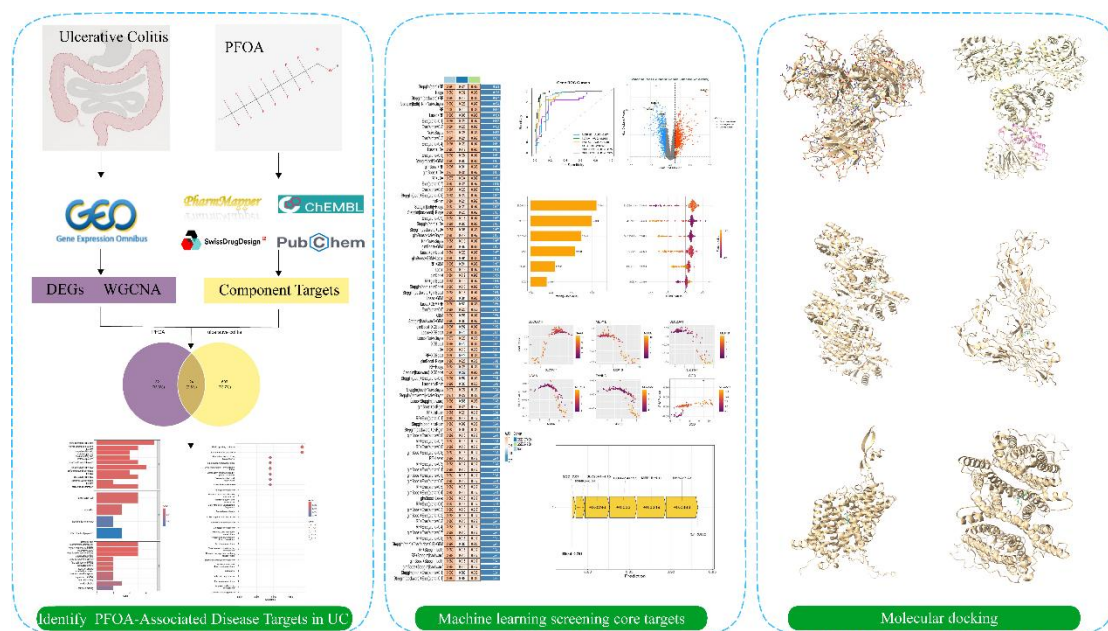


Figure 1 Flow-chart of datasets analysis in this paper.

2.2 PFOA Chemistry and Target Identification

Perfluorooctanoic acid (PFOA) was sourced from various databases, with its molecular

properties and biological activity systematically extracted from the PubMed database. The standard two-dimensional structural representation (SMILES code: C(=O)(C(C(C(C(C(C(C(F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)O) was retrieved from the PubChem database. Target prediction was conducted utilizing three databases: the ChEMBL database for molecular recognition and binding analyses; the SwissTargetPrediction database for chemogenomics-based forecasts (URL: <http://www.swisstargetprediction.ch>); and the PharmMapper database for three-dimensional pharmacophore matching (URL: <http://lilab-ecust.cn/pharmmapper>).

2.3 Differential Gene Expression Analysis

The limma package was employed for the analysis of transcriptomic data. The criteria established for identifying differentially expressed genes (DEGs) included an adjusted false discovery rate (FDR) value of $P < 0.05$ and an absolute log₂ fold change (log₂FC) greater than 0.585, which corresponds to a 1.5-fold change in expression. The resultant data were visualized employing the ggplot2 package.

2.4 The construction of a weighted gene co-expression network (WGCNA)

Using the WGCNA approach, we built a co-expression network of genes that exhibits a scale-free topology. Sample quality control involved the elimination of outliers via hierarchical clustering. The soft threshold power was selected by applying the dynamic tree cut method to achieve a scale-free topology fit index (R^2) greater than 0.85. Modules were identified through hierarchical clustering based on the Topological Overlap Matrix (TOM). The clustering was performed with a lower limit of 30 genes per module, and modules with a dissimilarity below 0.25 were subsequently merged. The association between module eigengenes and phenotypic traits was calculated, with a Pearson correlation coefficient threshold of $|r| > 0.5$ and significance set at $P < 0.05$. Hub genes were identified based on their intramodular connectivity ($kME > 0.8$).

2.5 Discovery of PFOA-Linked Targets

Core targets implicated in PFOA-induced ulcerative colitis (UC) were identified through an intersection analysis that combined differentially expressed genes (DEGs), highly connected genes identified by WGCNA and predicted PFOA targets, with the results illustrated using Venn diagrams.

2.6 Gene Ontology and Pathway Enrichment

Functional enrichment analysis based on Gene Ontology (covering Biological Process, Cellular Component, and Molecular Function) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis were performed using the clusterProfiler package ($P < 0.05$) to investigate the biological processes of PFOA in ulcerative colitis (UC) pathogenesis.

2.7 Leveraging Machine Learning for Core Gene Discovery

Following a systematic approach pinpoint diagnostic biomarkers for perfluorooctanoic acid (PFOA)-associated ulcerative colitis (UC), a novel prediction framework incorporating multiple machine learning algorithms was constructed. Utilizing the gene expression dataset from the training set, ten established machine learning algorithms were employed, including Least Absolute Shrinkage and Selection Operator (LASSO), Support Vector Machine (SVM), Random Forest (RF), Generalized Linear Boosting Model (glmBoost), Stepwise Generalized Linear Model (Stepglm), Ridge Regression (Ridge), Elastic Net (Enet), and Gradient Boosting Machine (GBM). Linear Discriminant Analysis (LDA), eXtreme Gradient Boosting (XGBoost), and Naive Bayes. A total of 127 predictive models were constructed. We optimized hyperparameters using five-fold cross-validation, and data partitioning into training and validation sets was carried out using stratified sampling. A comprehensive evaluation of model performance was carried out based on three criteria: the Area Under the Receiver Operating Characteristic Curve (AUC), accuracy, and F1 score. Ultimately, a stacked ensemble learning strategy was adopted to integrate the predictions from the top-performing individual models. Selection criteria required models to demonstrate high confidence ($AUC > 0.9$), and candidate core genes were identified through frequency ranking of feature genes. The pheatmap package was employed to visualize gene expression patterns.

2.8 Model Interpretation

To address the limited interpretability of machine learning "black box" models, the SHAP (SHapley Additive exPlanations) method was utilized to evaluate feature-specific contributions to predictive results. By calculating SHAP values for individual features, this technique provides a measurable understanding of how each variable influences the model's output.

2.9 Molecular Docking Simulation

To examine the binding of PFOA to the discovered key genes, computational docking simulations were carried out. The ligand structure in SDF format was obtained from the PubChem database, while the three-dimensional protein structures of the targets were retrieved from the UniProt database. Protein structures underwent preprocessing, including

water removal and hydrogen addition, while the ligand was energy-minimized with the MMFF force field. A docking grid was defined around the predicted active site and tailored to the ligand's geometry and binding mode. Docking was executed in AutoDock Vina, and results were processed and graphically represented with ChimeraX.

3. Results

3.1 Target Protein Discovery for PFOA

The molecular structure of PFOA was acquired from the PubChem database (Fig. 2A). To predict its possible biological targets, we utilized three synergistic databases: ChEMBL (an annotated collection of biologically active compounds), PharmMapper (a ligand-to-pharmacophore reverse screening platform), and SwissTargetPrediction (a ligand-based target identification system). Following data integration and elimination of redundant entries, 345 unique potential targets were ultimately identified (Fig. 2B).

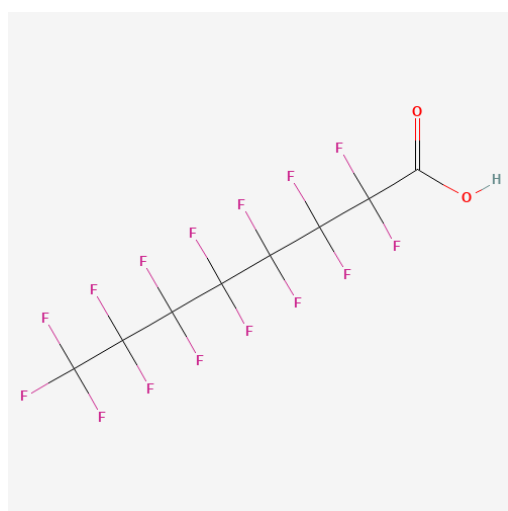


Figure 2A

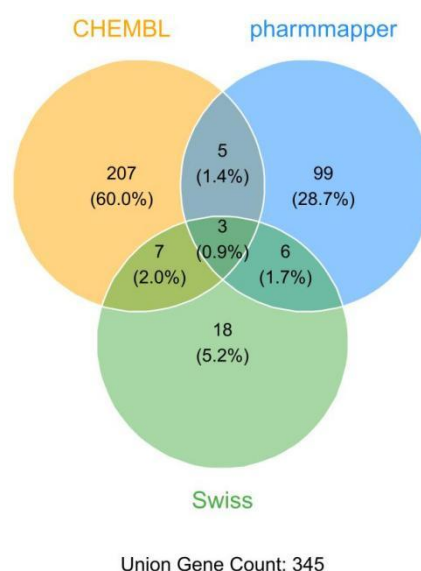


Figure 2B

Figure 2. Discovery of PFOA Target Proteins

(A) PFOA compound architecture.

(B) Multi-database target prediction.

3.2 Determination of target genes linked to UC

To mitigate batch effects, the GSE87473 and GGC38713 datasets were merged and underwent systematic normalization of the gene expression matrix. Principal component analysis (PCA) indicated a markedly enhanced distribution of post-normalization data, with distinct clustering patterns emerging in the normalized dataset (Fig. 3A-B). A total of 1,601

genes displaying significant alterations in ulcerative colitis (UC) were identified through differential expression analysis, and their expression profiles were graphically represented using volcano plots and heatmaps (Fig. 3C-D). For weighted gene co-expression network analysis (WGCNA), the minimal soft threshold power (β) enabling scale-free topology was established. The scale-free topology fit was assessed for power values from 1 to 20, leading to the identification of $\beta=5$ as the optimal value achieving $R^2 \geq 0.8$. Employing this value, a gene co-expression matrix (TOM) was generated and hierarchical cluster analysis for identifying co-expression modules, yielding four distinct color-labeled gene modules (Fig. 3E). Module-trait analysis uncovered statistically significant correlations between certain modules and UC ($P < 0.05$, Fig. 3F). Integration of differentially expressed genes from conventional analyses with module-derived genes from WGCNA—after removing duplicates—produced a final compendium of 629 genes associated with UC (Fig. 3G).

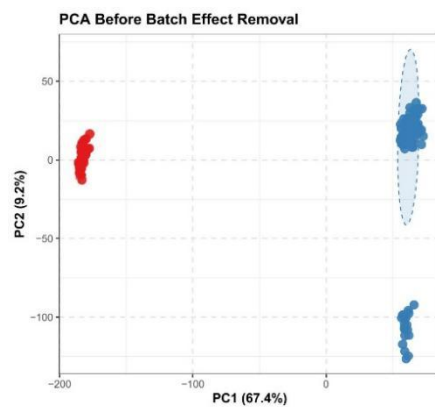


Figure 3A

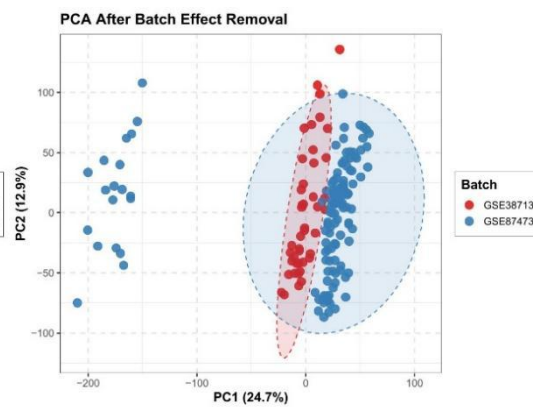


Figure 3B

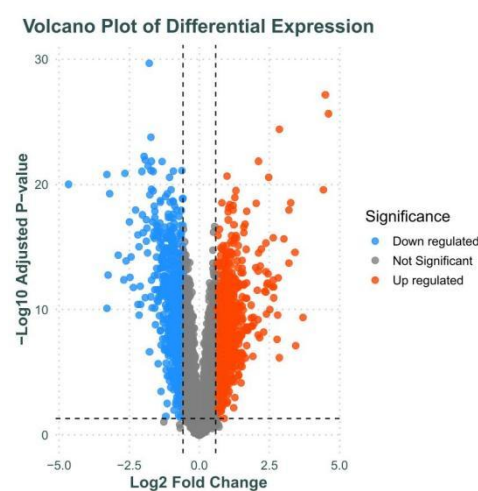


Figure 3C

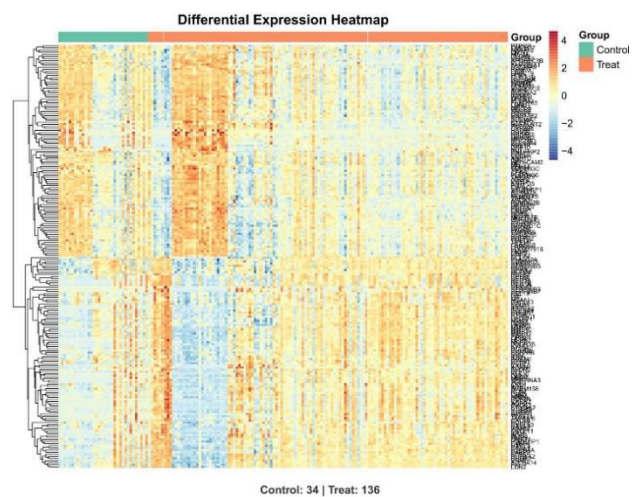


Figure 3D

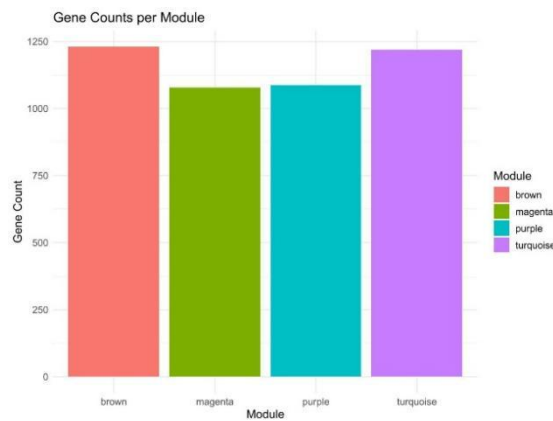


Figure 3E

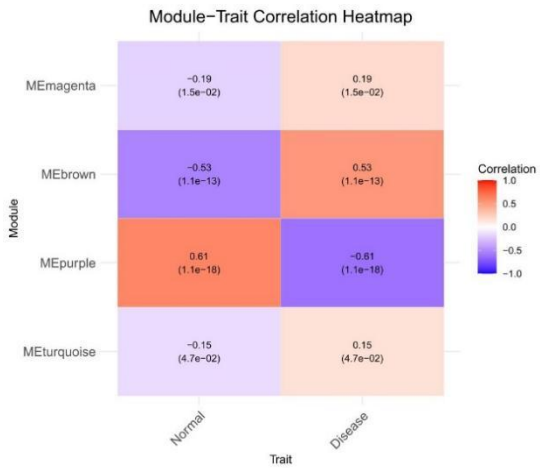


Figure 3F

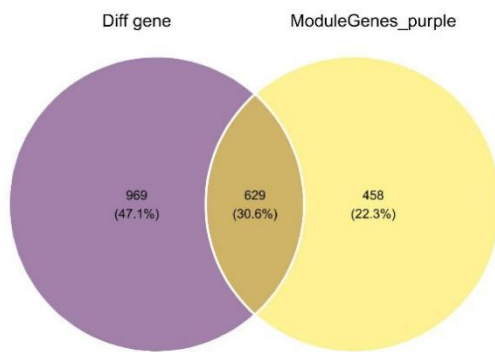


Figure 3G

Figure 3. Identification of UC-Associated Target Genes

- (A) The PCA scatter plot illustrates a clear distinction between the GSE87473 and GSE38713 datasets prior to batch correction, highlighting the existence of batch effects.
- (B) Following batch correction, the PCA scatter plot depicts the integration of the GSE87473 and GSE38713 datasets, signifying a reduction in batch effects.
- (C) The volcano plot displays gene expression changes (logFC and statistical significance), with upregulated (red), downregulated (blue), and non-significant (gray) genes.
- (D) Expression patterns of differentially expressed genes (DEGs) are visualized in the heatmap with red indicating upregulation and blue indicating downregulation.
- (E) Hierarchical clustering of co-expressed genes is displayed in the WGCNA dendrogram,

where differentially colored modules represent distinct gene groups.

(F) Correlations between gene modules identified by WGCNA and sample traits (control and experimental groups) are visualized in the module-trait relationship heatmap, where the numerical values within the boxes reflect correlation coefficients and their respective p-values.

(G) The Venn diagram illustrates the overlap between DEGs (purple) and WGCNA modules (yellow), with the intersecting area representing genes that are common to both methodologies.

3.3 Target Gene Enrichment

Cross-analysis of PFOA target proteins and UC-related genes identified 24 potential key targets involved in the pathogenesis of PFOA-induced ulcerative colitis (Fig. 4A-B). GO and KEGG enrichment analyses (Fig. 4C-D) uncovered comprehensive molecular mechanisms through functional characterization. GO analysis showed significant enrichment in fatty acid metabolic processes, alkene compound metabolic processes (biological processes), microvilli and cell apex (cellular components), as well as transmembrane transport function for carboxylic and organic acids (molecular functions). KEGG pathway analysis highlighted the significant involvement of cofactor and vitamin metabolism, biodegradation and metabolism of exogenous substances, as well as the digestive and endocrine systems.

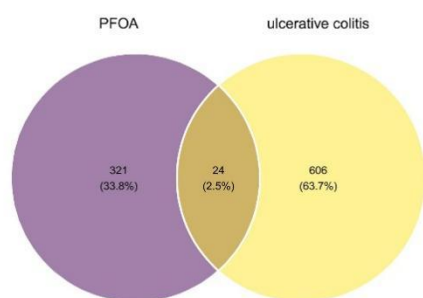


Figure 4A

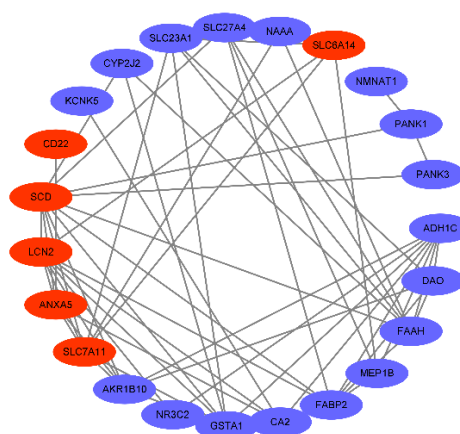


Figure 4B

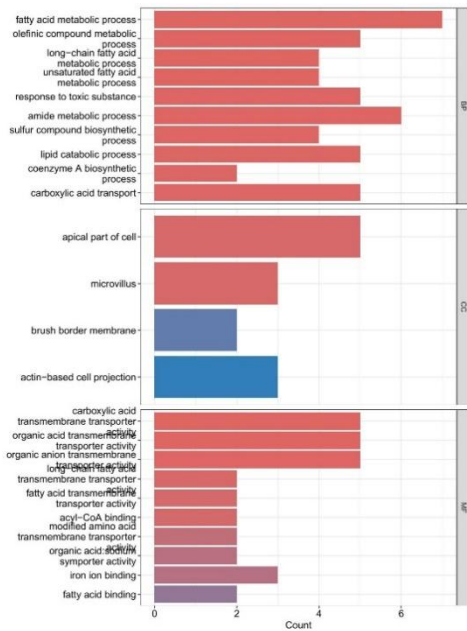


Figure 4C

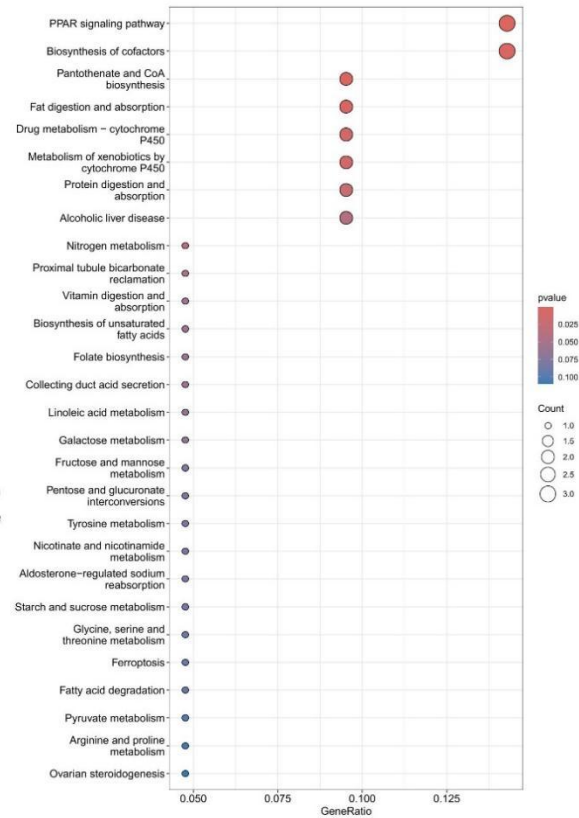


Figure 4D

Figure 4. Screening for PFOA-Related Disease Targets in UC

(A) The Venn diagram juxtaposes genes linked to PFOA exposure (purple) with those associated with UC (ulcerative colitis) (yellow), revealing 24 overlapping genes (2.5%).

(B) The PPI network of overlapping genes. Genes with increased expression are represented by red nodes, and those with decreased expression by blue nodes. Edges indicate predicted functional associations.

(C) GO enrichment analysis of overlapping genes across BP, CC, and MF categories. The bar plot shows gene counts on the X-axis, with the color gradient (red) representing the statistical significance based on adjusted p-values.

(D) The KEGG pathway analysis illustrates pathways enriched by the overlapping genes, with the X-axis showing gene ratios, dot sizes representing the number of genes, and a color gradient denoting adjusted p-values (red indicating heightened significance).

3.4 Validation of Core Genes

We developed 127 machine learning models from 24 candidate genes to identify core genes in

PFOA-induced ulcerative colitis. The Stepglm[both] + Random Forest (RF) combined model exhibited optimal performance during both the training and validation stages (refer to Fig. 5A), ultimately identifying six pivotal genes: MEP1B, NAAA, PANK3, SCD, SLC23A1, and SLC6A14. ROC analysis highlighted the considerable diagnostic potential of the core gene set ($AUC > 0.8$; Fig. 5B). The divergent gene expression profiles observed in UC tissues was illustrated in a volcano plot (Fig. 5C), and their functional roles were further clarified through SHAP interpretability analysis.

The genes SLC6A14 (SHAP value = 0.0644) and MEP1B (SHAP value = 0.0593) were identified as the most significant predictors of the target phenotype (refer to Fig. 5D). A notable trend was observed where high expression levels of all features (represented by orange dots) corresponded with positive SHAP values, indicating a facilitative role in predicting the target phenotype. Conversely, lower expression levels (illustrated by purple dots) typically resulted in negative or near-zero SHAP values, reflecting an inhibitory effect on the prediction of the target phenotype. Specifically, elevated expression of SLC6A14 yielded predominantly positive and closely clustered SHAP values, demonstrating a substantial promotive influence on predictions. In contrast, reduced expression led SHAP values to approach zero or even become negative, resulting in the loss of the promotive effect, thereby highlighting SLC6A14 as a pivotal regulatory gene (see Fig. 5E). The degree to which a gene contributes to the phenotype is contingent upon the expression levels of other genes; for example, the promotive effect of SLC6A14 was amplified in conjunction with high expression of NAAA. Gene interaction pairs, including SLC6A14-NAAA, MEP1B-NAAA, and SLC23A1-MEP1B, are proposed to form essential modules of gene regulation (illustrated in Fig. 5G). Force-directed analysis (depicted in Fig. 5F) further revealed that NAAA was the most significant contributor (+0.0383) and served as the primary factor propelling the prediction upward. Additionally, genes such as MEP1B (+0.0349), SLC6A14 (+0.033), and SLC23A1 (+0.0296) substantially enhanced the predicted value, steering the model's predictions towards a "high-value direction."

Stepglm[both]+RF	0.998	0.924	0.983	0.969
Ridge	0.989	0.932	0.983	0.985
Stepglm[backward]+RF	0.998	0.919	0.983	0.983
Stepglm[both]+RF+NaiveBayes	0.998	0.928	0.967	0.984
RF	1.000	0.910	0.983	0.964
Lasso+RF	0.998	0.906	0.983	0.963
Enet(alpha=0.1)	0.997	0.923	0.967	0.962
Enet(alpha=0.2)	0.994	0.925	0.967	0.962
NaiveBayes	0.975	0.926	0.983	0.962
Enet(alpha=0.3)	0.994	0.924	0.967	0.961
Enet(alpha=0.4)	0.997	0.920	0.967	0.961
Lasso+LDA	0.983	0.917	0.983	0.961
Enet(alpha=0.5)	0.993	0.924	0.967	0.961
Stepglm[both]+GBM	0.998	0.900	0.983	0.96
glmBoost+RF	0.997	0.901	0.983	0.96
glmBoost+LDA	0.978	0.934	0.967	0.96
RF+LDA	0.978	0.934	0.967	0.96
Enet(alpha=0.7)	0.998	0.921	0.967	0.959
Enet(alpha=0.8)	0.987	0.920	0.967	0.958
Stepglm[both]+Enet(alpha=0.8)	0.984	0.922	0.967	0.957
plsRglm	0.984	0.921	0.967	0.957
Stepglm[both]+Ridge	0.998	0.920	0.967	0.957
Stepglm[backward]+Ridge	0.998	0.920	0.967	0.957
Enet(alpha=0.9)	0.980	0.918	0.967	0.957
Stepglm[both]+LDA	0.981	0.906	0.983	0.957
Stepglm[backward]+LDA	0.981	0.906	0.983	0.957
glmBoost+NaiveBayes	0.985	0.919	0.967	0.957
RF+NaiveBayes	0.985	0.919	0.967	0.957
glmBoost+GBM	0.995	0.892	0.983	0.957
Lasso+glmBoost	0.984	0.920	0.967	0.957
glmBoost+GBM+Lasso	0.995	0.891	0.983	0.957
RF+GBM	0.995	0.891	0.983	0.957
Lasso	0.989	0.913	0.967	0.956
glmBoost	0.984	0.919	0.967	0.956
RF+glmBoost	0.984	0.919	0.967	0.956
Stepglm[both]+glmBoost	0.983	0.918	0.967	0.956
Stepglm[backward]+glmBoost	0.983	0.918	0.967	0.956
Lasso+GBM	1.000	0.881	0.983	0.955
Lasso+GBM+RF	0.999	0.880	0.983	0.954
Enet(alpha=0.6)	0.987	0.925	0.950	0.954
GBM	0.999	0.896	0.967	0.954
Stepglm[backward]+GBM	0.998	0.893	0.967	0.954
glmBoost+XGBoost	0.997	0.892	0.967	0.952
Lasso+XGBoost	0.996	0.875	0.983	0.951
Lasso+NaiveBayes	0.979	0.923	0.950	0.951
XGBoost	0.997	0.872	0.983	0.951
LDA	0.991	0.910	0.950	0.95
RF+XGBoost	0.995	0.873	0.983	0.95
glmBoost+Ridge	0.986	0.925	0.933	0.948
RF+Ridge	0.986	0.925	0.933	0.948
Stepglm[backward]+XGBoost	1.000	0.857	0.983	0.947
Stepglm[both]+Enet(alpha=0.5)	0.998	0.892	0.950	0.947
Lasso+plsRglm	0.992	0.896	0.950	0.946
Stepglm[both]+NaiveBayes	0.971	0.927	0.933	0.944
Stepglm[backward]+NaiveBayes	0.971	0.927	0.933	0.944
Lasso+Stepglm[forward]	0.993	0.898	0.960	0.943
glmBoost+plsRglm	0.988	0.921	0.917	0.941
RF+plsRglm	0.988	0.921	0.917	0.941
RF+Enet(alpha=0.1)	0.988	0.918	0.917	0.941
Stepglm[both]+plsRglm	0.980	0.909	0.933	0.941
Stepglm[backward]+plsRglm	0.980	0.909	0.933	0.941
glmBoost+Enet(alpha=0.9)	0.989	0.916	0.917	0.941
RF+Enet(alpha=0.7)	0.989	0.916	0.917	0.941
RF+Enet(alpha=0.9)	0.989	0.916	0.917	0.941
glmBoost+Enet(alpha=0.8)	0.989	0.916	0.917	0.941
RF+Lasso	0.989	0.916	0.917	0.941
RF+Enet(alpha=0.5)	0.988	0.917	0.917	0.941
glmBoost+Enet(alpha=0.1)	0.989	0.916	0.917	0.941
glmBoost+Enet(alpha=0.5)	0.989	0.916	0.917	0.941
RF+Enet(alpha=0.6)	0.989	0.916	0.917	0.941
glmBoost+Enet(alpha=0.4)	0.989	0.916	0.917	0.941
glmBoost+Lasso	0.989	0.916	0.917	0.941
RF+Enet(alpha=0.6)	0.989	0.916	0.917	0.941
glmBoost+Enet(alpha=0.3)	0.988	0.916	0.917	0.94
glmBoost+Enet(alpha=0.2)	0.988	0.916	0.917	0.94
RF+Enet(alpha=0.2)	0.988	0.916	0.917	0.94
RF+Enet(alpha=0.4)	0.988	0.916	0.917	0.94
glmBoost+Enet(alpha=0.7)	0.989	0.916	0.917	0.94
RF+Enet(alpha=0.8)	0.989	0.916	0.917	0.94
Stepglm[both]+Enet(alpha=0.2)+GBM	0.999	0.888	0.933	0.939
RF+Stepglm[both]	0.989	0.915	0.917	0.939
RF+Stepglm[backward]	0.989	0.915	0.917	0.939
glmBoost+Stepglm[both]	0.989	0.915	0.917	0.939
glmBoost+Stepglm[backward]	0.989	0.915	0.917	0.939
Stepglm[both]+Enet(alpha=0.1)	0.999	0.889	0.933	0.939
Stepglm[backward]+Enet(alpha=0.1)	0.999	0.888	0.933	0.939
Stepglm[backward]+Enet(alpha=0.3)	0.999	0.888	0.933	0.939
Stepglm[backward]+Enet(alpha=0.6)	0.999	0.887	0.933	0.939
Stepglm[both]+Enet(alpha=0.4)	0.999	0.887	0.933	0.939
Stepglm[both]+Enet(alpha=0.3)	0.999	0.887	0.933	0.939
Stepglm[backward]+Enet(alpha=0.7)	0.999	0.886	0.933	0.939
Stepglm[backward]+Enet(alpha=0.2)	1.000	0.886	0.933	0.939
Stepglm[backward]+Enet(alpha=0.4)	1.000	0.886	0.933	0.939
Stepglm[both]+Enet(alpha=0.6)+GBM	1.000	0.885	0.933	0.939
Stepglm[backward]+Lasso	0.999	0.886	0.933	0.939
Stepglm[backward]+Enet(alpha=0.2)+XGBoost	1.000	0.885	0.933	0.939
Stepglm[backward]+Enet(alpha=0.8)	1.000	0.885	0.933	0.939
Stepglm[both]+Enet(alpha=0.7)	1.000	0.884	0.933	0.939
Stepglm[backward]+Enet(alpha=0.3)	1.000	0.884	0.933	0.939
Stepglm[both]+Enet(alpha=0.9)	1.000	0.884	0.933	0.939
Stepglm[both]+Lasso+GBM	1.000	0.884	0.933	0.939
glmBoost+Stepglm[forward]	0.987	0.912	0.917	0.939
RF+Stepglm[forward]	0.987	0.912	0.917	0.939
xpglm[backward]+Enet(alpha=0.8)+glmBoost	1.000	0.881	0.933	0.938
Stepglm[both]+Enet(alpha=0.2)	1.000	0.881	0.933	0.938
Stepglm[both]+Lasso	1.000	0.881	0.933	0.938
Stepglm[backward]+Enet(alpha=0.9)	1.000	0.879	0.933	0.937
Lasso+Stepglm[both]	0.993	0.884	0.933	0.937
Lasso+Stepglm[backward]	0.993	0.884	0.933	0.937
Stepglm[both]+Enet(alpha=0.5)	0.993	0.884	0.933	0.937
Stepglm[both]+Enet(alpha=0.7)+GBM	1.000	0.879	0.917	0.932
Stepglm[both]+Enet(alpha=0.6)	1.000	0.879	0.917	0.932
Stepglm[backward]+Enet(alpha=0.5)	1.000	0.880	0.900	0.927
Stepglm[backward]+Lasso+RF	1.000	0.879	0.900	0.926
Stepglm[backward]+Enet(alpha=0.6)+GBM	1.000	0.877	0.900	0.925
Stepglm[both]+Enet(alpha=0.8)+XGBoost	1.000	0.877	0.900	0.926
Stepglm[both]+XGBoost	1.000	0.779	0.983	0.921
Stepglm[forward]	1.000	0.799	0.900	0.92
Lasso+SVM	0.834	0.800	0.900	0.878
Stepglm[both]	1.000	0.811	0.800	0.87
Stepglm[backward]	1.000	0.811	0.800	0.87
glmBoost+SVM	0.919	0.800	0.858	0.860
RF+SVM	0.919	0.800	0.858	0.860
SVM	0.930	0.811	0.800	0.857
Stepglm[both]+SVM	0.980	0.811	0.758	0.843
Stepglm[backward]+SVM	0.960	0.811	0.758	0.843

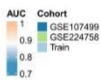
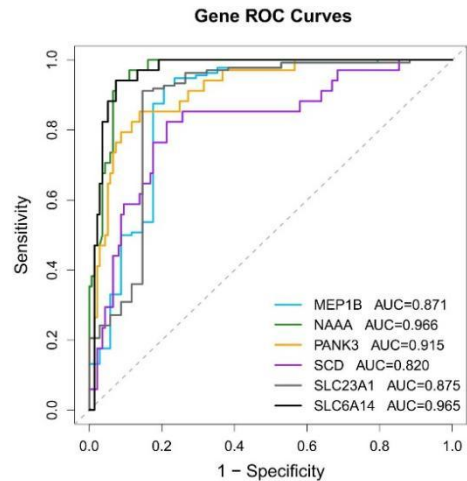


Figure 5A

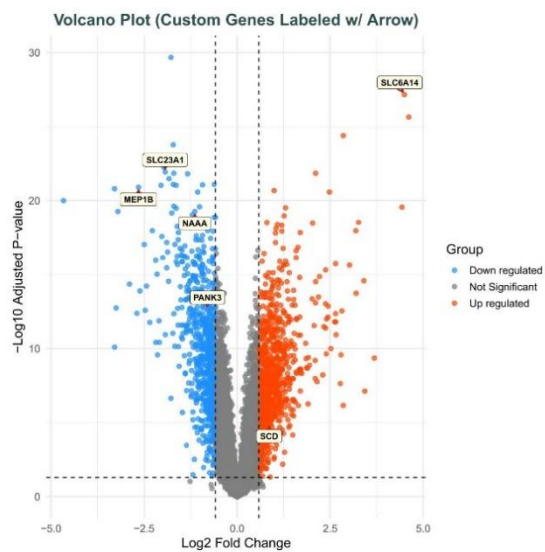


Figure 5B

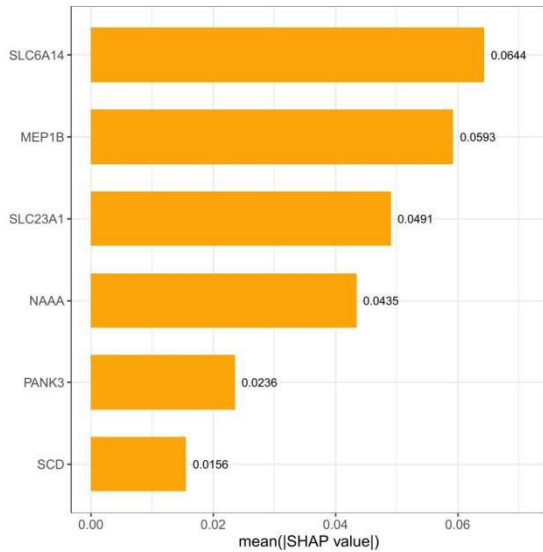


Figure 5C

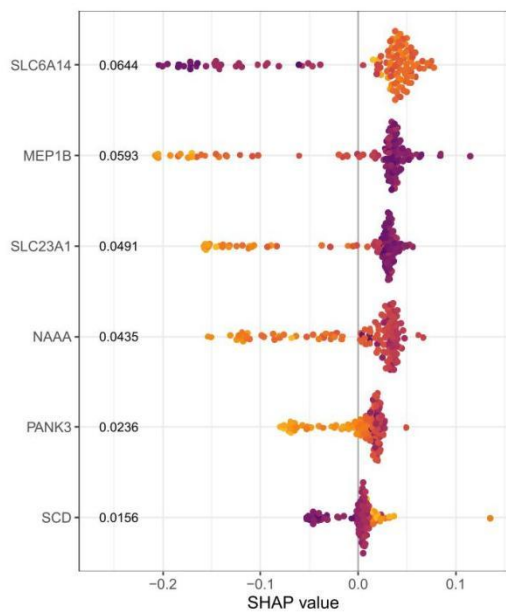


Figure 5D

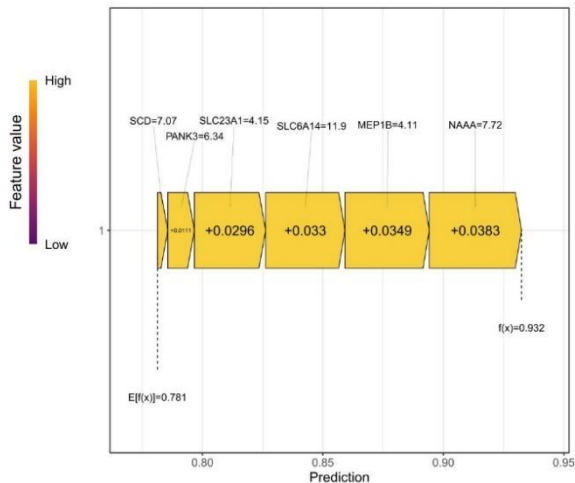


Figure 5E

Figure 5F

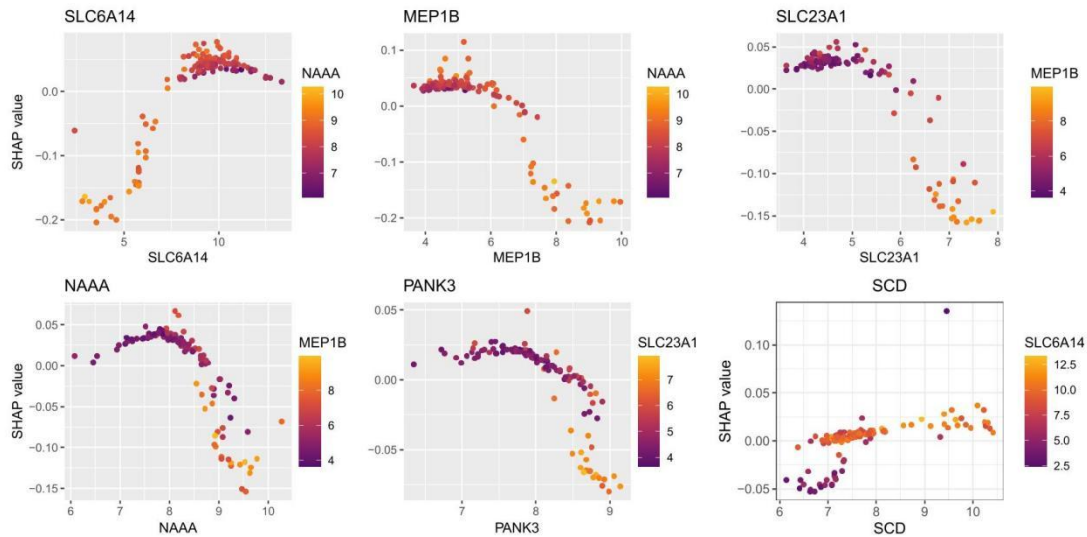


Figure 5G

Figure 5. Discovery of Key Genes in PFOA-Triggered UC Pathogenesis

(A) Comparative Analysis of Model Performance. This heatmap depicts the AUC values of various models across different cohorts. Model names and their corresponding AUC values are listed on the left and right, respectively, with higher values indicating better predictive performance. The color of each cell corresponds to the cohort source.

(B) ROC curves of candidate genes. ROC curves for MEP1B, NAAA, PANK3, SCD, SLC23A1, and SLC6A14, showing the relationship between sensitivity (Y-axis) and false positive rate (X-axis). AUC values are reported as a measure of predictive performance.

(C) Volcano plot of DEGs. Visualization of gene expression changes plotted by $\log_2$ fold change (X-axis) plotted against statistical significance ($-\log_{10}$ p-value; Y-axis). Red and blue points denote significantly up- and down-regulated genes, respectively. Select key genes are annotated.

(D) Feature Importance Ranking: A bar chart is presented to rank core genes based on their feature importance, where longer bars signify greater contributions to the model.

(E) Beeswarm Plot: This plot displays the distributions of gene expression levels under varying conditions, with width representing data density and color indicating expression levels.

(F) SHAP Summary Plot: This summarizes the directional contributions of each gene to the prediction outcome, where negative SHAP values denote reduced prediction risk and positive values signify increased prediction risk.

(G) SHAP Value Distribution: A scatter plot illustrating the SHAP values of key genes reflects

their influence on prediction outcomes, with a color gradient indicating the intensity of the impact of *NAAA* and *MEP1B* expression levels.

3.5 Computational Validation of PFOA Interactions with Core Genes via Docking Studies

To evaluate putative molecular interaction, molecular docking analyses were performed between PFOA and the core genes. PFOA demonstrated excellent binding affinity with every protein target in the study, with binding energies exceeding the -5.0 kcal/mol threshold indicative of strong, spontaneous interactions (Fig.6). Stable docking conformations were observed for all complexes (Fig. 7), corroborating the machine learning predictions by providing structural insights into the direct molecular engagement of PFOA with ulcerative colitis targets.

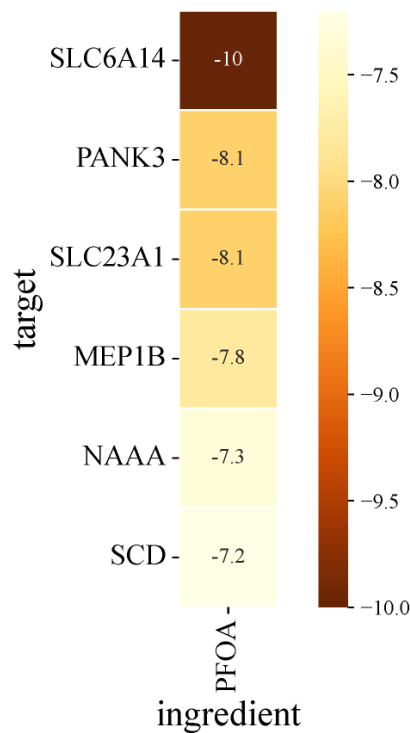


Figure 6. Binding Energy

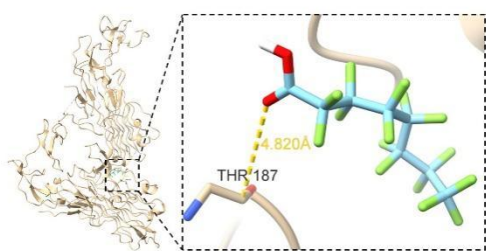


Figure 7A

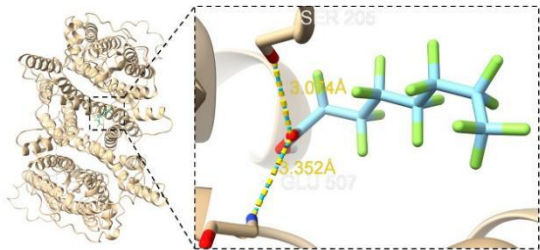


Figure 7B

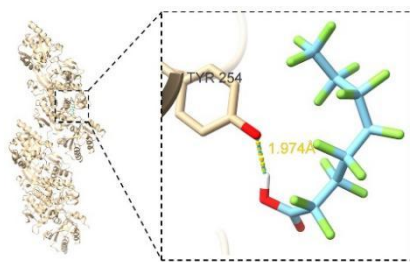


Figure 7C

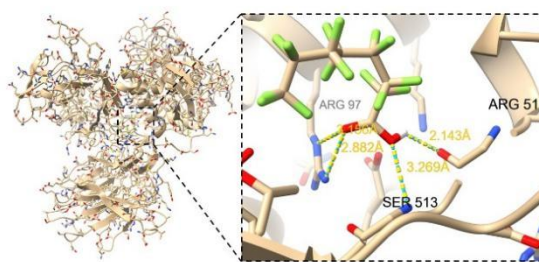


Figure 7D

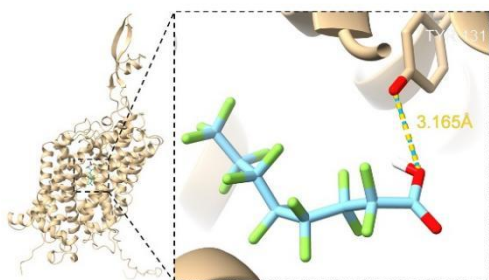


Figure 7E

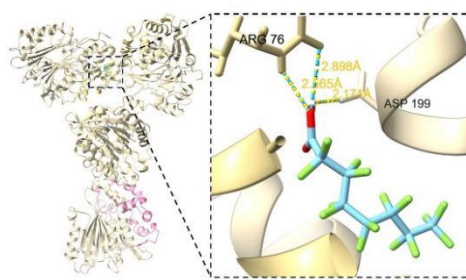


Figure 7F

Fig. 7A presents the docking results of SCD with PFOA.

Fig. 7B illustrates the docking results for SLC23A1 in interaction with PFOA.

Fig. 7C presents the docking findings for PANK3 with PFOA.

Fig. 7D shows the docking outcomes for MEP1B in association with PFOA.

Fig. 7E depicts the docking results of SLC6A14 and PFOA.

Fig. 7F displays the docking analysis for NAAA in relation to PFOA.

4. Discussion

The compromise of the intestinal barrier is recognized as a pivotal mechanism underlying the pathogenesis of inflammatory bowel disease (IBD). Research conducted by Fart et al. has reported significantly increased serum PFAS concentrations, notably PFOS and perfluorooctanoic (PFOA), in individuals diagnosed with late-onset ulcerative colitis (UC) [23]. Additional validation through ex vivo studies utilizing murine tissue provided support for the proposition that PFAS may play a role in the development of IBD: exposure to PFOA at concentrations of 10 and 100 μ M resulted in heightened permeability of the colon and ileum (G) and increased ion transport (Isc) across the colonic barrier upon stimulation with carbachol (CCh). These findings imply that PFOA not only exerts a direct impact on the intestinal barrier but also hinders its capacity for recovery. Moreover, these observations suggest that exposure to PFAS may facilitate the onset of UC by compromising the integrity

of the intestinal barrier or by indirectly modifying its functionality through interference with bile acid metabolism [24].

By integrating multi-omics data with machine learning and bioinformatic approaches, we uncovered pivotal molecular targets mediating perfluorooctanoic acid (PFOA) exposure to ulcerative colitis (UC) pathogenesis. Our network toxicology screening initially yielded 24 candidate genes, from which six core genes—*MEP1B*, *NAAA*, *PANK3*, *SCD*, *SLC23A1*, and *SLC6A14*—were prioritized. Expression profiling showed consistent upregulation of *SLC6A14* and *SCD* and marked downregulation of *MEP1B*, *NAAA*, *PANK3*, and *SLC23A1* in UC cohorts. Machine learning not only demonstrated the strong diagnostic potential of this gene panel but also, through SHAP analysis, highlighted *SLC6A14* (SHAP value = 0.0644) and *MEP1B* (SHAP value = 0.0593) as top predictive contributors. Molecular docking validated stable binding interactions between PFOA and these core targets, mediated by specific residue contacts. These findings collectively point to a mechanism whereby PFOA exacerbates UC by directly modulating these core genes.

SLC transporters regulate ferroptosis and cancer. SLC6A14, a member of the solute carrier family 6, appears to be implicated during UC progression [25,26]. The expression of SLC6A14 has the capacity to amplify the secretion of inflammatory cytokines induced by lipopolysaccharides. Inhibition of SLC6A14 expression has been shown to enhance cellular promote cell growth and suppress programmed cell death in LPS-treated FHC and NCM460 cell lines effectively preserving the integrity of colonic tissue and alleviating morphological alterations induced by DSS. This indicates its therapeutic promise for UC. The underlying mechanism may involve SLC6A14 promoting cellular pyroptosis in UC by modulating the NLRP3 inflammasome [27].

Meprin β (Mep1b), a metalloprotease, is known to cleave adhesion receptors across various tissues (e.g., skin, kidneys, and intestine) and is abnormally expressed in multiple disorders involving impaired barrier function, including inflammatory bowel disease [28]. This dysregulation adversely impacts the structural and functional integrity of the gut mucosal barrier. Metalloproteases meprin A and meprin B are abundantly expressed in intestinal epithelial cells, with their expression levels correlating with distinct stages of colorectal cancer progression. Due to their ability to cleave components of the extracellular matrix (ECM), these proteases are believed to play pro-tumorigenic roles. Furthermore, meprins are capable of cleaving cytokines and their receptors, thereby exerting pro-inflammatory effects that are closely associated with chronic intestinal inflammation. Conversely, meprin β has emerged as a pivotal enzyme involved in the turnover of the intestinal mucus layer, which is

crucial for preventing bacterial overgrowth and subsequent infections. Thus, its role extends to offering protection in preserving intestinal homeostasis[29].

Inflammatory bowel disease (IBD) is characterized by oxidative stress and chronic inflammation, with diminished levels of ascorbic acid (AA) observed in patients afflicted by this condition. TNF- α has been shown to inhibit the uptake of intestinal ascorbate in both in vitro and in vivo models. The suppression is partially mediated by reducing the transcription of *SLC23A1*, a gene responsible for producing the sodium-dependent vitamin C transporter-1 through the NF- κ B signaling pathway [30]. The enzyme NAAA is predominantly expressed in macrophages cells, it hydrolyzes palmitoylethanolamide (PEA), an endogenous lipid recognized for its notable anti-inflammatory and analgesic activities. [31]. The signaling of PEA through NAAA-regulated PPAR- α represents a critical mechanism in both the initiation and resolution of inflammation [32]. As such, inhibition of NAAA activity is posited as a viable therapeutic approach for diseases related to inflammation. Additionally, research has suggested that NAAA may mitigate intestinal fibrosis associated with complications of IBD, with its anti-fibrotic effects mediated by macrophages of the M1 and M2 phenotypes, which indirectly modulate collagen production by fibroblasts [33].

PANK3 is a pivotal enzyme controlling the production of intracellular coenzyme A (CoA). Primarily localized in the cytoplasm, it is likely to be a key factor in cellular metabolism, particularly in the metabolic pathways of cancer cells [34]. Furthermore, PANK3 may be utilized as a genetic marker for the early detection of colorectal cancer (CRC) [35]. Studies indicate that PANK3 is linked to *Akkermansia muciniphila* within the context of fecal microbiota transplantation therapy, suggesting its potential role in modulating gut microbiota composition and the expression of relevant genes through TLR4, thereby alleviating colitis and supporting intestinal homeostasis [36].

The SCD gene, known as the stearoyl-CoA desaturase gene, is a fundamental regulator of fatty acid metabolism in humans. The protein product of this gene, the SCD enzyme, plays an integral role in lipid synthesis, energy homeostasis, and cellular function [37].

Fecal calprotectin and Zonulin represent innovative biomarkers for the assessment of PFAS exposure. Calprotectin is recognized as a sensitive marker for evaluating mucosal inflammation, with its concentrations increasing in response to various ulcers, inflammatory conditions, or malignancies [4]. These relatively novel biomarkers for intestinal inflammation

and permeability open new avenues for investigating the pathways through which PFAS exposure impacts gastrointestinal function and may also serve as valuable tools for future research aimed at understanding PFAS elimination through the gastrointestinal tract.

Several limitations exist. Firstly, although the sample size was adequate for exploratory analysis, it may restrict the generalizability of the findings, thus posing challenges when extrapolating the results to larger populations. Secondly, despite the normalization processes applied to the datasets, residual batch effects may persist, potentially confounding the study outcomes. Thirdly, the computationally predicted outcomes necessitate further empirical validation to ascertain their biological significance and accuracy. Consequently, future research should aim to expand the cohort size and enhance sample diversity, with a particular emphasis on populations exposed to high levels of PFOA, to gain more profound insights into this specific demographic. The integration of diverse methodologies, including both in cell-based and animal-based studies, will be crucial for corroborating and validating the findings, thereby reinforcing the reliability and applicability of the study's conclusions.

The examination of the toxicity mechanisms associated with PFOA in this research carries considerable implications for public health. Future investigations should focus on the following priorities: 1) exploring the pathways through which PFOA exposure impacts intestinal health, utilizing laboratory studies and innovative biomarkers; 2) developing comprehensive risk assessment frameworks for PFOA exposure to formulate effective chemical prevention strategies.

References:

- [1]Du, Lillian, and Christina Ha. "Epidemiology and pathogenesis of ulcerative colitis." *Gastroenterology Clinics* 49.4 (2020): 643-654.
- [2]He YM, Mao R, Yuan G, et al. The hospitalization burden of inflammatory bowel disease in China: a nationwide study from 2013 to 2018[J]. *Therap Adv Gastroenterol*, 2022 (15) : 17562848221102307.
- [3]Tao Y, Li M, Gao H, et al. The hospitalization burden of inflammatory bowel disease in a southwestern highland region of China: a territory-wide study from 2015 to 2020[J]. *Front Med (Lausanne)*, 2024 (11) :
- [4]Xu, Yiyi, et al. "Inflammatory bowel disease and biomarkers of gut inflammation and permeability in a community with high exposure to perfluoroalkyl substances through drinking water." *Environmental research* 181 (2020): 108923.

- [5]Steenland, K., Zhao, L., Winqvist, A., Parks, C., 2013. Ulcerative colitis and perfluorooctanoic acid (pfoa) in a highly exposed population of community residents and workers in the mid Ohio valley. *Environ. Health Perspect.* 121, 900.
- [6]Steenland, K., Zhao, L., Winqvist, A., 2015. A cohort incidence study of workers exposed to perfluorooctanoic acid (pfoa). *Occup. Environ. Med.* 72, 373–380.
- [7]Eriksson, U., Kärrman, A., Rotander, A., Mikkelsen, B., Dam, M., 2013. Perfluoroalkyl substances (pfass) in food and water from Faroe Islands. *Environ. Sci. Pollut. Res. Int.* 20, 7940–7948.
- [8]Guelfo, J.L., Adamson, D.T., 2018. Evaluation of a national data set for insights into sources, composition, and concentrations of per-and polyfluoroalkyl substances
- [9] Hu, X.C., Andrews, D.Q., Lindstrom, A.B., Bruton, T.A., Schaider, L.A., Grandjean, P., et al., 2016. Detection of poly-and perfluoroalkyl substances (pfass) in us drinking water linked to industrial sites, military fire training areas, and wastewater treatment plants. *Environ. Sci. Technol. Lett.* 3, 344–350.
- [10]Domingo, J.L., 2012. Health risks of dietary exposure to perfluorinated compounds. *Environ. Int.* 40, 187–195.
- [11]European Food Safety Authority, 2018. Risk to Human Health Related to the Presence of Perfluorooctane Sulfonic Acid and Perfluorooctanoic Acid in Food. *EFSA Journal*. vol. 16, e05194.
- [12] HE Y L, CHI X D. Supramolecular macrocycles-based functional materials for the detection and adsorption of PFASs pollutant in water[J]. *Scientia sinica (chimica)*, 2023, 53(12): 2494-2508.
- [13] Research progress on environmental pollution and toxicity of PFOA and PFOS alternatives[J]. *Asian journal of ecotoxicology*, 2024, 19(5): 112-134.
- [14]Shi, Wenshan, et al. "Reproductive toxicity of PFOA, PFOS and their substitutes: A review based on epidemiological and toxicological evidence." *Environmental Research* 250 (2024): 118485.
- [15]Gou, **ao, et al. "A novel molecular pathway of lipid accumulation in human hepatocytes caused by PFOA and PFOS." *Environment international* 191 (2024): 108962.
- [16]Do PFCAs drive the establishment of thyroid cancer microenvironment? Effects of C6O4, PFOA and PFHxA exposure in two models of human thyroid cells in primary culture
- [17]Bartell, Scott M., and Verónica M. Vieira. "Critical review on PFOA, kidney cancer, and testicular cancer." *Journal of the Air & Waste Management Association* 71.6 (2021): 663-679.

- [18]DeWitt, Jamie C., et al. "Immunotoxicity of perfluorinated compounds: recent developments." *Toxicologic pathology* 40.2 (2012): 300-311.
- [19]Mannon, Peter, and Walter Reinisch. "Republished: Interleukin 13 and its role in gut defence and inflammation." *Postgraduate medical journal* 89.1054 (2013): 448-456.
- [20]Qazi, Mousumi R., et al. "High-dose, short-term exposure of mice to perfluorooctanesulfonate (PFOS) or perfluorooctanoate (PFOA) affects the number of circulating neutrophils differently, but enhances the inflammatory responses of macrophages to lipopolysaccharide (LPS) in a similar fashion." *Toxicology* 262.3 (2009): 207-214.
- [21]Liang, Luyun, et al. "Immunotoxicity mechanisms of perfluorinated compounds PFOA and PFOS." *Chemosphere* 291 (2022): 132892.
- [22]Han, **, et al. "The toxicological effects of perfluorooctanoic acid (PFOA) exposure in large yellow croaker (*Larimichthys crocea*): exploring the relationship between liver damage and gut microbiota dysbiosis." *Environmental Research* (2025): 121683.
- [23]Fart, Frida, et al. "Perfluoroalkyl substances are increased in patients with late-onset ulcerative colitis and induce intestinal barrier defects ex vivo in murine intestinal tissue." *Scandinavian journal of gastroenterology* 56.11 (2021): 1286-1295.
- [24]Fart, Frida, et al. "Perfluoroalkyl substances are increased in patients with late-onset ulcerative colitis and induce intestinal barrier defects ex vivo in murine intestinal tissue." *Scandinavian journal of gastroenterology* 56.11 (2021): 1286-1295.
- [25] Chen, Yanjun, et al. "SLC6A14 facilitates epithelial cell ferroptosis via the C/EBP β -PAK6 axis in ulcerative colitis." *Cellular and Molecular Life Sciences* 79.11 (2022): 563.
- [26] Lu, Jia Jie, et al. "Potential biomarkers associated with differential manifestations of ulcerative colitis (UC) in Uyghur and Han population in China." *Journal of Inflammation Research* (2021): 7431-7441.
- [27]Gu, Qing, et al. "SLC6A14 promotes ulcerative colitis progression by facilitating NLRP3 inflammasome-mediated pyroptosis." *World Journal of Gastroenterology* 30.3 (2024): 252.
- [28]Gindorf, Markus, et al. "Meprin β : a novel regulator of blood–brain barrier integrity." *Journal of Cerebral Blood Flow & Metabolism* 41.1 (2021): 31-44.
- [29]Peters, Florian, and Christoph Becker-Pauly. "Role of meprin metalloproteases in metastasis and tumor microenvironment." *Cancer and Metastasis Reviews* 38.3 (2019): 347-356.
- [30] Subramanian, Veedamali S., et al. "Tumor necrosis factor alpha reduces intestinal vitamin C uptake: a role for NF- κ B-mediated signaling." *American Journal of Physiology-*

Gastrointestinal and Liver Physiology 315.2 (2018): G241-G248.

[31]**, Wenhui, et al. "Anti-inflammatory effects of fucoxanthinol in LPS-induced RAW264.7 cells through the NAAA-PEA pathway." *Marine Drugs* 18.4 (2020): 222.

[32]Piomelli, Daniele, et al. "N-acylethanolamine acid amidase (NAAA): structure, function, and inhibition." *Journal of Medicinal Chemistry* 63.14 (2020): 7475-7490.

[33]Nani, Maria Francesca, et al. "Pharmacological inhibition of N-Acylethanolamine acid amidase (NAAA) mitigates intestinal fibrosis through modulation of macrophage activity." *Journal of Crohn's and Colitis* 19.2 (2025): jjæ132.

[34]Khadka, Sunada, et al. "Generation and validation of an anti-human PANK3 mouse monoclonal antibody." *Biomolecules* 12.9 (2022): 1323.

[35] Jiao, Meng, et al. "Potential key genes are expected to become biomarker for early diagnosis of colorectal cancer through bioinformatics analysis." *Biotechnology and Genetic Engineering Reviews* 40.1 (2024): 678-691.

[36]Wen, **n, et al. "Fecal microbiota transplantation alleviates experimental colitis through the toll-like receptor 4 signaling pathway." *World Journal of Gastroenterology* 29.30 (2023): 4657.

[37]Lin, Yiwei, et al. "Scd-1 deficiency promotes the differentiation of CD8+ T effector." *Frontiers in Cellular and Infection Microbiology* 14 (2024): 1325390.