frontiers | Frontiers in Pharmacology

TYPE Original Research
PUBLISHED 14 April 2025
DOI 10.3389/fphar.2025.1545392

OPEN ACCESS

EDITED BY
Shaoqiu Chen,
University of Hawaii at Mānoa, United States

REVIEWED BY
Xiaoyan Wang,
Shanghai Jiao Tong University, China
Rui Wang,
The First Affiliated Hospital of Xi'an Jiaotong University, China

*CORRESPONDENCE
Tao Huang,
24867509@qq.com

†These authors have contributed equally to this work

RECEIVED 14 December 2024
ACCEPTED 14 January 2025
PUBLISHED 14 April 2025

CITATION
Li L, Deng X, Wang S and Huang T (2025) Integrating traditional omics and machine learning approaches to identify microbial biomarkers and therapeutic targets in pediatric inflammatory bowel disease.
Front. Pharmacol. 16:1545392.
doi: 10.3389/fphar.2025.1545392

Integrating traditional omics and machine learning approaches to identify microbial biomarkers and therapeutic targets in pediatric inflammatory bowel disease

Lanlan Li[1†], XuZai Deng[1†], Shuge Wang[1†] and Tao Huang[2*]

[1]Department of Pediatrics, Tianyou Hospital Affiliated to Wuhan University of Science and Technology, Wuhan, China, [2]Department of Pediatrics, Maternal and Child Health Hospital of Hubei Province, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

Background: Pediatric inflammatory bowel disease (IBD), especially Crohn's disease, significantly affects gut health and quality of life. Although gut microbiome research has advanced, identifying reliable biomarkers remains difficult due to microbial complexity.

Methods: We used RNA-seq-based microbial profiling and machine learning (ML) to find robust biomarkers in pediatric IBD. Microbial taxa were profiled at phylum, genus, and species levels using kraken2 on Crohn's disease and non-IBD ileal biopsies. We performed abundance-based analyses and applied four ML models (Logistic Regression, Random Forest, Support Vector Machine, XGBoost) to detect discriminative taxa. An independent cohort of 36 pediatric stool samples assessed by 16S rRNA sequencing validated top ML results.

Results: Traditional abundance-based methods showed compositional shifts but identified few consistently significant taxa. ML models had better discriminatory performance, with XGBoost outperforming others and pinpointing Orthotospovirus and Vescimonas as key genera. These findings were confirmed in the validation cohort, where only one traditionally noted genus, *Actinomyces*, maintained significance.

Discussion: Integrating conventional omics with AI-driven analytics boosts reproducibility and clinical relevance of microbial biomarker discovery, opening new possibilities for targeted therapies and precision medicine in pediatric IBD.

KEYWORDS

pediatric IBD, biomarkers, RNA-Seq, machine learning, therapeutic targets, reproducibility

Introduction

Pediatric inflammatory bowel disease (IBD), encompassing Crohn's disease (CD) and ulcerative colitis (UC), remains a formidable clinical and research challenge due to its complex etiology, variable presentation, and substantial impact on growth and development. Unlike adult-onset IBD, pediatric IBD can have more extensive intestinal involvement and a higher disease burden, often manifesting as delayed puberty, impaired

growth, and reduced quality of life. Early and accurate diagnosis, alongside effective therapeutic interventions, is crucial for mitigating long-term complications and improving patient outcomes. However, the current diagnostic paradigm—primarily reliant on invasive endoscopic examinations and histopathological assessments—offers limited non-invasive biomarkers capable of reliably differentiating pediatric IBD from other gastrointestinal conditions.

Advancements in omics technologies, particularly RNA-sequencing (RNA-seq), have substantially enhanced our understanding of the gut microbiome's role in health and disease (Beaudry et al., 2016; Lloyd-Price et al., 2016). By capturing the functional and taxonomic composition of microbial communities, RNA-seq has facilitated the identification of candidate biomarkers and putative drug targets (Franzosa et al., 2014). Yet, while such abundance-based analyses have highlighted certain taxa, reproducibility and consistency across independent cohorts remain vexing issues. Many proposed microbial biomarkers fail to display stable, cross-study validation, hampering their clinical applicability and limiting insights into potential therapeutic mechanisms (Haberman et al., 2014; Shapiro et al., 2018).

Although advancements in omics-based technologies have yielded valuable insights into the gut microbiome's contribution to IBD, traditional abundance-based approaches often fail to identify reproducible biomarkers, resulting in a "reproducibility crisis." This reproducibility crisis underscores the need for integrative strategies that extend beyond conventional abundance-based approaches. Artificial intelligence (AI) and machine learning (ML) offer a powerful complement to traditional methods, capable of discerning subtle patterns and complex interactions within large, multi-dimensional datasets (Topol, 2019; Beam and Kohane, 2018). ML models, such as Random Forest, Support Vector Machines (SVM), and gradient-boosting algorithms like XGBoost, can sift through vast numbers of features—including microbial taxa at various taxonomic levels—and prioritize the most informative markers with high predictive value. Applying similar strategies to pediatric IBD could enhance early diagnosis, uncover novel therapeutic targets, and inform personalized interventions (Esteva et al., 2019).

The gut microbiome's complexity in pediatric IBD is further complicated by age-related factors. Children's microbiotas are dynamic, influenced by diet, early-life exposures, and ongoing maturation of the immune system (Yatsunenko et al., 2012). Such complexity demands robust analytical tools that can integrate biological knowledge with computational efficiency. Traditional omics approaches offer depth and mechanistic understanding, while ML provides scalability, pattern recognition, and improved predictive performance when confronted with heterogeneous and noisy data (Franzosa et al., 2019; Laukens et al., 2016).

Moreover, identifying biomarkers that translate into actionable drug targets requires a comprehensive approach that moves beyond static abundance measures. RNA-seq data can reveal microbial gene expression patterns, shedding light on metabolic pathways and potential therapeutic mechanisms (Paramsothy et al., 2017). By focusing on microbial taxa consistently linked to pediatric IBD, researchers may pinpoint targets for microbiome-modulating therapies—such as probiotics, prebiotics, fecal microbiota transplantation (FMT), or even metabolite-targeted interventions—that hold promise in complementing or enhancing existing pharmacological treatments (Sainath et al., 2020; Sinha et al., 2017).

Building on previous efforts to integrate computational and experimental methods in microbiome research, this study demonstrates the potential of combining RNA-seq-based microbial profiling with machine learning (ML) to enhance biomarker reproducibility and identify promising therapeutic targets in pediatric IBD. In this study, we combined traditional RNA-seq microbial profiling with ML-driven approaches to systematically identify reliable microbial signatures of pediatric IBD. By contrasting abundance-based statistical methods with multiple ML classifiers—including Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost—we aimed to address the synergistic value of integrating complementary methodologies. Ultimately, this approach seeks to enhance reproducibility in biomarker discovery and inform precision medicine strategies for pediatric IBD.

Materials and methods

Study cohorts and sample collection

We analyzed ileal biopsy RNA-seq data obtained from pediatric patients diagnosed with Crohn's disease (CD) and age-matched non-IBD controls. The primary dataset comprised 245 pediatric ileal biopsy samples originally described in publicly available repositories (GSE93624) (Schloss et al., 2011). In addition to this primary dataset, an independent validation cohort was assembled comprising 36 pediatric stool samples (19 IBD, 17 non-IBD), collected prospectively and processed using 16S rRNA gene sequencing to test the reproducibility of identified microbial biomarkers (Caporaso et al., 2012). Our protocols to include: (i) standardized fasting requirements (at least 8 h prior to biopsy) when clinically permissible, (ii) avoidance or documentation of probiotics and specific medications within 4 weeks prior to sampling, and (iii) adherence to uniform dietary guidelines where possible. Non-IBD controls included children undergoing ileal biopsy for reasons unrelated to IBD who exhibited normal histopathology and no signs of inflammatory disorders (Supplementary Table S1). Samples were immediately flash-frozen and stored at −80°C to preserve microbial integrity. Stool collection also followed standardized protocols, including instructions for home collection kits, immediate cooling, and rapid transfer to the lab, where samples were stored at −80°C before 16S rRNA gene sequencing. Patients were enrolled following informed consent and in accordance with the ethical guidelines and approval of the Tianyou Hospital Affiliated to Wuhan University of Science and Technology Review Board.

RNA-seq data processing and microbial profiling

Raw paired-end RNA-seq reads were quality-checked using FastQC and filtered to remove low-quality reads and adapter contamination with Trimmomatic (Bolger et al., 2014), applying default parameters for Illumina data. Host reads were removed by mapping against the human reference genome (GRCh38) using Hisat2 (Kim et al., 2015), retaining only non-host reads for downstream microbiome profiling. Microbial sequences were

taxonomically classified at phylum, genus, and species levels using kraken2 (Wood and Salzberg, 2014) with a comprehensive reference database, ensuring robust identification of bacterial, viral, and fungal taxa. To generate a normalized microbial abundance table, we calculated the relative abundance of each taxon by dividing raw counts by the total number of microbial reads per sample. Only taxa present in at least 10% of samples were retained to minimize the influence of rare and potentially spurious features.

Alpha and beta diversity analyses

To assess within-sample microbial diversity (alpha diversity), we computed the Shannon index using vegan R package functions (Oksanen 2020). Between-sample compositional differences (beta diversity) were evaluated using the Bray-Curtis dissimilarity metric, followed by Principal Coordinates Analysis (PCoA) for visualization. Group comparisons were performed using PERMANOVA (adonis function in vegan) to test for significant shifts in community structure (Anderson, 2017).

Traditional abundance-based comparisons

To identify differentially abundant taxa between pediatric CD and non-IBD controls using conventional approaches, we employed non-parametric tests (Wilcoxon rank-sum) on relative abundance data, adjusting for multiple comparisons with the Benjamini-Hochberg method (Benjamini and Hochberg, 1995). Taxa displaying a false discovery rate (FDR)-adjusted p-value <0.05 were considered statistically significant. Selected taxa were visualized with boxplots and stacked barplots to depict compositional differences at phylum and genus levels (Wickham, 2016).

Machine learning approaches for biomarker discovery

To enhance biomarker discovery, we applied four ML algorithms: Logistic Regression, Random Forest, SVM, and XGBoost. We performed hyperparameter tuning via grid search and 5-fold cross-validation for each model. For example, our SVM tested multiple kernel types (linear, polynomial, radial basis), while our XGBoost pipeline varied learning rates (0.01–0.3), max_depth (Franzosa et al., 2014; Haberman et al., 2014; Shapiro et al., 2018; Topol, 2019; Beam and Kohane, 2018; Esteva et al., 2019; Yatsunenko et al., 2012; Franzosa et al., 2019), and regularization parameters (lambda, alpha). Model performance was evaluated on a 70/30 train/test split, with Area Under the Receiver Operating Characteristic Curve (AUC), accuracy, sensitivity, and specificity as key metrics (Chen and Guestrin, 2016). Feature selection was conducted using a forward selection strategy, gradually adding taxa that improved classification performance until no further gain was observed. Models were evaluated on the testing set, and their performance was compared based on AUC, accuracy, sensitivity, and specificity (Kuhn, 2008). The best-performing model was then applied to identify the most informative taxa for distinguishing pediatric CD from non-IBD controls (Supplementary Table S2).

Independent validation and reproducibility testing

To validate the identified biomarkers, we analyzed the independent pediatric cohort of 36 stool samples processed via 16S rRNA amplicon sequencing. This cohort underwent similar quality control and taxonomic assignment steps using QIIME2 and DADA2 (Callahan et al., 2016), ensuring consistent methods for feature representation. The taxa highlighted by ML-based approaches were examined in this validation cohort to assess reproducibility. Taxa derived from traditional abundance-based methods were also tested, facilitating a direct comparison of stability and clinical relevance (Segata et al., 2011).

Statistical analyses and software

All statistical analyses were performed in R (version 4.2.2). Multiple packages, including phyloseq for microbiome data handling (McMurdie and Holmes, 2013), vegan for diversity analyses (Oksanen 2020), and caret for machine learning workflows (Kuhn and Johnson, 2013), were employed. Visualization was completed using python3 for differential abundance results. Unless otherwise stated, p-values were two-tailed, and $p < 0.05$ was considered significant.

Results

Microbial community composition at phylum and genus levels

Compositional profiling of pediatric Crohn's disease and non-IBD ileal biopsy samples revealed distinct microbiome signatures at both the phylum and genus levels (Figure 1). While the overall phylum distribution appeared dominated by a few core taxa in both groups, subtle shifts were evident, suggesting alterations in microbial community structures associated with disease status. At the genus level, stacked barplots indicated that certain taxa were enriched in either pediatric IBD or non-IBD controls, yet the magnitude and consistency of these differences varied across samples.

Alpha and beta diversity analyses at the genus level

Alpha diversity metrics, such as the Shannon index, showed no significant differences between pediatric IBD and non-IBD groups (Figure 2A). Both cohorts exhibited comparable within-sample microbial complexity, implying that reduced richness or evenness may not be a defining feature of pediatric IBD in this dataset. Conversely, beta diversity analyses using Bray-Curtis distances suggested mild compositional shifts (Figure 2B). Principal

10.3389/fphar.2025.1545392

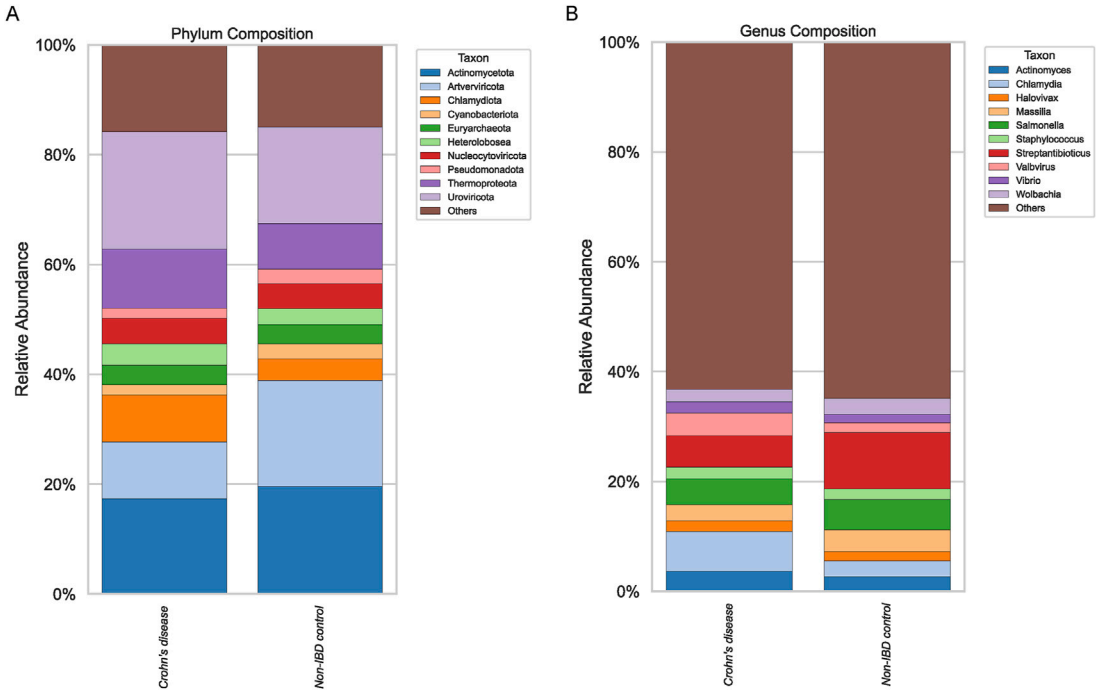

FIGURE 1
Phylum and Genus Composition. Stacked barplots comparing the relative abundance at phylum level **(A)** and genus level **(B)** between Crohn's disease and non-IBD control groups.

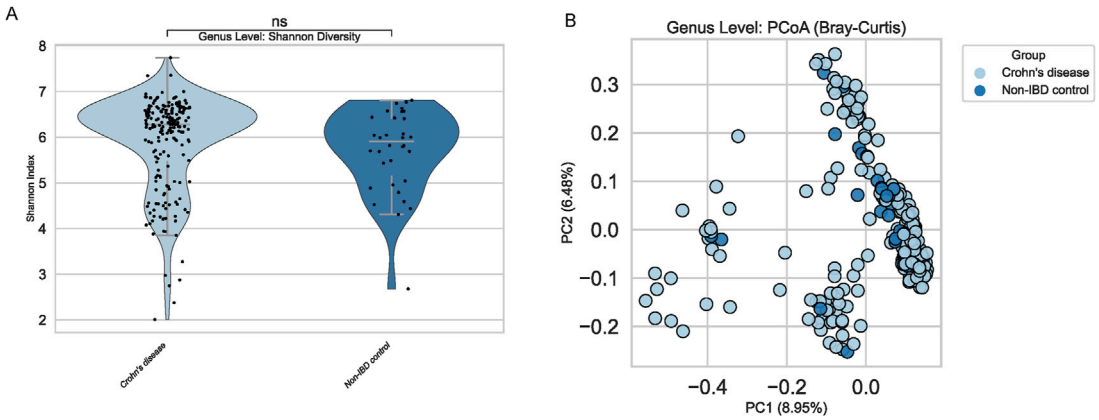

FIGURE 2
Alpha and Beta Diversity at Genus Level. **(A)** Violin plot of Shannon index shows no significant differences between groups. **(B)** PCoA plot (Bray-Curtis) indicates no clear clustering by disease status.

Coordinates Analysis (PCoA) showed a trend of clustering by disease status, albeit with notable overlap, highlighting the subtlety of the microbial distinctions.

Identifying taxa with traditional abundance-based methods

Traditional abundance-based comparisons identified a shortlist of genera that appeared differentially abundant between pediatric IBD and non-IBD samples. Among these, *Actinomyces* and Streptantibioticus emerged as candidates of interest (Figure 3). Boxplots of these genera revealed that *Actinomyces* displayed a more consistent pattern of enrichment in pediatric IBD, whereas Streptantibioticus showed differences but lacked robust statistical significance or consistency. These findings underscored the challenges inherent in relying solely on abundance-based methods, as initial signals may not always translate into stable biomarkers.

10.3389/fphar.2025.1545392

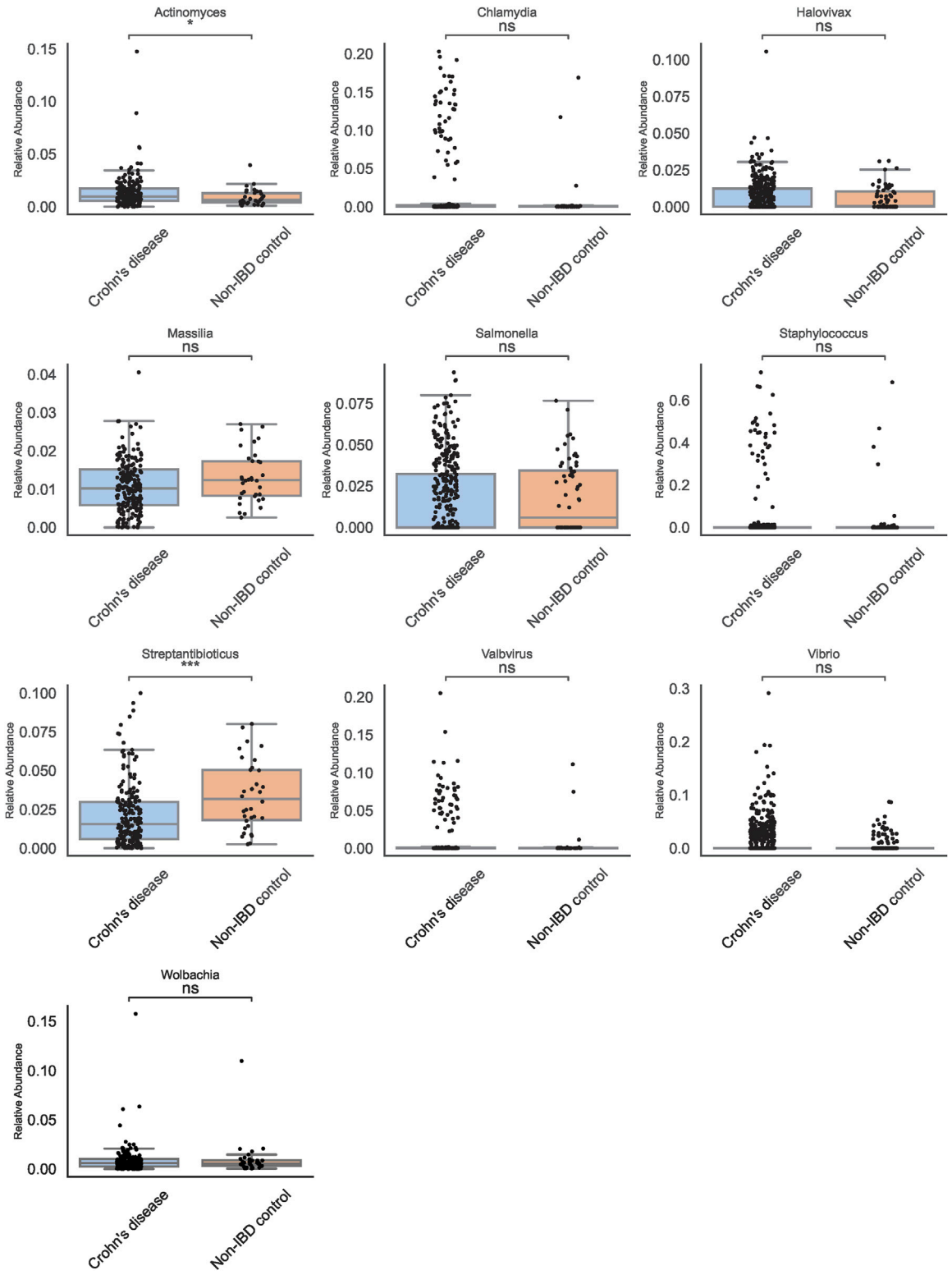

FIGURE 3
Genus-Level Boxplots from Traditional Abundance Analyses. Boxplots for selected genera reveal *Actinomyces* and Streptantibioticus as differentially abundant, but only *Actinomyces* is significant.

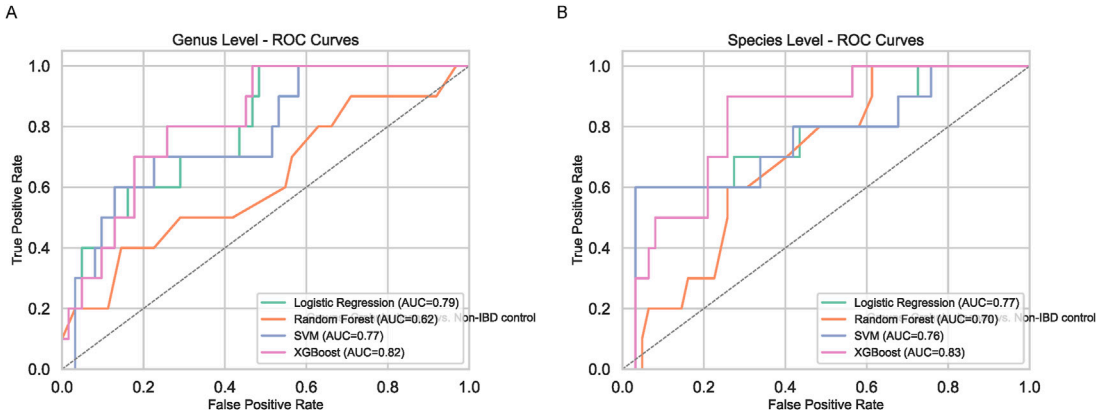

FIGURE 4
Machine Learning-Based Classification. ROC curves for Logistic Regression, Random Forest, SVM, and XGBoost at genus **(A)** and species **(B)** levels. XGBoost outperforms others.

Machine learning-based biomarker discovery

To enhance biomarker reproducibility, multiple machine learning (ML) classifiers were employed at both genus and species levels (Figure 4). While Logistic Regression, Random Forest, and SVM provided modest improvements over traditional methods, XGBoost outperformed all other models, demonstrating higher accuracy and a more pronounced ability to discriminate pediatric IBD from non-IBD controls. Feature importance analyses identified Orthotospovirus and Vescimonas at the genus level as key discriminators (Figure 5). Unlike many taxa highlighted by conventional abundance approaches, these XGBoost-selected genera exhibited robust and reproducible differences, suggesting they may serve as reliable biomarkers.

Independent validation in a separate pediatric cohort

To test the reproducibility and clinical relevance of the identified biomarkers, we examined an independent cohort of 36 pediatric stool samples using 16S rRNA sequencing (Figure 6). Orthotospovirus and Vescimonas maintained consistent trends in this external dataset, supporting their status as stable biomarkers. Notably, among the taxa initially suggested by abundance-based methods, only *Actinomyces* retained its observed pattern. This independent validation emphasized the value of combining traditional and ML-driven approaches to achieve reproducible and clinically meaningful results.

Discussion

This study demonstrates how integrating traditional omics analyses with AI-driven machine learning can refine the search for robust microbial biomarkers and potential therapeutic targets in pediatric IBD. While traditional abundance-based comparisons provided an initial shortlist of candidate taxa, their reproducibility proved limited. *Actinomyces* stood out as the sole traditional candidate maintaining consistent patterns across datasets, underscoring the rarity of stable biomarkers emerging from conventional methods alone. Our findings underscore XGBoost's advantages for this dataset, including its capacity to handle imbalanced classes, accommodate complex feature interactions, and integrate regularization to prevent overfitting. These attributes appear especially beneficial in microbiome studies where underlying microbial signals may be subtle or masked by high inter-individual variability. Identifying Orthotospovirus and Vescimonas as key discriminators speaks to the capacity of ML tools to capture complex interactions that might be overlooked by simpler statistical approaches. By systematically incorporating feature selection, cross-validation, and performance metrics, the ML framework not only improved classification accuracy but also enhanced the likelihood that identified taxa reflect underlying disease mechanisms rather than spurious associations.

The independent validation further corroborated the strength of ML-driven discoveries. While one genus from the traditional approach retained its trend, the consistent appearance of Orthotospovirus and Vescimonas across cohorts underscores their potential involvement in pediatric IBD pathogenesis. Orthotospovirus, traditionally known as a plant-infecting viral genus, may have unrecognized roles in human health through complex interactions within the gut ecosystem, while Vescimonas, a relatively understudied bacterial genus, could influence immune regulation or mucosal barrier integrity (De Oliveira et al., 2018; Huang et al., 2024). Such biomarkers can serve as entry points for investigating how shifts in the microbial community affect mucosal inflammation and immune modulation, ultimately guiding microbiome-targeted therapies—whether through dietary modifications, microbial transplantation, or metabolite-based interventions. By bridging traditional abundance-based analyses with cutting-edge computational tools, this integrated strategy refines biomarker discovery and underlines the importance of validating candidate biomarkers in independent cohorts. Harnessing the combined strengths of established biological

10.3389/fphar.2025.1545392

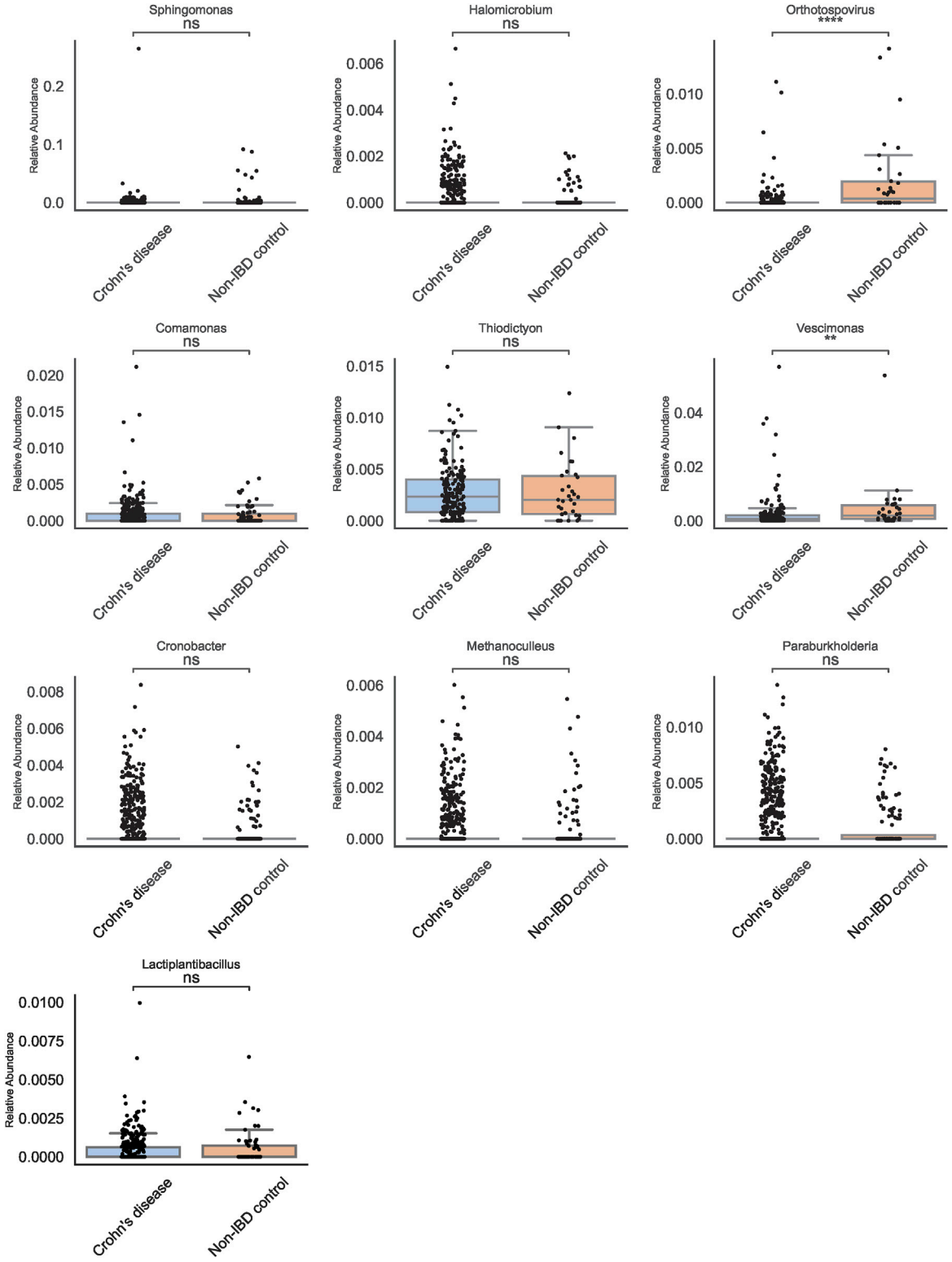

FIGURE 5
Key Genera Identified by XGBoost. Boxplots show Orthotospovirus (****) and Vescimonas (**) as top discriminators. These taxa exhibit robust differences between Crohn's disease and controls.

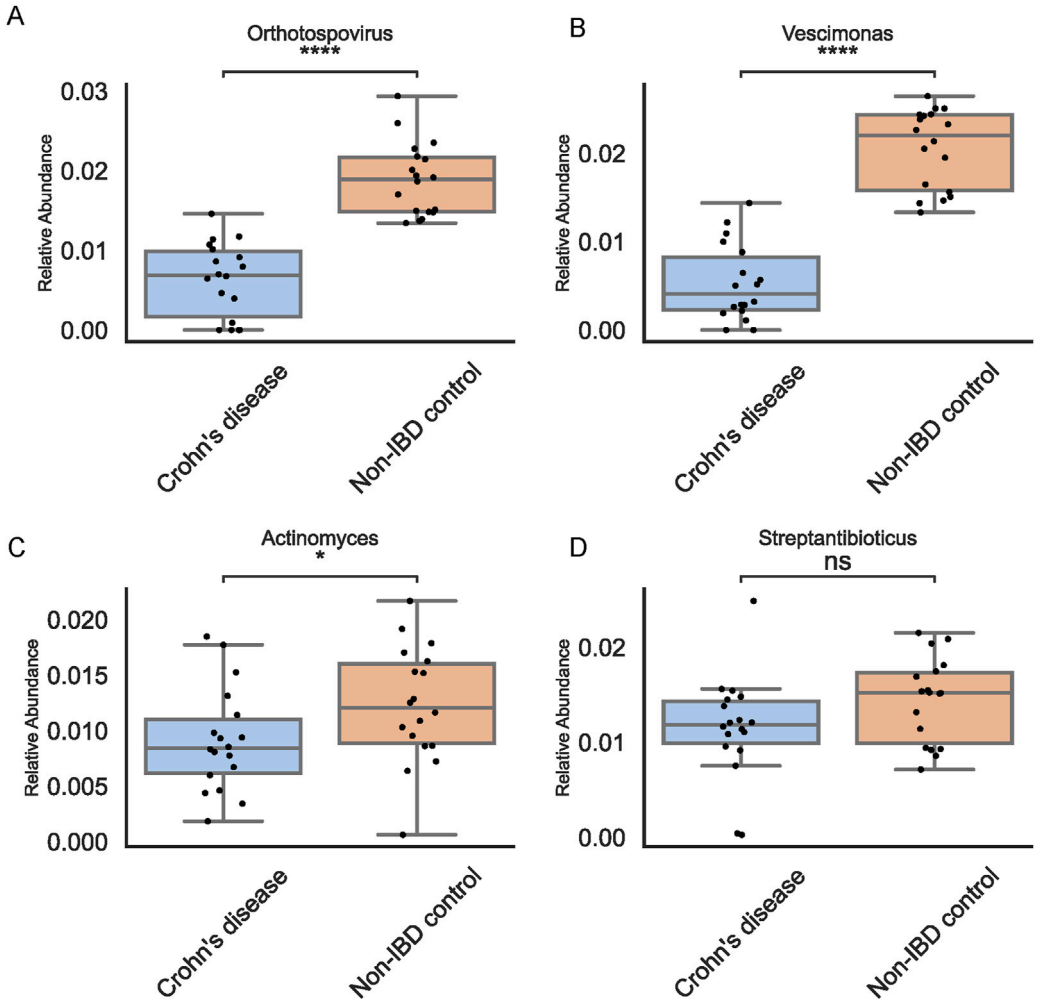

FIGURE 6
Independent Validation in 36 Pediatric Fecal Samples. Boxplots for top genera in the validation set confirm consistent trends for Orthotospovirus **(A)**, Vescimonas **(B)**, *Actinomyces* **(C)**, and Streptantibioticus **(D)** highlighting reproducibility of ML-based findings and indicating potential therapeutic targets.

methods and advanced machine learning offers a clearer path toward reproducible targets, promising improved early diagnosis, personalized treatments, and better outcomes for pediatric IBD patients.

As we continue to refine these computational methods, future work may delve deeper into functional characterizations of the identified taxa, examine temporal dynamics of the pediatric microbiome, and assess the efficacy of targeted interventions informed by these biomarkers. By doing so, we move closer to an era where precision medicine—fueled by robust microbial biomarkers and guided by integrative analytic frameworks—can transform the management of pediatric IBD.

Data availability statement

RAW data available at the link: https://ngdc.cncb.ac.cn/gsa-human/browse/HRA010328. According to local legal requirements and the need to protect minors, access to the raw data can be obtained through a request to the China National Information Center. Request link: https://ngdc.cncb.ac.cn/gsa-human/request/HRA010328.

Ethics statement

The studies involving humans were approved by the Tianyou Hospital Affiliated to Wuhan University of Science and Technology. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

LL: Conceptualization, Formal Analysis, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing–original draft, Writing–review and editing. XD: Data curation, Investigation, Methodology, Software, Writing–original

draft, Writing–review and editing. SW: Investigation, Methodology, Software, Writing–original draft, Writing–review and editing. TH: Conceptualization, Data curation, Funding acquisition, Methodology, Supervision, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This study was supported by Youth Talent Project of Hubei Provincial Department of Education (Q20201103).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2025.1545392/full#supplementary-material

References

Anderson, M. J. (2017). *Permutational multivariate Analysis of variance (PERMANOVA)*. Wiley StatsRef: Statistics Reference Online, 1–15. doi:10.1002/9781118445112.stat07841

Beam, A. L., and Kohane, I. S. (2018). Big data and machine learning in health care. *JAMA* 319, 1317–1318. doi:10.1001/jama.2017.18391

Beaudry, P., Campbell, M., Dang, N. H., Wen, J., Blote, K., and Weljie, A. M. (2016). A pilot study on the utility of serum metabolomics in neuroblastoma patients and xenograft models. *Pediatr. Blood Cancer* 63, 214–220. doi:10.1002/pbc.25784

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* 57, 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi:10.1093/bioinformatics/btu170

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Johnson, A. J. A., and Holmes, S. P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13, 581–583. doi:10.1038/nmeth.3869

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Huntley, J., Fierer, N., et al. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* 6, 1621–1624. doi:10.1038/ismej.2012.8

Chen, T., and Guestrin, C. (2016). "XGBoost: a scalable tree boosting system," in *Proc 22nd ACM SIGKDD int conf knowl discov data min*, 785–794. doi:10.1145/2939672.2939785

De Oliveira, A. S., Boiteux, L. S., Kormelink, R., and Resende, R. O. (2018). The Sw-5 gene cluster: tomato breeding and research toward orthotospovirus disease control. *Front. Plant Sci.* 9, 1055. doi:10.3389/fpls.2018.01055

Derikx, L., Lantinga, M. A., de Jong, D. J., et al. (2020). Clinical outcomes of fecal microbiota transplantation for chronic radiation enteritis: a prospective cohort study. *Am. J. Gastroenterol.* 115, 1411–1422. doi:10.14309/ajg.0000000000000657

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., et al. (2019). A guide to deep learning in healthcare. *Nat. Med.* 25, 24–29. doi:10.1038/s41591-018-0316-z

Franzosa, E. A., Morgan, X. C., Segata, N., Waldron, L., Reyes, J., Earl, A. M., et al. (2014). Relating the metatranscriptome and metagenome of the human gut. *Proc. Natl. Acad. Sci. U. S. A.* 111, E2329–E2338. doi:10.1073/pnas.1319284111

Franzosa, E. A., Sirota-Madi, A., Avila-Pacheco, J., Fornelos, N., Haiser, H. J., Reinker, S., et al. (2019). Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat. Microbiol.* 4, 293–305. doi:10.1038/s41564-018-0306-4

Haberman, Y., Timothy, L. T., Phillip, J. D., Mi-Ok, D., Dora, T., Rebekah, K., Robert, N. B., et al. (2014). Pediatric Crohn disease patients exhibit specific ileal transcriptome and microbiome signature. *J. Clin. Invest* 124, 3617–3633. doi:10.1172/JCI75436

Huang, H. J., Liu, C., Sun, X. W., Wei, R. Q., Liu, L. W., Chen, H. Y., et al. (2024). The rheumatoid arthritis gut microbial biobank reveals core microbial species that associate and effect on host inflammation and autoimmune responses. *Imeta* 3 (5), e242. doi:10.1002/imt2.242

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. doi:10.1038/nmeth.3317

Kuhn, M. (2008). Building predictive models in R using the caret package. *J. Stat. Softw.* 28, 1–26. doi:10.18637/jss.v028.i05

Kuhn, M., and Johnson, K. (2013). *Applied predictive modeling*. Springer. doi:10.1007/978-1-4614-6849-3

Laukens, D., Brinkman, B. M., Raes, J., De Vos, M., and Vandenabeele, P. (2016). Heterogeneity of the gut microbiome in inflammation and cancer: fueling disease progression. *Nat. Rev. Microbiol.* 14, 373–384. doi:10.1038/nrmicro.2016.63

Lloyd-Price, J., Abu-Ali, G., and Huttenhower, C. (2016). The healthy human microbiome. *Genome Med.* 8, 51. doi:10.1186/s13073-016-0307-y

McMurdie, P. J., and Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* 8, e61217. doi:10.1371/journal.pone.0061217

Oksanen, J. (2020). Vegan: community ecology package. Available at: http://vegan.r-forge.r-project. org/.

Paramsothy, S., Paramsothy, R., Rubin, D. T., Kamm, M. A., Kaakoush, N. O., Mitchell, H. M., et al. (2017). Faecal microbiota transplantation for inflammatory bowel disease: a systematic review and meta-analysis. *J. Crohns Colitis* 11, 1180–1199. doi:10.1093/ecco-jcc/jjx063

Sainath, K. G., Arun, V., and Virendra, S. (2020). Serum transferrin is an independent predictor of mortality in severe alcoholic hepatitis: upping the game or just upping the ante?. *Am. J. Gastroenterol.* 115, (7)1136. doi:10.14309/ajg.0000000000000657

Schloss, P. D., Gevers, D., and Westcott, S. L. (2011). Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS One* 6, e27310. doi:10.1371/journal.pone.0027310

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., et al. (2011). Metagenomic biomarker discovery and explanation. *Genome Biol.* 12, R60. doi:10.1186/gb-2011-12-6-r60

Shapiro, H., Kolodziejczyk, A. A., Halstuch, D., and Elinav, E. (2018). Bacteria and beyond: microbiome research in human health and disease. *Curr. Opin. Microbiol.* 43, 77–82. doi:10.1016/j.mib.2017.11.002

Sinha, R., Abu-Ali, G., Vogtmann, E., Fodor, A. A., Ren, B., Amir, A., et al. (2017). Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium. *Nat. Biotechnol.* 35, 1077–1086. doi:10.1038/nbt.3981

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* 25, 44–56. doi:10.1038/s41591-018-0300-7

Wickham, H. (2016). *ggplot2: elegant graphics for data Analysis*. New York: Springer-Verlag. doi:10.1007/978-0-387-98141-3

Wood, D. E., and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 15, R46. doi:10.1186/gb-2014-15-3-r46

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., et al. (2012). Human gut microbiome viewed across age and geography. *Nature* 486, 222–227. doi:10.1038/nature11053