frontiers | Research Topics

Untangle the broad connections and tight interactions between human microbiota and complex diseases through data-driven approaches

Edited by
Qi Zhao, Jian Li, Li Zhang and Liang Wang

Published in
Frontiers in Microbiology

ISSN 1664-8714
ISBN 978-2-83251-791-8
DOI 10.3389/978-2-83251-791-8

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Untangle the broad connections and tight interactions between human microbiota and complex diseases through data-driven approaches

Topic editors

Qi Zhao — University of Science and Technology Liaoning, China
Jian Li — Tulane University, United States
Li Zhang — University of New South Wales, Australia
Liang Wang — Guangdong Provincial People's Hospital, China

Topic coordinator

Zuobin Zhu — Xuzhou Medical University, China

Citation

Zhao, Q., Li, J., Zhang, L., Wang, L., eds. (2023). *Untangle the broad connections and tight interactions between human microbiota and complex diseases through data-driven approaches*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-83251-791-8

Table of contents

frontiers | Frontiers in Microbiology

TYPE Editorial
PUBLISHED 15 February 2023
DOI 10.3389/fmicb.2023.1157579

OPEN ACCESS

EDITED AND REVIEWED BY
Matthias Hess,
University of California, Davis, United States

*CORRESPONDENCE
Qi Zhao
✉ zhaoqi@lnu.edu.cn
Liang Wang
✉ wangliang@gdph.org.cn;
✉ liang.wang@ecu.edu.au

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

RECEIVED 02 February 2023
ACCEPTED 06 February 2023
PUBLISHED 15 February 2023

CITATION
Wang L, Zhu Z-B, Zhang L, Li J and Zhao Q (2023) Editorial: Untangle the broad connections and tight interactions between human microbiota and complex diseases through data-driven approaches. *Front. Microbiol.* 14:1157579. doi: 10.3389/fmicb.2023.1157579

Editorial: Untangle the broad connections and tight interactions between human microbiota and complex diseases through data-driven approaches

Liang Wang[1,2,3*†], Zuo-Bin Zhu[4†], Li Zhang[5], Jian Li[6] and Qi Zhao[7*]

[1]Laboratory Medicine, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou, Guangdong, China, [2]School of Medical and Health Sciences, Edith Cowan University, Perth, WA, Australia, [3]School of Medical Informatics and Engineering, Xuzhou Medical University, Xuzhou, Jiangsu, China, [4]Department of Genetics, School of Life Sciences, Xuzhou Medical University, Xuzhou, Jiangsu, China, [5]School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW, Australia, [6]School of Public Health and Tropical Medicine, Tulane University, New Orleans, LA, United States, [7]School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China

KEYWORDS

microbiome, human diseases, metagenomics, metaomics, computational tools

Editorial on the Research Topic
Untangle the broad connections and tight interactions between human microbiota and complex diseases through data-driven approaches

It is well-known that microorganisms are ubiquitous in the environment and occupy almost all habitats in animals and humans (Finlay and Clarke, 1999; Rosenberg, 2021). Traditionally, microorganisms are studied as individuals grown in isolation under artificial conditions; however, with the development of experimental techniques and computational methods, microbes are now frequently considered as a functional group in a particular niche and studied at the community level in order to best mimicking the real-world situations (American Academy of Microbiology, 2004). During the study of microbial communities, two terms are commonly used, that is, microbiota and microbiome. A microbiota is defined as the microorganisms present in a defined environment and consists of bacteria, fungi, viruses, archaea, protists, etc. (Malard et al., 2021), while a microbiome not only means the collection of genomes from all the microorganisms in a niche, but also includes the microbial structural elements, metabolites, and the environmental conditions (Berg et al., 2020). For the past two decades, both sequencing technologies and mass spectrometry techniques have been developing rapidly. With their in-depth applications in the dissection of the human microbiota, more and more evidence have shown that microorganisms play very important roles in physiological functions and are closely related to various complex diseases in human beings (Hou et al., 2022). This has led to an insightful understanding of underlying disease mechanisms from microbial perspectives. Therefore, elucidation of the microbiota-disease association will be of great help for understanding the pathogenesis of human diseases, promoting early diagnosis, and improving precision medicine.

Particularly, in the human gut, the vast majority of gut microbes not only synthesize essential amino acids and vitamins but also facilitate the digestion of indigestible

components of the human diet like plant polysaccharides (Rowland et al., 2017; Vernocchi et al., 2020). When gut microbial communities change, people are likely to suffer from related digestive system diseases, but when the abnormal gut microbial communities are restored to normality, disease symptoms could be alleviated (Gagliardi et al., 2018; Liu et al., 2021), though safety issues are still under intensive investigations (Daliri et al., 2018). If changes in intestinal microbes can be detected in time and given corresponding treatment, the workload of later clinical diagnosis and treatment will be greatly reduced (Zhang et al., 2015; Manor et al., 2020). Although current technologies have already helped us identify many previously unexpected connections between the microbiota and diseases, such as cancer, autoinflammatory diseases, metabolic syndromes, digestive system diseases, cardiovascular diseases, and central nervous system disorders, the present level of knowledge is still limited (Zhang et al., 2015). It is rather difficult to analyze the existing meta-omics data in-depth due to the lack of competent algorithms and bioinformatics tools, which leads to a narrow understanding of the microbiome-disease association and severely limits the development of the association mechanisms (Wang et al., 2022). Therefore, more efforts should be applied to the microbiota-disease association analysis, especially to promote the application of microbial analysis in the clinical settings for the diagnosis, treatment, and prevention of complex human diseases. In addition, downstream experimental validations of the microbiome discoveries in terms of the associations between microbiota and diseases are urgently needed to promote the real-world application of the meta-omics analysis. Therefore, studies with the combination of experimental and computational methods for interrogating the intriguing associations are also desirable.

In this Research Topic, all the collected articles could be divided into several groups that are either directly related to the topic by focusing on the interactions between microbiota and diseases or indirectly linked with the topic by focusing on models, tools, biomarkers and diseases, which are all summarized below to emphasize the core of the collection, that is, the broad connections and tight interactions between human microbiota and complex diseases. In specificity, Qin et al. compared the gut microbiome in 28 healthy people and 61 lung cancer patients that were classified into three types according to their histopathology, that is, Atypical Adenomatous Hyperplasia/Adenocarcinoma *in situ* (AAH/AIS), Minimally Invasive Adenocarcinoma (MIA), and Invasive Adenocarcinoma (IA). According to the results, categorized cancer patients had unique intestinal flora characteristics with comparatively lower density and flora diversity than healthy people. In addition, several flora markers were identified for the development of lung cancer, which held the potential for diagnosis, prognosis, prevention and treatment of lung cancer. Yang L. et al. from Sichuan University performed an integrative analysis of gut microbiota and fecal metabolites by comparing 32 metabolic associated fatty liver disease (MALFD) patients and 30 healthy individuals; according to the results, decreased species richness and diversity and altered β-diversity in feces were found in MALFD patients *via* 16S rRNA amplicon sequencing data, while metabolomic analysis identified overall changes in fecal and serum metabolites dominated by lipid molecules. Further associations between gut microbiota and fecal metabolites revealed that LPC 18:0 was positively correlated with *Christensenellaceae*_R-7_group, *Oscillospiraceae*_UCG-002 while neohesperidin was positively correlated with *Peptoniphilus, Phycicoccus,* and *Stomatobaculum*, which provided novel clues for understanding the molecular mechanisms of MALFD, and its diagnostic markers and therapeutic strategies. In addition, Wei et al. developed and validated an interpretable radiomic nomogram for severe radiation proctitis prediction in postoperative cervical cancer patients because radiation proctitis is a complex disease closely related to the microbiota but it is really time-consuming and expensive for analysis of gut microbiota. Therefore, this study emphasized the limitations of microbiota study and proposed a solution to solve the issue.

Except for the direct analysis of human microbiota and complex diseases, several studies focused on the constructions of models and development of tools for disease studies. For example, Yang B. et al. focused on the disease-ligand identification in the system of traditional Chinese medicine (TCM) based on a newly developed screening method termed as flexible neural tree (FNT) model, which were successfully applied to hypertension, diabetes, and COVID-19 for the identification of related compounds in TCM. It is also well-studied that hypertension, diabetes, and COVID-19 are closely associated with gut microbiota (Gurung et al., 2020; Mishima and Abe, 2021; Zhang F. et al., 2022). Therefore, the disease-ligand identification model has potential application in dissection the association between human microbiota and human diseases. In another study, Wang et al. tried to infer pan-cancer associated genes by examining the microbial model organism *Saccharomyces cerevisiae* by homology matching, which was based on the principle that the homologous genes of the common ancestor may have similarities in expression. According to the authors, their study holds the potential in revealing a link between microbiota and associated diseases, which is crucial to understand the molecular mechanisms of these diseases in the development of new microbiome-based therapies. In addition, Chen and Lei noticed that limitations of traditional medical experiments in the study of potential microbe-disease associations. Therefore, they proposed a method based on heterogeneous network and metapath aggregated graph neural network (MAGNN) to predict microbe-disease associations, which is termed as MATHNMDA. According to the results, their model could effectively predict microbe-disease associations in terms of case studies of asthma, inflammatory bowel disease, and COVID-19. In another study, Niu et al. studied an industrial yeast *Pichia pastoris* from the aspect of transcriptomic analysis, which aims to identify the regulation of foreign proteins with different stabilities expressed in *Pichia pastoris*. According to their results, the study shed a new light on the understanding of the regulatory mechanisms in yeast cells that responds to intracellular folding stress.

COVID-19 that is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has been widely spread worldwide since the end of 2019 (Liu et al., 2022; Zhang Y.-D. et al., 2022), which generated huge social and economic impact on human beings. Since COVID-19 infection is tightly associated with human microbiota, several studies also contributed to its diagnosis and therapy in this Research Topic collection. For example, Peng et al. developed a novel diagnostic analysis for CT scan images of COVID-19 pneumonia based on a deep ensemble framework

with DenseNet, Swin transformer, and RegNet, which achieved the best precision of 0.9833, recall of 0.9895, accuracy of 0.9894, F1-score of 0.9864, AUC of 0.9991, and AUPR of 0.9986 under binary classification problem by comparing with other classification methods. Moreover, Tian et al. constructed a deep ensemble learning-based automated detection of COVID-19 using lung CT images and Vision Transformer and ConvNeXt, which computed the best precision of 0.9668, an accuracy of 0.9696, and an F1-score of 0.9631 in the three-classification experiment. In addition, Chen et al. built a novel weighted reconstruction-based linear label propagation (WLLP) algorithm for predicting potential therapeutic agents for COVID-19, which exhibited excellent performance with an AUC of 0.8828 $\pm$ 0.0037 and an area under the precision-recall curve of 0.5277 $\pm$ 0.0053, showing that the algorithm could be used to suggest potential drugs for the treatment of COVID-19.

Finally, there are also two studies that are focusing on microRNAs in human diseases. According to previous studies, it was well-known that the interactions between gut microbiota and microRNA affected host pathophysiology such as intestinal, neurological, cardiovascular, and immune health and diseases (Li et al., 2020). Therefore, it is meaningful to include a couple of microRNAs studies in this Research Topic. In particular, Qu et al. investigated the spring-like effect of microRNA-31 in balancing inflammatory and regenerative responses in colitis, according to which MIR31 is able to alleviate inflammation *via* inhibiting inflammatory cytokine receptors and can promote epithelial regeneration by modulating the WNT and Hippo signaling pathways. In the other study, Yao et al. identified circRNA-miRNA interactions based on multi-biological interaction fusion by proposing a novel model termed as circRNA-miRNA interaction prediction model (IIMCCMA), which showed that the model could achieve excellent performance in predicting the rare interaction between circRNA and miRNA, which helped to understand the molecular mechanism and contributed to the diagnosis, treatment, and prognosis of human diseases. However, whether these microRNAs involve any interactions with gut microbiota require further studies.

Taken together, a total of 12 articles including research papers, methodologies, web server tools, and software were enclosed in this Research Topic, which were authored by 96 investigators from different countries and regions of the whole world. The Research Topic focuses on human diseases such as cancer, pneumoniae, liver disease, colitis, and proctitis mainly from the aspects of human microbiota and relevant factors that could greatly facilitate the understanding of complex diseases in human beings from a long-term perspective. In addition, we would also like to thank all the reviewers for their valuable, rigorous, and high-standard suggestions and comments during the tedious peer review process. We would like to express our sincere gratitude to the Specialty Chief Editor, Dr. George Tsiamis, and also the editorial office of Frontier in Microbiology, for providing us with this opportunity to hold this fascinating Research Topic issue successfully.

Author contributions

LW, Z-BZ, and QZ drafted the manuscript. All authors provided comments and feedbacks during the revision of the manuscript. All authors proposed the Research Topic theme, made a direct and intellectual contribution to the work, and approved the final version of the editorial for publication.

Funding

LW appreciated the financial support provided by Xuzhou Key R&D Plan Social Development Project (Grant No. KC22300, Year 2022) and Jiangsu Qinglan Project (Year 2020). QZ appreciated the financial support provided by Foundation of Education Department of Liaoning Province (Grant No. LJKZ0280).

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.

Publisher's note

References

American Academy of Microbiology (2004). *Microbiology in the 21st Century: Where Are We and Where Are We Going? This Report Is Based on a Colloquium Sponsored by the American Academy of Microbiology held September 5-7, 2003. in Charleston, South Carolina*. Washington, DC: American Academy of Microbiology.

Berg, G., Rybakova, D., Fischer, D., Cernava, T., Vergès, M. C., Charles, T., et al. (2020). Microbiome definition re-visited: old concepts and new challenges. *Microbiome* 8, 103. doi: 10.1186/s40168-020-00875-0

Daliri, E. B., Tango, C. N., Lee, B. H., and Oh, D. H. (2018). Human microbiome restoration and safety. *Int. J. Med. Microbiol.* 308, 487–497. doi: 10.1016/j.ijmm.2018.05.002

Finlay, B. J., and Clarke, K. J. (1999). Ubiquitous dispersal of microbial species. *Nature* 400, 828. doi: 10.1038/23616

Gagliardi, A., Totino, V., Cacciotti, F., Iebba,V., Neroni, B., Bonfiglio, G., et al. (2018). Rebuilding the gut microbiota ecosystem. *Int. J. Environ. Res. Public Health.* 15, 1679. doi: 10.3390/ijerph15081679

Gurung, M., Li, Z., You, H., Rodrigues, R., Jump, D. B., Morgun, A., et al. (2020). Role of gut microbiota in type 2 diabetes pathophysiology. *EBioMedicine* 51, 102590. doi: 10.1016/j.ebiom.2019.11.051

Hou, K., Wu, Z.-X., Chen, X.-Y., Wang, J.-Q., Zhang, D., Xiao, C., et al. (2022). Microbiota in health and diseases. *Signal. Transduct. Target Ther.* 7:135. doi: 10.1038/s41392-022-00974-4

Li, M., Chen, W.-D., and Wang, Y.-D. (2020). The roles of the gut microbiota–miRNA interaction in the host pathophysiology. *Mol. Med.* 26, 101. doi: 10.1186/s10020-020-00234-7

Liu, Z.-Z., Liu, Q.-H., Liu, Z., Tang, J. W., Chua, E. G., Li, F., et al. (2021). Ethanol extract of mulberry leaves partially restores the composition of intestinal microbiota and strengthens liver glycogen fragility in type 2 diabetic rats. *BMC Complement. Med. Ther.* 21, 172. doi: 10.1186/s12906-021-03342-x

Liu, Z.-Z., Wang, C.-Y., Yang, J.-Y., Liu, Q.-H., Zhang, X., and Wang, L. (2022). Epidemiological study and clinical characterization of COVID-19 cases in Xuzhou, China. *Pak. J. Zool.* 54, 1747–1755. doi: 10.17582/journal.pjz/20210112090155

Malard, F., Dore, J., Gaugler, B., and Mohty, M. (2021). Introduction to host microbiome symbiosis in health and disease. *Mucosal Immunol.* 14, 547–554. doi: 10.1038/s41385-020-00365-4

Manor, O., Dai, C. L., Kornilov, S. A., Smith, B., Price, N. D., Lovejoy, J. C., et al. (2020). Health and disease markers correlate with gut microbiome composition across thousands of people. *Nat. Commun.* 11, 5206. doi: 10.1038/s41467-020-18871-1

Mishima, E., and Abe, T. (2021). Role of the microbiota in hypertension and antihypertensive drug metabolism. *Hypertens. Res.* 45, 246–253. doi: 10.1038/s41440-021-00804-0

Rosenberg, E. (2021). *Microbiomes.* Cham: Springer. doi: 10.1007/978-3-030-65317-0

Rowland, I., Gibson, G., Heinken, A., Scott, K., Swann, J., Thiele, I., et al. (2017). Gut microbiota functions: metabolism of nutrients and other food components. *Eur. J. Nutr.* 57, 1–24. doi: 10.1007/s00394-017-1445-8

Vernocchi, P., Del Chierico, F., and Putignani, L. (2020). Gut microbiota metabolism and interaction with food components. *Int. J. Mol. Sci.* 21, 3688. doi: 10.3390/ijms21103688

Wang, L., Li, F., Gu, B., Qu, P., Liu, Q., Wang, J., et al. (2022). Metaomics in clinical laboratory: potential driving force for innovative disease diagnosis. *Front. Microbiol.* 13, 883734. doi: 10.3389/fmicb.2022.883734

Zhang, F., Lau, R. I., Liu, Q., Su, Q., Chan, F. K. L., and Ng, S. C. (2022). Gut microbiota in COVID-19: key microbial changes, potential mechanisms and clinical applications. *Nat. Rev. Gastroenterol. Hepatol.* 1–15. doi: 10.1038/s41575-022-00698-4. [Epub ahead of print].

Zhang, Y.-D., Chen, D., Hu, L., Shen, L., Wu, R.-Y., Cao, F.-M., et al. (2022). Epidemiological characteristics of COVID-19 outbreak in Yangzhou, China, 2021. *Front. Microbiol.* 13, 865963. doi: 10.3389/fmicb.2022.865963

Zhang, Y. J., Li, S., Gan, R. Y., Zhou, T., Xu, D. P., and Li, H. B. (2015). Impacts of gut bacteria on human health and diseases. *Int. J. Mol. Sci.* 16, 7493–7519. doi: 10.3390/ijms16047493

frontiers | Frontiers in Microbiology

ORIGINAL RESEARCH
published: 31 May 2022
doi: 10.3389/fmicb.2022.919380

Metapath Aggregated Graph Neural Network and Tripartite Heterogeneous Networks for Microbe-Disease Prediction

*Yali Chen and Xiujuan Lei**

School of Computer Science, Shaanxi Normal University, Xi'an, China

More and more studies have shown that understanding microbe-disease associations cannot only reveal the pathogenesis of diseases, but also promote the diagnosis and prognosis of diseases. Because traditional medical experiments are time-consuming and expensive, many computational methods have been proposed in recent years to identify potential microbe-disease associations. In this study, we propose a method based on heterogeneous network and metapath aggregated graph neural network (MAGNN) to predict microbe-disease associations, called MATHNMDA. First, we introduce microbe-drug interactions, drug-disease associations, and microbe-disease associations to construct a microbe-drug-disease heterogeneous network. Then we take the heterogeneous network as input to MAGNN. Second, for each layer of MAGNN, we carry out intra-metapath aggregation with a multi-head attention mechanism to learn the structural and semantic information embedded in the target node context, the metapath-based neighbor nodes, and the context between them, by encoding the metapath instances under the metapath definition mode. We then use inter-metapath aggregation with an attention mechanism to combine the semantic information of all different metapaths. Third, we can get the final embedding of microbe nodes and disease nodes based on the output of the last layer in the MAGNN. Finally, we predict potential microbe-disease associations by reconstructing the microbe-disease association matrix. In addition, we evaluated the performance of MATHNMDA by comparing it with that of its variants, some state-of-the-art methods, and different datasets. The results suggest that MATHNMDA is an effective prediction method. The case studies on asthma, inflammatory bowel disease (IBD), and coronavirus disease 2019 (COVID-19) further validate the effectiveness of MATHNMDA.

Keywords: microbe-disease associations, heterogeneous network, metapath aggregated graph neural network, multi-head attention mechanism, COVID-19

OPEN ACCESS

Edited by:
Qi Zhao,
University of Science and Technology Liaoning, China

Reviewed by:
Wei Peng,
Kunming University of Science and Technology, China
Xing Chen,
China University of Mining and Technology, China

***Correspondence:**
Xiujuan Lei
xjlei@snnu.edu.cn

Specialty section:
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

Received: *13 April 2022*
Accepted: *29 April 2022*
Published: *31 May 2022*

Citation:
Chen Y and Lei X (2022) Metapath Aggregated Graph Neural Network and Tripartite Heterogeneous Networks for Microbe-Disease Prediction. Front. Microbiol. 13:919380. doi: 10.3389/fmicb.2022.919380

INTRODUCTION

The microorganisms related to the human body mainly include eukaryotes, archaea, bacteria, fungi, and viruses [Human Microbiome Project (HMP), 2012]. These microorganisms form different microbial communities and parasitize in different parts of the human body, such as the skin, mouth, genitalia, intestinal tract, and other parts. Studies have shown that the number of microbes in the

adult intestine is equivalent to 10 times that of human cells (Sender et al., 2016), which indicates that the microbial community in the human body is relatively large. Microbes are generally beneficial to the human body. For example, by fermenting food ingredients that cannot be digested by the host, gut microbes can promote nutrient and energy absorption (Gill et al., 2006; Marco et al., 2017). The *Bifidobacteria* in the human intestine can produce lactic acid and acetic acid after fermentation, which can promote the absorption of iron and vitamin D. Therefore, a set of balanced microbes can keep the human body away from physiological disorders, but the imbalance or decline of the microbial community can harm the human host and cause diseases. For example, a study has found that compared to normal children, children with asthma would have a smaller number of *Faecalibacterium*, *Lachnospira*, *Veillonella*, and *Rothia* (Arrieta et al., 2015). Another study found that the relative abundance of *Enterococcus*, *Escherichia*/*Shigella*, *Klebsiella*, *Streptococcus,* and *Peptostreptococcus* in the intestinal flora of patients with colorectal cancer was increased (Wang et al., 2012). These studies have shown that identifying the relationship between microbes and diseases can help us understand the pathogenesis of the disease, so as to carry out more targeted treatment. Therefore, determining the relationship between microbes and diseases has become a key research topic in the current bioinformatics field.

Verifying the relationship between microbes and diseases through biological experiments is a time-consuming and expensive task. Therefore, many computational models have been proposed to predict the association between microbes and diseases. Wang et al. (2021) wrote a review on circular RNAs and complex diseases, which classified the prediction models of circRNA-disease associations. Inspired by this study, we can divide these computational models into four types according to the differences in the microbe-disease association prediction strategies based on heterogeneous networks: path-based methods, random walk methods, bipartite local models, and matrix decomposition methods (Wen et al., 2021). Path-based methods are widely used in association prediction (Zhang et al., 2021; Liu et al., 2022a). They make predictions by calculating path-based scores between microbe nodes and disease nodes. For example, Chen X. et al. (2016) proposed the first model KATZHMDA to predict microbe-disease associations, which calculated the predicted probability score according to the walking step length and walking times between the two nodes in the microbe-disease network. Huang et al. (2017) proposed a computational model PBHMDA based on the depth-first search to predict potential microbes associated with diseases. Fan et al. (2018) developed a new model MDPH_HMDA to predict microbe-disease associations by integrating multi-source data and path-based HeteSim score. The random walk has aroused extensive interest in the field of microbe-disease prediction. For instance, Yan et al. (2019) proposed a prediction model BRWMDA based on similarity and bi-random walk to predict potential microbe and disease associations. Luo and Long (2018) proposed a computational model NTSHMDA based on random walk and network topology similarity to predict the associations between microbes and diseases. Wu et al. (2018) developed a method named PRWHMDA, which attempted to infer potential microbe-disease pairs by random walk on the heterogeneous network with Particle Swarm Optimization (PSO). Bipartite local models are also common methods, which work independently on the basis of both sides of a microbe-disease pair and can be combined to yield a definitive prediction result. For example, Zou et al. (2018) proposed a method called NCPHMDA that utilized the network consistency projection to predict microbe-disease associations. Wang et al. (2017) constructed a semi-supervised computational model LRLSHMDA based on a Laplacian regularization least squares classifier to predict the associations between microbes and diseases. In addition, some prediction models for microbe-disease associations were developed based on matrix factorization techniques. For instance, He et al. (2018) presented a method called GRNMFHMDA, which incorporated weighted K-nearest known neighbors to predict microbe-disease associations. Shen et al. (2017) developed a computational model of CMFHMDA, which used collaborative matrix factorization to reconstruct the association matrices between diseases and microbes. Wang Y. et al. (2022) proposed a method HNGFL based on heterogeneous network and global graph feature learning to predict microbe-disease association. In addition to these computational models, several review articles on microbe-disease associations have been published. For example, Pan et al. (2022) developed a comprehensive approach to predict associations between genomics, proteinomics, transcriptomics, microbiome, metabolomics, pathomics, radiomics, drug, symptoms, environment factors, and disease networks. Wang L. et al. (2022) provided a comprehensive review on predicting pairwise relationships between human microbes, drugs, and diseases, from biological data to computational models. Wen et al. (2021) provided a survey on predicting microbe-disease associations based on biological data and computational methods.

Although the above-mentioned methods have achieved relatively stable prediction performance in the association prediction task of microbes and diseases, there are still some limitations and deficiencies. First, the vast majority of methods make predictions based on small-scale datasets, which makes them unable to obtain accurate predictions when it comes to new diseases (or new microbes) due to a lack of training data. Second, microbe imbalance (or the occurrence of disease) is not influenced by a single factor. Some studies have shown that microbes participate in drug absorption and metabolism, thereby regulating drug efficacy and drug toxicity for disease (Zimmermann et al., 2021). However, the above-mentioned methods are only based on microbes and diseases, which makes these models unable to obtain accurate prediction results due to the lack of more semantic information about microbes and diseases in the prediction process.

Therefore, with the discovery of multivariate biological data, the heterogeneous graph embedding method is increasingly applied to relational prediction. It can learn semantic and structural information between nodes to compensate for the poor prediction performance due to the small amount of known associated data. For example, Lei and Wang (2020)

proposed a method based on Node2vec and a heterogeneous network scoring mechanism, called LGRSH, to predict the association between microbes and diseases. Liu et al. (2022b) proposed a method to identify miRNA-disease associations via deep forest ensemble learning based on autoencoder. Yang et al. (2022) proposed a DeepWalk-based method to predict lncRNA-miRNA associations via a lncRNA-miRNA-disease-protein-drug graph. Zhu et al. (2018) proposed a method using Metapath2vec to predict drug-gene interactions. Lei et al. (2021) developed a method, called CDWBMS, to predict circRNA-disease associations based on an improved weighted biased meta-structure. Zhang et al. (2020) adopted metapath2vec++ and matrix factorization to predict circRNA-disease associations. All the heterogeneous graph embedding methods have some limitations when applied to association prediction, such as ignoring the information of multiple nodes, discarding all intermediate nodes on the metapath, or only using a single metapath. This will affect the predictive performance of the model.

To deal with the above-mentioned issues, we developed a novel method based on a metapath aggregated graph neural network (MAGNN) and tripartite heterogeneous network for microbe-disease association prediction named MATHNMDA. In particular, we integrate information from different sources, such as microbe-disease associations, microbe-drug interactions, and disease-drug interactions, to construct a tripartite heterogeneous network of microbe-drug-disease. Further, we feed the heterogeneous network to MAGNN. For each layer of MAGNN, we first use intra-metapath aggregation with a multi-head attention mechanism to extract the structural and semantic information of the metapath instance. After that, we further apply inter-metapath aggregation with an attention mechanism to fuse latent vectors of multiple metapaths. Finally, we take the output of the MAGNN as the final embedding features of the microbe node and disease node, and make predictions. In order to verify the predictive performance of MATHNMDA, we carried out cross-validation experiments, and the results indicate that MATHNMDA can effectively identify potential disease-related microbes.

Overall, our main contributions are as follows:

(1) We expand known microbe-disease association data by integrating multiple databases, and construct a tripartite heterogeneous network by introducing drug-disease associations and microbe-drug associations. We further apply MAGNN to predict microbe-disease associations.
(2) We use intra-metapath aggregation with the multi-head attention mechanism to learn the topological information and semantic information embedded in the internal nodes of metapath, so that the embedding learned by the target node is more comprehensive.
(3) We use inter-metapath aggregation with an attention mechanism to aggregate the embeddings of different metapaths for target nodes (microbe nodes or disease nodes).
(4) We conduct a case study of coronavirus disease 2019 (COVID-19) to verify the effectiveness of the MATHNMDA model.

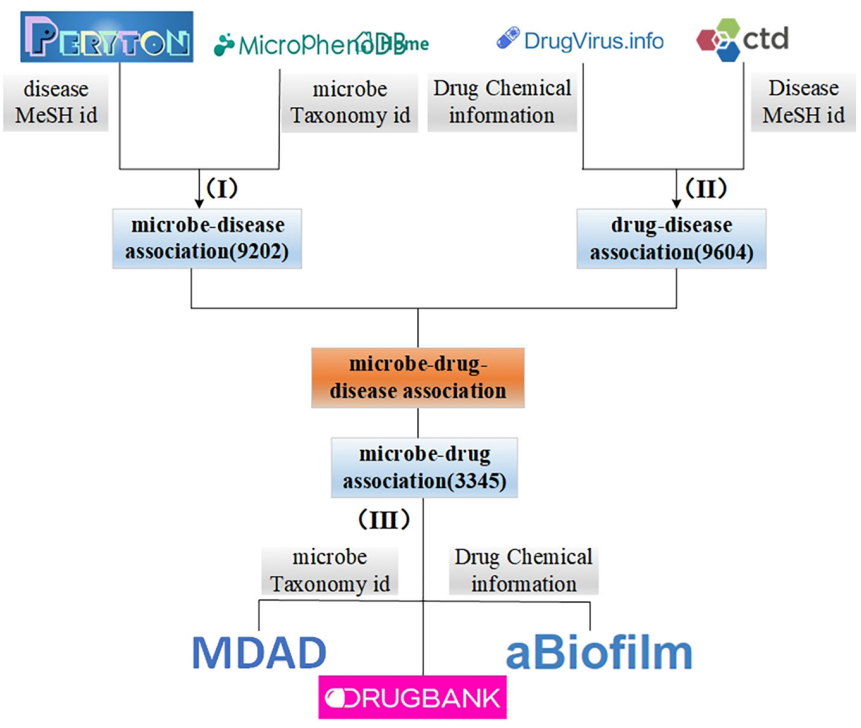

FIGURE 1 | The processing and integration process of microbe-drug-disease association data.

MATERIALS AND METHODS

Dataset

In this study, we integrate the information obtained from different sources. First, we collect microbe and disease association data from Peryton (Skoufos et al., 2020) and MicroPhenoDB (Yao et al., 2021). Among them, Peryton includes more than 7,900 relationships between 43 diseases and 1,396 microbes. The data in MicroPhenoDB are collected from some public datasets, such as Human Microbe-Disease Association Database (HMDAD; Kong et al., 2017), Disbiome (Yorick et al., 2018), Virulence Factor Database (VFDB; Chen L. et al., 2016), etc. MicroPhenoDB has 5,565 relationships between 515 diseases and 1,717 microbes. After eliminating redundancy for the same diseases and microbes, we obtain a total of 9,202 associations between 538 diseases and 2,491 microbes. Furthermore, we collect data about microbes and their related drugs from Microbe-Drug Association Database (MDAD; Sun et al., 2018), drugVirus (Andersen et al., 2020), and aBiofilm (Akanksha et al., 2017), and remove redundant records to obtain a total of 132 microbes and 1,933 drugs and 3,345 microbe-drug associations. Then, we download disease-drug interaction data from drugBank (Wishart et al., 2017) and Comparative Toxicogenomics Database (CTD; Davis et al., 2012) databases, and we obtain 9,604 interactions between 127 diseases and 247 drugs after de-redundancy. **Figure 1** illustrates the integration process of microbe-drug data, drug-disease data, and microbe-disease data. It is worth noting that in this study, we unified the disease, microbe, and drug according to the MESH id of the disease, the taxonomy id of the microbe and the chemical information of the drug, disease-related drugs, are included in drugs related to microbes.

Construction of Microbe-Drug-Disease Tripartite Heterogeneous Network

In this study, we use microbe-disease, microbe-drug, and disease-drug associations to build a tripartite network. The relationship between microbes, drugs, and diseases is shown in **Figure 2A**. A certain microbial imbalance can lead to certain diseases. and the pathogenesis of a certain disease will be affected by certain microbial communities. Some drugs can treat some diseases, and certain diseases can be treated with certain drugs. Microbes can regulate the activity and toxicity of drugs (Zimmermann M. et al., 2019). Drugs in turn can change the diversity and function of microbial communities. Suppose M, C, and D, respectively, represent all the sets of microbes, drugs, and diseases in the network, $m_i \in M$ represents a microbe, $i = 1, 2, 3..., n_m$; $c_j \in C$ represents a drug, $j = 1, 2, 3..., n_c$; and $d_k \in D$ represents a disease, $k = 1, 2, 3..., n_d$. Construct a tripartite heterogeneous network based on the relationship among microbes, drugs, and diseases. Here, we can simplify it to an undirected and unweighted network to represent the existence of associations, as shown in **Figure 2B**. We further construct the microbe-disease adjacency matrix $B \in R^{n_m \times n_d}$, where n_m represents the number of microbes and n_d represents the number of diseases. If there is a known association between a microbe node i and a disease node j, the value of $B(i,j)$ is 1, otherwise, it is 0.

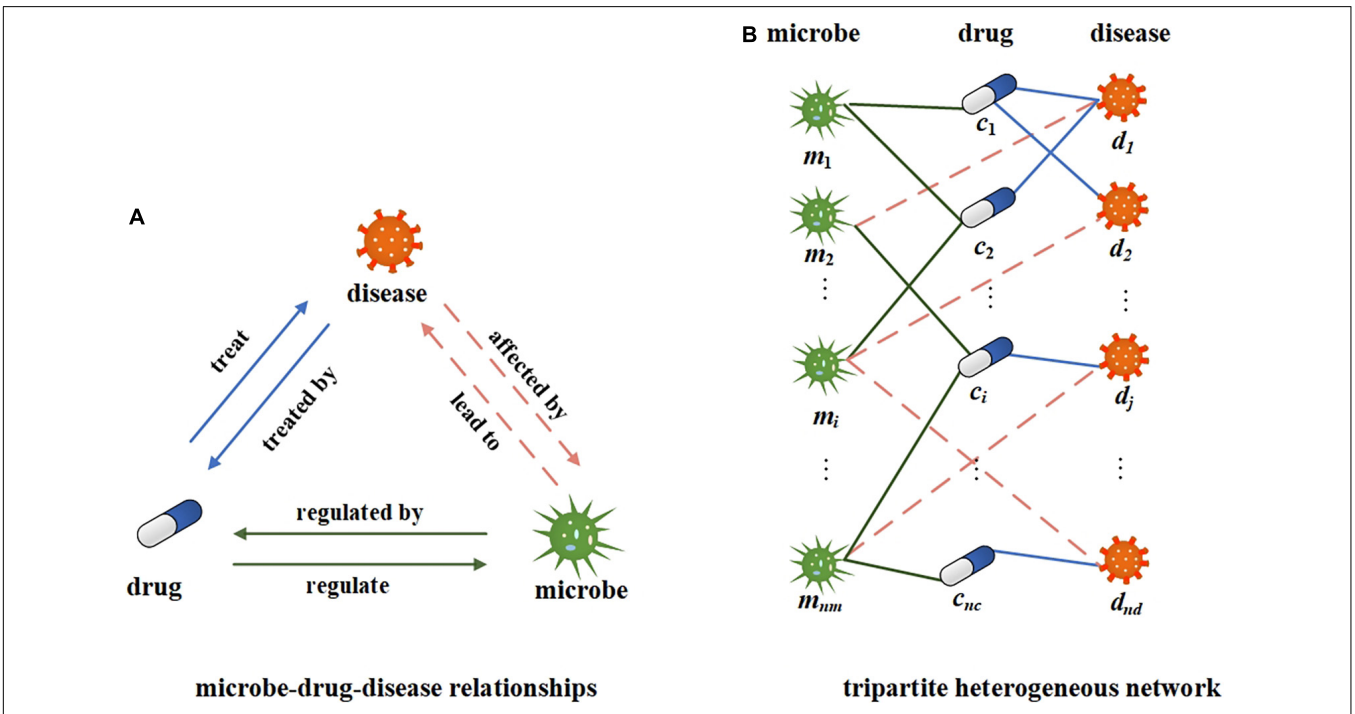

FIGURE 2 | Illustration of microbe-drug-disease relationship and heterogeneous networks. **(A)** The illustration of microbe-drug-diease relationships. **(B)** A tripartite network corresponding to the microbe-drug-disease relationships.

MATHNMDA

Our proposed MATHNMDA model consists of three main steps, as shown in **Figure 3**. The model takes heterogeneous microbe-drug-disease interaction network and MAGNN to predict microbe-disease associations. First, we take heterogeneous network as input of the MAGNN. Second, for each layer of the MAGNN, we use intra-metapath aggregation to learn the structural and semantic information embedded in the target node, metapath-based neighbor nodes, and the context between them. Third, we apply inter-metapath aggregation to combine the semantic information of all different metapaths. Finally, we take the output of the MAGNN as vector representations of microbe nodes and disease nodes, which can be used to predict potential microbe-disease associations.

Intra-Metapath Aggregation

In this study, we predict novel microbe-disease associations on the heterogeneous microbe-drug-disease interaction networks based on MAGNN. Given a heterogeneous network $G = (V, E)$, where V and E represent sets of nodes and edges, respectively, and the mapping functions of nodes and edges are δ: $V \rightarrow A$ and ψ: $E \rightarrow R$, A represents node types, R denotes edge types, and $|A| + |R| > 2$. Given a metapath M on the heterogeneous network G, we can define it as a path of the form $A_1 \rightarrow A_2 \rightarrow \ldots \rightarrow A_{n-1} \rightarrow A_n$, which can be abbreviated as $A_1A_2 \ldots A_{n-1}A_n$. The relationship between node types A_1 and A_n is $R = R_1 \circ R_2 \circ \ldots \circ R_{n-1}$, where $\circ$ represents the composite operation. That is to say, the relationship R is obtained by compositing the n–1 relationships of these $R_1, R_2, \ldots, R_{n-1}$. Therefore, a metapath can capture specific semantic information in the graph, and different metapaths represent different semantic information. For example, for the metapath microbe-drug-disease (abbreviated as *m-c-d*), drug *c* can act on microbe *m*, and drug *c* can be used to treat disease *d*, so microbe *m* may be associated with disease *d*. The key idea of intra-metapath aggregation is to learn structural and semantic information embedded in target nodes, metapath-based neighbors, and the context between them by encoding metapath instances under a certain metapath. Next, we introduce the process of intra-metapath aggregation in detail.

Given a metapath M, we define a sequence of nodes in G that follow the pattern of M as a metapath instance, defined as $M(i,j)$, which is represented as a metapath instance connecting the target node i and its neighbor node j based on the metapath. Here, $j \in N_i^M$, N_i^M represents the set of nodes connected to node i through the metapath instance $M(i,j)$. It is worth noting that if the metapath instance $M(i,j)$ is symmetric, $j \in N_i^M$ also includes node i itself. Then we define the intermediate node set of $M(i,j)$ as $TH(i,j)$, $TH(i,j) = M(i,j) / \{j, i\}$, where $\{j, i\}$ represents the set with elements i, j.

As mentioned before, intra-metapath aggregation learns structural and semantic information of target nodes by encoding metapath instances. Sun et al. (2019) proposed a method for knowledge graph embedding based on relational rotation in complex space, called RotatE. RotatE can model all relational patterns, so we use RotatE as the metapath instance encoder in this study. Given a metapath instance $M(i,j) = (i, th_2 \ldots, th_{n-1}, j)$, for convenience, let set $i = th_1$ and $j = th_n$. R_i represents the relationship between node th_i and node th_{i+1}, and the relationship vector is r_i. Therefore, for the metapath instance $M(i,j)$, RotatE can be defined as follows:

$$\begin{aligned} \Theta_1 &= \tilde{h}_{th_1} = \tilde{h}_i \\ \Theta_i &= \tilde{h}_{th_i} + \theta_{i-1} \odot r_i \\ h_{M(i,j)} &= \frac{\theta_n}{n} \end{aligned} \tag{1}$$

where $\tilde{h}_{th_i}$ and r_i are vectors of complex space, $\odot$ represents hadamard product, $hM(i,j) \in R^{d'}$, and d' is the dimension of $h_{M(i,j)}$. In which case, we get vector representation of the metapath instance $M(i,j)$. It is important to note that there may be multiple instances of the metapath M connecting nodes i and j, but we use $M(i,j)$ to represent a single instance here.

Graph attention network (GAT) is an effective graph representation learning tool, which represents the importance of neighbor nodes to the target node by assigning different weights to different neighbor nodes (Bian et al., 2021). Here, for target node i and metapath M related to i, we first use GAT to assign weights (attention coefficients) to metapath instances in M, thereby learning the importance of different metapath instances to target nodes. Then the features of different metapath instances are aggregated according to the obtained attention coefficients, which are represented as the feature vector of the target node i. Given a metapath instance $M(i,j)$, its attention coefficient can be defined as:

$$e_{ij}^M = LeakyReLU\left(\delta_M^T \left[\tilde{h}_i \parallel h_{M(i,j)}\right]\right) \tag{2}$$

where δ_M^T is the attention parameter of the metapath M, and $\parallel$ represents connection operation. To make the attention coefficients of different metapath instances comparable, we use the softmax function to normalize e_{ij}^M:

$$\alpha_{ij}^M = softmax\left(e_{ij}^M\right) = \frac{exp\left(e_{ij}^M\right)}{\sum_{k \in N_i^M} \exp\left(e_{ij}^M\right)} \tag{3}$$

Then, we aggregate the feature vectors of all metapath instances according to the activation function $\sigma(\cdot)$ to obtain the vector representation of node i based on the metapath M:

$$\alpha_{ij}^M = \sigma\left(\sum_{j \in N_i^M} \alpha_{ij}^M \cdot h_{M(i,j)}\right) \tag{4}$$

In this study, we further introduce a multi-head attention mechanism to stabilize the learning process of attention coefficients and reduce the influence of a single attention. Specifically, we independently repeat the attention mechanism K times and concatenate vector representation learned by each attention head. Therefore, the vector representation of node i can be further rewritten as follows:

$$h_i^M = \mathop{\Big\Vert}_{k=1}^{K} \sigma\left(\sum_{j \in N_i^M} \alpha_{ij}^M \cdot h_{M(i,j)}\right) \tag{5}$$

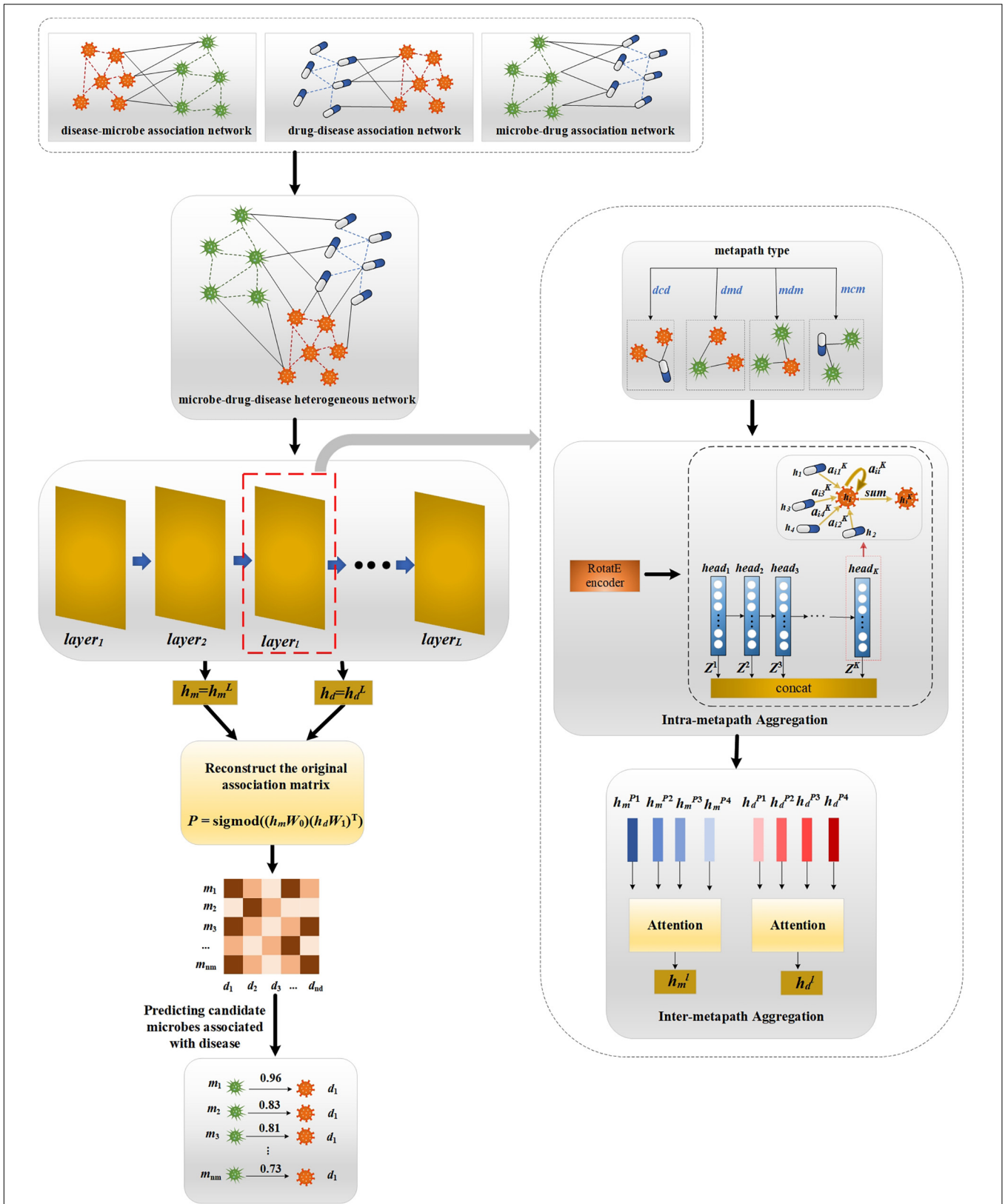

FIGURE 3 | The framework of MATHNMDA. *mcm*, *mdm*, *dmd,* and *dcd* are four metapaths, where *m* represents microbe, *c* represents drug, and *d* represents disease.

Since the metapath is undirected, each node in the metapath can be either a start node or an end node. Therefore, for the metapath set starting or ending with node type $a \in A$, it is denoted as $M_a = \{M_1, M_2, \cdots, M_S\}$, and S represents the number of metapaths. Intra-metapath aggregation obtains M metapath-specific vector representations of the target node $i \in V_a$, defined as $\left\{h_i^{M_1}, h_i^{M_2}, \cdots, h_i^{M_S}\right\}$.

Inter-Metapath Aggregation

After the intra-metapath aggregation of metapaths, we obtain the vector representation of a single metapath M for target node i. Then, we need to aggregate the semantic information and structural information of node i based on all metapaths of M_a, where S represents the number of metapaths. The node embedding set corresponding to these S metapaths is $\left\{h_i^{M_1}, h_i^{M_2}, \cdots, h_i^{M_S}\right\}$. A simple aggregation method between metapaths is to take the average of these node embeddings. However, because the importance of metapaths to node i in a heterogeneous network is different, we allocate weight for each metapath pattern through the attention mechanism, and then perform aggregation.

Specifically, given a metapath M_p, $M_p \in V_a$, we first transform these metapath-specific node vectors for all nodes $i \in V_a$ with the tanh function, and then take average value as feature of M_p:

$$S_{M_p} = \frac{1}{|V_a|} \sum_{i \in V_a} tanh\left(W_a \cdot h_i^{M_p} + b_a\right) \tag{6}$$

where W_a is the weight matrix of nonlinear transformation specific to node type a, and B_a is the corresponding bias, both of which are learnable parameters. V_a indicates all nodes of type a in the network.

Then we use the attention mechanism to calculate the importance of each metapath pattern for the target node i, and normalize the obtained attention coefficients by the softmax function. Then we fuse the corresponding vector representations of these metapaths to get the output of the target node i, as shown in Equation 7:

$$\begin{aligned} e_{M_p} &= c_a^T \cdot S_{M_p} \\ \beta_{M_p} &= softmax\left(e_{M_p}\right) = \frac{exp\left(e_{M_p}\right)}{\sum_{M_p \in M_a} \exp\left(e_{M_p}\right)} \\ h_i^{M_a} &= \sum_{M_p \in M_a} \beta_{M_p} \cdot h_i^{M_p} \end{aligned} \tag{7}$$

where c_a^T denotes the attention parameter, β_{M_P} denotes the normalized attention score, and M_P denotes the Pth metapath in M_a. $h_i^{M_a}$ represents the embedding vector of node i based on aggregation between metapaths.

MAGNN

The goal of a graph neural network (GNN) is to learn the low-dimensional vector representation of each node, which can be used for many downstream tasks, such as node clustering, node classification, and link prediction. Thus, we further apply an L-layer GNN to learn the low-dimensional representation vectors of microbe nodes and disease nodes. At each layer of the GNN, we use intra-metapath aggregation and inter-metapath aggregation to obtain vector representations of node-based metapath. In this way, we can define the low-dimensional representation for node i at the lth layer:

$$h_i^l = \sigma\left(W_o^l \cdot \left[h_i^{M_a}\right]^l\right) \tag{8}$$

where $\sigma(\cdot)$ is an activation function and W_o^l represents the weight vector at the lth layer. h_i^l represents the vector representation for node i at the lth layer, which is also the input of the $(l+1)$th layer. We define $h_i^0 = W_a \cdot X_i^a$, where W_a represents the linear transformation matrix of node type a and X_i is the original feature vector for node of type a. Here, we use one-hot encoding to initialize each type of node in the heterogeneous network.

Finally, we use vector representation of node i at the Lth layer to serve as the final embedding for nodes i:

$$h_i = h_i^L \tag{9}$$

where h_i^L represents vector representation of node i at the Lth layer.

Reconstruction of Microbe-Disease Association

After we get the final embeddings of all microbe nodes and disease nodes, we can predict new microbe-disease associations by reconstructing microbe-disease associations. Here we perform a simple inner product operation on the microbe and disease embeddings. In this case, each microbe-disease pair will receive a new score. Specifically, given a microbe node m and a disease node d, the predicted score C_{md} between them can be calculated as:

$$C_{md} = sigmoid\left(h_m^T \cdot h_d\right) \tag{10}$$

where h_m and h_d represent the final embeddings of microbes and diseases, respectively.

Optimization

Since our task is to predict microbe-disease associations, this is equivalent to a binary classification problem. So, here we use the cross-entropy function as the loss function and optimize through negative sampling:

$$L = -\sum_{(m,d) \in \mu} log\left(C_{md}\right) - \sum_{(m,d) \in \mu^-} log\left(-C_{md}\right) \tag{11}$$

where μ represents the set of positive samples, and μ^- represents the set of negative samples obtained by negative sampling.

RESULTS

In this section, we evaluate the performance of MATMNMDA through some experiments and analysis of the results. At the same time, we also analyze and adjust some parameters of the model in order to make better predictions.

Evaluation Metrics

In this study, we mainly use two metrics to evaluate the performance of the model, area under the receiver operating characteristic curve (AUC) and area under the precision–recall curve (AUPR), which are widely used in association prediction tasks.

AUC: This corresponds to the area of a planar graph bounded by the receiver operating characteristic (ROC) curve and horizontal axis, which can estimate the performance of binary classification models. The value of AUC is between 0 and 1. When it is closer to 1, the model performs better. In practical application, the advantages and disadvantages of different models can be compared by comparing the AUC values of different classification models.

AUPR: The precision–recall (PR) curve is also used to evaluate the classification ability of the model. In particular, the PR curve can collect more information when dealing with some imbalanced datasets. The area enclosed by the PR curve and the abscissa axis is called AUPR.

Baselines

In order to test the effectiveness of the MATMNMDA model, we compare it with six state-of-the-art methods based on the data processed in this study. Here, we calculate the AUC and AUPR values of these methods under the same conditions and analyze the results. The six baselines are as follows:

BRWMDA (Yan et al., 2019): It is a similarity-based and modified bi-random walk to predict associations between microbes and diseases.

KATZHMDA (Chen X. et al., 2016): It is a method to predict microbe-disease associations based on the katz metric.

LRLSHMDA (Wang et al., 2017): It is a semi-supervised model to predict microbe-disease associations by introducing a Gaussian kernel and Laplacian regularization.

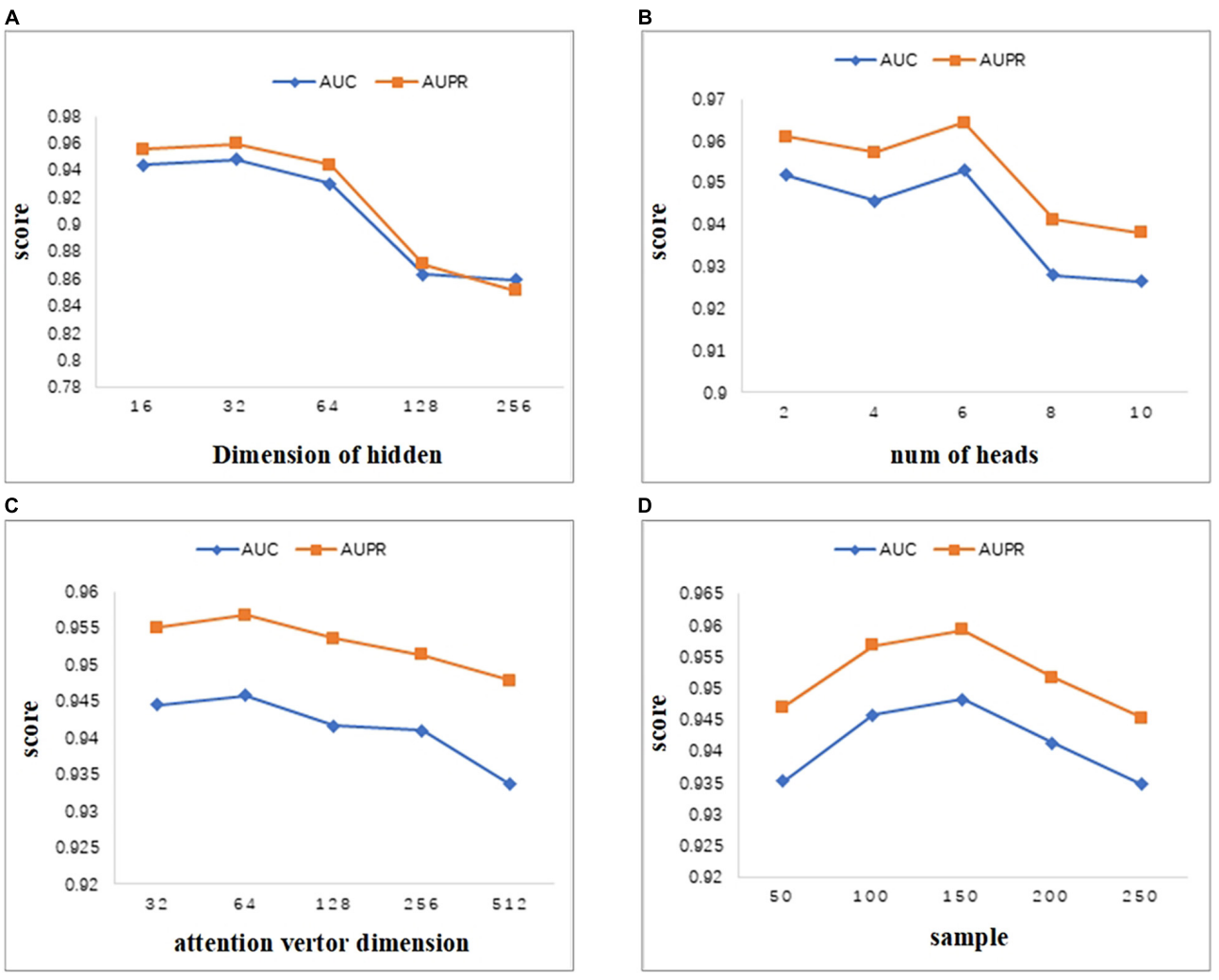

FIGURE 4 | Parameter analysis. **(A)** Comparison of AUC and AUPR for different hidden layer dimensions. **(B)** Comparison of AUC and AUPR of attention heads for different multi-attention mechanisms. **(C)** Comparison of AUC and AUPR for different attention vector dimensions. **(D)** Comparison of AUC and AUPR for different numbers of neighbors sampled by nodes.

NCPHMDA (Zou et al., 2018): It uses network consistent projections to predict microbe-disease associations.

NTSHMDA (Luo and Long, 2018): Predicting microbe-disease associations using heterogeneous network topological similarity and random walks.

CRPGCN (Ma et al., 2021): It is a method based on graph convolutional network (GCN) and random walk with restart (RWR), which was proposed for the cirRNA-disease association prediction task. Here, we use it as a baseline method for the prediction of microbe-disease association.

We compare MATMNMDA with these six baseline methods under the same conditions. For the CRPGCN method, the similarity of microbes and diseases is calculated in the same way as BRWMDA. For the MATMNMDA model, we first perform negative sampling on the microbe-drug-disease heterogeneous network. The positive and negative sample ratios of the training set, validation set, and test set are 1:1, and the proportion of the training set, validation set, and test set is 8:1:1, respectively. We randomly initialize vector representations of microbe nodes, drug nodes, and disease nodes. The Adam optimizer is used to optimize the model. The dropout and early stopping mechanisms are used to prevent overfitting. Here, according to the extensive literature (Phaisangittisagul, 2016), we set the value of dropout to 0.5. We train the model 100 times.

Parameter Analysis

In this section, we analyze the sensitivity of parameters. As we all know, important parameters will affect the performance of the model, so it is very necessary to conduct parameter analysis for the model. Some important parameters involved in the MATMNMDA model include the dimension of hidden layer, number of heads in the multi-head attention mechanism, dimension of attention vector, and number of neighbors sampled by the nodes in the experiment. We analyze these four parameters in turn and evaluate their impact on model performance.

As can be seen from **Figure 4A**, we set the dimension of the hidden layer to 16, 32, 64, 128, 256. As the dimension of the hidden layer increases, the performance of the model first increases. When the dimension reaches 32, both AUC and AUPR reach the maximum value. As the dimension continues to increase, the performance of the model begins to decrease gradually. Therefore, in this study, we set the embedding dimension of the hidden layer as 32. When the dimension changes between 16 and 256, the values of AUC and AUPR vary greatly. Thus, the MATMNMDA model is sensitive to the dimension of the hidden layer.

MATMNMDA model adopts a multi-attention mechanism to stabilize the process of attention coefficient learning. **Figure 4B** shows the influence of the number of attention heads in the multi-attention mechanism on model performance. We change the number of attention heads from 2 to 10 by step 2. It can be seen that when the number of attention heads is set to 6, the model has the best performance. **Figure 4C** shows the influence of the dimension of the attention vector. The dimension of the attention vector changes between 32 and 512. It can be observed that the vector dimension is too small or too large, which is not good for the performance of the model. Specifically, if the dimension of the attention vector is too large, it may lead to overfitting, which will degrade the performance of the model. When the dimension is set to 64, we can obtain better prediction ability.

In the MATMNMDA model, intra-metapath aggregation involves aggregating features of neighbor nodes to represent the representation of the current target node. Therefore, we analyze the number of neighbor nodes. In **Figure 4D**, the number of neighbor nodes is selected from {50,100,150,200,250}. It can be seen that when the number of neighbor nodes is too small or too large, the performance of the model is not very good. Specifically, if the number of neighbor nodes is too small, the structural information and semantic information of the target node may not be so comprehensive, while too large may cause noise. Therefore, we set the number of neighbor nodes to 150.

Ablation Study

As mentioned in the Introduction section, the previous heterogeneous network embedding methods have the following problems: (1) They only consider the neighbors based on the metapath, and do not consider the intermediate nodes inside the metapath. (2) In the metapath-based embedding, only the single best metapath is considered, and our model is proposed based on these problems. Therefore, in order to verify the effectiveness of each module of our model, we further conduct experiments on different variants of the MATMNMDA model. Taking MATMNMDA as a reference model, here we tested three variants of it.

MATMNMDA_nb: It only considers metapath-based neighbor nodes and does not consider intermediate nodes.

MATMNMDA_sm: It only considers the single best metapath.

MATMNMDA_avg: It replaces the RotatE with a mean encoder.

Figures 5, **6** show the comparison results of the MATMNMDA model and its variants. We can see that the MATMNMDA model has the highest AUC and AUPR. Followed by MATMNMDA_avg, MATMNMDA_sm has the worst performance. Comparing MATMNMDA and MATMNMDA_avg, we find that the MATMNMDA model performs better, which is because the mean encoder essentially treats metapath instances as a set and ignores the information embedded in the sequential structure of metapaths, while RotatE can be modeled according to the sequential structure of metapaths, thereby preserving the information embedded in the sequential structure of metapaths, so RotatE helps to improve the performance of the model by a small amount. Comparing MATMNMDA and MATMNMDA_nb, we can find that considering the intermediate nodes inside the metapath can help the model to obtain more structural information and thus improve the performance of the model. The results of MATMNMDA and MATMNMDA_sm show that the model performance can be significantly improved by combining multiple metapaths.

Comparison With Baselines

We run these baseline methods with default parameters. **Figures 7**, **8** show the performance of different methods. Our model achieves the highest prediction results on these two

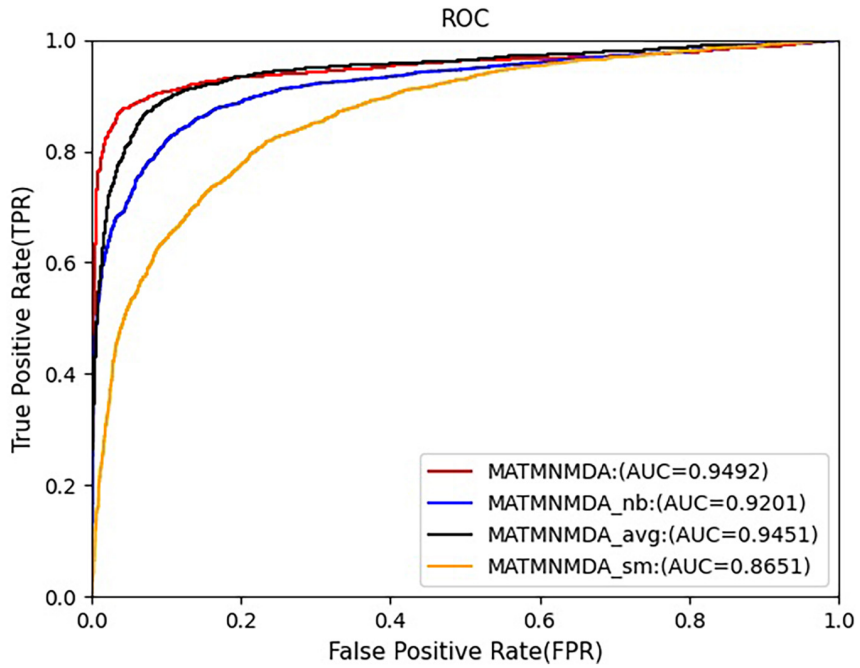

FIGURE 5 | Comparison of AUC for MATMNMDA and its variants.

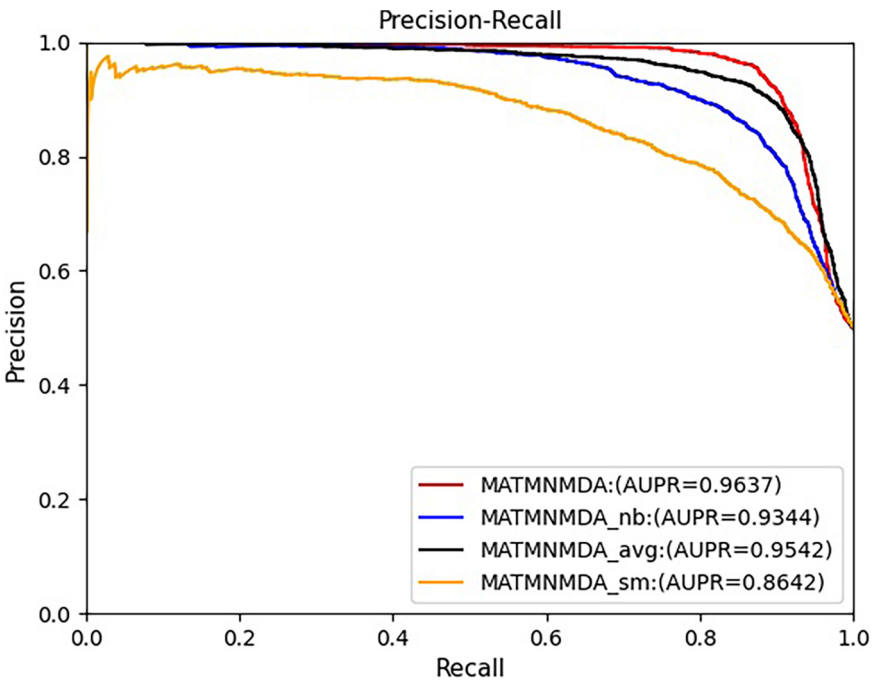

FIGURE 6 | Comparison of AUPR for MATMNMDA and its variants.

evaluation metrics, and its AUC and AUPR reach 0.9492 and 0.9637, respectively, which are better than all baseline methods. The CRPGCN model occupies the second position. It applies the RWR algorithm, which allows each calculated node to better fuse information from neighboring nodes with higher weights, so that GCN can learn features faster and get higher prediction scores. Next is the LRLSHMDA model, because the topological structure in the microbe-disease association network helps the model to effectively use the hidden information of vertices and edges, which helps to train the optimal classifier, so that microbe-disease

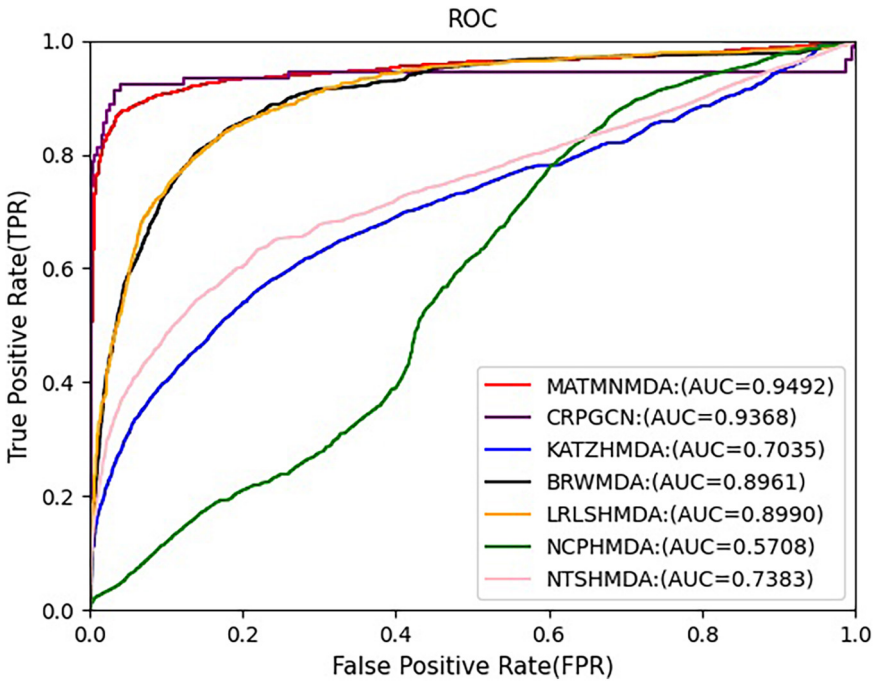

FIGURE 7 | Comparison of AUC for MATMNMDA and baselines.

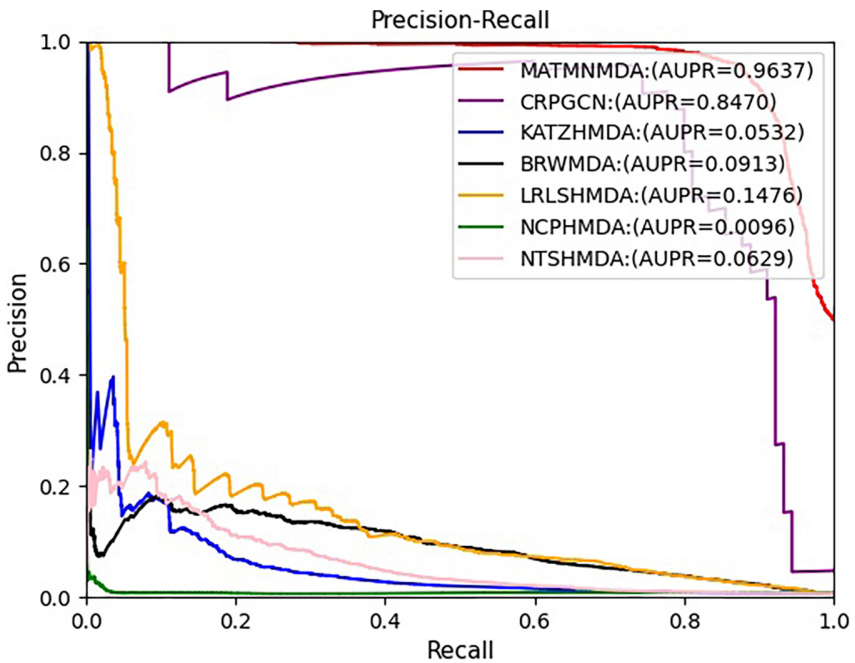

FIGURE 8 | Comparison of AUPR for MATMNMDA and baselines.

associations can be predicted more accurately. Next is the BRWMDA model, which also achieved good prediction results, because the BRWMDA model is based on similarity and bi-random walk, and it can model the topology information of the network well. However, NCPHMDA and NTSHMDA have poor prediction performance, because although we have obtained 9,202 known microbe-disease associations, they account for 0.7% of the whole microbe-disease association. The whole network is very sparse, and these two methods are based on network structure, so their performance is relatively poor.

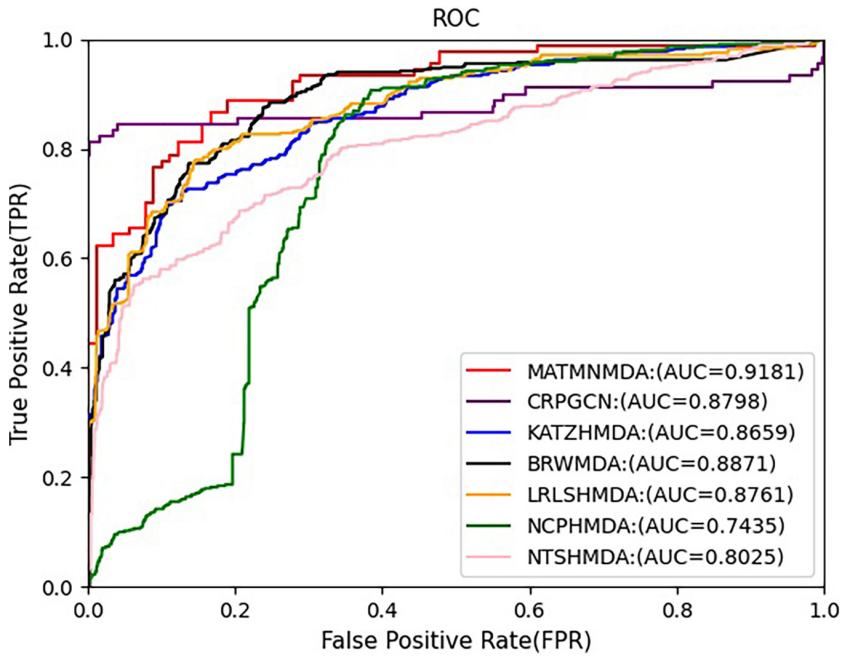

FIGURE 9 | Comparison of AUC for MATMNMDA and baselines on HMDAD dataset.

Comparison With Different Datasets

In this study, we augment the known microbe-disease association data. In order to verify the validity of the MATHNMDA model in our dataset, we also compare MATHNMDA with baseline methods on HMDAD and Disbiome, which are commonly used in microbe-disease prediction. The results are shown in **Figures 9**, **10**, and the comparison results of these methods on the three datasets are shown in **Table 1**. From **Table 1**, we can see that on each dataset, our model achieves the highest prediction value. It performs best on our dataset, so we can suggest that augmenting the known microbe-disease associations can help to improve the performance of the MATHNMDA. In addition, we can see that CRPGCN, LRLSHMDA, and BRWMDA methods perform well on these three datasets among the baseline methods. It also shows that these three methods are suitable for both large and small datasets, and the robustness of models is better. The remaining comparison methods are only suitable for small datasets.

Case Study

To further evaluate the predictive ability of the MATMNMDA model in identifying new microbe-disease associations, we conduct case studies on asthma, inflammatory bowel disease (IBD), and COVID-19. For each disease, microbes that have known associations with the disease are first removed. Then the predicted scores of candidate microbes are sorted in descending order according to the MATMNMDA model. Finally, we verify whether the top 10 microbes associated with the disease are confirmed by the relevant literature.

Asthma is a heterogeneous disease characterized by chronic airway inflammation (Lee and Kim, 2021). More than 300 million people worldwide suffer from asthma, and the incidence of asthma increased by 12.6% between 1990 and 2015 (Vasily, 2017). Therefore, it is necessary to study asthma deeply. With the development of 16rRNA sequencing technology, it has been found that there is an important relationship between asthma and microbe. In this study, when we employ the MATMNMDA model to predict potential microbe-disease associations, 7 of the top 10 candidate microbes are verified by relevant literature in PubMed (as shown in **Table 2**). For example, studies have shown that *Staphylococcus* (2nd) is linked to asthma attacks (Zhou et al., 2019), the relative abundance of *Bacteroidetes, Clostridium* (3rd), and *Enterobacteriaceae* were high, and the relative abundance of *Bifidobacterium* and *Lactobacteriaceae* were low, which is associated with allergies, eczema, or asthma (Zimmermann P. et al., 2019). An increased prevalence of *Staphylococcus aureus* (6th) colonization and sensitivity against its proteins are found in asthma (Tomassen et al., 2013). Bacterial dysbiosis and abundance within *Firmicutes* (4th) were significantly reduced in asthmatic children (Hufnagl et al., 2020). *Human parainfluenza virus* 1 (4th) was detected most frequently from patients with URI (3.74%, 47/1,257), followed by those with bronchitis (2.14%, 53/2,479), pneumonia (0.85%, 145/17,068), bronchiolitis (0.47%, 12/2,536), and asthma (0.43%, 2/462; Wang et al., 2015). *Herpesviruses* were the most abundant virus type in the asthma group (44.6 ± 4.6%), mainly *cytomegalovirus* (CMV; 9th) and EBV, which accounted for 24.5 ± 3.3 and 16.9 ± 3.5%, respectively (Choi et al., 2021). In healthy controls, the two viruses were 5.4 ± 2.5 and 7.1 ± 3.0%, respectively. Therefore, CMV and EBV are more abundant in patients with asthma exacerbations.

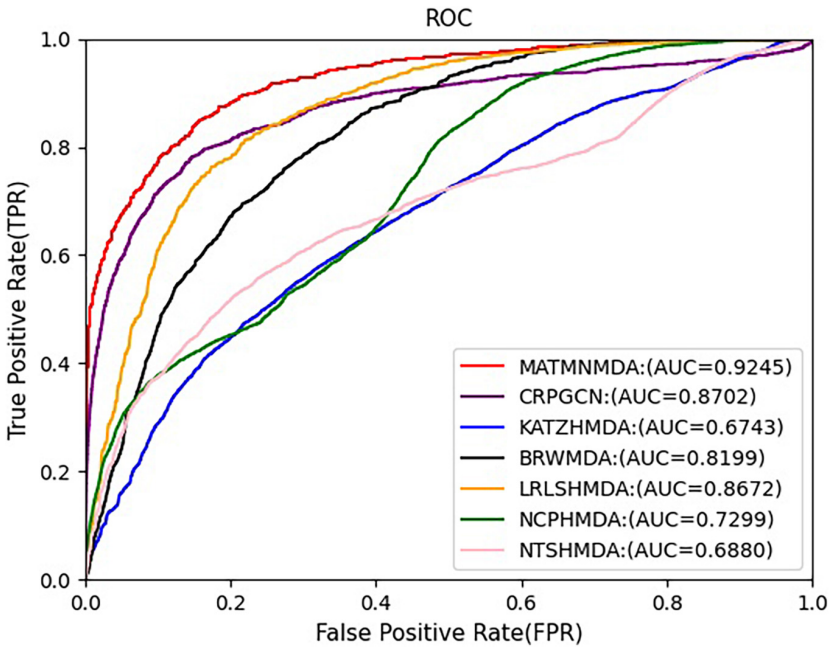

FIGURE 10 | Comparison of AUC for MATMNMDA and baselines on Disbiome dataset.

Inflammatory bowel disease (IBD) is an idiopathic intestinal inflammatory disease, mainly including ulcerative colitis (UC) and Crohn's disease (CD), with clinical manifestations of abdominal pain, diarrhea, and bloody stools. It is difficult to completely cure the disease, which is easy to recur, and there is a potential risk of cancer. Therefore, we perform a case study of IBD to evaluate the predictive ability of the MATMNMDA model for novel microbe-disease associations. The results are shown in **Table 3**, and 7 of the top 10 candidate microbes are verified by relevant literature. For example, *Fusobacterium* (2nd), *Halomonas*, *Acinetobacter*, *Shewanella*, and *Streptococcus* were enriched in the CD microbiota (Weng et al., 2019). Increased abundance of *Salmonella* sp., *Campylobacter* sp., *Helicobacter* sp., *Escherichia coli*, *Alcaligenes* sp., and *Mycobacterium* sp. (4th) was observed in IBD patients (Olejniczak-Staruch et al., 2021). IBD patients exhibit a lower abundance of butyrate-producing bacteria (6th; Gasaly et al., 2021) and butyrate content. Although some findings related to IBD dysbiosis have varied among the studies due to differences in sample type, survey method, patient profile, and drug treatment, the most consistent observation across these studies is that bacterial diversity decreased in IBD patients. For viruses infecting human cells, *Anelloviridae* (5th) showed a higher prevalence in very early-onset IBD compared to healthy controls (Liang et al., 2020). The population of *Firmicutes* decreased (7th) and that of *Proteobacteria* increased (Matsuoka and Kanai, 2015). Researchers observed a bias in the fungal microbiota in IBD compared to the normal control group, with an increased *Basidiomycota*/*Ascomycota* ratio (8th), a decreased *Saccharomyces cerevisiae* ratio, and an increased *Candida albicans* ratio (Sokol et al., 2017). There are experiments to verify that the intensity of both CMV and *human herpesvirus* 6 (HHV-6; 9th) correlated with endoscopic disease severity in IBD (CMV, $p = 0.010$ and HHV-6, $p = 0.048$; Sipponen et al., 2011).

Coronavirus disease 2019 (COVID-19) is a disease caused by severe respiratory syndrome coronavirus 2 (SARS-CoV-2). It has

TABLE 1 | Performance comparison of MATMNMDA and baselines on different datasets.

DATASET	HMDAD		Disbiome		Our dataset	
METHOD	AUC	AUPR	AUC	AUPR	AUC	AUPR
CRPGCN	0.8798	0.4533	0.8702	0.4965	0.9368	0.8470
KATZHMDA	0.8815	0.4828	0.6743	0.0508	0.7035	0.0532
BRWMDA	0.8748	0.3966	0.8199	0.0705	0.8961	0.0913
LRLSHMDA	0.8766	0.4960	0.8672	0.1370	0.8990	0.1476
NCPHMDA	0.7524	0.0795	0.7299	0.1024	0.5708	0.0092
NTSHMDA	0.8276	0.2975	0.6880	0.0630	0.7383	0.0629
MATMNMDA	0.9181	0.9297	0.9245	0.9322	0.9492	0.9637

TABLE 2 | Top 10 candidate microbes related to asthma.

Rank	Microbe	Evidence
1	Geotrichum sp.	PMID: 9376049
2	Staphylococcus	PMID: 24117882
3	Clostridiaceae	PMID: 30600099
4	Firmicutes	PMID: 32072252
5	Mitsuokella	Unconfirmed
6	Staphylococcus aureus	PMID: 30193937
7	Human parainfluenza virus 1	PMID: 26481737
8	Sphingobacteriia	Unconfirmed
9	Cytomegalovirus	PMID: 33757721
10	Aeromonas hydrophila	Unconfirmed

TABLE 3 | Top 10 candidate microbes related to IBD.

Rank	Microbe	Evidence
1	Holophagae	Unconfirmed
2	Fusobacterium	PMID: 31240835
3	Sneathia sanguinegens	Unconfirmed
4	Mycobacterium sp.	PMID: 33924414
5	Anelloviridae	PMID: 32406906
6	Butyrate-producing bacterium	PMID: 33802759
7	Firmicutes	PMID: 25420450
8	ascomycota	PMID: 26843508
9	Human herpesvirus 6	PMID: 21879802
10	Nitrososphaeraceae	Unconfirmed

TABLE 4 | Top 10 candidate microbes related to COVID-19.

Rank	Microbe	Evidence
1	Dyella	Unconfirmed
2	Acinetobacter calcoaceticus	Unconfirmed
3	Coriobacteriaceae bacterium	Unconfirmed
4	Bacteroides intestinalis	Unconfirmed
5	Bacteroides thetaiotaomicron	PMID: 32442562
6	Pisolithaceae	Unconfirmed
7	Pigmentiphaga	Unconfirmed
8	Mucor	PMID: 34009676
9	Prevotella disiens	PMID: 33577896
10	Blumeria graminis	Unconfirmed

been 3 years since its emergence and has become a pandemic threat to human health and the world economy. Although most cases of COVID-19 are mild or moderate, 3–4% of patients may be severe or critical, leading to hospitalization, respiratory failure, or death (Shen et al., 2020; Taleghani and Taghipour, 2021). Recent studies have found significant changes in the gut microbiome after infection with SARS-CoV-2. Therefore, this study conducts a case study on COVID-19 to evaluate the predictive ability of the model for COVID-19-related microbes, thereby helping researchers to conduct experimental verification purposefully, thus saving manpower and material resources. The results are presented in **Tables 3**, **4**, of the top 10 candidate microbes were verified by relevant literature. For example, the analysis of fecal samples from COVID-19 patients found that the populations of *Bacteroides dorei*, *Bacteroides thetaiotaomicron* (5th), *Bacteroides massiliensis*, and *Bacteroides ovatus* were negatively associated with SARS-CoV-2 viral load in the samples (Zuo et al., 2020). Mycological analysis revealed that 77.8 and 30.6% of patients were infected with *Mucor* (8th) and *Aspergillus*, respectively (El-Kholy and El-Fattah, 2021). *Staphylococcus haemolyticus*, *Prevotella disiens* (9th), and 2 *Corynebacterium*_1 unclassified amplicon sequence variants were more abundant in people with low SARS-CoV-2 viral load during COVID-19 infection (Rosas-Salazar et al., 2021).

CONCLUSION

Increasing studies have shown that microbes play a key role in human health and disease. Microbe-disease associations cannot only reveal disease pathogenesis, but also promote the diagnosis and prognosis of diseases. Therefore, research on microbe-disease associations has attracted wide attention. In this study, we propose a novel computational model, called MATMNMDA, to predict potential microbe-disease associations. In order to capture more semantic and structural information between microbe nodes and disease nodes, we introduce drugs to construct a tripartite heterogeneous network and apply MAGNN to learn low-dimensional embedded representations of microbe nodes and disease nodes. For each layer of MAGNN, we use intra-metapath aggregation to get the representation of the target node in each metapath, which is the input of inter-metapath aggregation layer. Then we aggregate the embedding representations between different metapaths related to the target node. Therefore, we can learn the embedding representation for the target node (microbe node or disease node) of the layer. Finally, we obtain vector representations of microbes and diseases based on the output of the last layer in the MAGNN, which is used for the prediction task. We designed multiple experiments to verify the effectiveness of the MATMNMDA model. By analyzing the experimental results, we found that: (1) Compared to the variants of our model, our model obtains the best prediction performance, which also indicates that our method could be better applied to microbe-disease prediction. (2) Under the same conditions, compared to the state-of-the-art methods, our method also obtains the best AUC and AUPR, which indicates that the MATMNMDA model can better identify potential disease-related microbes. (3) Compared to the state-of-the-art methods on different datasets, MATMNMDA achieves the best prediction performance on our enlarged dataset. It demonstrates that more known microbe-disease associations can help MATHMDA improve predictive performance. (4) Case studies on asthma, IBD, and COVID-19 further verified the effectiveness of MATMNMDA.

In future work, we will add more relational data, such as drug-drug interactions, drug-protein interactions, and protein-disease associations to achieve better predictive results.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

XL conceived, designed and managed the study, analyzed the results, and revised the manuscript. YC conducted the experiments, analyzed the results, and wrote the manuscript. Both authors read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (61972451 and 61902230).

REFERENCES

Akanksha, R., Anamika, T., Shivangi, S., and Manoj, K. (2017). A biofilm: a resource of anti-biofilm agents and their potential implications in targeting antibiotic drug resistance. *Nucleic Acids Res.* 46, D894–D900. doi: 10.1093/nar/gkx 1157

Andersen, P. I., Ianevski, A., Lysvand, H., Vitkauskiene, A., and Kainov, D. E. (2020). Discovery and development of safe-in-man broad-spectrum antiviral agents. *Int. J. Infect. Dis.* 93, 268–276. doi: 10.1016/j.ijid.2020.02.018

Arrieta, M., Stiemsma, L., Dimitriu, P., Thorson, L., Russell, S., Yurist-Doutsch, S., et al. (2015). Early infancy microbial and metabolic alterations affect risk of childhood asthma. *Sci. Transl. Med.* 7:307ra152. doi: 10.1126/scitranslmed.aab2271

Bian, C., Lei, X., and Wu, F. (2021). GATCDA: Predicting circRNA-Disease Associations Based on Graph Attention Network. *Cancers* 13:2595. doi: 10.3390/cancers13112595

Chen, L., Zheng, D., Liu, B., Jian, Y., and Jin, Q. (2016). VFDB 2016: hierarchical and refined dataset for big data analysis—10 years on. *Nucleic Acids Res.* 44, D694–D697. doi: 10.1093/nar/gkv1239

Chen, X., Huang, Y., You, Z., Yan, G., and Wang, X. (2016). A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases. *Bioinformatics* 33, 733–739.

Choi, S., Sohn, K. H., Jung, J. W., Kang, M. G., Yang, M. S., Kim, S., et al. (2021). Lung virome: new potential biomarkers for asthma severity and exacerbation. *J. Allergy Clin. Immun.* 148, 1007–1015. doi: 10.1016/j.jaci.2021.03.017

Davis, A. P., Murphy, C. G., Johnson, R., Lay, J. M., Lennon-Hopkins, K., Saraceni-Richards, C., et al. (2012). The Comparative Toxicogenomics Database: update 2013. *Nucleic Acids Res.* 41, D1104–D1114. doi: 10.1093/nar/gks994

El-Kholy, N. A., and El-Fattah, A. M. A. (2021). Invasive Fungal Sinusitis in Post COVID-19 Patients: A New Clinical Entity. *Laryngoscope* 131, 2652–2658. doi: 10.1002/lary.29632

Fan, C., Lei, X., and Guo, L. (2018). Predicting the associations between microbes and diseases by integrating multiple data sources and path-based HeteSim scores. *Neurocomputing* 323, 76–85.

Gasaly, N., Hermoso, M. A., and Gotteland, M. (2021). Butyrate and the Fine-Tuning of Colonic Homeostasis: Implication for Inflammatory Bowel Diseases. *Int. J. Mol. Sci*. 22:3061. doi: 10.3390/ijms22063061

Gill, S., Pop, M., DeBoy, R., Eckburg, P., Turnbaugh, P., Samuel, B., et al. (2006). Metagenomic analysis of the human distal gut microbiome. *Science* 312, 1355–1359. doi: 10.1126/science.1124234

He, B., Peng, L., and Li, Z. (2018). Human Microbe-Disease Association Prediction With Graph Regularized Non-Negative Matrix Factorization. *Front. Microbiol.* 9:2560. doi: 10.3389/fmicb.2018.02560

Huang, Z., Chen, X., Zhu, Z., Liu, H., and Wen, Z. (2017). PBHMDA: Path-Based Human Microbe-Disease Association prediction. *Front. Microbiol.* 8:233. doi: 10.3389/fmicb.2017.00233

Hufnagl, K., Pali-Schöll, I., Roth-Walter, F., and Jensen-Jarolim, E. (2020). Dysbiosis of the gut and lung microbiome has a role in asthma. *Semin. Immunopathol.* 42, 75–93. doi: 10.1007/s00281-019-00775-y

Human Microbiome Project (HMP), C. (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

Kong, W., Cui, Q., Zhou, X., Ma, Z., and Lu, Y. (2017). An analysis of human microbe-disease associations. *Brief. Bioinformatics.* 18, 85–97.

Lee, Y., and Kim, J. (2021). Urine Microbe-Derived Extracellular Vesicles in Children With Asthma. *Allergy Asthma Immunol. Res.*13, 75–87. doi: 10.4168/aair.2021.13.1.75

Lei, X., Bian, C., and Pan, Y. (2021). Predicting CircRNA-disease associations based on improved weighted biased meta-structure. *J. Comput. Sci. Technol.* 36, 288–298.

Lei, X., and Wang, Y. (2020). Predicting Microbe-Disease Association by Learning Graph Representations and Rule-Based Inference on the Heterogeneous Network. *Front. Microbiol.* 11:579. doi: 10.3389/fmicb.2020.00579

Liang, G., Conrad, M., Kelsen, J., Kessler, L., Breton, J., Albenberg, L., et al. (2020). Dynamics of the Stool Virome in Very Early-Onset Inflammatory Bowel Disease. *J. Crohns Colitis* 14, 1600–1610. doi: 10.1093/ecco-jcc/jjaa094

Liu, W., Jiang, Y., Peng, L., Sun, X., Gan, W., Zhao, Q., et al. (2022a). Inferring gene regulatory networks using the improved Markov blanket discovery algorithm. *Interdiscip. Sci.* 14, 168–181. doi: 10.1007/s12539-021-00478-9

Liu, W., Lin, H., Huang, L., Peng, L., Tang, T., Zhao, Q., et al. (2022b). Identification of miRNA–disease associations via deep forest ensemble learning based on autoencoder. *Brief. Bioinform.* [Epub ahead of print]. doi: 10.1093/bib/bbac104

Luo, J., and Long, Y. (2018). NTSHMDA: Prediction of Human Microbe-Disease Association based on Random Walk by Integrating Network Topological Similarity. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 17, 1341–1351. doi: 10.1109/TCBB.2018.2883041

Ma, Z., Kuang, Z., and Deng, L. (2021). CRPGCN: predicting circRNA-disease associations using graph convolutional network based on heterogeneous network. *BMC Bioinform.* 22:551. doi: 10.1186/s12859-021-04467-z

Marco, M. L., Heeney, D., Binda, S., Cifelli, C. J., Cotter, P. D., Foligne, B., et al. (2017). Health benefits of fermented foods: microbiota and beyond. *Curr. Opin. Biotech.* 44, 94–102. doi: 10.1016/j.copbio.2016.11.010

Matsuoka, K., and Kanai, T. (2015). The gut microbiota and inflammatory bowel disease. *Semin. Immunopathol.* 37, 47–55.

Olejniczak-Staruch, I., Ciążyńska, M., Sobolewska-Sztychny, D., Narbutt, J., Skibińska, M., and Lesiak, A. (2021). Alterations of the Skin and Gut Microbiome in Psoriasis and Psoriatic Arthritis. *Int. J. Mol. Sci.* 22:3998. doi: 10.3390/ijms22083998

Pan, Y., Lei, X., and Zhang, Y. (2022). Association predictions of genomics, proteinomics, transcriptomics, microbiome, metabolomics, pathomics, radiomics, drug, symptoms, environment factor, and disease networks: a comprehensive approach. *Med. Res. Rev.* 42, 441–461. doi: 10.1002/med.21847

Phaisangittisagul, E. (2016). "An Analysis of the Regularization Between L2 and Dropout in Single Hidden Layer Neural Network," in *International Conference on Intelligent Systems* (Bangkok: IEEE).

Rosas-Salazar, C., Kimura, K. S., Shilts, M. H., Strickland, B. A., Freeman, M. H., Wessinger, B. C., et al. (2021). SARS-CoV-2 infection and viral load are associated with the upper respiratory tract microbiome. *J. Allergy Clin. Immunol.* 147, 1226–1233.e2. doi: 10.1016/j.jaci.2021.02.001

Sender, R., Fuchs, S., and Milo, R. (2016). Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biol.* 14:e1002533. doi: 10.1371/journal.pbio.1002533

Shen, Y., Zheng, F., Sun, D., Ling, Y., and Chen, J. (2020). Epidemiology and clinical course of COVID-19 in Shanghai, China. *Emerg. Microbes Infect.* 9, 1537–1545. doi: 10.1080/22221751.2020.1787103

Shen, Z., Jiang, Z., and Bao, W. (2017). "CMFHMDA: Collaborative Matrix Factorization for Human Microbe-Disease Association Prediction," in *International Conference on Intelligent Computing* (Cham: Springer). doi: 10.3389/fmicb.2022.834982

Sipponen, T., Turunen, U., Lautenschlager, I., Nieminen, U., Arola, J., and Halme, L. (2011). Human herpesvirus 6 and cytomegalovirus in ileocolonic mucosa in inflammatory bowel disease. *Scand. J. Gastroenterol.* 46, 1324–1333. doi: 10.3109/00365521.2011.605466

Skoufos, G., Kardaras, F., Alexiou, A., Kavakiotis, I., and Hatzigeorgiou, A. (2020). Peryton: a manual collection of experimentally supported microbe-disease associations. *Nucleic Acids Res.* 49, D1328–D1333. doi: 10.1093/nar/gkaa902

Sokol, H., Leducq, V., Aschard, H., Pham, H., Jegou, S., Landman, C., et al. (2017). Fungal microbiota dysbiosis in IBD. *Gut* 66, 1039–1048. doi: 10.1136/gutjnl-2015-310746

Sun, Y., Zhang, D., Cai, S., Ming, Z., and Li, J. (2018). MDAD: A Special Resource for Microbe-Drug Associations. *Front. Cell. Infect. Microbiol.* 8:424. doi: 10.3389/fcimb.2018.00424

Sun, Z., Deng, Z., Nie, J., and Tang, J. (2019). "*RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space," in *7th International Conference on Learning Representations* (New Orleans, Louisiana, United States: ICLR 2019).

Taleghani, N., and Taghipour, F. (2021). Diagnosis of COVID-19 for controlling the pandemic: a review of the state-of-the-art. *Biosens. Bioelectron.* 174:112830. doi: 10.1016/j.bios.2020.112830

Tomassen, P., Jarvis, D., Newson, R., Van Ree, R., Forsberg, B., Howarth, P., et al. (2013). Staphylococcus aureus enterotoxin-specific IgE is associated with asthma in the general population: a GA(2)LEN study. *Allergy* 68, 1289–1297. doi: 10.1111/all.12230

Vasily, V. (2017). Global, regional, and national deaths, prevalence, disability-adjusted life years, and years lived with disability for chronic obstructive pulmonary disease and asthma, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet Respir. Med.* 5, 691–706. doi: 10.1016/S2213-2600(17)30293-X

Wang, C., Han, C., Zhao, Q., and Chen, X. (2021). Circular RNAs and complex diseases: from experimental results to computational models. *Brief. Bioinform.* 22:bbab286. doi: 10.1093/bib/bbab286

Wang, F., Huang, Z., Chen, X., Zhu, Z., Wen, Z., and Zhao, J. (2017). LRLSHMDA: Laplacian Regularized Least Squares for Human Microbe–Disease Association prediction. *Sci. Rep.* 7:7601. doi: 10.1038/s41598-017-08127-2

Wang, F., Zhao, L., Zhu, R., Deng, J., Sun, Y., Ding, Y., et al. (2015). Parainfluenza Virus Types 1, 2, and 3 in Pediatric Patients with Acute Respiratory Infections in Beijing During 2004 to 2012. *Chin. Med. J.* 128, 2726–2730. doi: 10.4103/0366-6999.167297

Wang, L., Tan, Y., Yang, X., Kuang, L., and Ping, P. (2022). Review on predicting pairwise relationships between human microbes, drugs and diseases: from biological data to computational models. *Brief. Bioinform.* [Epub ahead of print]. doi: 10.1093/bib/bbac080

Wang, T., Cai, G., Qiu, Y., Fei, N., Zhang, M., Pang, X., et al. (2012). Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. *ISME J.* 6, 320–329. doi: 10.1038/ismej.2011.109

Wang, Y., Lei, X., and Pan, Y. (2022). Predicting Microbe-disease Association Based on Heterogeneous Network and Global Graph Feature Learning. *Chin. J. Electron.* 31, 1–9.

Wen, Z., Yan, C., Duan, G., Li, S., Wu, F., and Wang, J. (2021). A survey on predicting microbe-disease associations: biological data and computational methods. *Brief. Bioinform.* 22:bbaa157. doi: 10.1093/bib/bbaa157

Weng, Y., Gan, H., Li, X., Huang, Y., Li, Z., Deng, H., et al. (2019). Correlation of diet, microbiota and metabolite networks in inflammatory bowel disease. *J. Dig. Dis.* 20, 447–459. doi: 10.1111/1751-2980.12795

Wishart, D. S., Feunang, Y. D., Guo, C. A., Lo, E. J., and Wilson, M. (2017). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 46, D1074–D1082. doi: 10.1093/nar/gkx1037

Wu, C., Gao, R., Zhang, D., Han, S., and Zhang, Y. (2018). PRWHMDA: Human Microbe-Disease Association Prediction by Random Walk on the Heterogeneous Network with PSO. *Int. J. Biol. Sci.* 14, 849–857. doi: 10.7150/ijbs.24539

Yan, C., Duan, G., Wu, F., Pan, Y., and Wang, J. (2019). BRWMDA:Predicting microbe-disease associations based on similarities and bi-random walk on disease and microbe networks. *IEEE ACM Trans. Comput. Biol. Bioinform*. 17, 1595–1604. doi: 10.1109/TCBB.2019.2907626

Yang, L., Li, L., and Yi, H. (2022). DeepWalk based method to predict lncRNA-miRNA associations via lncRNA-miRNA-disease-protein-drug graph. *BMC Bioinform.* 22:621. doi: 10.1186/s12859-022-04579-0

Yao, G., Zhang, W., Yang, M., Yang, H., and Li, W. (2021). MicroPhenoDB Associates Metagenomic Data with Pathogenic Microbes, Microbial Core Genes, and Human Disease Phenotypes. *Genom. Proteom. Bioinform*. 18, 760–772. doi: 10.1016/j.gpb.2020.11.001

Yorick, J., Joachim, N., Antoon, B., Nathan, D., Frederick, V., Evelien, W., et al. (2018). Disbiome database: linking the microbiome to disease. *BMC Microbiol.* 18:50. doi: 10.1186/s12866-018-1197-5

Zhang, L., Yang, P., Feng, H., Zhao, Q., and Liu, H. (2021). Using Network Distance Analysis to Predict lncRNA–miRNA Interactions. *Interdiscip. Sci.* 13, 535–545. doi: 10.1007/s12539-021-00458-z

Zhang, Y., Lei, X., Fang, Z., and Pan, Y. (2020). CircRNA-disease associations prediction based on metapath2vec++ and matrix factorization. *Big Data Mining Anal.* 3, 280–291.

Zhou, Y., Jackson, D., Bacharier, L. B., Mauger, D., Boushey, H., Castro, M., et al. (2019). The upper-airway microbiota and loss of asthma control among asthmatic children. *Nat. Commun.* 10:5714. doi: 10.1038/s41467-019-13698-x

Zhu, S., Bing, J., Min, X., Lin, C., and Zeng, X. (2018). Prediction of Drug-Gene Interaction by Using Metapath2vec. *Front. Genet.* 9:248. doi: 10.3389/fgene.2018.00248

Zimmermann, M., Patil, K. R., Typas, A., and Maier, L. (2021). Towards a mechanistic understanding of reciprocal drug-microbiome interactions. *Mol. Syst. Biol.* 17:e10116. doi: 10.15252/msb.202010116

Zimmermann, M., Zimmermann-Kogadeeva, M., Wegmann, R., and Goodman, A. L. (2019). Mapping human microbiome drug metabolism by gut bacteria and their genes. *Nature* 570:1. doi: 10.1038/s41586-019-1291-3

Zimmermann, P., Messina, N., Mohn, W. W., Finlay, B. B., and Curtis, N. (2019). Association between the intestinal microbiota and allergic sensitization, eczema, and asthma: a systematic review. *J. Allergy Clin. Immunol.* 143, 467–485. doi: 10.1016/j.jaci.2018.09.025

Zou, S., Zhang, J., and Zhang, Z. (2018). Novel human microbe-disease associations inference based on network consistency projection. *Sci. Rep.* 8:8034. doi: 10.1038/s41598-018-26448-8

Zuo, T., Zhang, F., Lui, G. C. Y., Yeoh, Y. K., Li, A. Y. L., Zhan, H., et al. (2020). Alterations in Gut Microbiota of Patients With COVID-19 During Time of Hospitalization. *Gastroenterology* 159, 944–955.e8. doi: 10.1053/j.gastro.2020.05.048

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Microbiology

ORIGINAL RESEARCH
published: 06 June 2022
doi: 10.3389/fmicb.2022.912145

Disease-Ligand Identification Based on Flexible Neural Tree

Bin Yang[1], Wenzheng Bao[2] and Baitong Chen[3]*

[1] School of Information Science and Engineering, Zaozhuang University, Zaozhuang, China, [2] School of Information and Electrical Engineering, Xuzhou University of Technology, Xuzhou, China, [3] Xuzhou No.1 People's Hospital, Xuzhou, China

In order to screen the disease-related compounds of a traditional Chinese medicine prescription in network pharmacology research accurately, a new virtual screening method based on flexible neural tree (FNT) model, hybrid evolutionary method and negative sample selection algorithm is proposed. A novel hybrid evolutionary algorithm based on the Grammar-guided genetic programming and salp swarm algorithm is proposed to infer the optimal FNT. According to hypertension, diabetes, and Corona Virus Disease 2019, disease-related compounds are collected from the up-to-date literatures. The unrelated compounds are chosen by negative sample selection algorithm. ECFP6, MACCS, Macrocycle, and RDKit are utilized to numerically characterize the chemical structure of each compound collected, respectively. The experiment results show that our proposed method performs better than classical classifiers [Support Vector Machine (SVM), random forest (RF), AdaBoost, decision tree (DT), Gradient Boosting Decision Tree (GBDT), KNN, logic regression (LR), and Naive Bayes (NB)], up-to-date classifier (gcForest), and deep learning method (forgeNet) in terms of AUC, ROC, TPR, FPR, Precision, Specificity, and F1. MACCS method is suitable for the maximum number of classifiers. All methods perform poorly with ECFP6 molecular descriptor.

Keywords: virtual screening, network pharmacology, flexible neural tree, grammar-guided genetic programming, salp swarm algorithm

OPEN ACCESS

Edited by:
Liang Wang,
Xuzhou Medical University, China

Reviewed by:
Chun-Chun Wang,
China University of Mining and Technology, China
Chandrabose Selvaraj,
Alagappa University, India

***Correspondence:**
Wenzheng Bao
baowz55555@126.com

Specialty section:
This article was submitted to Microbe and Virus Interactions with Plants, a section of the journal Frontiers in Microbiology

Received: *15 March 2022*
Accepted: *06 May 2022*
Published: *06 June 2022*

Citation:
Yang B, Bao W and Chen B (2022) Disease-Ligand Identification Based on Flexible Neural Tree. Front. Microbiol. 13:912145. doi: 10.3389/fmicb.2022.912145

INTRODUCTION

Computer-aided drug design (CADD) has gradually become an indispensable emerging technology in the research and development of a new drug (Leelananda and Steffen, 2016; Tong et al., 2019; Maia et al., 2020). CADD technology reduces the capital, time, and labor cost of drug development and greatly improves the efficiency of the research and development of new drug (Gomeni et al., 2001). Virtual screening is one of the important comprehensive technical means in CADD, which is a process of discovering new ligands on the basis of biological structure based on the computer methods (Guasch et al., 2016; Olubiyi et al., 2020; Rajguru et al., 2020). It is a new technology and method for innovative drug research. By using the high-speed computing of computer, a small number of potential active compounds are screened from a large number of candidate compounds, so as to greatly reduce the blindness of subsequent experimental verification. In the future, virtual screening technology will become an important means to explore the relevant biochemical space because of its many advantages, such as high efficiency, high speed, low cost, and so on (Zaslavskiy et al., 2019; Guo et al., 2021; Maddah et al., 2021; Selvaraj et al., 2021; Yang et al., 2021).

In the past decade, virtual screening has been applied to the medical and the pharmaceutical researches widely (Meng et al., 2011; Bajusz et al., 2017). The most commonly used virtual screening method is molecular docking, and the software involved contains AutoDock, SLIDE, DOCK,

Flex X, etc. (Morris et al., 1996; Kellenberger et al., 2004; Taufer et al., 2005). Fischer et al. (2021) utilized virtual screening method to screen 25, 56, 750 compounds in order to make the analysis about the binding of small molecules to translationally controlled tumor protein. Baxter et al. (2000) utilized molecular docking to screen ligand-receptor complexes in virtual database and tabu search method was utilized to assist this work. Talluri (2021) utilized Vina and SMINA to make molecular docking to predict potential drugs for the treatment of Corona Virus Disease 2019 (COVID-19). Zhou et al. (2016) screened the compounds of Chicory, which were bundled with concentrated nucleoside transporter 2 (CNT2) in order to validate that CNT2 as the potential target of chicory could reduce the absorption of purine nucleosides in the intestine. Meenakumari et al. (2019) made docking analysis between 17 coumarin derivatives and carbonic anhydrase IX (CAIX) to screen the ligands. Thiyagarajan et al. (2016) made molecular docking between the 3D structures of focal adhesion kinase and S6 kinase and 60 natural compounds to obtain the new specific inhibitors, and the findings could provide help for the treatment of tumorigenesis and metastasis.

In order to improve the time and accuracy of virtual screening, some machine learning methods have been utilized to assist or replace molecular docking (Berishvili et al., 2018; Zaki et al., 2021). Wang et al. (2016) proposed a new virtual screening based on ensemble learning and SVM to tackle with protein-ligand in action fingerprint. Zhang Y. et al. (2019) investigated the performances of 8 classifiers containing decision tree (DT), KNN, SVM, random forest (RF), extremely randomizer tree, AdaBoost, gradient boosting tree, and XGBoost with ACC inhibitor data for the researches of drug design and discovery. Zhang et al. (2017) proposed a new scoring function based on machine learning to screen the compounds targeting the viral neuraminidase protein so as to make anti-influenza therapy. Chen et al. (2011) proposed a ligand screening algorithm based SVM to discovery lead compounds. Bustamam et al. (2021) proposed a dipeptidyl peptidase-4 (DPP-4) inhibitors identification method based on Rotation Forest and Deep Neural Network with the fingerprint datasets for the treatment of type 2 diabetes mellitus. Zheng et al. (2020) utilized Naïve Bayesian and recursive partitioning to select the important active chemical components from many compounds in Xiaoshuan Tongluo formula with ECFP_6 and MACCS feature sets for treating stroke.

Virtual screening of disease-related compounds can narrow the scope of analysis in network pharmacology research. In this paper a new virtual screening method based on flexible neural tree (FNT) model is proposed to screen the disease-related active compounds. A novel hybrid evolutionary algorithm based on Grammar-guided genetic programming and salp swarm algorithm is proposed to infer the structure and parameters in each FNT model. The 3 diseases (hypertension, diabetes, and COVID-19) related compounds are searched from the up-to-date literatures. The unrelated compounds are selected by negative sample selection algorithm from DUD-E website. About 4 kinds of molecular descriptors (ECFP6, MACCS, Macrocycle, and RDKit) are utilized to numerically characterize the chemical structures of related and unrelated compounds of diseases, respectively. We make the investigation about the performances of these 4 molecular descriptors.

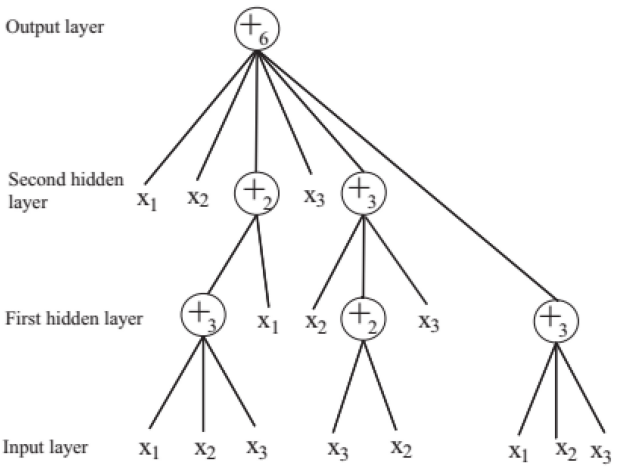

FIGURE 1 | An example of flexible neural tree.

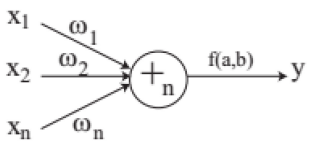

FIGURE 2 | A flexible neuron operator.

MATERIALS AND METHODS

Flexible Neural Tree Model

In order to solve the automatic design problem of artificial neural network, FNT was proposed, which is a hierarchical, multilayer, and irregular artificial neural network (Chen et al., 2012). FNT can transform a single and fixed neural network model into a special tree model that can change flexibly between various levels. It could overcome the difficulty of structural optimization of common neural network, have strong adaptive ability for various classification and prediction problems, and obtain high classification and prediction accuracy. In this paper, FNT is proposed to predict active disease-related compounds. An example of structure of FNT model is showed in **Figure 1**. AFNT includes input layer, several hidden layers and output layer. The nodes in the input layer are created randomly from terminal set $T = \{x_1, x_2, \ldots, x_n\}$. The nodes in the hidden layers are selected randomly from terminal set and operator set $F = \{+_2, +_3, \ldots, +_n\}$. The output layer contains one node.

In FNT, each layer is randomly generated according to the operation set and terminal set. The maximum depth of tree is set in advance. If an operator instruction $+_n$ is selected, n branches are created randomly from set T and F, which are terminal variables and operators. And n weights are generated randomly. If a terminal variable is selected, the corresponding branch is

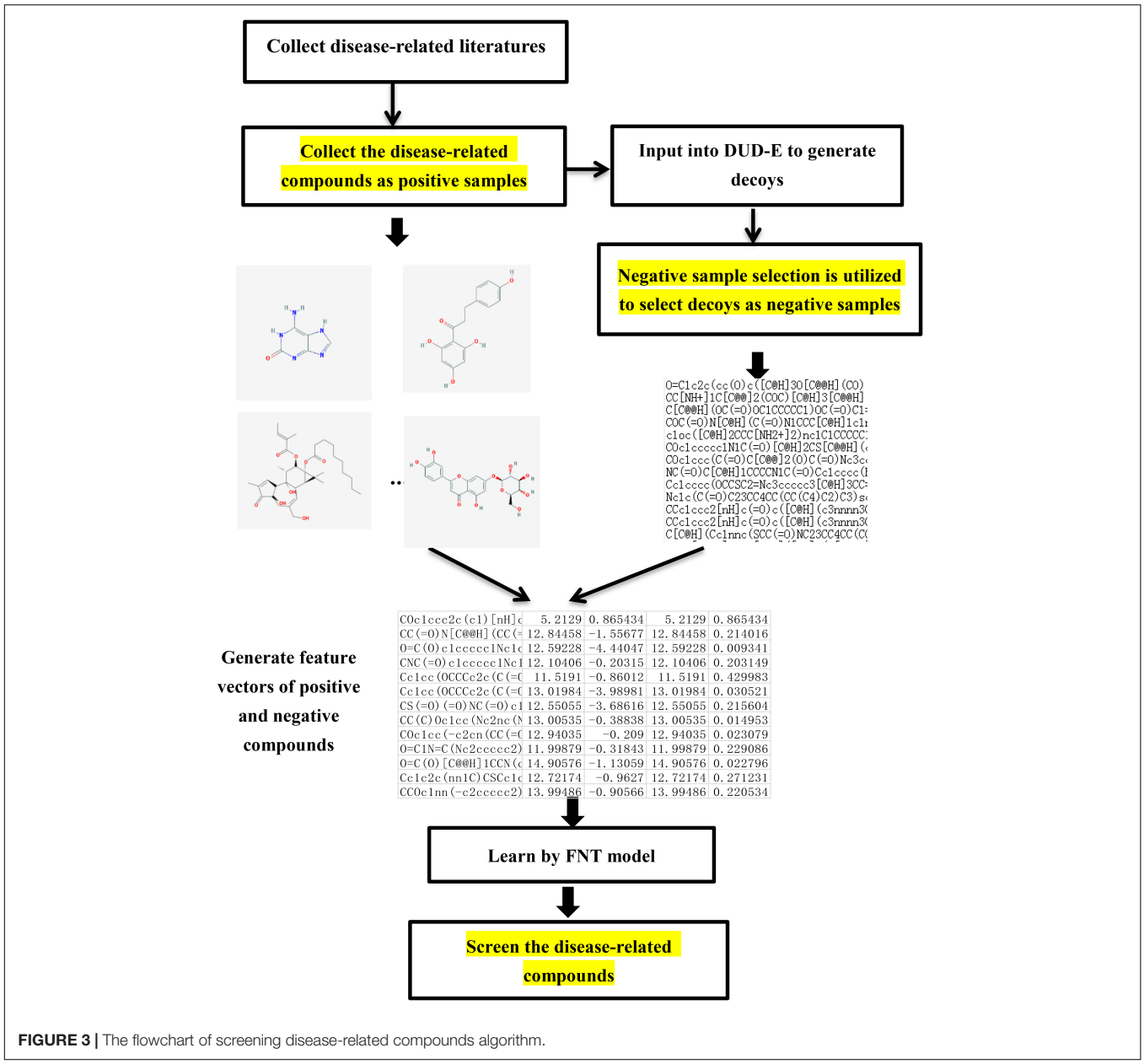

FIGURE 3 | The flowchart of screening disease-related compounds algorithm.

terminated. When FNT is created randomly, the depth of FNT could not exceed the maximum depth. $+_n$ is depicted in **Figure 2** and is calculated as follows.

$$net_n = \sum_{j=1}^{n} w_j x_j. \tag{1}$$

The final output of $+_n$ is calculated by activation function, which is given as follows.

$$y = f(net_n, a_n, b_n) = e^{-(\frac{net_n - a_n}{b_n})^2}. \tag{2}$$

Where a_n and b_n are parameters of activation function.

Model Optimization Algorithm

Grammar-Guided Genetic Programming

Grammar-guided genetic programming (GGGP) was proposed in order to overcome the shortcomings of genetic programming (Wu and Chen, 2007). In this paper, GGGP is utilized to search the optimal structure of FNT model. In GGGP, context-free grammar (CFG) model is utilized to guide the evolutionary process of GP in order to search the optimal solution faster.

The CFG model contains a quadruple, which is represented as $G = \{N, T, P, \sum\}$, where N is non-terminal symbol set, T is terminal symbol set, P is production rule set and $\sum$ is beginning symbol set. The 4 sets satisfy the conditions: $N \bigcap T = \phi$ and $\sum \in N$. An element in production rule set is represented as $x \to y$, where $x \in N$,

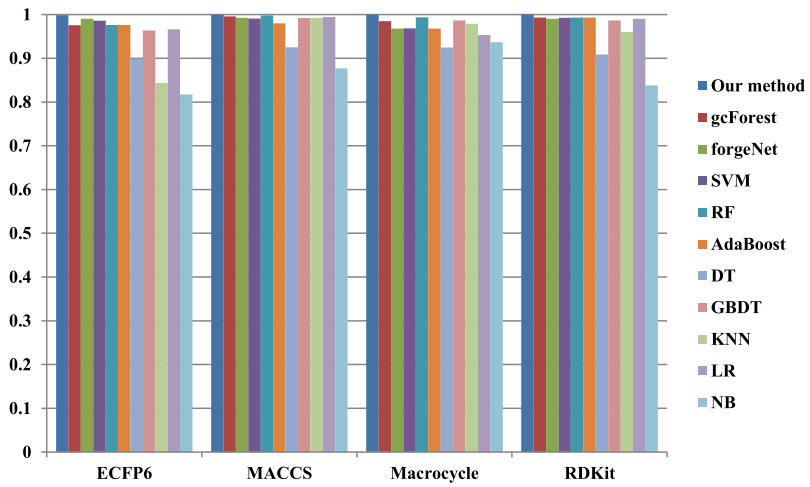

FIGURE 4 | AUC performances of 11 methods with hypertension dataset.

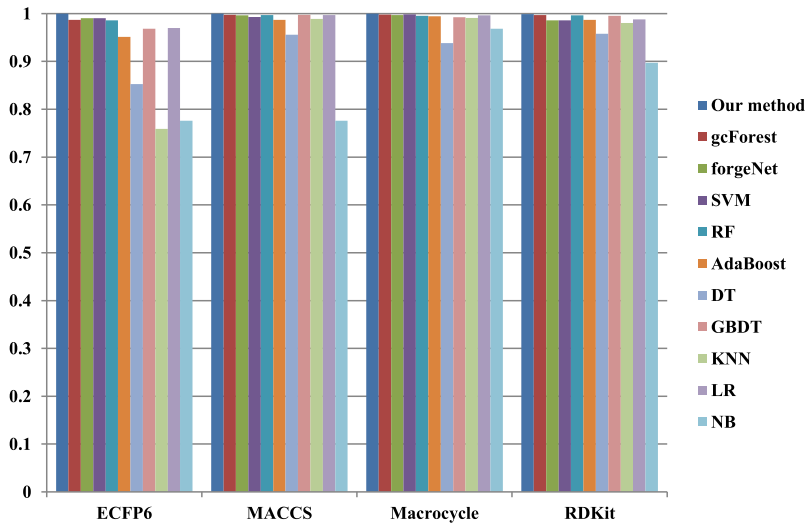

FIGURE 5 | AUC performances of 11 methods with diabetes dataset.

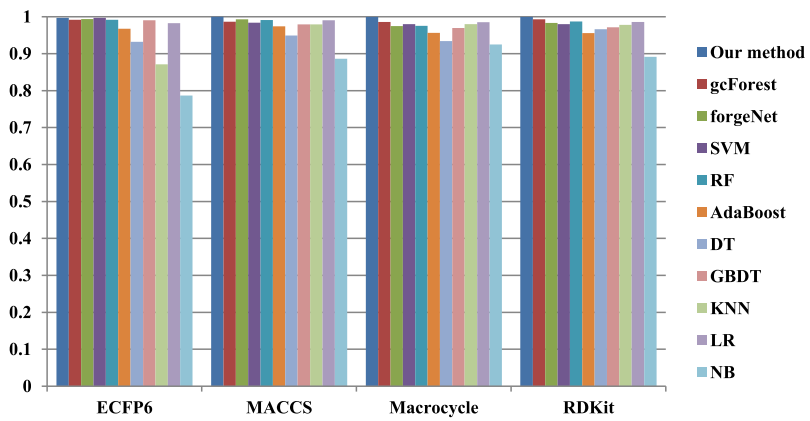

FIGURE 6 | AUC performances of 11 methods with COVID-19 dataset.

TABLE 1 | Prediction performances of 11 methods with hypertension dataset.

Molecular descriptors	Methods	TPR	FPR	Precision	Specificity	F1
ECFP6	Our method	**0.985075**	0.022222	0.956522	0.977778	**0.970588**
	gcForest	0.955224	0.155556	0.752941	0.844444	0.842105
	forgeNet	0.895522	**0**	**1**	**1**	0.944882
	SVM	0.880597	0.007407	0.983333	0.992593	0.929134
	RF	0.880597	**0**	**1**	**1**	0.936508
	AdaBoost	0.835821	0.037037	0.918033	0.962963	0.875
	DT	0.835821	0.044444	0.903226	0.955556	0.868217
	GBDT	0.850746	0.051852	0.890625	0.948148	0.870229
	KNN	0.686567	**0**	**1**	**1**	0.814159
	LR	0.970149	0.311111	0.607477	0.688889	0.747126
	NB	0.731343	0.096296	0.790323	0.903704	0.75969
MACCS	Our method	**1**	**0.007407**	**0.985294**	**0.992593**	**0.992593**
	gcForest	0.970149	0.051852	0.902778	0.948148	0.935252
	forgeNet	0.925373	0.018587	0.96124	0.981413	0.942966
	SVM	0.940299	0.02963	0.940299	0.97037	0.940299
	RF	0.940299	0.014815	0.969231	0.985185	0.954545
	AdaBoost	0.895522	0.044444	0.909091	0.955556	0.902256
	DT	0.895522	0.051852	0.895522	0.948148	0.895522
	GBDT	0.925373	0.014815	0.96875	0.985185	0.946565
	KNN	0.925373	0.02963	0.939394	0.97037	0.932331
	LR	0.970149	0.066667	0.878378	0.933333	0.921986
	NB	0.940299	0.192593	0.707865	0.807407	0.807692
Macrocycle	Our method	**0.984375**	**0**	**1**	**1**	**0.992126**
	gcForest	0.9375	0.09009	0.857143	0.90991	0.895522
	forgeNet	0.921875	0.018018	0.967213	0.981982	0.944
	SVM	0.890625	0.027027	0.95	0.972973	0.919355
	RF	0.90625	0.027027	0.95082	0.972973	0.928
	AdaBoost	0.953125	0.027027	0.953125	0.972973	0.953125
	DT	0.921875	0.072072	0.880597	0.927928	0.900763
	GBDT	0.90625	0.036036	0.935484	0.963964	0.920635
	KNN	0.921875	0.072072	0.880597	0.927928	0.900763
	LR	0.9375	0.153153	0.779221	0.846847	0.851064
	NB	0.9375	0.09009	0.857143	0.90991	0.895522
RDKit	Our method	**0.985075**	**0**	**1**	**1**	**0.992481**
	gcForest	0.955224	0.02963	0.941176	0.97037	0.948148
	forgeNet	0.895522	0.022222	0.952381	0.977778	0.923077
	SVM	0.940299	0.014815	0.969231	0.985185	0.954545
	RF	0.865672	0.014815	0.966667	0.985185	0.913386
	AdaBoost	0.925373	0.014815	0.96875	0.985185	0.946565
	DT	0.873134	0.055762	0.886364	0.944238	0.879699
	GBDT	0.895522	0.02963	0.9375	0.97037	0.916031
	KNN	0.865672	0.044444	0.90625	0.955556	0.885496
	LR	0.955224	0.02963	0.941176	0.97037	0.948148
	NB	0.895522	0.214815	0.674157	0.785185	0.769231

Bold values denote the best performances.

and $y \in N \bigcup T$. Assuming that terminal set and operator set are set as $T = \{x_1, x_2, \ldots, x_n\}$, and $F = \{+_2, +_3\}$, 4 sets of CFG model are defined:$N = \{s, exp, var, op2, op3\}$, $T = \{+_2, +_3, x_1, x_2, \ldots, x_n\}$, $\sum = \{s\}$, and P is represented with Eq. (3) or Eq. (4).

$$\begin{aligned} s &\rightarrow \exp \\ \exp &\rightarrow \exp\ op2\ \exp \\ \exp &\rightarrow op3\ \exp\ \exp\ \exp \\ \exp &\rightarrow var \\ op2 &\rightarrow +_2 \\ op3 &\rightarrow +_3 \\ var &\rightarrow x_1|x_2|\ldots|x_n \end{aligned} \tag{3}$$

$$\begin{aligned} s &\rightarrow \exp \\ \exp &\rightarrow op2\ \exp\exp \\ \exp &\rightarrow op3\ \exp\ \exp\ \exp \\ \exp &\rightarrow var \\ op2 &\rightarrow +_2 \\ op3 &\rightarrow +_3 \\ var &\rightarrow x_1|x_2|\ldots|x_n \end{aligned} \tag{4}$$

Generate the initial population randomly. When generating each individual tree, the non-terminal node S is started with. Then the subtree of each non-terminal node is derived in top-down and left-right order according to the rules of the syntax model. When all non-terminal nodes in the tree have sub-trees,

TABLE 2 | Prediction performances of 11 methods with diabetes dataset.

Molecular descriptors	Methods	TPR	FPR	Precision	Specificity	F1
ECFP6	Our method	0.991935	0.012048	0.97619	0.987952	**0.984**
	gcForest	0.967742	0.124498	0.794702	0.875502	0.872727
	forgeNet	0.916031	**0.007605**	**0.983607**	**0.992395**	0.948617
	SVM	0.935484	0.02008	0.958678	0.97992	0.946939
	RF	0.862903	0.008032	0.981651	0.991968	0.918455
	AdaBoost	0.879032	0.036145	0.923729	0.963855	0.900826
	DT	0.806452	0.100402	0.8	0.899598	0.803213
	GBDT	0.854839	0.02008	0.954955	0.97992	0.902128
	KNN	**1**	0.939759	0.346369	0.060241	0.514523
	LR	0.967742	0.15261	0.759494	0.84739	0.851064
	NB	0.604839	0.052209	0.852273	0.947791	0.707547
MACCS	Our method	**0.975806**	**0**	**1**	**1**	**0.987755**
	gcForest	**0.975806**	0.02008	0.960317	0.97992	0.968
	forgeNet	0.951613	0.024096	0.951613	0.975904	0.951613
	SVM	0.935484	0.024096	0.95082	0.975904	0.943089
	RF	0.943548	0.012048	0.975	0.987952	0.959016
	AdaBoost	0.943548	0.032129	0.936	0.967871	0.939759
	DT	0.951613	0.040161	0.921875	0.959839	0.936508
	GBDT	0.975806	0.02008	0.960317	0.97992	0.968
	KNN	0.951613	0.044177	0.914729	0.955823	0.932806
	LR	0.975806	0.02008	0.960317	0.97992	0.968
	NB	0.967742	0.417671	0.535714	0.582329	0.689655
Macrocycle	Our method	**0.991453**	**0**	**1**	**1**	**0.995708**
	gcForest	0.982906	0.028037	0.950413	0.971963	0.966387
	forgeNet	0.957265	0.009346	0.982456	0.990654	0.969697
	SVM	0.974359	0.018692	0.966102	0.981308	0.970213
	RF	0.957265	0.014019	0.973913	0.985981	0.965517
	AdaBoost	0.957265	0.018692	0.965517	0.981308	0.961373
	DT	0.91453	0.037383	0.930435	0.962617	0.922414
	GBDT	0.965812	0.046729	0.918699	0.953271	0.941667
	KNN	0.923077	0.018692	0.964286	0.981308	0.943231
	LR	0.982906	0.042056	0.927419	0.957944	0.954357
	NB	0.974359	0.042056	0.926829	0.957944	0.95
RDKit	Our method	0.959677	**0**	**1**	**1**	**0.979424**
	gcForest	0.959677	0.02008	0.959677	0.97992	0.959677
	forgeNet	**0.967742**	0.012048	0.97561	0.987952	0.97166
	SVM	0.951613	0.008032	0.983333	0.991968	0.967213
	RF	0.935484	0.012048	0.97479	0.987952	0.954733
	AdaBoost	0.943548	0.016064	0.966942	0.983936	0.955102
	DT	0.943548	0.028112	0.943548	0.971888	0.943548
	GBDT	0.943548	0.008032	0.983193	0.991968	0.962963
	KNN	0.903226	0.012048	0.973913	0.987952	0.937238
	LR	0.959677	0.024096	0.952	0.975904	0.955823
	NB	0.951613	0.204819	0.698225	0.795181	0.805461

stop the derivation process of the tree, and then judge the depth of the tree. If the depth is greater than the predefined maximum depth, the tree is considered invalid, and a tree is regenerated after deletion. If the depth is less than the maximum depth, the tree is considered and can be saved to the population. Then 3 genetic operators (replication, crossover, and mutation) are utilized to generate a new population in the iteration process.

Salp Swarm Algorithm

The Salp swarm algorithm (SSA) is a new swarm optimization algorithm proposed by Mirjalili et al. (2017). The main idea of SSA comes from simulating the group behavior of salp chain (Babaei et al., 2020; Ren et al., 2021). In this algorithm, salp chain is divided into 2 groups: leader and follower. The leader is at the head of the salp chain, and the followers are at the back of the chain. In each iteration, the leader directs the followers to move in a chain toward the food. In the process of moving, the leader makes global search, while the follower makes full local search, which greatly avoid falling into local optimization. The leader's leadership role for the followers behind will be weaker and weaker. The followers behind will not blindly move toward the leader, which could maintain the diversity of the population. Therefore, this movement mode makes the salp chain have a strong ability of global search and

local development. Because of its simple implementation, fast convergence speed, and easy computer implementation, SSA is utilized to optimize the parameters of FNT model. The SSA is given as follows in detailed.

(1) Initialize the population. Suppose that population size is m, the dimension is n, the upper bound of the search space is $X_{\max} = \{X^1_{\max}, X^2_{\max}, \ldots, X^n_{\max}\}$, the lower bound is $X_{\min} = \{X^1_{\min}, X^2_{\min}, \ldots, X^n_{\min}\}$. The positions of salp population are created randomly by the following equation.

$$X_i = rand() \times (X_{\max} - X_{\min}) + X_{\min}. \tag{5}$$

(2) Give the fitness values of population according to the fitness function defined in advanced. In the iteration process, the position of the food is not clear, so the fitness values of all individual salps are calculated and sorted. And the position of salp with the optimal fitness value is set as the current food position, which is set as $F = \{F^1, F^2, \ldots, F^n\}$.

(3) Positions of leader and followers are updated. The leader is responsible for searching food to lead the moving direction of the whole group. The position of the leader is updated as follows (Chen and Mu, 2021).

$$X^i_1 = \begin{cases} F^i + c_1 \times ((X^i_{\max} - X^i_{\min}) \times c_2 + X^i_{\min}) & c_3 \geq 0.5, \\ F^i - c_1 \times ((X^i_{\max} - X^i_{\min}) \times c_2 + X^i_{\min}) & c_3 < 0.5. \end{cases} \tag{6}$$

Where X^i_1 and F^i are the i-th positions of leader (the first salp) and food. c_2 and c_3 are random number. c_1 is the convergence factor in SSA, which could play the role of balancing global search and local development. c_1 is calculated as follows.

$$c_1 = 2e^{-(\frac{4t}{T})^2}. \tag{7}$$

Where t is the current generation and T is the maximum generation.

The positions of the followers are updated according to Newton's laws of motion, which is defined as follows.

$$X^j_i = 0.5 \times at^2 + v_0 t. \tag{8}$$

$$\begin{aligned} a &= \frac{v_{final} - v_0}{\Delta t}, \\ v_{final} &= \frac{X^j_i - X^{j-1}_i}{\Delta t}. \end{aligned} \tag{9}$$

Where a is acceleration. The difference between two adjacent iterations is 1 and $v_0 = 0$, so Eq. (8) could be defined as follows.

$$X^{'j}_i = \frac{X^j_i - X^{j-1}_i}{2}. \tag{10}$$

(4) Update the fitness values of new population and the position of food. If the end condition is satisfied, algorithm is stopped; otherwise go to step (3).

Screen Disease-Related Compounds by Our Proposed Method

Virtual screening is needed in the research of network pharmacology to select the disease-related compounds. In this paper, a novel virtual screening method based on FNT, hybrid evolutionary method and negative sample selection algorithm is proposed, which is depicted in **Figure 3**.

(1) Disease-related compound dataset collection. Search the up-to-date literatures for treating diseases according to the name of disease. By consulting these literatures with data mining method, the active compounds for the treatment of the disease are collected as the positive compound samples. In order to generate the unrelated compounds, the positive compounds are input into DUD-E database to generate the corresponding decoys, which are set as negative samples (Mysinger et al., 2012). There are too many decoys generated compared to the number of positive samples. In order to balance the proportion of positive samples and negative samples, negative sample selection based on Tanimoto index (**Algorithm 1**) is presented to choose a certain number of decoys that are quite different from the positive sample set. Tanimoto index could measure the distance between the 2 compounds, which can measure the similarity between 2 sets (Klekota et al., 2005), which can solve the relationship between 0 and 1 well. The greater Tanimoto index is, the higher the similarity of 2 sets is. The Tanimoto index of 2 sets A and B is calculated as followed.

$$T(A, B) = \frac{A \cap B}{A \cup B}. \tag{11}$$

Algorithm 1: Negative sample selection algorithm.

Input: disease-related compound set $[c_1, c_2, \ldots, c_m]$ (m is the number of compounds),
the generated decoy set $[g_1, g_2, \ldots, g_n]$ (n is the number of decoys)
Output: the selection negative compound set $[n_1, n_2, \ldots, n_{2m}]$
for $i = 1;\ i \leq n;\ i++$ do
  $sum_i = 0$;
  for $j = 1;\ j \leq m;\ j++$ do
    $T_{ij} = Tanimotoindex(g_i, c_j)$;
    $sum_i = sum_i + T_{ij}$;
  End
End
Sort the decoy set according to $[sum_1, sum_2, \ldots, sum_n]$;
Select the decoys with $2m$ smallest Tanimoto indexes as negative compound set;

(2) Screening process. The related and unrelated molecules collected are all chemical structures. To facilitate the compounds collected inputting into flexible neural tree model, 4 kinds of molecular descriptors (ECFP6, MACCS, Macrocycle, and RDKit) are utilized to numerically characterize the chemical structure of each compound (Todeschini and Consonni, 2009). ECFP6 contains 2,048 features, which denotes all possible molecular routes retrieved from the atom according to radius 3 and each bit denotes whether the special stator structure exists. MACCS contains 166 molecular characteristic sites, such as ISOTOPE, ATOMIC NO, 4M RING, and GROUP VIII. Macrocycle contains 1,613 features, which refer the information about the ring-size, sugars, and ester functional groups. RDK it contains 208 features, such as number of

TABLE 3 | Prediction performances of 11 methods with COVID-19 dataset.

Molecular descriptors	Methods	TPR	FPR	Precision	Specificity	F1
ECFP6	Our method	**0.965909**	**0**	**1**	**1**	**0.982659**
	gcForest	**0.965909**	0.101695	0.825243	0.898305	0.890052
	forgeNet	0.931818	0.00565	0.987952	0.99435	0.959064
	SVM	0.920455	0.011299	0.975904	0.988701	0.947368
	RF	0.931818	**0**	**1**	**1**	0.964706
	AdaBoost	0.896226	0.025882	0.945274	0.974118	0.920097
	DT	0.909091	0.045198	0.909091	0.954802	0.909091
	GBDT	0.886364	0.028249	0.939759	0.971751	0.912281
	KNN	0.897727	0.435028	0.50641	0.564972	0.647541
	LR	0.988636	0.214689	0.696	0.785311	0.816901
	NB	0.636364	0.062147	0.835821	0.937853	0.722581
MACCS	Our method	**1**	**0**	**1**	**1**	**1**
	gcForest	0.954545	0.011299	0.976744	0.988701	0.965517
	forgeNet	0.943182	0.008499	0.982249	0.991501	0.962319
	SVM	0.931818	0.011299	0.97619	0.988701	0.953488
	RF	0.954545	**0**	**1**	**1**	0.976744
	AdaBoost	0.886364	0.016949	0.962963	0.983051	0.923077
	DT	0.931818	0.033898	0.931818	0.966102	0.931818
	GBDT	0.931818	0.00565	0.987952	0.99435	0.959064
	KNN	0.954545	0.028249	0.94382	0.971751	0.949153
	LR	0.954545	0.016949	0.965517	0.983051	0.96
	NB	0.863636	0.090395	0.826087	0.909605	0.844444
Macrocycle	Our method	**0.965517**	**0**	**1**	**1**	**0.982456**
	gcForest	0.954023	0.006536	0.988095	0.993464	0.97076
	forgeNet	0.954023	**0**	**1**	**1**	0.976471
	SVM	0.942529	0.006536	0.987952	0.993464	0.964706
	RF	0.942529	0.006536	0.987952	0.993464	0.964706
	AdaBoost	0.954023	**0**	**1**	**1**	0.976471
	DT	0.908046	0.039216	0.929412	0.960784	0.918605
	GBDT	0.896552	0.03268	0.939759	0.96732	0.917647
	KNN	0.931034	0.019608	0.964286	0.980392	0.947368
	LR	0.954023	0.026144	0.954023	0.973856	0.954023
	NB	0.885057	0.039216	0.927711	0.960784	0.905882
RDKit	Our method	**0.965909**	**0**	**1**	**1**	**0.982659**
	gcForest	0.943182	0.022599	0.954023	0.977401	0.948571
	forgeNet	0.943182	0.011299	0.976471	0.988701	0.959538
	SVM	0.943182	0.011299	0.976471	0.988701	0.959538
	RF	0.931818	0.00565	0.987952	0.99435	0.959064
	AdaBoost	0.931818	0.016949	0.964706	0.983051	0.947977
	DT	0.943182	0.011299	0.976471	0.988701	0.959538
	GBDT	0.943182	0.011299	0.976471	0.988701	0.959538
	KNN	0.954545	0.016949	0.965517	0.983051	0.96
	LR	0.943182	0.028249	0.943182	0.971751	0.943182
	NB	0.897727	0.112994	0.79798	0.887006	0.84492

Bold values denote the best performances.

valence electros, number of radical electrons, charge information, and number of Aliphatic Carbocycles. Cross-validation method is utilized to divide the training and testing datasets to test the performance of our proposed method. With the feature vector of each compound in the training dataset as the input, flexible neural tree model is utilized to train with the feature datasets. A hybrid evolutionary method based on grammar-guided genetic programming and salp swarm algorithm is proposed to search the optimal structure and parameters of FNT model. For the unknown compounds of testing dataset, the feature vectors are used as the input of the optimal FNT model to obtain the output results. If the result is higher than 0.5, the compound is identified to be disease-related; otherwise, it is unrelated.

EXPERIMENT RESULTS AND DISCUSSION

In order to test the effectiveness of our method, the important compounds were collected, which were involved in the treatment of hypertension, diabetes, and COVID-19. The

related compounds of these 3 diseases are regarded as positive samples and the numbers of samples are 67, 124, and 88, respectively. Negative sample selection method is utilized to select the inactive compounds about hypertension, diabetes and COVID-19, and the numbers of negative samples are 134, 248, and 176, respectively. The 4 kinds of molecular descriptors (ECFP6, MACCS, Macrocycle, and RDKit) are utilized to numerically characterize related and unrelated compounds of diseases, respectively.

The 10-cross validation method is utilized to test the performance of our method. SVM (Hearst et al., 1998), RF (Breiman, 2001), AdaBoost (Collins et al., 2002), decision tree (DT) (Safavian and Landgrebe, 1991), GBDT (Zhang B. et al., 2019), KNN, logical regression (LR) (Collins et al., 2002), gc Forest (Zhou and Feng, 2017), forgeNet (Kong and Yu, 2020), and Naive Bayes (NB) (Kim et al., 2006)are also utilized to identify disease-related compounds of three diseases. In our method, operator set is set as $F = \{+_2, +_3, +_4, +_5\}$, population size is set as 30 and the maximum depth of tree is set as 5. In SVM, linear kernel function is selected. In RF, the number of trees is set as 100. In GBDT, the number of regression trees is set as 200. In DT, CART algorithm is utilized. The parameters of other algorithms are set by default. The AUC performances of 11 methods with the datasets about hypertension, diabetes, and COVID-19 are shown in **Figures 4–6**, respectively. From **Figure 4**, it could be seen that with ECFP6, Macrocycle, and RDKit methods, our method has the highest AUC performances among 11 methods. With MACCS method, the AUC values obtained by our method and RF are very close to 1.0, which are 0.999889 and 0.997772, respectively. For **Figure 5**, in terms of AUC, it could be clearly seen that our method performs best with ECFP6, MACCS, and RDKit methods. With Macrocycle feature method, our method, gcForest, and SVM could obtain the better AUC values than other 8 methods, which are 1, 0.99803, and 0.998435, respectively. By the comparison of these 3 methods, our method performs best, which show that our method is a good classifier for disease-compound identification problem. For **Figure 6**, with ECFP6molecular descriptor, our method and SVM could obtain the higher AUC values than other 9 methods, which are 0.996901 and 0.99703. With other molecular descriptors, our method could obtain the better performances, which are equal to or very close to 1.0.

TPR, FPR, Precision, Specificity, and F1 are also utilized to test the performances of 11 methods for compound identification about 3 diseases. TPR denotes the ratio of true disease-related compounds identified against all true disease-related ones. FPR denotes the ratio of disease-related compounds identified erroneously against all true disease-unrelated ones. Precision denotes the ratio of true disease-related compounds identified against all disease-related ones identified. Specificity is the ratio of true disease-unrelated compounds identified against all true disease-unrelated ones. F1 could evaluate a classifier comprehensively with Precision and Recall. TPR, FPR, Precision, Specificity, and F1performances of11 methods with the datasets about hypertension, diabetes and COVID-19 are listed in **Tables 1–3**, respectively. In **Table 1**, with ECFP6 method, our method has the highest TPR performance among 11 classifiers, which shows that our method could identify more true disease-related compounds. In terms of FPR, Precision and Specificity, forgeNet and RF perform best, which reveal that all the true disease-unrelated compounds are identified. But our method could obtain the highest F1 performance. Overall our method could obtain the more accurate identification results. With MACCS, Macrocycle, and RDKit, our method could obtain the best performances of TPR, FPR, Precision, Specificity, and F1.

In **Table 2**, with ECFP6 method, KNN has the highest TPR performance among 11 classifiers, which is 1.0. The result shows that KNN could identify all true disease-related compounds. In terms of FPR, Precision, and Specificity, forgeNet perform better than other 10 methods. But our method could also obtain the highest F1 performance. Overall our method could obtain the more accurate identification results. With MACCS and Macrocycle, our method could obtain the best performances of TPR, FPR, Precision, Specificity, and F1. With RDKit, our method performs best in terms of FPR, Precision, Specificity, and F1, while forgeNet could obtain the best TPR performance. For **Table 3**, our method performs best with 4 kinds of molecular descriptors in terms of 5 criterions. All results show that our method could predict disease-related compounds more accurately than gcForest, forgeNet, SVM, RF, AdaBoost, DT, GBDT, KNN, LR, and NB.

According to the performances of 11 methods with the datasets from 3 diseases and 4 molecular descriptors, 11 methods are ranked. For each molecular descriptor, the averaged ranking results of each method are listed in **Table 4**. From **Table 4**, we can see that our method, gcforest, forgenet, RF, GDBT, and LR perform best with MACCS feature set, while SVM and DT perform best with RDKit feature set. AdaBoost, KNN and NB perform better with Mordred feature set than the other 3 feature sets. All methods perform poorly with ECFP6 molecular descriptor. The results also show that the different molecular descriptors of compounds are suitable for the different classifiers and the ranking results can provide the guidance for each classifier to choose the appropriate molecular descriptor to solve the problem in the future. On the whole, MACCS method is suitable for the maximum number of classifiers. In future research, MACCS method can be preferred for a new classifier.

TABLE 4 | Averaged ranking scores of 11 methods with 3 datasets.

	ECFP6	MACCS	Macrocycle	RDKit
Our method	3.33	**1.67**	2	2.67
gcForest	3.67	**1.83**	2.33	2.17
forgeNet	2.5	**2.17**	2.33	3
SVM	2.83	2.5	2.5	**2.17**
RF	2.83	**1.33**	2.5	3.17
AdaBoost	3.5	2.5	**1.83**	2.17
DT	4	1.83	2.5	**1.67**
GBDT	3.5	**1.33**	2.83	2.33
KNN	4	1.83	**1.67**	2.5
LR	3.83	**1.17**	2.83	2.17
NB	3.67	2.83	**1**	2.33

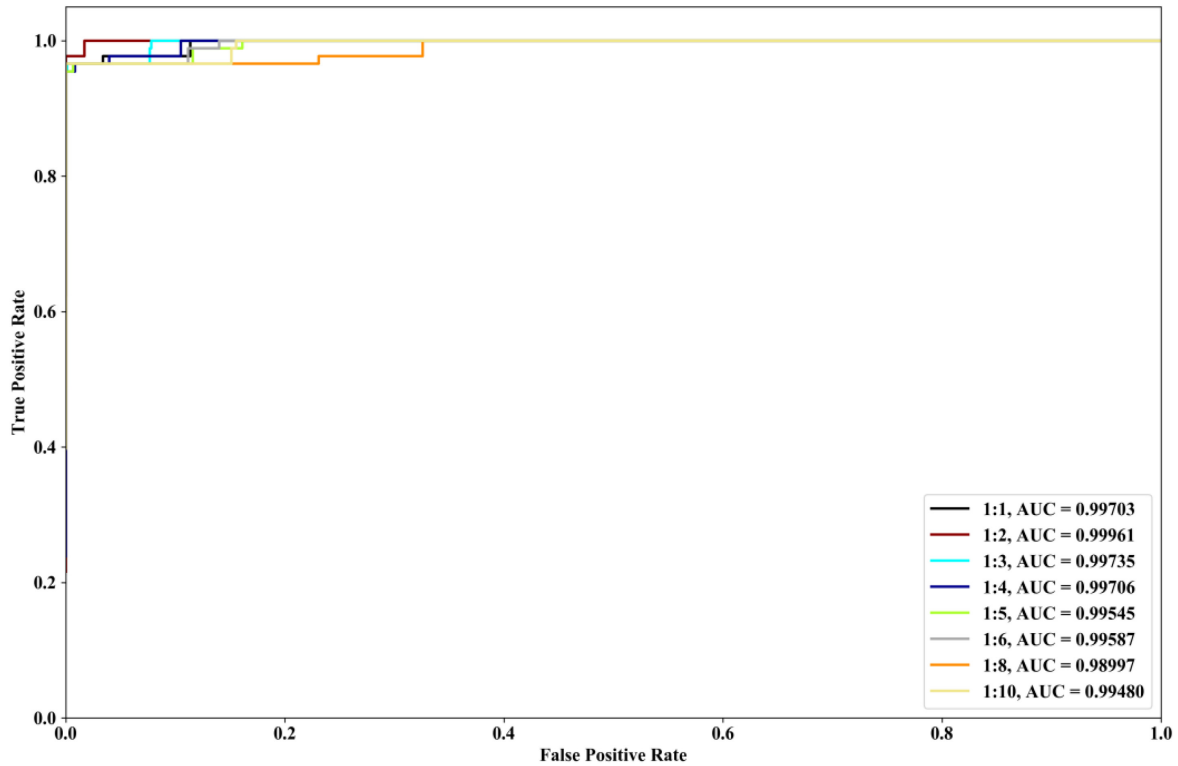

FIGURE 7 | Performances of our method with COVID-19 dataset and the different ratios of positive and negative samples.

We investigate the performances of our method with different ratios of positive and negative samples. The 8 kinds of ratios (1:1, 1:2, 1:3, 1:4, 1:5, 1:6, 1:8, and 1:10) are selected and COVID-19 dataset is utilized. The identification results are depicted in **Figure 7**. From **Figure 7**, it could be seen that when the ratios are 1:1, 1:2, 1:3, and 1:4, our method could have the better ROC and AUC performances. The excessive imbalance of data may affect the classification performance of the algorithm.

CONCLUSION

In order to sort the candidate compounds in a traditional Chinese medicine prescription and narrow the scope of analysis in network pharmacology research accurately, this paper proposes a new virtual screening method based on flexible neural tree (FNT) model, hybrid evolutionary method, and negative sample selection algorithm to screen the disease-related active compounds. 3 diseases (hypertension, diabetes, and Corona Virus Disease 2019) related compounds are collected from the up-to-date literatures. The unrelated compounds are selected by negative sample selection algorithm from DUD-E website. 4 kinds of molecular descriptors (ECFP6, MACCS, Macrocycle, and RDKit) are utilized to characterize the features of related and unrelated compounds of diseases, respectively. The experiment results show that our proposed method performs better than classical classifiers (SVM, RF, AdaBoost, DT, GBDT, KNN, LR, and NB), up-to-date classifier (gcForest) and deep learning method (forgeNet) in terms of AUC, ROC, TPR, FPR, Precision, Specificity, and F1.

We also investigate the performances of 11 methods with 4 kinds of molecular descriptors. The results show that our method, gcforest, forgenet, RF, GDBT, and LR perform best with MACCS feature set, while SVM and DT perform best with RDKit feature set, AdaBoost, KNN and NB perform best with Mordred feature set. With ECFP6 molecular descriptor all methods perform poorly.

In the paper, our proposed method has been successfully applied to hypertension, diabetes, and Corona Virus Disease. In the future, our method will be utilized to identify other chronic disorders related compounds, such as cancers, coronary heart disease, and rheumatoid disease.

DATA AVAILABILITY STATEMENT

The original contributions presented in this study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

WB conceived the method and wrote the main manuscript text. BY designed the method and conducted the experiments. All authors reviewed the manuscript.

FUNDING

This work was supported by the talent project of "Qingtan scholar" of Zaozhuang University, the Natural Science Foundation of China (Nos. 61702445 and 61902337), Jiangsu Provincial Natural Science Foundation (No. SBK2019040953), Natural Science Fund for Colleges and Universities in Jiangsu Province (No. 19KJB520016), Young talents of Science and Technology in Jiangsu, Youth Innovation Team of Scientific Research Foundation of the Higher Education Institutions of Shandong Province, China (No. 2019KJM006), the Key Research Program of the Science Foundation of Shandong Province (No. ZR2020KE001), the fundamental Research Funds for the Central Universities (2020QN89), Xuzhou Science and Technology Plan Project (KC19142 and KC21047), the Ph.D. research startup foundation of Zaozhuang University (No. 2014BS13), and Zaozhuang University Foundation (No. 2015YY02).

REFERENCES

Babaei, F., Lashkari, Z. B., Safari, A., Farrokhifar, M., Salehi, J., et al. (2020). Salp swarm algorithm-based fractional-order PID controller for LFC systems in the presence of delayed EV aggregators. *IET Electr. Syst. Transport.* 10, 259–267. doi: 10.1049/iet-est.2019.0076

Bajusz, D., Ferenczy, G. G., and Keser, G. M. (2017). Structure-Based Virtual Screening Approaches in Kinase-Directed Drug Discovery. *Curr. Topics Med. Chem.* 17, 2235–2259. doi: 10.2174/1568026617666170224121313

Baxter, C. A., Murray, C. W., Waszkowycz, B., Li, J., Sykes, R. A., Bone, R. G., et al. (2000). New approach to molecular docking and its application to virtual screening of chemical databases. *J. Chem. Inform. Comput. Sci.* 40, 254–262. doi: 10.1021/ci990440d

Berishvili, V. P., Voronkov, A. E., Radchenko, E. V., Palyulin, V. A., et al. (2018). Machine Learning Classification Models to Improve the Docking-based Screening: a Case of PI3K-Tankyrase Inhibitors. *QSAR Combinator. Sci.* 37:e1800030. doi: 10.1002/minf.201800030

Breiman, L. (2001). Random forest. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Bustamam, A., Hamzah, H., Husna, N. A., Syarofina, S., Dwimantara, N., Yanuar, A., et al. (2021). Artificial intelligence paradigm for ligand-based virtual screening on the drug discovery of type 2 diabetes mellitus. *J. Big Data* 8:74. doi: 10.1186/s40537-021-00465-3

Chen, L., and Mu, Y. (2021). Improved salp swarm algorithm. *Appl. Res. Comput.* 38, 1648–1652.

Chen, Y. F., Hsu, K. C., Lin, P. T., Hsu, D. F., Kristal, B. S., Yang, J. M., et al. (2011). LigSeeSVM: ligand-based virtual screening using support vector machines and data fusion. *Int. J. Comput. Biol. Drug Design* 4, 274–289. doi: 10.1504/IJCBDD.2011.041415

Chen, Y. H., Yang, B., and Meng, Q. (2012). Small-time scale network traffic prediction based on flexible neural tree. *Appl. Soft Comput.* 12, 274–279. doi: 10.1016/j.asoc.2011.08.045

Collins, M., Schapire, R. E., and Singer, Y. (2002). Logistic Regression, AdaBoost and Bregman Distances. *Mach. Learn.* 48, 253–285. doi: 10.1023/A:1013912006537

Fischer, N., Seo, E. J., Abdelfatah, S., Fleischer, E., Klinger, A., Efferth, T., et al. (2021). A novel ligand of the translationally controlled tumor protein (TCTP) identified by virtual drug screening for cancer differentiation therapy. *Invest. N. Drugs* 39, 914–927. doi: 10.1007/s10637-020-01042-w

Gomeni, R., BaniM, D., Angeli, C., Corsi, M., and Bye, A. (2001). Computer-assisted drug development (CADD): an emerging technology for designing first-time-in-man and proof-of-concept studies from preclinical experiments. *Eur. J. Pharmaceut. Sci.* 13, 261–270. doi: 10.1016/S0928-0987(01)00111-7

Guasch, L., Zakharov, A. V., Tarasova, O. A., Poroikov, V. V., Liao, C., Nicklaus, M. C., et al. (2016). Novel HIV-1 Integrase Inhibitor Development by Virtual Screening Based on QSAR Models. *Curr. Topics Med. Chem.* 16, 441–448. doi: 10.2174/1568026615666150813150433

Guo, S., Xie, H., Lei, Y., Liu, B., Zhang, L., Xu, Y., et al. (2021). Discovery of Novel Inhibitors Against Main Protease (Mpro) of SARS-CoV-2 *via* Virtual Screening and Biochemical Evaluation. *Bioorgan. Chem.* 110:104767. doi: 10.1016/j.bioorg.2021.104767

Hearst, M. A., Dumais, S. T., Osman, E., Platt, J., and Scholkopf, B. (1998). Support Vector Machines. *IEEE Intell. Syst.* 13, 18–28. doi: 10.1109/5254.708428

Kellenberger, E., Rodrigo, J., Muller, P., and Rognan, D. (2004). Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins.* 57, 225–242. doi: 10.1002/prot.20149

Kim, S. B., Han, K. S., Rim, H. C., and Myaeung, S. H. (2006). Some Effective Techniques for Naive Bayes Text Classification. *IEEE Transac. Knowledge Data Eng.* 18, 1457–1466. doi: 10.1109/TKDE.2006.180

Klekota, J., Brauner, E., and Schreiber, S. L. (2005). Identifying Biologically Active Compound Classes Using Phenotypic Screening Data and Sampling Statistics. *J. Chem. Inform. Modeling* 45, 1824–1836. doi: 10.1021/ci050087d

Kong, Y., and Yu, T. (2020). forgeNet: a graph deep neural network model using tree-based ensemble classifiers for feature graph construction. *Bioinformatics* 36, 3507–3515. doi: 10.1093/bioinformatics/btaa164

Leelananda, S. P., and Steffen, L. (2016). Computational methods in drug discovery. *Beilstein J. Organ. Chem.* 12, 2694–2718. doi: 10.3762/bjoc.12.267

Maddah, M., Bahramsoltani, R., Yekta, N. H., Rahimi, R., Aliabadi, R., Pourfath, M., et al. (2021). Proposing high-affinity inhibitors from Glycyrrhiza glabra L. against SARS-CoV-2 infection: virtual screening and computational analysis. *N. J. Chem.* 45, 15977–15995. doi: 10.1039/D1NJ02031E

Maia, E. H. B., Assis, L. C., de Oliveira, T. A., da Silva, A. M., and Taranto, A. G. (2020). Structure-Based Virtual Screening: from Classical to Artificial Intelligence. *Front. Chem.* 8:343. doi: 10.3389/fchem.2020.00343

Meenakumari, K., Bupesh, G., Vasanth, S., Vasu, C. A., Pandian, K., Prabhu, K., et al. (2019). Molecular docking based virtual screening of carbonic anhydrase IX with coumarin (a cinnamon compound) derived ligands. *Bioinformation* 15, 744–749. doi: 10.6026/97320630015744

Meng, X. Y., Zhang, H. X., Mezei, M., and Cui, M. (2011). Molecular Docking: a Powerful Approach for Structure-Based Drug Discovery. *Curr. Comput. Aided Drug Design* 7, 146–157. doi: 10.2174/157340911795677602

Mirjalili, S., Gandomi, A. H., Mirjalili, S. Z., Saremi, S., Faris, H., Mirajalili, M. S., et al. (2017). Salp swarm algorithm: a bio-inspired optimizer for engineering design problems. *Adv. Eng. Soft.* 114, 163–191. doi: 10.1016/j.advengsoft.2017.07.002

Morris, G. M., Goodsell, D. S., Huey, R., and Olson, A. J. (1996). Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4. *J. Mol. Recogn.* 10, 293–304. doi: 10.1007/BF00124499

Mysinger, M. M., Carchia, M., Irwin, J. J., and Shoichet, B. K. (2012). Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.* 55:6582. doi: 10.1021/jm300687e

Olubiyi, O. O., Olagunju, M., Keutmann, M., Loschwitz, J., and Strodel, B. (2020). High Throughput Virtual Screening to Discover Inhibitors of the Main Protease of the Coronavirus SARS-CoV-2. *Molecules* 25:3193. doi: 10.3390/molecules25143193

Rajguru, T., Bora, D., and Modi, M. K. (2020). Combined CADD and Virtual Screening to Identify Novel Nonpeptidic Falcipain-2 Inhibitors. *Curr. Comput. Drug Design* 17, 579–588. doi: 10.2174/1573409916666200701213526

Ren, H., Li, J., Chen, H., Li, C. Y., et al. (2021). Adaptive levy-assisted salp swarm algorithm: analysis and optimization case studies. *Mathemat. Comput. Simul.* 181, 380–409. doi: 10.1016/j.matcom.2020.09.027

Safavian, S. R., and Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Transac. Syst. Man, Cybernet.* 21, 660–674. doi: 10.1109/21.97458

Selvaraj, C., Panwar, U., Dinesh, D. C., Boura, E., Singh, P., Dubey, V. K., et al. (2021). Microsecond MD Simulation and Multiple-Conformation Virtual Screening to Identify Potential Anti-COVID-19 Inhibitors Against SARS-CoV-2 Main Protease. *Front. Chem.* 8:595273. doi: 10.3389/fchem.2020.595273

Talluri, S. (2021). Molecular Docking and Virtual Screening based prediction of drugs for COVID-19. *Comb Chem. High Throughput Screen* 24, 716–728. doi: 10.2174/1386207323666200814132149

Taufer, M., Crowley, M., Price, D. J., Chien, A. A., Brooks, C. L. III, et al. (2005). Study of a highly accurate and fast protein-ligand docking method based on molecular dynamics. *Concurr. Comput.* 14, 1627–1641. doi: 10.1002/cpe.949

Thiyagarajan, V., Lin, S. H., Chang, Y. C., Weng, C. F., et al. (2016). Identification of novel FAK and S6K1 dual inhibitors from natural compounds *via* ADMET screening and molecular docking. *Biomed. Pharmacother.* 80, 52–62. doi: 10.1016/j.biopha.2016.02.020

Todeschini, R., and Consonni, V. (2009). *Molecular Descriptors for Chemoinformatics*. Weinheim: Wiley-VCH. doi: 10.1002/9783527628766

Tong, J., Qin, S., and Jiang, G. (2019). 3D-QSAR Study of Melittin and Amoebapore Analogues by CoMFA and CoMSIA Methods. *Chin. J. Struct. Chem.* 2, 201–210.

Wang, M. Y., Peng, L., and Qiao, P. L. (2016). The Virtual Screening of the Drug Protein with a Few Crystal Structures Based on the Adaboost-SVM. *Comput. Math Methods Med.* 2016:4809831. doi: 10.1155/2016/4809831

Wu, P., and Chen, Y. (2007). "Grammar Guided Genetic Programming for Flexible Neural Trees Optimization," in *Advances in Knowledge Discovery and Data Mining. PAKDD 2007. Lecture Notes in Computer Science()*, Vol. 4426, eds Z. H. Zhou, H. Li, and Q. Yang (Berlin, Heidelber: Springer).

Yang, Z., Zhou, Y., and Zhong, L. (2021). Discovery of BAZ1A bromodomain inhibitors with the aid of virtual screening and activity evaluation. *Bioorganic Med. Chem. Lett.* 33:127745. doi: 10.1016/j.bmcl.2020.127745

Zaki, M. E. A., Alhussain, S. A., Masand, V. H., Akasapu, S., Bajaj, S. O., Ghosh, A., et al. (2021). Identification of Anti-SARS-CoV-2 Compounds from Food Using QSAR-Based Virtual Screening, Molecular Docking, and Molecular Dynamics Simulation Analysis. *Pharmaceuticals* 14:357. doi: 10.3390/ph14040357

Zaslavskiy, M., Jégou, S., and Tramel, E. W. (2019). ToxicBlend: virtual screening of toxic compounds with ensemble predictors. *Computat. Toxicol.* 10, 81–88. doi: 10.1016/j.comtox.2019.01.001

Zhang, B., Ren, J., Cheng, Y., Wang, B., Wei, Z., et al. (2019). Health Data Driven on Continuous Blood Pressure Prediction based on Gradient Boosting Decision Tree Algorithm. *IEEE ACCESS* 7, 32423–32433. doi: 10.1109/ACCESS.2019.2902217

Zhang, L., Ai, H. X., Li, S. M., Qi, M. Y., Zhao, J., Zhao, Q., et al. (2017). Virtual screening approach to identifying influenza virus neuraminidase inhibitors using molecular docking combined with machine-learning-based scoring function. *Oncotarget* 8, 83142–83154. doi: 10.18632/oncotarget.20915

Zhang, Y., Wang, Y., Zhou, W., Fan, Y., Zhao, J., Zhu, L., et al. (2019). A combined drug discovery strategy based on machine learning and molecular docking. *Chem. Biol. Drug Design* 93, 685–699. doi: 10.1111/cbdd.13494

Zheng, Y., Kong, L., Jia, H., Zhang, B., Wang, Z., Xu, L., et al. (2020). Network pharmacology study on anti-stroke of Xiaoshuan Tongluo formula based on systematic compound-target interaction prediction models. *Acta Pharmaceut. Sin.* 55, 256–264.

Zhou, Y., Zhang, B., Lin, Z. J., Zhang, X. M., Li, F., Wang, H. G., et al. (2016). Virtual screening for components in Chicory combined with CNT2 target based on molecular docking. *Zhongguo Zhong Yao Za Zhi* 41, 3962–3967.

Zhou, Z. H., and Feng, J. (2017). "Deep Forest: Towards An Alternative to Deep Neural Networks," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, (Nanjing: Nanjing University), </UB>3553–3559. doi: 10.24963/ijcai.2017/497

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Microbiology

ORIGINAL RESEARCH
published: 14 June 2022
doi: 10.3389/fmicb.2022.918823

Dysbiosis of the Gut Microbiome Is Associated With Histopathology of Lung Cancer

Xiong Qin[1†], Ling Bi[2†], Wenxiao Yang[2†], Yiyun He[2], Yifeng Gu[2], Yong Yang[1], Yabin Gong[2], Yichao Wang[2], Xiaoxia Yan[1], Ling Xu[2], Haibo Xiao[3] and Lijing Jiao[2,4*]*

[1] Department of Thoracic Surgery, Shanghai Pulmonary Hospital, Tongji University, Shanghai, China, [2] Department of Oncology, Yueyang Hospital of Integrated Traditional Chinese and Western Medicine, Shanghai University of Traditional Chinese Medicine, Shanghai, China, [3] Department of Cardiothoracic Surgery, Xinhua Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China, [4] Yueyang Hospital of Integrated Traditional Chinese and Western Medicine, Institute of Clinical Immunology, Shanghai University of Traditional Chinese Medicine, Shanghai, China

OPEN ACCESS

Edited by:
Liang Wang,
Xuzhou Medical University, China

Reviewed by:
Ranjith Kumavath,
Central University of Kerala, India
Chandrabose Selvaraj,
Alagappa University, India

***Correspondence:**
Haibo Xiao
xbcardio@163.com
Lijing Jiao
jiaolijing@shyueyanghospital.com

[†]These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

Received: *12 April 2022*
Accepted: *23 May 2022*
Published: *14 June 2022*

Citation:
Qin X, Bi L, Yang W, He Y, Gu Y, Yang Y, Gong Y, Wang Y, Yan X, Xu L, Xiao H and Jiao L (2022) Dysbiosis of the Gut Microbiome Is Associated With Histopathology of Lung Cancer. Front. Microbiol. 13:918823. doi: 10.3389/fmicb.2022.918823

Lung cancer is a malignancy with high incidence and mortality worldwide. Previous studies have shown that the gut microbiome plays an important role in the development and progression of metabolic cancers. However, data on the characteristics of the gut microbiome with different histopathology types of lung cancer remain scant. We collected stool samples from 28 healthy people (HP) and 61 lung cancer patients. The lung cancer patients were classified into three types according to their histopathology: Atypical Adenomatous Hyperplasia/Adenocarcinoma *in situ* (AAH/AIS), Minimally Invasive Adenocarcinoma (MIA), and Invasive Adenocarcinoma (IA). In addition, we employed 16S rRNA gene amplicon sequencing to analyze the characteristics of the gut microbiome in these patients. Our analysis revealed that the categorized cancer patients had unique intestinal flora characteristics, and had lower density and flora diversity compared to healthy people. Besides, the structure of the flora families and genera was more complex, and each group presented specific pathogenic microbiota. The patients in the AAH/AIS group and HP group had relatively similar flora structure compared with the IA and MIA groups. In addition, we identified several flora markers that showed significant changes with the development of lung cancer. Lung cancer gut microbiota showed a decrease in short-chain fatty acids (SCFAs) producing and anti-inflammatory bacteria compared to healthy people, while some pathogenic bacteria such as proinflammatory or tumor-promoting bacteria were more abundant in lung cancer patients. On the other hand, the Kyoto Encyclopedia of Genes and Genomes (KEGG) and Clusters of Orthologous Group (COG) annotation demonstrated suppression of some dominant metabolism-related pathways in lung cancer. These findings provide new biomarkers for the diagnosis and prognostic assessment of lung cancer and lay the basis for novel targeted therapeutic strategies for the prevention and treatment of lung cancer.

Clinical Trial Registration: [www.ClinicalTrials.gov], identifier [NCT03244605].

Keywords: 16S rRNA sequencing, lung cancer, gut microbiome, biomarkers, histopathology

INTRODUCTION

Lung cancer is one of the most aggressive and prevalent types of malignancy that leads to high morbidity and mortality (Allemani et al., 2018). Over 80% of lung cancer incidences are non-small cell lung cancer (NSCLC) (Wagner et al., 2020), which include adenocarcinoma (AC) and squamous cell carcinoma (SCC). However, with the development of individualized and targeted therapy for lung cancer, traditional pathological classification no longer meets the treatment requirements. It is, therefore, important to characterize the lung cancer subtypes based on the existing diagnostic criteria, coupled with more sensitive and specific diagnostic and prognostic markers.

Intestinal bacteria is a systemic metabolic product, which mediates disease resistance through metabolism, immunity, inflammation, and other mechanisms. Few studies have evaluated the interplay between the microbiome and lung cancer. Recent studies have also shown that intestinal flora has a unique population, which expresses in different cancers such as lung, breast, pancreatic, brain, and bone cancers (Nejman et al., 2020). In the treatment of lung cancer, intestinal flora can improve the efficacy and sensitivity of chemotherapy, radiotherapy, or immunotherapy, and reduce treatment-related toxicities (Cheng et al., 2020). In addition to carcinogenic effects, intestinal flora can also inhibit the development of cancer (Kadosh et al., 2020). The intestinal flora modulated cancer development by regulating its microenvironment, the host's immune system, as well as other metabolites (Finlay et al., 2020). Thus, the gut microbiome could correlate with the development of lung cancer, but evidence for the interplay between the microbiome and lung cancer is insufficient and cannot yet be used to predict tumor progression and prognosis.

An ideal diagnostic or prognostic index should have high specificity and sensitivity. Novel indexes such as intestinal flora have received considerable prospects for clinical application. To define biomarkers in the development of early lung adenocarcinoma, we explored the role played by intestinal flora changes using 16S rRNA sequencing and then attempted to correlate the intestinal flora changes with the development of infiltrating carcinoma. These data provided a theoretical basis for the accurate diagnosis and classification of early lung cancer.

MATERIALS AND METHODS

Samples

The 89 fecal samples for 16S rRNA sequencing were obtained from 28 healthy people and 61 lung cancer patients initially diagnosed by histopathology and computed tomography (CT). The lung cancer patients were further divided into 3 groups based on different histopathology as prescribed by WHO classification on Tumors of the Lung, Pleura, Thymus, and Heart in 2015, which include Atypical Adenomatous Hyperplasia/Adenocarcinoma *in situ* patients (AAH/AIS group, $n = 8$), minimally invasive adenocarcinoma patients (MIA group, $n = 18$), invasive adenocarcinoma patients (IA group, $n = 35$). None of the patients received therapy, such as chemotherapy, radiation therapy, targeted therapy, immunotherapy, or surgery before sample collection. We excluded patients who had one of the following conditions: congestive cardiac failure, respiratory failure, renal failure, severe liver dysfunction, consumption of probiotics or antibiotics within 1 month before admission. The control group was of 28 healthy people (HP group) who did not use any type of antibiotics or probiotics within 1 month before admission. Fresh fecal samples from all the participants were collected by the fecal sample collection kit (MGI Tech Co., Ltd., China) for intestinal microbial gene testing. The fecal samples were transferred into a sterilized tube containing stabilizer N-octylpyridine, which is a reliable reagent suitable for storage and transportation at room temperature. Then the fecal samples were frozen at −80°C immediately until DNA extraction. This study was conducted by the Declaration of Helsinki. The study was approved by the ethics committee of Yueyang Hospital of Integrated Traditional Chinese and Western Medicine Affiliated with Shanghai University of Traditional Chinese Medicine (NO.2016-059). Each patient gave signed informed consent before the study. The clinical trial registration date was August 9, 2017, and the registry number was NCT03244605.

Fecal DNA Extraction and 16S Sequencing

Microbial DNA was extracted from 89 fecal samples (61 fecal samples from lung cancer patients and 28 fecal samples from healthy people) by QIAamp® Fast DNA Stool Mini Kit following the manufacturer's protocol. Briefly, the V3–V4 variable regions of the bacterial 16S rRNA gene were amplified by polymerase chain reaction (PCR) using universal primers 338F: (ACTCCTACGGGAGGCAGCAG) 806R:GGACTACHVGGGTWTCTAAT). The extracted DNA was purified by silica gel and then quantified using a Quantus™ Fluorometer. The PCR cycle conditions included an initial denaturation at 95°C for 3 min; followed by 30 cycles at 95°C for 30 s, primer annealing at 52°C for 30 s, and extension at 72°C for 45 s; followed by a final elongation at 72°C for 10 min. The PCR products were then analyzed in 2% agarose gel. Subsequently, purified amplicons were pooled in equimolar amounts, and paired-end sequenced on Illumina HiSeq/MiniSeq for genome analysis.

Microbiome Data Analysis

The raw FASTQ files were first de-multiplexed, quality-filtered using chimera check, and then merged using FLASH (Magoč and Salzberg, 2011) with the sequences which were processed using the Cutadapt v1.3 and QIIME v1.8.0 (Cock et al., 2010). Briefly, forward, and reverse bacterial 16S rRNA reads were merged with a minimum length of 200 bps, and then we used the pick_open_reference method in the QIIME analysis to perform OTU clustering. The clustering algorithm used Uclust, and the database used the Greengenes 2013-08 release[1] version, and the similarity threshold was 80% for all sequences. Thereafter, we performed Operational Taxonomic Units (OTUs) division and statistical analysis, and the remaining parameters were the default

[1] http://greengenes.lbl.gov/Download/

parameters for QIIME. The index of observed species, Chao, Shannon, Sobs and Simpson were used to calculate alpha (α) diversity metrics. The beta (β) diversity measurements including Principal Component Analysis (PCA) and Principal Coordinates Analysis (PCoA) were used by the unweighted UniFrac metric. The PCA and PCoA were based on unweighted uniFrac distance. The statistical significance was evaluated using analysis of similarities (ANOSIM). In addition, the Linear Discriminant Analysis (LDA) Effect Size (LEfSe) method was used to evaluate the influence of each differentially abundant taxon. We further conducted an correlation network analysis to identify the co-occurring intestinal microbes under different histopathology types. To analyze the correlation network, we calculated the Spearman correlation between different groups of phylum using the R package cooccur. Subsequently, significant and robust correlations (P-value < 0.01, $|\rho| \geq 0.6$) were used to construct a network using the R package psych. Gephi (v0.9) was then used to construct network figures. Finally, pathway enrichment analysis was performed using the Kyoto Encyclopedia of Genes and Genomes (KEGG) and the Phylogenetic Investigation of Communities by Reconstruction of Unobserved States (PICRUSt) 2.0 database (Kanehisa et al., 2008; Langille et al., 2013; Douglas et al., 2020).

Statistical Analyses

Statistical tests were performed in R (3.0.2; R Foundation for Statistical Computing) and Prism software (Graph Prism7.0 Software Inc., CA, United States). Data were expressed as a mean ± standard deviation (SD) and the differences among the groups were evaluated by Wilcoxon rank-sum test. The Wilcoxon rank-sum test (for two groups) or Kruskal-Wallis test (for more than two groups) was used to analyze the diversity between multi-groups. Besides, Fisher's exact test was performed on categorical variables, whiles the chi-square test was used for categorical variables. A value $P < 0.05$ was considered statistically significant.

RESULTS

Patient Characteristics

Clinical characteristics of all the participants were listed in **Table 1**. No difference was observed in age, sex, disease stage, smoking status and family history ($P > 0.05$).

Clustering Analysis of Operational Taxonomic Unit

A total of 1,243 Operational Taxonomic Units (OTUs) were annotated for subsequent analysis, including 15 phyla, 81 families, 253 genera, and 555 species of gut microbes (**Figure 1A**). The coverage of 16S rRNA sequencing was 400–440 bp and the average length of these fragments was 415 bp (**Supplementary Figure 1A**). The data showed that the sobs index tended to be stable as sampling increased, which indicated that the depth of our sample sequencing met the analysis requirements for the diversity of intestinal flora (**Supplementary Figure 1B**).

Taxonomic Analysis of the 16S rRNA Sequence Data

To explore the features of the gut microbial community of the lung cancer patients, the relative microbiota taxon abundance in the lung cancer groups was compared with healthy people. The predominant genera were defined as those comprising greater than 1% of the total gut bacteria. Bacterial taxonomy distribution of the three lung cancer groups demonstrated increased density and clustering compared to the healthy controls group. In addition, a total of 605 OTUs were obtained for the HP group, 639 OTUs for the AAH/AIS group, 780 OTUs for the MIA group, and 944 OTUs for the IA group as shown by the Venn diagrams (**Figure 1A**). The number of unique OTUs in each group was 36, 38, 104, and 159 in AAH/AIS, MIA, IA, and HP groups, respectively. In addition, the HP and the lung cancer groups had a total of 446 shared OTUs, indicating that there was the high similarity between the structure of the intestinal flora of the healthy group and the lung cancer patients (**Figure 1A**). Rank-Abundance curves showed that the intestinal flora of the healthy group had higher abundance and diversity compared to the lung cancer groups (**Figure 1B**).

The Alpha Diversity of the Gut Microbiota

To investigate the diversity of the bacterial species in the gut ecosystem in each group, the microbial alpha diversity was measured as shown in **Figure 2**. Alpha diversity evaluates the diversity of microbial communities in a region, reflecting the richness and evenness. We obtained data such as species abundance by observation of various index values such as Chao, Shannon, Sobs, and Simpson index. Community richness can be measured by Chao index, while community diversity indices includes Shannon index and Simpson index. Sobs index represents the number of species observed in the sample (OTU number). The Chao, Shannon, Sobs index are positively correlated with the richness and diversity while the Simpson index is negatively correlated with them. We then employed a t-test to define the significance of the differences in the index values between the four groups. The Chao, Shannon, or Sobs index ($P < 0.05$) demonstrated that the diversity index of the HP group was significantly higher compared to the three lung cancer groups, while the Simpson index was lower compared to the three lung cancer groups ($P < 0.05$). Our results demonstrated that the intestinal flora of lung cancer patients was significantly different from in the HP group, and the gut microbiota abundance and diversity of the lung cancer patients were lower than the HP group. In addition, there was no significant differences in the indices of the different lung cancer groups ($P > 0.05$).

The Beta Diversity Analysis of the Gut Microbiota

The Beta (β) diversity was used to evaluate the similarities and differences of between-group diversity of each group, including principal component analysis (PCA) and principal coordinates analysis (PCoA) based on unweighted UniFrac distance. The more similar the community composition of the samples is, the closer they are to each other in the PCA or PCoA

TABLE 1 | Baseline characteristics of health people and non-small cell lung cancer (NSCLC) patients.

Characteristics	Total (*n* = 89)	HP (*n* = 28)	AAH/AIS (*n* = 8)	MIA (*n* = 18)	IA (*n* = 35)	*P-value*
Age, years Mean ± *SD*	55.88 ± 10.87	58.79 ± 11.16	49.00 ± 5.61	53.67 ± 12.34	56.26 ± 10.17	0.110
Sex, *n* (%)						
Male	30 (33.71)	11 (39.29)	2 (25.00)	7 (38.89)	10 (28.57)	0.731
Female	59 (66.29)	17 (60.71)	6 (75.00)	11 (61.11)	25 (71.43)	
Smoking status, *n* (%)						
Smoker	10 (11.24)	5 (17.86)	0 (0.00)	3 (16.67)	2 (5.71)	0.279
Non-smoker	79 (88.76)	23 (82.14)	8 (100.00)	15 (83.33)	33 (94.29)	
Family history, *n* (%)						
Yes	7 (7.87)	0 (0.00)	0 (0.00)	1 (5.56)	6 (17.14)	0.098
No	82 (92.13)	28 (100.00)	8 (100.00)	17 (94.44)	29 (82.86)	
Disease stage, *n* (%)						
IA	–	–	0 (0.00)	17 (94.44)	30 (85.71)	
IB	–	–	0 (0.00)	1 (5.56)	3 (8.57)	0.529
IIA	–	–	0 (0.00)	0 (0.00)	0 (0.00)	
IIB	–	–	0 (0.00)	0 (0.00)	2 (5.71)	
EGFR mutation, *n* (%)						
L858R	–	–	0 (0.00)	1 (5.56)	8 (22.86)	
19-del	–	–	0 (0.00)	0 (0.00)	4 (11.43)	0.074
Unknown	–	–	8 (100.00)	17 (94.44)	23 (65.71)	
Solitary/multiple nodule, *n* (%)						
Solitary	–	–	2 (25.00)	8 (44.44)	18 (51.43)	0.396
Multiple	–	–	6 (75.00)	10 (55.56)	17 (48.57)	
Defecation, *n* (%)						
Normal	62 (69.66)	28 (100.00)	7 (87.50)	15 (83.33)	27 (77.14)	0.067
Abnormal	27 (30.34)	0 (0.00)	1 (12.50)	3 (16.67)	8 (22.86)	

HP, healthy people; AAH, atypical adenomatous hyperplasia; AIS, adenocarcinoma in situ; MIA, minimally invasive adenocarcinoma; IA, invasive adenocarcinoma.

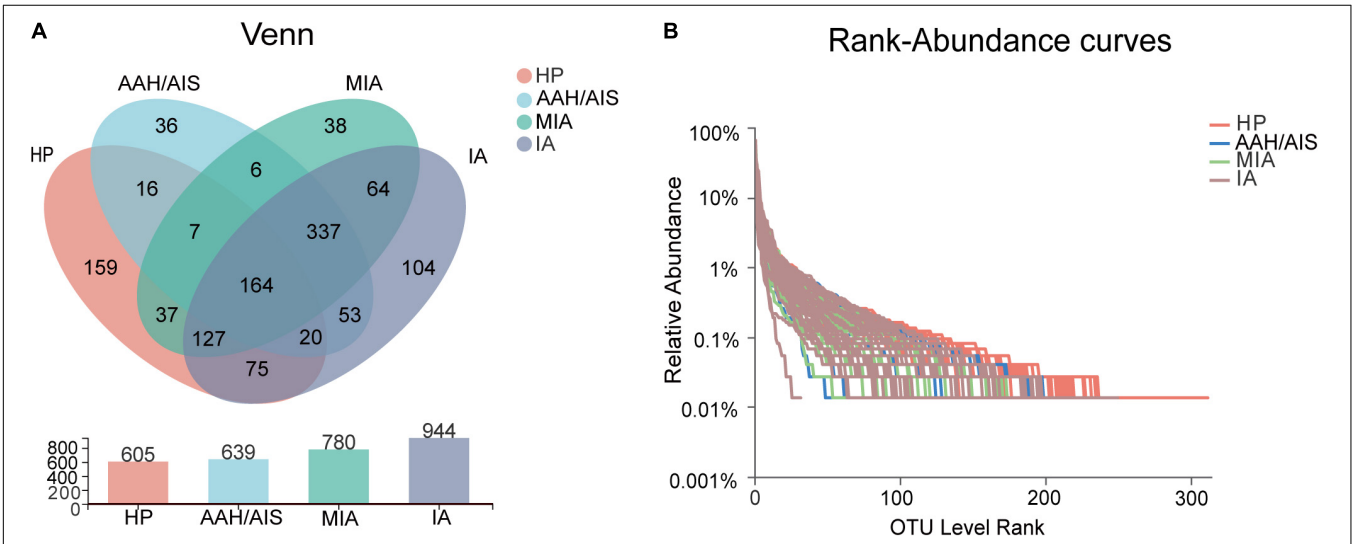

FIGURE 1 | Taxonomic analysis of 16S rRNA sequence data. **(A)** Venn diagram of OTU shared among the four groups. **(B)** Rank-Abundance curves of intestinal flora in four groups of samples.

diagram. Therefore, samples with high similarity in community structure tend to cluster together, while those with very different communities are far apart. We performed the PCA analysis between the four groups as shown in **Figure 3A**. When PC1 (35.09%) and PC2 (25.33%) were taken as the abscissa and ordinate, respectively, the four groups were well distinguished ($P = 0.007$), demonstrating that the four groups had significant differences in the composition of the intestinal bacteria. Besides,

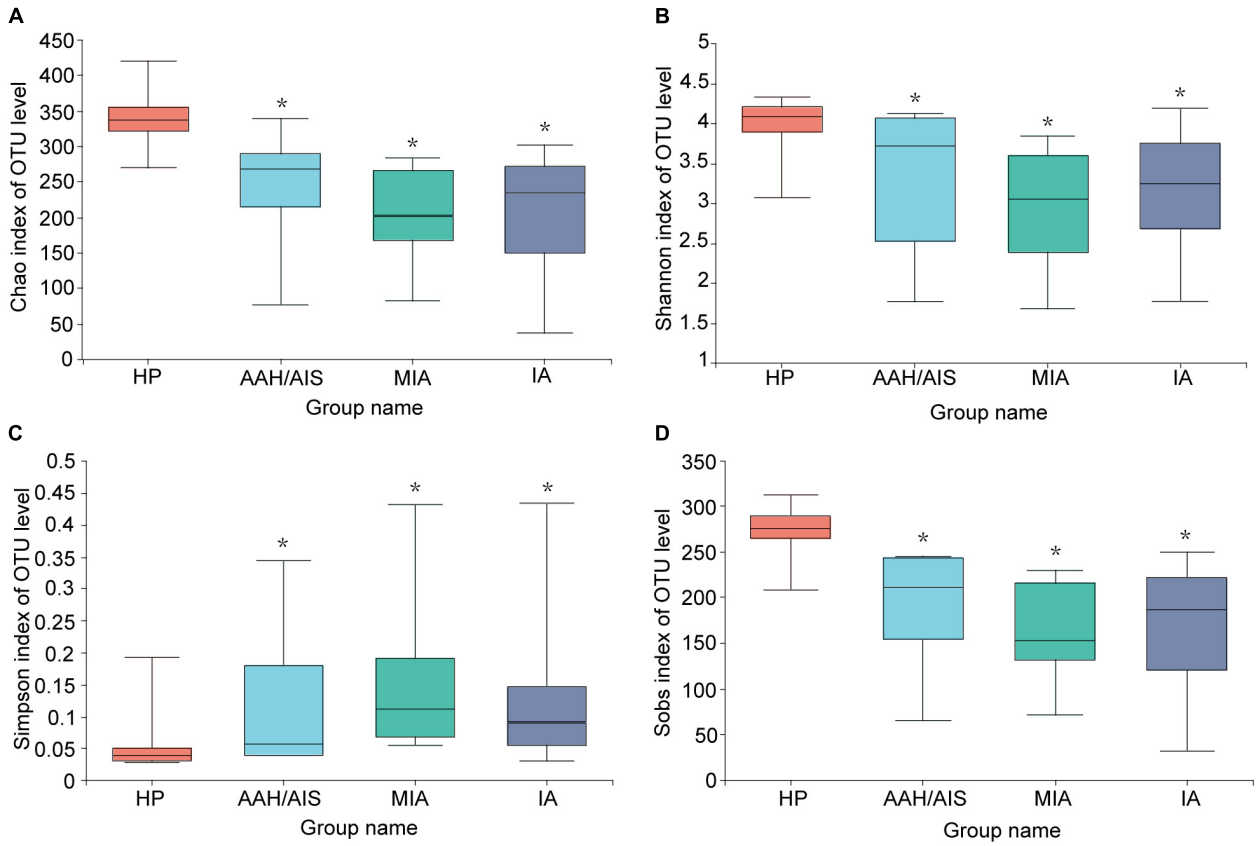

FIGURE 2 | Comparison of Alpha diversity index of intestinal flora among the four groups. **(A)** Microbial alpha diversity of Chao index. **(B)** Shannon index. **(C)** Simpson index. **(D)** Sobs index. *Indicates $P < 0.05$ compared to the HP group.

in the PCoA analysis (**Figure 3B**), when PC1 (19.27%) and PC2 (11.75%) were taken as the abscissa and ordinate, respectively, the four groups were farther apart in the coordinate chart ($P = 0.001$), which indicated that there was a significant difference in species diversity between the four groups.

In summary, our data showed that there were significant differences in the species diversity and community composition of the intestinal flora between the lung cancer patients and healthy controls, as well as certain differences in the diversity and structure of the intestinal flora between the three different pathological subgroups of lung cancer. However, the results of Beta diversity can only illustrate the general similarities and differences of diversity between each group. Therefore, the clear information on detailed differences between the four groups were further reflected by subsequent species taxonomic profiling at different levels of biological classification.

Variation Analysis

Species Specificity in Multi-Level Tests

At the phylum level, Firmicutes, Bacteroidetes, and Proteobacteria were the most common phyla identified in the three lung cancer groups, contributing 87.27% (AAH/AIS), 93.53% (MIA), and 93.09% (IA) of the gut bacteria, respectively. Firmicutes, Bacteroidetes, Proteobacteria, and Acidobacteria contributed to 98.95% of the gut bacteria in the HP group (**Figure 4A**). The lung cancer groups especially the MIA group had a significantly lower abundance of Firmicutes, a relatively higher abundance of Proteobacteria, Bacteroidetes, and Fusobacteria compared to the HP group. On the other hand, the AAH/AIS group showed a relatively low abundance of Acidobacteria (**Figures 4B–D**). The ratio of Firmicutes to Bacteroidetes can reflect the homeostasis of intestinal flora. The Firmicutes/Bacteroidetes ratio in the HP group was 1.88, while in the lung cancer group, the ratio was 1.12 (AAH/AIS), 0.48 (MIA), and 0.95 (IA), respectively.

In addition, analysis of relative abundance showed a clear difference between the taxa with high and low abundance were distinguished, and the color gradient were used to reflect the similarity and difference of the composition of multiple samples at each classification level. As shown in **Figure 5**, the difference between the four groups of samples can be seen intuitively according to the change in the color gradient.

Gut Microbial Signature in Lung Cancer Patients

The multi-level LEfSe was used to analyze biomarkers between the lung cancer patients with different histopathology and the healthy controls. Our results showed that dominant fecal gut microbiota was specific to the histopathological types of lung

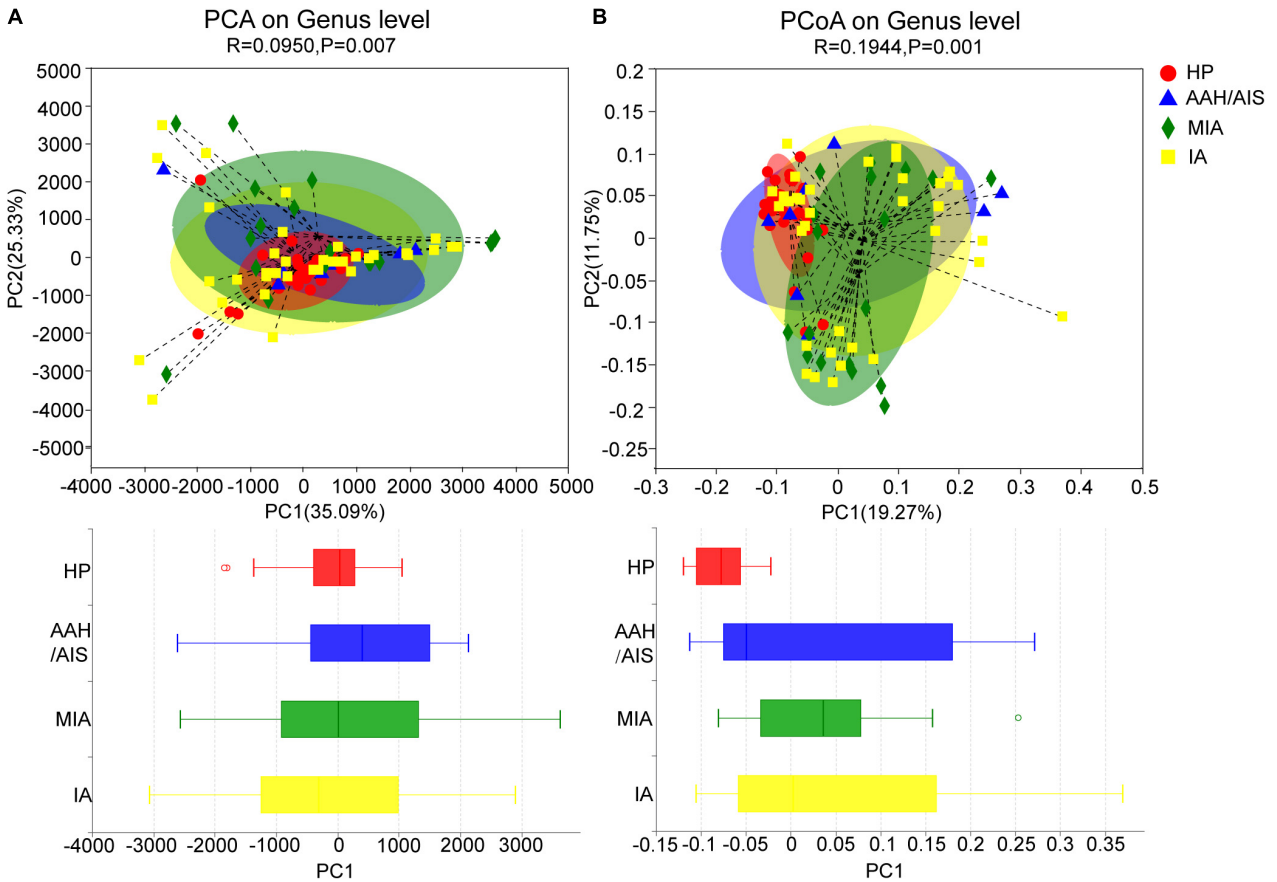

FIGURE 3 | Beta diversity analysis of intestinal flora. **(A)** PCA analysis of intestinal flora in four groups of samples. **(B)** PCoA analysis of intestinal flora in four groups of samples.

cancer. There were 74, 20, 15, and 15 bacterial taxonomic clades that were significantly different in HP, AAH/AIS, MIA, IA groups, respectively [log10 (LDA score) > 2] (**Figure 6A**).

We also analyzed the evolutionary relatedness of the intestinal flora species as shown in **Figure 6B**. The data showed that the species were divergent, which was in sync with the LDA value distribution data. The data showed that the dominant flora in each group of lung cancer patients was significantly different from the healthy people, and there were also significant differences in the characteristic flora in the lung cancer patients based on the different pathological types.

The flora evolution analysis showed the relative content of these dominant bacteria (**Figure 7**). In the HP group, p_Firmicutes, c_Clostridia, and o_Clostridiales were shown to be the most significant, while in AAH/AIS group, g_Lachnoclostridium, g_Parasutterella, and g_Eubacterium_coprostanoligenes had the highest abundance. On the other hand, in the MIA group, p_Bacteroidetes, o_Bacteroidale, and c_Bacteroidia were shown to be the most significant genus, while in the IA group, g_Prevotella_9, g_Klebsiella, and g_Eubacterium_eligens were most represented. Besides, in the HP group, the dominant bacteria group was classified at a high level, while the different lung cancer groups were significantly reflected in the low-level classification. Further analysis showed that o_Bacteroidales, o_Clostridiales, f_Lachnospiraceae, f_Ruminococcaceae, g_Anaerotruncus, g_Faecalibacterium, g_Prevotella_9, g_Roseburia, and g_Subdoligranulum in HP group was significantly different from MIA, IA, but not with AAH/AIS group (**Figure 7A**). On the other hand, f__Peptostreptococcaceae, f_Christensenellaceae, f_Veillonellaceae, g_Blautia, g_Christensenellaceae_R-7_group, g_Haemophilus, g_Lachnospira, g_Lachnospiraceae_NK4A136_group, and g_Lachnospiraceae_UCG-001 were significantly different from the other three groups (**Figure 7B**). These florae features may be related to the development of lung cancer.

Moreover, several specific genera were presented in both lung cancer patients and healthy people. According to the LEfSe analysis, the genera of Lachnospiraceae, Ruminococcaceae, and Eubacterium were predominantly identified in both cancer patients and healthy people. Specifically speaking, the genera of Lachnospiraceae were in both healthy people and IA group. The genera of Ruminococcaceae were both enriched in healthy people and AAH/AIS group. Eubacterium genera were simultaneously identified in healthy people and three lung cancer subgroups AAH/AIS, MIA, and IA group.

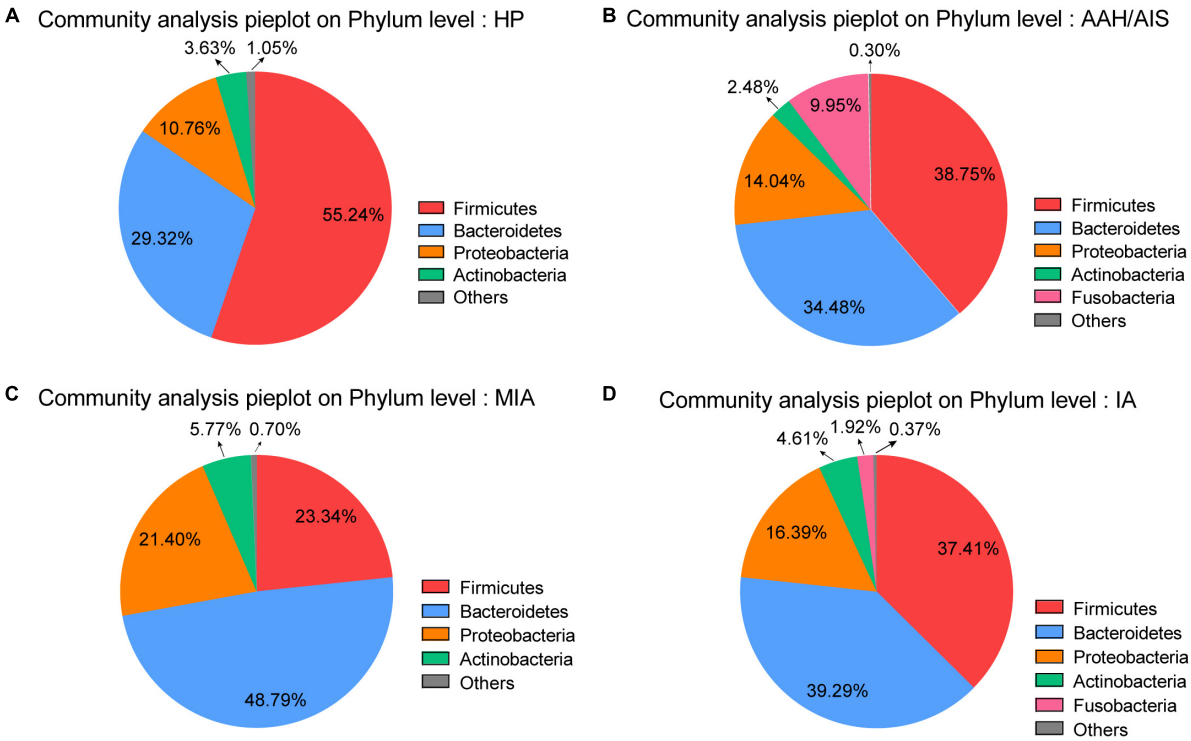

FIGURE 4 | Major OTUs at Phylum level in the HP group vs. the three lung cancer subgroups. **(A)** Major OTUs at Phylum level in the HP group. **(B)** Major OTUs at Phylum level in the AAH/AIS group. **(C)** Major OTUs at Phylum level in the MIA group. **(D)** Major OTUs at Phylum level in the IA group.

Constructed networks revealed that samples from the HP had fewer edges, a lower average degree and lower nodes than those from the lung cancer group, which indicated that there were fewer significant correlations of phylum (**Supplementary Table 1**). In AAH/AIS group, average weighted degree, density and clustering coefficient were higher than the other three groups, demonstrating a elevation in the network complexity. Co-occurrence was also found among species of the Proteobacteria in AAH/AIS, MIA, IA environments (**Figures 8B–D**), however, such co-occurrence was missing in the healthy environment (**Figure 8A**).

Functional Profile of the Gut Microbiome in Non-small Cell Lung Cancer

The KEGG and COG pathway analyses were performed to explore potential differences in the functions of the microbiome in lung cancer patients vs. healthy individuals.

Although the functional analyses showed significant similarity between the lung cancer patients and the control group, the microbiome of the lung cancer patients was abundant in pathways such as carbohydrate digestion and absorption, which was proportional to the development of lung cancer. On the other hand, the KEGG analysis showed clustering of valine, leucine, and isoleucine biosynthesis, arginine biosynthesis, and glutamatergic synapse, which showed lower abundance in the lung cancer patients than the healthy controls (**Figure 9A**). In addition, diguanylate cyclase (COG2199) and RNA-binding protein (COG1534) of the ABC (ATP-binding cassette) transporter system were significantly downregulated in lung cancer patients compared to the healthy controls, which might be promoting utilization of glucose or ribose/galactoside to regulate energy. In addition, exported protein (COG2911) ortholog was upregulated in the lung cancer patients compared to the healthy controls ($P < 0.05$) (**Figure 9B**).

DISCUSSION

Adenocarcinoma is the most common type of pathology in NSCLC. With the development of imaging techniques like High Resolution Computed Tomography (HRCT), CT imaging, Positron Emission Tomography-Computed Tomography (PET-CT), and Magnetic Resonance Imaging (MRI), the detection rate of early lung adenocarcinoma has significantly improved. However, how to analyze the prognosis of patients with early lung adenocarcinoma is particularly critical.

There are three major types of pathology in early lung adenocarcinoma, including adenocarcinoma *in situ* (AIS), minimally invasive adenocarcinoma (MIA), and invasive adenocarcinoma (IA) with a maximum tumor diameter of ≤ 3 cm (Travis et al., 2015). Besides the AIS, there is atypical adenomatous hyperplasia (AAH) with very similar morphology. Although the CT imaging of the above lesions is mainly in the form of ground-glass nodules, their prognosis is quite different. AAH can be observed and followed up without surgery for years;

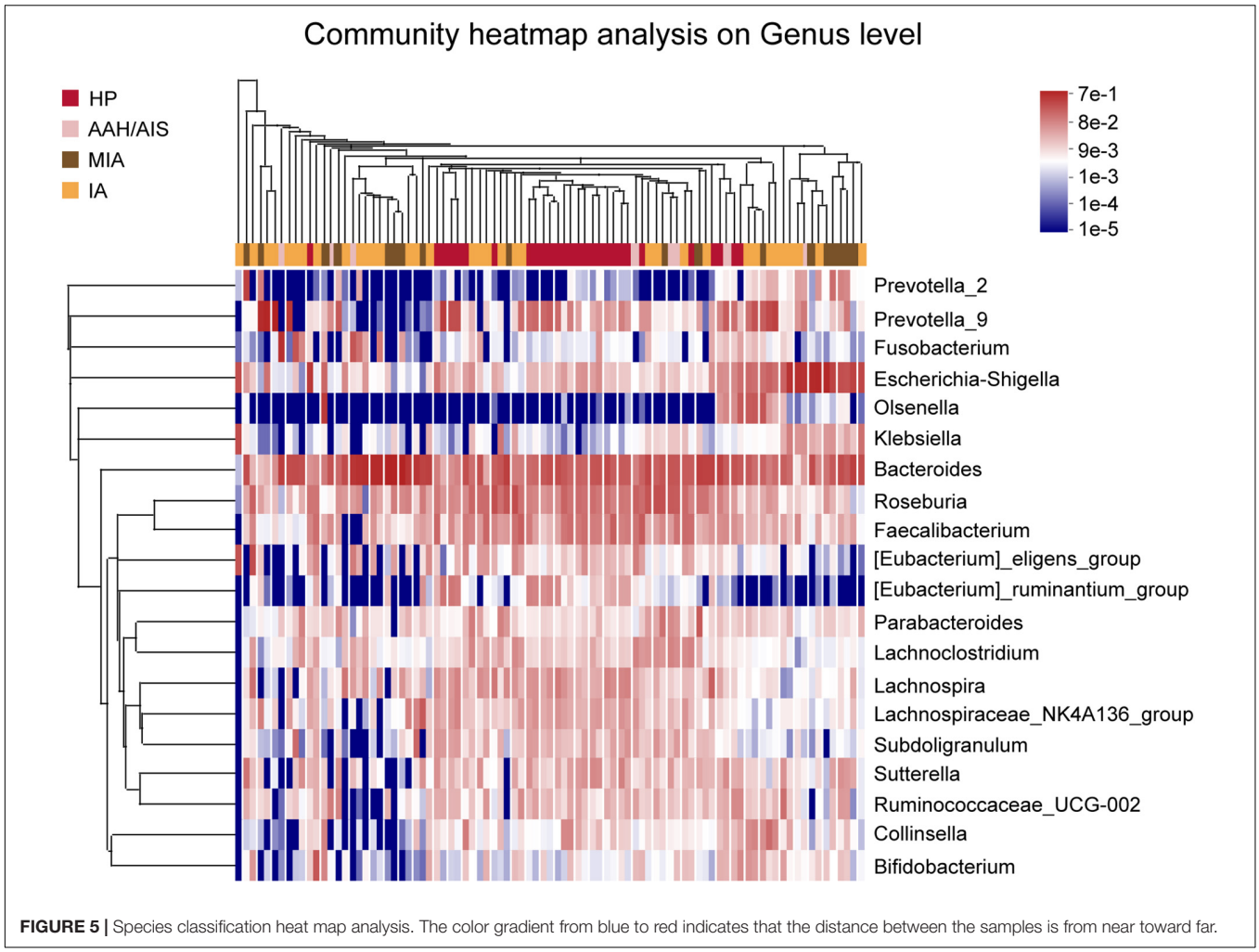

FIGURE 5 | Species classification heat map analysis. The color gradient from blue to red indicates that the distance between the samples is from near toward far.

AIS and MIA could be treated by lobectomy without regional lymph node dissection, with a 5-year survival rate of 100%; while submerged IA with predominantly appendicular growth requires lobectomy and regional lymph node dissection, with a 5-year survival rate of 67% (Detterbeck et al., 2017). With the increase in the detection rate of pulmonary ground-glass nodules, it is essential to classify the degree of malignancy of the nodules. Unfortunately, the ground glass nodules have similarities and overlap in histomorphology, which blocks accurate diagnosis and treatment. Presently, experienced pathologists identified the types of early lung adenocarcinoma based on infiltrating carcinoma components in the lesion, but there are no specific biological markers of infiltrating carcinoma components in lesions, especially for the early lung adenocarcinoma patients. Therefore, it is urgent to explore non-invasive and economical screening modalities which could easily detect samples with high positive rates.

Recent studies have shown that intestinal flora can be used in the diagnosis of human diseases such as tumors (Zheng et al., 2020; Leng et al., 2021). Intestinal flora is a large group of microorganisms that colonize the intestines, and their homeostasis plays an important role in regulating the development of human diseases and is referred to as the "second genome" (Qin et al., 2010) or "a new organ" (Donaldson et al., 2016). Previous data has demonstrated a pathogenic association between the microorganisms and the gut-lung axis (Gut-lung axis) (Budden et al., 2017), which is the basis for the regulation of lung cancer by the intestinal flora microenvironment. The intestines and the lung regulate each other through the gut-lung axis, which relies on various biological structures such as embryonic homology, mucosal immune channels, and neurological channels. Besides, the intestinal microenvironment could influence the occurrence, development, treatment, and prognosis of lung cancer through various pathways. In our study, we showed that the lung cancer group had significant differences from the healthy group, which is consistent with previous reports (Liu et al., 2019a). Thus, the microbiota has high sensitivity in early lung adenocarcinoma compared to blood tumor markers such as carcinoembryonic antigen (CEA), carbohydrate antigen 125 (CA125), and squamous cell carcinoma (SCC) antigen.

The α- and β-diversity results of lung cancer patients with different histopathology types did not show any significant differences, but the HP and AAH/AIS groups showed high similarity, while the IA group was similar to the MIA group.

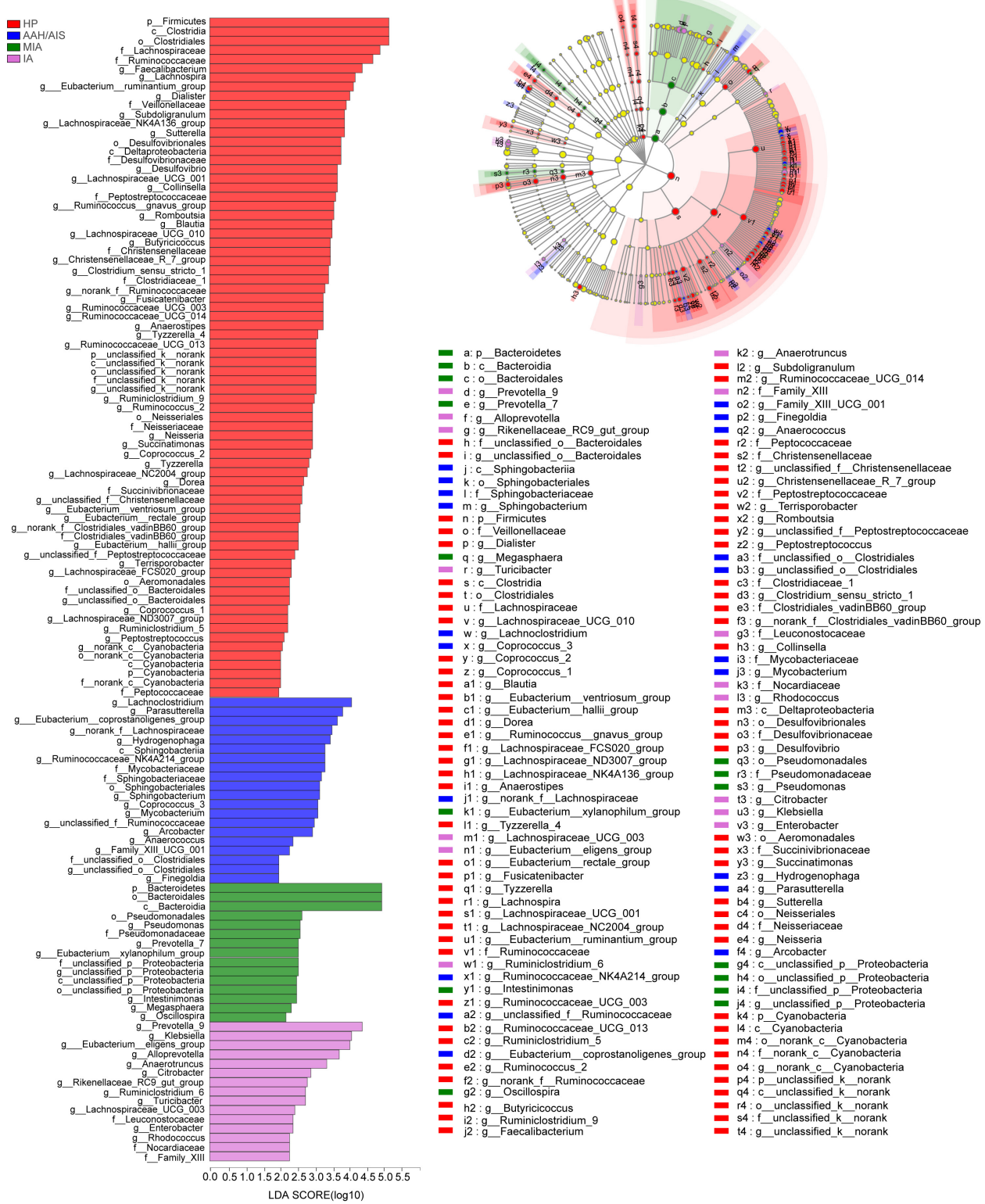

FIGURE 6 | Identification of gut microbiota composition and abundance across the four groups. **(A)** Histogram of the distribution of LDA values for LEfSe analysis of intestinal flora in four groups of samples. **(B)** Evolutionary map of species branching for LefSe analysis of intestinal flora in four groups of samples.

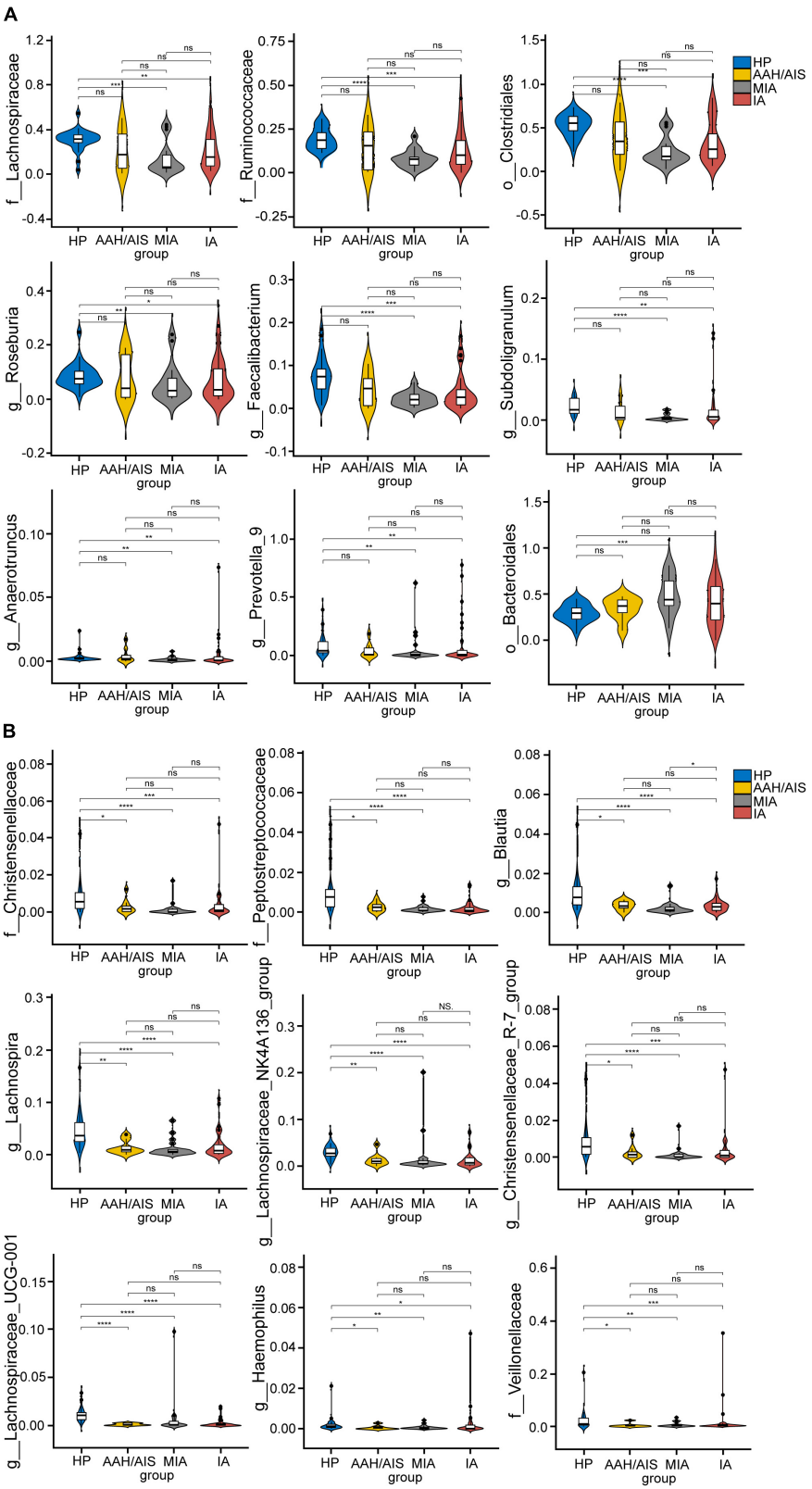

FIGURE 7 | The comparison of relative abundant microbiome between each group. **(A)** The Characteristic flora which are significantly HP group is significantly different from MIA, IA, and no difference. **(B)** The Characteristic flora which are significantly different between HP and the other three group. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, ****$P < 0.0001$, ns, no significance.

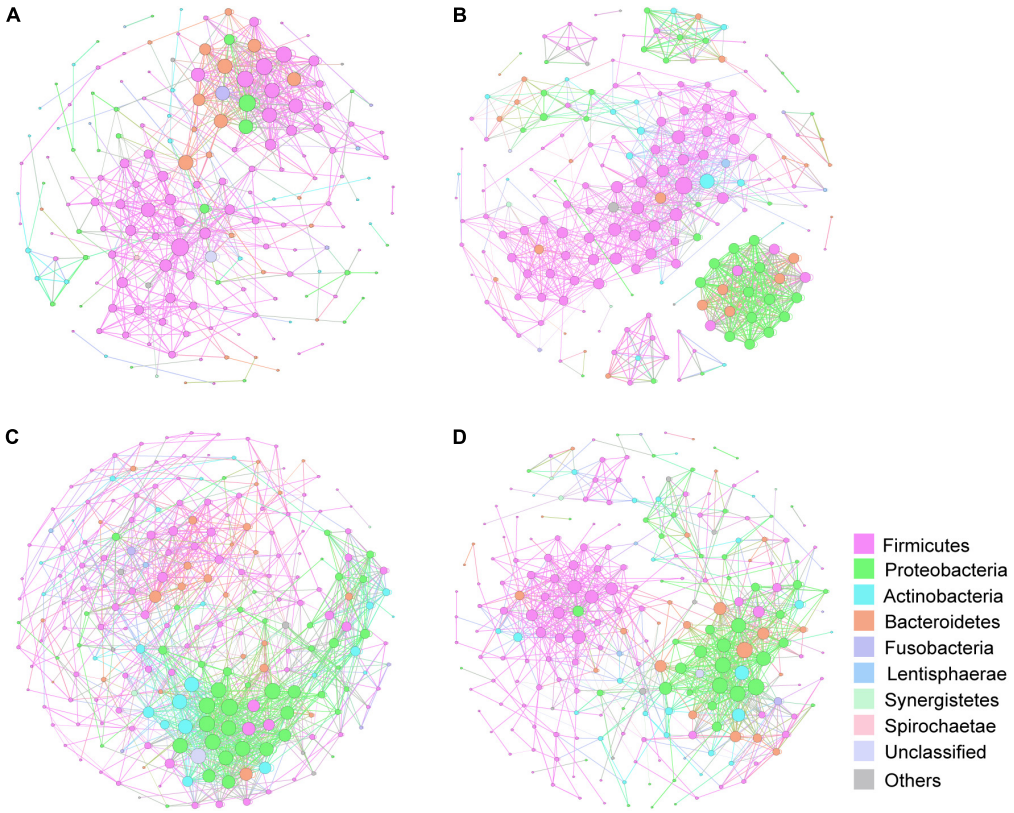

FIGURE 8 | Correlation network of the gut microbiome in the four groups. The correlation coefficient was calculated with Spearman rank correlation test ($|\rho| \geq 0.6$). Gephi (v0.9) was used for network construction. **(A)** Correlation networks in HP. **(B)** Correlation networks in AAH/AIS. **(C)** Correlation networks in MIA. **(D)** Correlation networks in IA. Each circle represents the average relative abundance of a microbial species in that state. Node sizes are scaled according to their degrees of connections.

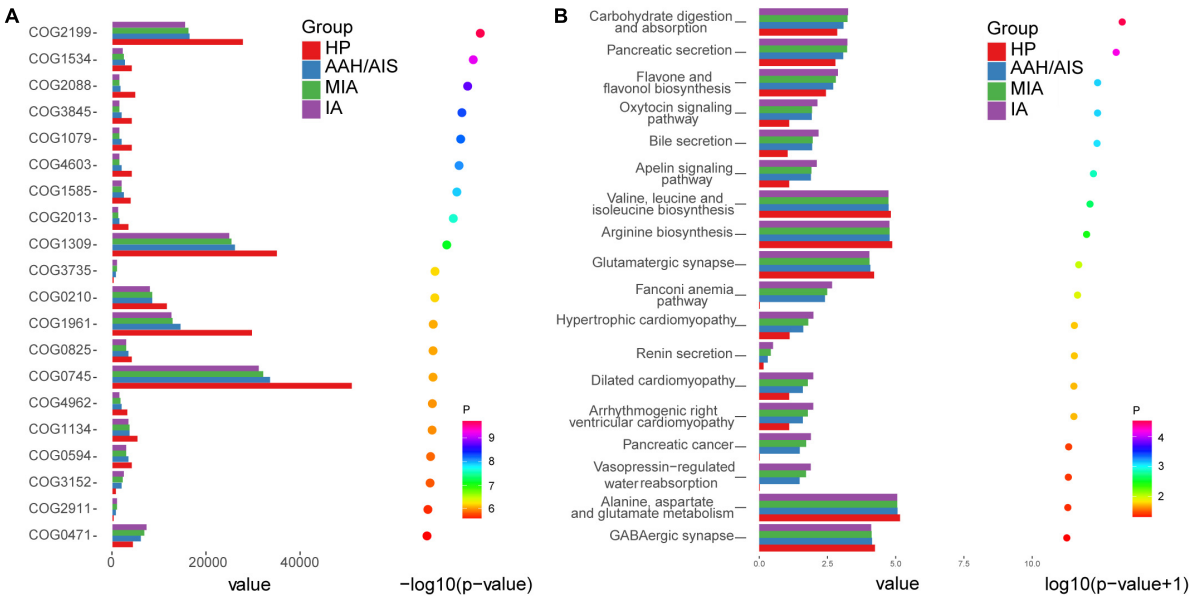

FIGURE 9 | COG pathway and KEGG analysis in the four groups. **(A)** The relative abundance of COG pathway differentially enriched in the four groups. **(B)** The relative abundance of KEGG pathway differentially enriched in the four groups.

Together, these differences were not statistically significant but was confirmed by specific flora structure.

Moreover, at the phylum level, Firmicutes were significantly higher in the HP group compared to the AAH/AIS, MIA, and IA groups, while the ratio of Firmicutes to Bacteroidetes was lower than in the HP group. Previous data demonstrated that all butyrate-producing bacteria belong to the Firmicutes. Besides, butyrate is one of the most important fatty acids associated with anti-inflammatory activity, cell proliferation, induction of regulatory T cell differentiation, and apoptosis through activation of signaling pathways (O'Keefe, 2016; Feng et al., 2018). High rates of Firmicutes/Bacteroidetes phylum are frequently observed in healthy adults, as previously demonstrated using a large gut microbiome cohort study (Zhong et al., 2019). Reduced Firmicutes/Bacteroidetes ratio has been shown to be associated with dysbiosis of gastrointestinal tract metabolism, which results in low concentration of circulating short-chain fatty acids, and then influenced elements for host systemic immunity and systemic inflammation (Liu et al., 2019b). This data shows that there is a disrupted balance of gut microbiota in lung cancer patients and the presence of distinct microbiota profiles from those of precancerous lesions.

The characterizations in family and genus levels were more complex and significantly varied from each group, presenting a more diverse pathogenic population. Our results showed that the Lachnospiraceae and Blautia genera were suppressed in lung cancer patients, which was in agreement with previous studies (Liu et al., 2019a; Zhang et al., 2019). The Lachnospiraceae genera of the Clostridium family belongs to Firmicutes phylum, which was suppressed in each lung cancer group compared to the HP. Lachnospiraceae can protect the host against cancer by producing butyric acid which plays an important role in the suppression of tumor growth, regulation of immunity, and participation in anti-inflammatory reactions (Daniel et al., 2017). Each lung cancer group exhibited a decreased abundance of the Blautia genus belonging to the Firmicutes phylum, which has a role in digesting complex carbohydrates. The suppression of the Blautia genus was also seen in irritable bowel syndrome, non-alcoholic fatty liver diseases, Crohn's disease, and diabetes. However, the specific roles of these common specific florae and their importance need further confirmation (Zhang et al., 2019). Our results indicated that the composition and development of bacterial communities varied in lung cancer with a different course. Therefore, it is feasible to speculate that some microbiome might be used for diagnosis, prognosis, therapeutics or fecal microbiota transplantation in lung cancer.

Our data also showed that there was a lower abundance of Faecalibacterium, Prevotella, Roseburia, and Subdoligranulum, Anaerotruncus genera in lung cancer patients in IA and MIA groups compared with HP, but no difference with AAH/AIS group. Faecalibacterium was reported as a "favorable" gut microbiome, which can enhance systemic and anti-tumor immune responses mediated by increased antigen presentation, and improved effector T cell functions as well as the tumor microenvironment, which modulates the response of melanoma patients to anti-programmed death-1(PD-1) immunotherapy (Gopalakrishnan et al., 2018). It was also shown that patients on Cytotoxic T Lymphocyte-associated Antigen-4 (CTLA-4) blockade with a higher abundance of Faecalibacterium had a prolonged PFS compared to those with a higher abundance of Bacteroidales in the gut microbiome (Chaput et al., 2017). Thus, these findings demonstrated that Faecalibacterium plays an important role in immunotherapy. Prevotella belongs to the Prevotaceae family of Bacteroides. It has a diverse bacterial species and is a dominant genus in the human intestine. It is negatively associated with metabolic diseases such as obesity and diabetes (Lukens et al., 2014). In our study, we showed that Prevotella decreased with lung cancer progression. Roseburia genus has been shown to produce short-chain fatty acids, especially butyric acid, which affects colon movement with anti-inflammatory properties and has the potential of being a probiotic (Sanders et al., 2019). Studies have shown that the occurrence of colorectal cancer may be related to the reduction of the Roseburia (Bisht et al., 2021). Our findings showed that the reduction of Roseburia genus was associated with the occurrence and progression of lung cancer.

In the healthy people group, the majority of gut bacteria were associated with the production of short-chain fatty acids (SCFAs), the regulation of the immune system, and the modulation of metabolism. The microbial genera in healthy people were characterized by a higher abundance of beneficial bacteria that promote the restoration of gut microenvironment balance, and some of them were identified as the next-generation probiotics (Singh and Natraj, 2021). However, these beneficial gut microbiota were not significantly observed in either of the three subgroups of lung cancer patients. On the contrary, most of the beneficial gut microbiota were significantly decreased in lung cancer patients, and some pathogenic bacteria such as proinflammatory or tumor-promoting bacteria were more abundant in lung cancer patients. Lachnoclostridium (Liang et al., 2020), Pseudomonas (Rathje et al., 2020), Eubacterium_xylanophilum_group (Zhang et al., 2020), Megasphaera (Lee et al., 2016), Klebsiella (Jian et al., 2020), Citrobacter (Mullineaux-Sanders et al., 2019), and Enterobacter (Yurdakul et al., 2015) were regarded as pathogenic bacteria involved in inducing inflammation or generating cancer development. Further studies will be conducted to investigate the mechanisms of how these gut microbiota influence lung cancer occurrence, progression and prognosis.

Gut microbiota interaction is a key factor of the microbial equilibrium. Our correlation networks results demonstrated that the microbial network was complexed in the early stage of lung cancer. These results suggest changes in gut microbial homeostasis in the early stages of lung cancer. Our results also showed the network indices including network density, clustering coefficient and average degree was significantly different between HP and lung cancer. Whether they could be used as quantitative parameters to assess cancer risk and homeostasis of the lung microbiome requires further study.

In addition, the predicted 16S functions showed that there were significant differences between the different groups. These results were in agreement with our hypothesis which showed that in the early stages of lung carcinogenesis, there was no significant disease progression in the AAH/AIS group compared

with the HP group. Therefore, the structure of the intestinal flora was closer compared to that of healthy individuals. In contrast, patients in the IA and MIA groups were at a later stage of lung cancer development and had a more altered flora structure compared to the healthy individuals. We thought that the tumor cells may produce metabolites and exhibit different characteristics, and the metabolic disorders and tumor abnormalities may progressively worsen as the disease progresses. On the other hand, the harmful flora in the lung cancer group was also reduced. To a certain extent, this was also a manifestation of the imbalance of the intestinal flora. The Anaerotruncus genus belongs to the Clostridium and participates in the carbohydrate metabolism pathway. The final metabolites are beneficial acetic and butyric acids. A previous study demonstrated that the abundance of Anaerotruncus was significantly increased in the intestinal flora of a mouse model with non-alcoholic fatty liver-related cancer fed on high diet cholesterol (Zhang et al., 2021). Besides, Anaerotruncus was significantly enriched in the uterine microbiome of patients with endometrial cancer (Walther-António et al., 2016).

The KEGG and COG analysis also showed significant differences in the intestinal flora between lung cancer patients and healthy individuals. Further functional analysis of the intestinal bacteria revealed that the flora in lung cancer patients was associated with carbohydrate digestion and absorption. Our findings showed the same metabolic disorders and tumor abnormalities in the intestinal flora. These bacteria may shed different microbial bioactive molecules and affect the utilization of valine, leucine and isoleucine biosynthesis, arginine biosynthesis, glutamate synthesis, glucose, ribose/galactoside by the host. Firmicutes could alter undigested carbohydrates and proteins into acetate, which then produces energy for the organism (Liu et al., 2019a). Furthermore, the reduced abundance of the ABC (ATP-binding cassette) transporter system suggested the potential for energetic and metabolic alterations in the microbiota in lung cancer. This observation is consistent with the hypothesis that lung cancer is fundamentally a metabolic disease and that lung cancer patients often exhibit coexisting metabolic disorder phenotypes and pathologies.

Existing data focused on comparative analysis of intestinal flora changes, which investigated the characteristics of the changes in intestinal flora in different lung cancer histopathology. However, to our knowledge, there are no studies on the relationship between intestinal flora and the development of different histopathological lung cancers. Our study compared the structure of intestinal flora in healthy individuals and patients with different histopathology types in early stage lung cancer. These findings may provide new insights into the development of lung cancer, suggesting that the intestinal flora may be closely related to the progression of lung cancer which can help determine the stage of the disease. Using various bioinformatics methods, such as α-diversity and β-diversity analysis, we identified intestinal flora in lung cancer patients. The population structure of the lung cancer patients was different from the healthy population, which was consistent with previous results. However, there was no overall imbalance in the structure of the intestinal flora in patients with early lung cancer, indicating that the imbalance does not significantly affect the occurrence and development of lung cancer. Meanwhile, the observation of dynamic observation with larger scale were needed in the future.

CONCLUSION

We classified lung cancer patients with different histopathology types and performed a detailed study to characterize the structure of intestinal flora. Our results revealed that the different histopathology types of lung cancer were associated with structural changes in the intestinal flora. AAH/AIS group had a more similar structure to the HP group, while the IA and MIA groups showed a greater change in the colony structure. Lung cancer gut microbiome showed a decrease in SCFA-producing and anti-inflammatory bacteria compared to healthy people, while some pathogenic bacteria such as proinflammatory or tumor-promoting bacteria were more abundant in lung cancer patients. Our findings would provide clues for the use of intestinal flora as a biomarker in the assessment of lung cancer progression and the effective development of targeted therapy.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/, PRJNA772805.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of Yueyang Hospital of Integrated Traditional Chinese and Western Medicine Affiliated to Shanghai University of Traditional Chinese Medicine. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

XQ, HX, and LJ contributed to the conception and design of the study. LB, WY, YH, and XY performed the experiments. YGu, YY, and YW performed the statistical analysis. WY, LB, and LJ wrote the first draft of the manuscript. LX and YGo edited the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This work was supported by grants from Pilot project of Chinese and Western medicine clinical cooperation of Shanghai Municipal Health Commission (ZXYXZ-201901), the National

Natural Science Foundation of China (82173382 and 81973810) and the Shanghai Sailing Program (20YF1450800).

ACKNOWLEDGMENTS

We thank all the patients and healthy volunteers for participating in this study providing the fecal samples and completing the data collection.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.918823/full#supplementary-material

Supplementary Figure 1 | Sequencing data. **(A)** Quality sequence length distribution map. **(B)** OTU rarefaction curves of intestinal flora in four groups of stool samples.

REFERENCES

Allemani, C., Matsuda, T., Di Carlo, V., Harewood, R., Matz, M., Niksic, M., et al. (2018). Global surveillance of trends in cancer survival 2000-14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. *Lancet* 391, 1023–1075. doi: 10.1016/s0140-6736(17)33326-3

Bisht, V., Nash, K., Xu, Y., Agarwal, P., Bosch, S., Gkoutos, G. V., et al. (2021). Integration of the Microbiome, Metabolome and Transcriptomics Data Identified Novel Metabolic Pathway Regulation in Colorectal Cancer. *Int. J. Mol. Sci.* 22:5763. doi: 10.3390/ijms22115763

Budden, K. F., Gellatly, S. L., Wood, D. L., Cooper, M. A., Morrison, M., Hugenholtz, P., et al. (2017). Emerging pathogenic links between microbiota and the gut-lung axis. *Nat. Rev. Microbiol.* 15, 55–63. doi: 10.1038/nrmicro.2016.142

Chaput, N., Lepage, P., Coutzac, C., Soularue, E., Le Roux, K., Monot, C., et al. (2017). Baseline gut microbiota predicts clinical response and colitis in metastatic melanoma patients treated with ipilimumab. *Ann. Oncol.* 28, 1368–1379. doi: 10.1093/annonc/mdx108

Cheng, W. Y., Wu, C. Y., and Yu, J. (2020). The role of gut microbiota in cancer treatment: friend or foe? *Gut* 69, 1867–1876. doi: 10.1136/gutjnl-2020-321153

Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L., and Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* 38, 1767–1771. doi: 10.1093/nar/gkp1137

Daniel, S. G., Ball, C. L., Besselsen, D. G., Doetschman, T., and Hurwitz, B. L. (2017). Functional Changes in the Gut Microbiome Contribute to Transforming Growth Factor β-Deficient Colon Cancer. *mSystems* 2, e00065–17. doi: 10.1128/mSystems.00065-17

Detterbeck, F. C., Boffa, D. J., Kim, A. W., and Tanoue, L. T. (2017). The Eighth Edition Lung Cancer Stage Classification. *Chest* 151, 193–203. doi: 10.1016/j.chest.2016.10.010

Donaldson, G. P., Lee, S. M., and Mazmanian, S. K. (2016). Gut biogeography of the bacterial microbiota. *Nat. Rev. Microbiol.* 14, 20–32. doi: 10.1038/nrmicro3552

Douglas, G. M., Maffei, V. J., Zaneveld, J. R., Yurgel, S. N., Brown, J. R., Taylor, C. M., et al. (2020). PICRUSt2 for prediction of metagenome functions. *Nat. Biotechnol.* 38, 685–688. doi: 10.1038/s41587-020-0548-6

Feng, Q., Chen, W. D., and Wang, Y. D. (2018). Gut Microbiota: An Integral Moderator in Health and Disease. *Front. Microbiol.* 9:151. doi: 10.3389/fmicb.2018.00151

Finlay, B. B., Goldszmid, R., Honda, K., Trinchieri, G., Wargo, J., and Zitvogel, L. (2020). Can we harness the microbiota to enhance the efficacy of cancer immunotherapy? *Nat. Rev. Immunol.* 20, 522–528. doi: 10.1038/s41577-020-0374-6

Gopalakrishnan, V., Spencer, C. N., Nezi, L., Reuben, A., Andrews, M. C., Karpinets, T. V., et al. (2018). Gut microbiome modulates response to anti-PD-1 immunotherapy in melanoma patients. *Science* 359, 97–103. doi: 10.1126/science.aan4236

Jian, X., Zhu, Y., Ouyang, J., Wang, Y., Lei, Q., Xia, J., et al. (2020). Alterations of gut microbiome accelerate multiple myeloma progression by increasing the relative abundances of nitrogen-recycling bacteria. *Microbiome* 8:74. doi: 10.1186/s40168-020-00854-5

Kadosh, E., Snir-Alkalay, I., Venkatachalam, A., May, S., Lasry, A., Elyada, E., et al. (2020). The gut microbiome switches mutant p53 from tumour-suppressive to oncogenic. *Nature* 586, 133–138. doi: 10.1038/s41586-020-2541-0

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., et al. (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* 36, D480–D484. doi: 10.1093/nar/gkm882

Langille, M. G., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., et al. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* 31, 814–821. doi: 10.1038/nbt.2676

Lee, S. H., Sung, J. Y., Yong, D., Chun, J., Kim, S. Y., Song, J. H., et al. (2016). Characterization of microbiome in bronchoalveolar lavage fluid of patients with lung cancer comparing with benign mass like lesions. *Lung Cancer* 102, 89–95. doi: 10.1016/j.lungcan.2016.10.016

Leng, Q., Holden, V. K., Deepak, J., Todd, N. W., and Jiang, F. (2021). Microbiota Biomarkers for Lung Cancer. *Diagnostics* 11:407. doi: 10.3390/diagnostics11030407

Liang, J. Q., Li, T., Nakatsu, G., Chen, Y. X., Yau, T. O., Chu, E., et al. (2020). A novel faecal Lachnoclostridium marker for the non-invasive diagnosis of colorectal adenoma and cancer. *Gut* 69, 1248–1257. doi: 10.1136/gutjnl-2019-318532

Liu, F., Li, J., Guan, Y., Lou, Y., Chen, H., Xu, M., et al. (2019a). Dysbiosis of the Gut Microbiome is associated with Tumor Biomarkers in Lung Cancer. *Int. J. Biol. Sci.* 15, 2381–2392. doi: 10.7150/ijbs.35980

Liu, S., Li, E., Sun, Z., Fu, D., Duan, G., Jiang, M., et al. (2019b). Altered gut microbiota and short chain fatty acids in Chinese children with autism spectrum disorder. *Sci. Rep.* 9:287. doi: 10.1038/s41598-018-36430-z

Lukens, J. R., Gurung, P., Vogel, P., Johnson, G. R., Carter, R. A., McGoldrick, D. J., et al. (2014). Dietary modulation of the microbiome affects autoinflammatory disease. *Nature* 516, 246–249. doi: 10.1038/nature13788

Magoč, T., and Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 27, 2957–2963. doi: 10.1093/bioinformatics/btr507

Mullineaux-Sanders, C., Sanchez-Garrido, J., Hopkins, E. G. D., Shenoy, A. R., Barry, R., and Frankel, G. (2019). Citrobacter rodentium-host-microbiota interactions: immunity, bioenergetics and metabolism. *Nat. Rev. Microbiol.* 17, 701–715. doi: 10.1038/s41579-019-0252-z

Nejman, D., Livyatan, I., Fuks, G., Gavert, N., Zwang, Y., Geller, L. T., et al. (2020). The human tumor microbiome is composed of tumor type-specific intracellular bacteria. *Science* 368, 973–980. doi: 10.1126/science.aay9189

O'Keefe, S. J. (2016). Diet, microorganisms and their metabolites, and colon cancer. *Nat. Rev. Gastroenterol. Hepatol.* 13, 691–706. doi: 10.1038/nrgastro.2016.165

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59–65. doi: 10.1038/nature08821

Rathje, K., Mortzfeld, B., Hoeppner, M. P., Taubenheim, J., Bosch, T. C. G., and Klimovich, A. (2020). Dynamic interactions within the host-associated microbiota cause tumor formation in the basal metazoan Hydra. *PLoS Pathog.* 16:e1008375. doi: 10.1371/journal.ppat.1008375

Sanders, M. E., Merenstein, D. J., Reid, G., Gibson, G. R., and Rastall, R. A. (2019). Probiotics and prebiotics in intestinal health and disease: from biology to the clinic. *Nat. Rev. Gastroenterol. Hepatol.* 16, 605–616. doi: 10.1038/s41575-019-0173-3

Singh, T. P., and Natraj, B. H. (2021). Next-generation probiotics: a promising approach towards designing personalized medicine. *Crit. Rev. Microbiol.* 47, 479–498. doi: 10.1080/1040841x.2021.1902940

Travis, W. D., Brambilla, E., Burke, A. P., Marx, A., and Nicholson, A. G. (2015). Introduction to The 2015 World Health Organization Classification of Tumors of the Lung. Pleura, Thymus, and Heart. *J. Thorac. Oncol.* 10, 1240–1242. doi: 10.1097/jto.0000000000000663

Wagner, G., Stollenwerk, H. K., Klerings, I., Pecherstorfer, M., Gartlehner, G., and Singer, J. (2020). Efficacy and safety of immune checkpoint inhibitors in patients with advanced non-small cell lung cancer (NSCLC): a systematic literature review. *Oncoimmunology* 9:1774314. doi: 10.1080/2162402x.2020.1774314

Walther-António, M. R., Chen, J., Multinu, F., Hokenstad, A., Distad, T. J., Cheek, E. H., et al. (2016). Potential contribution of the uterine microbiome in the development of endometrial cancer. *Genome Med.* 8:122. doi: 10.1186/s13073-016-0368-y

Yurdakul, D., Yazgan-Karataş, A., and Şahin, F. (2015). *Enterobacter* Strains Might Promote Colon Cancer. *Curr. Microbiol.* 71, 403–411. doi: 10.1007/s00284-015-0867-x

Zhang, L., Yue, Y., Shi, M., Tian, M., Ji, J., Liao, X., et al. (2020). Dietary Luffa cylindrica (L.) Roem promotes branched-chain amino acid catabolism in the circulation system via gut microbiota in diet-induced obese mice. *Food Chem.* 320:126648. doi: 10.1016/j.foodchem.2020.126648

Zhang, W., Luo, J., Dong, X., Zhao, S., Hao, Y., Peng, C., et al. (2019). Salivary Microbial Dysbiosis is Associated with Systemic Inflammatory Markers and Predicted Oral Metabolites in Non-Small Cell Lung Cancer Patients. *J. Cancer* 10, 1651–1662. doi: 10.7150/jca.28077

Zhang, X., Coker, O. O., Chu, E. S., Fu, K., Lau, H. C. H., Wang, Y. X., et al. (2021). Dietary cholesterol drives fatty liver-associated liver cancer by modulating gut microbiota and metabolites. *Gut* 70, 761–774. doi: 10.1136/gutjnl-2019-319664

Zheng, Y., Fang, Z., Xue, Y., Zhang, J., Zhu, J., Gao, R., et al. (2020). Specific gut microbiome signature predicts the early-stage lung cancer. *Gut Microbes* 11, 1030–1042. doi: 10.1080/19490976.2020.1737487

Zhong, H., Penders, J., Shi, Z., Ren, H., Cai, K., Fang, C., et al. (2019). Impact of early events and lifestyle on the gut microbiota and metabolic phenotypes in young school-age children. *Microbiome* 7: 2. doi: 10.1186/s40168-018-0608-z

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 22 August 2022
DOI 10.3389/fmicb.2022.969757

OPEN ACCESS

EDITED BY
Liang Wang,
Guangdong Provincial People's Hospital, China

REVIEWED BY
Rosa Del Campo,
Ramón y Cajal Institute for Health Research, Spain
George Grant,
University of Aberdeen, United Kingdom
Sergio Perez-Burillo,
Public University of Navarre, Spain
Douglas Maya Miles,
Institute of Biomedicine of Seville (CSIC), Spain

*CORRESPONDENCE
Zhenmei An
azmhxnfm@163.com
Ga Liao
Liaoga.scu@qq.com

SPECIALTY SECTION
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

RECEIVED 15 June 2022
ACCEPTED 05 August 2022
PUBLISHED 22 August 2022

CITATION
Yang L, Dai Y, He H, Liu Z, Liao S, Zhang Y, Liao G and An Z (2022) Integrative analysis of gut microbiota and fecal metabolites in metabolic associated fatty liver disease patients.
Front. Microbiol. 13:969757.
doi: 10.3389/fmicb.2022.969757

Integrative analysis of gut microbiota and fecal metabolites in metabolic associated fatty liver disease patients

Lidan Yang[1], Yuzhao Dai[2], He He[1], Zhi Liu[1], Shenling Liao[1], Yu Zhang[2], Ga Liao[3,4]* and Zhenmei An[2]*

[1]Department of Laboratory Medicine, West China Hospital, Sichuan University, Chengdu, China, [2]Department of Endocrinology and Metabolism, West China Hospital, Sichuan University, Chengdu, China, [3]State Key Laboratory of Oral Diseases, National Clinical Research Center for Oral Diseases, West China Hospital of Stomatology, Sichuan University, Chengdu, China, [4]Department of Information Management, Department of Stomatology Informatics, West China Hospital of Stomatology, Sichuan University, Chengdu, China

Objective: Metabolic associated fatty liver disease (MAFLD) affects nearly a quarter of the world's population. Our study aimed to characterize the gut microbiome and overall changes in the fecal and serum metabolomes in MAFLD patients.

Methods: Thirty-two patients diagnosed with MAFLD and 30 healthy individuals (control group, CG) were included in this study, the basic clinical characteristics and laboratory test results including routine biochemistry, etc. were recorded for all, and their serum and fecal samples were collected. A portion of the fecal samples was subjected to 16S rDNA sequencing, and the other portion of the fecal samples and serum samples were subjected to non-targeted metabolomic detection based on liquid chromatography-mass spectrometry (LC–MS). Statistical analysis of clinical data was performed using SPSS software package version 25.0 (SPSS Inc., Chicago, IL, United States). The analysis of 16S rDNA sequencing results was mainly performed by R software (V. 2.15.3), and the metabolomics data analysis was mainly performed by CD 3.1 software. Two-tailed p value<0.05 was considered statistically significant.

Results: The 16S sequencing data suggested that the species richness and diversity of MAFLD patients were reduced compared with controls. At the phylum level, the relative abundance of *Bacteroidota*, *Pseudomonadota*, and *Fusobacteriota* increased and *Bacillota* decreased in MAFLD patients. At the genus level, the relative abundances of *Prevotella*, *Bacteroides*, *Escherichia-Shigella*, etc. increased. 2,770 metabolites were detected in stool samples and 1,245 metabolites were detected in serum samples. The proportion of differential lipid metabolites in serum (49%) was higher than that in feces (21%). There were 22 differential metabolites shared in feces and serum. And the association analysis indicated that LPC 18:0 was positively correlated with *Christensenellaceae_R-7_group*, *Oscillospiraceae_UCG-002*; neohesperidin was also positively correlated with *Peptoniphilus*, *Phycicoccus*, and *Stomatobaculum*.

Conclusion: Microbial sequencing data suggested decreased species richness and diversity and altered β-diversity in feces. Metabolomic analysis identified

overall changes in fecal and serum metabolites dominated by lipid molecules. And the association analysis with gut microbes provided potentially pivotal gut microbiota-metabolite combinations in MAFLD patients, which might provide new clues for further research on the disease mechanism and the development of new diagnostic markers and treatments.

KEYWORDS

metabolic associated fatty liver disease, intestinal microflora, metabolomics, non-alcoholic fatty liver disease, lipid metabolites

Introduction

Metabolic associated fatty liver disease (MAFLD), a new definition of fatty liver officially proposed by an international expert group in 2020 after the non-alcoholic fatty liver disease (NAFLD; Eslam et al., 2020a), affects at least one-quarter of the adult population worldwide (Powell et al., 2021). MAFLD is closely associated with metabolic disease and its complications, and it has rapidly become one of the leading causes of hepatocellular carcinoma and cirrhosis in Western countries (Younossi et al., 2019). The main complication causing death in patients with MAFLD is CVD. However, liver-related complications are more common in patients with advanced fibrosis or cirrhosis and account for the majority of deaths (Lin et al., 2021). The currently widely accepted treatment recommendations for MAFLD are lifestyle changes aimed at weight loss, and there are no drugs approved for the therapy of MAFLD at this stage (Eslam et al., 2020b; Fouad et al., 2022). Therefore, it is crucial to continue to explore the mechanisms associated with MAFLD and develop new treatments.

In the process of hepatic steatosis and its progression to liver inflammation and liver fibrosis, MAFLD involves the interaction of multiple metabolic, environmental, genetic, and microbial factors (Friedman et al., 2018; Lin et al., 2020). Altered gut-liver axis, increased susceptibility to hepatic triglyceride accumulation, altered lipid metabolism, dyslipidemia, and insulin resistance are key components of the pathophysiology of MAFLD. Notably, multiple studies have shown that the gut-liver axis is closely related to metabolic syndrome, obesity, and type 2 diabetes (Leung et al., 2016; Mardinoglu et al., 2019; Yuan et al., 2019). A high-fat, high-sugar diet and a sedentary lifestyle promote adipogenesis and subclinical inflammation in the intestines, adipose tissue, and liver. Furthermore, this metabolic inflammation in adipose tissue and intestine can promote hepatic adipogenesis and aggravate inflammation through cytokines, fatty acids, dysbiosis of gut flora, and gut barrier disruption (Friedman et al., 2018). The intestinal microbiota is considered to be a new metabolic organ involved in the regulation of host metabolism. The association between microbiota and the pathogenesis of MAFLD has placed those small organisms as a critical focus in MAFLD research. However, the relationship between the gut microbiome and metabolism in MAFLD patients has not been established. Therefore, a comprehensive analysis of the gut microbiome and metabolome may help us uncover the complexity of MAFLD.

Here, we performed 16S gut microbiome sequencing and untargeted metabolomics studies in MAFLD patients and healthy volunteers. We revealed the disruption of gut microbiota homeostasis and the changes in fecal and serum metabolism in patients with MAFLD. In addition, we constructed a map showing the correlation of gut microbiota with fecal and blood metabolism, revealing possible key gut microbe-metabolite combinations, and laying a foundation for further study of the disease mechanism of MAFLD.

Materials and methods

Subject enrollment

Thirty-two NAFLD patients and 30 healthy volunteers were recruited from July 2019 to February 2020 at the West China Hospital of Sichuan University (Sichuan Province, China). NAFLD patients were newly diagnosed outpatients and were diagnosed according to the clinical diagnostic criteria recommended by the Chinese Association for the Study of Liver Diseases and the American Association for the Study of Liver Diseases (Fan et al., 2011; Chalasani et al., 2018). The detailed inclusion and exclusion criteria of the case and control groups can be found in the Supplementary material. Patients' basic clinical characteristics and relevant laboratory test results were recorded, including age, sex, height, weight, body mass index (BMI), and routine biochemical test results (TBIL, DBIL, IBIL, ALT, AST, TP, ALB, GLB, fasting glucose, UREA, CREA, eGFR, URIC, TG, CHOL, HDL-C, LDL-C, ALP, and GGT), thyroid function test results (TSH, FT3, and FT4), etc. This study was approved by the Institutional Review Board of West China Hospital, Sichuan University, exempted from informed consent, and conducted following the Declaration of Helsinki.

Sample processing

Fecal and serum samples from each volunteer were collected on the same day. Volunteers self-collected samples after defecation in the hospital and immediately transferred the samples to a

laboratory freezer at −80°C for cryopreservation. Blood samples were collected from fasting venous blood, placed at room temperature to stratify, and centrifuged at 3,000 rpm for 10 min. Serum samples were collected and frozen in a −80 freezer.

DNA extraction from stool samples was performed using the sodium dodecyl sulfate (SDS) method. Before amplification, check the purity and concentration of DNA by electrophoresis, and dilute the sample DNA with sterile water to obtain the amplification template (target concentration was 1 ng/μl). The 16S V3-V4 region was selected as the amplified region in this study (Abellan-Schneyder et al., 2021). The PCR was completed using specific primers with "barcode short sequences" (used to distinguish each sample), buffers that provide GC bases (Phusion® High-Fidelity PCR Master Mix, New England Biolabs), and high-efficiency, high-fidelity enzymes. The samples were mixed in the same volume according to the PCR product concentration. Purification was carried out using 2% agarose gel electrophoresis, and finally, the target band was recovered with a gel recovery kit (Qiagen). The PCR-free library was constructed using the TruSeq® DNA PCR-Free Sample Preparation Kit based on the Illumina Nova sequencing platform, followed by paired-end sequencing (Caporaso et al., 2012). And Qubit and Q-PCR quantitative detection were used to judge whether the library was qualified or not before running on the computer (NovaS eq6000).

For metabolomics sample processing, firstly 100 mg of liquid nitrogen-ground fecal samples were placed in an EP tube, and 500 μl of 80% methanol in water was added. 100 μl of serum sample was placed in an EP tube, and 400 μl of 80% methanol in water was added. Vortex and shake, stand in an ice bath for 5 min, centrifuge at 15,000 rpm and 4°C for 10 min, take a certain amount of supernatant and add mass spectrometry-grade water to dilute to 53% methanol, and place it in a centrifuge tube at 15,000 *g* and centrifuge at 4°C 10 min. The supernatant was collected and analyzed by liquid chromatography-mass spectrometry technology (LC–MS; Alseekh et al., 2021). In addition, an equal volume of samples was taken from each experimental sample and mixed well as a quality control sample for equilibrating the chromatography-mass spectrometry system and monitoring the instrument status, and evaluating the system stability throughout the experimental process. At the same time, a blank sample was set, which was a 53% methanol aqueous solution containing 0.1% formic acid. The pretreatment process was the same as that of the experimental sample and was mainly used to remove background ions.

Statistical analysis

Statistical analysis of clinical data was performed using the SPSS software package version 25.0 (SPSS Inc., Chicago, IL, United States). The continuous variables were tested for normality first. Variables with homogeneity of normal variance were expressed as mean ± SD, and a *t*-test was used for comparison between groups; variables that were normal but with unequal variance were expressed as mean ± SD, and the Wilcoxon rank-sum test was used for comparison between groups; non-normal variables were expressed as medians (upper and lower quartiles), and the Wilcoxon rank-sum test was used for comparison between groups. Categorical variables were expressed as frequency (percentage), and the chi-square test was used for comparison between groups. Two-tailed value of $p < 0.05$ was considered statistically significant.

The analysis of the results of 16S rDNA sequencing was mainly done using R software (V. 2.15.3). Using Uparse v7.0.1001 software to cluster effective sequences into operational taxonomic units (OTUs) with 97% consistency, and then performed species annotation analysis according to the SILVA132 SSUrRNA database. The data with the least amount of data in the sample were used as the standard to normalize the data to obtain the relative abundance value of the species. Using Qiime software (V. 1.9.1) to calculate the alpha diversity index (including Observed species, Good's coverage, Chao1, ACE, Shannon, and Simpson index) and beta diversity index (Unifrac distance and Bray-Curtis distance). T-test and Wilcox test were used for inter-group difference analysis of diversity index. Using R software for principal component analysis (PCA) and principal coordinates analysis (PCoA). Finally, R software was used for a routine *t*-test to obtain taxons with significant differences between groups (value of $p < 0.05$); Furtherly, using LEfSe software, taxons with significant differences between the two groups [linear discriminant analysis (LDA) index > 4] were screened.

Data analysis of non-targeted metabolic results used CD 3.1 software, combined with the mzCloud, mzVault, and MassList database for identification and processing to obtain metabolite qualitative and quantitative results. The final identification results were selected from the compounds with a coefficient of variation value of less than 30% in the quality control samples. Compounds were functionally and taxonomically annotated with the KEGG, Human Metabolome Database (HMDB), and LIPID MAPS databases. The partial least squares discriminant analysis model (PLS-DA) was obtained by multivariate statistical analysis. In order to evaluate the reliability of the model, the PLS-DA model of each group was first established, and the model evaluation parameters (R2, Q2) were obtained through 7-fold cross-validation. The closer the values of R2 and Q2 were to 1, the more stable and reliable the model was. Then, the grouping marks of each sample were randomly scrambled, and further modeling and prediction were performed to determine whether the model was "overfitting." Each modeling corresponded to a set of R2 and Q2 values, and their regression lines were drawn based on the Q2 and R2 values after 200 scrambles and modeling. When R2 was greater than Q2 and the Q2 regression line and the *Y*-axis intercept were less than 0, it could indicate that the model was not "overfitting." By calculating the variable importance in projection (VIP) value and fold change (FC) of the first principal component, and combining it with a T-test to find differentially expressed metabolites, setting the screening threshold to VIP > 1.0, FC > 1.5, or FC<0.667, and value of $p < 0.05$. The correlation analysis of

differential metabolites and differential flora was performed using Pearson correlation analysis. Based on the RandomForest analysis, the genus-level taxons and metabolome data were separately divided into a test set and validation set (7:3), and then the test set was used to build a random forest model. Important taxons or metabolites were screened out according to MeanDecreaseAccuracy and MeanDecreaseGin, and then each model was cross-validated (10-fold) and ROC curves were drawn.

Results

Characterization of participants

A total of 30 healthy controls (control group, CG) and 32 MAFLD volunteers (MAFLD group) were included in this study. The two groups had comparable ages [MAFLD group, 38.50 (33.00–51.75) years; CG, 35.33 (32.50–51.25) years], and the difference was not statistically significant. The basic screen of the participants showed that the BMI of the MAFLD group (26.21 ± 3.80) was significantly higher than that of the CG (23.83 ± 3.28), and the difference was statistically significant ($p < 0.05$). Among laboratory indicators, serum AST, ALT, ALP, GGT, fast glucose, TG, and URIC levels in the MAFLD group were higher than those in the CG, and HDL-C was lower than that in the CG, and the differences were statistically significant. In addition, the serum levels of TB, TP, ALB, and CREA in the MAFLD group were higher than those in the CG, and the difference was not statistically significant. We calculated the Fibrosis-4 index (FIB-4) of all people, and the results showed that the results of the MAFLD group [1.25 (0.70–1.92)] were higher than those of the control group [0.65 (0.33–0.85)], but the difference was not statistically significant. In the MAFLD group, FIB-4 < 1.3 and FIB-4 between 1.3 and 2.67 each accounted for 50%. The results of the serum thyroid function test showed that compared with the CG, the MAFLD patients had increased TSH and decreased FT4 and FT3, but the differences were not statistically significant (Supplementary Table S1; Supplementary Figure S1).

Altered gut microbiota diversity in MAFLD patients

An average of 104,138 tags was detected per sample by splicing reads. After quality control, an average of 97,013 pieces of effective data was obtained, and the effective rate of quality control was 61.89%. 1,882 OTUs were obtained by clustering the sequences with 97% identity. According to the rarefaction curve (Figure 1A) and species accumulation boxplot (Supplementary Figure S2A), the current amount of sequencing data and the sample size were reasonable. In addition, the rank abundance curve (Supplementary Figure S2B) and the analysis results of alpha diversity indices (Shannon index, Simpson index, etc.) showed that the species richness and diversity of MAFLD patients were reduced compared with the CG (Figures 1C,D). The difference in beta diversity was observed by PCoA analysis of unifrac distance (Figures 1E,F). In addition, the results of MRPP analysis ($p < 0.001$) and ANOSIM analysis ($p < 0.001$) indicated significant differences in community structure between the MAFLD group and the CG.

The number of OTUs that could be annotated into the database was 1,866 (99.15%). The proportions of annotations at the kingdom level, phylum level, class level, order level, family level, genus level, and species level were 99.15, 91.29, 90.12, 85.44, 79.17, 54.89, and 18.07%, respectively. At the phylum level, we found that the dominant taxa included *Bacillota* (previous name: *Firmicutes*), *Pseudomonadota* (previous name: *Proteobacteria*), *Actinomycetota* (previous name: *Actinobacteria*), and *Bacteroidota* (previous name: *Bacteroidetes*; Supplementary Figure S3A); The dominant genera were *Escherichia-Shigella*, *Bifidobacterium*, and *Prevotella* et al. (Figure 1B); the dominant species were *Escherichia_coli*, *Raoultella_ornithinolytica*, and *Bacteroides_vulgatus* et al. (Supplementary Figure S3B). Through LEfSe analysis, there were 39 taxa (including six grading levels) with LDA value > 4 between the two groups (Figure 2A), and their evolutionary branch diagram was shown in Figure 2B. At the phylum level, the relative abundance of *Bacteroidota*, *Pseudomonadota*, and *Fusobacteriota* increased and *Bacillota* decreased in MAFLD patients. At the genus level, the relative abundances of *Prevotella*, *Bacteroides*, *Escherichia-Shigella*, *Megamonas*, *Fusobacterium*, and *Lachnoclostridium* increased, while *Clostridium_sensu_stricto_1*, *Agathobacter*, *Romboutsia*, *Faecalibacterium*, *Blautia* decreased. Species with increasing relative abundance were *Escherichia_coli*, *Bacteroides_vulgatus*, and species with decreasing relative abundance were *Romboutsia_ilealis*.

Serum and fecal metabolite profiling in MAFLD patients

A total of 2,770 metabolites were identified in fecal samples, and a total of 1,245 metabolites were identified in serum samples. The classification results of 997 metabolites in fecal samples and 400 in serum samples were obtained through the HMDB (Supplementary Figures S4A,B, S5A,B), of which Lipids and lipid-like molecules were the most classified. Then, we obtained the classification and annotation results of 305 lipid metabolites in fecal samples and 212 in serum samples through the LIPID MAPS database (Supplementary Figures S6A,B, S7A,B), among which Fatty Acids metabolites accounted for the most.

For differential metabolites screening, we performed PLS-DA on the resulting data (Figures 3A,B; Supplementary Figures S8A,B), and the ranking validation results show that the PLS-DA model was not "overfit" (Supplementary Figures S9A–D). Then, we screened out differential metabolites with VIP > 1.0, FC > 1.5 or FC < 0.667, and value of $p < 0.05$ (Table 1). 34% of differential metabolites in

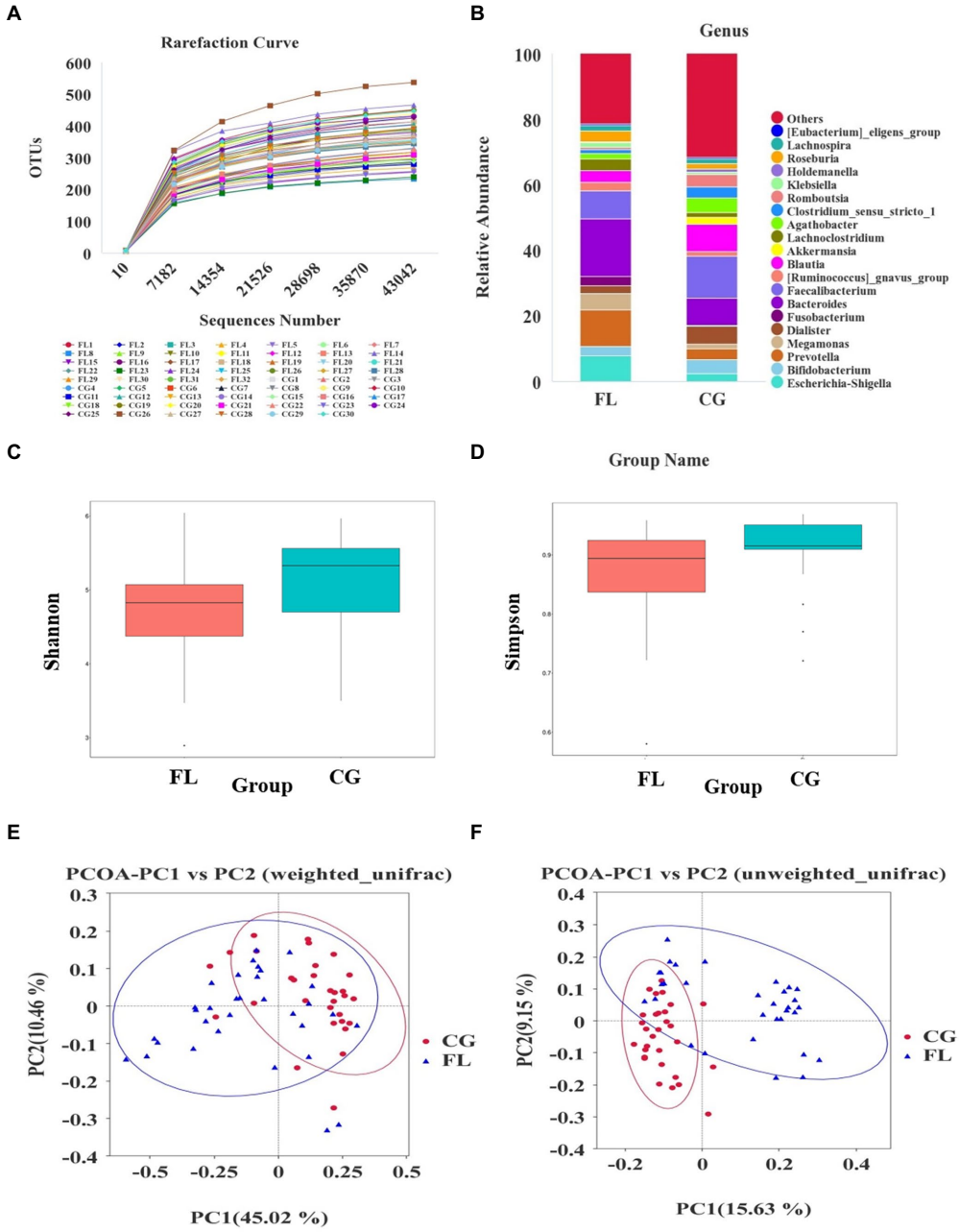

FIGURE 1
Altered gut microbiota diversity in metabolic associated fatty liver disease (MAFLD) patients. **(A)** Rarefaction curve based on OUT count in control groups (CGs) and MAFLD patients. **(B)** The relative abundance of dominant taxa at the genus level in each group. **(C,D)** The analysis results of alpha diversity indices (Shannon index and Simpson index), both $p<0.05$. **(E,F)** Principal coordinates analysis (PCoA) analysis based on weighted unifrac distance and unweighted unifrac distance.

fecal and 40% in serum were classified in HBDM and/or LIPID MAPS database, which mainly included: amino acids, peptides, and analogs; Lipids and lipid-related molecules; Nucleotides and analogs; carbohydrates and carbohydrate conjugates; Benzene and substituted derivatives, etc. Notably, lipids accounted for a large fraction of the significantly variable metabolites in serum and feces, especially in serum, which suggested a disruption of lipid homeostasis in MAFLD patients (Figure 3C). KEGG pathway enrichment results showed that fecal differential metabolites were more enriched in the biosynthesis of amino acids, purine metabolism, pantothenate and CoA biosynthesis, pyrimidine metabolism, nicotinate, and nicotinamide metabolism pathway (Figures 3D,E); serum differential metabolites were more enriched in purine metabolism, pyrimidine metabolism, bile secretion, and pentose phosphate pathway (Supplementary Figures S8C,D). Further analysis found

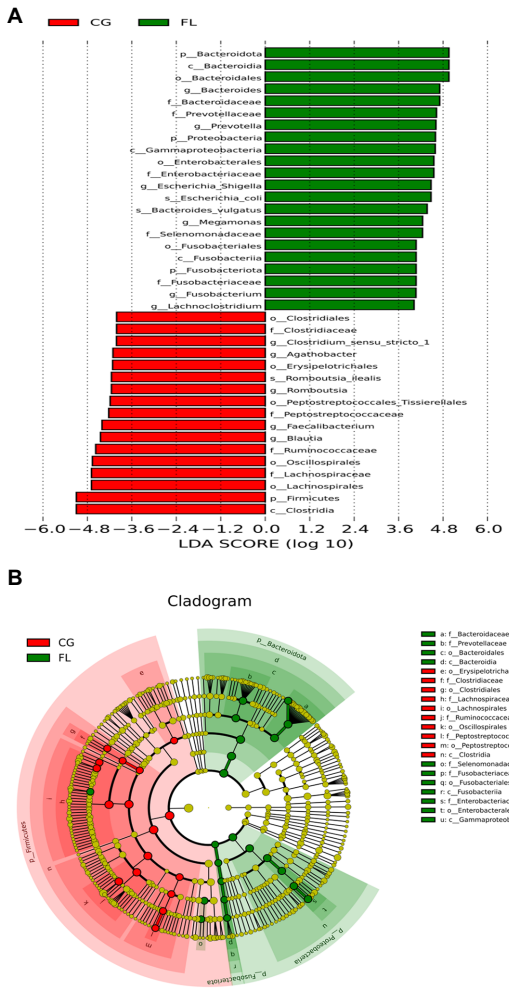

FIGURE 2
Linear discriminant analysis (LDA) effect size analysis.
(A) Histogram of the LDA scores for different abundant taxa. Green, enriched in MAFLD patients; red, enriched in CG.
(B) Cladogram of LEfSe linear discriminant analysis. Red and green circles represent the differences of the most abundant microbiome class. The diameter of each circle is proportional to the relative abundance of the taxon.

that there were 22 common differential metabolites in feces and serum (Supplementary Table S2), which mainly included purines and purine derivatives: hypoxanthine; amino acids, peptides, and analogues: methionine, gamma-glu-leu, and tyrosylalanine; fatty esters: propionylcarnitine; glycerophosphocholines: lysophosphatidylcholine (LPC 16:0, LPC 18:0); and flavanones: hesperetin and neohesperidin. In particular, during the differential metabolite analysis, we found some bile acids and derivatives: lithocholic acid and taurocholic acid decreased in the serum of MAFLD patients; glycocholic acid increased in the serum of MAFLD patients; and taurodeoxycholic acid, 7-Ketolithocholic acid, allolithocholic acid, and dehydrocholic acid were decreased in the feces of MAFLD patients.

Correlation between differential bacteria and differential metabolites

Pearson correlation analysis was performed between the top 10 differential bacterial genera in relative abundance and the top 20 differential metabolites in relative abundance, and the differential species-metabolite combinations satisfying $|rho| \geq 0.5$ and $p \leq 0.05$ were screened out (Table 2). Then, we performed a correlation analysis between the common differential metabolites in serum and feces (number = 22) and all differential species in feces (number = 74), and the differential species-metabolite combinations satisfying $|rho| \geq 0.5$ and $p \leq 0.05$ were screened out (Supplementary Table S3). In addition, we correlated the differential bile acids and derivatives with all differential bacterial genera (Supplementary Table S4). Among them, *Erysipelotrichaceae_UCG-003* had a weak positive correlation with serum taurocholic acid (rho = 0.563, $p < 0.05$); allolithocholic acid in feces was associated with *Prevotellaceae_NK3B31_group* (rho = 0.723, $p < 0.05$) and *unidentified_Ruminococcaceae* (rho = 0.797, $p < 0.05$). Finally, we also correlated all differential lipids and lipid-related molecules (number = 44) in serum with fecal differential bacteria (Supplementary Table S5). Among them, the metabolite-genus combinations with strong correlation were: L-Leucyl-L-alanine Hydrate-*Lachnoanaerobaculum* (rho = 0.889, $p < 0.05$), 6-Keto-prostaglandin f1alpha-*Fusicatenibacter* (rho = 0.743, $p < 0.05$), and 6-Keto-prostaglandin f1alpha-*Anaerostipes* (rho = 0.730, $p < 0.05$). The Random Forest analysis results of the genus-level taxons and metabolome data were shown in Supplementary Figures S10–S12.

Discussion

Evidence accumulated from many preclinical and clinical studies had indicated that the communication between the gut microbiota, its metabolites, and the liver plays a crucial role in the pathogenesis of MAFLD. Here, we recruited 32 patients with MAFLD and investigated overall changes in the gut microbiome in feces and the metabolome in serum and feces. In addition, we identified alterations in several gut microbiota-produced metabolites that may influence the pathogenesis of MAFLD.

The human gut microbiota are mainly composed of four phyla—*Bacteroidota, Bacillota, Pseudomonadota,* and *Actinomycetes*, of which *Bacteroidota* and *Bacillota* dominate the gut (Eckburg et al., 2005; Mokhtari et al., 2017). In the present study, we observed decreased species richness and diversity and altered β-diversity in the feces of MAFLD patients, confirming the development of dysbiosis. Specifically, the relative abundance of *Bacteroidota* and *Pseudomonadota* increased and *Bacillota* decreased in MAFLD patients, which is consistent with previous reports (Boursier et al., 2016; Wang et al., 2016). Under *Bacteroidota*, differential taxa analysis showed that the relative abundance of *Prevotella* and *Bacteroides* increased in MAFLD patients, and the relative abundance of *Bacteroides_vulgatus*,

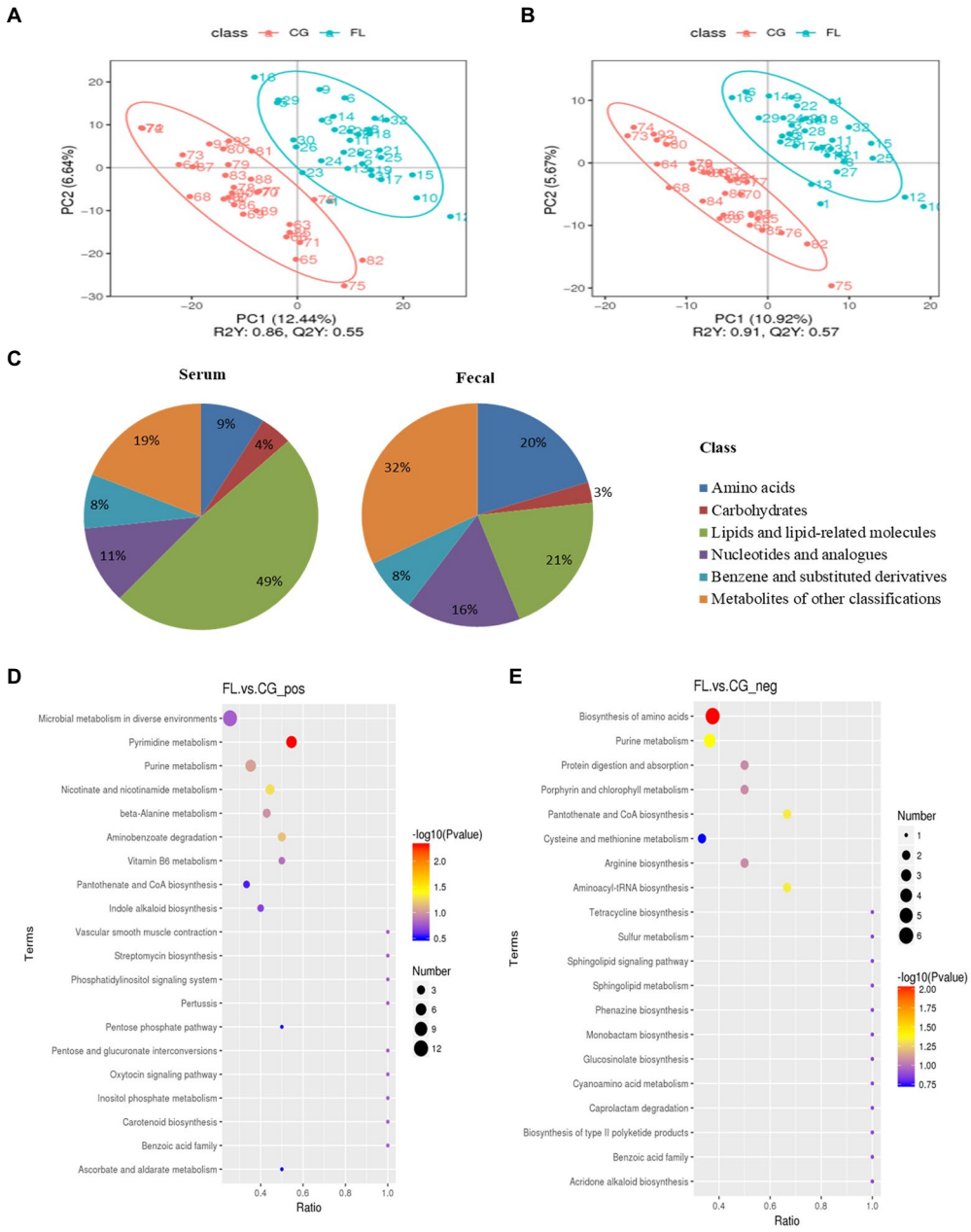

FIGURE 3
(A,B) are the Scatter plot of partial least squares discriminant analysis model (PLS-DA) scores in positive and negative ion mode for fecal metabolites, respectively. The abscissa is the score of the sample on the first principal component, and the ordinate is the score of the sample on the second principal component. R2Y represents the interpretation rate of the model, Q2Y is used to evaluate the predictive ability of the PLS-DA model, and when R2Y is greater than Q2Y, the model is well established. **(C)** The proportion of fecal/serum differential metabolite classification. **(D,E)** are bubble plots of fecal differential metabolite pathway enrichment in positive and negative ion modes, respectively. The abscissa in the figure is the number of differential metabolites in the corresponding metabolic pathway/the total number of metabolites identified in the pathway. The larger the value, the higher the enrichment of differential metabolites in the pathway. The color of the dots represents the value of *p* of the hypergeometric test, and the smaller the value, the greater the reliability of the test. The size of the dots represents the number of differential metabolites in the corresponding pathway.

which belonged to *Bacteroides* at the species level, was also increased. Among the increased *Pseudomonadota*, the relative abundances of *Escherichia coli-Shigella* and *Escherichia_coli* increased. Among the reduced *Bacillota*, the taxa with decreased relative abundance at the genus level include: *Clostridium_sensu_stricto_1*, *Agathobacter*, *Faecalibacterium*, *Blautia*, and *Romboutsia*, and the increased ones include: *Megamonas* and *Lachnoclostridium*. The changes in some genera were consistent with the statistical results of a recent meta-analysis that included 1,265 NAFLD patients (from eight countries;

TABLE 1 Metabolite differential screening results.

Sample type	Number of Total Ident.	Number of Total Sig.	Number of Sig. Up	Number of Sig. down
faeces_pos.	1,888	362	47	315
faeces_neg.	882	138	34	104
serum_pos.	731	143	36	117
serum_neg.	414	82	13	69

(1) pos.: positive ion mode; neg.: negative ion mode; (2) Num of Total Ident: Total identification results of metabolites; (3) Num of Total Sig: The total number of metabolites with significant differences; (4) Num of Sig Up: The total number of metabolites significantly upregulated; and (5) Num of Sig down: The total number of metabolites significantly downregulated.

TABLE 2 Correlation analysis results of fecal differential bacteria and differential metabolites (feces and serum).

	Differential bacteria	Differential metabolites	rho	*p*
Faeces	*Cutibacterium*	Adenosine	0.584	<0.05
	Intestinibacter	YMK	0.514	<0.05
	Intestinibacter	tert-Butyl N-[1-(aminocarbonyl)-3-methylbutyl] carbamate	0.562	<0.05
	Intestinibacter	L-Alanyl-L-proline	0.521	<0.05
	Intestinibacter	3'-Hydroxystanozolol	0.601	<0.05
	Intestinibacter	FQH	0.521	<0.05
	Monoglobus	tert-Butyl N-[1-(aminocarbonyl)-3-methylbutyl] carbamate	0.538	<0.05
	Intestinibacter	1,5-Anhydro-D-glucitol	0.525	<0.05
	Intestinibacter	N-(1-benzothiophen-3-yl)- N′-(1-benzyl-4-piperidinyl)urea	0.560	<0.05
	Lachnospiraceae_UCG-004	3'-Dephospho-CoA	0.525	<0.05
Serum	*Neisseria*	Cnidioside A	0.540	<0.05
	Lachnospiraceae_FCS020_group	PA (16:0/18:2)	0.503	<0.05
	Staphylococcus	LPA 18:2	0.507	<0.05

Li et al., 2021b). In addition, the relative abundance of *Romboutsia_ilealis* (belonging to *Romboutsia*) also decreased at the species level. In particular, among the differential phyla, we also observed an increase of *Fusobacteriota* in MAFLD patients and an increase of *Fusobacterium* below it.

Previous studies have shown that gut dysbiosis usually led to elevated levels of SCFAs in the gut, with acetate and propionate mainly produced by *Bacteroidota* and butyrate by *Bacillota* (Morrison and Preston, 2016; Feng et al., 2018). Elevated SCFAs promoted the transport of monosaccharides to the liver, while increased hepatic acetate (substrate for fatty acid synthesis) led to the accumulation of triglycerides, and elevated hepatic propionate promoted gluconeogenesis, eventually leading to weight gain (den Besten et al., 2013; Alves-Bezerra and Cohen, 2017). Further, supplementation with SCFAs could also alter the composition of the gut microbiome and prevent the occurrence and progression of NAFLD through multiple mechanisms (Zhou et al., 2017; Zhai et al., 2019; Deng et al., 2020). On the other hand, our study found that the decrease of serum taurocholic acid content was related to *Erysipelotrichaceae_UCG-003*, and the decrease of fecal taurodeoxycholic acid, allolithocholic acid, and dehydrocholic acid content was related to *Subdoligranulum, Prevotellaceae_NK3B31_group,* and *Parvibacter*, respectively. The gut microbiota has a direct impact on bile acid composition and concentration and contributes to NAFLD progression (Ridlon et al., 2014). It was found that NAFLD patients with advanced fibrosis had elevated serum glycocholic acid and fecal deoxycholic acid concentrations, which were associated with increased abundances of *Bacteroidota* and *Lachnospiraceae*, compared with non-NAFLD controls (Adams et al., 2020). Increased secondary bile acid production in the NAFLD gut was associated with *Escherichia* and *Bilophila* (Jiao et al., 2018). *Bacteroides, Bifidobacterium, Clostridium, Lactobacillus,* and *Listeria* can convert bound bile acids to free bile acids *via* bile salt hydrolases, which are subsequently converted to secondary bile acids by *Clostridium* and *Eubacterium* under *Bacillota via* 7αdihydroxylation (Gérard, 2013). Furthermore, *Eggerthella* and *Ruminococcus* were also directly involved in bile acid metabolism (Jia et al., 2018). Thus, our findings suggested that the increase of underlying pathological *Fusobacteriota* and *Pseudomonadota* in MAFLD patients may contribute to the occurrence and development of the disease.

Gut microbiota-related metabolites, such as choline and tryptophan metabolites, SCFAs, bile acids, endogenous ethanol, and lipopolysaccharides, were involved in the pathogenesis of MAFLD (Vallianou et al., 2021). In this study, we performed an overall analysis of fecal and serum metabolites in MAFLD patients, and we identified more metabolites in feces. Although lipid molecules were the most abundant in both, the proportion of differential lipid metabolites in serum (49%) was higher than that in feces (21%), which further confirmed that lipid homeostasis in MAFLD patients was disrupted. At the same time, we also found some other metabolites that may be associated with the pathogenesis of MAFLD. We found that the following metabolites were simultaneously decreased in feces and serum of MAFLD patients: hypoxanthine, propionylcarnitine, tyrosylalanine, hesperetin, methionine, gamma-Glu-Leu, propylparaben, and neohesperidin. However, LPC 16:0, which belongs to glycerophosphocholine, increased in fecal and serum; LPC 18:0 decreased in feces and increased in serum. Studies have shown that the increased concentrations of hypoxanthine and uric acid in hepatocytes contribute to the accumulation of intracellular lipids, which in turn causes the occurrence of oxidative stress associated with the establishment of fatty liver-related diseases, laying the foundation for the development of fibrosis (Stirpe et al., 2002; Taylor et al., 2020). Accumulation of

hypoxanthine in the liver established a link between hyperuricemia and NAFLD (Toledo-Ibelles et al., 2021). Hesperetin is a citrus flavonoid found mainly in citrus fruits (oranges, grapefruits, and lemons) with various pharmacological properties, including anticancer, anti-Alzheimer's disease, and antidiabetic effects (Rekha et al., 2019). Hesperetin can alleviate hepatic steatosis, oxidative stress, inflammatory cell infiltration, and fibrosis in a high-fat diet (HFD)-induced rat model of NAFLD (Li et al., 2021a). Another flavonoid, neohesperidin, can reduce body weight, low-grade inflammation, and insulin resistance by altering the composition of the gut microbiota in mice fed a high-fat diet (Lu et al., 2020). Another study found that neohesperidin enhanced PGC-1α-mediated mitochondrial biosynthesis to alleviate hepatic steatosis in high-fat diet-fed mice (Wang et al., 2020). It was worth noting that our association analysis results suggested that LPC 18:0 was positively correlated with feces *Christensenellaceae_R-7_group*, *Oscillospiraceae_UCG-002*; Propylparaben was correlated with *Erysipelotrichaceae_UCG-003*; neohesperidin was also positively correlated with *Peptoniphilus*, *Phycicoccus*, and *Stomatobaculum* (Supplementary Table S2). However, the discovery and confirmation of the specific role relationship and related mechanisms require further follow-up research. For multi-omics data obtained through designed experiments, the ANOVA simultaneous component analysis (ASCA) and the group-wise ANOVA-simultaneous component analysis (GASCA) were considered to have certain advantages for analyzing the variations ascribable to the main experimental factors and their interactions (Saccenti et al., 2018; Raimondi et al., 2021).

Where we fall short is that due to the inherent worldwide variability in the composition of the gut microbiota (inter-individual and inter-population) it is unclear if our data apply to other areas of the world. Furthermore, previous studies have shown that diet was essential for gut microbial composition and function, and diet, gut microbiome, and metabolome were all interconnected (David et al., 2014; Tang et al., 2019). Although our study excluded patients with "abnormal" dietary habits (e.g., vegetarian food) within the past 12 months, we did not strictly require all participants to adjust their diets but retained their daily dietary habits. Therefore, while our study suggests differences in microbiota and metabolome due to the disease, the study cannot positively tell whether the findings were actually due to disease or diet. Further studies based on the patient's dietary structure are needed, which may help promote the development of individualized treatments. Finally, there was no significant difference in the FIB-4 index between the MAFLD group [1.25 (0.70–1.92)] and the control group [0.65 (0.33–0.85)], which may be due to the limitation of the sample size. Second, FIB-4 may not be sensitive enough to reflect differences between MAFLD patients and healthy controls when MAFLD patients are in an early stage of the disease. This suggests that we may be able to discover patients with MAFLD in more sensitive ways, such as gut microbiota and metabolites.

In conclusion, the human metabolome consists of interactions of host and microbiota-produced metabolites, and current functional metabolomics studies have focused on determining the role of individual metabolites or individual microbial taxa in MAFLD progression. Characterizing the complex interplay between the gut microbiota, its metabolites, and NAFLD progression remains a challenge. Our data provided a profile of alterations in gut microbes and metabolites in MAFLD patient systems, which may contribute to further studies of MAFLD disease mechanisms and the development of new diagnostic markers and therapeutics.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: NCBI—PRJNA851946.

Ethics statement

The studies involving human participants were reviewed and approved by the Institutional Review Board of West China Hospital, Sichuan University. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

LY was responsible for the study design, sample collection, data analysis, and manuscript writing. YD participated in the study design and sample collection. HH participated in the study design and manuscript revision. ZL, YZ, and SL participated in the sample collection and data analysis. GL and ZA revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This research was supported by the National Natural Science Foundation of China (NSFC; 32071462) and the Science and Technology Department of Sichuan Province, China (2021YFH0167).

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.969757/full#supplementary-material

References

Abellan-Schneyder, I., Matchado, M. S., Reitmeier, S., Sommer, A., Sewald, Z., Baumbach, J., et al. (2021). Primer, pipelines, parameters: issues in 16S rRNA gene sequencing. *mSphere* 6, 222–224. doi: 10.1128/mSphere.01202-20

Adams, L. A., Wang, Z., Liddle, C., Melton, P. E., Ariff, A., Chandraratna, H., et al. (2020). Bile acids associate with specific gut microbiota, low-level alcohol consumption and liver fibrosis in patients with non-alcoholic fatty liver disease. *Liver Int.* 40, 1356–1365. doi: 10.1111/liv.14453

Alseekh, S., Aharoni, A., Brotman, Y., Contrepois, K., D'Auria, J., Ewald, J., et al. (2021). Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices. *Nat. Methods* 18, 747–756. doi: 10.1038/s41592-021-01197-1

Alves-Bezerra, M., and Cohen, D. E. (2017). Triglyceride metabolism in the liver. *Compr. Physiol.* 8, 1–8. doi: 10.1002/cphy.c170012

Boursier, J., Mueller, O., Barret, M., Machado, M., Fizanne, L., Araujo-Perez, F., et al. (2016). The severity of nonalcoholic fatty liver disease is associated with gut dysbiosis and shift in the metabolic function of the gut microbiota. *Hepatology* 63, 764–775. doi: 10.1002/hep.28356

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

Chalasani, N., Younossi, Z., Lavine, J. E., Charlton, M., Cusi, K., Rinella, M., et al. (2018). The diagnosis and management of nonalcoholic fatty liver disease: practice guidance from the American Association for the Study of Liver Diseases. *Hepatology* 67, 328–357. doi: 10.1002/hep.29367

David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., et al. (2014). Diet rapidly and reproducibly alters the human gut microbiome. *Nature* 505, 559–563. doi: 10.1038/nature12820

den Besten, G., Lange, K., Havinga, R., van Dijk, T. H., Gerding, A., van Eunen, K., et al. (2013). Gut-derived short-chain fatty acids are vividly assimilated into host carbohydrates and lipids. *Am. J. Physiol. Gastrointest. Liver Physiol.* 305, G900–G910. doi: 10.1152/ajpgi.00265.2013

Deng, M., Qu, F., Chen, L., Liu, C., Zhang, M., Ren, F., et al. (2020). SCFAs alleviated steatosis and inflammation in mice with NASH induced by MCD. *J. Endocrinol.* 245, 425–437. doi: 10.1530/JOE-20-0018

Eckburg, P. B., Bik, E. M., Bernstein, C. N., Purdom, E., Dethlefsen, L., Sargent, M., et al. (2005). Diversity of the human intestinal microbial flora. *Science* 308, 1635–1638. doi: 10.1126/science.1110591

Eslam, M., Newsome, P. N., Sarin, S. K., Anstee, Q. M., Targher, G., Romero-Gomez, M., et al. (2020a). A new definition for metabolic dysfunction-associated fatty liver disease: an international expert consensus statement. *J. Hepatol.* 73, 202–209. doi: 10.1016/j.jhep.2020.03.039

Eslam, M., Sarin, S. K., Wong, V. W., Fan, J. G., Kawaguchi, T., Ahn, S. H., et al. (2020b). The Asian Pacific Association for the Study of the liver clinical practice guidelines for the diagnosis and management of metabolic associated fatty liver disease. *Hepatol. Int.* 14, 889–919. doi: 10.1007/s12072-020-10094-2

Fan, J. G., Jia, J. D., Li, Y. M., Wang, B. Y., Lu, L. G., Shi, J. P., et al. (2011). Guidelines for the diagnosis and management of nonalcoholic fatty liver disease: update 2010: (published in Chinese on Chinese journal of Hepatology 2010; 18:163-166). *J. Dig. Dis.* 12, 38–44. doi: 10.1111/j.1751-2980.2010.00476.x

Feng, W., Ao, H., and Peng, C. (2018). Gut microbiota, short-chain fatty acids, and herbal medicines. *Front. Pharmacol.* 9:1354. doi: 10.3389/fphar.2018.01354

Fouad, Y., Esmat, G., Elwakil, R., Zakaria, S., Yosry, A., Waked, I., et al. (2022). The egyptian clinical practice guidelines for the diagnosis and management of metabolic associated fatty liver disease. *Saudi J. Gastroenterol.* 28, 3–20. doi: 10.4103/sjg.sjg_357_21

Friedman, S. L., Neuschwander-Tetri, B. A., Rinella, M., and Sanyal, A. J. (2018). Mechanisms of NAFLD development and therapeutic strategies. *Nat. Med.* 24, 908–922. doi: 10.1038/s41591-018-0104-9

Gérard, P. (2013). Metabolism of cholesterol and bile acids by the gut microbiota. *Pathogens* 3, 14–24. doi: 10.3390/pathogens3010014

Jia, W., Xie, G., and Jia, W. (2018). Bile acid-microbiota crosstalk in gastrointestinal inflammation and carcinogenesis. *Nat. Rev. Gastroenterol. Hepatol.* 15, 111–128. doi: 10.1038/nrgastro.2017.119

Jiao, N., Baker, S. S., Chapa-Rodriguez, A., Liu, W., Nugent, C. A., Tsompana, M., et al. (2018). Suppressed hepatic bile acid signalling despite elevated production of primary and secondary bile acids in NAFLD. *Gut* 67, 1881–1891. doi: 10.1136/gutjnl-2017-314307

Leung, C., Rivera, L., Furness, J. B., and Angus, P. W. (2016). The role of the gut microbiota in NAFLD. *Nat. Rev. Gastroenterol. Hepatol.* 13, 412–425. doi: 10.1038/nrgastro.2016.85

Li, J., Wang, T., Liu, P., Yang, F., Wang, X., Zheng, W., et al. (2021a). Hesperetin ameliorates hepatic oxidative stress and inflammation via the PI3K/AKT-Nrf2-ARE pathway in oleic acid-induced HepG2 cells and a rat model of high-fat diet-induced NAFLD. *Food Funct.* 12, 3898–3918. doi: 10.1039/D0FO02736G

Li, F., Ye, J., Shao, C., and Zhong, B. (2021b). Compositional alterations of gut microbiota in nonalcoholic fatty liver disease patients: a systematic review and meta-analysis. *Lipids Health Dis.* 20:22. doi: 10.1186/s12944-021-01440-w

Lin, Y. C., Wu, C. C., and Ni, Y. H. (2020). New perspectives on genetic prediction for pediatric metabolic associated fatty liver disease. *Front. Pediatr.* 8:603654. doi: 10.3389/fped.2020.603654

Lin, H., Zhang, X., Li, G., Wong, G. L., and Wong, V. W. (2021). Epidemiology and clinical outcomes of metabolic (dysfunction)-associated fatty liver disease. *J. Clin. Transl. Hepatol.* 9, 972–982. doi: 10.14218/JCTH.2021.00201

Lu, J. F., Zhu, M. Q., Zhang, H., Liu, H., Xia, B., Wang, Y. L., et al. (2020). Neohesperidin attenuates obesity by altering the composition of the gut microbiota in high-fat diet-fed mice. *FASEB J.* 34, 12053–12071. doi: 10.1096/fj.201903102RR

Mardinoglu, A., Ural, D., Zeybel, M., Yuksel, H. H., Uhlén, M., and Borén, J. (2019). The potential use of metabolic cofactors in treatment of NAFLD. *Nutrients* 11:1578. doi: 10.3390/nu11071578

Mokhtari, Z., Gibson, D. L., and Hekmatdoost, A. (2017). Nonalcoholic fatty liver disease, the gut microbiome, and diet. *Adv. Nutr.* 8, 240–252. doi: 10.3945/an.116.013151

Morrison, D. J., and Preston, T. (2016). Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism. *Gut Microbes* 7, 189–200. doi: 10.1080/19490976.2015.1134082

Powell, E. E., Wong, V. W., and Rinella, M. (2021). Non-alcoholic fatty liver disease. *Lancet* 397, 2212–2224. doi: 10.1016/S0140-6736(20)32511-3

Raimondi, S., Calvini, R., Candeliere, F., Leonardi, A., Ulrici, A., Rossi, M., et al. (2021). Multivariate analysis in microbiome description: correlation of human gut protein degraders, metabolites, and predicted metabolic functions. *Front. Microbiol.* 12:723479. doi: 10.3389/fmicb.2021.723479

Rekha, S. S., Pradeepkiran, J. A., and Bhaskar, M. (2019). Bioflavonoid hesperidin possesses the anti-hyperglycemic and hypolipidemic property in STZ induced diabetic myocardial infarction (DMI) in male Wister rats. *J. Nutr. Intermed. Metabol.* 15, 58–64. doi: 10.1016/j.jnim.2018.12.004

Ridlon, J. M., Kang, D. J., Hylemon, P. B., and Bajaj, J. S. (2014). Bile acids and the gut microbiome. *Curr. Opin. Gastroenterol.* 30, 332–338. doi: 10.1097/MOG.0000000000000057

Saccenti, E., Smilde, A. K., and Camacho, J. (2018). Group-wise ANOVA simultaneous component analysis for designed omics experiments. Metabolomics: official journal of the Metabolomic. *Society* 14:73. doi: 10.1007/s11306-018-1369-1

Stirpe, F., Ravaioli, M., Battelli, M. G., Musiani, S., and Grazi, G. L. (2002). Xanthine oxidoreductase activity in human liver disease. *Am. J. Gastroenterol.* 97, 2079–2085. doi: 10.1111/j.1572-0241.2002.05925.x

Tang, Z. Z., Chen, G., Hong, Q., Huang, S., Smith, H. M., Shah, R. D., et al. (2019). Multi-Omic analysis of the microbiome and metabolome in healthy subjects reveals microbiome-dependent relationships between diet and metabolites. *Front. Genet.* 10:454. doi: 10.3389/fgene.2019.00454

Taylor, R. S., Taylor, R. J., Bayliss, S., Hagström, H., Nasr, P., Schattenberg, J. M., et al. (2020). Association Between fibrosis stage and outcomes of patients With nonalcoholic fatty liver disease: a systematic review and meta-analysis. *Gastroenterology* 158, 1611–25.e12. doi: 10.1053/j.gastro.2020.01.043

Toledo-Ibelles, P., Gutiérrez-Vidal, R., Calixto-Tlacomulco, S., Delgado-Coello, B., and Mas-Oliva, J. (2021). Hepatic accumulation of hypoxanthine: a link between Hyperuricemia and nonalcoholic fatty liver disease. *Arch. Med. Res.* 52, 692–702. doi: 10.1016/j.arcmed.2021.04.005

Vallianou, N., Christodoulatos, G. S., Karampela, I., Tsilingiris, D., Magkos, F., Stratigou, T., et al. (2021). Understanding the role of the gut microbiome and microbial metabolites in non-alcoholic fatty liver disease: current evidence and perspectives. *Biomol. Ther.* 12:56. doi: 10.3390/biom12010056

Wang, B., Jiang, X., Cao, M., Ge, J., Bao, Q., Tang, L., et al. (2016). Altered fecal microbiota correlates with liver biochemistry in nonobese patients with non-alcoholic fatty liver disease. *Sci. Rep.* 6:32002. doi: 10.1038/srep32002

Wang, S. W., Sheng, H., Bai, Y. F., Weng, Y. Y., Fan, X. Y., Lou, L. J., et al. (2020). Neohesperidin enhances PGC-1α-mediated mitochondrial biogenesis and alleviates hepatic steatosis in high fat diet fed mice. *Nutr. Diabetes* 10:27. doi: 10.1038/s41387-020-00130-3

Younossi, Z., Stepanova, M., Ong, J. P., Jacobson, I. M., Bugianesi, E., Duseja, A., et al. (2019). Nonalcoholic steatohepatitis is the fastest growing cause of hepatocellular carcinoma in liver transplant candidates. *Clin. Gastroenterol. Hepatol.* 17, 748–755.e3. doi: 10.1016/j.cgh.2018.05.057

Yuan, J., Chen, C., Cui, J., Lu, J., Yan, C., Wei, X., et al. (2019). Fatty liver disease caused by high-alcohol-producing *Klebsiella pneumoniae*. *Cell Metab.* 30, 675–88.e7. doi: 10.1016/j.cmet.2019.08.018

Zhai, S., Qin, S., Li, L., Zhu, L., Zou, Z., and Wang, L. (2019). Dietary butyrate suppresses inflammation through modulating gut microbiota in high-fat diet-fed mice. *FEMS Microbiol. Lett.* 366:fnz153. doi: 10.1093/femsle/fnz153

Zhou, D., Pan, Q., Xin, F. Z., Zhang, R. N., He, C. X., Chen, G. Y., et al. (2017). Sodium butyrate attenuates high-fat diet-induced steatohepatitis in mice by improving gut microbiota and gastrointestinal barrier. *World J. Gastroenterol.* 23, 60–75. doi: 10.3748/wjg.v23.i1.60

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 23 September 2022
DOI 10.3389/fmicb.2022.995323

OPEN ACCESS

EDITED BY
Qi Zhao,
University of Science and Technology Liaoning, China

REVIEWED BY
Cangzhi Jia,
Dalian Maritime University, China
Bing Wang,
Anhui University of Technology, China

*CORRESPONDENCE
Min Chen
chenmin@hnit.edu.cn
Zejun Li
lzjfox@hnit.edu.cn

SPECIALTY SECTION
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

RECEIVED 15 July 2022
ACCEPTED 22 August 2022
PUBLISHED 23 September 2022

CITATION
Peng L, Wang C, Tian G, Liu G, Li G, Lu Y, Yang J, Chen M and Li Z (2022) Analysis of CT scan images for COVID-19 pneumonia based on a deep ensemble framework with DenseNet, Swin transformer, and RegNet. *Front. Microbiol.* 13:995323. doi: 10.3389/fmicb.2022.995323

Analysis of CT scan images for COVID-19 pneumonia based on a deep ensemble framework with DenseNet, Swin transformer, and RegNet

Lihong Peng[1,2], Chang Wang[1], Geng Tian[3], Guangyi Liu[1], Gan Li[1], Yuankang Lu[1], Jialiang Yang[3], Min Chen[4]* and Zejun Li[4]*

[1]School of Computer Science, Hunan University of Technology, Zhuzhou, China, [2]College of Life Sciences and Chemistry, Hunan University of Technology, Zhuzhou, China, [3]Geneis (Beijing) Co., Ltd., Beijing, China, [4]School of Computer Science, Hunan Institute of Technology, Hengyang, China

COVID-19 has caused enormous challenges to global economy and public health. The identification of patients with the COVID-19 infection by CT scan images helps prevent its pandemic. Manual screening COVID-19-related CT images spends a lot of time and resources. Artificial intelligence techniques including deep learning can effectively aid doctors and medical workers to screen the COVID-19 patients. In this study, we developed an ensemble deep learning framework, DeepDSR, by combining DenseNet, Swin transformer, and RegNet for COVID-19 image identification. First, we integrate three available COVID-19-related CT image datasets to one larger dataset. Second, we pretrain weights of DenseNet, Swin Transformer, and RegNet on the ImageNet dataset based on transformer learning. Third, we continue to train DenseNet, Swin Transformer, and RegNet on the integrated larger image dataset. Finally, the classification results are obtained by integrating results from the above three models and the soft voting approach. The proposed DeepDSR model is compared to three state-of-the-art deep learning models (EfficientNetV2, ResNet, and Vision transformer) and three individual models (DenseNet, Swin transformer, and RegNet) for binary classification and three-classification problems. The results show that DeepDSR computes the best precision of 0.9833, recall of 0.9895, accuracy of 0.9894, F1-score of 0.9864, AUC of 0.9991 and AUPR of 0.9986 under binary classification problem, and significantly outperforms other methods. Furthermore, DeepDSR obtains the best precision of 0.9740, recall of 0.9653, accuracy of 0.9737, and F1-score of 0.9695 under three-classification problem, further suggesting its powerful image identification ability. We anticipate that the proposed DeepDSR framework contributes to the diagnosis of COVID-19.

KEYWORDS

COVID-19 pneumonia, CT scan image, deep ensemble, DenseNet, Swin transformer, RegNet

Introduction

In December 2019, a novel acute atypical respiratory disease, COVID-19, has broken in Wuhan, China (Ksiazek et al., 2003; Zhou et al., 2020). COVID-19 was defined as a global pandemic by the World Health Organization on 3 November 2020. Till 26 June 2022, this disease has infected over 541 million individuals and caused over 6.3 million deaths (COVID Live—Coronavirus Statistics—Worldometer, 2022). COVID-19 has exacerbated human suffering, damaged the global economy, and seriously affected the health, environmental and social fields worldwide (Mofijur et al., 2021). It has still indirectly affected the global educational and religions level. Moreover, it has caused healthcare service resources to the brink in many countries and regions and will deeply affect medical research (Harper et al., 2020). Furthermore, middle-income countries especially low-income countries remain more vulnerable in preventing COVID-19 and need to face more serious challenges (Peters et al., 2020).

The COVID-19 pandemic has caused severe challenges to global public health (Wang et al., 2020; Sun et al., 2022a). The screening of massive samples each day overwhelms laboratories worldwide (Agaoglu et al., 2022). Detection of SARS-CoV-2 through RT-PCR from a nasopharyngeal swab sample is the most common avenue to diagnose COVID-19. However, RT-PCR does not demonstrate powerful sensitivity and specificity (Pu et al., 2022). Moreover, it need spend about 6 h for sampling and consecutive tests to distinguish false positives and false negatives (Lee et al., 2022). Multiple patients demonstrate clinical, laboratorial, and radiological features related to COVID-19, however, their RT-PCR test results are negative (Saad Menezes et al., 2022).

Many evidences have suggested that chest Computer Tomography (CT) is an accurate and efficient COVID-19 diagnosis avenue (Chung et al., 2020; Pan et al., 2020; Wang C C et al., 2021; Wang B et al., 2021). It has high sensitivity and low misdiagnosis rate, thus is an efficient complement to RT-PCR (Fields et al., 2021). Although it is vital to rapidly detect patients with the COVID-19 infection by CT images, expert thoracic radiologists are not likely to immediately diagnose positive cases at all times, which may not only cause treatment delay, but also urge further transmission of COVID-19 because the COVID-19 patients are not promptly isolated (Jin et al., 2020; Shorten et al., 2021; Afshar et al., 2022). In this situation, it is especially important to aid doctors and health care workers to distinguish COVID-19-related CT images from non-COVID-19-realted CT images using artificial intelligence techniques.

Many studies have suggested that artificial intelligence (AI) techniques including machine learning obtained enormous success in bioinformatics and medical image analysis (Chen et al., 2018a,b, 2019; Wang B et al., 2021; Wang C C et al., 2021; Zhang et al., 2021; Yang et al., 2022; Liu et al., 2022a). Over the last decade years, deep learning techniques have outperformed numerous state-of-the-art machine learning algorithms and demonstrated excellent learning ability in many fields including image recognition (Voulodimos et al., 2018; Wang B et al., 2021; Wang C C et al., 2021;Sun et al., 2022; Liu et al., 2022a,b).

Under the situation of no standardization, artificial intelligence technologies, especially deep learning, have been widely applied to data collection and performance evaluation for COVID-19 (Roberts et al., 2021). Abbas et al. (2021) proposed a novel convolutional neural network (CNN) model, DeTraC, to classify COVID-19-related chest X-ray images based on feature extraction, decomposition and class composition. Shalbaf and Vafaeezadeh (2021) designed a deep transfer learning-based ensemble model with different pre-trained CNN architectures to detect CT images for novel COVID-19 diagnosis. Zhang et al. (2020) developed a deep learning-based anomaly detection system to screen COVID-19 from chest x-ray images. Zhou et al. (2021) explored an ensemble deep learning framework to detect COVID-19 from CT images. Karbhari et al. (2021) introduced an auxiliary classifier generative adversarial network to generate synthetic chest X-ray images and further detect COVID-19 based on custom-made deep learning model. Chouat et al. (2022) exploited deep transfer learning algorithm to screen COVID-19 positive patients based on CT scan and chest X ray images. Fan et al. (2022) proposed a branch network model by combining CNN and transformer structure for the identification of COVID-19 using CT scan images. Ter-Sarkisov (2022) built a COVID-CT-Mask-Net model to diagnose COVID-19 through regional features from chest CT scan images. Chieregato et al. (2022) presented a deep learning-based COVID-19 prognostic hybrid model to support clinical decision making.

These models are mainly based on CNN and attention mechanism and effectively classify COVID-19-related images and non-COVID-19-related ones. However, they remain to improve the classification performance. In this study, we developed an ensemble deep learning framework (DeepDSR) by integrating three state-of-the-art neural networks including DenseNet, Swin transformer, and RegNet for the COVID-19 diagnosis.

Materials and methods

Materials

We use three available CT image datasets related to COVID-19 to investigate the performance of our proposed DeepDSR model. Dataset 1 can be downloaded from https://www.kaggle.com/datasets/plameneduardo/a-covid-multiclass-dataset-of-ct-scans. It contains publicly available 4,173 CT scan images from 210 different patients, out of which 2,168 images are from 80 patients infected by COVID-19 and confirmed by RT-PCR in hospitals from Sao Paulo, Brazil (Soares et al., 2020). Dataset 2 can be downloaded from https://www.kaggle.com/datasets/plameneduardo/sarscov2-ctscan-dataset. It contains 1,252 CT scan images from patients infected by COVID-19 and 1,230 CT scan images for patients non-infected by COVID-19 in hospitals

from Sao Paulo, Brazil (Soares et al., 2020). Dataset 3 can be downloaded from https://github.com/UCSD-AI4H/COVID-CT. It contains 349 COVID-19 CT images from 216 patients and 463 non-COVID-19 CT images (Zhao et al., 2020).

To boost the generalization ability of our proposed DeepDSR model, we integrate the above three datasets to one larger dataset. Consequently, DeepDSR can be used to effectively classify CT images in both individual datasets and other datasets. And we remove images with poor imaging and ones nonconforming to specifications. Finally, we obtain one dataset with 7,398 pulmonary CT images, which include 3,768 CT images from patients with the COVID-19 infections, 1,247 ones with other pneumonia infections, and 2,383 ones from normal lungs. We use 3,768 COVID-19-related images and 2,383 normal CT images to train the models for binary classification problems and use all 7,398 images for three classification problems. As shown in Figure 1, Lines 1–3 show pulmonary CT images from patients with COVID-19 infections, normal lungs, and patients with other pneumonia infections, respectively.

The pipeline of DeepDSR

It is difficult to obtain the best prediction accuracy when only thousands of images are trained. Thus, we design an ensemble model to reduce the limitation of lack of data through transfer learning. The ensemble model integrates three state-of-the-art and different network architectures, that is, DenseNet, Swin transformer and RegNet. The pipeline is shown as Figure 2. As shown in Figure 2, first, we preprocess data by integrating three available COVID-19-related CT image datasets to one larger dataset. Second, we pretrain weights of DenseNet, Swin transformer, and RegNet on the ImageNet dataset based on transformer learning. Third, we continue to train DenseNet, Swin Transformer, and RegNet on the integrated larger dataset. Finally, the classification results are obtained by integrating results from the above three models and the soft voting approach.

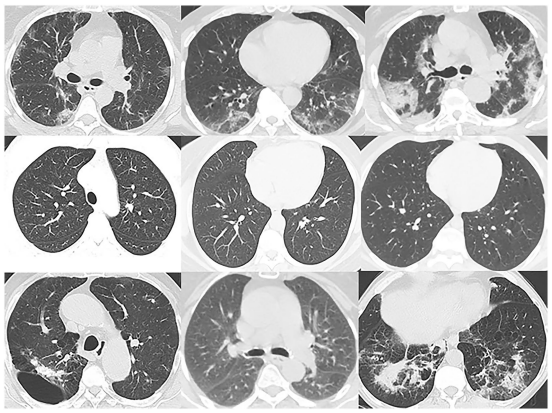

FIGURE 1
Image examples in dataset.

DenseNet

CNNs can implement accurate and efficient train when they contain shorter connections between layers close to the input and those close to the output. Traditional convolutional networks composed of L layers connect each layer to its subsequent layer. Inspired by the model proposed by Huang et al. (2017), we introduced a Dense convolutional Network, DenseNet, to classify COVID-19-related CT scan images. DenseNet implements connection between each layer in a feed-forward fashion to accurately and efficiently train the model. DenseNet with L layers has $\frac{1}{2}L(L+1)$ direct connections. At each layer, as shown in Figure 3A, the CT image feature maps from all previous layers are taken as its inputs and its outputs are taken as the inputs at next layer. For ResNet (Radosavovic et al., 2020), the original features and the new features are added by element by element to achieve the sample features. Differed from ResNet, DenseNet obtains shortcut through direct concatenation. DenseNet reduces the vanishing-gradient problem, boosts feature propagation, advances feature reuse while greatly decrease the number of parameters.

Swin transformer

Transformer has difficulty in application from language to vision because of differences between the two areas. Thus, Liu et al. developed a hierarchical transformer to obtain data representation by shifted windows (Liu et al., 2021). For an image, first, transformer splits it into fixed-size patches. Second, the patches are linearly embedded and added position embeddings. Third, the embedded results are feed to a standard Transformer encoder. Finally, an extra learnable "classification token" is added to the sequence to classify images. Inspired by model proposed by Liu et al. (2021), we use the window-shift technique and design a Swin transformer to classify COVID-19-related CT scan images.

The window-shift technique and the sliding window approach are similar in modeling ability, but the former is beneficial for all-MLP architectures and has much lower latency than the latter. Swin transformer focuses on shifting window partition between consecutive self-attention layers. As shown in Figure 3B, the shifted windows connect with the windows in the previous layer, thus significantly enhancing the modeling ability. The window-shift technique limits self-attention computation to non-overlapping local windows as well as supports cross-window connection, thereby effectively improving the image classification ability of models. Furthermore, Swin transformer utilizes the window-shift technique and demonstrates the flexibility when

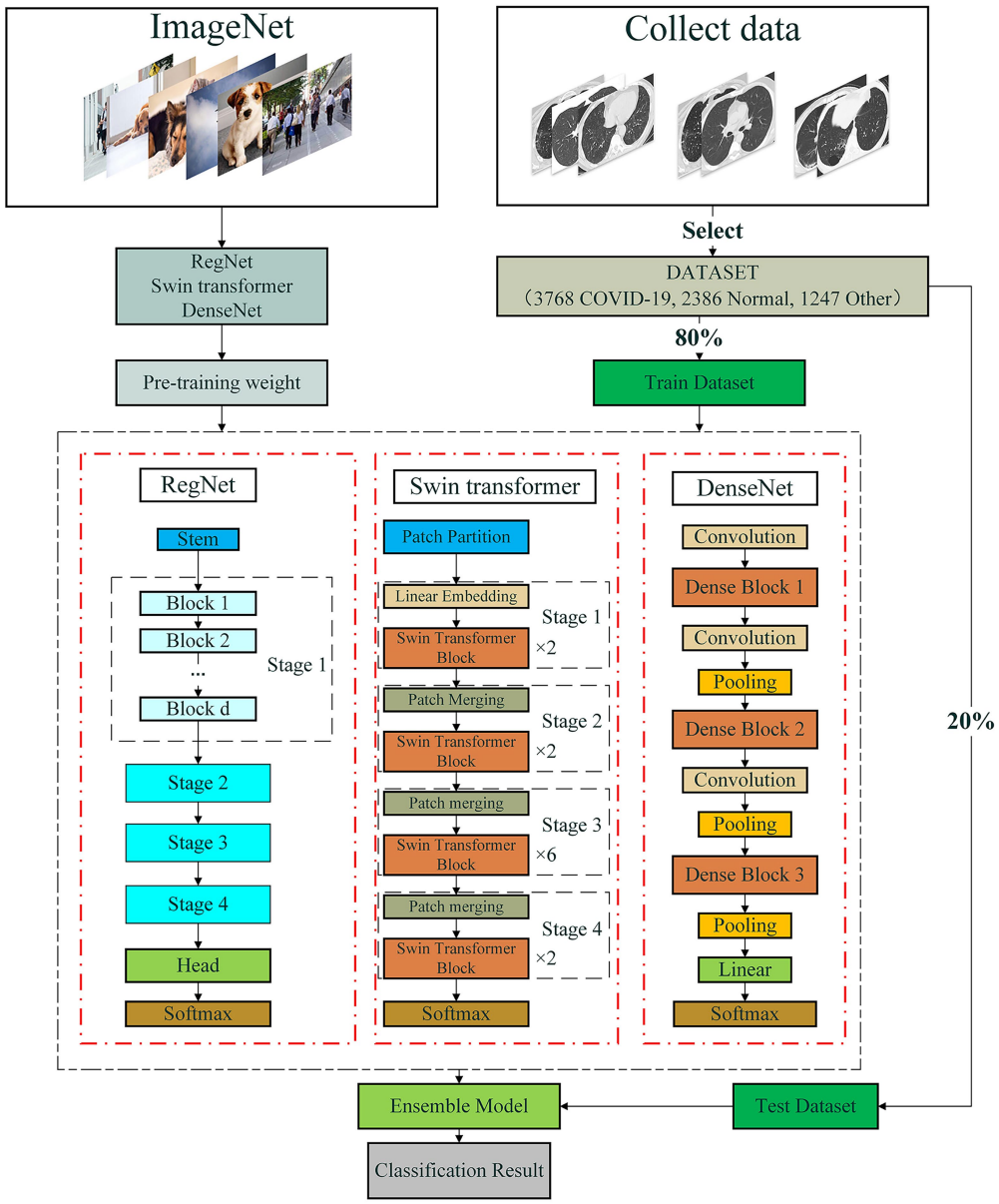

FIGURE 2
The pipeline for COVID-19-related CT image classification based on an ensemble of DenseNet, RegNet, and Swin transformer.

modeling on COVID-19-related image identification as well as computational complexity linearly with image size.

RegNet

Neural architecture search and RegNet are two representative neural network design paradigms. The two complementary design paradigms can improve the efficiency of search algorithms while develop better models. Neural architecture search mainly focuses on the search strategy to more efficiently find the best network instances in a fixed and manually designed search space. In contrast, RegNet (Radosavovic et al., 2020) more focuses on designing paradigms on novel design spaces.

RegNet is a novel neural network design paradigm. It used a residual network to simplify the deeper network training. It can boost the understanding of network design and further investigate design principles with strong generalize abilities across different settings. Instead of concentrating on individual network instance design, RegNet designs network design spaces that can parameterize network populations. The design process is similar to manually design network while advances the design space level. Consequently, we can obtain a low-dimensional design space composed of multiple simple and regular networks.

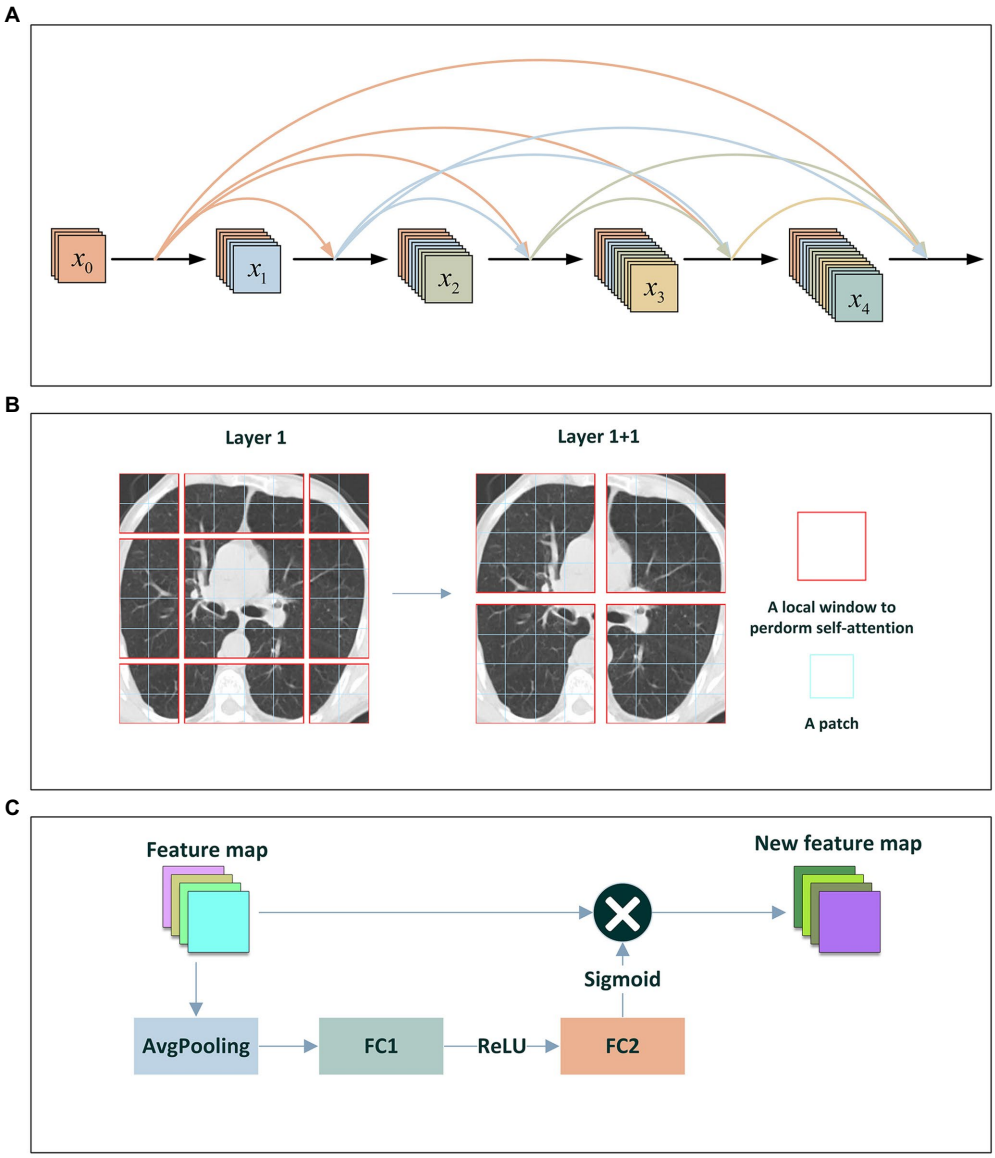

FIGURE 3
(A) The DenseNet Block; **(B)** Shifted-Window technique; **(C)** The Squeeze-and-Excitation network.

In this study, RegNet composes a stem with the stride of 2 and 32 3×3 convolution kernels, followed by a network body composed of a series of stages, and finally a head. In the network body, each stage operates at gradually reduced resolution. It consists of multiple identical blocks with the stride of 1 except that the first block uses stride-two convolution kernel. The head is composed of an average pooling layer and a fully connected layer. It is used to output n classes.

In addition, RegNet contains RegNetX and RegNetY composed of RegNetX and squeeze-and-excitation network. As shown in Figure 3C, the squeeze-and-excitation network generally composed of one global average pooling layer and two fully connection layers that separately use ReLU and sigmoid as activation functions.

Ensemble

Although machine learning techniques achieve significant successes in knowledge discovery, they fail to obtain powerful performance because of imbalanced, high-dimensional and noisy features of data. Consequently, ensemble learning, which effectively integrates the prediction results from multiple individual classifiers, has been widely applied to image processing (Sagi and Rokach, 2018).

Ensemble learning methods first generate multiple weak predictive results using different machine learning models, and obtain better performance by ensemble of the results from each individual model with different voting strategies. It composes of five main types: bagging, AdaBoost, gradient

boosting, random forest, and random sub-space (Dong et al., 2020). Bagging generates sample subsets based on the random sampling approach, and train basic learners in a parallel manner (Breiman, 1996). AdaBoost concentrates on improving classification ability of individual models *via* iteratively adjusting weights for all misclassified samples (Hastie et al., 2009). Gradient boosting achieves sample subsets based on the random sampling approach, and trains each classifier to alleviate the residuals caused by the previous model. Thus, gradient boosting better fits the real data (Friedman, 2002). Random forest takes decision trees as predictors and separately trains multiple models to reduce the overfitting problem (Breiman, 2001). Random subspace constructs a set of feature subspaces based on the random sampling approach, and trains learners on the feature subspace set. Finally, it obtains the final classification by combining the results from each individual classifier (Ho, 1995).

Ensemble learning utilizes different ensemble strategies to ensemble results from individual models. For regression estimation, it gains the final results *via* averaging all predictions. For classification, ensemble learning uses the voting method to achieve the final classification by combining each individual classifier. The absolute majority voting approach takes the same classification result as one from more than half of individual classifiers as the final result, and the relative majority voting approach takes the classification result where the number of individual predictors involved in a certain prediction is the largest as the final result. Therefore, we combine DenseNet, Swin transformer, and RegNet and develop an ensemble deep learning model, DeepDSR, to improve the COVID-19 classification performance of the model.

The classification scores from the three individual classifiers are integrated based on the soft voting approach. Given a query image, for a binary classification problem, suppose that its scores classified to COVID-19-related image by DenseNet, Swin transformer, and RegNet are $\boldsymbol{S}_1^{covid-19}$, $\boldsymbol{S}_2^{covid-19}$, and $\boldsymbol{S}_3^{covid-19}$, respectively, its final score $\boldsymbol{S}_{final}^{covid-19}$ classified to COVID-19-related sample can be represented by Eq. (1):

$$\boldsymbol{S}_{final}^{covid-19} = \boldsymbol{S}_1^{covid-19} + \boldsymbol{S}_2^{covid-19} + \boldsymbol{S}_3^{covid-19} \tag{1}$$

Similarly, its final score $\boldsymbol{S}_{final}^{non-covid-19}$ classified to non-COVID-19-related image can be represented by Eq. (2):

$$\boldsymbol{S}_{final}^{non-covid-19} = \boldsymbol{S}_1^{non-covid-19} + \boldsymbol{S}_2^{non-covid-19} + \boldsymbol{S}_3^{non-covid-19} \tag{2}$$

The image will be taken as COVID-19 related when $\boldsymbol{S}_{final}^{covid-19} > \boldsymbol{S}_{final}^{non-covid-19}$, it will be taken as non-COVID-19 related, otherwise.

Furthermore, for a three-classification problem, suppose that its scores classified to COVID-19 related by DenseNet, Swin transformer, and RegNet are $\boldsymbol{S}_1^{covid-19}$, $\boldsymbol{S}_2^{covid-19}$, and $\boldsymbol{S}_3^{covid-19}$, respectively, its final score $\boldsymbol{S}_{final}^{covid-19}$ classified to positive sample can be computed by Eq. (3):

$$\boldsymbol{S}_{final}^{covid-19} = \boldsymbol{S}_1^{covid-19} + \boldsymbol{S}_2^{covid-19} + \boldsymbol{S}_3^{pcovid-19} \tag{3}$$

Similarly, its final score $\boldsymbol{S}_{final}^{other}$ classified to other pneumonia can be computed by Eq. (4):

$$\boldsymbol{S}_{final}^{other} = \boldsymbol{S}_1^{other} + \boldsymbol{S}_2^{other} + \boldsymbol{S}_3^{other} \tag{4}$$

And its final score $\boldsymbol{S}_{final}^{normal}$ from normal lung can be computed by Eq. (5):

$$\boldsymbol{S}_{final}^{normal} = \boldsymbol{S}_1^{normal} + \boldsymbol{S}_2^{normal} + \boldsymbol{S}_3^{normal} \tag{5}$$

Finally, the image will be taken as COVID-19 related when $\boldsymbol{S}_{final}^{covid-19}$ is larger than $\boldsymbol{S}_{final}^{other}$ and $\boldsymbol{S}_{final}^{normal}$; it will be taken as other pneumonia when $\boldsymbol{S}_{final}^{other}$ is larger than other two values; it is from normal lung otherwise.

Transfer learning and pre-training

CNNs usually need to train a mass of parameters. However, it is almost impossible to learn such massive parameters only through a few training images (Zhuang et al., 2020; Zhu et al., 2021). In particular, transfer learning can utilize existing knowledge and transfer knowledge from source domains to the target domain and thus has been widely applied to solve problems in different while relevant fields (Pan and Yang, 2009; Weiss et al., 2016). It usually pretrains weights on a large-scale dataset using a standard neural architecture and then fine-tunes the weights on a target dataset. It has been successfully applied to medical image classification, for instance, cancer classification, pneumonia detection, and skin lesion identification (Chang et al., 2017; Deepak and Ameer, 2019; Khalifa et al., 2019; Chouhan et al., 2020).

Furthermore, existing lung CT scan images do not satisfy the need of a powerful image identification model because most of lung CT images are not publicly available. In addition, a image processed by random affine transformation, random crop or flip may not be a complete lung CT image because of the specificity of CT scanning techniques. The above two situations may easily produce the overfitting problem of image classification models in small datasets. Therefore, we want to pretrain the proposed

TABLE 1 The confusion matrix.

		True results	
		Positive	**Negative**
Predicted results	Positive	TP	FP
	Negative	FN	TN

TABLE 2 Parameter settings.

Model	Parameter setting
Swin transformer	epochs = 100, batch_size = 8, lr = 0.0001
RegNet	epochs = 100, batch_size = 16, lr = 0.001, lrf = 0.01
DenseNet	epochs = 100, batch_size = 16, lr = 0.001, lrf = 0.01

DeepDSR model by transfer learning to advance the training speed, reduce overfitting, alleviate problems produced by insufficient data, and further improve the classification performance (Hijab et al., 2019; Cherti and Jitsev, 2021; Mustafa et al., 2021).

Finally, we developed an ensemble deep model (DeepDSR) to analyze COVID-19 CT images by combining DenseNet, Swin transformer, and RegNet. First, we integrate three COVID-19 image dataset to one larger dataset. Second, we pretrain weights of DenseNet, Swin Transformer, and RegNet on the ImageNet dataset. Third, we repeatedly select 80% of CT images from the integrated larger dataset as the training set and the remaining 20% as the test set. Fourth, the training set is used to train DenseNet, Swin transformer, and RegNet, respectively. The test set is used to test the performance of DenseNet, Swin transformer, and RegNet, respectively. Finally, the final classification results are obtained by integrating the results from the above three models.

Results

Experimental evaluation and parameter settings

To evaluate the performance of the proposed DeepDSR framework, we use six measurement metrics: precision, recall, accuracy, F1-score, AUC and AUPR. Suppose that True Positive (TP), True Negative (TN), False Negative (FN), and False Negative (FN) are defined as Table 1. We can compute precision, recall, accuracy, F1-score, True Positive Rate (TPR), and False Positive Rate (FGR) as follows:

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$F_1 - score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{9}$$

$$TPR = \frac{TP}{TP + FN} \tag{10}$$

$$FPR = \frac{FP}{TN + FP} \tag{11}$$

And AUC is the area under the TPR-FPR curve, and AUPR is the area under the precision-recall curve. For each sample (image), its classification scores from three individual networks (DenseNet, Swin transformer, and RegNet) are computed by the softmax layer, respectively. Its final classification probability is obtained by averaging the scores from the three single models. AUC and AUPR can be computed based on the obtained final classification probability.

Moreover, the six metrics are not equally important to COVID-19 CT image classification. The results caused by false negatives are more severe than ones caused by false positives for medical image classification. Therefore, recall and AUPR are more important compared to the other four evaluation metrics.

The experiments are performed for 100 epochs to obtain the optimal parameter settings. In addition, DenseNet and RegNet use stochastic gradient descent algorithm and Swim transformer uses AdamW as optimizer to update parameters. The detailed parameters are set in Table 2. In Table 2 and the following Tables 2–5, the bold font in each column represents the best performance computed by corresponding method.

The performance comparison of DeepDSR with other three models for COVID-19 image binary classification

We compare the proposed DeepDSR method to three state-of-the-art deep learning models (efficientNetV2, ResNet, and Vision transformer) when classifying CT scan images to two classes: COVID-19 related or non-COVID-19 related. EfficientNetV2 (Tan and Le, 2021) aims to solve the problem of slow training when the size of the training image is large in efficientNetV1. Moreover, it replaced some MBConv structures in shallow layers with Fused-MBConv structures and found the optimal combination through neural architecture search technology to improve the network training speed. Finally, efficientNetV2 used a non-uniform scaling strategy to scale the model and thus make the model more reasonable.

ResNet (He et al., 2016) aims to solve the vanishing gradient and network degradation problems in traditional neural networks.

ResNet solved the vanishing gradient problem through data preprocessing and batch normalization layer, and reduced the network degradation problem through a residual structure. ResNet used a connection model of shortcut to add interlayers in the feature matrix and thus greatly improve the depth of the network.

Transformer (Vaswani et al., 2017) has been broadly used in the natural language processing field. Attention mechanism has been widely used in the computer vision field. Inspired by the transformer mechanism, Vaswani et al. divided each image into patches, and took the linear embedded sequence of these image blocks as the input of the transformer. The processing method of image patches is the same as marks in NLP applications. Vision transformer (Dosovitskiy et al., 2020) achieved excellent results when both pretraining on a sufficient scale dataset and migrating to tasks with fewer data points.

We first selected 80% images as training set and 20% as test set from the integrated COVID-19-related CT scan images. We then train DeepDSR, efficientNetV2 (Tan and Le, 2021), ResNet (He et al., 2016), and Vision transformer (Dosovitskiy et al., 2020) for 100 epochs, respectively. The results are shown in Table 3 and Figure 4A. We can find that DeepDSR significantly outperforms efficientNetV2 in terms of precision, recall, accuracy, F1-score, AUC and AUPR. For examples, DeepDSR outperforms 21.93% and 33.42% compared to efficientNetV2 based on AUC and AUPR, respectively. DeepDSR also performs better than ResNet and Vision transformer although the improvement is slight. Figures 4B,C illustrate the AUC and AUPR values of DeepDSR and other models when classifying COVID-19-related CT images to two classes. The above results show that DeepDSR can efficiently identify CT scan images for patients infected by COVID-19.

The performance comparison of DeepDSR and three individual models for COVID-19 image binary classification

To investigate the image classification performance of the proposed DeepDSR model with DenseNet, Swin transformer, and RegNet, we conduct experiment for 100 epochs. At each epoch, we select 80% samples to train DeepDSR, DenseNet, Swin transformer, and RegNet and the remaining 20% to test their performance. Table 4 and Figure 5A demonstrate the prediction results of the above four models. The results show that the proposed ensemble model, DeepDSR, outperforms other three individual models in terms of precision, recall, accuracy, F1-score, AUC, and AUPR. Figures 5B,C illustrate the AUC and AUPR values obtained from the above four models. We find that DeepDSR, ensemble of DenseNet, Swin transformer, and RegNet, can more effectively classify CT images to two classes: COVID-19-related or not.

Statics of true positives/negatives and false positives/negatives

We investigate the classification results on 1,231 COVID-19-related CT images from the test set to more intuitively illustrate the affect of DeepDSR on CT image identification performance. Table 5 and Figure 6 give the number of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) computed by DeepDSR, DenseNet, Swin transformer, and RegNet, respectively.

The results show that DeepDSR, DenseNet, Swin transformer, and RegNet misclassify a few samples. DeepDSR computes the most TPs and TNs while the least FPs and FNs. Furthermore, efficientNetV2, ResNet, and Vision transformer compute much more FPs and FNs compared with DeepDSR, demonstrating higher error rates. Moreover, DeepDSR, ensemble of DenseNet, Swin transformer, and RegNet, outperforms all other three individual models. Thus, the neural network, combining the predictions obtained from all the base models, can significantly improve the CT image classification performance of models. In addition, the stacking ensemble consisting of all three base models outperforms all other combinations. DeepDSR is tuned to utilize those predictions that help improve the classification performance and ignore the wrong predictions made by the base models.

TABLE 4 The performance comparison of DeepDSR and three individual models for binary classification problem.

	Precision	Recall	Accuracy	F1-score	AUC	AUPR
Swin transformer	0.9619	0.9539	0.9675	0.9579	0.9943	0.9924
RegNet	0.9571	0.9832	0.9764	0.9700	0.9963	0.9949
DenseNet	0.9770	0.9790	0.9829	0.9780	0.9981	0.9973
DeepDSR	**0.9833**	**0.9895**	**0.9894**	**0.9864**	**0.9991**	**0.9986**

The bold fonts represent the best performance in each column.

TABLE 3 The performance comparison of DeepDSR and other models for COVID-19 image binary classification.

	Precision	Recall	Accuracy	F1-score	AUC	AUPR
EfficientNetV2	0.5077	0.9015	0.6231	0.6495	0.7800	0.6649
ResNet	0.9786	0.9602	0.9764	0.9693	0.9960	0.9943
Vision transformer	0.9811	0.9769	0.9838	0.9790	0.9982	0.9975
DeepDSR	**0.9833**	**0.9895**	**0.9894**	**0.9864**	**0.9991**	**0.9986**

The bold fonts represent the best performance in each column.

The affect of transfer learning on the performance

In the above sections, we pretrain the weights of DenseNet, Swin transformer, and RegNet on the ImageNet dataset and continue to train the three models on the integrated larger dataset for 100 epochs. We set up a group of control experiments without pretraining (100 epochs and 200 epochs) to validate the importance of pretraining weights of the models for 100 epochs. The results are shown in Table 6 and Figure 7.

From Table 6 and Figure 7, we can observe that the performance of network architecture with the pretrained weights is much better than that of the network without pretraining weights for 100 epochs and 200 epochs. For example, under 100 epochs, the pretrained network computes accuracy of 0.9894, AUC of 0.9991, and AUPR of 0.9986, outperforming 7.88%, 2.83%, and 5.61% than the network without pretraining, respectively. In addition, we also investigate the performance of DeepDSR with pretraining for 100 epochs and ones without pretraining for 200 epochs. The results show that the pretrained network significantly outperforms the network without pretraining even for 200 epochs. Accuracy, AUC, and AUPR

TABLE 5 Statistical analyses of four models on 1,231 images.

	DenseNet	Swin transformer	RegNet	DeepDSR
TN	743	736	733	**746**
FN	10	22	8	**5**
FP	11	18	21	**8**
TP	467	455	469	**472**

The bold fonts represent the best performance in each column.

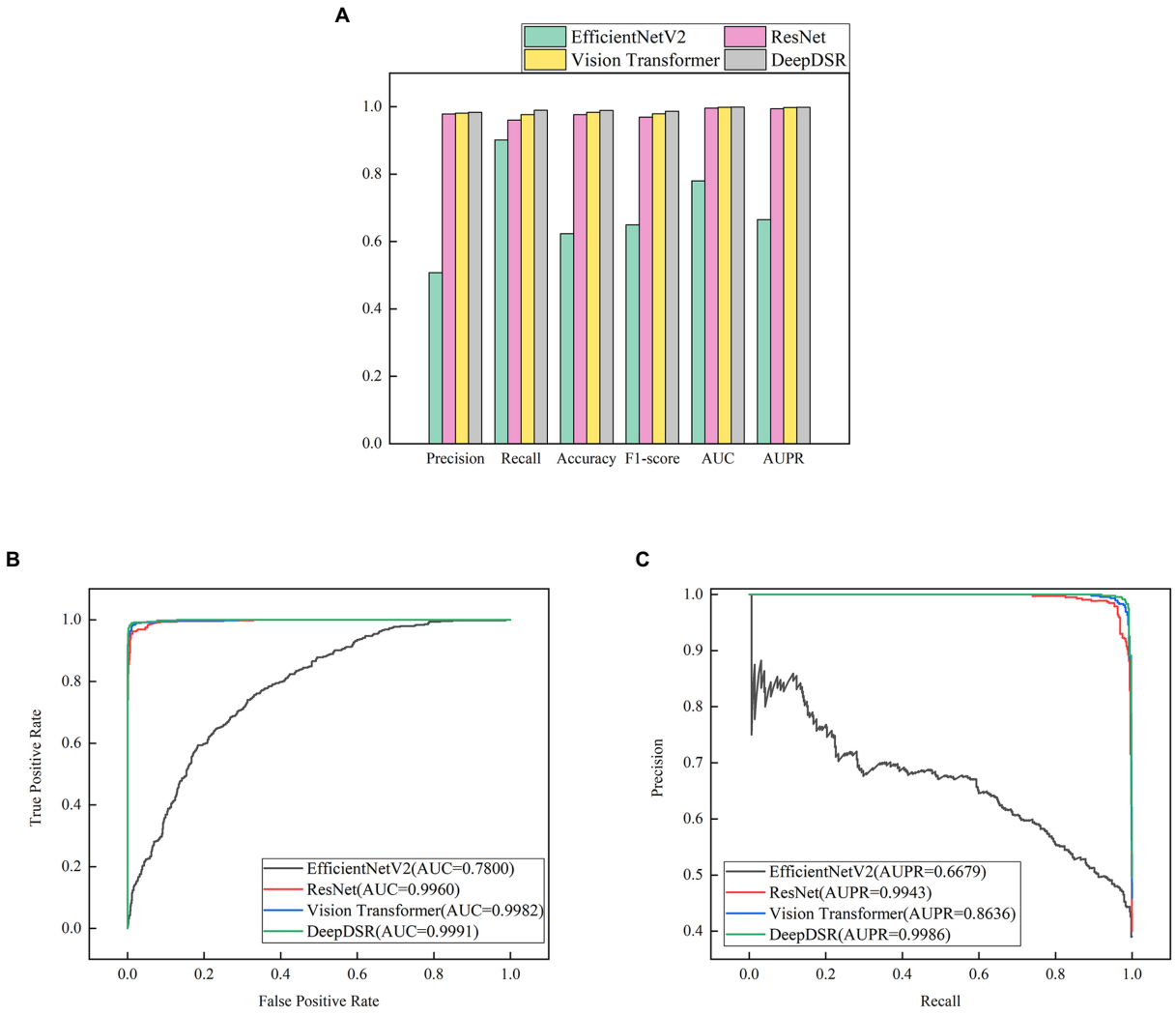

FIGURE 4
(A) The performance comparison of DeepDSR and other models for COVID-19 image binary classification. **(B,C)** The AUC and AUPR values of DeepDSR and other models for COVID-19 image binary classification.

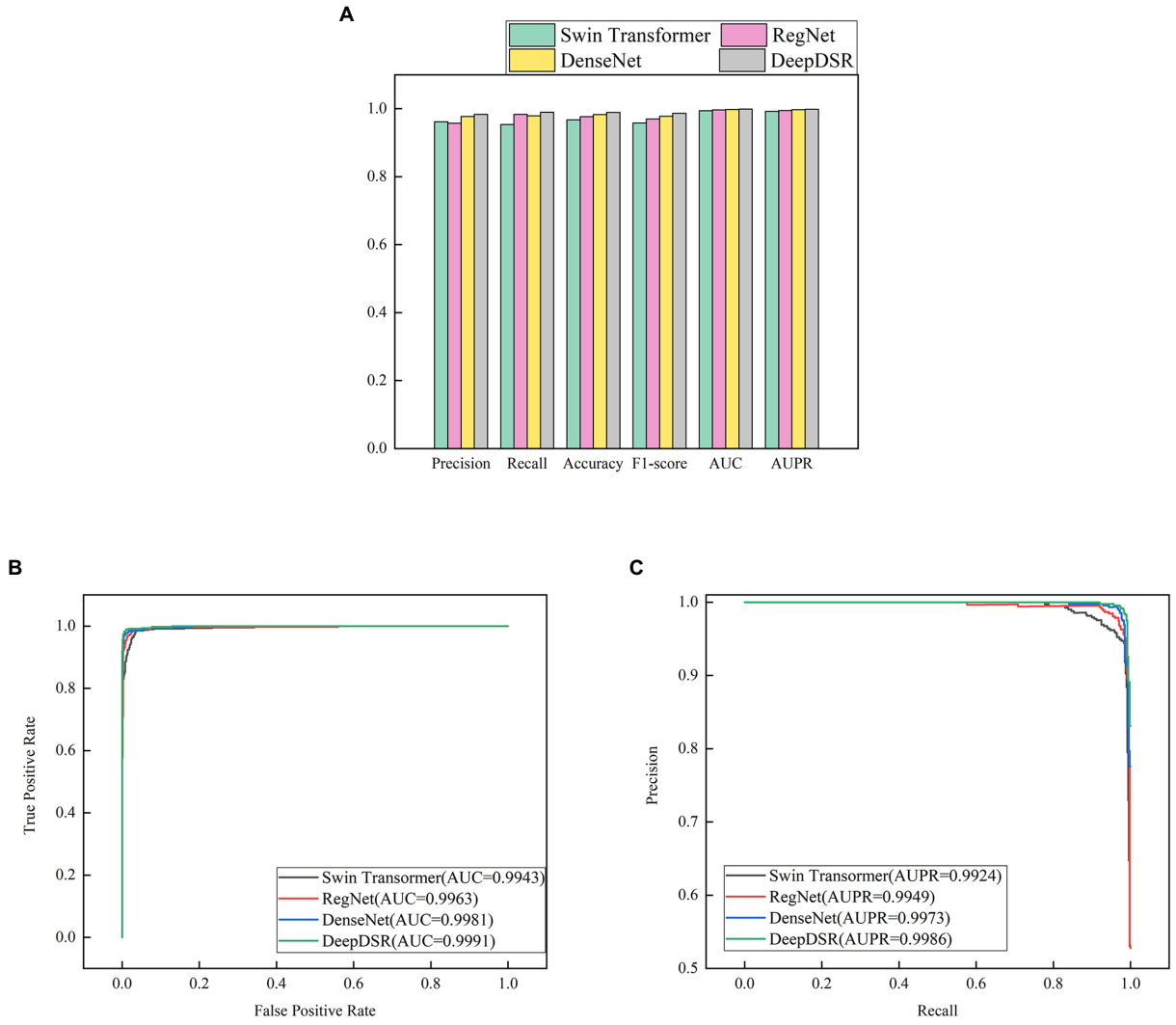

FIGURE 5
(A) The performance comparison of DeepDSR and three individual models for COVID-19 binary classification problem; **(B,C)** The AUC and AUPR values of DeepDSR and three individual models for COVID-19 binary classification problem.

computed by the pretrained network are better 3.83%, 1.27%, and 1.68% than ones without pretraining for 200 epochs, respectively. The results demonstrate that pretraining based on transfer learning can reduce the training time while improve the classification performance. Finally, when adding epochs on the pretrained network, however, the performance improvement is not obvious. On the contrary, it even produces drifts and thus causes poorer performance.

Performance comparison for three-classification problem

Finally, we classify CT scan images to three classes to further evaluate the robustness and credibility of DeepDSR. We use 7,398 lung CT scan images, which contain 3,768 lung CT scan images from patients infected by COVID-19, 2,383 ones from normal lung, and 1,247 ones from patients infected by other pneumonia. And 80% images are selected the training set and the remaining images are the test set. We repeatedly conduct the three-classification experiments on the obtained 7,398 images for 100 epochs. Table 7 and Figure 8 give precision, recall, accuracy, and F1-socre of DeepDSR, other three comparative methods, and three individual models.

The results from Table 7 and Figure 8 show that the proposed DeepDSR framework significantly outperforms efficientNet-V2 and Vision transformer in terms of precision, recall, accuracy, and F1-score. DeepDSR is also better than ResNet and three individual models based on the above measurement metrics. For example, DeepDSR computes the best precision of 0.9740, recall of 0.9653, accuracy of 0.9737, and F1-score of 0.9695, outperforming 1.93%, 1.27%, 1.31%,

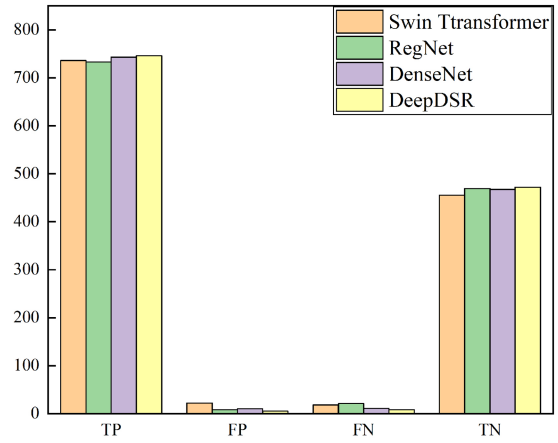

FIGURE 6
Statistical analysis of four methods on 1,231 images.

TABLE 6 The affect of transfer learning on the performance.

	Precision	Recall	Accuracy	F1-score	AUC	AUPR
With pre-train	**0.9833**	**0.9895**	**0.9894**	**0.9864**	**0.9991**	**0.9986**
Without pre-train	0.8773	0.914	0.9171	0.8953	0.9716	0.9455
Without pre-train (200 epoch)	0.9544	0.9224	0.9529	0.9382	0.9866	0.9821

The bold fonts represent the best performance in each column.

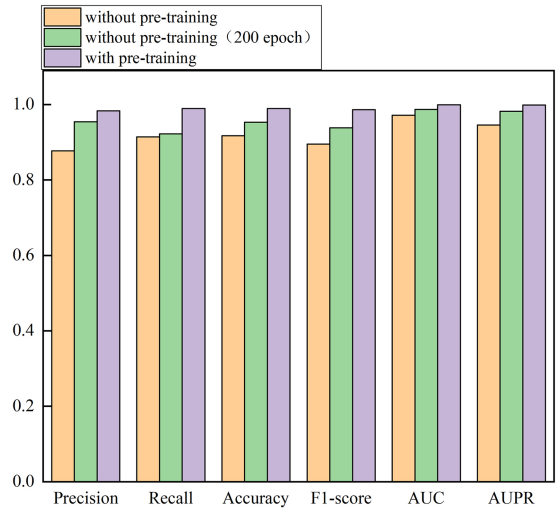

FIGURE 7
The affect of transfer learning on the performance.

TABLE 7 The performance of DeepDSR and other models for three-classification problem.

	Precision	Recall	Accuracy	F1-score
EfficientNet V2	0.4023	0.4479	0.5132	0.3736
ResNet	0.9487	0.9397	0.9541	0.9439
Vision transformer	0.7112	0.6264	0.7373	0.6301
Swin transformer	0.9488	0.9371	0.9548	0.9424
RegNet	0.9492	0.9463	0.9568	0.9476
DenseNet	0.9552	0.953	0.9608	0.9541
DeepDSR	**0.974**	**0.9653**	**0.9737**	**0.9695**

The bold fonts represent the best performance in each column.

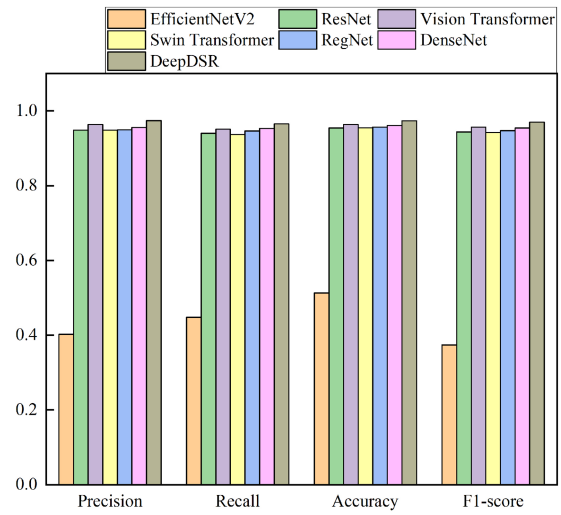

FIGURE 8
The performance of DeepDSR and other models for three-classification problem.

and 1.59 compared the second-best methods (DenseNet, DenseNet, RegNet, and DenseNet), respectively. The results demonstrate that DeepDSR has better generalization ability and can thus be applied to classify COVID-19-related CT scan images.

Conclusion

COVID-19 detection through CT scan images has the characteristics of high sensitivity, low misdiagnosis rate, and high commercial practicability. Therefore, it has been a research hotspot to detect COVID-19 through CT scan images based on deep learning. In this study, we developed a deep ensemble model, DeepDSR to identify CT scan images for patients infected by COVID-19. DeepDSR combined three different state-of-the-art network architectures, DenseNet, Swin transformer, and RegNet.

It obtained the best performance compared to three classical deep learning models (efficientNetV2, ResNet, and Vision transformer) as well as three individual models when classifying CT images to two classes (COVID-19-related or non-COVID-19-related) or three classes (COVID-19-related, normal pneumonia, and healthy lung).

EfficientNetV2, ResNet, and Vision transformer are three state-of-the-art deep learning models with different network architectures. The proposed DeepDSR model computed the best measurement values compared with the three network architectures, demonstrating its optimal image classification ability. Moreover, DeepDSR aggregated three individual deep models, DenseNet, Swin transformer, and RegNet. Lower correlations between the three individual models more obviously reduced the variance of DeepDSR. In addition, DeepDSR also reduced its variance due to the ensemble nature. Therefore, DeepDSR, ensemble of different single models, significantly outperforms the three individual models, thereby suggesting its powerful performance.

Our proposed DeepDSR has three advantages: first, three COVID-19-related CT image datasets were fused to boost the generalization ability of DeepDSR. Moreover, multiple methods including batch normalization were adopted to prevent overfitting. Finally, DeepDSR, ensemble of DenseNet, Swin transformer, and RegNet, can more accurately classify CT images and thus improve the classification performance. However, the training of DeepDSR was more complex than single model, it also spend more time to train and test the model, and more parameters need to be adjusted, thereby requiring more computing resources. In the future, we will design more robust ensemble deep learning models to accurately classify images for query diseases including COVID-19 and cancer. In particular, we will further consider deep heterogeneous ensemble framework to accurately identify images for related diseases by ensemble of deep learning model and supervised learning model.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material. Source code is freely downloadable at: https://github.com/plhhnu/DeepDSR/. Datasets 1-3 can be downloaded from the following three links: https://www.kaggle.com/datasets/plameneduardo/a-covid-multiclass-dataset-of-ct-scans; https://www.kaggle.com/datasets/plameneduardo/sarscov2-ctscan-dataset; https://github.com/UCSD-AI4H/COVID-CT.

Author contributions

LP, CW, MC, and ZL: conceptualization. LP, CW, and ZL: methodology. CW and MC: software. LP, CW, GT, GLiu: validation. LP, MC, and ZL: investigation. CW, GLi, and YL: data curation. LP and CW: writing—original draft preparation. LP, GT, and JY: writing—review and editing. LP: supervision. LP, CW, and MC: project administration. LP, and MC: funding acquisition. All authors contributed to the article and approved the submitted version.

Funding

ZL was supported by National Natural Science Foundation of China under grant no. 62172158. LP was supported by the National Natural Science Foundation of China under grant no. 61803151. GLiu and YL were supported by the Innovation and Entrepreneurship Training Program for College Students of Hunan Province under grant no. S202111535031 and the Innovation and Entrepreneurship Training Program for College Students of Hunan University of Technology under grant no. 20408610119.

Conflict of interest

GT and JY were employed by Geneis (Beijing) Co. Ltd.

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

Publisher's note

References

Abbas, A., Abdelsamea, M. M., and Gaber, M. M. (2021). Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network. *Appl. Intell.* 51, 854–864. doi: 10.1007/s10489-020-01829-7

Afshar, P., Rafiee, M. J., Naderkhani, F., Heidarian, S., Enshaei, N., Oikonomou, A., et al. (2022). Human-level COVID-19 diagnosis from low-dose CT scans using a two-stage time-distributed capsule network. *Sci. Rep.* 12, 1–11. doi: 10.1038/s41598-022-08796-8

Agaoglu, N. B., Yildiz, J., Dogan, O. A., Kose, B., Alkurt, G., Demirkol, Y. K., et al. (2022). COVID-19 PCR test performance on samples stored at ambient temperature. *J. Virol. Methods* 301:114404. doi: 10.1016/j.jviromet.2021.114404

Breiman, L. (1996). Bagging predictors. *Mach. Learn.* 24, 123–140. doi: 10.1007/BF00058655

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Chang, J., Yu, J., Han, T., Chang, H. J., and Park, E. (2017). A method for classifying medical images using transfer learning: A pilot study on histopathology of breast cancer. In *2017 IEEE 19th international conference on e-health networking, applications and services (Healthcom)* (pp. 1–4). IEEE.

Chen, X., Xie, D., Wang, L., Zhao, Q., You, Z. H., and Liu, H. (2018b). BNPMDA: bipartite network projection for MiRNA–disease association prediction[J]. *Bioinformatics* 34, 3178–3186. doi: 10.1093/bioinformatics/bty333

Chen, X., Yin, J., Qu, J., and Huang, L. (2018a). MDHGI: matrix decomposition and heterogeneous graph inference for miRNA-disease association prediction[J]. *PLoS Comput. Biol.* 14:e1006418. doi: 10.1371/journal.pcbi.1006418

Chen, X., Zhu, C. C., and Yin, J. (2019). Ensemble of decision tree reveals potential miRNA-disease associations[J]. *PLoS Comput. Biol.* 15:e1007209. doi: 10.1371/journal.pcbi.1007209

Cherti, M., and Jitsev, J. (2021). Effect of Pre-Training Scale on Intra-and Inter-Domain Full and Few-Shot Transfer Learning for Natural and Medical X-Ray Chest Images. arXiv [Preprint] arXiv:2106.00116.

Chieregato, M., Frangiamore, F., Morassi, M., Baresi, C., Nici, S., Bassetti, C., et al. (2022). A hybrid machine learning/deep learning COVID-19 severity predictive model from CT images and clinical data. *Sci. Rep.* 12, 1–15. doi: 10.1038/s41598-022-07890-1

Chouat, I., Echtioui, A., Khemakhem, R., Zouch, W., Ghorbel, M., and Hamida, A. B. (2022). COVID-19 detection in CT and CXR images using deep learning models. *Biogerontology* 23, 65–84. doi: 10.1007/s10522-021-09946-7

Chouhan, V., Singh, S. K., Khamparia, A., Gupta, D., Tiwari, P., Moreira, C., et al. (2020). A novel transfer learning based approach for pneumonia detection in chest X-ray images. *Appl. Sci.* 10:559. doi: 10.3390/app10020559

Chung, M., Bernheim, A., Mei, X., Zhang, N., Huang, M., Zeng, X., et al. (2020). CT imaging features of 2019 novel coronavirus (2019-nCoV). *Radiology* 295, 202–207. doi: 10.1148/radiol.2020200230

Deepak, S., and Ameer, P. M. (2019). Brain tumor classification using deep CNN features via transfer learning. *Comput. Biol. Med.* 111:103345. doi: 10.1016/j.compbiomed.2019.103345

Dong, X., Yu, Z., Cao, W., Shi, Y., and Ma, Q. (2020). A survey on ensemble learning. *Front. Comp. Sci.* 14, 241–258. doi: 10.1007/s11704-019-8208-z

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: transformers for image recognition at scale. arXiv [Preprint] arXiv:2010.11929.

Fan, X., Feng, X., Dong, Y., and Hou, H. (2022). COVID-19 CT Image Recognition Algorithm based on Transformer and CNN. *Displays* 72:102150. doi: 10.1016/j.displa.2022.102150

Fields, B. K., Demirjian, N. L., Dadgar, H., and Gholamrezanezhad, A. (2021). Imaging of COVID-19: CT, MRI, and PET. *Semin. Nucl. Med.* 51, 312–320. doi: 10.1053/j.semnuclmed.2020.11.003

Friedman, J. H. (2002). Stochastic gradient boosting. *Comput. Stat. Data Analy.* 38, 367–378. doi: 10.1016/S0167-9473(01)00065-2

Harper, L., Kalfa, N., Beckers, G. M. A., Kaefer, M., Nieuwhof-Leppink, A. J., Fossum, M., et al. (2020). The impact of COVID-19 on research[J]. *J. Pediatr. Urol.* 16, 715–716. doi: 10.1016/j.jpurol.2020.07.002

Hastie, T., Rosset, S., Zhu, J., and Zou, H. (2009). Multi-class adaboost. Statistics and its. *Interface* 2, 349–360. doi: 10.4310/SII.2009.v2.n3.a8

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hijab, A., Rushdi, M. A., Gomaa, M. M., and Eldeib, A. (2019). Breast cancer Classification in Ultrasound Images using Transfer Learning. In *2019 Fifth international conference on advances in biomedical engineering (ICABME)* (pp. 1–4). IEEE.

Ho, T. K. (1995). Random Decision Forests. In *Proceedings of 3rd international conference on document analysis and recognition* (*Vol. 1*, pp. 278–282). IEEE.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700–4708).

Jin, S., Wang, B., Xu, H., Luo, C., Wei, L., Zhao, W., et al. (2020). AI-assisted CT imaging analysis for COVID-19 screening: building and deploying a medical AI system in four weeks. medRxiv [Preprint].

Karbhari, Y., Basu, A., Geem, Z. W., Han, G. T., and Sarkar, R. (2021). Generation of synthetic chest X-ray images and detection of COVID-19: A deep learning based approach. *Diagnostics* 11:895. doi: 10.3390/diagnostics11050895

Khalifa, N. E. M., Loey, M., Taha, M. H. N., and Mohamed, H. N. E. T. (2019). Deep transfer learning models for medical diabetic retinopathy detection. *Acta Informatica Medica* 27, 327–332. doi: 10.5455/aim.2019.27.327-332

Ksiazek, T. G., Erdman, D., Goldsmith, C. S., Zaki, S. R., Peret, T., Emery, S., et al. (2003). A novel coronavirus associated with severe acute respiratory syndrome. *N. Engl. J. Med.* 348, 1953–1966. doi: 10.1056/NEJMoa030781

Lee, Y., Kim, Y. S., Lee, D. I., Jeong, S., Kang, G. H., Jang, Y. S., et al. (2022). The application of a deep learning system developed to reduce the time for RT-PCR in COVID-19 detection. *Sci. Rep.* 12, 1–10. doi: 10.1038/s41598-022-05069-2

Liu, W., Jiang, Y., Peng, L., Sun, X., Gan, W., Zhao, Q., et al. (2022a). Inferring gene regulatory networks using the improved Markov blanket discovery algorithm. *Interdiscipl. Sci. Comput. Life Sci.* 14, 168–181. doi: 10.1007/s12539-021-00478-9

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). Swin transformer: hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 10012–10022).

Liu, W., Lin, H., Huang, L., Peng, L., Tang, T., Zhao, Q., et al. (2022b). Identification of miRNA-disease associations via deep forest ensemble learning based on autoencoder. *Brief. Bioinform.* 23:bbac104. doi: 10.1093/bib/bbac104

Mofijur, M., Fattah, I. R., Alam, M. A., Islam, A. S., Ong, H. C., Rahman, S. A., et al. (2021). Impact of COVID-19 on the social, economic, environmental and energy domains: lessons learnt from a global pandemic. *Sustain. Product. Consumpt.* 26, 343–359. doi: 10.1016/j.spc.2020.10.016

Mustafa, B., Loh, A., Freyberg, J., MacWilliams, P., Wilson, M., McKinney, S. M., et al. (2021). Supervised Transfer Learning at scale for Medical Imaging. arXiv [Preprint] arXiv:2101.05913.

Pan, S. J., and Yang, Q. (2009). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* 22, 1345–1359. doi: 10.1109/TKDE.2009.191

Pan, F., Ye, T., Sun, P., Gui, S., Liang, B., Li, L., et al. (2020). Time course of lung changes on chest CT during recovery from 2019 novel coronavirus (COVID-19) pneumonia. *Radiology* 295, 715–721. doi: 10.1148/radiol.2020200370

Peters, A., Vetter, P., Guitart, C., Lotfinejad, N., and Pittet, D. (2020). Understanding the emerging coronavirus: what it means for health security and infection prevention. *J. Hosp. Infect.* 104, 440–448. doi: 10.1016/j.jhin.2020.02.023

Pu, R., Liu, S., Ren, X., Shi, D., Ba, Y., Huo, Y., et al. (2022). The screening value of RT-LAMP and RT-PCR in the diagnosis of COVID-19: systematic review and meta-analysis. *J. Virol. Methods* 300:114392. doi: 10.1016/j.jviromet.2021.114392

Radosavovic, I., Kosaraju, R. P., Girshick, R., He, K., and Dollár, P. (2020). Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10428–10436).

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., et al. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat. Machine Intell.* 3, 199–217. doi: 10.1038/s42256-021-00307-0

Saad Menezes, M. C., Santinelli Pestana, D. V., Ferreira, J. C., Ribeiro de Carvalho, C. R., Felix, M. C., Marcilio, I. O., et al. (2022). Distinct outcomes in COVID-19 patients with positive or negative RT-PCR test. *Viruses* 14:175. doi: 10.3390/v14020175

Sagi, O., and Rokach, L. (2018). "Ensemble learning: A survey," in *Data Mining and Knowledge Discovery*, *Vol. 8*. ed. W. Pedrycz (Hoboken, New Jersey: Wiley Interdisciplinary Reviews) e1249

Shalbaf, A., and Vafaeezadeh, M. (2021). Automated detection of COVID-19 using ensemble of transfer learning with deep convolutional neural network based on CT scans. *Int. J. Comput. Assist. Radiol. Surg.* 16, 115–123. doi: 10.1007/s11548-020-02286-w

Shorten, C., Khoshgoftaar, T. M., and Furht, B. (2021). Deep learning applications for COVID-19. *J. Big Data* 8, 1–54. doi: 10.1186/s40537-020-00392-9

Soares, E., Angelov, P., Biaso, S., Froes, M. H., and Abe, D. K. (2020). SARS-CoV-2 CT-scan dataset: A large dataset of real patients CT scans for SARS-CoV-2 identification. medRxiv [Preprint].

Sun, F., Sun, J., and Zhao, Q. (2022). A deep learning method for predicting metabolite-disease associations via graph neural network. *Brief. Bioinform.* 23, 1–11. doi: 10.1093/bib/bbac266

Sun, H., Wang, A., Wang, L., Wang, B., Tian, G., Yang, J., et al. (2022a). Systematic tracing of susceptible animals to SARS-CoV-2 by a bioinformatics framework. *Front. Microbiol.* 13:781770. doi: 10.3389/fmicb.2022.781770

Tan, M., and Le, Q. (2021). Efficientnetv2: Smaller models and faster training. In *International Conference on Machine Learning* (pp. 10096–10106). PMLR.

Ter-Sarkisov, A. (2022). Covid-ct-mask-net: prediction of covid-19 from CT scans using regional features. *Appl. Intell.* 52, 1–12. doi: 10.1007/s10489-021-02731-6

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Adv. Neural Inf. Proces. Syst.* 30, 5998–6008. doi: 10.48550/arXiv.1706.03762

Voulodimos, A., Doulamis, N., Doulamis, A., and Protopapadakis, E. (2018). Deep learning for computer vision: A brief review. *Comput. Intell. Neurosci.* 2018, 1–13. doi: 10.1155/2018/7068349

Wang, C. C., Han, C. D., Zhao, Q., and Chen, X. (2021). Circular RNAs and complex diseases: from experimental results to computational models. *Brief. Bioinform.* 22:bbab286. doi: 10.1093/bib/bbab286

Wang, D., Hu, B., Hu, C., Zhu, F., Liu, X., Zhang, J., et al. (2020). Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China. *JAMA* 323, 1061–1069. doi: 10.1001/jama.2020.1585

Wang, B., Jin, S., Yan, Q., Xu, H., Luo, C., Wei, L., et al. (2021). AI-assisted CT imaging analysis for COVID-19 screening: building and deploying a medical AI system. *Appl. Soft Comput.* 98:106897. doi: 10.1016/j.asoc.2020.106897

Weiss, K., Khoshgoftaar, T. M., and Wang, D. (2016). A survey of transfer learning. *J. Big Data* 3, 1–40. doi: 10.1186/s40537-016-0043-6

Worldometer (2022). COVID live - coronavirus statistics - Worldometer. Available at: https://www.worldometers.info/coronavirus/ (Accessed July 11, 2022).

Yang, M., Yang, H., Ji, L., Hu, X., Tian, G., Wang, B., et al. (2022). A multi-omics machine learning framework in predicting the survival of colorectal cancer patients[J]. *Comput. Biol. Med.* 146:105516. doi: 10.1016/j.compbiomed.2022.105516

Zhang, J., Xie, Y., Li, Y., Shen, C., and Xia, Y. (2020). Covid-19 Screening on chest x-ray Images using deep Learning based Anomaly Detection. arXiv [Preprint] arXiv:2003.12338, 27.

Zhang, L., Yang, P., Feng, H., Zhao, Q., and Liu, H. (2021). Using network distance analysis to predict lncRNA-miRNA interactions. *Interdiscipl. Sci. Comput. Life Sci.* 13, 535–545. doi: 10.1007/s12539-021-00458-z

Zhao, J., Zhang, Y., He, X., and Xie, P. (2020). Covid-CT-Dataset: a CT scan Dataset about Covid-19. arXiv [Preprint] arXiv:2003.13865, 490.

Zhou, T., Lu, H., Yang, Z., Qiu, S., Huo, B., and Dong, Y. (2021). The ensemble deep learning model for novel COVID-19 on CT images. *Appl. Soft Comput.* 98:106885. doi: 10.1016/j.asoc.2020.106885

Zhou, P., Yang, X. L., Wang, X. G., Hu, B., Zhang, L., Zhang, W., et al. (2020). A pneumonia Outbreak Associated with a new Coronavirus of Probable bat origin. *Nature* 579, 270–273. doi: 10.1038/s41586-020-2012-7

Zhu, W., Braun, B., Chiang, L. H., and Romagnoli, J. A. (2021). Investigation of transfer learning for image classification and impact on training sample size. *Chemom. Intel. Lab. Syst.* 211:104269. doi: 10.1016/j.chemolab.2021.104269

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., et al. (2020). A comprehensive survey on transfer learning. *Proc. IEEE* 109, 43–76. doi: 10.1109/JPROC.2020.3004555

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 04 October 2022
DOI 10.3389/fmicb.2022.963704

OPEN ACCESS

EDITED BY
Qi Zhao,
University of Science and Technology Liaoning, China

REVIEWED BY
Dingjie Wang,
The Ohio State University, United States
Mingzhi Liao,
Northwest A&F University Apple Research Center, China

*CORRESPONDENCE
Yuan Zhu
zhuyuan@cug.edu.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

RECEIVED 07 June 2022
ACCEPTED 16 August 2022
PUBLISHED 04 October 2022

CITATION
Wang C, Zhang H, Ma H, Wang Y, Cai K, Guo T, Yang Y, Li Z and Zhu Y (2022) Inference of pan-cancer related genes by orthologs matching based on enhanced LSTM model. *Front. Microbiol.* 13:963704. doi: 10.3389/fmicb.2022.963704

Inference of pan-cancer related genes by orthologs matching based on enhanced LSTM model

Chao Wang[1†], Houwang Zhang[2†], Haishu Ma[3,4,5], Yawen Wang[6], Ke Cai[3,4,5], Tingrui Guo[3,4,5], Yuanhang Yang[6], Zhen Li[6] and Yuan Zhu[3,4,5,7*]

[1]Department of Surgery, Hepatic Surgery Center, Institute of Hepato-Pancreato-Biliary Surgery, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China, [2]Department of Electrical Engineering, City University of Hong Kong, Kowloon, Hong Kong SAR, China, [3]School of Automation, China University of Geosciences, Wuhan, China, [4]Hubei Key Laboratory of Advanced Control and Intelligent Automation for Complex Systems, Wuhan, China, [5]Engineering Research Center of Intelligent Technology for Geo-Exploration, Wuhan, China, [6]School of Mathematics and Physics, China University of Geosciences, Wuhan, China, [7]Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence, Shanghai, China

Many disease-related genes have been found to be associated with cancer diagnosis, which is useful for understanding the pathophysiology of cancer, generating targeted drugs, and developing new diagnostic and treatment techniques. With the development of the pan-cancer project and the ongoing expansion of sequencing technology, many scientists are focusing on mining common genes from The Cancer Genome Atlas (TCGA) across various cancer types. In this study, we attempted to infer pan-cancer associated genes by examining the microbial model organism Saccharomyces Cerevisiae (Yeast) by homology matching, which was motivated by the benefits of reverse genetics. First, a background network of protein-protein interactions and a pathogenic gene set involving several cancer types in humans and yeast were created. The homology between the human gene and yeast gene was then discovered by homology matching, and its interaction sub-network was obtained. This was undertaken following the principle that the homologous genes of the common ancestor may have similarities in expression. Then, using bidirectional long short-term memory (BiLSTM) in combination with adaptive integration of heterogeneous information, we further explored the topological characteristics of the yeast protein interaction network and presented a node representation score to evaluate the node ability in graphs. Finally, homologous mapping for human genes matched the important genes identified by ensemble classifiers for yeast, which may be thought of as genes connected to all types of cancer. One way to assess the performance of the BiLSTM model is through experiments on the database. On the other hand, enrichment analysis, survival analysis, and other outcomes can be used to confirm the biological importance of the prediction results. You may access the whole experimental protocols and programs at https://github.com/zhuyuan-cug/AI-BiLSTM/tree/master.

KEYWORDS

microbe-disease, orthologs, essential proteins, deep learning, BiLSTM model

1. Introduction

Cancer is a malignant and complex kind of disease that seriously endangers human existence. Because of its rapid spread, early onset, and high death rate, cancer is a disease that is challenging to cure. According to the American Cancer Society, there will be 608,570 cancer-related deaths and 1,898,160 new cases of cancer in the nation in 2021 (Xia et al., 2022). The prevention and treatment of cancer have evolved into a public health issue that requires collective human effort. A growing number of scholars are dedicating themselves to pan-cancer research as it is a hot topic globally. The therapeutic treatment of viral diseases, genetic diseases, and other diseases may be improved by the use of gene therapy (Ma et al., 2020). Therefore, accurate detection of pan-cancer genes is essential for understanding cancer and provides better benefits for its prevention, treatment, and development of anti-cancer drugs, which is relevant from a social and economic perspective (Aromolaran et al., 2021).

Currently, the identification of essential genes is the main source of the issue with pan-cancer associated genes prediction. In previous decades, biological experiments including single gene knockout, conditional knockout, and RNA interference were used as the typical methods for identifying essential proteins. These experimental techniques require lengthy and expensive procedures, and the experimental settings frequently affect the outcomes. The same organism may respond differently to different experimental settings (Zhong et al., 2021). An enormous number of protein-protein interactions (PPI) enriched with gene expression data have been available in recent years benefiting from the advancement of high-throughput technology (Li et al., 2017).

According to the two sides, studies on cancer-related genes can be roughly split into two categories. It is intended to investigate the tissue-specific driver genes, on the one hand. The ideas pertaining to complex network analysis were transferred and utilized to biological network analysis by merging cancer sample data onto biological networks. Each node in the network structure had its level of importance evaluated, and the genes with the highest value were found to be the cancer driver genes. Since genes only selectively express proteins, essential proteins can be used to discover essential genes. Numerous effective network-based techniques have been put forth over years to identify crucial proteins from PIN. The most well-known and straightforward one is degree centrality (DC) (Jeong et al., 2001). According to a molecular theory known as the centrality-lethality rule, the highly linked nodes within the PIN serve as its fundamental structural components and are generally more significant than other nodes (Jeong et al., 2001; Zotenko et al., 2008). Other node topological feature-based methods, such as subgraph centrality (SC) (Estrada and Rodriguez-Velazquez, 2005), eigenvector centrality (EC) (Bonacich, 1987), betweenness centrality (BC) (Joy et al., 2005), closeness centrality (CC) (Wuchty and Stadler, 2003), information centrality (IC) (Stephenson and Zelen, 1989) and others, are also used to identify proteins in addition to DC. These techniques assess each node according to its topological structure. In general, network-based approaches are extensively employed in the early stages since they can predict important proteins directly without the need for further information. However, these techniques feature low recall rates and identification precision due to the abundance of false positive and false negative data in PPI networks (Li et al., 2016). The intrinsic biological importance of necessary proteins is also disregarded by these techniques, which ignores essential proteins with low connectivity (Li et al., 2016). Recent research has attempted to incorporate biological knowledge into network-based techniques, which not only reduce the impact of false positives in PPI data but also significantly increase the prediction accuracy of essential proteins (Li et al., 2012; Zhang et al., 2019; Wang et al., 2021). Ess-NEXG (Wang et al., 2020) and DeepEP (Zeng et al., 2019; Liu et al., 2022b) are two related algorithms for finding essential proteins that have been developed as a result of the rapid growth of deep learning. Other algorithms have also been presented to predict other associations (Zhang et al., 2021; Liu et al., 2022a,b).

On the other side, it seeks to identify potential disease-related genes across a variety of malignancies. Several computational methods have been proposed to uncover pan-cancer related genes or driver module types by integrating multi-omics data across various malignancies (Cao and Zhang, 2016; Zhang and Zhang, 2016, 2017; Yang et al., 2017; Li et al., 2020), which is motivated by the objectives of the cancer genome program named The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013). By combining existing information on cancer from various types of tumors, potential patterns and biological processes are investigated. For example, Park et al. (2016) proposed an algorithm called NTriPath based on matrix decomposition to identify and complement pathogenic gene pathways, which overcomes the limitation of studying a single cancer and can complement the existing set of pathogenic pathway genes in multiple cancers. In order to identify possible pan-cancer related genes, Zhu et al. (2022) combined the network representation method with differential expression analysis.

Geneticists have long noted that functional relationships frequently exist between mutations that result in the same biological manifestation. Utilizing these predictions to connect particular genes to phenotypes opens the door to using similar techniques to directly find new disease genes in the study of human genes. In reverse genetics, it is feasible to infer linked phenotypes based on linkages in functional gene networks (Sommer, 2008). Homologous genes are genes found in several species that descended vertically from a single gene found in the last common ancestor, which is how organisms evolved from a common ancestor. Highly identical DNA sequences

between two homologous genes, which may also have the same function, are extremely likely to be found in two animals with very close affinity (Müller, 2003). The concept of homology allows us to more easily study human genes with gene sequences from other species. Similar structures and functionalities are shared by genes that are crucial for life's functions in model organisms. Furthermore, there is mounting evidence that model species are essential for addressing issues connected to the gene variations that underlie human disease. Using model organisms for homology mapping can help us understand human pathogenic genes (Bleackley and MacGillivray, 2011).

Due to its genetic flexibility, small genome size, and manipulability, yeast is one of the model organisms with the highest genetic adaptability. Yeast is a single-cell eukaryote that helps to uncover many fundamental concepts in biology and reveals the activity of human cells. Consequently, yeast is essential for identifying genetic variations in human genes related to illnesses and encoding genetic variations in proteins engaged in multiple pathways. The study revealed a link between the microbiota and associated diseases, and it is crucial to understand the molecular mechanisms of these diseases in order to develop new microbiome-based therapies. Microbiota is the microbial population colonizing multiple organ systems in humans and impacting the outcomes of microbiota-related diseases (Belkaid and Hand, 2014; Sun et al., 2022). Among them, gut microbiota, a dense microbial community in human intestines, has been found closely associated to acute kidney injury (Lei et al., 2022), atherosclerosis (Anto and Blesso, 2022), reduced bone mineral density (Wan et al., 2022), age-related neuroinflammation and cognitive decline (Alsegiani and Shah, 2022), carcinogenesis and cancer immunotherapy resistance (Hersi et al., 2022), and metabolic disorders such as hyperlipidemia, hyperglycemia, hypertension, obesity and diabetes (Beg et al., 2022). Manipulation of the gut microbiota has broad application prospects on diseases. Fecal microbiota transplant (FMT) is one of the microbiome-based therapeutics with clinical application potential in clostridioides difficile colitis, graft-vs.-host disease, and inflammatory bowel disease (Sorbara and Pamer, 2022). In addition, engineered bacteria, postbiotics, and phages are also used as precision microbiome-centered therapies (Bajaj et al., 2022).

Multiple biological data are currently available due to the advancement of sequencing technologies, enabling it to integrate multi-omics data from various tumors to uncover genes associated to pan-cancer. In this study, we use the yeast network to predict human disease genes. We gathered a pathogenic gene set from multiple cancers. Homologous mapping is then utilized to locate the homologs integrating all of the pathogenic genes of ten tumors. We propose a parameter adaptive model for characterizing node representation ability by merging Subcellular localization information, Gene expression data, and Protein Complexes data with the specifically designed topological properties of the PPI network, which is called PSGN score for short. Additionally, the BiLSTM, a LSTM model with adjacency constraint and multiple features, is proposed for the prediction of essential proteins. The yeast genes that are similar to the seed genes are screened as candidate genes using the BiLSTM algorithm. In order to identify the final predicted human pan-cancer associated genes, homolog mapping of these candidate genes was performed.

Comparative experiments were conducted on the publicly accessible PPI data of Yeast, in order to validate the effectiveness of the proposed evaluation PSGN score and the classification results of BiLSTM. We verified the efficacy of the new proposed score by contrasting the performance of PSGN with classic unsupervised approaches including DC, BC, CC, EC NC, LAC, PeC, and WDC. Further, we compared our BiLSTM model to established machine learning techniques like SVM, decision tree, ensemble learning-based methods, and the most recent deep learning-based approach put forth by Zeng et al. (2019). According to the experimental findings, BiLSTM may identify essential proteins with superior overall outcomes than other cutting-edge techniques. Besides, some biological significance experiments were conducted on real datasets, the results validated the effectiveness of the new proposed algorithm from the reverse genetics perspective. The remaining parts are organized as follows. Section 2 presents the material and methods of the new proposed method. Experimental results and discussions are illustrated in Section 3. Finally, Section 4 concludes the work.

2. Materials and methods

2.1. Datasets

PPI networks: among other species, the PPI network dataset of yeast is the most reliable and complete, making it popular for use in evaluating and identifying essential proteins. Therefore, in this study, we also selected the yeast PPI network dataset. The DIP database is used to gather the PPI data of yeast (Xenarios et al., 2002). There are 5,093 proteins and 24,743 interactions in total after subtracting self-interactions and repetitive interactions.

Essential protein datasets: A list of essential proteins of yeast were collected from the following databases: MIPS (Mewes et al., 2006), SGD (Cherry et al., 2012), and DEG (Zhang and Lin, 2009). A protein in the yeast protein interaction network is considered as an essential protein if it is marked as essential at least in one database. This data has 1,285 essential proteins, 1,167 of which are included in the PPI network constructed from the DIP database. Hence, we take the 1,167 proteins as essential proteins and the rest 3,926 proteins as non-essential ones.

Subcellular localization dataset: the dataset is available in the knowledge channel of COMPARTMENTS database (Binder et al., 2014), which combines the UniProtKB (Magrane, 2011),

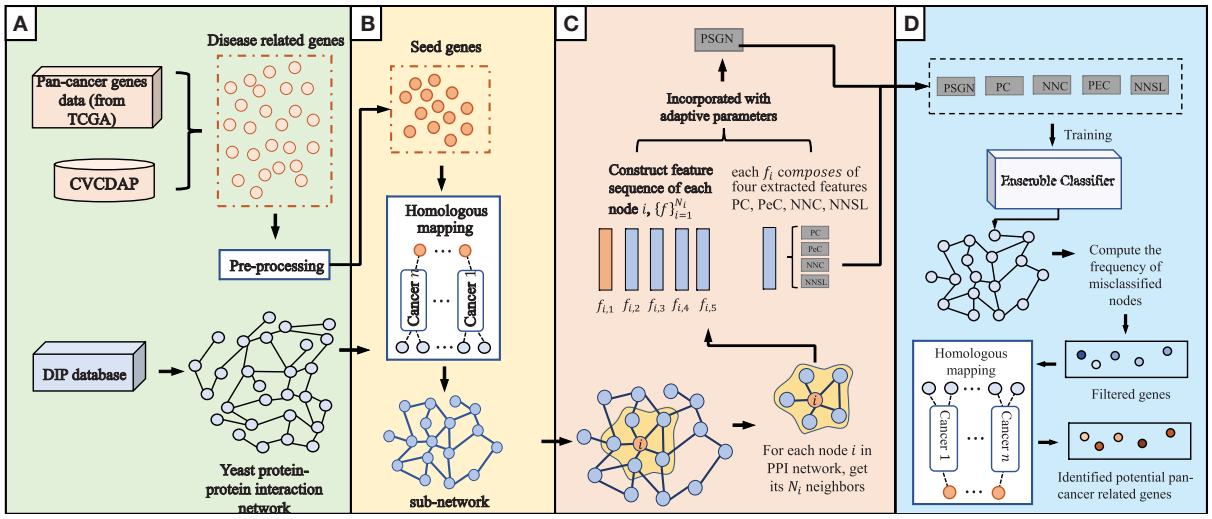

FIGURE 1
An overview of our proposed method to predict pan-cancer related genes *via* orthologs matching. The main algorithm consists of **(A–D)** four parts. **(A)** data integration and pre-processing. **(B)** Homologous mapping. **(C)** Evaluation of node representation ability in graph. **(D)** Prediction of pan-cancer related genes.

MGD (Eppig et al., 2012), SGD (Cherry et al., 2012), FlyBase (Mcquilton et al., 2012), and WormBase datasets (Harris et al., 2010). There are 206,831 subcellular localization records in this dataset, which can be further subdivided into 830 categories.

Protein complex datasets (Luo and Qi, 2015): it is comprised of four real protein complex sets (CM270, CM425, CYC408, and CYC428). Seven hundred and forty-five protein complexes are included in the consolidated dataset.

Gene Expression Omnibus (GEO) dataset: GSE3431 derives from GEO and samples 12 time points during each of three yeast successive metabolic cycles (the interval between two time points is 25 min). The dataset contains 36 samples with 6,777 genes.

Online Mendelian Inheritance in Man (OMIM) dataset: we retained only disease-related variants linked to a genetic disorder listed in the OMIM database. Cross-references were used to directly access annotations for each OMIM disease by downloading the DO (Human Disease Ontology) OBO (Open Biological and Biomedical Ontology) file release. Each retrieved leaf DO term connected to a single OMIM was expanded to include all ancestors and the ontological root term. Term expansion was calculated by parsing the OBO file with an impromptu script.

The Cancer Genome Atlas (TCGA) Database: the Human Genome Research Institute (HGRI) and National Cancer Institute (NCI) launched the Cancer Genome Mapping Project in 2006. The database contains more than 20,000 samples from 33 cancer types, including transcriptome expression data, genome variation data, methylation data, clinical data, and others, which can be accessed *via* https://portal.gdc.cancer.gov/exploration.

2.2. Overview of the new proposed method

The current proposed method, which consists of four main steps of data integration and pre-processing, homologous mapping, evaluation of node representation ability, and prediction of pan-cancer related genes, is shown in detail in Figure 1.

2.2.1. Data integration and pre-processing

The gene expression data of 10 cancers were obtained from TCGA database, including esophageal carcinoma, pancreatic cancer, lung cancer (lung adenocarcinoma, lung squamous cell carcinoma), breast invasive carcinoma, colon adenocarcinoma, rectum adenocarcinoma, cholangiocarcinoma, gastric cancer and ovarian cancer. Due to the duplications and deletions in the pathogenic genes of each cancer, they are used as experimental data after sorting and deletion. We uploaded the TCGA data of 10 cancers selected in the CVCDAP database (https://omics.bjcancer.org/cvcdap/home.do), successfully generated the pan-cancer related pathogenic gene set, and completed the analysis of the pan-cancer network driving genes with the help of the analysis tool of CVCDAP database. We obtained the data of Yeast protein interaction network on DIP database (https://dip.doe-mbi.ucla.edu/dip/Main.cgi) and

downloaded the connection information between Yeast protein nodes directly.

2.2.2. Homologous mapping

We believe that the homologous genes of co-ancestors express themselves similarly. The NCBI Homologene database (https://www.ncbi.nlm.nih.gov/homologene/) compiles homologous gene data for species with complete genome sequencing. In this section, we used Homologene package in the R language, the imported human pathogenic genes were annotated by homology mapping, and the homologous genes of human and yeast genes were taken as seed genes. After identifying the proteins expressed by the seed genes, the interaction network among these yeast proteins can be determined by the STRING database (https://cn.string-db.org/).

2.2.3. Evaluation of node representation ability

Yeast is one of the most genetically model organism. In this study, we firstly explore the essential proteins in the yeast PPI network to further find potential disease related genes. Thus, a new score is defined to evaluate the node representation ability. The PPI network is denoted as graph $G = (V, E)$, where $V = \{v_1, \cdots, v_m\}$ and $E = \{e_{i,j}, 1 \leq i, j \leq m\}$ represent the node set and edge set of the graph, respectively. Specifically, v_i denotes the i-th protein while e_{ij} denotes protein-protein interaction linkage between protein v_i and v_j. $|V| = m$ represents the number of total proteins within G.

The features of our new proposed score considers node-aided biological information, edge-aided biological information and network topological features. We'll go through how to use and integrate this data to create the attributes needed to determine a protein's essentiality in the subsections that follow. The establishment process requires three specific steps.

Step 1: Construction of node represented features

1) Protein complexes score: previous studies indicated that intracellular proteins always tend to connect with their neighbors to form densely connected modules, which are called protein complexes and by this way proteins could take part in more complex and diverse biological activities and functions (Luo and Qi, 2015). Given that essential proteins are crucial in maintaining the main structure and functions in protein complexes (Zotenko et al., 2008), protein complexes data could be used for the identification of essential proteins (Lei et al., 2018).

For the protein v_i, the essentiality tends to be higher if it is found in more protein complexes. In order to calculate the protein complexes (PC) score, we do the following:

$$PC(i) = |Complex(i)| \tag{1}$$

where $Complex(i)$ denotes the sets of protein complexes including v_i, and $|Complex(i)|$ is the number of protein complexes including v_i.

2) Subcellular localization score: it has been proved that proteins must be localized at their appropriate subcellular compartments to perform their desired functions and thus the subcellular localization information is beneficial for the identification of essential proteins (Peng et al., 2015). To ensure the relationships of subcellular localization with the topological features of PPI network, refer to Li et al. (2016), we firstly use the previous feature NNC to sort the proteins within the PPI network, and then calculate the numbers of subcellular location l where the top k% proteins appear and where the bottom k% proteins appear, respectively.

Given the data's false positives, counting proteins at higher rates may result in more errors; as a result, we use $k = 5$ in this work as Li et al. (2016) sets, i.e., that the top/bottom 5% proteins are selected. Besides, we define T_l as the frequency of the localization l where the top k% proteins appear and B_l as the frequency of the localization l where the bottom k% proteins appear. Subcellular localization correlation coefficient SLCC(l) can be calculated by Equation (2)

$$SLCC(l) = \begin{cases} 1 - \dfrac{B_l}{T_l}, & T_l < B_l; \\ \dfrac{T_l}{B_l} - 1, & otherwise, \end{cases} \tag{2}$$

when $T_l < B_l$, it means that more proteins with low NNC values tend to appear in the location l and it is assumed that the relationship between the location l and the essentiality of proteins is negative. On the other hand, when $T_l \geq B_l$, there should also be a positive correlation between the location l and the essentiality of proteins. When $T_l = 0$, we set $SLCC(l)$ as the maximum of $1 - \dfrac{B_l}{T_l}$ with $T_l \neq 0$. And when $B_l = 0$, we set $SLCC(l)$ as the maximum of $\dfrac{T_l}{B_l} - 1$ with $B_l \neq 0$.

Besides, considering that a protein may appear in multiple subcellular locations, take protein v_i for instance, its subcellular localization score $SL(i)$ could be calculated as the sum of $SLCC(l)$ of all the subcellular locations where it appears. Moreover, the normalized value $NSL(i)$ of SL for each protein v_i is used by Equation (3)

$$NSL(i) = \frac{SL(i) + max\_SL}{Max(SL(i) + max\_SL)}, \tag{3}$$

where max_SL represents the maximum value of $SL(i)$ for all the proteins within the PPI network. Maxin the denominator takes for all the nodes within the PPI network.

In order to strengthen the identification precision of subcellular localization, we combine the NSL score with a network topological feature NNEC that is proposed in Zhu and Wu (2018) and has a good compatibility with biological

information. The combined feature is called NNSL for short, for each protein v_i, its $NNSL(i)$ score can be calculated by Equation (4)

$$NNSL(i) = NSL(i) \times NNEC(i), \tag{4}$$

where $NNEC(i) = \sum_{j \in \mathbb{N}(i)} NECC(i,j)$ and $NECC$ can be obtained by Equation (5)

$$NECC(i,j) = \frac{T(i,j)^3 \times C(j)}{\prod_{t=\{i,j\}}(d(t)-1)}, \tag{5}$$

where $T(i,j)$ denotes the number of triangles made up of proteins v_i and v_j, $C(j) = \frac{2\mathbb{E}_j}{d(j)(d(j)-1)}$ is the clustering coefficient of protein v_j, $\mathbb{E}_j$ is the number of non-repetitive edges consisting of all nearest neighbors of v_j. $d(t)$ denotes the degree for protein t, for $t = i$ or j.

Step 2: Construction of edge represented features

Gene expression data is a type of biological information that has been utilized for a long time to compute edge correlations and identify essential proteins. *PeC* is a method that combines gene expression data with edge clustering coefficient *ECC* in order to reduce the impact of false positives on the PPI network. As a result, we apply *PeC* in this study to extract pertinent information from gene expression data. For a protein v_i, its PeC score $PeC(i)$ can be computed by Equation (6)

$$PeC(i) = \sum_{j \in \mathbb{N}(i)} ECC(i,j) \times PCC(i,j), \tag{6}$$

where $ECC(i,j)$ is the edge coefficient between edge $e_{i,j}$, $PCC(i,j)$ is the Pearson's correlation coefficient of a pair of proteins (v_i and v_j). s denotes the length of the gene expression data, which can be calculated by Equation (7)

$$PCC(i,j) = \frac{1}{s-1}\sum_{t=1}^{s}[\frac{g(i,t)-\bar{g}(i)}{\sigma(i)}] \times [\frac{g(j,t)-\bar{g}(j)}{\sigma(j)}], \tag{7}$$

where $g(i,t)$ and $g(j,t)$ are the expression level of v_i and v_j in the sample time t under a specific condition, $\bar{g}(i)$ and $\bar{g}(j)$ represent the mean expression level of v_i and v_j, and $\sigma(i)$ and $\sigma(j)$ represent the standard deviation of expression level of v_i and v_j, respectively.

To extract the topological information of proteins within the PPI network, it is necessary to construct an effective feature representing the network structures of the nodes and connections with neighbors. Network centrality (NC) is a representative topology based method widely used for predicting essential proteins (Wang et al., 2012). Hence, we choose it for network topological feature construction. For the protein v_i, its network centrality $NC(i)$ can be calculated as the sum of edge clustering coefficients $ECC(i,j)$ of each edge $e_{i,j}$ connected with v_i by Equation (8)

$$\begin{aligned} NC(i) &= \sum_{j \in \mathbb{N}(i)} ECC(i,j) \\ &= \sum_{j \in \mathbb{N}(i)} \frac{T(i,j)}{min(d_i - 1, d_j - 1)}, \end{aligned} \tag{8}$$

where $\mathbb{N}(i)$ is the set of nodes which directly connect with protein v_i.

In order to match other features based on biological information, here we use the normalized NC value (denoted as NNC) for each protein. Then for v_i, its normalized value $NNC(i)$ is defined by Equation (9)

$$NNC(i) = \frac{NC(i)}{max(NC(i))}, \tag{9}$$

where $max(NC(i))$ denotes the maximum NC value of all the proteins in the graph G, and the value of $NNC(i)$ will be normalized between 0 and 1.

Step3 : Feature integration by liniear model with adaptive parameters

The structure of hetergeous feature integration involves Protein complex $PC(i)$, Subcellular localization $NNSL(i)$, Gene expression $PeC(i)$ and Network topology $NNC(i)$ multiple information (PSGN). Here, we reconcile these features using a linear model in order to fully integrate this information. Take protein v_i within the PPI network for instance, its evaluation score could be calculated by PSGN score presented in Equation (10)

$$\begin{aligned} PSGN(i) = ((PC(i) + NNSL(i) \times a + PeC(i) \times (1-a)) \\ \times b + NNC(i) \times (1-b) \end{aligned} \tag{10}$$

where a and b are two weights to balance these heterogeneous features. And a is utilized for combining the node based and edge based biological features, b is set to integrate the topological features and biological features.

When integrating numerous pieces of information, several methods for identifying essential proteins require for the adjustment of parameters and the setting of an optimal one for feature combinations. In contrast, our approach proposes an adaptable parameter strategy to deal with various information based on the unique number of essential proteins that must be identified. These are the concepts: depending on the number of essential proteins we need to identify, the adaptive domain of each piece of information varies.

For example, we use PC and NNC two features to identify essential proteins of Yeast PPI dataset respectively. Through

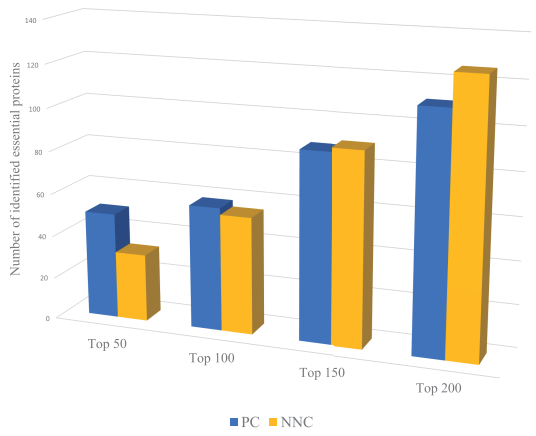

FIGURE 2
The number of essential proteins predicted by PC and NNC in different scales.

Figure 2, PC can capture more essential proteins compared with NNC when dealing with proteins with higher ranking positions. And for proteins with lower positions in rank, the effect of PC is not so significant as NNC. That means when we need to identify the proteins with higher ranking positions (like Top 50, Top 100), we need to assign larger weights on PC. On the contrary, to predict essential proteins with lower ranking positions (like Top 150, Top 200), NNC should be assigned with larger weights. However, most methods will give constant parameters which ignore the variation of functions of different biological information for identification when the number of essential proteins needed to be predicted changes.

In general, the effect of biological information is more reliable than topological features of network when dealing with proteins with higher ranking positions. Therefore, the weight should be adjusted adaptively according to the number of essential proteins needed to be identified. The parameter adaptive model is proposed by Equation (11)

$$P = \alpha_i + \beta_i \times input, \tag{11}$$

where input is the expected number of essential proteins needed to be identified. In this research, $i = 1$ or 2, when $i = 1$, $P = a$, by test, we take $\alpha_1 = 0.49$, $\beta_1 = -0.0005$, when $i = 2$, $P = b$, by test, we take $\alpha_2 = 1$, $\beta_2 = -0.0003$. This parameter model means that, the weights of biological information are greater when calculating the top ranked essential proteins, especially the node-aided biological information (PC and NNSL). With the increase of input, the weight of network-based topological feature (NNC) gradually increases, and the weight of edge-aided biological information (PeC) also increases gradually.

2.3. Prediction of potential pan-cancer related genes

As we discussed above, for proteins in the PPI network, the proteins' feature can be represented by NNC, NNSL, PC, PeC and PSGN. As it is shown in Figure 1B, the seed proteins are labeled as 1 and other proteins in yeast PPI network are labeled as 0. The final prediction results *via* enhanced BiLSTM model *via* repeated experiments as shown in Figure 3. Then the representation can be divided into training dataset and testing dataset, we sample the data from the embedding vector of pan-cancer network based on cross-validation. As shown in Figure 1, this process is trained by multiple classifiers on the sampled data. After obtaining the trained classifiers, we use them to pre- dict pan-cancer-related genes. For each predicted node, the frequency of the node can be considered as the decision metric in the training processes. Finally, the final node representation ability can be calculated by counting the frequency. We take nodes with proper frequencies as potential candidate pan-cancer genes. The whole procedures of our proposed approach AI-BiLSTM are presented in Algorithm 1.

3. Results and discussion

In this research, we investigate the interaction network of the model microbial Yeast, and find potential pan-cancer related genes by homologous mapping. Firstly, the LSTM model was used to categorize the essential genes in the Yeast interaction network, and then homology matching was used to further mine the disease genes. Therefore, the experimental analysis was carried out from two aspects. On the PPI datasets for yeast, we compared the performance of the novel proposed BiLSTM model with several conventional approaches. Secondly, we validated biological significance of the predicted genes through GO enrichment analysis, pathway analysis, survival analysis, clustering analysis and so forth. All of the approaches that are compared in this study adopt their default parameters. All the experiments are run on a personal computer with Windows 10 OS, Intel Core i7 2.3GHz CPU, and 16GB memory.

3.1. Effectiveness of the new proposed BiLSTM model

3.1.1. Evaluation of PSGN

For PSGN, similar to most of validation methods for the identification of essential proteins, we also ranked all proteins by using each essential protein identification method in a descending order. And then we selected a certain number of top

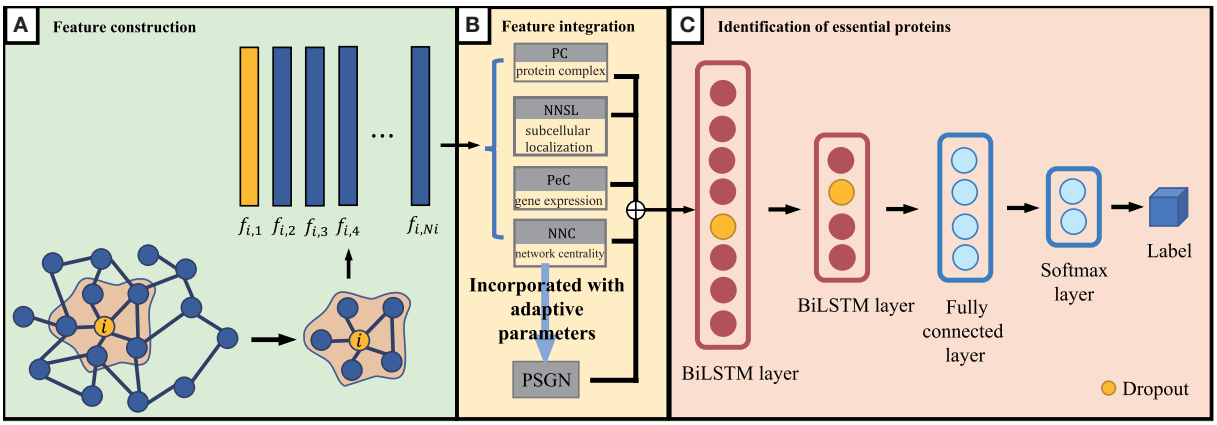

FIGURE 3
An overview of our proposed BiLSTM model. **(A)** Feature construction, **(B)** Feature integration, and **(C)** Identification of essential proteins.

Input: The PPI network $G = (V, E)$, protein complex, subcellular localization, gene expression data, threshold
Output: The classification label for proteins;
Calculate PC for each protein by using Equation (1);
Calculate NNSL for each protein by using Equation (4);
Calculate PeC for each protein by using Equation (6);
Calculate NNC for each protein by using Equation (9);
Incorporate PC, NNSL, PeC and NNC by using the adaptive parameters through Equation 10 to obtain PSGN score;
Integrate protein feature representation enhanced by [NNC, NNSL, PC, PeC, PSGN];
for i $(1 \rightarrow n)$ do
Random select (1/n)% data as training dataset, others as test dataset;
Classification and fix the protein label by BiLSTM;
end
Count the frequency of the predicted essential gene (labeled 1)
return The genes with frequencies greater than the threshold.

Algorithm 1. BiLSTM for prediction of potential pan-cancer related genes.

ranked proteins as the essential protein candidates (like top 100), after that the accuracy of identification could be computed by counting the number of true essential proteins.

Figure 4 gives a specific comparison of the results of identification of essential proteins. As shown in the figure, PSGN can identify more essential proteins compared with the other eight methods. The number of true essential proteins identified by PSGN is higher than other methods in the top 100, top 200, top 300, top 400, top 500, and top 600 proteins. In addition, by observing the results of the top 100 proteins, we find that PSGN can obtain a prediction precision of 90%, which is much higher than other methods.

For better comparison, the precision-recall (PR) curve, a common methodology for evaluating the performance of essential proteins identification methods, is used in this paper. The comparison of our method with the other methods for predicting essential proteins on the Yeast PPI network by using the PR curve is shown in Figure 5. The PR curve of PSGN obtains the better result compared to the PR curves of other methods. Our method significantly exceeds other methods with the largest AUC value, illustrating the effectiveness of our method.

To further evaluate its effectiveness, we take the jackknife curve to compare the prediction results of our proposed method PSGN with other methods. The results are shown in Figure 6. The x-axis denotes the number of proteins ranked by each essential protein identification method and the y-axis is the number of truly identified essential proteins of each method. The areas under the jackknife curves can measure the performances of the method for identifying essential proteins. As shown in Figure 6, the jackknife curve of our proposed method PSGN can identify more essential proteins from the Yeast PPI network compared with other methods, demonstrating that PSGN is more effective and can get better results than other state-of-art methods.

For interpreting the advantages of our method in deeper levels, we also choose 5 widely used metrics (sensitive, specificity,

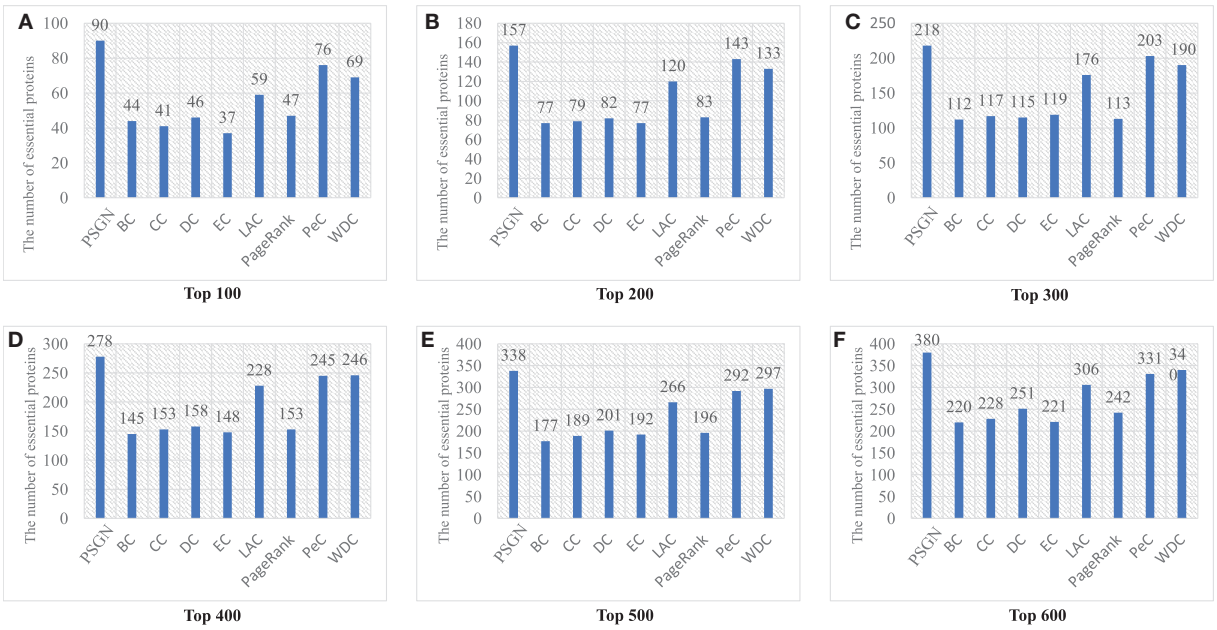

FIGURE 4
The number of essential proteins predicted by PSGN, BC, CC, DC, EC, LAC, PageRank, PeC, and WDC. **(A–F)** show the results of these methods when selecting top 100 to 600 ranked proteins as candidates of essential proteins.

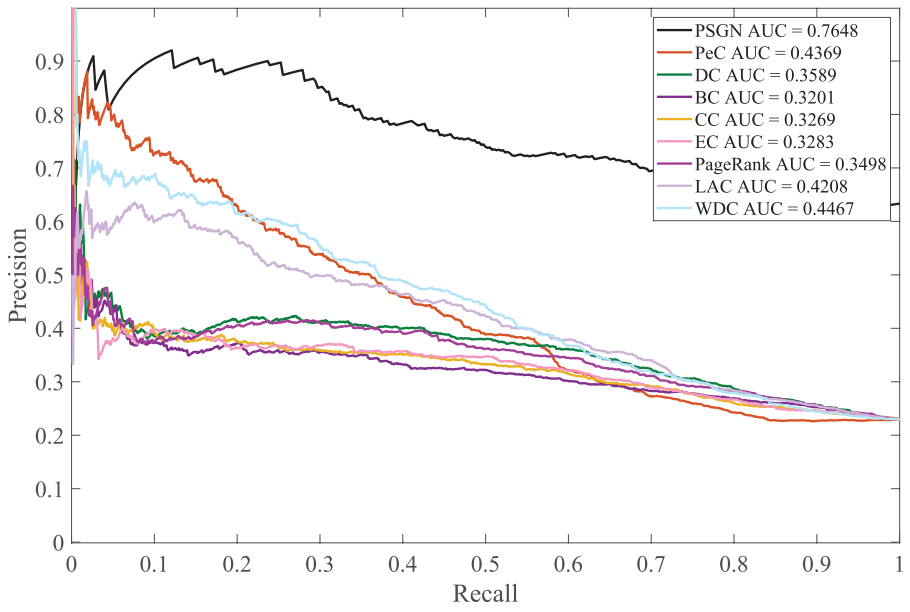

FIGURE 5
Comparison of PSGN, BC, CC, DC, EC, LAC, PageRank, PeC, and WDC using precision-recall (PR) curve method.

precision, F-measure, and accuracy) to evaluate all the methods. Figure 7 shows the results of 5 evaluation metrics obtained by all identification methods on the PPI network of Yeast. As shown in the figure, it is obvious that our proposed PSGN can outperform other methods significantly in terms of all 5 evaluation metrics.

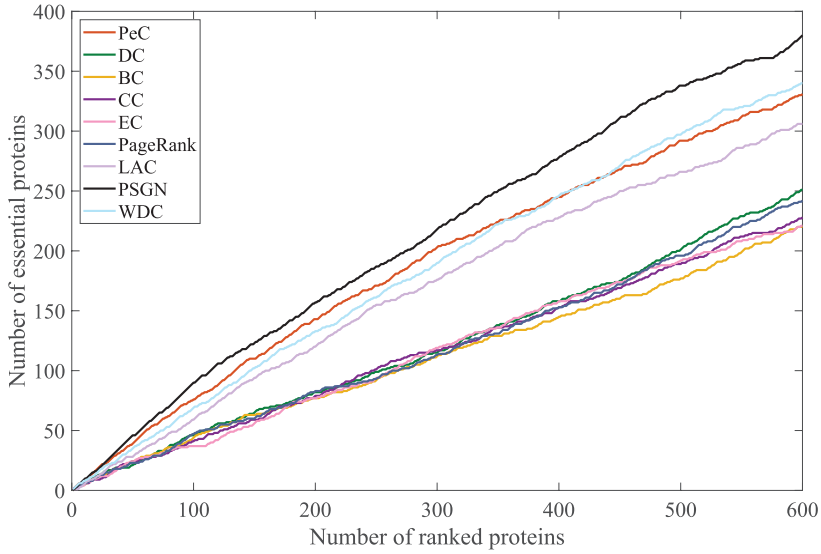

FIGURE 6
Comparison of PSGN, BC, CC, DC, EC, LAC, PageRank, PeC, and WDC using Jackknife method.

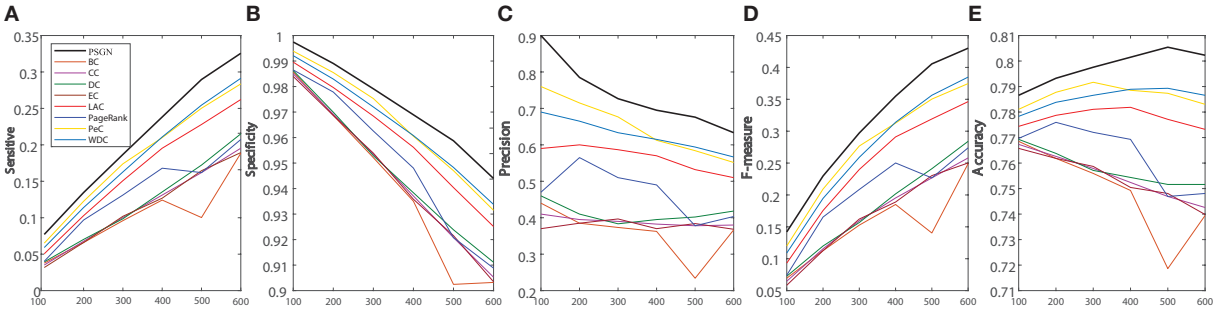

FIGURE 7
Comparative experiments on the Yeast PPI networks in terms of sensitive **(A)**, specificity **(B)**, precision **(C)**, F-measure **(D)**, and accuracy **(E)** obtained by PSGN, BC, CC, DC, EC, LAC, PageRank, PeC, and WDC.

3.1.2. Evaluate of the classified performance of BiLSTM

Machine learning algorithms like SVM, decision tree (DT), random forest (RF) and adaboost are widely used in the tasks of bioinformatics. For fair comparison with these machine learning methods, as the setting in the work of Zeng et al. (2019), we use the sequences composed of integrated biological features PC, PeC, NNSL, the topological feature NNC and the integrate feature PSGN as the input of these machine learning algorithms for training and testing. Besides, we also compared with the algorithm proposed by Zeng et al. (2019).

AI-BiLSTM proposed in this research achieved improved performance compared with other state-of-the-art algorithms with the highest value marked in bold in Table 1. Our model obtains recall, F-measure and AUC with values of 0.5674, 0.5134, and 0.6781, respectively, which are better than SVM, decision tree, random forest, Adaboost, and Zeng et al. (2019). Although

TABLE 1 Comparison of performance between our model and other machine learning algorithms.

Classifier	Accuracy	Precision	Recall	F-measure	AUC
SVM	**0.7654**	**0.4931**	0.3037	0.3759	0.6045
RF	0.7252	0.4295	0.5527	0.4833	0.6651
DT	0.7134	0.3809	0.3713	0.3760	0.5942
Adaboost	0.7409	0.4347	0.3797	0.4054	0.6150
Zeng et al. (2019)	0.7055	0.3802	0.4219	0.3999	0.6067
BiLSTM	0.7369	0.4803	**0.5742**	**0.5231**	**0.6829**

our model does not show the highest values in terms of accuracy and precision and the performance is slightly weaker than SVM in these two assessments, our model owns much better recall, F-measure and AUC. In general, BiLSTM is superior to all other methods.

Besides, for verifying the significance of each feature, we make an ablation test on the features including PC, PeC, NNC, and NNSL. In the ablation experiments, we remove a feature to observe its effect on the identification of essential proteins.

Table 2 shows that NNSL takes the most crucial role in prediction of essential proteins (lowest value marked in bold). The score of accuracy, F-measure, and AUC will drop dramatically without NNSL.

In this section, we compared our BiLSTM with the traditional methods like DC, CC, BC, EC, NC, LAC, PeC, WDC, and PSGN. For fair comparision, 20% of top ranked proteins scored by classical methods are treated as the essential proteins, the rest are regarded as non-essential proteins. Comparing with the list of essential proteins, we can calculate the scores of accuracy, precision, recall, F-measure and AUC of each method.

As the experimental results shown in Table 3, we can find that the scores of our BiLSTM in terms of precision, recall, F-measure, and AUC are significantly higher than the results of DC, BC, CC, EC, NC, LAC, PeC, WDC, and PSGN, which also illustrates the remarkable performance of our method for identifying essential proteins.

TABLE 2 Experimental results for ablation test.

Features	Accuracy	Precision	Recall	F-measure	AUC
Without PC	0.7409	0.4615	**0.4918**	0.476	0.6556
Without NNSL	**0.7222**	0.4111	0.5086	**0.4547**	**0.6469**
Without PeC	0.7242	**0.3833**	0.5769	0.4606	0.6694
Without NNC	0.7311	0.4491	0.5637	0.5000	0.6736
Without PSGN	0.7340	0.4688	0.5674	0.5134	0.6781
BiLSTM	0.7369	0.4803	0.5742	0.5231	0.6829

TABLE 3 Comparison of performance between our proposed non-local GNN and other classical methods.

Method	Accuracy	Precision	Recall	F-measure	AUC
DC	0.7335	0.4050	0.3470	0.3737	0.5977
CC	0.7150	0.3580	0.3067	0.3304	0.5716
BC	0.7139	0.3550	0.3041	0.3276	0.5699
EC	0.7194	0.3690	0.3161	0.3405	0.5777
LAC	0.7563	0.4630	0.3967	0.4273	0.6299
NC	0.7469	0.4390	0.3761	0.4051	0.6166
PeC	0.7555	0.4610	0.3950	0.4254	0.6288
WDC	**0.7630**	0.4800	0.4113	0.4430	0.6394
PSGN	0.7614	0.4771	0.4301	0.4524	0.6450
LSTM-AM	0.7340	0.4688	0.5674	0.5134	0.6781
BiLSTM	0.7369	**0.4803**	**0.5742**	**0.5231**	**0.6829**

Bold values mean best scores.

3.2. Analyze biology significance of the new proposed method

Human disease phenotypes share corresponding orthologs in Yeast gene sets. The BiLSTM model, which was firstly established based on Yeast gene sets, has been further validated in human disease gene sets. In order to reasonably extrapolating the proposed model in microbiota-diseases, genes known to be associated formed a seed set. For the test of human disease gene prediction, we collected sets of Yeast genes whose human orthologs were linked to the same OMIM disease. Human disease phenotypes from OMIM were collapsed into major categories.

3.2.1. Identification of pan-cancer related genes

In the experiments, we selected 10 kinds of cancers as the research objects, including esophageal carcinoma, pancreatic cancer, lung cancer (lung adenocarcinoma, lung squamous cell carcinoma), breast invasive carcinoma, colon adenocarcinoma, rectum adenocarcinoma, cholangiocarcinoma, gastric cancer and ovarian cancer, which can be obtained from the TCGA dataset. Due to the duplications of pathogenic genes between cancers, a total of 17,126 pathogenic genes were obtained after weight removal.

We believed that the common ancestor genes were similar in expression, so we did homology mapping on the background PPI network to find the homologous genes of human genes and Yeast genes. Then we take these genes as seed genes, a total of 1,166 homologous genes were found. Besides, we collected a total of 1,166 proteins expressed by seed genes and obtained the protein-protein interaction network using the STRING database. As it is shown in Figure 8A, it can be found that the corresponding Yeast proteins have a strong correlation with each other, which lays a foundation for our subsequent experiments. Through inputting the seed genes combined with the constructed PSGN features into the proposed BiLSTM algorithm, potential genes which are similar to seed genes will be predicted with corresponding scores. Predicted genes with score greater than 8 were screened out and regarded as candidate genes. By homologous mapping candidate genes, the homologous genes of these genes in human were found as the final predicted genes, and a total of 365 final predicted genes were obtained which is shown in Figure 8B. To further validate the biological significance of the predicted cancer related genes, we conducted a series of biological analysis like GO enrichment

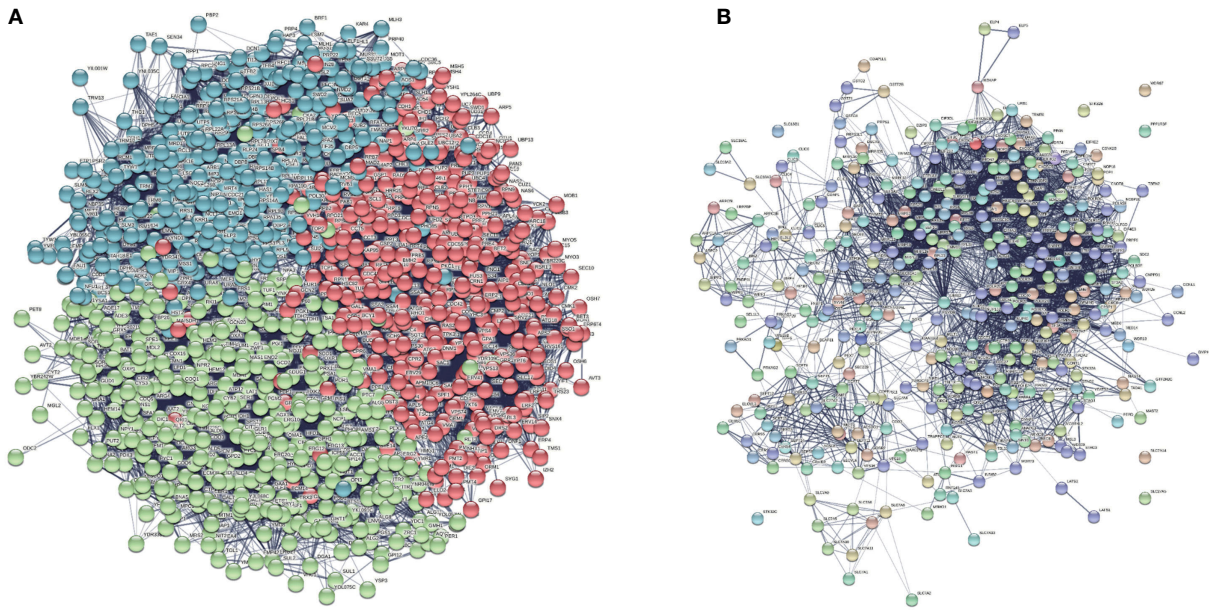

FIGURE 8
We found a total of 1,166 proteins expressed by seed genes and obtained the interaction network between these Yeast proteins screening in the background Yeast PPI network. **(A)** Generated Yeast protein interaction network. **(B)** Clustered protein interaction network with nine communities.

analysis, KEGG pathway analysis, clustering analysis in the following sections.

3.2.2. GO enrichment analysis

For the GO items, we analyzed the relationships of final predicted genes with pan-cancers. According to the ranking of the error rate (FDP), 10 functional annotations with the largest statistical significance were obtained from Biological Process (BP), Cellular Component (CC) and Molecular Function (MF) three branches of GO datasets. As is shown in Table 4, we can find that genes are highly correlated with several important biological processes such as transcription, mRNA splicing, rRNA binding and processing, and cytokinesis, which proved the inner correlations with these predicted genes. What's more, the occurrence sites also involve several cellular sites such as nucleoplasm, ribosome and cytosols, which indicates that these predicted genes are highly related to cell development and possibility with the growth of tumors.

Besides, during clustering analysis, eight modular subneworks M_0 to M_7 enriched in much more CGC genes with higher compactness structures are showed in Figure 9A. Specifically, we find that six of our predicted pan-cancer related genes are enriched in these modulars. Besides, for each module of gene lists, pathway and process enrichment analysis has been carried out with the ontology sources. The results are showed in Figure 9B.

3.2.3. KEGG pathway enrichment analysis

By KEGG pathway enrichment analysis of the predicted genes, we obtained five pathways with the highest correlation with these genes like Proteasome (map03050), Valine, leucine and isoleucine degradation (map00280), Terpenoid backbone biosynthesis (map00900), Mismatch repair (map03430), and Glutathione metabolism (map00480). Among these pathways, the proteasome pathway was the most enriched pathway, which are usually used as an inhibitor in the cancer therapy.

3.2.4. Survival analysis

To verify the biological significance of the experimental results, we conducted further survival analysis. As shown in Figure 10, EIF4A3, NHP2L1, and UBA52 are the three genes with the highest moderate prediction of human genes, which are closely related to RNA metabolic function. Here, we carried out a survival analysis of these three genes, respectively, and it can be seen from the results that all these three genes have a significant impact on the survival time of Bladder urothelial carcinoma (BLCA) patients, which verifies the performance of

TABLE 4 Ten functional annotations with the largest statistical significance for three branches in GO database.

Category	Term	Function	*p*-value	FDR
BP	GO:0006412	Translation	4.59573E-17	3.92797E-20
	GO:0000398	mRNA splicing *via* spliceosome	1.06421E-14	1.81916E-17
	GO:0061640	Cytoskeleton-dependent cytokinesis	2.2982E-11	5.89282E-14
	GO:0006749	Glutathione metabolic process	7.84992E-10	2.68373E-12
	GO:0006364	rRNA processing	1.48788E-09	6.35848E-12
	GO:0002181	Cytoplasmic translation	2.12271E-09	1.25302E-11
	GO:0034613	Cellular protein localization	2.12271E-09	1.27E-11
	GO:1903241	U2-type prespliceosome assembly	4.36481E-08	2.98449E-10
	GO:0006351	Transcription, DNA-templated	2.9991E-07	2.307E-09
	GO:0016575	Histone deacetylation	4.39654E-07	3.75773E-09
CC	GO:0005654	Nucleoplasm	1.91891E-33	5.1171E-36
	GO:0005829	Cytosol	1.29113E-15	8.91902E-18
	GO:0005940	Septin ring	1.29113E-15	1.72151E-17
	GO:0032153	Cell division site	1.29113E-15	1.72151E-17
	GO:0031105	Septin complex	1.29113E-15	1.72151E-17
	GO:0071005	U2-type precatalytic spliceosome	1.28732E-14	2.05972E-16
	GO:0005681	Spliceosomal complex	3.15875E-13	5.89633E-15
	GO:0005840	Ribosome	3.95972E-12	8.44741E-14
	GO:0046540	U4/U6 × U5 tri-snRNP complex	5.04467E-11	1.21072E-12
	GO:0005666	DNA-directed RNA polymerase III complex	4.88524E-10	1.30273E-11
MF	GO:0003735	Structural constituent of ribosome	5.78912E-16	3.91157E-18
	GO:0003899	DNA-directed 5'-3' RNA polymerase activity	7.15936E-13	7.2561E-15
	GO:0005515	Protein binding	1.02098E-11	1.3797E-13
	GO:0060090	Binding, bridging	5.14905E-09	8.69772E-11
	GO:0046933	Proton-transporting ATP synthase activity, Rotational mechanism	1.20677E-06	2.44616E-08
	GO:0003743	Translation initiation factor activity	1.45514E-06	3.44121E-08
	GO:0000340	RNA 7-methylguanosine cap binding	1.51118E-06	4.08426E-08
	GO:0050291	Sphingosine N-acyltransferase activity	1.71826E-06	5.22443E-08
	GO:0015179	L-amino acid transmembrane transporter activity	2.08402E-06	7.04059E-08
	GO:0019843	rRNA binding	2.57108E-05	9.55469E-07

the new proposed prediction method from the respective of homologous matching.

4. Conclusion

High-throughput techniques and machine learning approaches, combined with an increasing understanding of the microbiota and their collective genome from preclinical and large-scale clinical studies, offer exciting opportunities for modernizing microbe-based strategies from untargeted to precision microbiome-centered therapies. Essential proteins have drawn attention for their crucial roles in controlling signal transduction, individual variation in treatment response, and a wide range of other microbiome-related processes. The properties and purposes of biological data used to identify critical proteins are explored in this study. In light of the findings, we suggest a linear adaptive model PSGN, which may adaptively modify the weights for balancing each type of biological or topological property. We have demonstrated that the NNSL feature is significantly more important than other features through experimental validation. Moreover, the new algorithm PSGN improved the ability to represent features discriminatively. In the experiments, we first contrasted the PSGN with established methods including PageRank, DC, BC, CC, EC NC, LAC, PeC, and WDC. The results demonstrated that PSGN outperforms the other approaches in terms of overall performance. Furthermore, we evaluate our BiLSTM with machine learning methods and the most recent deep learning-based methods. The results of experiments

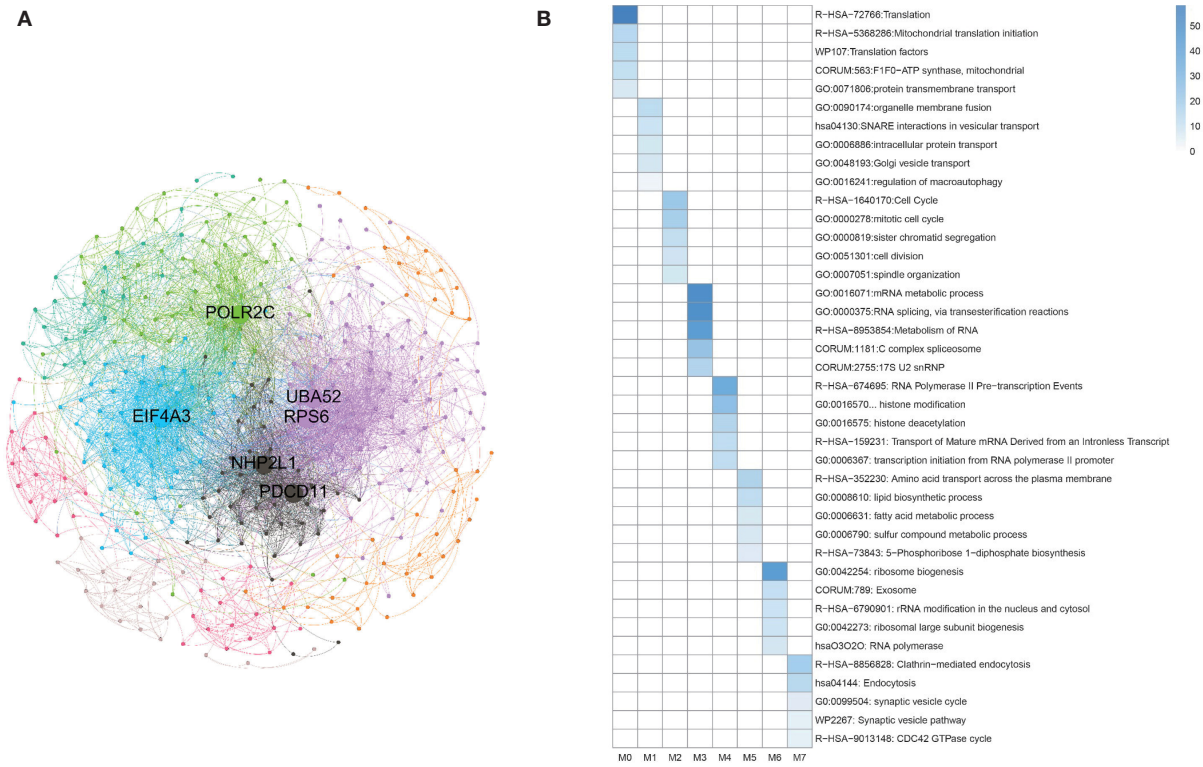

FIGURE 9
Clustering analysis for matched homologous genes in human. **(A)** Topological organization for eight modular sub-networks marked with different colors. **(B)** Enrichment biological functions of pan-cancer sub-networks. Each row represents a GO BP term and each column corresponds to a pan-cancer sub-network for each subnetwork.

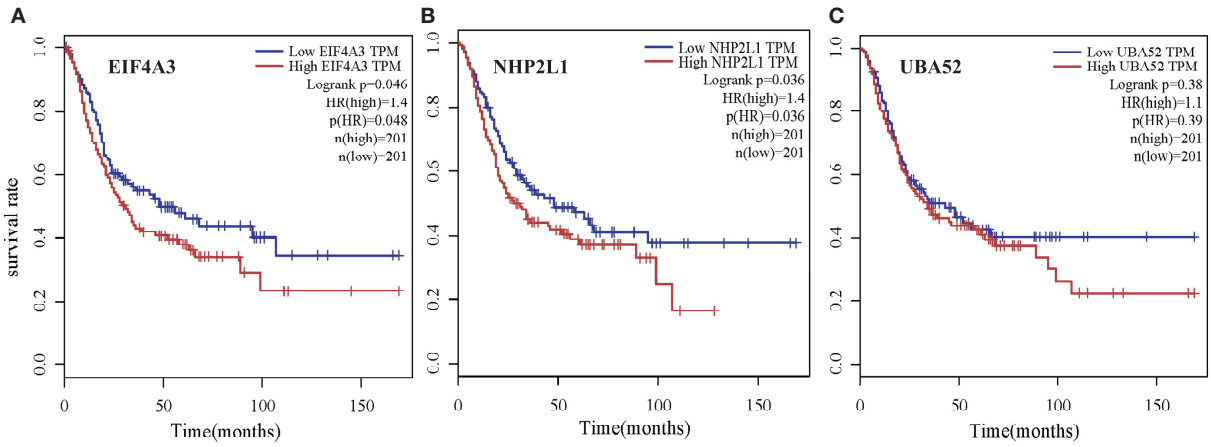

FIGURE 10
Survival analysis of three genes with the highest moderate prediction of human genes, which are closely related to RNA metabolic function for Bladder urothelial carcinoma (BLCA) patients. **(A)** EIF4A3, **(B)** NHP2L1, and **(C)** UBA52.

may potentially establish the capability of the new proposed BiLSTM. Our suggested models for biological information have considerable generality, making them suitable for integrating almost all biological features. In the future, we will continue to test and search for more suitable biological information for identifying essential proteins in more species.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

CW: conceptualization. HZ and HM: methodology. HZ and KC: software. CW, TG, YW, YY, and ZL: validation. HZ and YZ: writing—original draft preparation. All authors have read and agreed to the published version of the manuscript.

Funding

This work was partially supported by the National Natural Science Foundation of China (12126367 and 12126305), Chen Xiao-Ping Foundation for the Development of Science and Technology of Hubei Province (CXPJJH12000002-2020058), the Hubei Provincial Natural Science Foundation of China (2015CFA010), Fundamental Research Funds for the Central Universities, China University of Geosciences (Wuhan) (CUGGC02), and Shanghai Municipal Science and Technology Major Project (2018SHZDZX01), Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence (LCNBI), and ZJLab.

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.

Publisher's note

References

Alsegiani, A. S., and Shah, Z. A. (2022). The influence of gut microbiota alteration on age-related neuroinflammation and cognitive decline. *Neural Regenerat. Res.* 17, 2407. doi: 10.4103/1673-5374.335837

Anto, L., and Blesso, C. N. (2022). Interplay between diet, the gut microbiome, and atherosclerosis: role of dysbiosis and microbial metabolites on inflammation and disordered lipid metabolism. *J. Nutr. Biochem.* 105, 108991. doi: 10.1016/j.jnutbio.2022.108991

Aromolaran, O., Aromolaran, D., Isewon, I., and Oyelade, J. (2021). Machine learning approach to gene essentiality prediction: a review. *Brief. Bioinform.* 22, bbab128. doi: 10.1093/bib/bbab128

Bajaj, J. S., Ng, S. C., and Schnabl, B. (2022). Promises of microbiome-based therapies. *J. Hepatol.* 76, 1379–1391. doi: 10.1016/j.jhep.2021.12.003

Beg, R., Gonzalez, K., and Martinez-Guryn, K. (2022). Implications of microbe-mediated crosstalk in the gut: impact on metabolic diseases. *Bioch. Biophys. Acta* 1867, 159180. doi: 10.1016/j.bbalip.2022.159180

Belkaid, Y., and Hand, T. W. (2014). Role of the microbiota in immunity and inflammation. *Cell* 157, 121–141. doi: 10.1016/j.cell.2014.03.011

Binder, J. X., Pletscher-Frankild, S., Tsafou, K., Stolte, C., O'Donoghue, S. I., et al. (2014). Compartments: unification and visualization of protein subcellular localization evidence. *Database* 2014, bau012. doi: 10.1093/database/bau012

Bleackley, M. R., and MacGillivray, R. T. (2011). Transition metal homeostasis: from yeast to human disease. *Biometals* 24, 785–809. doi: 10.1007/s10534-011-9451-4

Bonacich, P. (1987). Power and centrality: a family of measures. *Am. J. Sociol.* 92, 1170–1182. doi: 10.1086/228631

Cao, Z., and Zhang, S. (2016). An integrative and comparative study of pan-cancer transcriptomes reveals distinct cancer common and specific signatures. *Sci. Rep.* 6, 1–13. doi: 10.1038/srep33398

Cherry, J. M., Hong, E. L., Amundsen, C., Balakrishnan, R., Binkley, G., et al. (2012). Saccharomyces genome database: the genomics resource of budding yeast. *Nucleic Acids Res.* 40, 700–705. doi: 10.1093/nar/gkr1029

Eppig, J. T., Blake, J. A., Bult, C. J., Kadin, J. A., and Richardson, J. E. (2012). The mouse genome database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. *Nucleic Acids Res.* 40, 881–886. doi: 10.1093/nar/gkr974

Estrada, E., and Rodriguez-Velazquez, J. A. (2005). Subgraph centrality in complex networks. *Phys. Rev. E* 71, 056103. doi: 10.1103/PhysRevE.71.056103

Harris, T. W., Antoshechkin, I., Bieri, T., Blasiar, D., Chan, J., et al. (2010). WormBase: a comprehensive resource for nematode research. *Nucleic Acids Res.* 38, 463–467. doi: 10.1093/nar/gkp952

Hersi, F., Elgendy, S. M., Al Shamma, S. A., Atell, R. T., Sadiek, O., and Omar, H. A. (2022). Cancer immunotherapy resistance: the impact of microbiome-derived short-chain fatty acids and other emerging metabolites. *Life Sci.* 300, 120573. doi: 10.1016/j.lfs.2022.120573

Jeong, H., Mason, S. P., Barabási, A. L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature* 411, 41–42. doi: 10.1038/35075138

Joy, M. P., Brock, A., Ingber, D. E., and Huang, S. (2005). High-betweenness proteins in the yeast protein interaction network. *J. Biomed. Biotechnol.* 2005, 96. doi: 10.1155/JBB.2005.96

Lei, J., Xie, Y., Sheng, J., and Song, J. (2022). Intestinal microbiota dysbiosis in acute kidney injury: novel insights into mechanisms and promising therapeutic strategies. *Ren. Fail.* 44, 571–580. doi: 10.1080/0886022X.2022.2056054

Lei, X., Fang, M., Wu, F., and Chen, L. (2018). Improved flower pollination algorithm for identifying essential proteins. *BMC Syst. Biol.* 12, 129–140. doi: 10.1186/s12918-018-0573-y

Li, G., Li, M., Wang, J., Wu, J., Wu, F., and Pan, Y. (2016). Predicting essential proteins based on subcellular localization, orthology and PPI networks. *BMC Bioinform.* 17, 571–581. doi: 10.1186/s12859-016-1115-5

Li, M., Ni, P., Chen, X., Wang, J., Wu, F., et al. (2017). Construction of refined protein interaction network for predicting essential proteins. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 16, 1386–1397. doi: 10.1109/TCBB.2017.2665482

Li, M., Zhang, H., Wang, J., and Pan, Y. (2012). A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data. *BMC Syst. Biol.* 6, 1–9. doi: 10.1186/1752-0509-6-15

Li, Y., Jiang, T., Zhou, W., Li, J., Li, X., et al. (2020). Pan-cancer characterization of immune-related lncRNAs identifies potential

oncogenic biomarkers. *Nat. Commun.* 11, 1–13. doi: 10.1038/s41467-020-14802-2

Liu, W., Jiang, Y., Li, P., Sun, X., Gan, W., et al. (2022a). Inferring gene regulatory networks using the improved Markov blanket discovery algorithm. *Interdisc. Sci. Comput. Life Sci.* 14, 168–181. doi: 10.1007/s12539-021-00478-9

Liu, W., Lin, H., Huang, L., Peng, L., Tang, T., Zhao, Q., et al. (2022b). Identification of miRNA-disease associations *via* deep forest ensemble learning based on autoencoder. *Brief. in Bioinform.* 23, bbac104. doi: 10.1093/bib/bbac104

Luo, J., and Qi, Y. (2015). Identification of essential proteins based on a new combination of local interaction density and protein complexes. *PLoS ONE* 10, e0131418. doi: 10.1145/2818302

Ma, C., Wang, Z., Xu, T., He, Z., and Wei, Y. (2020). The approved gene therapy drugs worldwide: from 1998 to 2019. *Biotechnol. Adv.* 40, 107502. doi: 10.1016/j.biotechadv.2019.107502

Magrane, M. (2011). Uniprot knowledgebase: a hub of integrated protein data. *Database* 2011, bar009. doi: 10.1093/database/bar009

Mcquilton, P., Pierre, S. E. S., and Thurmond, J. (2012). Flybase 101-the basics of navigating flybase. *Nucleic Acids Res.* 40, D706-D714. doi: 10.1093/nar/gkr1030

Mewes, H., Frishman, D., Mayer, K. F. X., Munsterkotter, M., Noubibou, O., Pagel, P., et al. (2006). MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res.* 34, 169–172. doi: 10.1093/nar/gkj148

Müller, G. B. (2003). Homology: the evolution of morphological organization. *Origin. Organ. Beyond Gene Dev. Evolut. Biol.* 2, 51. doi: 10.7551/mitpress/5182.001.0001

Park, S., Kim, S.-J., Yu, D., Pena-Llopis, S., Gao, J., Park, J. S., et al. (2016). An integrative somatic mutation analysis to identify pathways linked with survival outcomes across 19 cancer types. *Bioinformatics* 32, 1643–1651. doi: 10.1093/bioinformatics/btv692

Peng, X., Wang, J., Wang, J., Wu, F., and Pan, Y. (2015). Rechecking the centrality-lethality rule in the scope of protein subcellular localization interaction networks. *PLoS ONE* 10, e0130743. doi: 10.1371/journal.pone.0130743

Sommer, R. J. (2008). Homology and the hierarchy of biological systems. *Bioessays* 30, 653–658. doi: 10.1002/bies.20776

Sorbara, M. T., and Pamer, E. G. (2022). Microbiome-based therapeutics. *Nat. Rev. Microbiol.* 20, 365–380. doi: 10.1038/s41579-021-00667-9

Stephenson, K., and Zelen, M. (1989). Rethinking centrality: methods and examples. *Soc. Networks* 11, 1–37. doi: 10.1016/0378-8733(89)90016-6

Sun, F., Sun, J., and Zhao, Q. (2022). A deep learning method for predicting metabolite-disease associations *via* graph neural network. *Brief. Bioinform.* 23, bbac266. doi: 10.1093/bib/bbac266

Wan, X., Eguchi, A., Fujita, Y., Ma, L., Wang, X., Yang, Y., et al. (2022). Effects of (R)-ketamine on reduced bone mineral density in ovariectomized mice: a role of gut microbiota. *Neuropharmacology* 213, 109139. doi: 10.1016/j.neuropharm.2022.109139

Wang, C., Han, C., Zhao, Q., and Chen, X. (2021). Circular RNAs and complex diseases: from experimental results to computational models. *Brief. Bioinform.* 22, bbab286. doi: 10.1093/bib/bbab286

Wang, J., Li, M., Wang, H., and Pan, Y. (2012). Identification of essential proteins based on edge clustering coefficient. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 9, 1070–1080. doi: 10.1109/TCBB.2011.147

Wang, N., Zeng, M., Zhang, J., Li, Y., and Li, M. (2020). Ess-"NEXG: predict essential proteins by constructing a weighted protein interaction network based on node embedding and xgboost," in *International Symposium on Bioinformatics Research and Applications* (Moscow: Springer), 95–104.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* 45, 1113–1120. doi: 10.1038/ng.2764

Wuchty, S., and Stadler, P. F. (2003). Centers of complex networks. *J. Theor. Biol.* 223, 45–53. doi: 10.1016/S0022-5193(03)00071-7

Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S., and Eisenberg, D. (2002). DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* 30, 303–305. doi: 10.1093/nar/30.1.303

Xia, C., Dong, X., Li, H., Cao, M., Sun, D., et al. (2022). Cancer statistics in china and united states, 2022: profiles, trends, and determinants. *Chin. Med. J.* 135, 584–590. doi: 10.1097/CM9.0000000000002108

Yang, X., Gao, L., and Zhang, S. (2017). Comparative pan-cancer DNA methylation analysis reveals cancer common and specific patterns. *Brief. Bioinform.* 18, 761–773. doi: 10.1093/bib/bbw063

Zeng, M., Li, M., Wu, F.-X., Li, Y., and Pan, Y. (2019). DeepEP: a deep learning framework for identifying essential proteins. *BMC Bioinform.* 20, 506–510. doi: 10.1186/s12859-019-3076-y

Zhang, J., and Zhang, S. (2016). The discovery of mutated driver pathways in cancer: models and algorithms. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 15, 988–998. doi: 10.1109/TCBB.2016.2640963

Zhang, J., and Zhang, S. (2017). Discovery of cancer common and specific driver gene sets. *Nucleic Acids Res.* 45, e86-e86. doi: 10.1093/nar/gkx089

Zhang, L., Yang, P., Feng, H., Zhao, Q., and Liu, H. (2021). Using network distance analysis to predict lncRNA-miRNA lnteractions. *Interdisc. Sci. Comput. Life Sci.* 13, 535–545. doi: 10.1007/s12539-021-00458-z

Zhang, R., and Lin, Y. (2009). Deg 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic Acids Res.* 37, 455–458. doi: 10.1093/nar/gkn858

Zhang, W., Xu, J., and Zou, X. (2019). Predicting essential proteins by integrating network topology, subcellular localization information, gene expression profile and go annotation data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 17, 2053–2061. doi: 10.1109/TCBB.2019.2916038

Zhong, J., Tang, C., Peng, W., Xie, M., Sun, Y., Tang, Q., et al. (2021). A novel essential protein identification method based on ppi networks and gene expression data. *BMC Bioinform.* 22, 1–21. doi: 10.1186/s12859-021-04175-8

Zhu, Y., and Wu, C. (2018). "Identification of essential proteins using improved node and edge clustering coefficient," in *The 37th Chinese Control Conference (CCC)* (Wuhan), 1543–1547.

Zhu, Y., Zhang, H., Yang, Y., Zhang, C., Ou-Yang, L., et al. (2022). Discovery of pan-cancer related genes *via* integrative network analysis. *Brief. Funct. Genomics* 21, 325–338. doi: 10.1093/bfgp/elac012

Zotenko, E., Mestre, J., O'Leary, D. P., and Przytycka, T. M. (2008). Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS Comput. Biol.* 4, e1000140. doi: 10.1371/journal.pcbi.1000140

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 04 November 2022
DOI 10.3389/fmicb.2022.1024104

OPEN ACCESS

EDITED BY
Qi Zhao,
University of Science and Technology Liaoning, China

REVIEWED BY
Guangzhou Xiong,
Huazhong University of Science and Technology, China
Xueying Zeng,
Ocean University of China, China

*CORRESPONDENCE
Zejun Li
lzjfox@hnit.edu.cn
Xueming Luo
lionver@hut.edu.cn
Lihong Peng
plhhnu@163.com

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

RECEIVED 21 August 2022
ACCEPTED 16 September 2022
PUBLISHED 04 November 2022

CITATION
Tian G, Wang Z, Wang C, Chen J, Liu G, Xu H, Lu Y, Han Z, Zhao Y, Li Z, Luo X and Peng L (2022) A deep ensemble learning-based automated detection of COVID-19 using lung CT images and Vision Transformer and ConvNeXt. *Front. Microbiol.* 13:1024104. doi: 10.3389/fmicb.2022.1024104

A deep ensemble learning-based automated detection of COVID-19 using lung CT images and Vision Transformer and ConvNeXt

Geng Tian[1,2†], Ziwei Wang[1†], Chang Wang[1], Jianhua Chen[3], Guangyi Liu[1], He Xu[1], Yuankang Lu[1], Zhuoran Han[4], Yubo Zhao[5], Zejun Li[6*], Xueming Luo[1*] and Lihong Peng[1,7*]

[1]School of Computer Science, Hunan University of Technology, Zhuzhou, China, [2]Geneis (Beijing) Co., Ltd., Beijing, China, [3]Hunan Storm Information Technology Co., Ltd., Changsha, China, [4]High School Attached to Northeast Normal University, Changchun, China, [5]No. 2 Middle School of Shijiazhuang, Shijiazhuang, China, [6]School of Computer Science, Hunan Institute of Technology, Hengyang, China, [7]College of Life Sciences and Chemistry, Hunan University of Technology, Zhuzhou, China

Since the outbreak of COVID-19, hundreds of millions of people have been infected, causing millions of deaths, and resulting in a heavy impact on the daily life of countless people. Accurately identifying patients and taking timely isolation measures are necessary ways to stop the spread of COVID-19. Besides the nucleic acid test, lung CT image detection is also a path to quickly identify COVID-19 patients. In this context, deep learning technology can help radiologists identify COVID-19 patients from CT images rapidly. In this paper, we propose a deep learning ensemble framework called VitCNX which combines Vision Transformer and ConvNeXt for COVID-19 CT image identification. We compared our proposed model VitCNX with EfficientNetV2, DenseNet, ResNet-50, and Swin-Transformer which are state-of-the-art deep learning models in the field of image classification, and two individual models which we used for the ensemble (Vision Transformer and ConvNeXt) in binary and three-classification experiments. In the binary classification experiment, VitCNX achieves the best recall of 0.9907, accuracy of 0.9821, F1-score of 0.9855, AUC of 0.9985, and AUPR of 0.9991, which outperforms the other six models. Equally, in the three-classification experiment, VitCNX computes the best precision of 0.9668, an accuracy of 0.9696, and an F1-score of 0.9631, further demonstrating its excellent image classification capability. We hope our proposed VitCNX model could contribute to the recognition of COVID-19 patients.

KEYWORDS

COVID-19, CT scan image, deep ensemble, Vision Transformer, ConvNeXt

Introduction

In March 2020, the World Health Organization declared COVID-19 as an international pandemic disease due to its rapid and strong transmission (Cascella et al., 2022). Until 22 April 2022, the pandemic has caused about 6.213 million deaths worldwide, over 505.8 million people have been infected with this virus, and there are up to ∼700 thousand new cases within 24 h of that time (Geneva: World Health Organization, 2020; Wang et al., 2021). Different from SARS, the new coronavirus did not disappear quickly or cause limited losses (Stadler et al., 2003). On the contrary, its Delta and Omicron variants induced new pandemics worldwide after multiple mutations (Vasireddy et al., 2021; V'kovski et al., 2021; Yu et al., 2021; Del Rio et al., 2022). It has also caused a sustained impact on the global economy. Long-term shutdowns left many people unemployed. Many countries enforced lockdowns during periodical outbreaks, which resulted in a global economic recession (Alshater et al., 2021; Padhan and Prabheesh, 2021). Although vaccines have been researched and developed to prevent COVID-19 transmission to a certain extent, there is still a need to adopt various methods to detect the virus and prevent its spread.

As a highly contagious respiratory disease, the clinical symptoms of COVID-19 are similar to the common flu and common pneumonia, for instance, coughing, dyspnea, dizziness, and some mild symptoms (Zhang et al., 2020). But the patient infected by the novel coronavirus may deteriorate into fatal acute respiratory distress syndrome in a very short period of time (Guan, 2020). As a result, it greatly increases the difficulty of its early detection and places higher demands on the healthcare system for its treatment. Therefore, the efficient and accurate identification of COVID-19 in patients has become a key to preventing its spread. The nucleic acid test is currently the most widely used due to its high accuracy, simple operation, and low cost (Tahamtan and Ardebili, 2020). But the paucity of standard laboratory environments with specially trained staff has limited the entire testing process.

As an alternative, the non-invasive detection technology, Computed Tomography (CT) provides a new rapid detection method for detecting COVID-19. After the patient has undergone a lung CT scan, experienced radiologists can quickly find typical lesions in the patient's lungs, such as ground-glass opacity, consolidation, and interlobular interstitial thickening by reading the CT images (Chung et al., 2020; Xu et al., 2020). We can also detect COVID-19 in a short time by combining patients' clinical symptoms and investigating recent social situations using epidemiological survey methods. It can help medical workers and epidemic management departments to quickly deal with patients and deploy new prevention and control strategies, and thus intervene in the treatment of patients as early as possible to control its contagion.

However, during the initial stage of the epidemic outbreak, the massive influx of patients often means medical staff and healthcare professionals have to work 24 h a day, which has a bad effect on the physical and mental health of doctors and affects the accuracy and efficiency of the medical diagnosis (Zhan et al., 2021). Alternatively, artificial intelligence technology is a quite efficient strategy and obtains wide application in various fields (Chen et al., 2019; Liu et al., 2021a, 2022a,b; Tang et al., 2021; Wang et al., 2021; Zhang et al., 2021; Liang et al., 2022; Sun et al., 2022; Yang et al., 2022), and can be used to complement the work of radiologists. It can efficiently assist medical staff in judging symptoms, for example, pre-classifying pathological images or predicting sampling results, and thus can greatly reduce their working intensity. Particularly, deep learning has achieved optimal performance in medical image processing (Munir et al., 2019). For instance, Sohail et al. (2021) used a modified deep residual neural network to detect pathological tissue images of breast cancer and implemented automated tumor grading by detecting cell mitosis. Similarly, Codella et al. (2017) introduced a deep ensemble model for pathological image segmentation of skin cancer and the detection of melanoma to improve the detection efficiency of skin cancer. Dou et al. (2016) established a three-dimensional multi-layer convolution model to detect pulmonary nodules in lung stereoscopic CT images, thereby reducing the false positive rate of automated pulmonary nodule detection. Farooq and Hafeez (2020) proposed a ResNet-based COVID-19 screening system to assist radiologists to diagnose. Aslan et al. (2021) developed a new type of COVID-19 infection detection system based on convolutional neural networks (CNN) by combining the long short-term memory (LSTM) network model. These methods effectively improved the identification performance of COVID-19-related CT images. In this paper, we propose a deep-learning ensemble model by integrating Vision Transformer (Dou et al., 2016) and ConvNeXt (Liu et al., 2022c) to effectively improve the prediction accuracy of COVID-19-related CT images.

Materials and methods

Materials

We constructed a comprehensive dataset by integrating and screening data from three lung CT datasets (Soares et al., 2020; Yang et al., 2020). Dataset 1 contained a total of 4,171 images, where 2,167 images were from COVID-19 patients, 757 were from healthy people, and 1,247 were from other pneumonia patients. Dataset 2 contained a total of 2,481 images, where 1,252 images were from COVID-19 patients, and 1,229 were from healthy people; both datasets 1 and 2 were from São Paulo, Brazil. Dataset 3 was from Wuhan, China, and included 746 CT

images, of which 349 were from COVID-19 patients and 397 were from healthy people. Using these datasets we constructed an integrated dataset with a total of 7,398 CT images, which had 3,768 CT images of COVID-19 patients, 2,383 healthy CT images, and 1,247 CT images of other pneumonia patients.

Methods

We investigated various CNN and transformer models and chose Vision Transformer and ConvNeXt as the basic classifier of the ensemble model.

Vision transformer

Transformers have been widely used in the natural language processing field since it was proposed in 2017 (Vaswani et al., 2017). It constructs basic decoder units by connecting the feed-forward neural network and the self-attention mechanism (Bahdanau et al., 2014), as well as adding an encoder-decoder self-attention layer between the two network structures. It creates a brand-new structure that differs from CNN while obtaining relatively high accuracy. The self-attention mechanism used in the transformer first converts the input text into an embedding vector based on word embedding progress. Next, the obtained embedding vectors are used as inputs (named Queries, Keys, and Values) of the self-attention mechanism by a series of multiplication operations. Finally, the output of the self-attention layer is computed using Equation (1) and is fed to the next fully connected layer.

$$\text{Attention}\,(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

$$d_k = \dim\,(K)$$

In 2020, Dosovitskiy et al. built Vision Transformer for image classification. It achieved powerful classification ability comparable to the top CNN models on multiple datasets (CIFAR-100, ImageNet, etc.) (Dosovitskiy et al., 2020).

As shown in Figure 1, the main architecture of the Vision Transformer model is mainly composed of three parts: First is the embedding layer which is used to convert an image into a vector that the transformer encoder can recognize. It also plays a role in embedding position information. The second is the transformer encoder layer which is used to extract features. Finally, a multi-layer perceptron head is used to feature dimension reduction and classify images.

The embedding layer

We used Vision Transformer-B/16-224 to classify COVID-19-related images. The procedure for embedding the layer is shown in Figure 2. First, an original image is resized to the following dimensions: 224*224*3. Second, the image is segmented into blocks of 16*16*3 according to the VIT-B/16-224 configuration, thereby generating 14*14 = 196 (224/16 = 14) blocks. Third, each block is mapped on a 768-dimensional vector through linear mapping. Finally, a matrix of 196*768 size is obtained as the basic input token.

In the original transformer model, all vectors need to embed position vectors to represent the spatiotemporal information of the original input. Similarly, Vision Transformer takes the location information as a trainable parameter and adds it to the token after the image is converted into a vector. The token is extended by one dimension, and a trainable parameter that represents the class or label is added to this new dimension to represent the original class or label of the token for training. The obtained final vector is input into the Transformer Encoder as a token.

Transformer encoder layer

As shown in Figure 3, the encoder layer mainly includes layer normalization (LN), multi-head attention (MHA) block, dropout, and multi-layer perceptron (MLP) block. The core of this structure is the parallel attention mechanism processing layer called multi-head attention. First, the input token matrix is normalized through layer normalization. Second, three matrices Q, K, and V are obtained by multiplying $W^Q$and W^K, *which* are the same as the self-attention module. Third, Q, K, and V are divided into a matrix equal to the number of heads h by multiples of W_i^Q, W_i^K, W_i^V. The corresponding Q_i, K_i, V_i matrix of each head is then used to compute the respective attention score using Equation (2):

$$\text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \qquad (2)$$

$$W_i^Q \in \mathbb{R}^{d_{model}\times d_q}, W_i^K \in \mathbb{R}^{d_{model}\times d_k}, W_i^V \in \mathbb{R}^{d_{model}\times d_v},$$
$$d_q = d_k = d_v = d_{model}/h$$

Finally, the output of the MHA layer is obtained by concatenating all heads and multiplying a matrix-like full connection using Equation (3):

$$\text{MultiHead}(Q, K, V) = \text{Concat}\left(\text{head}_1, \ldots, \text{head}_h\right)W^o \qquad (3)$$
$$W^o \in \mathbb{R}^{hd_v\times d_{model}}$$

The output of the entire transformer encoder layer can be obtained through a residual connection both before and after the MHA and MLP layers. And the encoder layer of the entire model is usually formed by stacking multiple transformer encoders.

MLP head

The main role of the MLP head is to obtain the high-dimensional features and obtain the final classification result.

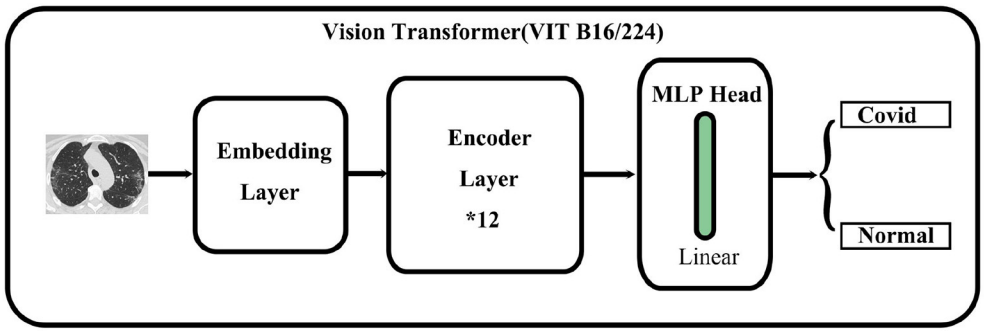

FIGURE 1
Concise structure of Vision Transformer.

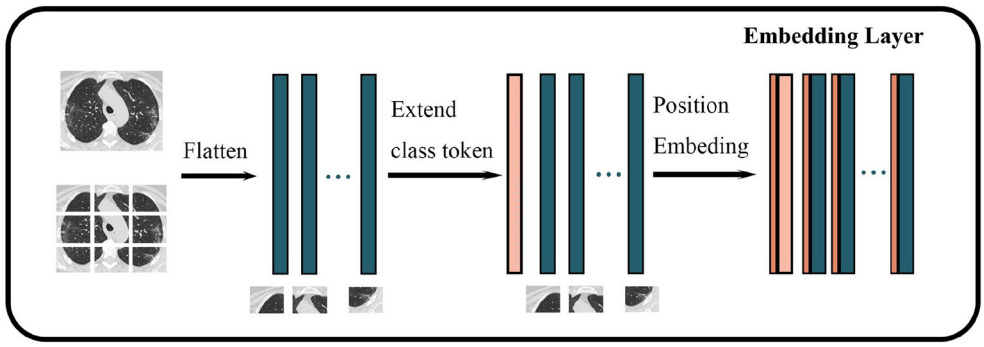

FIGURE 2
Structure of embedding layer in Vision Transformer. The darker green wider rectangles represent the flattened feature vector of each block of an image, while the pink wider rectangles represent the feature vectors corresponding to classes, and the brown narrower rectangle represents the spatiotemporal information of the image.

The outputs of the transformer encoder layer (197*768 in VIT B16/224) are used to compute the classification probability of an image. That is, the output of the transformer encoder layer is a 197*768 matrix, whose sizes are the same as the input of the transformer encoder layer. Finally, only one 768-dimensional vector is used as the input for the MLP head to obtain the classification result of an image corresponding to the matrix.

ConvNeXt

CNN is a classic neural network structure. Lenet was used for handwritten digit recognition as the earliest convolutional neural network model (LeCun et al., 1989). Due to the limitation of the lack of computer performance and the difficulty of collecting large-scale datasets in the 1990s, CNN did not achieve outstanding results in the 20 years that followed. In 2012, Krizhevsky et al. (2012) proposed the AlexNet CNN model, which defeated all image classification models at the ILSVRC2012 competition (Russakovsky et al., 2015). The following CNN models, for instance, VGGNet (Simonyan and Zisserman, 2014) and GoogleNet (Szegedy, 2015), have become prevalent in many AI application fields. The concept of residual and bottleneck layer proposed by the ResNet (He et al., 2016) model in 2015 again improved the performance of CNN. It effectively avoids the gradient problem caused by deeper layers. The generative adversarial network (GAN) proposed by Goodfellow et al. (2014) divided the network into two parts including generation and discriminator based on game theory to achieve better performance through iterative evolutions.

Since the transformer structure came into being in 2020, CNN has not become obsolete. On the contrary, the ConvNeXt network was introduced. ConvNeXt absorbs the advantages of multiple transformer structures in the network structure setting and parameter selection. It outperformed the most powerful transformer model named swin-transformer (Liu et al., 2021a,b) on the ImageNet-1K dataset by adjusting training parameter settings, optimizer, and convolution kernel sizes.

As shown in Figure 4, ConvNeXt has a pretty concise structure. Its performance is greatly improved to the original ResNet although it is quite similar to ResNet. Moreover, it not

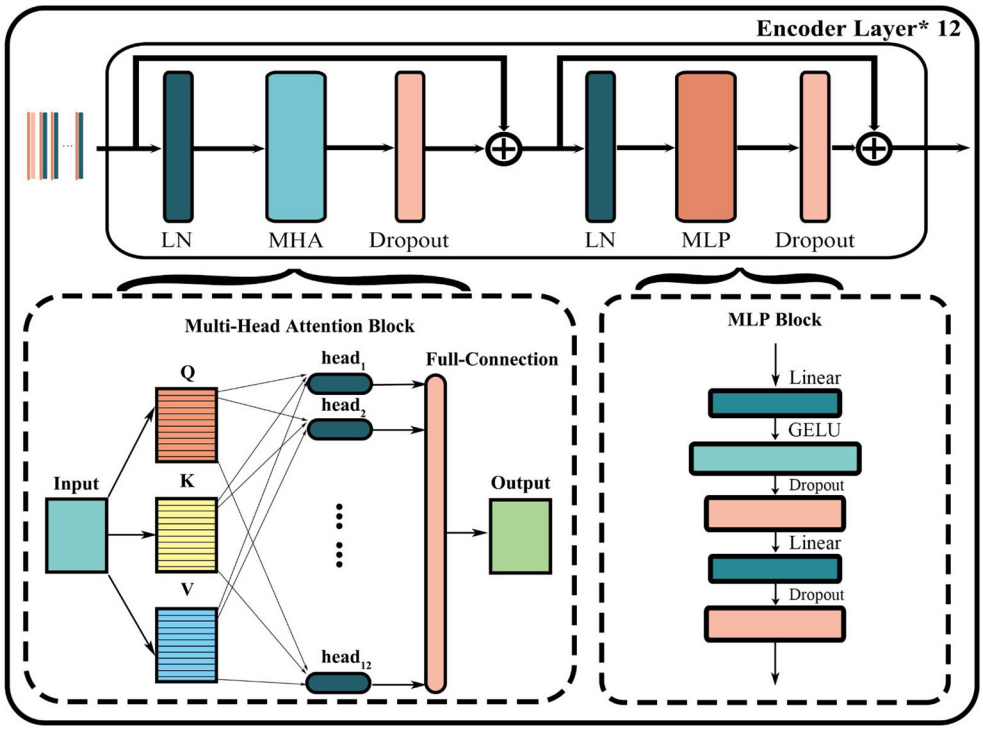

FIGURE 3
Structure of encoder layer in Vision Transformer.

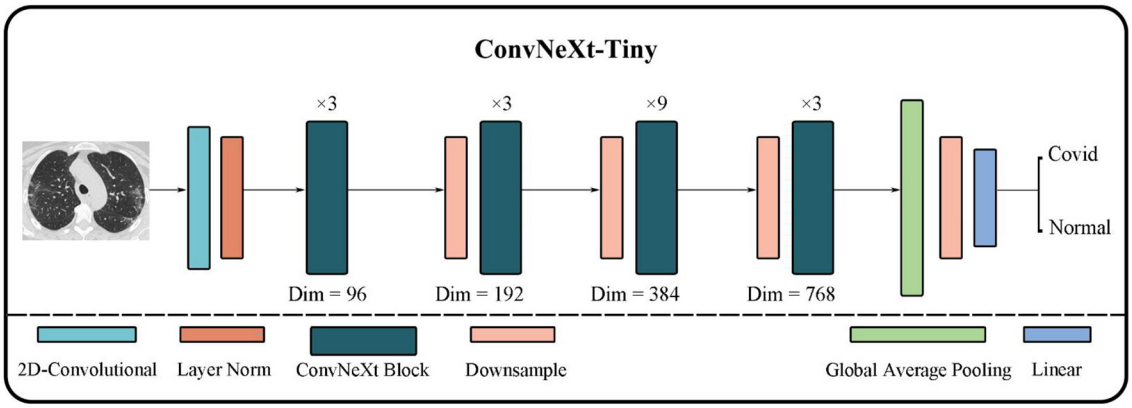

FIGURE 4
Concise structure of ConvNeXt.

only demonstrates better performance than many classic CNN models but also outperforms many transformer models.

First, ConvNeXt starts training ResNet-50 using techniques similar to training transformer models, such as better optimizers, more efficient hyper-parameter settings, and new data augmentation methods. Second, various new optimization strategies are gradually applied to optimize the model, for instance, setting new layer numbers and larger convolution kernels. And eventually, ConvNeXt outperforms the transformer model on the ImageNet-1K dataset.

The overall structure of ConvNeXt is very similar to ResNet-50. It includes the feature extraction layer of the head, the

middle layer where the bottleneck structure of four different dimensions is separately stacked, and the final high-dimensional feature classification layer. However, the strategy of stacking and the interior of each layer has undergone several changes. The changes include: (i) In each stage of the original ResNet-50, the stacking number of each block is 3:4:6:3; in ConvNeXt this has been revised to 3:3:9:3, which is similar to the block stacking of the transformer model. (ii) In the block of ResNet-50, the bottleneck design is to reduce the dimension first, then feature extraction, and finally increase the dimension. However, as shown in Figure 5, the bottleneck in ConvNeXt is designed to run feature extraction first, then reduce the dimension, and finally increase the dimension. (iii) It has modified the size of the convolution kernel to 7*7 from the ResNet 3*3. (iv) Its activation function has also been replaced from ReLU to GELU, and cut back the usage count of activation functions. (v) Its normalization has changed to layer normalization from batch normalization as well as reduced usage count of normalization. The performance of ConvNeXt has gradually improved and even outperforms the VIT through the above five strategies and a few other settings including new parameters, structures, and functions.

Ensemble

As shown in the pipeline in Figure 6, we can obtain the final classification results by integrating the results of the Vision Transformer and ConvNeXt based on the soft voting mechanism using Equation (4):

$$S_f = \alpha S_v + (1 - \alpha)\, S_c \tag{4}$$

Where S_v and S_c denote the classification scores from Vision Transformer and ConvNeXt for all images, respectively.

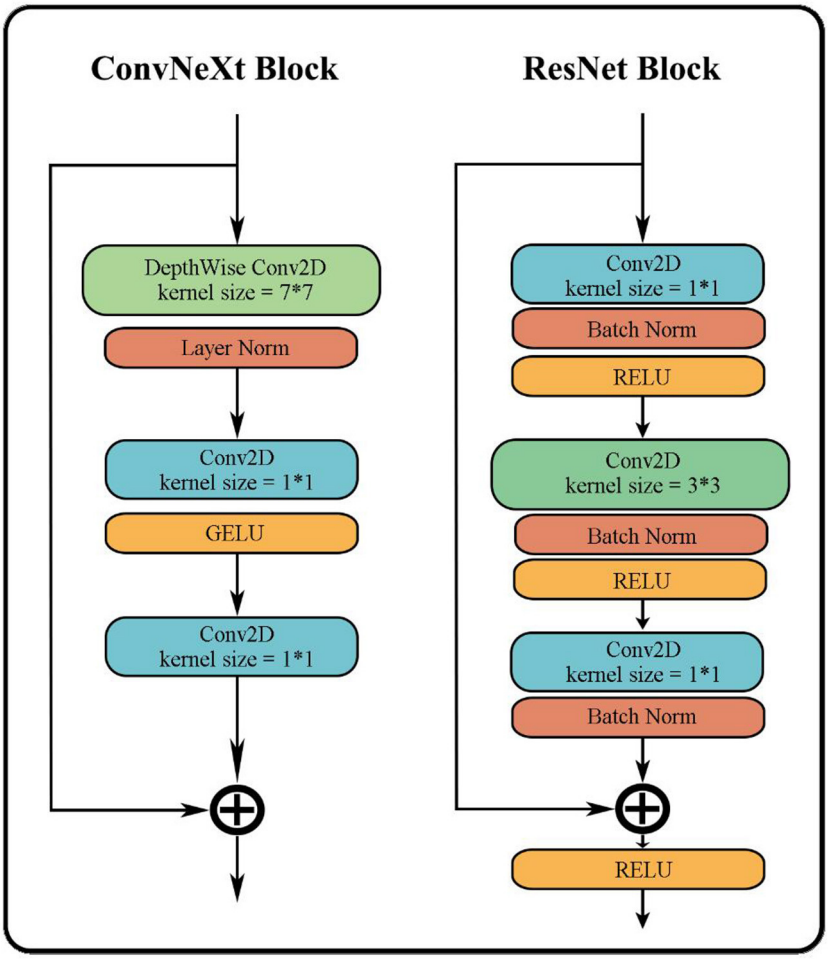

FIGURE 5
Differences between ConvNeXt and ResNet in bottleneck.

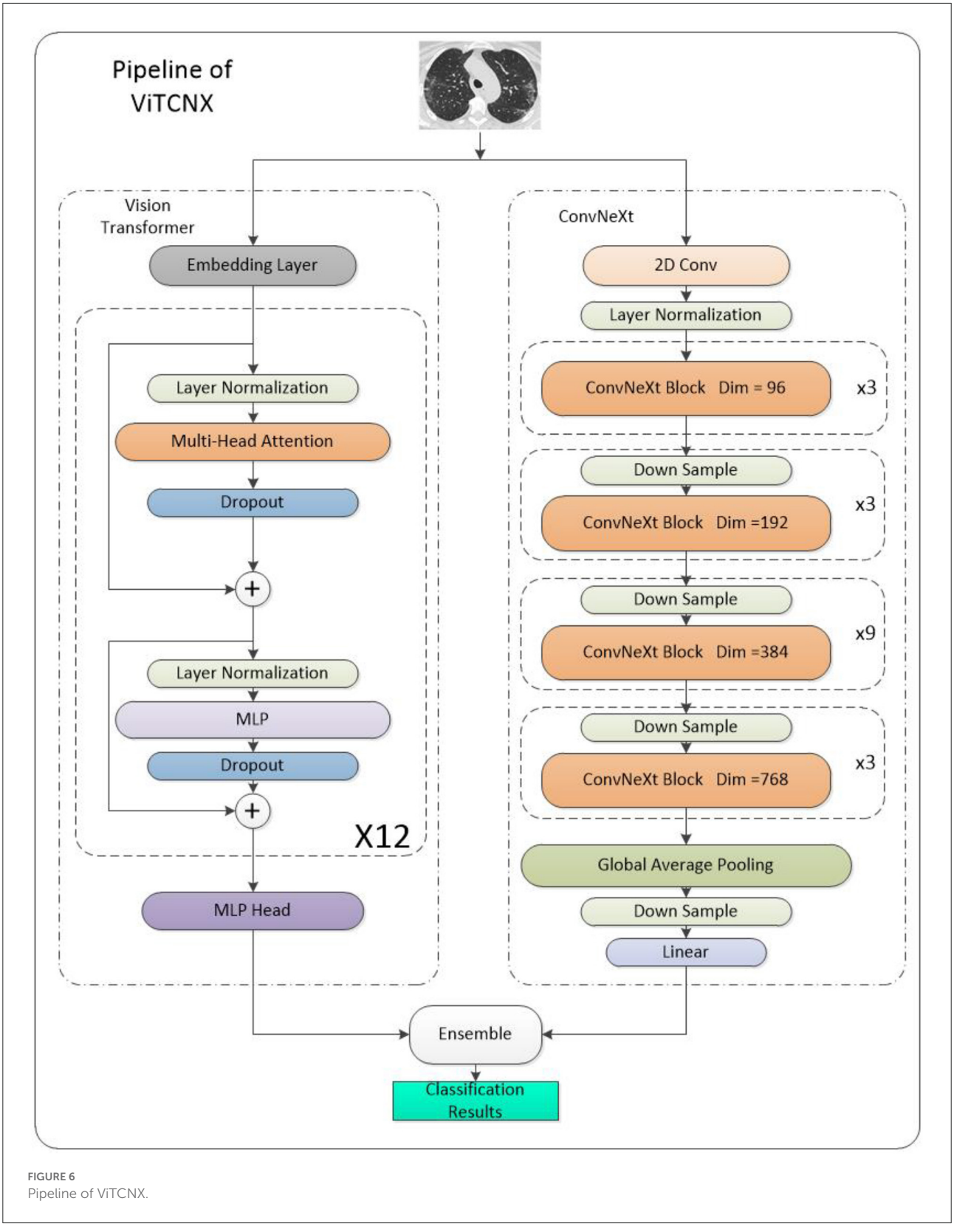

FIGURE 6
Pipeline of ViTCNX.

Results

Experimental evaluation and parameter settings

We used six metrics to evaluate the performance of all classification models, that is, precision, recall, accuracy, F1-score, AUC, and AUPR. These six evaluation metrics are defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{7}$$

$$\text{F1} - \text{Score} = \frac{2^{*}\text{Precision}^{*}\text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

$$\text{TPR (Ture Positive Rate)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

$$\text{FPR}\left(\text{False Positive Rate}\right) = \frac{\text{FP}}{\text{TN} + \text{FP}} \tag{10}$$

AUC is the area under the TPR-FPR curve. AUPR is the area under the precision-recall curve. For COVID-19-related image binary classification, precision means the proportion of images that are COVID-19-related images in the dataset and are predicted to be COVID-19-related among all the predicted COVID-19 images. Recall represents the proportion of images that are COVID-19-related images in the dataset and are predicted to be COVID-19-related among all COVID-19-related images in the dataset. Accuracy represents the proportion that is correctly predicted. F1-Score, AUC, and AUPR are comprehensive metrics that consider precision, recall, and FPR.

To investigate the performance of our proposed ViTCNX model in different classification situations, we conducted experiments under binary classification and three-class classification, respectively. In the ViTCNX, the dataset was randomly initialized with seed = 8. ConvNeXt uses ConvNeXt_tiny to construct and initialize parameters, and its initial learning rate was set to 5e-4, and the initial weight adopted the convnext_tiny_1k_224_ema. The Vision Transformer uses vit_base_patch16 to construct and initialize parameters, and its initial learning rate was set to 1e-3. It adopted the initial weight vit_base_patch16_224_in21k. In all image classification algorithms, the training epoch and the batch size were set to 100 and 8, respectively. DenseNet, ResNet-50, Swin Transformer, and EfficinetNetV2 used densenet121, resnet50-pre, swin_tiny_patch4_window7_224, and pre_efficientnetv2-s to initialize their weight parameters, respectively. The corresponding learning rates were 1e-3, 1e-4, 1e-4, and 1e-3, respectively. ViTCNX used the same parameter settings as individual Vision Transformer and ConvNeXt. After comparing the image classification ability under different values of α, we set $\alpha = 0.6$ where ViTCNX computed the best performance.

TABLE 1 Performance of ViTCNX and the other six models under the binary classification.

Metrics	Precision	Recall	Accuracy	F1-score	AUC	AUPR
EfficientNetV2	**0.9920**	0.3293	0.5875	0.4945	0.9609	0.9738
ConvNeXt	0.9650	0.9894	0.9715	0.9770	0.9952	0.9968
DenseNet	0.9788	0.9814	0.9756	0.9801	0.9973	0.9983
Swin Transformer	0.9587	0.9548	0.9471	0.9568	0.9911	0.9945
ResNet-50	0.9892	0.9695	0.9748	0.9792	0.9970	0.9979
Vision Transformer	0.9815	0.9854	0.9797	0.9834	0.9985	0.9990
ViTCNX	0.9803	**0.9907**	**0.9821**	**0.9855**	**0.9985**	**0.9991**

Bold values means the highest score under this metric.

Binary classification for CT images

Under the binary classification of images, there were a total of 6,151 CT images, including 3,768 CT images from COVID-19 patients and 2,383 CT images from healthy individuals. The 6,151 images were divided into a ratio of 0.8:0.2. Consequently, 4,922 images were used as the training set, including 3,015 COVID-19-related images and 1,907 CT images from healthy individuals. The remaining 1,229 images were used as the test set, including 753 COVID-19-related CT images and 476 healthy images. We compared our proposed ViTCNX model with four state-of-the-art image classification algorithms, that is, DenseNet (Huang et al., 2017), ResNet-50, Swin Transformer, and EfficinetNetV2 (Tan and Le, 2021). In addition, ViTCNX was also compared with the two individual models it was comprised of, that is, Vision Transformer and ConvNeXt. The results are shown in Table 1. The bold font in each column represents the best performance computed by the corresponding method among the above seven methods. Table 1 and Figure 7 show the precision, recall, accuracy, F1-score, AUC, and AUPR values and curves of these models.

From Table 1 and Figure 7, we can find that ViTCNX obtained the best recall, accuracy, F1-score, AUC, and AUPR, significantly outperforming the other six methods. EfficientNetV2 achieved the best score of precision. This result is consistent with the prediction results on the confusion matrix. In the experiments, EfficientNetV2 computed higher precision than ViTCNX. The reasons may be that different models perform very differently on different parameter settings, different datasets, and different sizes, which have a significant impact on the classification performance of the model. In particular, ViTCNX outperforms its two individual models, Vision Transformer and ConvNeXt, demonstrating that an ensemble of single classification models can improve image identification performance. Figures 7B,C show the AUC

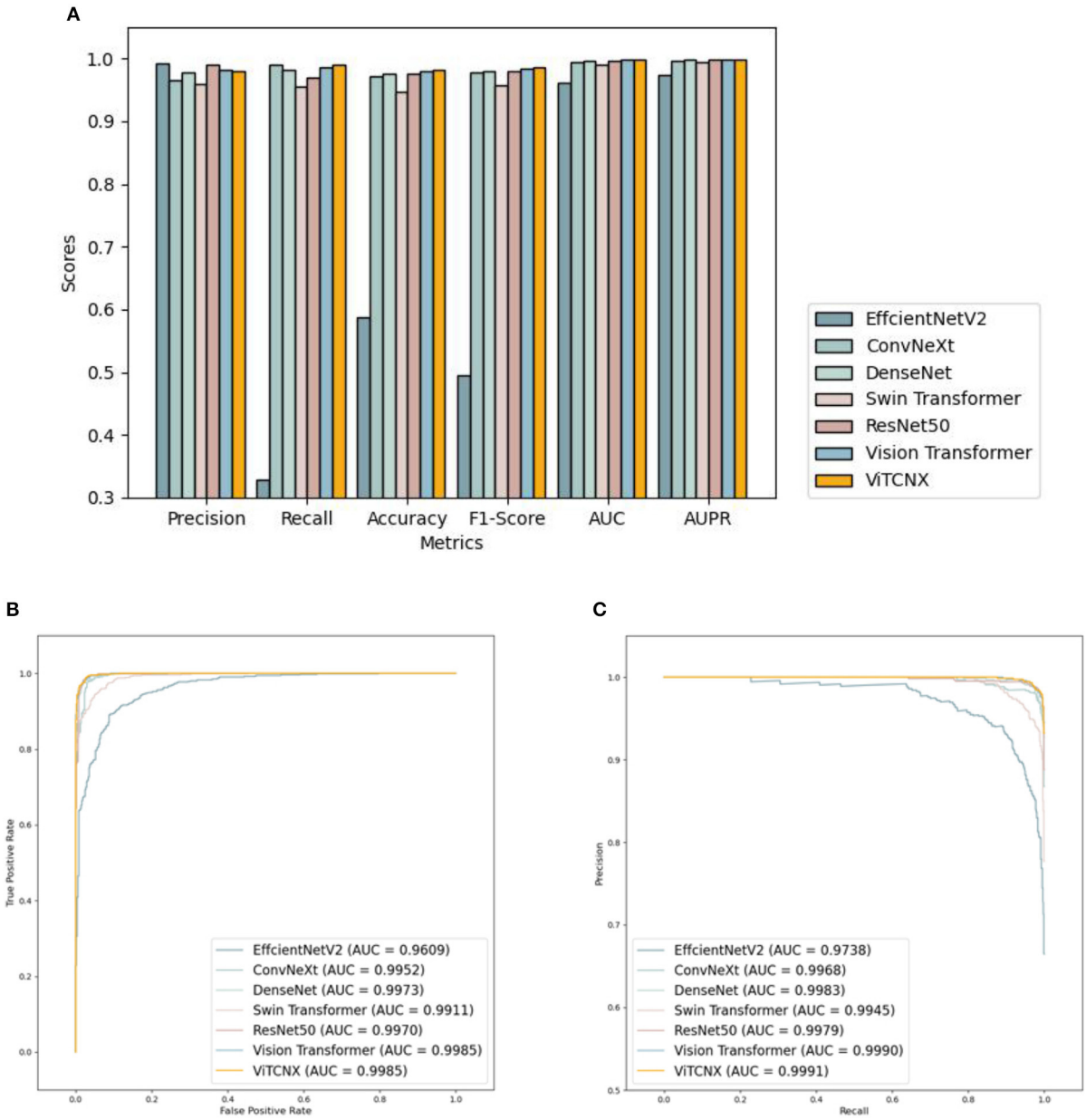

FIGURE 7
(A) The performance comparison of VitCNX and six other models for COVID-19 in binary classification problems; **(B,C)** The AUC and AUPR values of VitCNX and six other models for COVID-19 in binary classification problems.

and AUPR values obtained by the seven models. ViTCNX outperforms the other six models, elucidating that it can effectively classify related CT images as COVID-19-related or not.

Three-classification for CT images

To further investigate the performance of the seven models under the three-classification challenge, we considered a total of 7,398 CT images, including 3,768 images from COVID-19 patients, 2,383 from healthy individuals, and 1,247 from other pneumonia patients. The 7,398 images were divided in a ratio of 0.8:0.2, resulting in 5,920 images in the training set and 1,478 images in the test set. The 5,920 images in the training set consisted of 3,015, 1,907, and 998 images from COVID-19 patients, healthy individuals, and other pneumonia patients, respectively. The 1,478 images in the test set consisted of 753, 476, and 249 images from COVID-19 patients, healthy individuals, and other pneumonia patients, respectively. We

trained ViTCNX and the other comparable models using the training set and then evaluated their performance using the test set. Table 2 and Figure 8 show the precision, recall, accuracy, and F1-score values of ViTCNX and the other six models for the three-classification situation.

From Table 2 and Figure 8, we can observe that ViTCNX computed the best precision, accuracy, and F1-score, greatly outperforming the other six models. Although it calculated a relatively lower recall of 0.9597 than Vision Transformer with a recall of 0.9599, the difference is very minor. Particularly, compared with Vision Transformer, ConvNeXt, DenseNet, ResNet-50, Swin Transformer, and EfficientNetV2, ViTCNX computed a F1-score of 0.9631, better by 0.04, 1.58, 1.89, 5.32, 6.74, and 64.11% than the six models, respectively. These results demonstrate that ViTCNX can more accurately classify CT images from COVID-19, from other pneumonia cases, and healthy individuals.

The confusion matrix analysis

We further evaluated the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) obtained by Vision Transformer, ConvNeXt, DenseNet, ResNet-50, Swin Transformer, EfficientNetV2, and ViTCNX under binary classification. Table 3 and Figure 9 present the statistical data of TP, TN, FP, and FN from the above seven models for binary classification. The importance of these four evaluation metrics is not equal. For COVID-19 image recognition, TP denotes the number of images that are COVID-19 images in the dataset and are predicted to be COVID-19-related. FN denotes the number of images that are COVID-19 images but are predicted to be non-COVID-19-related. FN denotes that there are undetected COVID-19 patients, which may cause the spread of the pandemic. TP and FN are more important than the other two metrics. Higher TP and lower FN represent the better performance of ViTCNX.

TABLE 2 Performance of ViTCNX and the other six models under three classification.

Metrics	Precision	Recall	Accuracy	F1-Score
EfficientNetV2	0.7783	0.4188	0.4526	0.3221
ConvNeXt	0.9562	0.9397	0.9574	0.9473
DenseNet	0.9487	0.9402	0.9560	0.9442
Swin Transformer	0.9259	0.8754	0.9127	0.8957
ResNet-50	0.9369	0.8936	0.9317	0.9100
Vision Transformer	0.9657	**0.9599**	0.9689	0.9627
ViTCNX	**0.9668**	0.9597	**0.9696**	**0.9631**

Bold values means the highest score under this metric.

TABLE 3 Statistics of ViTCNX and other six models for binary classification.

Metrics	TP	TN	FP	FN
EfficientNetV2	248	**474**	**2**	505
ConvNeXt	745	449	27	8
DenseNet	739	460	16	14
Swin Transformer	719	445	31	34
ResNet-50	730	468	8	23
Vision Transformer	742	462	14	11
ViTCNX	**746**	461	15	**7**

Bold values means the highest score under this metric.

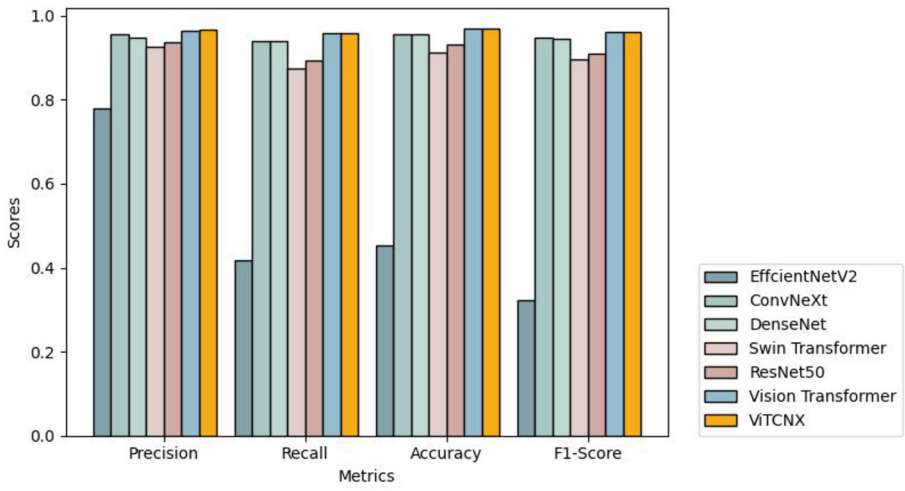

FIGURE 8
The performance of VitCNX and six other models for three-classification problem.

10.3389/fmicb.2022.1024104

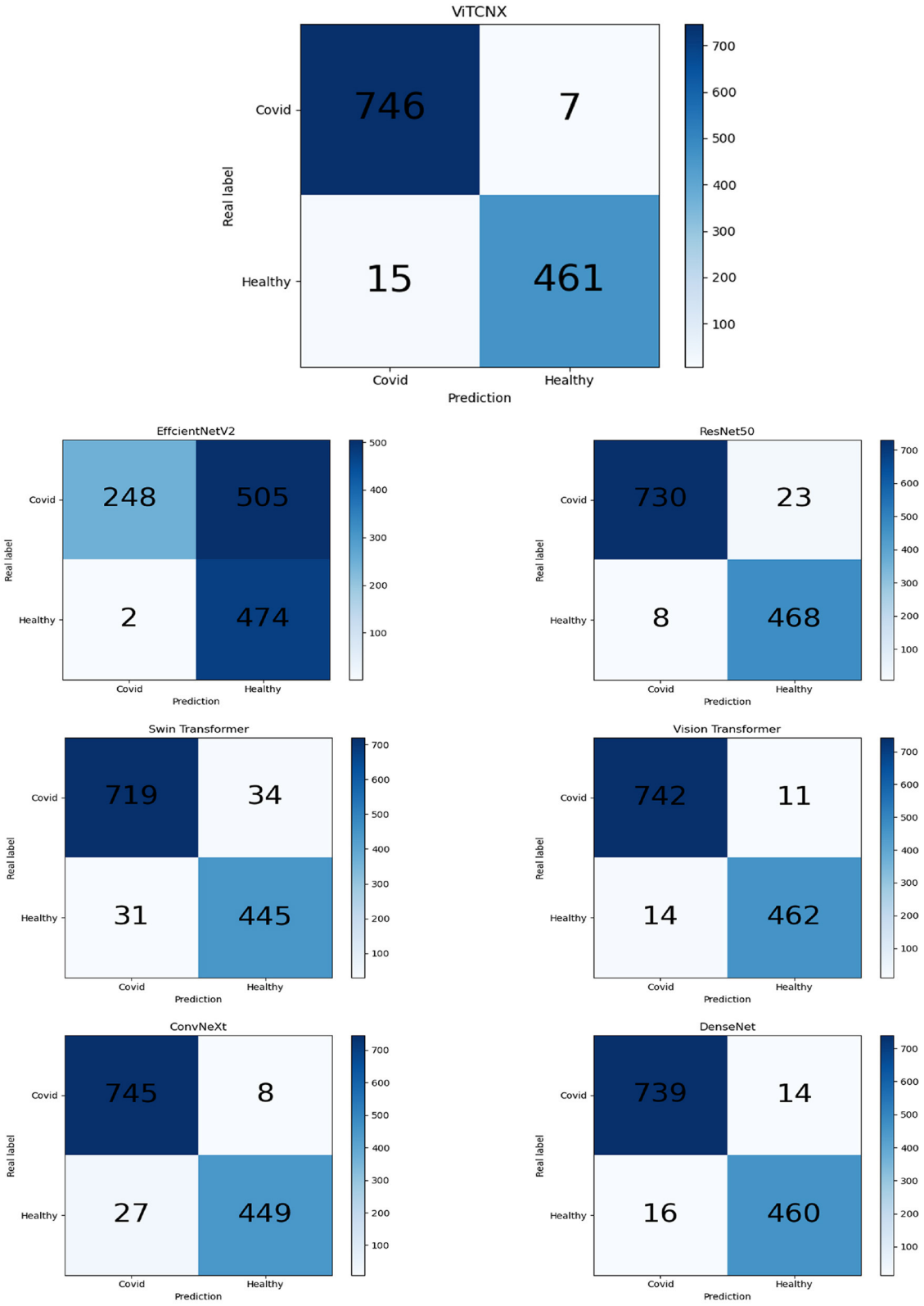

FIGURE 9
The confusion matrix of results of VitCNX and six other models.

From Table 3 and Figure 9, we can observe that our proposed ViTCNX model screens the most TP, and the least FN compared to the other six models. Our proposed ViTCNX model computes the highest TP of 746 and the lowest FN of 7 among 1,229 test samples, demonstrating that it can most efficiently recognize COVID-19-related images of COVID-19 patients.

Discussion and conclusion

With the rapid development of AI technology and high-performance computing platforms, using deep learning models to detect COVID-19 through lung CT images has become a research hotspot. Not only because this method has a higher performance and faster speed, but also lower time and economic cost. In this paper, we proposed an ensemble deep learning model (ViTCNX) to recognize COVID-19-related CT images by combining Vision Transformer and ConvNeXt. We compared ViTCNX with six other state-of-the-art deep learning models (Vision Transformer, ConvNeXt, DenseNet, ResNet-50, Swin Transformer, and EfficientNetV2). We conducted a series of comparative experiments to evaluate the performance of ViTCNX. The results show that ViTCNX computed the best recall, accuracy, F1-score, AUC, and AUPR under binary classification and the best precision, accuracy, and F1-score under three-classification tests. Moreover, ViTCNX obtained the highest TP and the lowest FN in binary classification. The results show that our proposed ViTCNX model has powerful COVID-19-related image recognition ability.

We adopted several techniques to reduce over-fitting. First, we used three different datasets of COVID-19 to evaluate the performance of ViTCNX. The three datasets were collected from two different places (Wuhan, China, and São Paulo, Brazil). We integrated the three different datasets into one dataset to increase the differences in datasets and further enhance the generalization performance of ViTCNX. Additionally, we used techniques including layer normalization and dropout to prevent over-fitting. The ensemble learning strategies also helped to improve the model's generalization ability and reduce over-fitting.

There are two advantages of the proposed ViTCNX model: First, the variance is reduced through the ensemble of multiple models, thereby improving the robustness and generalization ability of the model. Second, Vision Transformer and ConvNeXt are greatly different in structure. An ensemble of them can lower their correlation and further reduce the classification error. Although ViTCNX obtains better performance, it does increase a large number of training parameters, which increases the training and testing time of the model and requires higher computational resources.

In the future, we will continuously update data to build larger COVID-19 datasets to enhance the generalization ability of ViTCNX. We will also design a new deep learning framework, adopt efficient training methods, and optimize parameter settings to improve the prediction ability of the model. Additionally, we will establish an automatic annotation model to autonomously label hot spots. We anticipate that our proposed ViTCNX model can contribute to the clinical detection of COVID-19.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

Author contributions

GT, LP, ZW, XL, and ZL: conceptualization. GT, LP, ZW, CW, and ZL: methodology. ZW, GL, CW, ZH, YZ, and JC: software. GT, LP, ZW, YL, HX, ZH, YZ, and GL: validation. GT, LP, XL, JC, and ZL: investigation. ZW, CW, GL, and HX: data curation. LP and ZW: writing-original draft preparation and project administration. GT and LP: writing-review and editing. GT, LP, and ZL: supervision and funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding

ZL was supported by the National Natural Science Foundation of China under Grant No. 62172158. LP was supported by the National Natural Science Foundation of China under Grant No. 61803151. GL and YL were supported by the Innovation and Entrepreneurship Training Program for College Students of Hunan Province under Grant No. S202111535031 and the Innovation and Entrepreneurship Training Program for College Students of the Hunan University of Technology under Grant No. 20408610119.

Conflict of interest

Author GT was employed by the company Geneis (Beijing) Co., Ltd. Author JC was employed by Hunan Storm Information Technology Co., Ltd.

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

Publisher's note

References

Alshater, M.M., Atayah, O.F., and Khan, A. (2021). What do we know about business and economics research during COVID-19: a bibliometric review. *Econ. Res.* 35, 1–29. doi: 10.1080/1331677X.2021.1927786

Aslan, M.F., Unlersen, M.F., Sabanci, K., and Durdu, A. (2021). CNN-based transfer learning–BiLSTM network: A novel approach for COVID-19 infection detection. *Appl. Soft Comput.* 98, 106912. doi: 10.1016/j.asoc.2020.106912

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv [Preprint]*. arXiv: 1409.0473. Available online at: https://arxiv.org/pdf/1409.0473.pdf

Cascella, M., Rajnik, M., Aleem, A., Dulebohn, S.C., and Di Napoli, R. (2022). *Features, Evaluation, and Treatment of Coronavirus (COVID-19)*. Statpearls.

Chen, X., Xie, D., Zhao, Q., and You, Z.H. (2019). MicroRNAs and complex diseases: from experimental results to computational models. *Brief. Bioinform.* 20, 515–539. doi: 10.1093/bib/bbx130

Chung, M., Bernheim, A., Mei, X., Zhang, N., Huang, M., Zeng, X., et al. (2020). CT imaging features of 2019 novel coronavirus (2019-nCoV). *Radiology*. 295, 202–207. doi: 10.1148/radiol.2020200230

Codella, N.C., Nguyen, Q.B., Pankanti, S., Gutman, D.A., Helba, B., Halpern, A.C., et al. (2017). Deep learning ensembles for melanoma recognition in dermoscopy images. *IBM J. Res. Dev.* 61, 5. doi: 10.1147/JRD.2017.2708299

Del Rio, C., Omer, S.B., and Malani, P.N. (2022). Winter of Omicron—the evolving COVID-19 pandemic. *JAMA* 327, 319–320. doi: 10.1001/jama.2021.24315

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv [Preprint]*. arXiv: 2010.11929. Available online at: https://arxiv.org/pdf/2010.11929.pdf

Dou, Q., Chen, H., Yu, L., Qin, J., and Heng, P.A. (2016). Multilevel contextual 3-D CNNs for false positive reduction in pulmonary nodule detection. *IEEE Transact. Biomed. Eng.* 64, 1558–1567. doi: 10.1109/TBME.2016.2613502

Farooq, M., and Hafeez, A. (2020). Covid-resnet: a deep learning framework for screening of covid19 from radiographs. *arXiv [Preprint]*. arXiv: 2003.14395. Available online at: https://arxiv.org/pdf/2003.14395.pdf

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* 63, 139–144. doi: 10.1145/3422622

Guan, W.J., Ni, Z.Y., Hu, Y., Liang, W.H., Ou, C.Q., He, J.X., et al. (2020). Clinical characteristics of 2019 novel coronavirus infection in China. *MedRxiv [Preprint]*. doi: 10.1101/2020.02.06.20020974

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas), 770–778.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K.Q. (2017). "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu), 4700–4708.

Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* 60, 84–90. doi: 10.1145/3065386

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., et al. (1989). Handwritten digit recognition with a back-propagation network. *Adv. Neural Inf. Process. Syst.* 2, 396–404.

Liang, Y., Zhang, Z.Q., Liu, N.N., Wu, Y.N., Gu, C.L., and Wang, Y.L. (2022). MAGCNSE: predicting lncRNA-disease associations using multi-view attention graph convolutional network and stacking ensemble model. *BMC Bioinform.* 23, 1–22. doi: 10.1186/s12859-022-04715-w

Liu, H., Qiu, C., Wang, B., Bing, P., Tian, G., Zhang, X., et al. (2021a). Evaluating DNA methylation, gene expression, somatic mutation, and their combinations in inferring tumor tissue-of-origin. *Front. Cell Dev. Biol.* 9, 619330. doi: 10.3389/fcell.2021.619330

Liu, W., Jiang, Y., Peng, L., Sun, X., Gan, W., Zhao, Q., et al. (2022a). Inferring gene regulatory networks using the improved Markov blanket discovery algorithm. *Interdiscipl. Sci. Comp. Life Sci.* 14, 168–181. doi: 10.1007/s12539-021-00478-9

Liu, W., Lin, H., Huang, L., Peng, L., Tang, T., Zhao, Q., et al. (2022b). Identification of miRNA–disease associations via deep forest ensemble learning based on autoencoder. *Brief. Bioinformat.* 23, bbac104. doi: 10.1093/bib/bbac104

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021b). "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer* Vision, 10012–10022.

Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022c). "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (New Orleans), 11976–11986.

Munir, K., Elahi, H., Ayub, A., Frezza, F., and Rizzi, A. (2019). Cancer diagnosis using deep learning: a bibliographic review. *Cancers* 11, 235. doi: 10.3390/cancers11091235

Padhan, R., and Prabheesh, K.P. (2021). The economics of COVID-19 pandemic: a survey. *Econ. Anal. Policy* 70, 220–237. doi: 10.1016/j.eap.2021.02.012

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* 115, 211–252. doi: 10.1007/s11263-015-0816-y

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv [Preprint]*. arXiv: 1409.1556. Available online at: https://arxiv.org/pdf/1409.1556.pdf

Soares, E., Angelov, P., Biaso, S., Froes, M. H., and Abe, D. K. (2020). SARSCoV-2 CT-scan dataset: a large dataset of real patients CT scans for SARS-CoV-2 identification. *MedRxiv [Preprint]*. doi: 10.1101/2020.04.24.20078584

Sohail, A., Khan, A., Wahab, N., Zameer, A., and Khan, S. (2021). A multi-phase deep CNN based mitosis detection framework for breast cancer histopathological images. *Sci. Rep.* 11, 1–18. doi: 10.1038/s41598-021-85652-1

Stadler, K., Masignani, V., Eickmann, M., Becker, S., Abrignani, S., Klenk, H.D., et al. (2003). SARS—beginning to understand a new virus. *Nat. Rev. Microbiol.* 1, 209–218. doi: 10.1038/nrmicro775

Sun, F., Sun, J., and Zhao, Q. (2022). A deep learning method for predicting metabolite–disease associations via graph neural network. *Brief. Bioinform.* 23, bbac266. doi: 10.1093/bib/bbac266

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA), 1–9.

Tahamtan, A., and Ardebili, A. (2020). Real-time RT-PCR in COVID-19 detection: issues affecting the results. *Expert Rev. Mol. Diagn.* 20, 453–454. doi: 10.1080/14737159.2020.1757437

Tan, M., and Le, Q. (2021). "Efficientnetv2: smaller models and faster training," in *International Conference on Machine Learning* (PMLR), 10096–10106.

Tang, X., Cai, L., Meng, Y., Xu, J., Lu, C., and Yang, J. (2021). Indicator regularized non-negative matrix factorization method-based drug repurposing for COVID-19. *Front. Immunol.* 11, 603615. doi: 10.3389/fimmu.2020.603615

Vasireddy, D., Vanaparthy, R., Mohan, G., Malayala, S.V., and Atluri, P. (2021). Review of COVID-19 variants and COVID-19 vaccine efficacy: what the clinician should know?. *J. Clin. Med. Res.* 13, 317. doi: 10.14740/jocmr4518

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., et al. (2017). Attention is all you need. *Adv. Neural Inf. Process. Syst.* 30.

V'kovski, P., Kratzel, A., Steiner, S., Stalder, H., and Thiel, V. (2021). Coronavirus biology and replication: implications for SARS-CoV-2. *Nat. Rev. Microbiol.* 19, 155–170. doi: 10.1038/s41579-020-00468-6

Wang, C.C., Han, C.D., Zhao, Q., and Chen, X. (2021). Circular RNAs and complex diseases: from experimental results to computational models. *Brief. Bioinform.* 22, bbab286. doi: 10.1093/bib/bbab286

World Health Organization (2020). *WHO COVID-19 Dashboard*. Geneva: World Health Organization. Available online at: covid19.who.int (accessed April 25, 2022).

Xu, X., Yu, C., Qu, J., Zhang, L., Jiang, S., Huang, D., et al. (2020). Imaging and clinical features of patients with 2019 novel coronavirus SARS-CoV-2. *Eur. J. Nucl. Med. Mol. Imaging* 47, 1275–1280. doi: 10.1007/s00259-020-04735-9

Yang, J., Ju, J., Guo, L., Ji, B., Shi, S., Yang, Z., et al. (2022). Prediction of HER2-positive breast cancer recurrence and metastasis risk from histopathological images and clinical information via multimodal deep learning. *Comput. Struct. Biotechnol. J.* 20, 333–342. doi: 10.1016/j.csbj.2021.12.028

Yang, X., He, X., Zhao, J., Zhang, Y., Zhang, S., and Xie, P. (2020). COVID-CT dataset: a CT scan dataset about COVID-19. *arXiv [Preprint]*. arXiv: 2003.13865. Available online at: https://arxiv.org/pdf/2003.13865.pdf

Yu, F., Lau, L.T., Fok, M., Lau, J.Y.N., and Zhang, K. (2021). COVID-19 Delta variants—Current status and implications as of August 2021. *Precis. Clin. Med.* 4, 287–292. doi: 10.1093/pcmedi/pbab024

Zhan, H., Schartz, K., Zygmont, M.E., Johnson, J.O., and Krupinski, E.A. (2021). The impact of fatigue on complex CT case interpretation by radiology residents. *Acad. Radiol.* 28, 424–432. doi: 10.1016/j.acra.2020.06.005

Zhang, K., Liu, X., Shen, J., Li, Z., Sang, Y., Wu, X., et al. (2020). Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. *Cell* 181, 1423–1433. doi: 10.1016/j.cell.2020.04.045

Zhang, L., Yang, P., Feng, H., Zhao, Q., and Liu, H. (2021). Using network distance analysis to predict lncRNA–miRNA interactions. *Interdiscipl. Sci.* 13, 535–545. doi: 10.1007/s12539-021-00458-z

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 17 November 2022
DOI 10.3389/fmicb.2022.1040252

OPEN ACCESS

EDITED BY
Qi Zhao,
University of Science and Technology Liaoning, China

REVIEWED BY
Yi Xiong,
Shanghai Jiao Tong University, China
Li Zhang,
China University of Mining and Technology, China

*CORRESPONDENCE
Lieqing Lin
tiger@gdut.edu.cn

SPECIALTY SECTION
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

RECEIVED 09 September 2022
ACCEPTED 06 October 2022
PUBLISHED 17 November 2022

CITATION
Chen L, Lin D, Xu H, Li J and Lin L (2022) WLLP: A weighted reconstruction-based linear label propagation algorithm for predicting potential therapeutic agents for COVID-19.
Front. Microbiol. 13:1040252.
doi: 10.3389/fmicb.2022.1040252

WLLP: A weighted reconstruction-based linear label propagation algorithm for predicting potential therapeutic agents for COVID-19

Langcheng Chen[1], Dongying Lin[2], Haojie Xu[2], Jianming Li[2] and Lieqing Lin[1]*

[1]Center of Campus Network and Modern Educational Technology, Guangdong University of Technology, Guangzhou, China, [2]School of Computer Science, Guangdong University of Technology, Guangzhou, China

The global coronavirus disease 2019 (COVID-19) pandemic caused by the severe acute respiratory syndrome coronavirus-2 (SARS-CoV) has led to a huge health and economic crises. However, the research required to develop new drugs and vaccines is very expensive in terms of labor, money, and time. Owing to recent advances in data science, drug-repositioning technologies have become one of the most promising strategies available for developing effective treatment options. Using the previously reported human drug virus database (HDVD), we proposed a model to predict possible drug regimens based on a weighted reconstruction-based linear label propagation algorithm (WLLP). For the drug–virus association matrix, we used the weighted *K*-nearest known neighbors method for preprocessing and label propagation of the network based on the linear neighborhood similarity of drugs and viruses to obtain the final prediction results. In the framework of 10 times 10-fold cross-validated area under the receiver operating characteristic (ROC) curve (AUC), WLLP exhibited excellent performance with an AUC of 0.8828 ± 0.0037 and an area under the precision-recall curve of 0.5277 ± 0.0053, outperforming the other four models used for comparison. We also predicted effective drug regimens against SARS-CoV-2, and this case study showed that WLLP can be used to suggest potential drugs for the treatment of COVID-19.

KEYWORDS

COVID-19, drug repositioning, linear neighborhood similarity, label propagation, WKNKN

1. Introduction

In November 2019, a novel coronavirus disease broke out in Wuhan, China, for unknown reasons, which was named coronavirus disease 2019 (COVID-19) by the World Health Organization (WHO) (Zhu et al., 2020). COVID-19 is caused by severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2). To date, seven human coronaviruses (HCoV) have been identified, namely HCoV-229E, HCoV-OC43,

HCoV-NL63, HCoV-HKU1, SARS-CoV, Middle East respiratory syndrome (MERS) coronavirus (MERS-CoV), and SARS-CoV-2. Specifically, HCoV-229E, HCoV-OC43, HCoV-NL63, and HCoV-HKU1 are frequently found and have low pathogenicity, generally causing only common cold symptoms, whereas MERS-CoV and SARS-CoV are zoonotic viruses that are first reported in the twenty-first century (Sohrabi et al., 2020). SARS-CoV-2 is recognized as the most pathogenic human coronavirus ever discovered (Guan et al., 2020). As of September 2022, 613 million confirmed SARS-CoV-2 infections were reported around the world, with nearly 6 million deaths (Organization, 2020). Until now, there is no cure for COVID-19.

Despite substantial increases in investment by pharmaceutical companies in response to COVID-19, the successful development and approval of a new drug typically requires billions of dollars and an average of 10 years (Liu S. et al., 2020), with the disadvantages of being time consuming (Pushpakom et al., 2019), expensive, and risky. Therefore, drug repositioning (drug repurposing) has been identified as a viable solution to improve the overall process of drug development, especially following recent advances in information technology and data science. The primary goal of drug repositioning is the use of existing drugs to treat new symptoms. Compared with traditional drug development methods, drug repositioning can significantly reduce research and development time and costs while minimizing risks. In short, drug repositioning is considered a promising strategy to accelerate the development of COVID-19 therapeutics.

As Xue et al. (2018) described, current work on drug repositioning is supported by various prediction models, among which the association prediction models for computational drug repositioning applicable to COVID-19 can be broadly classified into the following three categories (Dotolo et al., 2021): (I) network-based models, (II) artificial intelligence algorithms, and (III) matrix completion.

Network-based approaches construct heterogeneous networks by integrating multiple data to predict drug–virus associations. Such approaches are mostly based on the assumption that drugs with similar functions are often associated with viruses having similar phenotypes (Chen et al., 2018b). Prediction approaches based on complex networks (Liu et al., 2022b) have important and widespread applications in drug repositioning because of their ability to integrate multiple datasets of interest (Fan et al., 2020; Zhou et al., 2020). More specifically, network nodes represent drugs, diseases, viruses, or genes, while edges represent interactions or relationships between nodes (Re and Valentini, 2013; Chen et al., 2015). The obtained predictions may contribute to the process of structure-directed drug and diagnostic research and help to identify new potential biological targets (Barlow et al., 2020). In this regard, there are two network-based approaches applicable to drug repositioning for COVID-19: the network-based clustering approach and the network-based propagation approach. Macropol et al. (2009) proposed the repeated random walks (RRW) method that uses RRW on the protein–protein interaction (PPI) network for local clustering of the network and then predicts some protein complexes. Although this was found to be a precise and general approach, it requires a great deal of time and memory overhead and cannot detect overlapping clusters. King et al. (2012) introduced a new model named restricted neighborhood search clustering (RNSC), which is a global network algorithm for identifying protein clusters on PPI networks. It considers both global and local information from the network and can also detect overlapping clusters, but some information may be lost if the cluster size is too small. Luo et al. (2016) proposed the bidirectional random walk (BiRW) algorithm for predicting relationships between diseases and drugs. It uses the similarity of diseases and drugs with the original correlation matrix to form a heterogeneous network and then clusters this network by a double random walk. The resulting prediction is accurate, but more biological information is needed to improve the confidence of the similarity metric. In addition to the network clustering approach, Vanunu et al. (2010) proposed an overall propagation algorithm called PRINCE, which combines weighted PPI and disease similarity networks for overall disease gene ranking and protein complex association inference. An integrated propagation method for predicting propagation strategies in different sub-networks was proposed by Martinez et al. (2015) and named DrugNet. Zhang et al. (2017b) developed the linear neighborhood similarity (LNS) method to calculate drug–drug similarities in the drug characteristic space. Peng et al. (2021), in response to COVID-19, combined the virus–drug association network topology and a random walk with restart method (VDA-RWR) to predict potential drug–virus associations using a 2×2 similarity matrix and known associations between drugs and viruses. Zhang et al. (2021b) developed a network distance analysis model for the prediction of lncRNA–miRNA association (NDALMA). It is worth mentioning that the primary approach in recent years has been to update the network mainly by similarity and network inference (Zhang et al., 2021a; Liu et al., 2022a).

For drug repositioning, artificial intelligence-based models mainly use machine learning methods. Numerous common machine learning algorithms have been applied to predict potential therapeutic agents, such as decision trees (Chen et al., 2019b) and Laplacian regularization (Chen and Huang, 2017). The influence of deep learning models that belong to machine learning has been particularly remarkable (Chen et al., 2019a, 2021a; Keshavarzi Arshadi et al., 2020). In terms of prediction, graph convolutional neural networks (GCNNs) are the most popular tools for drug discovery applications because they can process graphs and extract features by encoding adjacency information within features to learn representations from molecules. Based on drug-target interactions in this model, Torng and Altman (2019) made correlation predictions. In recent years, sequence-based models,

such as genomics, proteomics, and transcriptomics, have also attracted considerable attention. Vaswani et al. (2017) and Devlin et al. (2018) advanced a transformer model for extracting features from sequences through the attention mechanism and self-supervision, which are widely used in the field of natural language processing. Moreover, Shin et al. (2019) demonstrated that drug-target interactions can be predicted by using the transformer model. Pollastri et al. (2002) demonstrated that recurrent neural networks (RNNs) and long short-term memory (LSTM) networks can predict the secondary structure of molecular or protein sequences. Through an ensemble strategy of three mainstream machine learning algorithms, Hu et al. (2018) proposed a model named HLPI-Ensemble that was specifically designed for human lncRNA–protein interactions. Matrix completion mainly relies on the matrix decomposition algorithms (Chen et al., 2018a,c). Specifically, these algorithms decompose a matrix into two lower-order potential factor matrices based on known association matrices of diseases and drugs (Liu H. et al., 2020). Gönen (Gönen, 2012) put forward a prediction method by using Bayesian probabilistic matrix factorization (BPMF) based on chemical and nuclear genomes. Yang et al. (2019) developed a model based on bounded nuclear norm regularization for drug repositioning. Considering the similarity information between drugs and diseases, Meng et al. (2021) proposed a method called similarity-constrained PMF (SCPMF) to examine the potential value of existing drugs. Liu et al. (2022b) proposed a new computational method *via* deep forest ensemble learning based on an autoencoder (DFELMDA) to predict miRNA–disease associations.

The novel similarity measure of LNS proposed by Zhang et al. has been successfully applied to several bioinformatics problems (Zhang et al., 2017a, 2018a). In this method, the data points are reconstructed by linear neighborhood information and are used to measure the similarity between two points in the association network. Inspired by this, we applied this similarity measure to our model. In recent years, label propagation has been widely used for biological association prediction owing to its various advantages, such as simple logic algorithm and fast optimization. Thus, we adopted the label propagation method for network propagation of the drug–virus association matrix.

Herein, we reported on the development of a method termed label propagation through linear neighborhood similarity for the prediction of undetected drug–virus associations. More specifically, we represented drugs or viruses as feature vectors and treated them as data points in the feature space, from which we computed pairwise linear neighborhood similarities between drugs and drugs or between viruses and viruses. The computed drug and virus similarities and the known disease–virus association networks were treated as a weighted directed graph, which was then input to the label propagation algorithm. Each drug–virus interaction was scored using the label propagation method. Experiments showed that the WLLP model offered superior prediction results when compared with other models, with an area under the receiver operating characteristic (ROC) curve (AUC) of 0.8828 in the framework of 10 times 10-fold cross-validated.

2. Materials and methods

2.1. Experimental data

2.1.1. Human drug virus database

The collection of data concerning viruses, drugs, and drug–virus associations is a crucial precursor to using bioinformatics methods to predict novel drug–virus associations. Moreover, systematic collection and management of relevant information are important for further studying the mechanism of virus action (Wang et al., 2021). Meng et al. (2021) collected a large number of experimentally validated drug–virus interaction entries from the literature by using text mining techniques and then constructed the HDVD, which is a database of human drug–virus associations. The HDVD includes 34 viruses, 219 drugs, and 455 confirmed human drug–virus interactions.

2.1.2. Construction of the drug–virus interaction network

From the HDVD dataset, we constructed an association network using known drug–virus interactions, where the points represent the drugs and viruses and the edges represent drug–virus associations. Let $G = (D, V, I)$ represents the drug–virus association network, where $D = \{d_1, d_2, \ldots, d_n\}$ represents the known drugs in the dataset, $V = \{v_1, v_2, \ldots, v_m\}$ represents the known viruses in the dataset, and I represents the interaction relationship between D and V. Let $A_{n\times m}$ represents the adjacency matrix of graph G. If d_i and v_j are related, $A_{ij} = 1$; otherwise, $A_{ij} = 0$. Also, let A^T represent the inversion of $A_{n\times m}$.

2.1.3. Chemical structure similarity of drug pairs

The chemical structure similarity between two drugs can be calculated from their molecular structure information. In the current study, we downloaded the chemical structure information of drugs from the DrugBank database in the SMILES format (Öztürk et al., 2016) and then calculated their molecular access system (MACCS) fingerprints (O'Boyle et al., 2011). Finally, we used the Tanimoto index to measure the absolute similarity between two molecules (Bajusz et al., 2015). Specifically, we set two drug molecules as A and B, respectively, a is the number of bits in molecule A, and b is number of bits in molecule B. c is the number of bits that are in both molecules. The formula is as follows:

$$T = c/(a + b - c) \quad (1)$$

We used this formula to construct the drug chemical structure similarity matrix $DD_{n\times n}$. This is a two-dimensional

matrix whose values represent the chemical fingerprint scores between drugs. In general, the size of this score is between 0 and 1, with larger values representing greater drug–drug similarity.

2.1.4. Genomic sequence similarity of virus pairs

The sequence similarity between viruses can be calculated from their genomic nucleotide sequences. We downloaded the

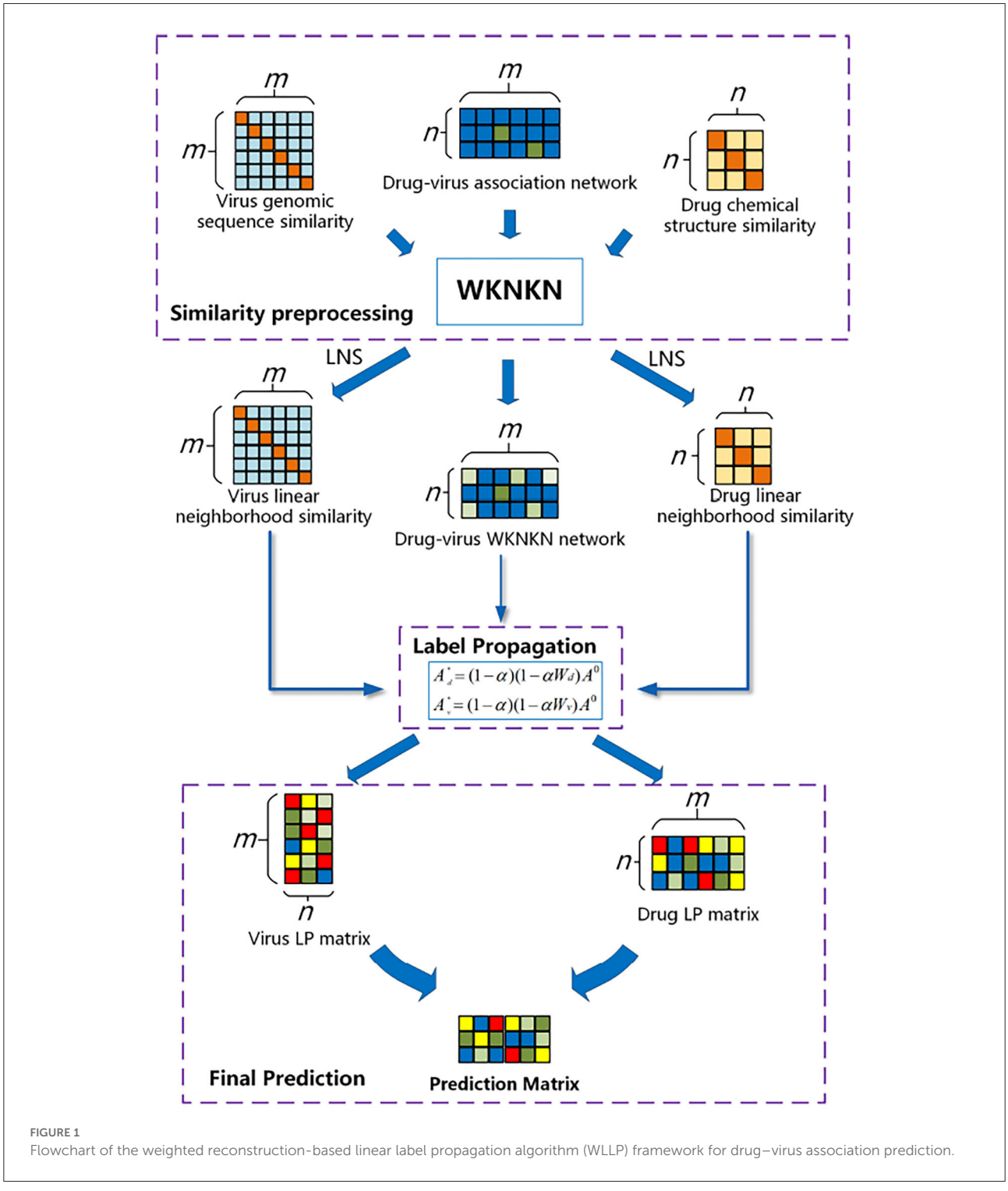

FIGURE 1
Flowchart of the weighted reconstruction-based linear label propagation algorithm (WLLP) framework for drug–virus association prediction.

genomic nucleotide sequences of viruses from the National Center for Biotechnology Information (Wheeler et al., 2002). To calculate the sequence similarity, we used the multiple sequence alignment program MAFFT on account of its high performance (Katoh and Standley, 2013). Finally, the virus sequence similarity matrix $VV_{m\times m}$ was constructed, which is a two-dimensional matrix whose values represent the sequence similarity between viruses. In general, the value of this matrix is between 0 and 1, and larger values represent greater virus–virus similarity.

2.2. Methods

2.2.1. Overview of WLLP

In this study, we developed the WLLP framework for predicting disease–virus associations based on LNS in conjunction with label propagation. As shown in Figure 1, the framework consists of three main steps: (I) Label set preprocessing: considering the sparse nature of the drug–virus interaction matrix, we introduced the weighted K-nearest known neighbors (WKNKN) algorithm to make a correction for the potential interactions between the drugs and viruses. (II) LNS information for the drugs and viruses was mined separately based on drug–virus interaction information. (III) Label propagation: a weighted directed graph consisting of known association information, drug–drug LNS, and virus–virus LNS matrices was constructed, and the drug label information was iteratively updated by the label propagation algorithm to reveal unknown potential drug–virus associations.

The flowchart of the WLLP algorithm is shown in Algorithm 1. The details of the principle and process of each WLLP module are described in the following sections.

Input: Matrices $A_{n\times m}$, $DD_{n\times n}$, and $VV_{m\times m}$; Number of neighbors K and decay factor r; LNS size parameters dN and DN; Probability retention factors for drugs and viruses, α and β, respectively;
Output: Predictive association matrix $A^*_{m\times n}$
Step 1: Reconstruct the association matrix
for $i = 1$ to n **do**
  construct A_d using Equation (2)
for $i = 1$ to n **do**
  construct A_v using Equation (3)
$A_{dv} = \left(A_d + A_v^{\mathrm{T}}\right)/2$
$A = max(A, A_{dv})$
Step 2: Construct the drug LNS matrix W_d
for $i = 1$ to DN **do**
  construct refactoring weight w_i for each drug using Equation (5)
Step 3: Similarly, construct the virus LNS matrix W_v
Step 4: Update the associated network by label propagation
Predict new association matrix A_d^* using Equation (8) and weight W_d
Similarly, predict the new association matrix A_v^*
$A^* = \left(A_d^* + A_v^{*\ \mathrm{T}}\right)/2$
return A^*

Algorithm 1. WLLP.

2.2.2. WKNKN

Because it is hard to construct expression datasets, coming up with datasets that contain a large number of samples is generally difficult. A small number of samples complicates the knowledge discovery task (Sirin et al., 2016). The unknown nature of a large part of the information makes the drug–virus association matrix very sparse. Here, we used the WKNKN algorithm to preprocess the original association matrix (Ezzat et al., 2016). Specifically, WKNKN replaces A_{0ij}=0 with the interaction likelihood in the following three steps:

Step 1. For each known drug, the chemical structure similarities of the closest K known drugs are calculated by the k-nearest neighbors (KNN) method and their corresponding interaction profiles are used to estimate the interaction likelihood profiles. The derived formula is

$$A_d(i,:) = \frac{\Sigma_{k=1}^{K} T^{k-1} DD(i, D_k) A(D_k, :)}{\Sigma_{k=1}^{K} DD(i, D_k)}, \tag{2}$$

where i denotes the drug index, T is the decay factor, and in general, $T \leq 1$. D_k denotes the k-th drug index that is most similar to drug i. It is worthwhile to mention that the denominator part is the normalization term.

Step 2. For each known virus, the sequence similarities of the closest K known viruses are calculated by the KNN method and their corresponding interaction profiles are used to estimate the interaction likelihood profiles:

$$A_v(:,j) = \frac{\Sigma_{k=1}^{K} T^{k-1} VV(j, V_k) A^T(V_k, :)}{\Sigma_{k=1}^{K} VV(j, V_k)}, \tag{3}$$

where j denotes the virus index, T is the decay factor, and in general, $T \leq 1$. V_k denotes the k-th virus index that is most similar to virus j. Similarly, the denominator part is the normalization term.

Step 3. If $A_{ij} = 0$, then we average the interaction likelihood values calculated by Equations (2) and (3) and replace the original values. Using WKNKN, we finally calculate a weighted nearest neighbor interaction spectrum, which we will substitute into the prediction model later.

2.2.3. LNS

Previous studies have demonstrated that each data point can be perfectly reconstructed with linear neighborhood information (Wang and Zhang, 2006; Chen et al., 2021b). Based on these studies, we used the known drug–virus interactions to update the degree of drug–virus similarity. Inspired by Zhang et al. (2018b), we established linear neighborhood similarity. In the following, we analyzed the drugs as an example. We take the association matrix of drugs as $X = \{x_1, x_2, \ldots, x_n\}$, and each vector x_i is reconstructed from a linear combination of its neighboring data points. The objective function is to minimize the reconstruction loss with the following expression:

$$\min_{w_i} L_i = \left| x_i - \sum_{i_j : x_{i_j} \in N(x_i)} w_{i,i_j} x_{i_j} \right|^2 = \omega_i^T G^i \omega_i$$
$$s.t. \sum_{i_j; x_{i_j} \in N(x_i)} \omega_{i,i_j} x_{i_j} = 1, \omega_i \geq 0, \mathrm{j} = 1, \ldots, \mathrm{DN}, \tag{4}$$

where $N(x_i)$ denotes the set of DN nearest neighbors and $DN(0 < DN < n)$ is a conditioning parameter that indicates the number of neighbors. x_{i_j} denotes the j-th neighbor of the vector x_i. $w_i = \{w_{i,i_1}, w_{i,i_2}, \ldots, w_{i,i_{DN}}\}$is a vector whose size is $DN \times 1$ representing the weight size of the k nearest neighbors of x_i and also indicates the similarity between x_{i_j} and x_i. G_i denotes the gram matrix whose size is $DN \times DN$, where $G^i_{i_p,i_q} = (x_i - x_{i_p})(x_i - x_{i_q})^T$. To prevent overfitting, we incorporated the Tikhonov regularization term, which makes the minimization reconstruction loss normalized. The formula is as follows:

$$\min_{w_i} L_i = \omega_i^{\mathrm{T}} G^i \omega_i + \mu ||\omega_i||_1^2 = \omega_i^{\mathrm{T}} \left(G^i + \mu I \right) \omega_i,$$
$$s.t. \sum_{i_j : x_{i_j} \in N(x_i)} \omega_{i,i_j} x_{i_j} = 1, \omega_i \geq 0, j = 1, \ldots, \mathrm{DN}, \tag{5}$$

where μ is the regularization factor. For simplicity, we set μ to 1. Finally, we used the standard quadratic programming method to solve the objective function, and the result can be regarded as the reconstruction weight of each data point x_i. We thus obtained two weight matrices, $W_d \in R^{n \times n}$ and $W_v \in R^{m \times m}$, which were the LNS matrices for the drugs and viruses, respectively.

2.2.4. Label propagation

From the previous calculation steps, we finally obtained three matrices: the drug–virus association matrix $A_{n \times m}$ after WKNKN processing, the drug–drug LNS matrix W_d, and the virus–virus LNS matrix W_v. In the following, as a representative example, we considered the drug–drug LNS matrix as a directed weighted graph, with drugs as the nodes and the degree of similarity as the weights of the lines. It is worth noting that the similarity matrix is not diagonally symmetric, i.e., $w_{ij} \neq w_{ji}$. Based on this, we used a label propagation approach to circularly and iteratively propagate the label information of the drugs to reveal potential drug–virus associations. On the association network, the neighboring edge information of each drug was computed and updated at each label propagation. Meanwhile, we set a probability parameter α to retain its updated state and retain its initial state with a probability of $1 - \alpha$. The specific updated equation is as follows:

$$A_j^{t+1} = \alpha W_d A_j^t + (1 - \alpha) A_j^0 \tag{6}$$

where, for the exact virus v_j, A_j^0 denotes all known original drug interaction relationships and A_j^t denotes the predicted label at iteration t. For all viruses, we expressed the prediction matrix as $A^t = \{A_1^t, A_2^t, \ldots, A_m^t\}$ and represented the equation further by the following matrix form:

$$A^{t+1} = \alpha W_d A^t + (1 - \alpha).A^0 \tag{7}$$

As t tends to infinity, the expression converges to the following form:

$$A^* = (1 - \alpha) \left(I - \alpha W_d \right)^{-1} A^0 \tag{8}$$

where $I \in R^{n \times n}$is the identity matrix and A^* is the association score matrix. For more details on the convergence analysis of label propagation, please refer to the analysis (Wang and Zhang, 2006).

3. Results

3.1. Experimental setting

In this study, we used 10 times 10-fold cross-validation to evaluate the performance of our proposed WLLP method. Specifically, 90% of the interaction data was used as the training set, and the remainder was used as the test set. For the evaluation results of the 10 prediction matrices, we averaged them. The true positive rate (TPR or recall), false positive rate (FPR), precision, AUC, and area under the precision-recall curve (AUPR) were used as evaluation metrics. The TPR and FPR indicate the ability of the model to correctly predict positive and negative labels. Precision is the ratio of correctly predicted positive labels to all predicted positive labels, and greater precision indicates better prediction performance. The formulas for the TPR, FPR, and precision are as follows:

$$TPR = \frac{TP}{TP + FN}, \tag{9}$$
$$FPR = \frac{FP}{TN + FP}, \text{ and} \tag{10}$$
$$Precision = \frac{TP}{TP + FP}, \tag{11}$$

where *TP* denotes the number of labels correctly predicted as positive, *TN* denotes the number of labels correctly predicted as negative, *FP* denotes the number of labels incorrectly predicted as positive, and *FN* denotes the number of labels incorrectly predicted as negative.

Area under the receiver operating characteristic curve and AUPR are widely used to evaluate the performance of binary classifiers. We constructed the ROC curve and the precision-recall (PR) curve by calculating the TPR, FPR, and precision. The ROC curve is a probability curve with FPR on the x-axis and TPR on the y-axis at various thresholds (Kumar and Indrayan, 2011; Pegoraro et al., 2021; Sun et al., 2022). The AUC is then the area under the ROC curve, which is primarily used to describe the global prediction performance, where larger values indicate better performance (Tang et al., 2022). An AUC of 1 indicates excellent performance and an AUC of 0.5 indicates stochastic performance (Peng et al., 2020). In addition, the PR curve is more effective than the ROC curve for representing highly unbalanced data, thus we also used the AUPR to fully evaluate the performance of the WLLP model. Similar to the AUC, a larger AUPR corresponds to better prediction performance.

3.2. Model comparison

In this study, we compared the WLLP model with four other models, namely SCPMF (Meng et al., 2021), NTSIM (Zhang et al., 2018c), TP-NRWRH (Liu et al., 2016), and VDA-RWR (Peng et al., 2021), for the same HDVD dataset. SCPMF is a drug–virus interaction prediction algorithm based on a novel SCPMF. NTSIM is a drug–disease association prediction method that considers only LNS and label propagation. TP-NRWRH uses the bipartite network projection to enhance similarity and propagates it over a heterogeneous network of drugs and diseases with the help of RWR. VDA-RWR applies RWR to the prediction of the newest drug-coronavirus association.

Table 1 shows a comparison of the results obtained from the five prediction models for the HDVD dataset with 10 times 10-fold cross-validation. Figure 2 shows the corresponding ROC and PR curves for the five models. The experimental results demonstrated that the ROC and PR curves of our WLLP model were higher than those of the other four models. It was also apparent that our proposed model offered the best performance in terms of the average AUC and AUPR values. More concretely, the AUC value of WLLP was 0.8828, which was higher than that of the other four approaches (SCPMF: 0.8596; NTSIM: 0.8552; TP-NRWRH: 0.8090; and VDA-RWR: 0.7999). Meanwhile, the AUPR value of WLLP was 0.5277, which was also higher than the other four methods (SCPMF: 0.4958; NTSIM: 0.4778; TP-NRWRH: 0.4929; and VDA-RWR:0.4781). It was not difficult to find

TABLE 1 Performances of the five prediction methods on the human drug virus database (HDVD) dataset.

Method	10 times 10-fold CV AUC	10 times 10-fold CV AUPR
WLLP	**0.8828 ± 0.0037**	**0.5277 ± 0.0053**
SCPMF	0.8596 ± 0.0011	0.4958 ± 0.0010
NTSIM	0.8552 ± 0.0051	0.4778 ± 0.0110
TP-NRWRH	0.8090 ± 0.0079	0.4929 ± 0.0175
VDA-RWR	0.7999 ± 0.0071	0.4781 ± 0.0143

that the NTSIM model produced much better results on AUC than the TP-NRWRH and VDA-RWR models, which implied that using LNS was superior to using the original similarity alone, and indicated that using more complex and effective similarity performance provided more important information for association prediction. Due to the effect of the WKNKN pre-training method on the sparsity of the original interaction matrix, the WLLP model produced better prediction results than the NTSIM model, and it also supported the usefulness of the preprocessing procedure (WKNKN) by comparing with the SCPMF model. In summary, the WLLP model exhibited excellent performance.

4. Discussion

4.1. Ablation experiments

To investigate the plausibility of the WLLP structure, we also tested the model with ablation experiments. We again applied 10 times 10-fold cross-validation to calculate the AUC and AUPR values of the compared models, and the average results were used as the final evaluation indices. The WLLP model comprises three components: WKNKN, LNS, and label propagation (LP). As shown in Table 2, model 1 uses only LNS to set the weights between the nodes on the original label graph and uses label propagation for network diffusion, while model 2 directly applies label propagation to the association network.

The results presented in Table 2 demonstrated that the WLLP model resulted in better AUC and AUPR values for the HDVD dataset than the other two models. Specifically, for model 1, owing to the sparsity of the original drug–virus association matrix, the lack of diffusion channels without preprocessing using the WKNKN algorithm made the nodes with blank labels received little or no resources during network diffusion, and the propagated information was concentrated on the nodes with high association probability in the global prediction. The introduction of WKNKN alleviated the sparsity of the matrix, and the association prediction of blank labels by WLLP became very simple. Therefore, the

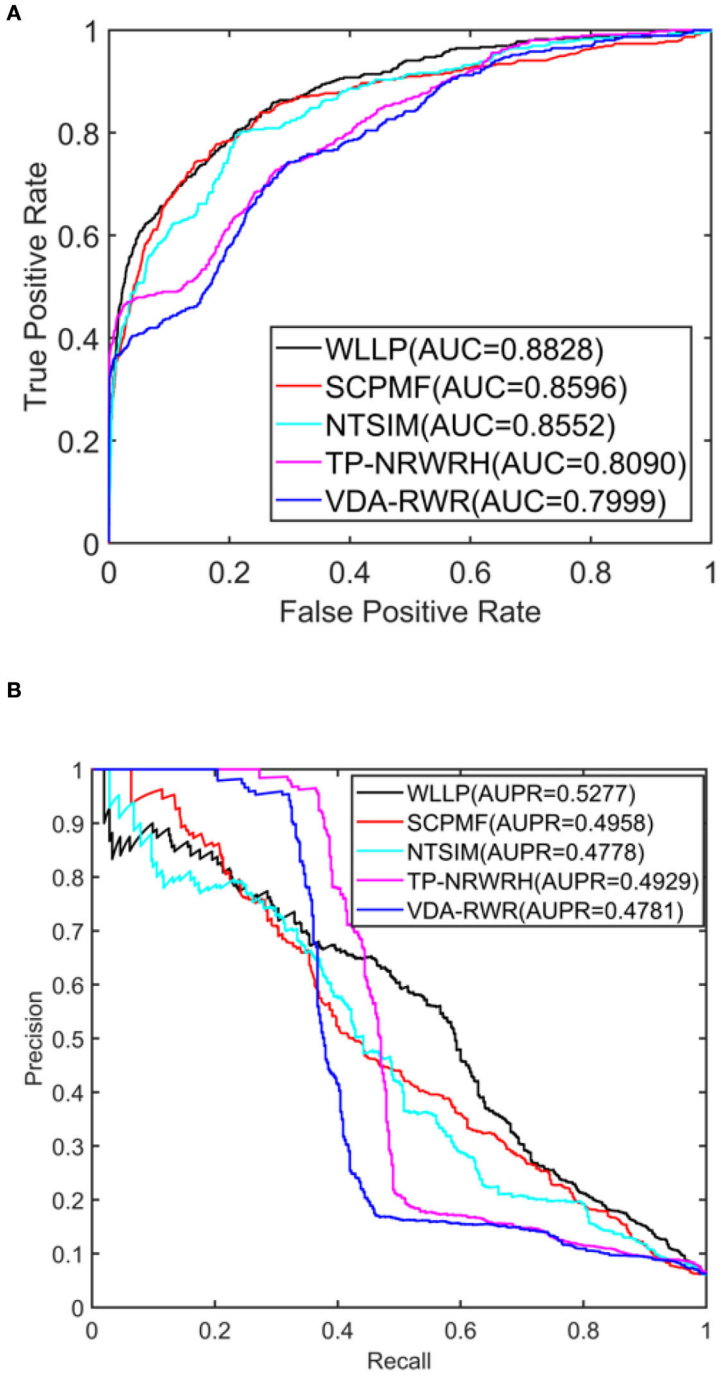

FIGURE 2
Area under the receiver operating characteristic curve (AUC) and area under the precision-recall curve (AUPR) values of the five prediction methods on the human drug virus database (HDVD) dataset. **(A)** AUC values of the five prediction methods. **(B)** AUPR values of the five prediction methods.

WKNKN algorithm can be considered an indispensable part of WLLP. Furthermore, a comparison of model 2 and model 1 clearly revealed that the label propagation algorithm in conjunction with LNS took more information into account than using the chemical structure and sequence similarity alone. The lack of a linear relationship between nodes can make the connections less compact, which in turn leads to poor association prediction for highly unbalanced samples, which is the main reason why the AUPR of model 2 was only 0.1028.

TABLE 2 Results of ablation experiments for the weighted reconstruction-based linear label propagation algorithm (WLLP) model.

Model	WKNKN	LNS	LP	10 times 10-fold CV AUC	10 times 10-fold CV AUPR
WLLP	✓	✓	✓	0.8828 ± 0.0037	0.5277 ± 0.0053
Model		✓	✓	0.8552 ± 0.0051	0.4778 ± 0.0110
Model			✓	0.7886 ± 0.0045	0.1028 ± 0.0005

4.2. Parameter settings

We conducted experiments to analyze the effect of parameters on model WLLP. To determine the optimal combination of parameters, we used the grid search method. The WLLP model used seven parameters, namely K, T, DN, dN, α, β, and w, where K and T are the parameters appearing in the WKNKN algorithm. K denotes the maximum neighborhood value in the KNN function, while T denotes the decay factor. The adjustment range of parameter K is from 1 to 10, while the adjustment range of parameter T is from 1 to 0.1. We end up with K set to 8 and T set to 1 (Figure 3). DN and dN correspond to the number of elements in the set of nearest neighbors for the drugs and viruses in the LNS calculation process. The number of drug neighbors DN should be less than the number of all drugs, and the same is true for the number of virus neighbors dN based on previous experience (Chen et al., 2021b). We varied the values from 10 to 100, increasing by 10 each time. In Figure 4, for the label propagation algorithm, we used α and β to represent the retention probability of the update status for drugs and viruses. Thus, we set the different values of α and β from 0.1 to 1 with step 0.1 (Figure 5). Meanwhile, w is the label fusion parameter for the final matrix from 0.9 to 0.1 with step 0.1. The effect of the parameter selection of w is shown in Figure 6, where we observed that good performance is achieved at $w = 0.4$. The optimal parameter values for the best model performance were found to be as follows: $K = 8$, $T = 1$, $DN = 100$, $dN = 6$, $\alpha = 0.2$, $\beta = 0.5$, and $w = 0.4$.

4.3. Case study

The overall aim of this work was to identify possible clues for the treatment of COVID-19 after confirming the performance of the WLLP model. Table 3 lists the top 15 drugs predicted from the HDVD dataset, showing the ranking, drug name, DrugBank ID, and literature evidence for each drug. It can be observed that a majority (80%) of the predicted drugs were supported by a variety of literature evidence. Ribavirin was initially recommended for clinical use in China 2019-nCoV Pneumonia Treatment Plan Version 5-Revised (Khalili et al., 2020). It is the eight predicted drug candidate for the potential treatment of COVID-19. Remdesivir is a nucleotide analog precursor drug with a broad viral spectrum that includes filoviruses, pneumoviruses, parvoviruses, and coronaviruses (Al-Tawfiq et al., 2020; Grein et al., 2020). Remdesivir inhibits viral RNA polymerase and displays *in vitro* activity against COVID-19 (Al-Tawfiq et al., 2020; De Wit et al., 2020; Grein et al., 2020). The combination of remdesivir with emetine may provide better clinical efficacy (Touret and de Lamballerie, 2020). Chloroquine is an inexpensive, safe, and widely administered antimalarial drug that has been used for more than 70 years and is very effective in controlling COVID-19 infection *in vitro* and therefore may be used for the clinical treatment of COVID-19 (Choy et al., 2020). The combination of chloroquine and remdesivir was reported to be very effective in controlling COVID-19 infection *in vitro* (Wang et al., 2020). Based on their combined pathophysiological and pharmacological potential, camostat and nitazoxanide may be recommended for early evaluation and clinical trials against COVID-19 (Khatri and Mago, 2020). Another study provided preliminary evidence for the use of favipiravir in the treatment of SARS-CoV-2 infection (Cai et al., 2020). Umifenovir is a broad-spectrum antiviral drug. In recent years, clinical trials of umifenovir have been initiated in China (O'Boyle et al., 2011). Sodium lauryl sulfate, an anionic surfactant with protein denaturing ability, effectively inhibits the infectivity of several enveloped viruses through denaturation of the viral envelope. Mouthwash containing sodium lauryl sulfate may be effective in preventing SARS-CoV-2 infection through the oral cavity (Sawa et al., 2021). The 18-kDa cytoplasmic protein procyclin A is an important cellular biomolecule required for RNA virus replication, and recent studies have shown that non-immunosuppressive analogs, such as alisporivir, inhibit the activity of procyclins (Almasi and Mohammadipanah, 2020). Saracatinib, sirolimus, and suramin have also been indicated as therapeutic agents for COVID-19 in recent studies (Romanelli and Mascolo, 2020; Salgado-Benvindo et al., 2020; Tatar et al., 2021).

For hexachlorophene, rifamycin, and tacrolimus, there are no studies proving their activity against COVID-19. However, hexachlorophene is a common detergent additive used for hand washing and disinfection, while rifamycin is an anti-tuberculosis agent that exhibits antiviral properties against various infectious viruses. Tacrolimus, an immunosuppressant, is commonly used in immunotherapy. Although no studies

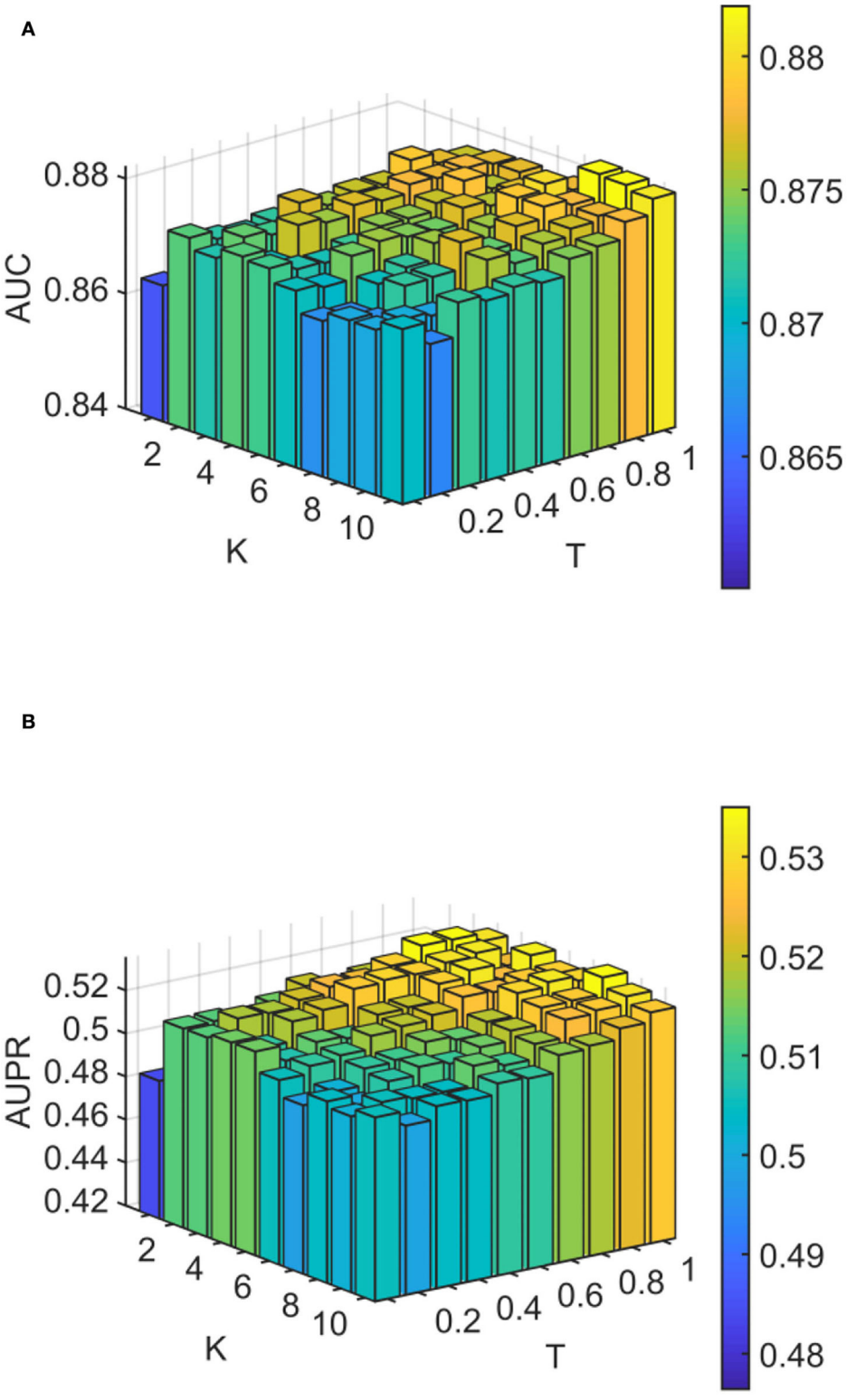

FIGURE 3
Analytical plots of AUC and AUPR for *K* and *T* in the weighted K-nearest known neighbors (WKNKN) algorithm. **(A)** Analytical plots of AUC for *K* and *T*. **(B)** Analytical plots of AUPR for *K* and *T*.

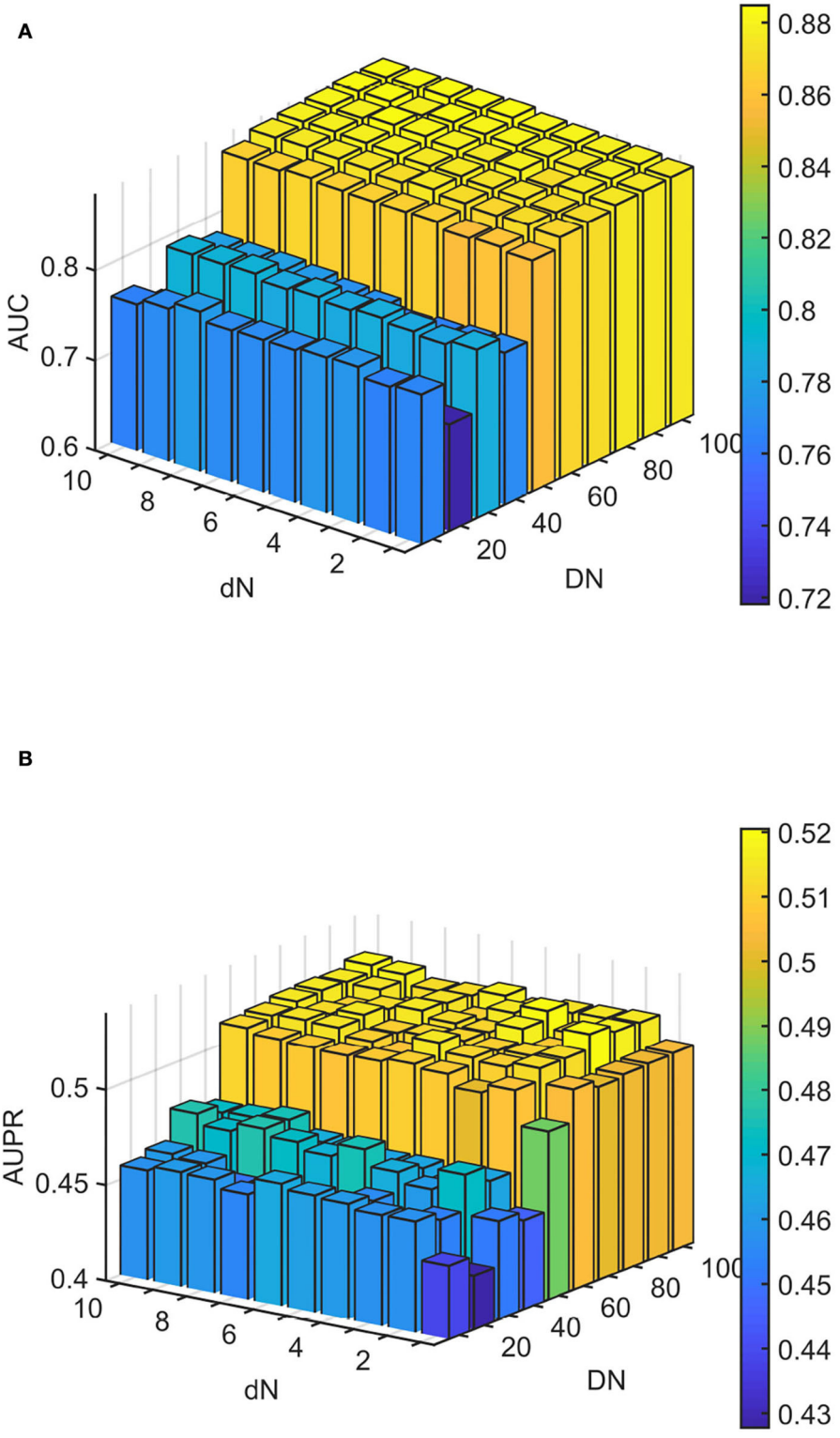

FIGURE 4
Analytical plots of AUC and AUPR for DN and dN in the linear neighborhood similarity (LNS) algorithm. **(A)** Analytical plots of AUC for *DN* and *dN*. **(B)** Analytical plots of AUPR for *DN* and *dN*.

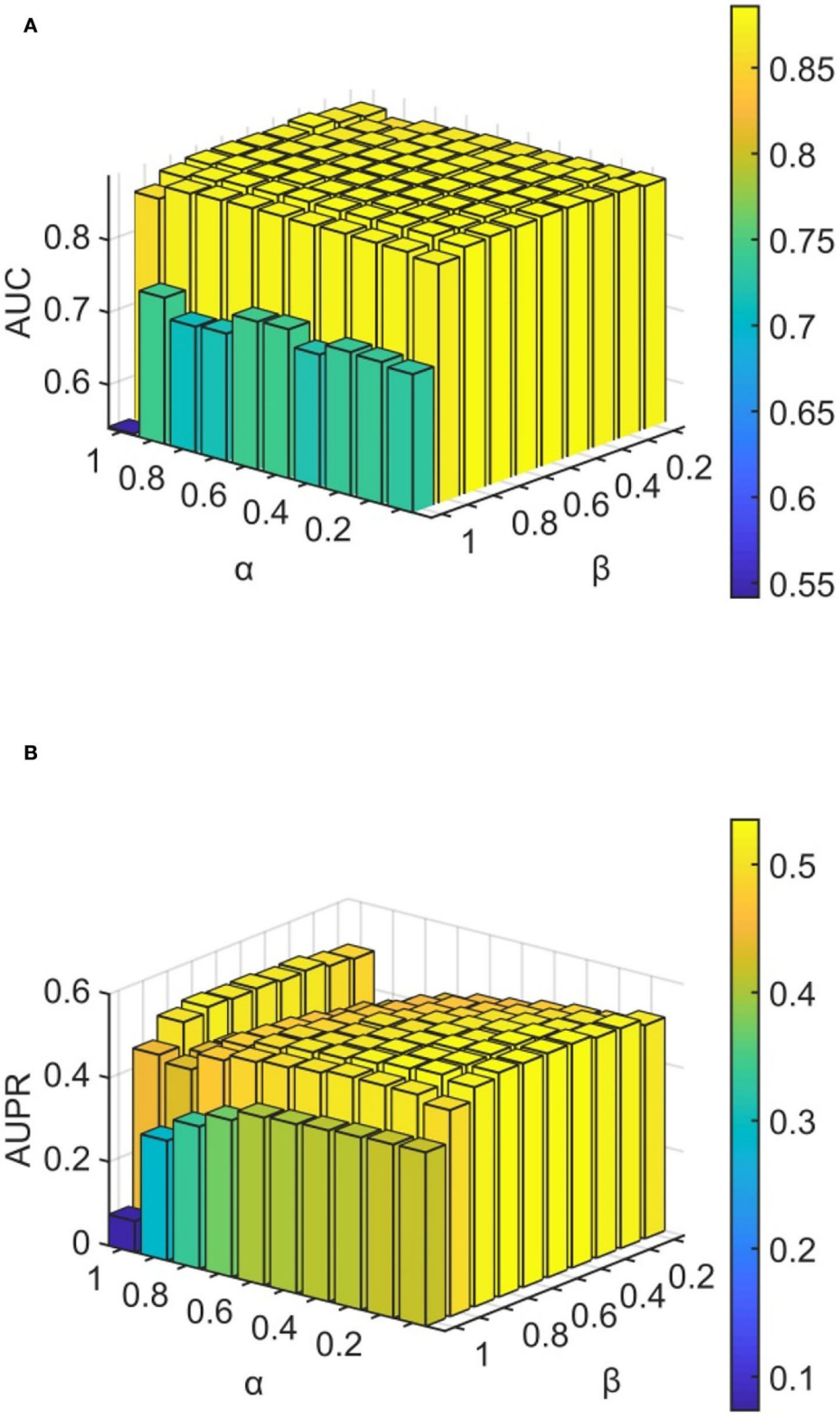

FIGURE 5
Analytical plots of AUC and AUPR for α and β in the LP algorithm. **(A)** Analytical plots of AUC for α and β. **(B)** Analytical plots of AUPR for α and β.

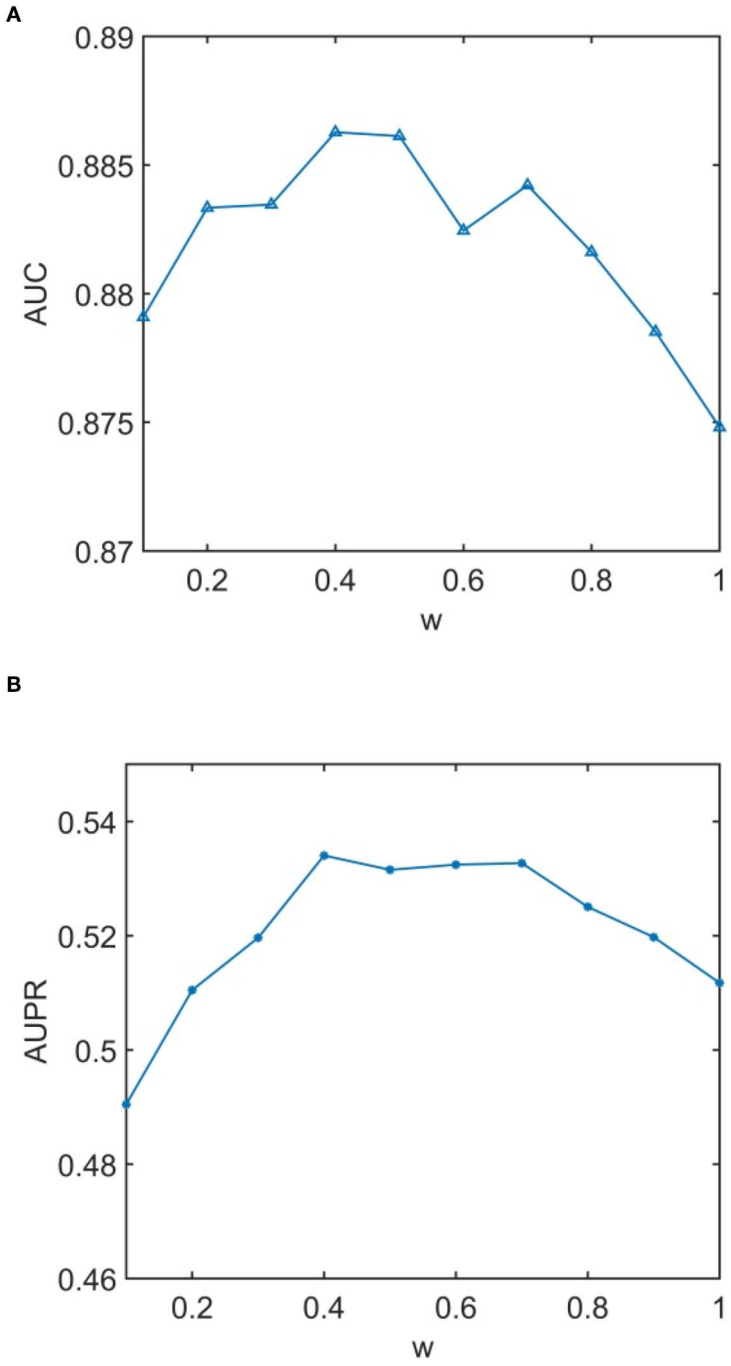

FIGURE 6
Analytical plots of AUC and AUPR for *w* in the label propagation (LP) algorithm. **(A)** Analytical plots of AUC for *w*. **(B)** Analytical plots of AUPR for *w*.

have been conducted to demonstrate the efficacy of these three drugs against COVID-19, they still have considerable potential, which remains to be further validated by subsequent work of drug developers.

5. Summary

To prevent the spread of SARS-CoV-2, it is critical to deepening our understanding of the association between the

TABLE 3 Top 15 drugs predicted from the HDVD dataset.

Rank	Drug name (DrugBank ID)	Evidence
1	Chloroquine (DB00608)	Choy et al., 2020
2	Hexachlorophene (DB00756)	Unknown
3	Nitazoxanide (DB00507)	Khatri and Mago, 2020
4	Rifamycin (DB11753)	Unknown
5	Remdesivir (DB14761)	Al-Tawfiq et al., 2020; Grein et al., 2020; Meng et al., 2021
6	Odium lauryl sulfate (DB00815)	Sawa et al., 2021
7	Camostat (DB13729)	Zhou et al., 2015; Hoffmann et al., 2020
8	Ribavirin (DB00811)	Khalili et al., 2020
9	Saracatinib (DB11805)	Tatar et al., 2021
10	Alisporivir (DB12139)	Almasi and Mohammadipanah, 2020
11	Tacrolimus (DB00864)	Unknown
12	Favipiravir (DB12466)	Cai et al., 2020
13	Sirolimus (DB00877)	Romanelli and Mascolo, 2020
14	Suramin (DB04786)	Salgado-Benvindo et al., 2020
15	Umifenovir (DB13609)	McKee et al., 2020

virus, target proteins, and potential drugs. In the short term, it may be unrealistic to rely on conventional laboratory techniques to develop new drugs against COVID-19, and drug repositioning may represent a more powerful approach. Drug repositioning provides an effective method for prioritizing chemical agents associated with SARS-CoV-2. In this study, a WLLP approach was used to predict the relevance of unknown associations based on drug-virus heterogeneous association networks by combining LNS with LP. The algorithm performs LP on the drug–virus association network, the drug–drug LNS network, and the virus–virus LNS network to diffuse the existing information. With 10 times 10-fold cross-validation, our model achieved an AUC of 0.8828 and an AUPR of 0.5277, both of which were higher than the other methods used for comparison. Furthermore, the information and feasibility of the first 15 drugs were determined by a case study of SARS-CoV-2. Even so, our model still has room for improvement. The predictive performance of the proposed method is limited owing to the current scarcity of data. In the future, we will attempt to tap into drug library and pharmacological resources, and with the addition and integration of more data from recent studies, the prediction results of our model should be improved.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

Author contributions

LC and DL: conceptualization, investigation, and project administration. LC: data curation and resources. LC, DL, HX, JL, and LL: formal analysis, funding acquisition, and writing—review and editing. LC, DL, and HX: methodology. HX, JL, and LL: supervision. HX and JL: validation and writing draft. JL: visualization. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Natural Science Foundation of China (72001202 and 62002070), the Opening Project of Guangdong Province Key Laboratory of Computational Science at Sun Yat-sen University (2021013), and the Science and Technology Plan Project of Guangzhou City (202102021236).

Acknowledgments

We thank reviewers for their valuable suggestions.

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.

Publisher's note

References

Almasi, F., and Mohammadipanah, F. (2020). Potential targets and plausible drugs of coronavirus infection caused by 2019-ncov. *Authorea [Preprints]*. doi: 10.22541/au.158766083.33108969

Al-Tawfiq, J. A., Al-Homoud, A. H., and Memish, Z. A. (2020). Remdesivir as a possible therapeutic option for the COVID-19. *Travel. Med. Infect. Dis.* 34, 101615. doi: 10.1016/j.tmaid.2020.101615

Bajusz, D., Rácz, A., and Héberger, K. (2015). Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.* 7, 1–13. doi: 10.1186/s13321-015-0069-3

Barlow, A., Landolf, K. M., Barlow, B., Yeung, S. Y. A., Heavner, J. J., Claassen, C. W., et al. (2020). Review of emerging pharmacotherapy for the treatment of coronavirus disease 2019. *Pharmacotherapy* 40, 416–437. doi: 10.1002/phar.2398

Cai, Q., Yang, M., Liu, D., Chen, J., Shu, D., Xia, J., et al. (2020). Experimental treatment with favipiravir for COVID-19: an open-label control study. *Engineering* 6, 1192–1198. doi: 10.1016/j.eng.2020.03.007

Chen, H., Zhang, H., Zhang, Z., Cao, Y., and Tang, W. (2015). Network-based inference methods for drug repositioning. *Comput. Math. Methods Med.* 2015, 130620. doi: 10.1155/2015/130620

Chen, X., and Huang, L. (2017). Lrsslmda: Laplacian regularized sparse subspace learning for mirna-disease association prediction. *PLoS Comput Biol.* 13, e1005912. doi: 10.1371/journal.pcbi.1005912

Chen, X., Li, T.-H., Zhao, Y., Wang, C.-C., and Zhu, C.-C. (2021a). Deep-belief network for predicting potential mirna-disease associations. *Brief. Bioinform.* 22, bbaa186. doi: 10.1093/bib/bbaa186

Chen, X., Sun, L.-G., and Zhao, Y. (2021b). Ncmcmda: mirna-disease association prediction through neighborhood constraint matrix completion. *Brief. Bioinform.* 22, 485–496. doi: 10.1093/bib/bbz159

Chen, X., Wang, L., Qu, J., Guan, N.-N., and Li, J.-Q. (2018a). Predicting mirna-disease association based on inductive matrix completion. *Bioinformatics* 34, 4256–4265. doi: 10.1093/bioinformatics/bty503

Chen, X., Xie, D., Wang, L., Zhao, Q., You, Z.-H., and Liu, H. (2018b). Bnpmda: bipartite network projection for mirna-disease association prediction. *Bioinformatics* 34, 3178–3186. doi: 10.1093/bioinformatics/bty333

Chen, X., Xie, D., Zhao, Q., and You, Z.-H. (2019a). Micrornas and complex diseases: from experimental results to computational models. *Brief. Bioinform.* 20, 515–539. doi: 10.1093/bib/bbx130

Chen, X., Yin, J., Qu, J., and Huang, L. (2018c). Mdhgi: matrix decomposition and heterogeneous graph inference for mirna-disease association prediction. *PLoS Comput. Biol.* 14, e1006418. doi: 10.1371/journal.pcbi.1006418

Chen, X., Zhu, C.-C., and Yin, J. (2019b). Ensemble of decision tree reveals potential mirna-disease associations. *PLoS Comput. Biol.* 15, e1007209. doi: 10.1371/journal.pcbi.1007209

Choy, K.-T., Wong, A. Y.-L., Kaewpreedee, P., Sia, S. F., Chen, D., Hui, K. P. Y., et al. (2020). Remdesivir, lopinavir, emetine, and homoharringtonine inhibit sars-cov-2 replication *in vitro*. *Antiviral Res.* 178, 104786. doi: 10.1016/j.antiviral.2020.104786

De Wit, E., Feldmann, F., Cronin, J., Jordan, R., Okumura, A., Thomas, T., et al. (2020). Prophylactic and therapeutic remdesivir (gs-5734) treatment in the rhesus macaque model of mers-cov infection. *Proc. Natl. Acad. Sci. U.S.A.* 117, 6771–6776. doi: 10.1073/pnas.1922083117

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. doi: 10.48550/arXiv.1810.04805

Dotolo, S., Marabotti, A., Facchiano, A., and Tagliaferri, R. (2021). A review on drug repurposing applicable to COVID-19. *Brief. Bioinform.* 22, 726–741. doi: 10.1093/bib/bbaa288

Ezzat, A., Zhao, P., Wu, M., Li, X.-L., and Kwoh, C.-K. (2016). Drug-target interaction prediction with graph regularized matrix factorization. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 14, 646–656. doi: 10.1109/TCBB.2016.2530062

Fan, H.-H., Wang, L.-Q., Liu, W.-L., An, X.-P., Liu, Z.-D., He, X.-Q., et al. (2020). Repurposing of clinically approved drugs for treatment of coronavirus disease 2019 in a 2019-novel coronavirus-related coronavirus model. *Chin. Med. J.* 133, 1051–1056. doi: 10.1097/CM9.0000000000000797

Gönen, M. (2012). Predicting drug-target interactions from chemical and genomic kernels using bayesian matrix factorization. *Bioinformatics* 28, 2304–2310. doi: 10.1093/bioinformatics/bts360

Grein, J., Ohmagari, N., Shin, D., Diaz, G., Asperges, E., Castagna, A., et al. (2020). Compassionate use of remdesivir for patients with severe COVID-19. *N. Engl. J. Med.* 382, 2327–2336. doi: 10.1056/NEJMoa2007016

Guan, W.-J., Ni, Z.-Y., Hu, Y., Liang, W.-H., Ou, C.-Q., He, J.-X., et al. (2020). Clinical characteristics of coronavirus disease 2019 in China. *N. Engl. J. Med.* 382, 1708–1720. doi: 10.1056/NEJMoa2002032

Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., et al. (2020). Sars-cov-2 cell entry depends on ace2 and tmprss2 and is blocked by a clinically proven protease inhibitor. *Cell* 181, 271–280. doi: 10.1016/j.cell.2020.02.052

Hu, H., Zhang, L., Ai, H., Zhang, H., Fan, Y., Zhao, Q., et al. (2018). Hlpi-ensemble: prediction of human lncrna-protein interactions based on ensemble strategy. *RNA Biol.* 15, 797–806. doi: 10.1080/15476286.2018.1457935

Katoh, K., and Standley, D. M. (2013). Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Keshavarzi Arshadi, A., Webb, J., Salem, M., Cruz, E., Calad-Thomson, S., Ghadirian, N., et al. (2020). Artificial intelligence for COVID-19 drug discovery and vaccine development. *Front. Artif. Intell.* 3, 65. doi: 10.3389/frai.2020.00065

Khalili, J. S., Zhu, H., Mak, N. S. A., Yan, Y., and Zhu, Y. (2020). Novel coronavirus treatment with ribavirin: groundwork for an evaluation concerning COVID-19. *J. Med. Virol.* 92, 740–746. doi: 10.1002/jmv.25798

Khatri, M., and Mago, P. (2020). Nitazoxanide/camostat combination for COVID-19: an unexplored potential therapy. *Chem. Biol. Lett.* 7, 192–196. Available online at: http://pubs.iscience.in/journal/index.php/cbl/article/view/1085

King, A. D., Pržulj, N., and Jurisica, I. (2012). Protein complex prediction with RNSC. *Methods Mol. Biol.* 804, 297–312. doi: 10.1007/978-1-61779-361-5_16

Kumar, R., and Indrayan, A. (2011). Receiver operating characteristic (roc) curve for medical researchers. *Indian Pediatr.* 48, 277–287. doi: 10.1007/s13312-011-0055-4

Liu, H., Ren, G., Chen, H., Liu, Q., Yang, Y., and Zhao, Q. (2020). Predicting lncrna-mirna interactions based on logistic matrix factorization with neighborhood regularized. *Knowl. Based Syst.* 191, 105261. doi: 10.1016/j.knosys.2019.105261

Liu, H., Song, Y., Guan, J., Luo, L., and Zhuang, Z. (2016). Inferring new indications for approved drugs via random walk on drug-disease heterogenous networks. *BMC Bioinform.* 17, 269–277. doi: 10.1186/s12859-016-1336-7

Liu, S., Zheng, Q., and Wang, Z. (2020). Potential covalent drugs targeting the main protease of the sars-cov-2 coronavirus. *Bioinformatics* 36, 3295–3298. doi: 10.1093/bioinformatics/btaa224

Liu, W., Jiang, Y., Peng, L., Sun, X., Gan, W., Zhao, Q., et al. (2022a). Inferring gene regulatory networks using the improved markov blanket discovery algorithm. *Interdisc. Sci.* 14, 168–181. doi: 10.1007/s12539-021-00478-9

Liu, W., Lin, H., Huang, L., Peng, L., Tang, T., Zhao, Q., et al. (2022b). Identification of mirna-disease associations via deep forest ensemble learning based on autoencoder. *Brief. Bioinform.* 23, bbac104. doi: 10.1093/bib/bbac104

Luo, H., Wang, J., Li, M., Luo, J., Peng, X., Wu, F.-X., et al. (2016). Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm. *Bioinformatics* 32, 2664–2671. doi: 10.1093/bioinformatics/btw228

Macropol, K., Can, T., and Singh, A. K. (2009). Rrw: repeated random walks on genome-scale protein networks for local cluster discovery. *BMC Bioinformatics* 10, 1–10. doi: 10.1186/1471-2105-10-283

Martinez, V., Navarro, C., Cano, C., Fajardo, W., and Blanco, A. (2015). Drugnet: network-based drug-disease prioritization by integrating heterogeneous data. *Artif. Intell. Med.* 63, 41–49. doi: 10.1016/j.artmed.2014.11.003

McKee, D. L., Sternberg, A., Stange, U., Laufer, S., and Naujokat, C. (2020). Candidate drugs against sars-cov-2 and covid-19. *Pharmacol. Res.* 157, 104859. doi: 10.1016/j.phrs.2020.104859

Meng, Y., Jin, M., Tang, X., and Xu, J. (2021). Drug repositioning based on similarity constrained probabilistic matrix factorization: COVID-19 as a case study. *Appl. Soft Comput.* 103, 107135. doi: 10.1016/j.asoc.2021.107135

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R. (2011). Open babel: An open chemical toolbox. *J. Cheminform.* 3, 1–14. doi: 10.1186/1758-2946-3-33

Organization, W. H. (2020). *Global surveillance for covid-19 disease caused by human infection with novel coronavirus (covid-19)*:

interim guidance, 27 february 2020. Technical report, World Health Organization.

Öztürk, H., Ozkirimli, E., and Özgür, A. (2016). A comparative study of smiles-based compound similarity functions for drug-target interaction prediction. *BMC Bioinform*. 17, 1–11. doi: 10.1186/s12859-016-0977-x

Pegoraro, J. A., Lavault, S., Wattiez, N., Similowski, T., Gonzalez-Bermejo, J., and Birmelé, E. (2021). Machine-learning based feature selection for a non-invasive breathing change detection. *BioData Min*. 14, 1–16. doi: 10.1186/s13040-021-00265-8

Peng, L., Shen, L., Xu, J., Tian, X., Liu, F., Wang, J., et al. (2021). Prioritizing antiviral drugs against sars-cov-2 by integrating viral complete genome sequences and drug chemical structures. *Sci. Rep*. 11, 1–11. doi: 10.1038/s41598-021-83737-5

Peng, L.-H., Zhou, L.-Q., Chen, X., and Piao, X. (2020). A computational study of potential mirna-disease association inference based on ensemble learning and kernel ridge regression. *Front. Bioeng. Biotechnol*. 8, 40. doi: 10.3389/fbioe.2020.00040

Pollastri, G., Przybylski, D., Rost, B., and Baldi, P. (2002). Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins* 47, 228–235. doi: 10.1002/prot.10082

Pushpakom, S., Iorio, F., Eyers, P. A., Escott, K. J., Hopper, S., Wells, A., et al. (2019). Drug repurposing: progress, challenges and recommendations. *Nat. Rev. Drug Discov*. 18, 41–58. doi: 10.1038/nrd.2018.168

Re, M., and Valentini, G. (2013). Network-based drug ranking and repositioning with respect to drugbank therapeutic categories. *IEEE/ACM Trans. Comput. Biol. Bioinform*. 10, 1359–1371. doi: 10.1109/TCBB.2013.62

Romanelli, A., and Mascolo, S. (2020). Sirolimus to treat sars-cov-2 infection: an old drug for a new disease. *Respir. Med*. 8, 420–422. doi: 10.34172/jrcm.2020.044

Salgado-Benvindo, C., Thaler, M., Tas, A., Ogando, N. S., Bredenbeek, P. J., Ninaber, D. K., et al. (2020). Suramin inhibits sars-cov-2 infection in cell culture by interfering with early steps of the replication cycle. *Antimicrob. Agents Chemother*. 64, e00900-e00920. doi: 10.1128/AAC.00900-20

Sawa, Y., Ibaragi, S., Okui, T., Yamashita, J., Ikebe, T., and Harada, H. (2021). Expression of sars-cov-2 entry factors in human oral tissue. *J. Anat*. 238, 1341–1354. doi: 10.1111/joa.13391

Shin, B., Park, S., Kang, K., and Ho, J. C. (2019). "Self-attention based molecule representation for predicting drug-target interaction," in *Machine Learning for Healthcare Conference* (PMLR), 230–248. Available online at: http://proceedings.mlr.press/v106/shin19a.html?ref=https://githubhelp.com

Sirin, U., Erdogdu, U., Polat, F., Tan, M., and Alhajj, R. (2016). Effective gene expression data generation framework based on multi-model approach. *Artif. Intell. Med*. 70, 41–61. doi: 10.1016/j.artmed.2016.05.003

Sohrabi, C., Alsafi, Z., O'neill, N., Khan, M., Kerwan, A., Al-Jabir, A., et al. (2020). World health organization declares global emergency: a review of the 2019 novel coronavirus (COVID-19). *Int. J. Surgery* 76, 71–76. doi: 10.1016/j.ijsu.2020.02.034

Sun, F., Sun, J., and Zhao, Q. (2022). A deep learning method for predicting metabolite-disease associations via graph neural network. *Brief. Bioinform*. 23, bbac266. doi: 10.1093/bib/bbac266

Tang, Q., Nie, F., Zhao, Q., and Chen, W. (2022). A merged molecular representation deep learning method for blood-brain barrier permeability prediction. *Brief. Bioinform*. 23, bbac357. doi: 10.1093/bib/bbac357

Tatar, G., Ozyurt, E., and Turhan, K. (2021). Computational drug repurposing study of the rna binding domain of sars-cov-2 nucleocapsid protein with antiviral agents. *Biotechnol. Prog*. 37, e3110. doi: 10.1002/btpr.3110

Torng, W., and Altman, R. B. (2019). Graph convolutional neural networks for predicting drug-target interactions. *J. Chem. Inf. Model*. 59, 4131–4149. doi: 10.1021/acs.jcim.9b00628

Touret, F., and de Lamballerie, X. (2020). Of chloroquine and COVID-19. *Antiviral Res*. 177, 104762. doi: 10.1016/j.antiviral.2020.104762

Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., and Sharan, R. (2010). Associating genes and protein complexes with disease via network propagation. *PLoS Comput. Biol*. 6, e1000641. doi: 10.1371/journal.pcbi.1000641

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems, Vol. 30*. Available online at: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Wang, C.-C., Han, C.-D., Zhao, Q., and Chen, X. (2021). Circular rnas and complex diseases: from experimental results to computational models. *Brief. Bioinform*. 22, bbab286. doi: 10.1093/bib/bbab286

Wang, F., and Zhang, C. (2006). "Label propagation through linear neighborhoods," in *Proceedings of the 23rd International Conference on Machine Learning*, 985–992. doi: 10.1145/1143844

Wang, M., Cao, R., Zhang, L., Yang, X., Liu, J., Xu, M., et al. (2020). Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-ncov) in vitro. *Cell Res*. 30, 269–271. doi: 10.1038/s41422-020-0282-0

Wheeler, D. L., Church, D. M., Lash, A. E., Leipe, D. D., Madden, T. L., Pontius, J. U., et al. (2002). Database resources of the national center for biotechnology information: 2002 update. *Nucleic Acids Res*. 30, 13–16. doi: 10.1093/nar/30.1.13

Xue, H., Li, J., Xie, H., and Wang, Y. (2018). Review of drug repositioning approaches and resources. *Int. J. Biol. Sci*. 14, 1232. doi: 10.7150/ijbs.24612

Yang, M., Luo, H., Li, Y., and Wang, J. (2019). Drug repositioning based on bounded nuclear norm regularization. *Bioinformatics* 35, i455-i463. doi: 10.1093/bioinformatics/btz331

Zhang, L., Liu, T., Chen, H., Zhao, Q., and Liu, H. (2021a). Predicting lncrna-mirna interactions based on interactome network and graphlet interaction. *Genomics* 113, 874–880. doi: 10.1016/j.ygeno.2021.02.002

Zhang, L., Yang, P., Feng, H., Zhao, Q., and Liu, H. (2021b). Using network distance analysis to predict lncrna-mirna interactions. *Interdisc. Sci*. 13, 535–545. doi: 10.1007/s12539-021-00458-z

Zhang, W., Chen, Y., and Li, D. (2017a). Drug-target interaction prediction through label propagation with linear neighborhood information. *Molecules* 22, 2056. doi: 10.3390/molecules22122056

Zhang, W., Chen, Y., Li, D., and Yue, X. (2018a). Manifold regularized matrix factorization for drug-drug interaction prediction. *J. Biomed. Inform*. 88, 90–97. doi: 10.1016/j.jbi.2018.11.005

Zhang, W., Qu, Q., Zhang, Y., and Wang, W. (2018b). The linear neighborhood propagation method for predicting long non-coding rna-protein interactions. *Neurocomputing* 273, 526–534. doi: 10.1016/j.neucom.2017.07.065

Zhang, W., Yue, X., Huang, F., Liu, R., Chen, Y., and Ruan, C. (2018c). Predicting drug-disease associations and their therapeutic function based on the drug-disease association bipartite network. *Methods* 145, 51–59. doi: 10.1016/j.ymeth.2018.06.001

Zhang, W., Yue, X., Liu, F., Chen, Y., Tu, S., and Zhang, X. (2017b). A unified frame of predicting side effects of drugs by using linear neighborhood similarity. *BMC Syst. Biol*. 11, 23–34. doi: 10.1186/s12918-017-0477-2

Zhou, Y., Hou, Y., Shen, J., Huang, Y., Martin, W., and Cheng, F. (2020). Network-based drug repurposing for novel coronavirus 2019-ncov/sars-cov-2. *Cell Discov*. 6, 1–18. doi: 10.1038/s41421-020-0153-3

Zhou, Y., Vedantham, P., Lu, K., Agudelo, J., Carrion Jr, R., Nunneley, J. W., et al. (2015). Protease inhibitors targeting coronavirus and filovirus entry. *Antiviral Res*. 116, 76–84. doi: 10.1016/j.antiviral.2015.01.011

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., et al. (2020). A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J. Med*. 82, 727–733. doi: 10.1056/NEJMoa2001017

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 16 December 2022
DOI 10.3389/fmicb.2022.1089729

OPEN ACCESS

EDITED BY
Liang Wang,
Guangdong Provincial People's Hospital, China

REVIEWED BY
Yang Jiao,
Arizona State University, United States
Chuansheng Shen,
Anqing Normal University, China

*CORRESPONDENCE
Zhengquan Yu
✉ zyu@cau.edu.cn
Xiang Li
✉ xianglibp@xmu.edu.cn

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

RECEIVED 04 November 2022
ACCEPTED 05 December 2022
PUBLISHED 16 December 2022

CITATION
Qu J, Shao C, Ying Y, Wu Y, Liu W, Tian Y, Yin Z, Li X, Yu Z and Shuai J (2022) The spring-like effect of microRNA-31 in balancing inflammatory and regenerative responses in colitis. *Front. Microbiol.* 13:1089729. doi: 10.3389/fmicb.2022.1089729

The spring-like effect of microRNA-31 in balancing inflammatory and regenerative responses in colitis

Jing Qu[1†], Chunlei Shao[2†], Yongfa Ying[1], Yuning Wu[3], Wen Liu[1], Yuhua Tian[2], Zhiyong Yin[1], Xiang Li[1,4,5*], Zhengquan Yu[2*] and Jianwei Shuai[1,4,5,6,7]

[1]Department of Physics, and Fujian Provincial Key Laboratory for Soft Functional Materials Research, Xiamen University, Xiamen, China, [2]State Key Laboratories for Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing, China, [3]Department of Mathematics and Physics, Fujian Jiangxia University, Fuzhou, China, [4]National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, China, [5]State Key Laboratory of Cellular Stress Biology, Innovation Center for Cell Signaling Network, School of Life Sciences, Xiamen University, Xiamen, China, [6]Oujiang Laboratory (Zhejiang Lab for Regenerative Medicine, Vision and Brain Health), University of Chinese Academy of Sciences, Wenzhou, China, [7]Wenzhou Institute, Wenzhou Key Laboratory of Biophysics, University of Chinese Academy of Sciences, Wenzhou, China

Inflammatory bowel diseases (IBDs) are chronic inflammatory disorders caused by the disruption of immune tolerance to the gut microbiota. MicroRNA-31 (MIR31) has been proven to be up-regulated in intestinal tissues from patients with IBDs and colitis-associated neoplasias. While the functional role of MIR31 in colitis and related diseases remain elusive. Combining mathematical modeling and experimental analysis, we systematically explored the regulatory mechanism of MIR31 in inflammatory and epithelial regeneration responses in colitis. Level of MIR31 presents an "adaptation" behavior in dextran sulfate sodium (DSS)-induced colitis, and the similar behavior is also observed for the key cytokines of p65 and STAT3. Simulation analysis predicts MIR31 suppresses the activation of p65 and STAT3 but accelerates the recovery of epithelia in colitis, which are validated by our experimental observations. Further analysis reveals that the number of proliferative epithelial cells, which characterizes the inflammatory process and the recovery of epithelia in colitis, is mainly determined by the inhibition of MIR31 on IL17RA. MIR31 promotes epithelial regeneration in low levels of DSS-induced colitis but inhibits inflammation with high DSS levels, which is dominated by the competition for MIR31 to either inhibit inflammation or promote epithelial regeneration by binding to different targets. The binding probability determines the functional transformation of MIR31, but the functional strength is determined by MIR31 levels. Thus, the role of MIR31 in the inflammatory response can be described as the "spring-like effect," where DSS, MIR31 action strength, and proliferative epithelial cell number are regarded as external force, intrinsic spring force, and spring length, respectively. Overall, our study uncovers the vital roles of MIR31 in balancing inflammation and the recovery of epithelia in colitis, providing potential clues for the development of therapeutic targets in drug design.

KEYWORDS

MIR31, inflammatory response, epithelial regeneration, spring-like effect, network modeling

Introduction

Inflammatory bowel diseases (IBDs), including ulcerative colitis (UC) and Crohn's disease (CD), are chronic inflammatory disorders that impact gastrointestinal tract (Kaser et al., 2010). Chronic inflammatory disorders are characterized by submucosal accumulation of immune cells, resulting in damage to the epithelial layer (Maloy and Powrie, 2011; Cader and Kaser, 2013). The prevalence and incidence of IBD have continued to increase over the past few decades around the world (Xavier and Podolsky, 2007; Kaplan and Ng, 2017; Ng et al., 2017), but the precise etiology and pathogenesis of IBD have not been fully revealed. Recent studies have shown the significant role of gut microbiota in IBD (Seksik, 2010; Dalal and Chang, 2014; Sheehan et al., 2015). A widely accepted pathogenesis is that environmental or genetic factors trigger an abnormal immune response to the gut microbiota in a genetically susceptible host (Fava and Danese, 2011; Becker et al., 2015; Matsuoka and Kanai, 2015). IBD is associated with marked changes in gene expression and protein level (Neurath, 2014; Kumar et al., 2016). Increasing studies show that microRNAs (miRNAs) play vital roles in the regulation of IBD (Dalal and Kwon, 2010; Kalla et al., 2015; Xu and Zhang, 2016; Soroosh et al., 2018). MiRNAs are a class of small noncoding RNAs with a length of approximately 18–25 nucleotides (Bartel, 2004; Emde and Hornstein, 2014), which are widely found in nematodes, fruit flies, plants and eukaryotes (Bartel, 2009; Carthew and Sontheimer, 2009; Fabian et al., 2010). MiRNAs have been estimated to regulate over 60% of protein-coding genes (Garzon et al., 2010; Lin et al., 2013a; Hammond, 2015; Treiber et al., 2018) by base pairing with target mRNAs and repressing translation (Krol et al., 2010; Ha and Kim, 2014). Abnormal expression of miRNAs is highly correlated with many diseases including cancer (Wu et al., 2008; Schaefer et al., 2015; Tili et al., 2017) and neurodevelopmental disorders (Ha and Kim, 2014).

Among miRNAs, microRNA-31 (MIR31) is increased in colorectal cancer (Wang et al., 2009) and patients with IBD (Béres et al., 2017). MIR31 is also proven to be up-regulated during IBD-associated neoplastic transformation (Olaru et al., 2011). Targeting MIR31 pathways involved in the inflammatory response paves the way for disease treatments (Yu et al., 2021). MIR31 can promote epithelial regeneration during skin wound healing by mediating inflammatory signaling (Shi et al., 2018), and target IL-25 to regulate IL-12/23-mediated Th1/Th17 inflammatory responses during colitis (Shi et al., 2016). We previously showed that MIR31 promotes the self-renewal of mammary stem cells and mammary tumor growth by regulating the WNT signaling pathway (Lv et al., 2017). Our recent study also indicated that MIR31 can reduce inflammatory responses in colonic epithelium by inhibiting inflammatory cytokines receptors, and promote epithelial regeneration through regulating WNT and Hippo signaling pathways (Tian et al., 2019). However, the mechanism underlying the regulation of MIR31 in these pathways is not yet clear. More importantly, the questions of whether and how these two functions (i.e., inflammatory response inhibition and epithelial regeneration promotion) compete for MIR31 during colitis require further clarification.

Experiment-based network modeling is a powerful approach to investigate the biological dynamics in animals (Li et al., 2021b, 2022), plants (Wu et al., 2021), and bacteria (Liu et al., 2021), and is also widely applied to investigate the role of miRNAs (Lai et al., 2018). To systematically analyze the regulatory mechanism of MIR31 in inflammation and epithelial regeneration, experimental analysis is performed and a corresponding phenomenological model is proposed. Experimental observations suggest that the expression of MIR31, phosphorylated p65 (p-65), and phosphorylated STAT3 (p-STAT3) present an "adaptation" behavior in dextran sulfate sodium (DSS)-induced mouse colitis (Ma et al., 2009), which are well reproduced by our model. Dynamics of the number of proliferative epithelial cells and the expression of p-65 and p-STAT3 in MIR31 knockout (KO) mice are also predicted and validated, indicating that MIR31 restrains the activation of p65 and STAT3 but promotes epithelial regeneration. Further analysis suggests the behavior of the epithelial cell number exhibits the "spring-like effect." Acting as an external force, DSS drives the system to a "spring compressing process" by reducing the epithelial cell number which is analogous to spring length. MIR31 acts as the intrinsic force and fine-tunes the epithelial cell number, propelling the system to a "spring compression state" with a small cell number and a high level of MIR31. Highly expressed MIR31 then accelerates epithelial regeneration by promoting cell proliferation after the withdrawal of DSS, corresponding to "spring recovery process." Overall, this study provides quantitative new insights into the regulatory mechanism of MIR31, offering possible therapeutic strategies for colitis and related diseases.

Materials and methods

Animal experiments

Wild-type (WT) C57BL/6 mice were purchased from Beijing Vital River Laboratory Animal Technology Company (Beijing). MIR31 knockout (MIR31-KO) and control mice have been previously described (Tian Y. et al., 2017). About three to four 8-week-old mice were used at each time points for analysis in this

study. All mice were fed under specific pathogen-free conditions. All experiments were approved by the guidelines of the Institutional Animal Care and Use Committee of China Agricultural University.

DSS treatment

Adult mice were fed 3.5% wt/vol DSS with molecular weight 36,000 to 50,000 (MP Biochemicals, Santa Ana, CA) in drinking water for 5 days, and then DSS was withdrawn for recovery for 3 days. Tissues were harvested at the indicated time points. The colonic tissues were fixed in 4% paraformaldehyde for 24 h, embedded in paraffin, sectioned, and stained with hematoxylin and eosin.

Immunofluorescence staining

The paraffin-embedded 5 μm sections were dehydrated with graded alcohol, and antigen retrieval was performed by heating slides for 20 min in 0.01 M citrate buffer (pH 6.0) with a microwave oven. The sections were blocked for 1 h at room temperature with blocking buffer (Beyotime) and incubated with primary antibodies at 4°C overnight. Next, the sections were washed with PBS three times, each time for 5 min, incubated with secondary antibodies for 1 h at room temperature and counterstained with DAPI. For immunohistochemistry staining, antigen retrieval was performed by heating slides for 20 min in 0.01 M citrate buffer (pH 6.0) with a microwave oven. Then, the sections were stained according to the SP Kit (ZSGB-Bio) manufacturer's instructions. The following primary antibodies were used: Ki67 (1:1,000, ab15580, Abcam), β-catenin (1:500, sc-7,963, Santa Cruz), pStat3 (1:400, 9,145, CST), and p65 (1:1,000, 8,242, CST). The following secondary antibodies were used: Alexa Fluor 488 goat anti-mouse IgG (H + L) and Alexa Fluor 594 goat anti-rabbit IgG (H + L).

In situ hybridization

The MIR31 *in situ* hybridization assay was performed as described previously with modifications (Tian Y. et al., 2017). Digoxigenin-labeled LNA probes (Exiqon, Vedbaek, Denmark) were used following the manufacturer's protocol. Both digoxigenin-labeled MIR31 and scrambled LNA probes (Exiqon) were hybridized at 61°C. The U6 probe was used as a positive control. *In situ* signals were detected by staining with anti-digoxigenin-AP antibody (Roche, Basel, Switzerland) and developed using BM purple substrate (Roche).

Model construction

In WT model, the DSS-induced inflammatory response is composed of four processes: DSS-induced inflammatory cytokines production, MIR31 induction, MIR31-inhibited inflammatory cytokines production and MIR31-promoted epithelial regeneration. While in KO model, MIR31 is knocked out and the correlated processes of MIR31 exist no more. We only consider the epithelial regeneration by intrinsic cell proliferation. A system of ordinary differential equations (ODEs) is a common approach to describe dynamics of biochemical reactions and interactions of signaling molecules (Li et al., 2021a). Based on the Hill equation, the evolution of molecular concentrations with time in the model can be described

$$\frac{dY_i}{dt} = \sum_{j}^{s} k_{ij} \times \frac{{Y_j}^n}{l_{ij}^n + {Y_j}^n}, i = 1, \ldots m$$

where d*Yi*/*dt* is the rate at which the concentration of molecule i changes over time. *m* represents the number of molecules with concentration *dYi*. *s* denotes the number of reactions with rate k_{ij}, the half-saturation constant l_{ij} and the Hill coefficient n. *Yj* is the concentration of molecule involved in the reaction. The ODEs that describe the reactions of different modules in the signaling model are shown in Supplementary Text.

Parameter values and initial amount selection

All parameters in the signaling model are first limited to the typical biological ranges depending on the types of reaction (Alon, 2006) and further estimated based on the experimental results (Tian et al., 2019). The initial values of the parameters are random selected to avoid convergence to local minimums, and then mainly determined by a global optimization method to minimize the deviations between simulation results and experimental results, including the expressions of MIR31, p-p65 and p-STAT3, as well as the number of proliferative cells. The descriptions, values and units of all parameters in the signaling model are given in Supplementary Tables S1, S2.

Results

"Adaptation" behavior of MIR31 In DSS-induced colitis

A feasible theory of the pathogenesis of IBD is that the barrier of intestinal epithelial cells loses its function, with luminal organisms or their products entering the lamina propria. One of the most widely used IBD animal models is the DSS-induced colitis mice model (containing a simple microbiota), which is similar to UC in terms of pathology, pathogenesis and other aspects (Wirtz and Neurath, 2007; Solomon et al., 2010; Mizoguchi, 2012; Nguyen et al., 2015; Eichele and Kharbanda, 2017). A common mechanism for DSS-induced colitis involves damage to the intestinal epithelial

barrier, which allows luminal bacteria and associated antigens to enter the mucosa (Figure 1A). The entry induces immune responses from immune cells (e.g., macrophages and T cells) in the epithelial lamina propria and the release of inflammatory cytokines, triggering acute inflammation (Kiesler et al., 2015). Receptors (Gp130 and IL17RA) of inflammatory cytokines localize to the colonic epithelium (Zhang et al., 2006; Ernst et al., 2014) and activate epithelial STAT3 and NF-κB signaling pathways (De Robertis et al., 2011), inducing MIR31 activation. In addition, MIR31 promotes epithelial regeneration by activating the WNT signaling pathway and inhibiting the Hippo signaling pathway through several target genes, such as Axin1 and Lats1/2 (Tian et al., 2019).

DSS-induced inflammatory response is a complicated biological process that involves different cells, such as macrophages and intestinal epithelial cells (IECs; Saleh and Trinchieri, 2011), and the major biological network is shown in Figure 1B. The signaling network is composed of four modules, including the production of DSS-induced inflammatory cytokine (green background), the induction of MIR31 (yellow background), the inhibition of inflammatory cytokine production by MIR31 (blue background), and the epithelial regeneration promoted by MIR31 (purple background). After DSS administration, immune cells can recognize the agents and rapidly release inflammatory cytokines IL1β and IL6 (Francescone et al., 2015), accompanied by increased TNFα expression (Zelov and Hošek, 2013; Kany et al., 2019). When TNFα reaches high concentrations, the NF-κB signaling pathway is activated in the form of phosphorylated p65 (p-p65) *via* the canonical pathway (Giridharan and Srinivasan, 2018), which promotes the production of IL1β (Kelley et al., 2019) and further exacerbates inflammation. The NF-κB signaling pathway also induces IL6 expression, which can activate the STAT3 signaling pathway (Ernst et al., 2014). Then, the activated STAT3 (p-STAT3) promotes the secretion of the inflammatory cytokine IL17, in turn facilitating the activation of the NF-κB pathway (Razi et al., 2019). In addition, IL1β can slightly accelerate the activation of STAT3 (Parker et al., 2015; van de Wetering et al., 2020). We previously identified one STAT3 and two NF-kB binding sites in the promoter of MIR31 (Lv et al., 2017; Tian Y. et al., 2017) and proved that the induction of MIR31 is due to the activation of NF-κB and STAT3 signaling pathways (Tian et al., 2019). Moreover, our former study also demonstrated that MIR31 directly inhibits inflammation through the suppression of receptor Gp130 and receptor IL17RA (Tian et al., 2019). The canonical WNT signaling pathway is a key regulator of epithelial regeneration (Moparthi and Koch, 2019). Hippo signaling pathway also drives epithelial regeneration in colon after DSS-induced injury (Deng et al., 2018; Xie et al., 2021). We previously revealed that MIR31 promotes epithelia regeneration by modulating the WNT and Hippo pathways, restoring the ability of epithelial cells to resist inflammation (Tian et al., 2019). Besides, Axin1 and β-catenin, as well as Lats1/2 and Yap, are the two groups of important transducers in WNT and Hippo pathways, respectively.

To qualitatively investigate the regulatory mechanism of MIR31 in colitis, experimental analysis is performed to explore the dynamics of the core transducers, i.e., MIR31, p-p65, and p-STAT3, in response to the DSS-induced colitis in WT mice. Figure 1C shows the normal colonic tissues (0 days) and colonic tissues at 5 days of DSS treatment from WT mice. *In situ* hybridization for MIR31 in colons in Figures 1D,E is obtained from untreated WT mice and WT mice after 5 days of DSS treatment, respectively. As the quantified experimental data shows, an obvious up-regulation of MIR31 in the colonic epithelium is observed (Figure 1F). After 5-day DSS administration, MIR31 expression increased to a high peak and then rapidly decreased to the baseline of pretreatment after DSS withdrawal, presenting an "adaptation" behavior (Figure 1F). A phenomenological network model is also developed based on the signaling pathways shown in Figure 1B, which can provide a more quantitative diagram to further explore the role of the signaling pathways in the pathogenesis of colitis. The model is described by a cast of ordinary differential equations and the complete model descriptions are presented in the Supplementary material. Simulation results suggest that our model can well reproduce the "adaptation" behavior of MIR31 expression after DSS treatment (Figure 1G).

Dynamical expressions of the two core transducers p-p65 and p-STAT3 that directly facilitate the activation of MIR31 are quantified as well. Abnormal increases in p-p65 are associated with many chronic diseases such as rheumatoid arthritis and IBD (Simmonds and Foxwell, 2008; Giridharan and Srinivasan, 2018). With DSS treatment, the NF-κB signaling pathway is activated, accompanied by a sustained increase in p-p65, which can be detected in the nucleus by immunohistochemistry (Figure 2A). Then, p-p65 expression is gradually restored to its initial level along with the reduction of inflammation. The quantified experimental data shown in Figure 2B indicates that p-p65 also presents an "adaptation" behavior, which can be reproduced by our model (Figure 2C). The p-STAT3 mediated inflammation is shown by immunohistochemistry in Figure 2D and the experimental result suggests the "adaptation" behavior of p-STAT3 (Figure 2E). Simulation results of p-STAT3 and p-p65 expressions suggest that the "adaptation" behavior of p-STAT3 and p-p65 are slightly different. After 5 days of DSS treatment, p-STAT3 seems to decrease at a constant rate (Figure 2F), while p-p65 declines at first rapidly and then gradually (Figure 2C). When quantifying the experimental results, ilastik, an interactive machine learning for (bio)image analysis, was used to binarily classify the images firstly (Berg et al., 2019). Then the objects in the processed images were statistically analyzed with ImageJ which was widely used in the biological sciences and other projects (Schindelin et al., 2012).

MIR31 suppresses p-p65 and p-STAT3, but accelerates the recovery of epithelia in colitis

Dysregulation of the intestinal epithelium homeostasis has been detected in IBD (Podolsky, 2002). The intestinal epithelium

10.3389/fmicb.2022.1089729

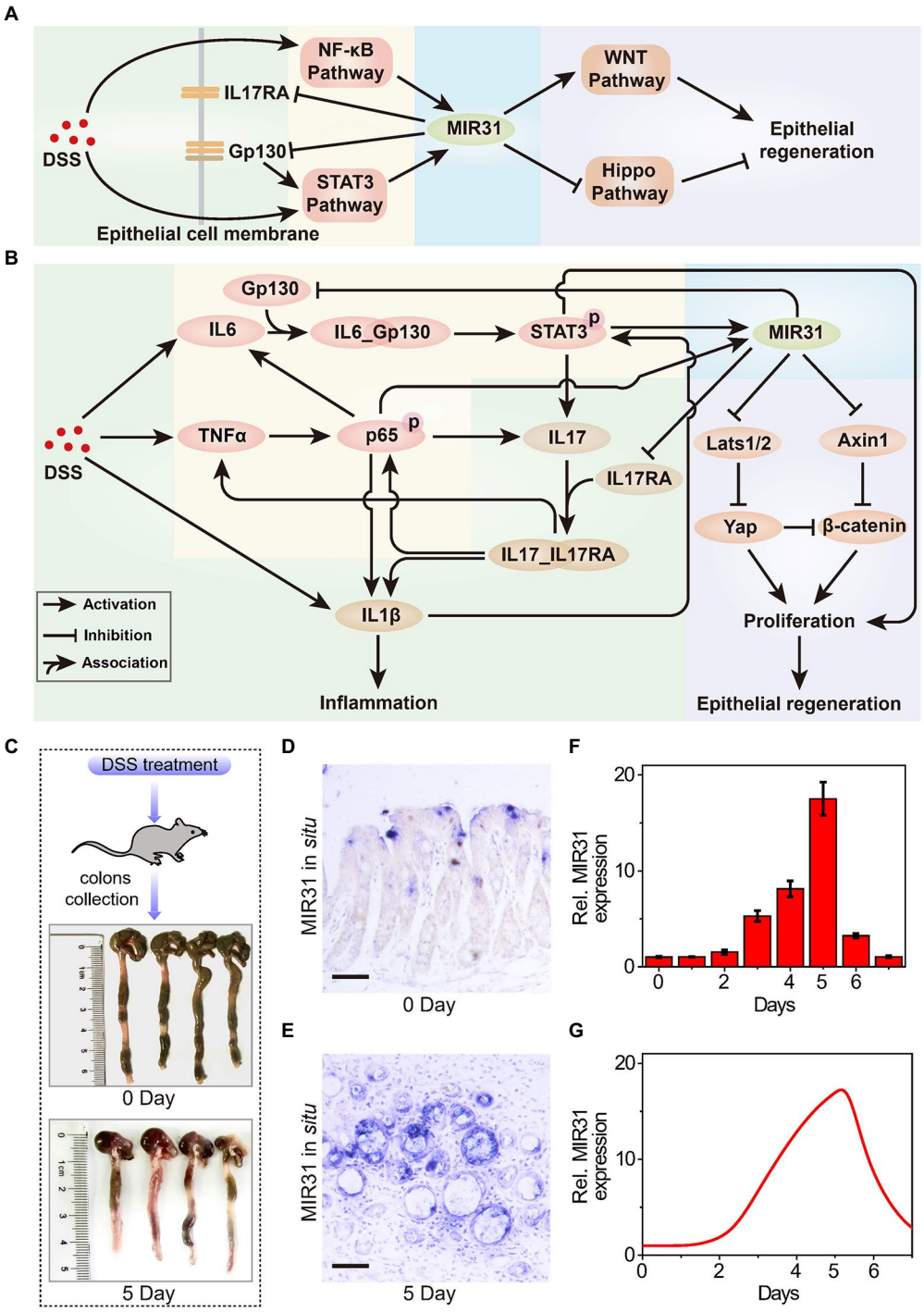

FIGURE 1
The role of MIR31 in DSS-induced colitis signaling network. **(A)** Simplified signal transduction network of DSS-induced colitis to show the main interactions between MIR31 and four pathways of NF-κB, STAT3, WNT and Hippo signals. **(B)** Detailed signal transduction network. The network is composed of four modules, which are highlighted by different backgrounds. The green background is for the module of DSS-induced inflammatory cytokine production, the yellow for the module of MIR31 induction, the blue for the module of MIR31-inhibited inflammatory cytokine production, and the purple for the module of MIR31-promoted epithelial regeneration. **(C)** WT mice used in our experiments and colonic tissues collected from WT mice at 0day and 5day of DSS treatment. **(D)** *In situ* hybridization for MIR31 in normal colons without DSS treatment. **(E)** *In situ* hybridization for MIR31 in colons from mice treated with DSS for 5days. **(F)** qRT–PCR analysis showing MIR31 expression levels in the colonic epithelium from DSS-treated mice at the indicated time points. n=4 at each time point. **(G)** Simulation results of the MIR31 expression level over time in DSS-induced WT model.

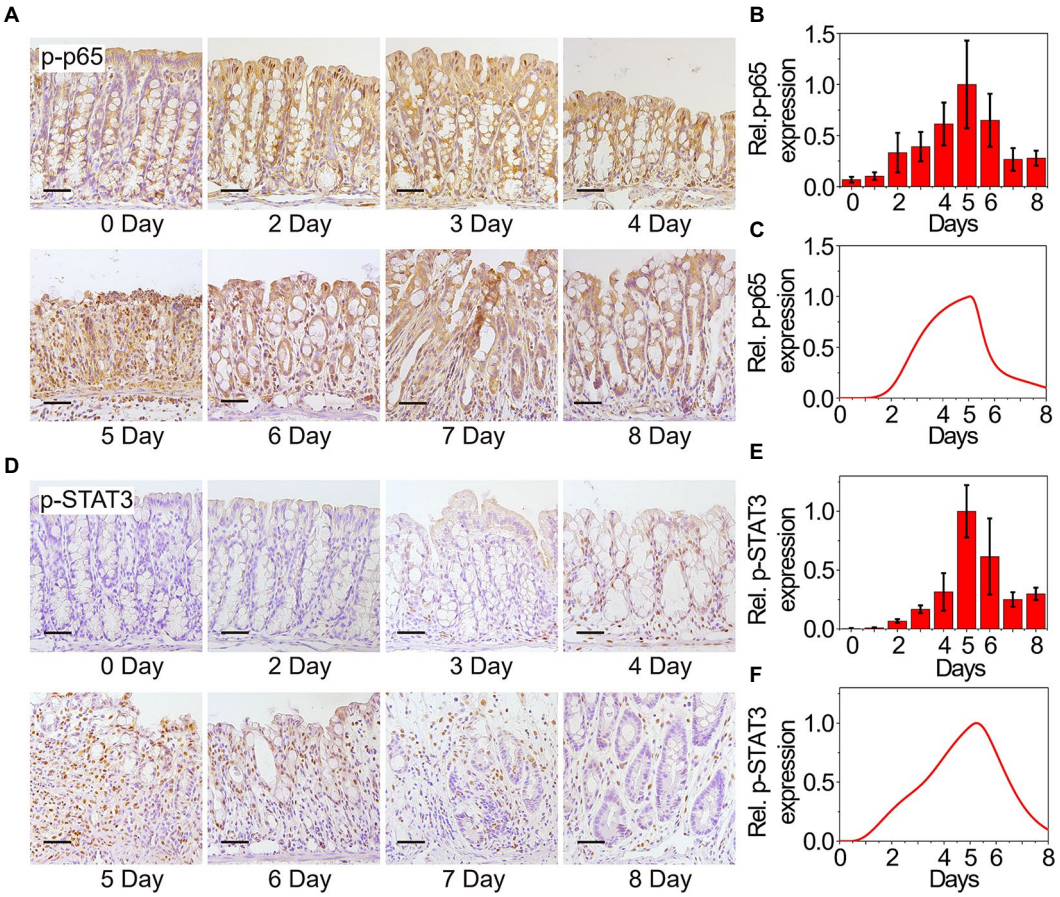

FIGURE 2
Experimental and modeling results of p65 and STAT3 activation. **(A)** Immunohistochemistry for p65 in the colons from WT mice at the indicated time points following DSS treatment. n=3 at each time point. **(B)** Statistical histogram of experimental immunohistochemistry results for p65 in panel **(A)**. **(C)** Simulation results of p-p65 expression in WT model over time after DSS treatment. P65 is phosphorylated when it is transferred to the nucleus. **(D)** Immunohistochemistry for p-STAT3 in the colons from WT mice at the indicated time points following DSS treatment. *n*=3 at each time point. **(E)** Statistical histogram of experimental immunohistochemistry results in panel **(D)**. **(F)** Simulation results of p-STAT3 expression in WT model over time after DSS treatment. Here, the results in **(B,C,E,F)** are normalized to the corresponding maximum of p-p65 and p-STAT3, respectively.

homeostasis depends on the IECs, the intestinal microbiota, and the intestinal immune system (Hooper and Macpherson, 2010), in which the IECs provide a mucosal barrier to segregate host immune system and commensal bacteria (Roda et al., 2010; Peterson and Artis, 2014; Okumura and Takeda, 2017). The proliferation and apoptosis of IECs are also studied to evaluate the clinical symptoms in colitis (Renes et al., 2002). Thus, the number change of proliferative cells is considered to characterize the inflammatory process and the recovery of epithelia in colitis. Figure 3A shows the double immunofluorescence of the proliferative cells at different time points in WT mice upon DSS treatment. For double immunofluorescence of Ki67 and β-catenin in the colons, we count the number of proliferative cells per crypt. The quantified data indicates that the number of proliferative cells declines first and then recovers to the initial state (Figure 3B). Dynamics of the proliferative cells can also be well reproduced by our model (Figure 3C), confirming that our model has the potential for exploring the signaling properties and giving mechanistic insights of MIR31 in colitis.

We next apply our model to quantitatively dissect the functional roles of MIR31 in colitis. Simulation results predict that deletion of MIR31 in mice increases the peak value of p-p65 compared to the WT mice (Figure 4A). Immunofluorescence experimental analysis for p-p65 is performed in MIR31-KO mice to validate the prediction (Figure 4B). As the quantified results shown (Figure 4C), deletion of MIR31 indeed amplifies the "adaptation" behavior of p-p65, revealing that MIR31 suppresses the activation of p65 in colitis. Similar predictions and experimental validations of p-STAT3 suppressed by MIR31 are also determined (Figures 4D,F). Thus, both the activation of p65 and STAT3 are restrained by MIR31 in colitis. The role of MIR31 in mediating the proliferative cell number is investigated as well. As shown in Figure 4G, our model predicts that the proliferative cell number decreases faster

10.3389/fmicb.2022.1089729

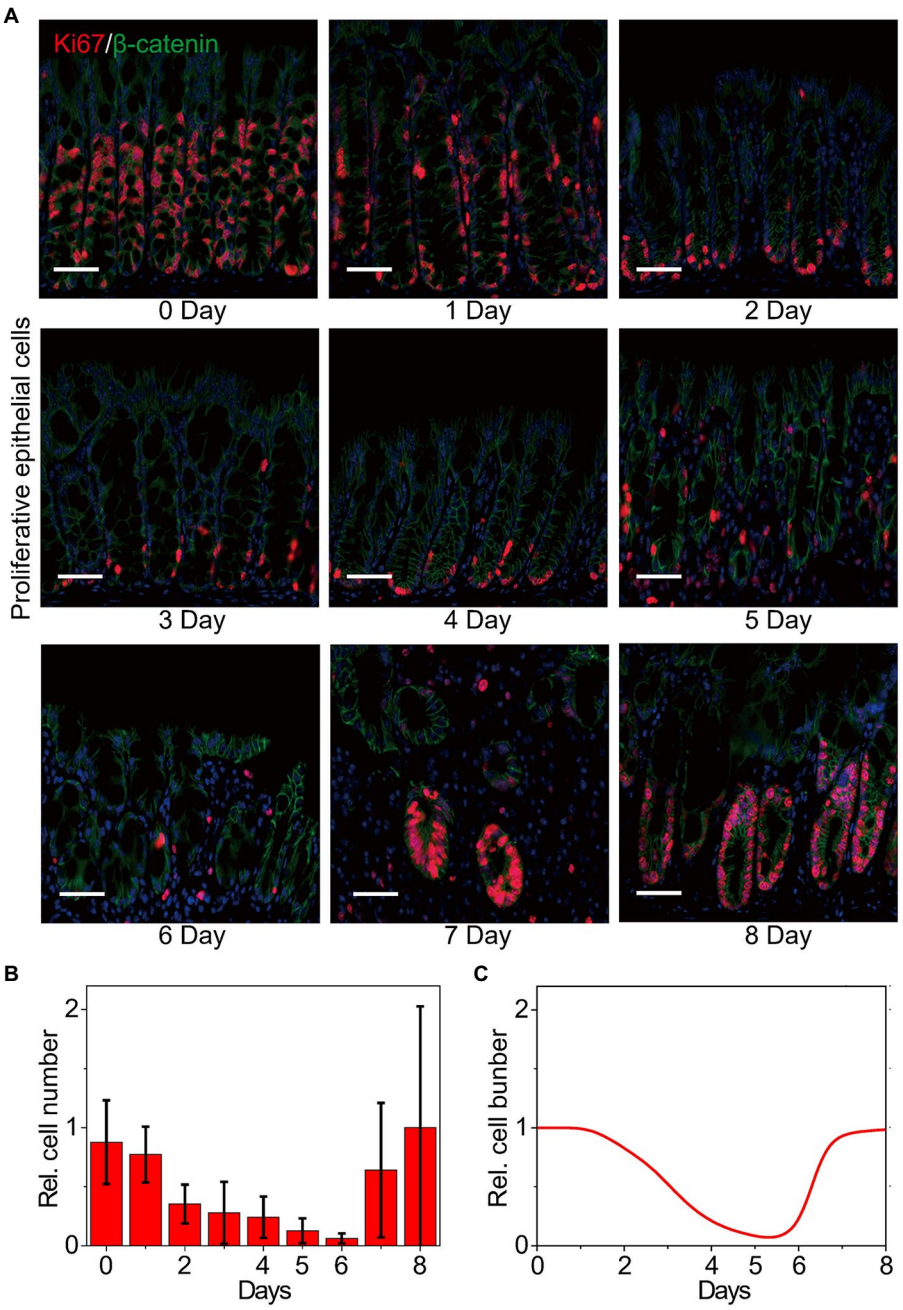

FIGURE 3
Dynamics of the number of proliferative epithelial cells. **(A)** Double immunofluorescence for Ki67 and β-catenin in colons from WT littermates at the indicated time points following DSS treatment. *n*=3 at each time point. **(B)** Statistical histogram of experimental results for the proliferative cell number in panel **(A)**. **(C)** Simulation results of the proliferative cell number in WT model over time after DSS treatment. Here, the cell numbers of proliferative epithelial cells are normalized by the maximum value.

and recovers slower in the absence of MIR31. Analysis of immunofluorescence for Ki67 and β-catenin in MIR31-KO and WT mice is performed (Figure 4H) and our prediction matches well with the quantified experimental results (Figure 4I). Hence, above observations determine that MIR31 regulates inflammatory response through suppressing p65 and STAT3 activation, but promoting the recovery of epithelia during colitis.

MIR31-involved reactions in determining proliferative epithelial cell number

To dissect whether and how MIR31 mediate or is mediated by the transducers in colitis, we further analyze the effect of MIR31-involved reactions on the proliferative epithelial cell number, including the inhibitions of MIR31 on Gp130, IL17RA, Axin1 and

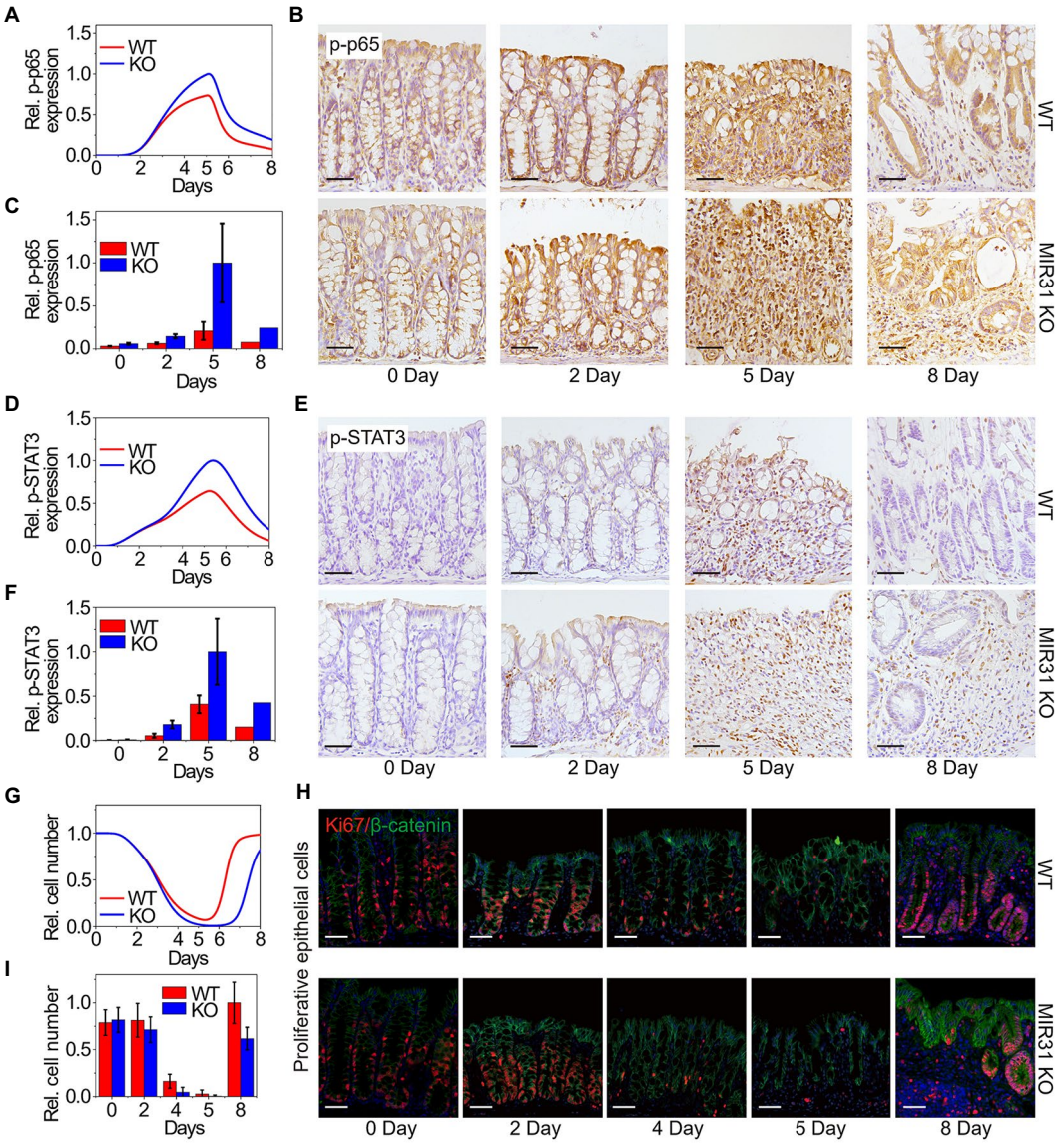

FIGURE 4
Modeling predictions and experimental confirmations in WT and MIR31-KO mice. **(A,D,G)** Model predictions of p65 activation **(A)**, STAT3 activation **(D)** and the proliferative cell number **(G)** for the WT and MIR31-KO models over time after DSS treatment. **(B,E,H)** Immunofluorescence for p65 **(B)**, p-STAT3 **(E)**, and Ki67/β-catenin **(H)** in colons from WT and MIR31-KO mice at the indicated time points following DSS treatment. *n*=3 at each time point. **(C,F,I)** Corresponding statistical histograms of experimental results for p65 in panel **(B)**, for p-STAT3 in panel **(E)**, and for the proliferative cell number in panel **(H)**, respectively. All the results in **(A,C,D,F,G,I)** are normalized to the corresponding maximum values.

Lats1/2 (represented by parameters γ_{GpMIR}, $\gamma_{I17RMIR}$, γ_{AxiMIR} and γ_{LatMIR}, respectively), and the promotions of p-p65 and p-STAT3 on MIR31 (represented by k_{MIRp65} and k_{MIRpST}, respectively) as shown in Figure 5A. We scale the parameters by multiplying the factor of λ_i and define the change ratio $\delta_i(t)$ of the proliferative cell number as $\delta_i(t) = |N_{\lambda i}(t)\text{-}N(t)|/N(t)$. $N_{\lambda i}(t)$ is the proliferative cell number in the modified model with the scaling factor of λ_i, i represents the corresponding reaction parameters, and $N(t)$ is the proliferative cell number in the WT model.

We first discuss the effect of $\gamma_{I17RMIR}$ (the inhibition strength of MIR31 on IL17RA) on the proliferative cell number. As shown in Figure 5B, when the inhibition strength $\gamma_{I17RMIR}$ is increased by 10 times, the proliferative cell number increases significantly, indicating that the inhibition of MIR31 on IL17RA exhibits a strong impact on the inflammatory process. Effects of all the six MIR31-involved reactions on the proliferative cell number are discussed. The corresponding strengths change from 0.01 fold to 100 fold by tuning the scaling factor λ_i from 0.01 to 100, and the

variations of $\delta_i(t)$ at 5 days of DSS treatment are shown in Figure 5C. One can see that the reaction of MIR31 inhibiting IL17RA ($\gamma_{I17RMIR}$), and the reactions of p-p65 (k_{MIRp65}) and p-STAT3 (k_{MIRpST}) promoting MIR31 exhibit significant effects on the changes of proliferative epithelial cell number. The impacts of enhanced strengths on the proliferative cell number are greater than those of decreased strengths. While the variation of the other reactions, i.e., the inhibitions of MIR31 on Gp130, Axin1, and Lats1/2 (γ_{GpMIR}, γ_{AxiMIR} and γ_{LatMIR}), barely influence the proliferative cell number (Figure 5C). Dynamic evolutions in the change ratio of the proliferative cell number $\delta_i(t)$ as a function of time when the scaling factor λ_i is varied continuously are studied and shown in Figures 5D–I, which further indicate that the parameters such as $\gamma_{I17RMIR}$, k_{MIRp65} and k_{MIRpST} are still important on time scales, while the parameters of γ_{GpMIR}, γ_{AxiMIR} and γ_{LatMIR} still exhibit little impacts. These results further display the significant roles of the reaction of MIR31 inhibiting IL17RA (Figure 5D), and the reactions of p-p65 (Figure 5E) and p-STAT3 (Figure 5F) promotion on MIR31 on the proliferative epithelial cells.

Competition of MIR31 for inflammation inhibition and regeneration promotion

Previous studies indicate that the DSS-induced inflammation response is concentration dependent (Naito et al., 2003; Perše and Cerar, 2012), while the underlying regulatory mechanism remains unclear. To address this issue, quantitative analysis of the influence of DSS concentration in colitis is performed. Figures 6A–C show the dynamics of MIR31, IL1β, and the proliferative cell number, under continuous variation in DSS concentrations. Low concentrations (<2.5% wt/vol) of DSS treatment barely affect the signaling dynamics (Figures 6A–C). With high DSS concentrations, MIR31 is increased rapidly (Figure 6A), leading to the secretion of the inflammatory cytokine IL1β and inflammation induction during the first 5 days (Figure 6B). The induction of inflammation results in a significant decrease of the proliferative cell number (Figure 6C).

Since MIR31 can both suppress inflammation and promote epithelial regeneration (Figure 4), the competition of the two

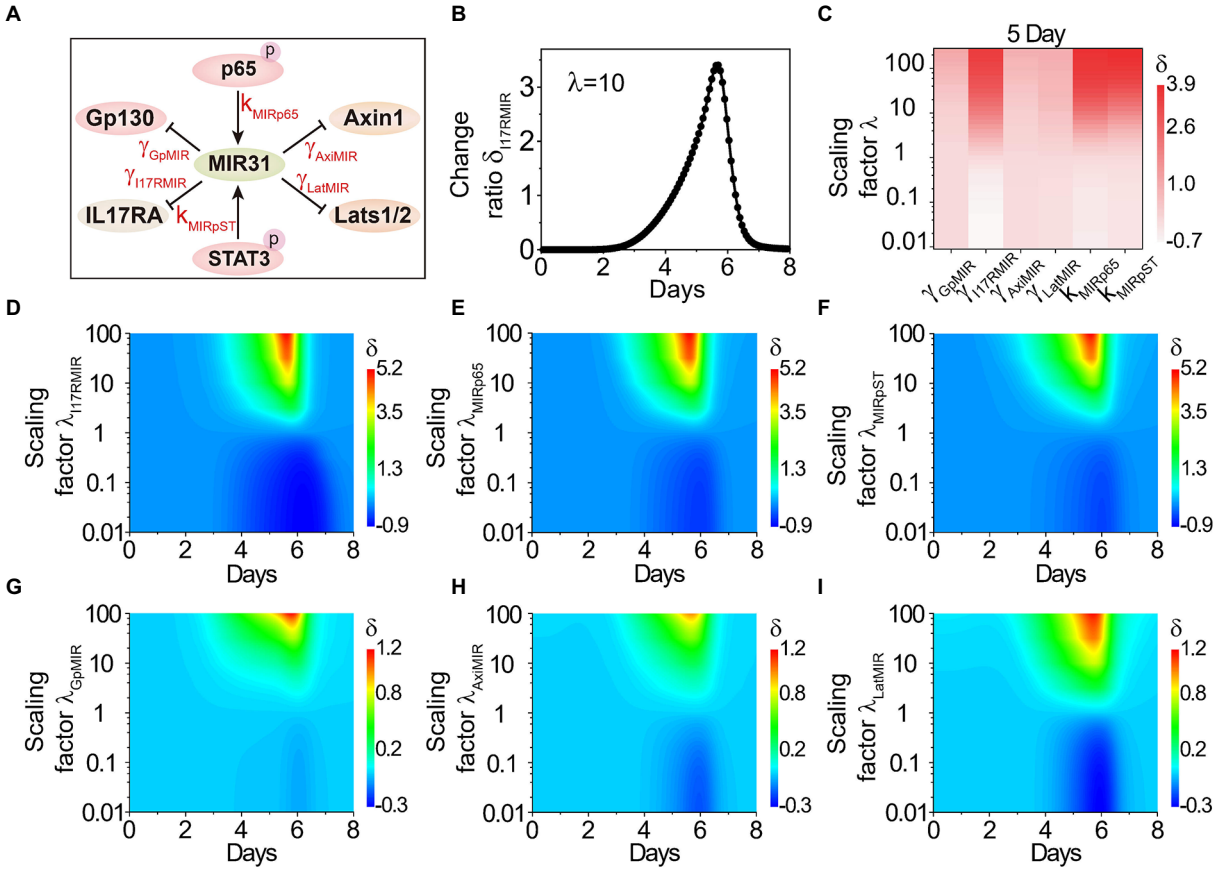

FIGURE 5
The effects of MIR31-involved reactions on the proliferative cell number. **(A)** The relationship between MIR31 and the related six proteins Gp130, IL17RA, Axin1, Lats1/2, p-p65 and p-STAT3. **(B)** The change ratio δ of the proliferative cell number varies with the scaling factor λ=10 for the inhibition of MIR31 on IL17RA (i.e., $\gamma_{I17RMIR}$ increases with 10 times). **(C)** The variations δ_i**(t)** for the inhibitions of MIR31 on Gp130, IL17RA, Axin1, and Lats1/2 (γ_{GpMIR}, $\gamma_{I17RMIR}$, γ_{AxiMIR}, γ_{LatMIR}), and the promotions of p-p65 and p-STAT3(k_{MIRp65} and k_{MIRpST}) on MIR31 with the scaling factor λ at 5days of DSS treatment. **(D–I)** The change ratio $\delta_i(t)$ as a function of time under continuous changes in the scaling factors λ_i for all six parameters.

functional roles for MIR31 as well as the influence of DSS concentration on such competition are subsequently discussed. There are four major MIR31 target proteins, Gp130, IL17RA, Axin1, and Lats1/2, with certain probabilities of acting on inflammation and regeneration in colitis (Figure 1B). The probability of MIR31 binding to Gp130 and IL17RA are defined as *p1* and *p2*, which mainly contribute to inflammatory responses (Figure 6D). The probability of binding to Axin1 and Lats1/2 are defined as *p3* and *p4*, which mediates the epithelial regeneration. The constraint is $p1(t)+p2(t)+p3(t)+p4(t)=1\text{-}p0(t)$, where *p0* represents the probability of MIR31 in a resting state without any binding. MIR31 inhibits inflammation with a probability of $p1+p2$ and promotes regeneration with a probability of $p3+p4$. Dynamics of the MIR31 binding probability assignment for inhibiting inflammation and promoting regeneration are plotted in Figure 6E.

In healthy conditions without DSS (0 days), the main function of MIR31 is to promote regeneration. With DSS administration, the binding probability of MIR31-inhibited inflammation gradually increases (Figure 6E, red line) and the binding probability of MIR31-promoted regeneration conversely decreases (Figure 6E, blue line), indicating that the major function of MIR31 changes from regeneration promotion to inflammation inhibition. After DSS is withdrawn, the binding probability of MIR31-inhibited inflammation decreases, and correspondingly, the binding probability of MIR31-promoted regeneration increases. As shown in Figure 6E, the two binding probabilities intersect at the 3rd day and 10th day with intersections D1 and D2. This indicates a larger binding probability of MIR31 for inflammation inhibition than for regeneration promotion during the 3rd to the 10th days to prevent the inflammation caused by DSS.

To discuss the influence of DSS concentration on the competition of MIR31-inhibited inflammation and MIR31-promoted regeneration, the time corresponding to D1 and D2 with the change in DSS concentration is studied. As the result shows in Figure 6F, when the DSS concentration is smaller than 1.1% wt/vol, the binding probability of MIR31-promoted regeneration is typically larger than the binding probability of MIR31-inhibited inflammation without intersection, suggesting that the major function of MIR31 is to promote regeneration in weak DSS-induced colitis. The probability of MIR31-inhibited inflammation increases gradually as the DSS concentration increases. In the case of strong DSS-induced colitis (DSS > 3.0% wt/vol), the major function of MIR31 changes to inhibiting inflammation rather than to promoting regeneration (Figure 6F).

MIR31 expression ([MIR31]) increases obviously after DSS treatment (Figure 1F). Thus, the action strength of MIR31-inhibited inflammation can be defined as $F1(t)=[MIR31(t)]\times(p1(t)+p2(t))$, and the action strength of MIR31-promoted regeneration as $F2(t)=[MIR31(t)]\times(p3(t)+p4(t))$, giving the total action of MIR31 on the system as $F(t)=F0(t)+F1(t)+F2(t)$, where $F0(t)=[MIR31(t)]\times p0(t)$ represents the action strength of MIR31 in a resting state without any binding. Interestingly, the action strength of MIR31-inhibited inflammation (*F1*) increases significantly after DSS treatment, while the action strength of MIR31-promoted regeneration (*F2*) shows less enhancement in Figure 6G. The maximal action strength is obtained at approximately the 5th day.

Considering the influence of DSS concentration on the action strengths of MIR31, as shown in Figure 6H, the maximal action strengths of MIR31 remain at low levels when the DSS concentration is small, while they increase rapidly with increasing DSS concentration when DSS is larger than 3.0% wt/vol, especially for the maximal action strength of MIR31-inhibited inflammation ($F1_{max}$). Surprisingly, we found that the probability of MIR31-promoted regeneration decreases in a stepwise manner with increasing DSS concentration (Figure 6I, blue line), which is contrary to the trend of the action strength of MIR31-promoted regeneration (Figure 6H, blue line). Further analysis determines the reason for this to be the corresponding increase in MIR31 expression as shown in Figure 6J being much greater than the decrease in the probability of MIR31-promoted regeneration (Figure 6I). Thus, the changes in the binding probabilities of MIR31 determine the transformation of its functions, while the expression level of MIR31 determines the action strengths of its functions for inhibiting inflammation and promoting regeneration.

Discussion

Previous studies demonstrated that miRNAs are associated with various diseases including COVID-19 (Li et al., 2020; de Gonzalo-Calvo et al., 2021) and have the potential to be therapeutic targets (Guterres et al., 2020; Hum et al., 2021). MIR31 is identified as a key regulator in diseases (Valastyan and Weinberg, 2010; Liu et al., 2010a; Laurila and Kallioniemi, 2013; Stepicheva and Song, 2016). MIR31 acts as an oncogenic miRNA in lung cancer by targeting specific tumor suppressors for repression (Liu et al., 2010b). In addition, MIR31 is proven to be a target for inhibiting tumor growth and metastasis (Tian C. et al., 2017). The down-regulation of MIR31 disrupts cellular homeostasis and promotes the evolution and progression of prostate cancer (Lin et al., 2013b). The pathogenesis of colitis involves many complex signaling pathways that are related to various types of cells at tissue level (Perše and Cerar, 2012). Understanding the mechanism of MIR31 in colitis is therefore urgently needed for developing therapies for diseases.

Combining experimental analysis and a proposed phenomenological network model, we quantitatively explored the important roles of MIR31 in modulating inflammation and epithelial regeneration, and identified effective targets for clinical treatment of inflammation. Our study indicates that MIR31 exhibits an "adaptation" behavior in WT model of DSS-induced colitis and similar "adaptation" behavior also occurs in p-p65 and p-STAT3. The number of proliferative cells decreases gradually and then recovery to normal state after DSS treatment (Figure 3). In MIR31 KO model, the "adaptation" behavior of p-p65 and p-STAT3 is magnified, indicating the suppression of MIR31 on

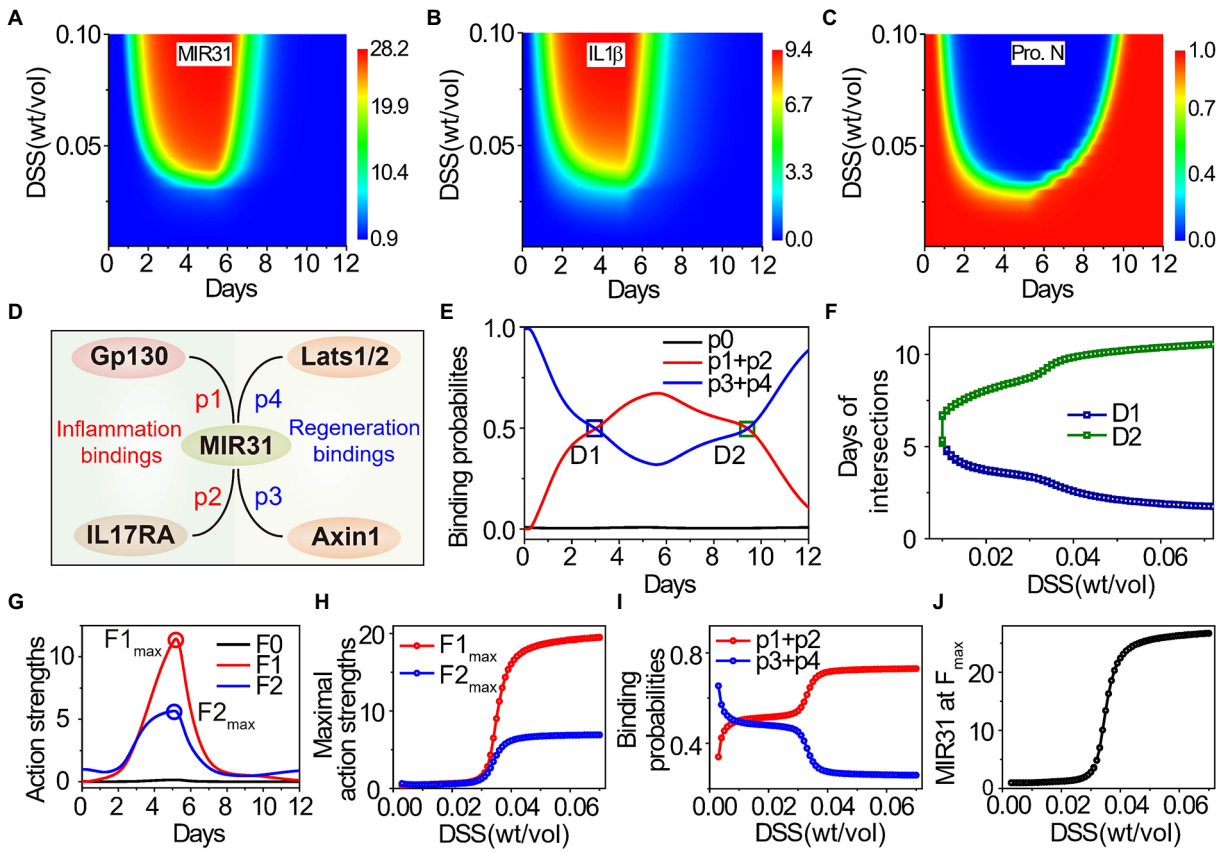

FIGURE 6
Effects of DSS concentration on inflammatory response and epithelial regeneration. **(A–C)** Dynamic results of MIR31 expression **(A)**, IL1β expression **(B)**, and the proliferative cell number **(C)** under continuous changes in DSS concentration. **(D)** The probabilities of MIR31 binding to Gp130 (*p1*), IL17RA (*p2*), Axin1 (*p3*), and Lats1/2 (*p4*). **(E)** Dynamics of the probability assignment of MIR31, where *p1+p2* and *p3+p4* correspond to MIR31-inhibited inflammation probability and MIR31-promoted regeneration probability, respectively, and *p0* represents the probability that MIR31 is in a resting state. **(F)** The influence of DSS concentration on two intersections D1 and D2 of the two probability curves in **(E)**. **(G)** Dynamics of the action strengths of MIR31-inhibited inflammation and MIR31-promoted regeneration. The influence of DSS concentration on the maximal action strengths **(H)** and on the probabilities of MIR31-inhibited inflammation and MIR31-promoted regeneration **(I)** and the MIR31 expression **(J)** corresponding to the two maximal action strengths in **(G)**. The results of **(C)** are compared with the maximum proliferative cell number. In **(E,G)**, MIR31 is divided into three categories: MIR31 in the resting state (black), MIR31 inhibiting inflammation (red) and MIR31 promoting regeneration (blue).

the activation of NF-κB and STAT3 signaling pathways. The number of proliferative cells decreases more quickly with fewer survival cells and recovers slower, suggesting a promotion of MIR31 on epithelial regeneration (Figure 4). As a novel therapeutic target, MIR31 has been extensively studied in various diseases such as colorectal cancer (Zhao et al., 2020), nasopharyngeal carcinoma (Wu et al., 2016). PEX5, a novel target of MIR31, is also proven to be a therapeutic option in hepatocellular carcinoma (Wen et al., 2020). Our analysis shows that the inhibition of MIR31 on IL17R, the promotions of p-p65 and p-STAT3 on MIR31 exhibit virtual influences on the number of proliferative cells, which can be considered as a potential therapeutic target in future studies.

To intuitively present the mechanism of MIR31, we propose that the MIR31 response process in colitis can be characterized by the "spring-like effect" (Figure 7A). In this analogy, DSS acts as the external pressure on the spring, the number of prolifertive cells is the spring length, and the MIR31 action strength, which involves MIR31-promoted regeneration and MIR31-inhibited inflammation, is the intrinsic spring force. With such a view, DSS drives the system into the "spring compressing process" and thus the cell number decreases. Meanwhile, MIR31 expression increases gradually to prevent DSS-induced colitis (Figure 7B). When DSS remains high, the action strength of MIR31 on inflammation also remains strong with a small cell number. Even immediately after DSS is withdrawn, the residual DSS in the system is still strong, which leads to low cell numbers. We name this state the "spring compression state." Then, DSS becomes attenuated while the action strength of MIR31 on inflammation inhibition still holds dominant, resulting in a rapid increase in cell number. This process corresponds to the "spring recovery process," in which MIR31 expression decreases, leading to a gradual reduction in the action strength of MIR31. Finally, the system returns to the normal state. Note that in the MIR31-KO model,

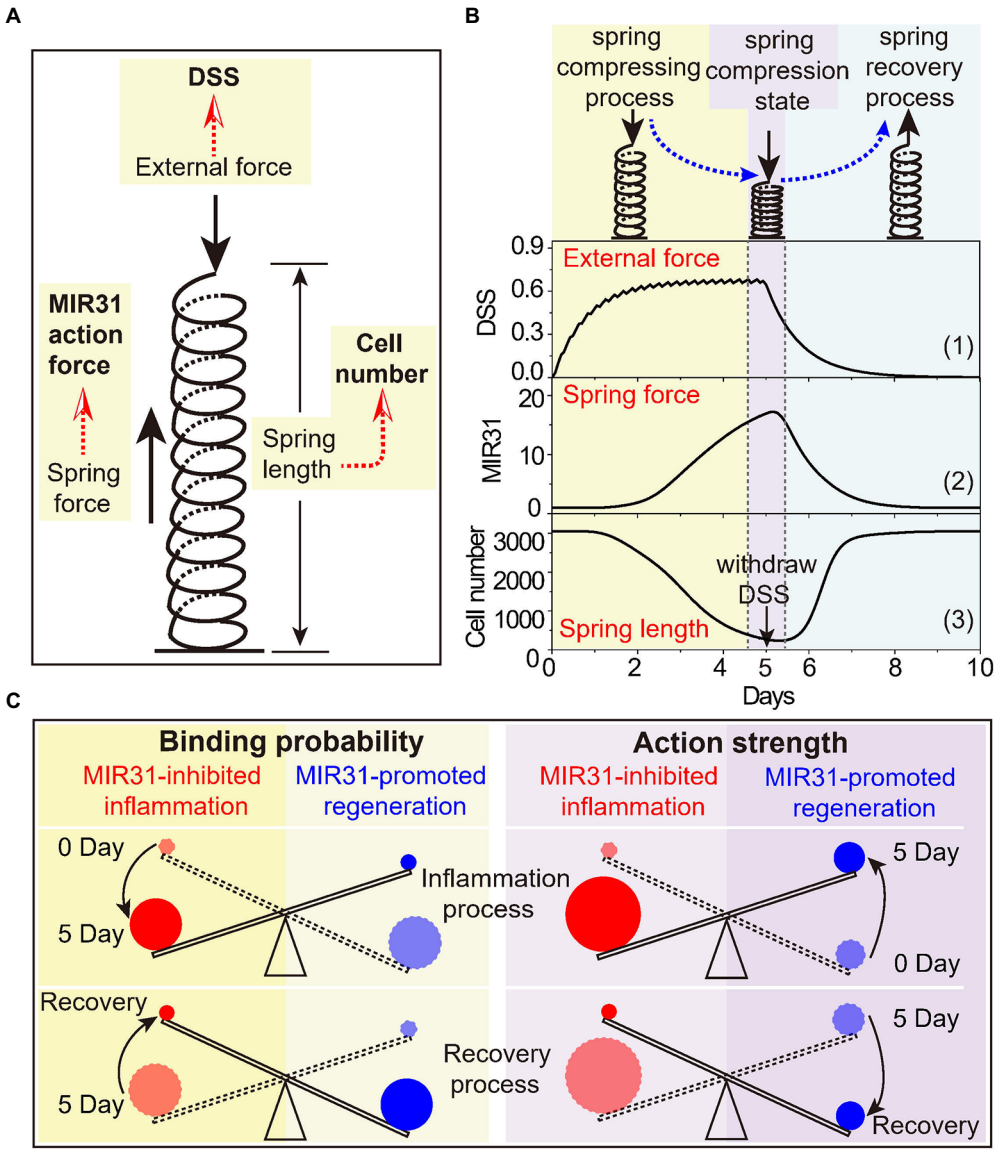

FIGURE 7
The spring-like effect and the seesaw model of MIR31 in balancing inflammatory and regenerative responses. **(A)** Schematic diagram of the spring-like effect of inflammatory response to DSS treatment. **(B)** The specific correspondence between the inflammatory response process to DSS and the spring system. Considering the trend of MIR31 expression trend was consistent with that of MIR31 action strength, it was used to characterize the trend of the MIR31 action strength. **(C)** The seesaw model of MIR31 competition mechanism in the WT model from the viewpoints of the binding probability (left) and the action strength of MIR31 (right).

the inflammatory response process can also be described by the "spring-like effect," while the spring compressing process (compression state and recovery process) occurs on a time scale shorter (longer) than that in the WT model.

In our study, we simply compared the inflammatory response to "spring-like effect" as the response presents a process similar to that of a spring from compression to recovery. The inflammatory response is a nonlinear process that depends on time and external force. Actually, the process can also be regarded as a visco-elastic system, such as the "spring-dashpot model." Recently, a visco-elastic system has been proposed in which the cytoplasm contributes to mitotic spindle positioning through its visco-elastic property (Xie et al., 2022). Besides, a commentary also defines this visco-elastic property as the "spring-like behavior" (Bai and Mitchison, 2022).

Due to the competitive binding mechanism for MIR31 to either inhibit inflammation or promote regeneration, as well as the influence of MIR31 level on the competition, we suggest that such a competition mechanism can be understood by a "seesaw model" (Figure 7C). The seesaw is balanced by MIR31-inhibited inflammation and MIR31-promoted regeneration. Based on the

concept of the binding probabilities, the change in the seesaw is induced by the change of MIR31 binding probabilities to different proteins (Figure 6E). In detail, MIR31 displays an increased binding probability for inflammation inhibition and a decreased binding probability for regeneration promotion during the inflammation process (0 day–5 day). During the recovery process, there is a diminution in MIR31 binding probability for inflammation inhibition and an enhancement of MIR31 binding probability for regeneration promotion (5 day–10 day). Notably, a different mechanism can be obtained in the seesaw from the viewpoint of the MIR31 action strength. Both the action strengths of MIR31-promoted regeneration and MIR31-inhibited inflammation increase with time during the inflammation process, with the action strength of MIR31-inhibited inflammation gradually becoming dominant (0 day–5 day). However, during the recovery process, both the action strengths of MIR31-promoted regeneration and MIR31-inhibited inflammation decrease with time, with the action strength of MIR31-promoted regeneration finally becoming dominant (5 day–10 day).

In summary, the seesaw model and the spring-like effect for MIR31 functions highlight the importance of MIR31 in the inflammatory response process. MIR31 can effectively alleviate inflammation by inhibiting inflammatory cytokine receptors and can promote epithelial regeneration by modulating the WNT and Hippo signaling pathways. With the model, we suggest that the inhibition of MIR31 on cytokine receptors is crucial to inflammation control and can be regarded as a therapeutic target for drug design.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding authors.

Ethics statement

The animal study was reviewed and approved by the Institutional Animal Care and Use Committee of China Agricultural University.

Author contributions

JQ developed the model and performed simulation. JQ and XL wrote the manuscript. CS and YT designed the experiments. CS implemented the experiments. JQ, CS, XL, and YY performed discussion of experimental data and modeling results. YY, YW, WL, and ZhiY helped for model discussion. XL, ZheY, and JS revised the paper, conceived the idea, and supervised the project. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Natural Science Foundation of China (Grant Nos. 12090052, 11874310, 81772984, and 82025006), the National Science and Technology Major Project of the Ministry of Science and Technology of China (Grant Nos. 2021ZD0201900 and 2021ZD0201904), and the Fujian Province Foundation (Grant No. 2020Y4001).

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.1089729/full#supplementary-material

References

Alon, U. (2006). *An Introduction to Systems Biology: Design Principles of Biological Circuits*, London: Chapman and Hall/CRC.

Bai, L., and Mitchison, T. J. (2022). Spring-like behavior of cytoplasm holds the mitotic spindle in place. *Proc. Natl. Acad. Sci. U. S. A.* 119:e2203036119. doi: 10.1073/pnas.2203036119

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cells* 116, 281–297. doi: 10.1016/S0092-8674(04)00045-5

Bartel, D. P. (2009). MicroRNAs: target recognition and regulatory functions. *Cells* 136, 215–233. doi: 10.1016/j.cell.2009.01.002

Becker, C., Neurath, M. F., and Wirtz, S. (2015). The intestinal microbiota in inflammatory bowel disease. *ILAR J.* 56, 192–204. doi: 10.1093/ilar/ilv030

Béres, N. J., Kiss, Z., Sztupinszki, Z., Lendvai, G., Arat, A., Sziksz, E., et al. (2017). Altered mucosal expression of microRNAs in pediatric patients with inflammatory bowel disease. *Dig. Liver Dis.* 49, 378–387. doi: 10.1016/j.dld.2016.12.022

Berg, S., Kutra, D., Kroeger, T., Straehle, C. N., Kausler, B. X., Haubold, C., et al. (2019). Ilastik: interactive machine learning for (bio)image analysis. *Nat. Methods* 16, 1226–1232. doi: 10.1038/s41592-019-0582-9

Cader, M. Z., and Kaser, A. (2013). Recent advances in inflammatory bowel disease: mucosal immune cells in intestinal inflammation. *Gut* 62, 1653–1664. doi: 10.1136/gutjnl-2012-303955

Carthew, R. W., and Sontheimer, E. J. (2009). Origins and mechanisms of miRNAs and siRNAs. *Cells* 136, 642–655. doi: 10.1016/j.cell.2009.01.035

Dalal, S. R., and Chang, E. B. (2014). The microbial basis of inflammatory bowel diseases. *J. Clin. Invest.* 124, 4190–4196. doi: 10.1172/JCI72330

Dalal, S. R., and Kwon, J. H. (2010). The role of MicroRNA in inflammatory bowel disease. *Gastroenterol Hepatol (N Y)* 6, 714–722. doi: 10.7150/ijbs.59904

De Gonzalo-Calvo, D., Benítez, I. D., Pinilla, L., Carratalá, A., Moncusí-Moix, A., Gort-Paniello, C., et al. (2021). Circulating microRNA profiles predict the severity of COVID-19 in hospitalized patients. *Transl. Res.* 236, 147–159. doi: 10.1016/j.trsl.2021.05.004

De Robertis, M., Massi, E., Poeta, M. L., Carotti, S., Morini, S., Cecchetelli, L., et al. (2011). The AOM/DSS murine model for the study of colon carcinogenesis: from pathways to diagnosis and therapy studies. *J. Carcinog.* 10:9. doi: 10.4103/1477-3163.78279

Deng, F., Peng, L., Li, Z., Tan, G., Liang, E., Chen, S., et al. (2018). YAP triggers the Wnt/β-catenin signalling pathway and promotes enterocyte self-renewal, regeneration and tumorigenesis after DSS-induced injury. *Cell Death Dis.* 9:153. doi: 10.1038/s41419-017-0244-8

Eichele, D. D., and Kharbanda, K. K. (2017). Dextran sodium sulfate colitis murine model: an indispensable tool for advancing our understanding of inflammatory bowel diseases pathogenesis. *World J. Gastroenterol.* 23, 6016–6029. doi: 10.3748/wjg.v23.i33.6016

Emde, A., and Hornstein, E. (2014). miRNAs at the interface of cellular stress and disease. *EMBO J.* 33, 1428–1437. doi: 10.15252/embj.201488142

Ernst, M., Thiem, S., Nguyen, P. M., Eissmann, M., and Putoczki, T. L. (2014). Epithelial gp130/Stat3 functions: an intestinal signaling node in health and disease. *Semin. Immunol.* 26, 29–37. doi: 10.1016/j.smim.2013.12.006

Fabian, M. R., Sonenberg, N., and Filipowicz, W. (2010). Regulation of mRNA translation and stability by microRNAs. *Annu. Rev. Biochem.* 79, 351–379. doi: 10.1146/annurev-biochem-060308-103103

Fava, F., and Danese, S. (2011). Intestinal microbiota in inflammatory bowel disease: friend of foe? *World J. Gastroenterol.* 17, 557–566. doi: 10.3748/wjg.v17.i5.557

Francescone, R., Hou, V., and Grivennikov, S. I. (2015). Cytokines, IBD, and colitis-associated cancer. *Inflamm. Bowel Dis.* 21, 409–418. doi: 10.1097/MIB.0000000000000236

Garzon, R., Marcucci, G., and Croce, C. M. (2010). Targeting microRNAs in cancer: rationale, strategies and challenges. *Nat. Rev. Drug Discov.* 9, 775–789. doi: 10.1038/nrd3179

Giridharan, S., and Srinivasan, M. (2018). Mechanisms of NF-κB p65 and strategies for therapeutic manipulation. *J. Inflamm. Res.* 11, 407–419. doi: 10.2147/JIR.S140188

Guterres, A., De Azeredo Lima, C. H., Miranda, R. L., and Gadelha, M. R. (2020). What is the potential function of microRNAs as biomarkers and therapeutic targets in COVID-19? Infection. *Genet. Evol.* 85:104417. doi: 10.1016/j.meegid.2020.104417

Ha, M., and Kim, V. N. (2014). Regulation of microRNA biogenesis. *Nat. Rev. Mol. Cell Biol.* 15, 509–524. doi: 10.1038/nrm3838

Hammond, S. M. (2015). An overview of microRNAs. *Adv. Drug Deliv. Rev.* 87, 3–14. doi: 10.1016/j.addr.2015.05.001

Hooper, L. V., and Macpherson, A. J. (2010). Immune adaptations that maintain homeostasis with the intestinal microbiota. *Nat. Rev. Immunol.* 10, 159–169. doi: 10.1038/nri2710

Hum, C., Loiselle, J., Ahmed, N., Shaw, T. A., Toudic, C., and Pezacki, J. P. (2021). MicroRNA mimics or inhibitors as antiviral therapeutic approaches against COVID-19. *Drugs* 81, 517–531. doi: 10.1007/s40265-021-01474-5

Kalla, R., Ventham, N. T., Kennedy, N. A., Quintana, J. F., Nimmo, E. R., Buck, A. H., et al. (2015). MicroRNAs: new players in IBD. *Gut* 64, 504–513. doi: 10.1136/gutjnl-2014-307891

Kany, S., Vollrath, J. T., and Relja, B. (2019). Cytokines in inflammatory disease. *Int. J. Mol. Sci.* 20:6008. doi: 10.3390/ijms20236008

Kaplan, G. G., and Ng, S. C. (2017). Understanding and preventing the global increase of inflammatory bowel disease. *Gastroenterology* 152, 313–321.e2. doi: 10.1053/j.gastro.2016.10.020

Kaser, A., Zeissig, S., and Blumberg, R. S. (2010). Inflammatory bowel disease. *Annu. Rev. Immunol.* 28, 573–621. doi: 10.1146/annurev-immunol-030409-101225

Kelley, N., Jeltema, D., Duan, Y., and He, Y. (2019). The NLRP3 Inflammasome: an overview of mechanisms of activation and regulation. *Int. J. Mol. Sci.* 20:3328. doi: 10.3390/ijms20133328

Kiesler, P., Fuss, I. J., and Strober, W. (2015). Experimental models of inflammatory bowel diseases. *Cell. Mol. Gastroenterol. Hepatol.* 1, 154–170. doi: 10.1016/j.jcmgh.2015.01.006

Krol, J., Loedige, I., and Filipowicz, W. (2010). The widespread regulation of microRNA biogenesis, function and decay. *Nat. Rev. Genet.* 11, 597–610. doi: 10.1038/nrg2843

Kumar, P., Monin, L., Castillo, P., Elsegeiny, W., Horne, W., Eddens, T., et al. (2016). Intestinal Interleukin-17 receptor signaling mediates reciprocal control of the gut microbiota and autoimmune inflammation. *Immunity* 44, 659–671. doi: 10.1016/j.immuni.2016.02.007

Lai, X., Gupta, S. K., Schmitz, U., Marquardt, S., Knoll, S., Spitschak, A., et al. (2018). MiR-205-5p and miR-342-3p cooperate in the repression of the E2F1 transcription factor in the context of anticancer chemotherapy resistance. *Theranostics* 8, 1106–1120. doi: 10.7150/thno.19904

Laurila, E. M., and Kallioniemi, A. (2013). The diverse role of miR-31 in regulating cancer associated phenotypes. *Genes Chromosom. Cancer* 52, 1103–1113. doi: 10.1002/gcc.22107

Li, C., Hu, X., Li, L., and Li, J.-H. (2020). Differential microRNA expression in the peripheral blood from human patients with COVID-19. *J. Clin. Lab. Anal.* 34:e23590. doi: 10.1002/jcla.23590

Li, X., Jin, J., Zhang, X., Xu, F., Zhong, J., Yin, Z., et al. (2021a). Quantifying the optimal strategy of population control of quorum sensing network in *Escherichia coli*. *NPJ Syst. Biol. Appl.* 7:35. doi: 10.1038/s41540-021-00196-4

Li, X., Zhang, P., Yin, Z., Xu, F., Yang, Z.-H., Jin, J., et al. (2022). Caspase-1 and Gasdermin D afford the optimal targets with distinct switching strategies in NLRP1b Inflammasome-induced cell death. *Research* 2022:9838341. doi: 10.34133/2022/9838341

Li, X., Zhong, C.-Q., Wu, R., Xu, X., Yang, Z.-H., Cai, S., et al. (2021b). RIP1-dependent linear and nonlinear recruitments of caspase-8 and RIP3 respectively to necrosome specify distinct cell death outcomes. *Protein Cell* 12, 858–876. doi: 10.1007/s13238-020-00810-x

Lin, P.-C., Chiu, Y.-L., Banerjee, S., Park, K., Mosquera, J. M., Giannopoulou, E., et al. (2013b). Epigenetic repression of miR-31 disrupts androgen receptor homeostasis and contributes to prostate cancer progression. *Cancer Res.* 73, 1232–1244. doi: 10.1158/0008-5472.CAN-12-2968

Lin, J., Welker, N. C., Zhao, Z., Li, Y., Zhang, J., Reuss, S. A., et al. (2013a). Novel specific microRNA biomarkers in idiopathic inflammatory bowel disease unrelated to disease activity. *Mod. Pathol.* 27, 602–608. doi: 10.1038/modpathol.2013.152

Liu, C. J., Kao, S. Y., Tu, H. F., Tsai, M. M., Chang, K. W., and Lin, S. C. (2010a). Increase of microRNA miR-31 level in plasma could be a potential marker of oral cancer. *Oral Dis.* 16, 360–364. doi: 10.1111/j.1601-0825.2009.01646.x

Liu, W., Li, X., Qi, H., Wu, Y., Qu, J., Yin, Z., et al. (2021). Biphasic regulation of transcriptional surge generated by the gene feedback loop in a two-component system. *Bioinformatics* 37, 2682–2690. doi: 10.1093/bioinformatics/btab138

Liu, X., Sempere, L. F., Ouyang, H., Memoli, V. A., Andrew, A. S., Luo, Y., et al. (2010b). MicroRNA-31 functions as an oncogenic microRNA in mouse and human lung cancer cells by repressing specific tumor suppressors. *J. Clin. Invest.* 120, 1298–1309. doi: 10.1172/JCI39566

Lv, C., Li, F., Li, X., Tian, Y., Zhang, Y., Sheng, X., et al. (2017). MiR-31 promotes mammary stem cell expansion and breast tumorigenesis by suppressing Wnt signaling antagonists. *Nat. Commun.* 8. doi: 10.1038/s41467-017-01059-5

Ma, W., Trusina, A., El-Samad, H., Lim, W. A., and Tang, C. (2009). Defining network topologies that can achieve biochemical adaptation. *Cells* 138, 760–773. doi: 10.1016/j.cell.2009.06.013

Maloy, K. J., and Powrie, F. (2011). Intestinal homeostasis and its breakdown in inflammatory bowel disease. *Nature* 474, 298–306. doi: 10.1038/nature10208

Matsuoka, K., and Kanai, T. (2015). The gut microbiota and inflammatory bowel disease. *Semin. Immunopathol.* 37, 47–55. doi: 10.1007/s00281-014-0454-4

Mizoguchi, A. (2012). "Animal models of inflammatory bowel disease" in *Progress in Molecular Biology and Translational Science*. ed. P. M. Conn (Cambridge, MA: Academic Press)

Moparthi, L., and Koch, S. (2019). Wnt signaling in intestinal inflammation. *Differentiation* 108, 24–32. doi: 10.1016/j.diff.2019.01.002

Naito, Y., Takagi, T., Handa, O., Ishikawa, T., Nakagawa, S., Yamaguchi, T., et al. (2003). Enhanced intestinal inflammation induced by dextran sulfate sodium in tumor necrosis factor-alpha deficient mice. *J. Gastroenterol. Hepatol.* 18, 560–569. doi: 10.1046/j.1440-1746.2003.03034.x

Neurath, M. F. (2014). Cytokines in inflammatory bowel disease. *Nat. Rev. Immunol.* 14, 329–342. doi: 10.1038/nri3661

Ng, S. C., Shi, H. Y., Hamidi, N., Underwood, F. E., Tang, W., Benchimol, E. I., et al. (2017). Worldwide incidence and prevalence of inflammatory bowel disease

in the 21st century: a systematic review of population-based studies. *Lancet* 390, 2769–2778. doi: 10.1016/S0140-6736(17)32448-0

Nguyen, T. L. A., Vieira-Silva, S., Liston, A., and Raes, J. (2015). How informative is the mouse for human gut microbiota research? *Dis. Model. Mech.* 8, 1–16. doi: 10.1242/dmm.017400

Okumura, R., and Takeda, K. (2017). Roles of intestinal epithelial cells in the maintenance of gut homeostasis. *Exp. Mol. Med.* 49:e338. doi: 10.1038/emm.2017.20

Olaru, A. V., Selaru, F. M., Mori, Y., Vazquez, C., David, S., Paun, B., et al. (2011). Dynamic changes in the expression of MicroRNA-31 during inflammatory bowel disease-associated neoplastic transformation. *Inflamm. Bowel Dis.* 17, 221–231. doi: 10.1002/ibd.21359

Parker, K. H., Beury, D. W., and Ostrand-Rosenberg, S. (2015). Myeloid-derived suppressor cells: critical cells driving immune suppression in the tumor microenvironment. *Adv. Cancer Res.* 128, 95–139. doi: 10.1016/bs.acr.2015.04.002

Perše, M., and Cerar, A. (2012). Dextran sodium Sulphate colitis mouse model: traps and tricks. *J. Biomed. Biotechnol.* 2012:718617. doi: 10.1155/2012/718617

Peterson, L. W., and Artis, D. (2014). Intestinal epithelial cells: regulators of barrier function and immune homeostasis. *Nat. Rev. Immunol.* 14, 141–153. doi: 10.1038/nri3608

Podolsky, D. K. (2002). The current future understanding of inflammatory bowel disease. *Best Pract. Res. Clin. Gastroenterol.* 16, 933–943. doi: 10.1053/bega.2002.0354

Razi, S., Baradaran Noveiry, B., Keshavarz-Fathi, M., and Rezaei, N. (2019). IL-17 and colorectal cancer: from carcinogenesis to treatment. *Cytokine* 116, 7–12. doi: 10.1016/j.cyto.2018.12.021

Renes, I. B., Verburg, M., Van Nispen, D. J., Taminiau, J. A., BLler, H. A., Dekker, J., et al. (2002). Epithelial proliferation, cell death, and gene expression in experimental colitis: alterations in carbonic anhydrase I, mucin MUC2, and trefoil factor 3 expression. *Int. J. Color. Dis.* 17, 317–326. doi: 10.1007/s00384-002-0409-4

Roda, G., Sartini, A., Zambon, E., Calafiore, A., Marocchi, M., Caponi, A., et al. (2010). Intestinal epithelial cells in inflammatory bowel diseases. *World J. Gastroenterol.* 16, 4264–4271. doi: 10.3748/wjg.v16.i34.4264

Saleh, M., and Trinchieri, G. (2011). Innate immune mechanisms of colitis and colitis-associated colorectal cancer. *Nat. Rev. Immunol.* 11, 9–20. doi: 10.1038/nri2891

Schaefer, J. S., Attumi, T., Opekun, A. R., Abraham, B., Hou, J., Shelby, H., et al. (2015). MicroRNA signatures differentiate Crohn's disease from ulcerative colitis. *BMC Immunol.* 16:5. doi: 10.1186/s12865-015-0069-0

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., et al. (2012). Fiji: an open-source platform for biological-image analysis. *Nat. Methods* 9, 676–682. doi: 10.1038/nmeth.2019

Seksik, P. (2010). Microbiote intestinal et MICI. *Gastroenterologie Clinique et Biologique* 34, 48–55. doi: 10.1016/S0399-8320(10)70007-5

Sheehan, D., Moran, C., and Shanahan, F. (2015). The microbiota in inflammatory bowel disease. *J. Gastroenterol.* 50, 495–507. doi: 10.1007/s00535-015-1064-1

Shi, J., Ma, X., Su, Y., Song, Y., Tian, Y., Yuan, S., et al. (2018). MiR-31 mediates inflammatory signaling to promote re-epithelialization during skin wound healing. *J. Investig. Dermatol.* 138, 2253–2263. doi: 10.1016/j.jid.2018.03.1521

Shi, T., Xie, Y., Fu, Y., Zhou, Q., Ma, Z., Ma, J., et al. (2016). The signaling axis of microRNA-31/interleukin-25 regulates Th1/Th17-mediated inflammation response in colitis. *Mucosal Immunol.* 10, 983–995. doi: 10.1038/mi.2016.102

Simmonds, R. E., and Foxwell, B. M. (2008). Signalling, inflammation and arthritis: NF-κB and its relevance to arthritis and inflammation. *Rheumatology* 47, 584–590. doi: 10.1093/rheumatology/kem298

Solomon, L., Mansor, S., Mallon, P., Donnelly, E., Hoper, M., Loughrey, M., et al. (2010). The dextran sulphate sodium (DSS) model of colitis: an overview. *Comp. Clin. Pathol.* 19, 235–239. doi: 10.1007/s00580-010-0979-4

Soroosh, A., Koutsioumpa, M., Pothoulakis, C., and Iliopoulos, D. (2018). Functional role and therapeutic targeting of microRNAs in inflammatory bowel disease. *Am. J. Physiol. Gastrointest. Liver Physiol.* 314, G256–G262. doi: 10.1152/ajpgi.00268.2017

Stepicheva, N. A., and Song, J. L. (2016). Function and regulation of microRNA-31 in development and disease. *Mol. Reprod. Dev.* 83, 654–674. doi: 10.1002/mrd.22678

Tian, Y., Ma, X., Lv, C., Sheng, X., Li, X., Zhao, R., et al. (2017). Stress responsive miR-31 is a major modulator of mouse intestinal stem cells during regeneration and tumorigenesis. *elife* 6:e29538. doi: 10.7554/eLife.29538

Tian, Y., Xu, J., Li, Y., Zhao, R., Du, S., Lv, C., et al. (2019). MicroRNA-31 reduces inflammatory signaling and promotes regeneration in colon epithelium, and delivery of mimics in microspheres reduces colitis in mice. *Gastroenterology* 156, 2281–2296.e6. doi: 10.1053/j.gastro.2019.02.023

Tian, C., Yao, S., Liu, L., Ding, Y., Ye, Q., Dong, X., et al. (2017). Klf4 inhibits tumor growth and metastasis by targeting microRNA-31 in human hepatocellular carcinoma. *Int. J. Mol. Med.* 39, 47–56. doi: 10.3892/ijmm.2016.2812

Tili, E., Michaille, J.-J., Piurowski, V., Rigot, B., and Croce, C. M. (2017). MicroRNAs in intestinal barrier function, inflammatory bowel disease and related cancers—their effects and therapeutic potentials. *Curr. Opin. Pharmacol.* 37, 142–150. doi: 10.1016/j.coph.2017.10.010

Treiber, T., Treiber, N., and Meister, G. (2018). Regulation of microRNA biogenesis and its crosstalk with other cellular pathways. *Nat. Rev. Mol. Cell Biol.* 20, 5–20. doi: 10.1038/s41580-018-0059-1

Valastyan, S., and Weinberg, R. A. (2010). miR-31: a crucial overseer of tumor metastasis and other emerging roles. *Cell Cycle* 9, 2124–2129. doi: 10.4161/cc.9.11.11843

Van De Wetering, C., Aboushousha, R., Manuel, A. M., Chia, S. B., Erickson, C., Macpherson, M. B., et al. (2020). Pyruvate kinase M2 promotes expression of Proinflammatory mediators in house dust mite–induced allergic airways disease. *J. Immunol.* 204, 763–774. doi: 10.4049/jimmunol.1901086

Wang, C.-J., Zhou, Z.-G., Wang, L., Yang, L., Zhou, B., Gu, J., et al. (2009). Clinicopathological significance of microRNA-31, -143 and-145 expression in colorectal cancer. *Dis. Markers* 26:921907. doi: 10.1042/BSR20211280

Wen, J., Xiong, K., Aili, A., Wang, H., Zhu, Y., Yu, Z., et al. (2020). PEX5, a novel target of microRNA-31-5p, increases radioresistance in hepatocellular carcinoma by activating Wnt/β-catenin signaling and homologous recombination. *Theranostics* 10, 5322–5340. doi: 10.7150/thno.42371

Wirtz, S., and Neurath, M. (2007). Mouse models of inflammatory bowel disease☆. *Adv. Drug Deliv. Rev.* 59, 1073–1083. doi: 10.1016/j.addr.2007.07.003

Wu, J., Tan, X., Lin, J., Yuan, L., Chen, J., Qiu, L., et al. (2016). Minicircle-oriP-miR-31 as a novel EBNA1-specific miRNA therapy approach for nasopharyngeal carcinoma. *Hum. Gene Ther.* 28, 415–427. doi: 10.1089/hum.2016.136

Wu, Y., Wang, Q., Qu, J., Liu, W., Gao, X., Li, X., et al. (2021). Different response modes and cooperation modulations of blue-light receptors in photomorphogenesis. *Plant Cell Environ.* 44, 1802–1815. doi: 10.1111/pce.14038

Wu, F., Zikusoka, M., Trindade, A., Dassopoulos, T., Harris, M. L., Bayless, T. M., et al. (2008). MicroRNAs are differentially expressed in ulcerative colitis and Alter expression of macrophage inflammatory peptide-2α. *Gastroenterology* 135, 1624–1635.e24. doi: 10.1053/j.gastro.2008.07.068

Xavier, R. J., and Podolsky, D. K. (2007). Unravelling the pathogenesis of inflammatory bowel disease. *Nature* 448, 427–434. doi: 10.1038/nature06005

Xie, J., Najafi, J., Le Borgne, R., Verbavatz, J.-M., Durieu, C., Sall, J., et al. (2022). Contribution of cytoplasm viscoelastic properties to mitotic spindle positioning. *Proc. Natl. Acad. Sci.* 119:e2115593119. doi: 10.1073/pnas.2115593119

Xie, Z., Wang, Y., Yang, G., Han, J., Zhu, L., Li, L., et al. (2021). The role of the hippo pathway in the pathogenesis of inflammatory bowel disease. *Cell Death Dis.* 12:79. doi: 10.1038/s41419-021-03395-3

Xu, X. M., and Zhang, H. J. (2016). miRNAs as new molecular insights into inflammatory bowel disease: crucial regulators in autoimmunity and inflammation. *World J. Gastroenterol.* 22, 2206–2218. doi: 10.3748/wjg.v22.i7.2206

Yu, Y., Zhang, X., Liu, F., Zhu, P., Zhang, L., Peng, Y., et al. (2021). A stress-induced miR-31–CLOCK–ERK pathway is a key driver and therapeutic target for skin aging. *Nat. Aging* 1, 795–809. doi: 10.1038/s43587-021-00094-8

Zelov, H., and Hošek, J. (2013). TNF-α signalling and inflammation: interactions between old acquaintances. *Inflamm. Res.* 62, 641–651. doi: 10.1007/s00011-013-0633-0

Zhang, Z., Zheng, M., Bindas, J., Schwarzenberger, P., and Kolls, J. K. (2006). Critical role of IL-17 receptor signaling in acute TNBS-induced colitis. *Inflamm. Bowel Dis.* 12, 382–388. doi: 10.1097/01.MIB.0000218764.06959.91

Zhao, R., Du, S., Liu, Y., Lv, C., Song, Y., Chen, X., et al. (2020). Mucoadhesive-to-penetrating controllable peptosomes-in-microspheres co-loaded with anti-miR-31 oligonucleotide and Curcumin for targeted colorectal cancer therapy. *Theranostics* 10, 3594–3611. doi: 10.7150/thno.40318

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 22 December 2022
DOI 10.3389/fmicb.2022.987930

OPEN ACCESS

EDITED BY
George Tsiamis,
University of Patras,
Greece

REVIEWED BY
Mete Yılmaz,
Bursa Technical University,
Turkey
ChangQing Yu,
Xijing University,
China

*CORRESPONDENCE
Shunhan Yao
✉ 2128402005@st.gxu.edu.cn

SPECIALTY SECTION
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

RECEIVED 06 July 2022
ACCEPTED 30 November 2022
PUBLISHED 22 December 2022

CITATION
Yao D, Nong L, Qin M, Wu S and Yao S (2022) Identifying circRNA-miRNA interaction based on multi-biological interaction fusion.
Front. Microbiol. 13:987930.
doi: 10.3389/fmicb.2022.987930

Identifying circRNA-miRNA interaction based on multi-biological interaction fusion

Dunwei Yao[1,2], Lidan Nong[3], Minzhen Qin[1,2], Shengbin Wu[2,4] and Shunhan Yao[5]*

[1]Department of Gastroenterology, The People's Hospital of Baise, Baise, China, [2]The Southwest Affiliated Hospital of Youjiang Medical University for Nationalities, Baise, China, [3]Department of Child Healthcare, Baise Maternal and Child Hospital, Baise, China, [4]Department of Pulmonary and Critical Care Medicine, The People's Hospital of Baise, Baise, China, [5]Medical College of Guangxi University, Nanning, China

CircRNA is a new type of non-coding RNA with a closed loop structure. More and more biological experiments show that circRNA plays important roles in many diseases by regulating the target genes of miRNA. Therefore, correct identification of the potential interaction between circRNA and miRNA not only helps to understand the mechanism of the disease, but also contributes to the diagnosis, treatment, and prognosis of the disease. In this study, we propose a model (IIMCCMA) by using network embedding and matrix completion to predict the potential interaction of circRNA-miRNA. Firstly, the corresponding adjacency matrix is constructed based on the experimentally verified circRNA-miRNA interaction, circRNA-cancer interaction, and miRNA-cancer interaction. Then, the Gaussian kernel function and the cosine function are used to calculate the circRNA Gaussian interaction profile kernel similarity, circRNA functional similarity, miRNA Gaussian interaction profile kernel similarity, and miRNA functional similarity. In order to reduce the influence of noise and redundant information in known interactions, this model uses network embedding to extract the potential feature vectors of circRNA and miRNA, respectively. Finally, an improved inductive matrix completion algorithm based on the feature vectors of circRNA and miRNA is used to identify potential interactions between circRNAs and miRNAs. The 10-fold cross-validation experiment is utilized to prove the predictive ability of the IIMCCMA. The experimental results show that the AUC value and AUPR value of the IIMCCMA model are higher than other state-of-the-art algorithms. In addition, case studies show that the IIMCCMA model can correctly identify the potential interactions between circRNAs and miRNAs.

KEYWORDS

circRNA-miRNA interaction, multi-biological interaction fusion, inductive matrix completion, network embedding, computational method

1. Introduction

Different from traditional linear non-coding RNA, circRNA is a new type of non-coding RNA with a closed loop structure (3′ and 5′ in circRNA are connected together; Wilusz and Sharp, 2013; Lan et al., 2015b). The unique molecular structure of circRNA ensures that it cannot be affected by RNA exonuclease. In addition, the expression of circRNA is more stable and not easily degraded than other linear non-coding RNA. Further experiments proved that circRNA is rich in miRNA binding sites, which can act as a miRNA sponge in cells to splice, transcribe, and modify the expression of parental genes (Qu et al., 2015; Rybak-Wolf et al., 2015).

Recent experimental results show that circRNA plays an important role in many diseases. For example, quantitative real-time PCR (qRT-PCR) detection found that circRNA BCRC-3 is low expressed in bladder cancer tissue cells. Moreover, cricRNA BCRC-3 can directly bind to miRNA miR-182-5p, and then act as a sponge for miRNA miR-182-5p to promote the activity of its target genes. Therefore, circRNA BCRC3 can be used as a tumor suppressor to inhibit the proliferation of bladder cancer cells (Xie et al., 2018). The expression of circRNA hsa_circ_0008068 is significantly down-regulated in prostate cancer cells. There are multiple binding sites between the circRNA and the anticancer miRNA miR-145-3p. CircRNA hsa_circ_0008068 can play an anti-cancer role in prostate cancer cells by regulating miR-145-3p and its target gene WISP1. Therefore, circRNA hsa_circ_000806 may be an important target for the diagnosis and treatment of prostate cancer (Zheng et al., 2020).

With the continuous development of high-throughput sequencing technology, more and more circRNA-miRNA-disease interactions have been confirmed. At the same time, a large number of databases have been developed to store the basic information of circRNA and interactions related to circRNA such as circBase (Glažar et al., 2014), circBank (Liu et al., 2019), circad (Rophina et al., 2020), and circR2Cancer (Lan et al., 2020c). As a benchmark database in the circRNA field, the circBase database stores basic information related to circRNA such as the position of circRNA, the genomic length, the spliced sequence length, and the gene symbol (Glažar et al., 2014). circBank is a professional database dedicated to standardizing circRNA naming (Liu et al., 2019). This database not only provides basic information about circRNA, but also names some newly discovered circRNAs uniformly. The Circad database collects 1,388 experimentally verified circRNA-disease interactions from five different species (Homo sapiens, mice, rats, chickens, and wild boars; Rophina et al., 2020). CircR2Cancer is a new database that stores circRNA-cancer interactions. This database not only stores experimentally verified circRNA-cancer interactions but also circRNA-miRNA interactions and miRNA-cancer interactions (Lan et al., 2020c). In addition to storing experimentally verified interactions, the circR2Cancer database also stores basic information about circRNA and diseases.

The emergence of circRNA-related databases provides a data basis for circRNA-related interaction prediction based on computational methods. Compared with traditional biological identification method, the interaction prediction model based on the computational method has higher accuracy and less time consumption. Guo et al. (2022) presented a computational model to predict circRNA-miRNA interactions by using Word2vec, Structural Deep Network Embedding, Convolutional Neural Network, and Deep Neural Network. Qian et al. (2022) proposed a computational model (CMASG) for circRNA-miRNA interactions prediction based on graph neural network and singular value decomposition. It utilized the graph neural network to learn feature representations of nodes and the lightGBM to predict circRNA-miRNA association. Lan et al. (2021b) developed a computational framework (NECMA) to identify interactions between circRNAs and miRNAs by using network embedding. It extracted features of circRNA and miRNA based on network embedding and predict circRNA-miRNA associations based on neighborhood regularization logic matrix decomposition and inner product. He et al. (2022) proposed a computational approach (GCNCMI) to predict the potential interactions between circRNAs and miRNAs based on graph convolutional neural network. It used the graph convolutional neural network to exact the potential interactions of adjacent nodes and then utilized the embedded representations generated by each layer to predict the final score. Qian et al. (2021) introduced a computational framework (CMIVGSD), to predict circRNA-miRNA interaction by using singular value decomposition and graph variational auto-encoders. Yu et al. (2022) proposed a computational model (SGCNCMI) to identify circRNA-miRNA interactions by combining multimodal information and graph convolutional neural network. Wang et al. (2022) presented a computing method (KGDCMI) to predict the interactions between circRNA and miRNA based on multi-source information fusion. It exacts RNA attribute information from sequence and similarity and captures the behavior information in RNA association based on graph-embedding algorithm. Then, the principal component analysis is used to obtain feature vector, and further the deep neural network is utilized to identify potential circRNA-miRNA interactions. Fang and Lei (2019) fused circRNA-miRNA interaction network, circRNA functional similarity network, and miRNA functional similarity network to construct a circRNA-miRNA heterogeneous network. Then use the K-nearest neighbor algorithm based on restart random walk to predict the potential interaction of circRNA and miRNA.

In this paper, we propose a circRNA-miRNA interaction prediction model (IIMCCMA) based on multi-biological interaction data. This model uses experimentally verified circRNA-miRNA interaction, circRNA-cancer interaction, and miRNA-cancer interaction to construct circRNA-miRNA adjacency matrix, circRNA-cancer adjacency matrix, and miRNA-cancer adjacency matrix, respectively. On the basis of the above adjacency matrix, this model uses Gaussian kernel function and cosine function to calculate circRNA GIP kernel similarity and circRNA functional similarity, as well as miRNA GIP kernel similarity and miRNA functional similarity. In order to reduce the negative impact of noise or redundant information in the known circRNA-miRNA interaction

on the prediction model, the IIMCCMA model first uses the known circRNA-miRNA interaction to construct a heterogeneous network. Then we use the network embedding algorithm to extract the potential feature vectors of circRNA and miRNA in heterogeneous networks. In order to make full use of the information contained in different data sources, this model uses a feature fusion method to integrate the similarity features and topological features of entities in the interaction network to form circRNA fusion features and miRNA fusion features, respectively. Finally, on the basis of circRNA fusion features and miRNA fusion features, an improved inductive matrix completion algorithm is used to predict the potential interaction of circRNA and miRNA. The 10-fold cross-validation experiment was used to evaluate the predictive performance of the IIMCCMA model. The experimental results show that the IIMCCMA model achieves better performance than other advanced interaction prediction models. In addition, the case study results show that the IIMCCMA model can correctly identify the potential interaction between circRNA and miRNA.

2. Materials and methods

2.1 Materials

We use two datasets as gold standard set inhere which is downloaded from circR2Cancer (Lan et al., 2020c) and KGNACDA (Lan et al., 2022a). In dataset 1, there are 756 interactions between 514 circRNAs and 461 miRNAs, 647 interactions between 514 circRNAs and 62 cancers, and 732 interactions between 461 miRNAs and 62 cancers. In dataset 2, there are 330 circRNAs, 79 diseases and 245 miRNAs, 346 circRNA-disease interactions, 146 circRNA-miRNA interactions, and 106 miRNA-disease interactions. Further, we construct an adjacency matrix to represent the above-mentioned interaction network. The adjacency matrix *CM* represents the circRNA-miRNA interactions. If circRNA CM_i is related to miRNA CM_j, $CM(i,j)=1$, otherwise, $CM(i,j)=0$. Similarly, the adjacency matrix *CC* represents the circRNA-cancer interactions. If circRNA CC_i is related to cancer CC_j, $CC(i,j)=1$, otherwise, $CC(i,j)=0$. The adjacency matrix *MC* represents the miRNA-cancer interactions. If miRNA is related to cancer MC_j, $MC(i,j)=1$, otherwise, $MC(i,j)=0$.

2.2 circRNA and miRNA similarity calculation

Based on the assumption that circRNAs with similar functions are often associated with similar miRNAs (Lan et al., 2020a, 2021a, 2022c), circRNA GIP kernel similarity and miRNA GIP kernel similarity are calculated based on the circRNA-miRNA interaction network, respectively. We define *GCS* to represent the Gaussian interaction profile kernel similarity network of circRNA.

The definition of GIP kernel similarity between circRNA c_i and circRNA c_j is as follows:

$$GCS(c_i,c_j)=\exp\left(-\gamma_{cs}CM(i,:)-CM(j,:)^2\right)$$

$$\gamma_{cs}=1/\left(\frac{1}{n_{circ}}\sum_{i=1}^{n_{circ}}CM(i,:)^2\right)$$

where, $CM(i,:)$ represents the *i-th* row of the circRNA-miRNA interaction network *CM*. n_{circ} represents the number of rows of the interaction network *CM*. γ_{cs} represents the kernel bandwidth.

Similarly, we define *GMS* to represent the Gaussian interaction profile kernel similarity network of miRNA. The definition of GIP kernel similarity between miRNA m_i and miRNA m_j is as follows:

$$GMS(m_i,m_j)=\exp\left(-\gamma_{ms}CM(:,i)-CM(:,j)^2\right)$$

$$\gamma_{ms}=1/\left(\frac{1}{n_{mi}}\sum_{i=1}^{n_{mi}}CM(:,i)^2\right)$$

where, $CM(:,i)$ represents the *i-th* column of the circRNA-miRNA interaction network *CM*. n_{mi} represents the number of columns of the interaction network *CM*. γ_{ms} represents the kernel bandwidth.

In addition, we also use the cosine function to calculate the circRNA functional similarity and the miRNA functional similarity on the basis of circRNA-cancer interaction network and miRNA-cancer interaction network. The cosine similarity measures the similarity between two vectors by the angle between two vectors in a two-dimensional space. If the two vectors point in the same direction, it means that the two vectors are more similar, otherwise, the similarity is lower. Therefore, according to the above cosine similarity theory, the circRNA functional similarity and miRNA functional similarity are defined as follows:

$$CCS(c_i,c_j)=\frac{\sum_{i=1}^{n_{circ}}CC(i,:)\times CC(j,:)}{\sqrt{\sum_{i=1}^{n_{circ}}CC(i,:)^2}\times\sqrt{\sum_{i=1}^{n_{circ}}CC(j,:)^2}}$$

$$CMS(m_i,m_j)=\frac{\sum_{i=1}^{n_{mi}}MC(i,:)\times MC(j,:)}{\sqrt{\sum_{i=1}^{n_{mi}}MC(i,:)^2}\times\sqrt{\sum_{i=1}^{n_{mi}}MC(j,:)^2}}$$

where *CCS* and *CMS* represent the circRNA functional similarity network and the miRNA functional similarity network,

respectively. $CCS(c_i,c_j)$ represents the functional similarity between circRNA c_i and circRNA c_j. $CC(i,:)$ represents the *i-th* row in the circRNA-cancer network *CC*. n_{circ} represents the number of rows in the network *CC*. $CMS(m_i,m_j)$ represents the functional similarity between miRNA m_i and miRNA m_i. $MC(i,:)$ represents the *i-th* row in the miRNA-cancer network *MC*. n_{mi} represents the number of rows in the network *MC*.

In order to make better use of the circRNA and the miRNA similarity characteristics, we integrate the above two similarities to obtain the circRNA similarity $circ_{sim}$ and miRNA similarity mi_{sim}, which are defined as follows:

$$circ_{sim}(c_i,c_j)=\begin{cases}GCS(c_i,c_j), if\ CCS(c_i,c_j)=0\\ CCS(c_i,c_j), otherwise\end{cases}$$

$$mi_{sim}(m_i,m_j)=\begin{cases}GMS(m_i,m_j), if\ CMS(m_i,m_j)=0\\ CMS(m_i,m_j), otherwise\end{cases}$$

where $circ_{sim}(c_i,c_j)$ represents the integrated similarity between circRNA c_i and circRNA c_j. $mi_{sim}(m_i,m_j)$ represents the integrated similarity between miRNA m_i and miRNA m_j. *GCS* represents circRNA GIP kernel similarity. *CCS* represents the circRNA functional similarity. In the same way, *GMS* represents miRNA GIP kernel similarity. *CMS* represents the miRNA functional similarity.

2.3 Potential feature extraction and fusion of circRNA and miRNA

In order to reduce the influence of noise or redundant information in the known circRNA-miRNA interaction network, we construct the heterogeneous network $H_{circ-mi}$. The heterogeneous network is composed of the circRNA-miRNA interaction adjacency matrix *CM* and the transposed matrix CM^T of the circRNA-miRNA adjacency matrix. It is defined as follows:

$$H_{circ-mi}=\begin{bmatrix}0 & CM\\ CM^T & 0\end{bmatrix}$$

After obtaining the heterogeneous network, the NetMF algorithm (Qiu et al., 2018) is used to obtain the circRNA-miRNA latent feature matrix with size equals to $(m+n)\times d$. Among them, m represents the number of circRNA in the heterogeneous network and $H_{circ-mi}$. n represents the number of miRNAs. d represents the dimensions of circRNA and miRNA low-dimensional space vectors. Experiments have verified that the model has the best prediction effect when the dimension of the low-dimensional space vector of circRNA and miRNA is set to 16.

In order to make full use of the information of different interaction, we use a fusion method to fuse the circRNA and miRNA topological features ($circ_{Net}$, mi_{Net}) obtained through the NetMF algorithm with the integrated circRNA similarity features and miRNA similarity features, respectively. The fused information can not only describe the characteristics of different data sources, but also describe the complex relationship between circRNA and miRNA more comprehensively. The fusion feature of circRNA $circ\_feature$ and the fusion feature of miRNA $mi\_feature$ are defined as follows:

$$circ\_f=[circ_{Net},circ_{sim}]$$

$$mi\_f=[mi_{Net},mi_{sim}]$$

where $circ_{Net}$ and mi_{Net} represent the topological characteristics of circRNA and miRNA based on the NetMF algorithm, respectively. $circ_{sim}$ and mi_{sim} represents the circRNA integrated similarity and miRNA integrated similarity, respectively.

2.4 Prediction of potential interaction between circRNA and miRNA

In this paper, we propose a circRNA-miRNA interaction prediction model (IIMCCMA) based on an improved inductive matrix completion algorithm. This model is implemented based on the known circRNA-miRNA interaction, the fusion feature of circRNA and the fusion feature of miRNA. The specific implementation process of the IIMCCMA model is shown in Figures 1A,B.

Many studies have found that the sparsity problem of biological interaction networks is very serious. Taking the circRNA-miRNA interaction network used in this paper as an example, the circRNA-miRNA interaction network *CM* is composed of 756 interactions between 514 circRNAs and 461 miRNAs. Obviously, the interaction network *CM* is very sparse (the matrix density is 0.0032). In addition, in the calculation process of the inductive matrix completion algorithm (Jain and Dhillon, 2013; Lan et al., 2015a; Si et al., 2016; Nazarov et al., 2018), due to the high sparsity of the known interaction matrix, a relatively large amount of effective information will be lost in the process of low-dimensional mapping, which will affect the prediction effect of the circRNA-miRNA potential interaction prediction model. Therefore, in order to alleviate the negative impact of the high sparsity of the interaction network on the model, we modify the mapping method of the low-rank matrix in the inductive matrix completion algorithm. Specifically, in order to better protect the structural information in the sparse matrix, we perform multiple low-dimensional

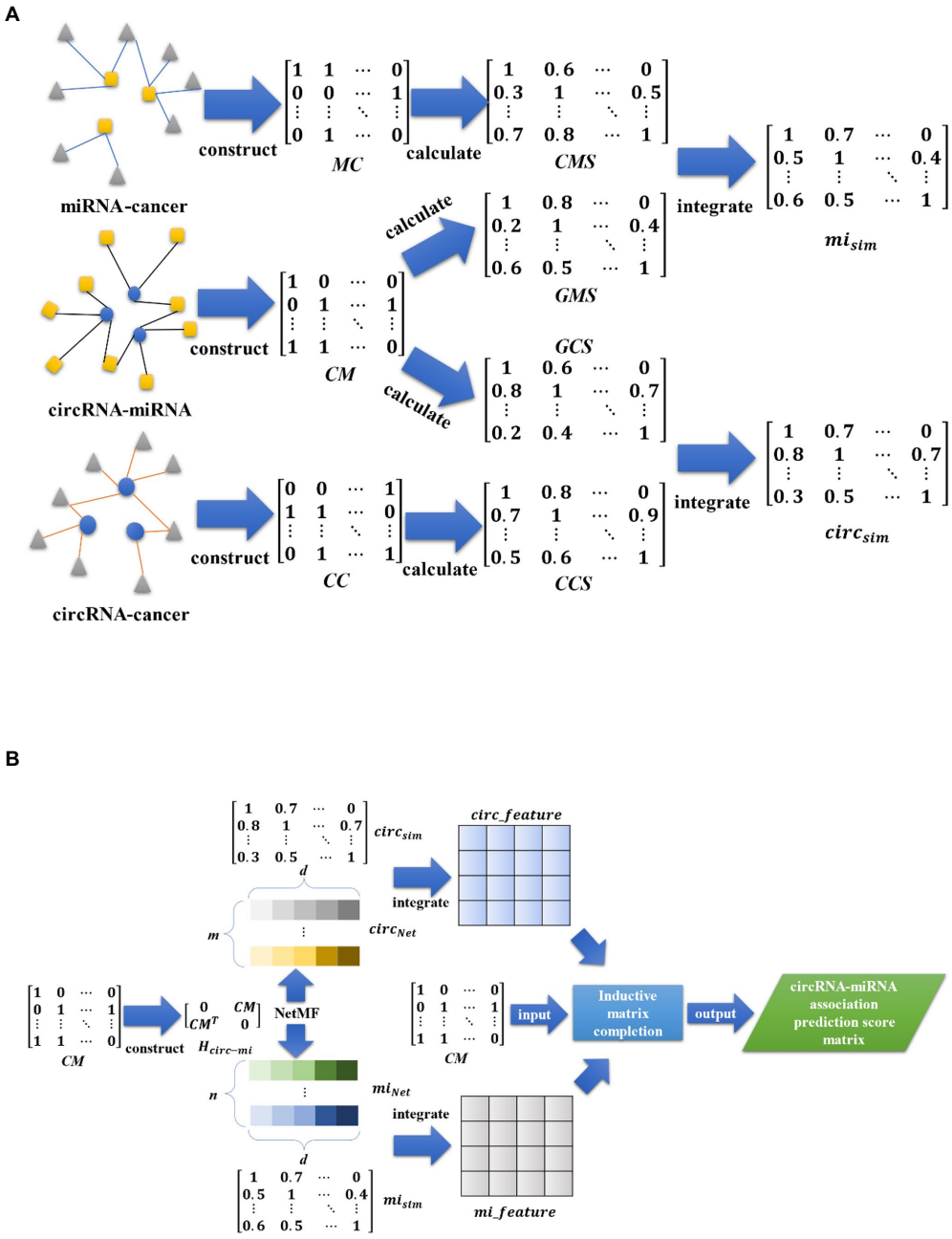

FIGURE 1
(A) Overview of interaction prediction model for circRNA and miRNA based on multi-biological interaction (1). **(A)** Mainly shows the construction of the incidence matrix, the calculation of similarity, and the fusion of similarity. **(B)** Overview of interaction prediction model for circRNA and miRNA based on multi-biological interaction (2). **(B)** Mainly shows the construction of heterogeneous networks, feature extraction based on NetMF algorithm, feature fusion, and calculation of interaction prediction scores.

mapping operations to obtain multiple low-rank matrices with different dimensions. Then we use low-rank matrices of different dimensions to calculate the potential interaction prediction scores of circRNA and miRNA. Finally, the prediction score matrix calculated from the low-rank matrix of different dimensions is integrated to realize the potential interaction prediction of circRNA and miRNA.

In summary, the objective function of the circRNA-miRNA potential interaction prediction model based on the fusion feature and the improved inductive matrix completion algorithm is as follows:

$$\min_{W,H} \frac{1}{2} \| CM - CM_{Pre} \|_F^2 + \frac{\theta_1}{2} \| W_i^{d_i} \|_F^2 + \frac{\theta_2}{2} \| H_i^{d_i} \|_F^2, W_i^{d_i} \geq 0, H_i^{d_i} \geq 0$$

$$CM_{Pre} = mi\_fW_i^{d_i} H_i^{d_i T} circ\_f^T$$

where CM represents the known circRNA-miRNA interaction matrix. CM_{pre} represents the predicted circRNA-miRNA interaction matrix. $W_i^{d_i}$ and $H_i^{d_i}$ represent the d-dimensional low-rank matrix obtained through the i-th complete iteration of the circRNA-miRNA interaction matrix. $\theta 1$ and $\theta 2$ represent the regularization parameters. According to the previous research, we set $\theta_1 = \theta_2 = 1$ $\|\cdot\|F$ represents the Frobenius norm of the matrix (F-norm). $\frac{\theta_1}{2}\|W_i^{d_i}\|_F^2$ and $\frac{\theta_2}{2}\|H_i^{d_i}\|_F^2$ are used to prevent overfitting. In order to find the minimum value of the objective function, we first set up the random dense matrices of $W_i^{d_i}$ and $H_i^{d_i}$, and then update the matrices $W_i^{d_i}$ and $H_i^{d_i}$ through iterative equations. When the convergence condition is met, we will stop iteration. The iterative equation are defined as follows:

$$W_i^{d_i} \leftarrow W_i^{d_i} \frac{mi\_f^T * CM * circ\_fH_{ini}}{mi\_f^T mi\_fW_{ini} H_{ini}{}^T circ\_f^T circ\_fH_{ini} + \theta_1 W_{ini}}$$

$$H_i^{d_i} \leftarrow H_i^{d_i} \frac{circ\_feature^T CM^T mi\_featureW_{ini}}{circ\_feature^T circ\_featureH_{ini} W_{ini}{}^T mi\_feature^T mi\_featureW_{ini} + \theta_2 H_{ini}}$$

where $circ\_feature$ and $mi\_feature$ represent the fusion characteristics of circRNA and the fusion characteristics of miRNA, respectively. $circ\_feature^T$ and $mi\_feature^T$ represent the transposition matrix of the circRNA fusion feature matrix and the transposition matrix of the miRNA fusion feature matrix, respectively. CM^T represents the transposed matrix of the known circRNA-miRNA interaction matrix. W_{ini} and H_{ini} represent the initial random dense matrix of the low-rank matrix $W_i^{d_i}$ and $H_i^{d_i}$, respectively.

Finally, the calculation method of the circRNA and miRNA correlation prediction score matrix is defined as follows:

$$Pre_{circRNA-miRNA} = \frac{\sum_i^k mi\_featureW_i^{d_i} H_i^{d_i T} circ\_feature}{k}$$

where $Pre_{circRNA-miRNA}$ represents the final circRNA-miRNA potential interaction prediction score matrix. Each item in the matrix represents the interaction probability score between circRNA and miRNA. The higher the score, the greater the probability that there is an exact interaction between circRNA and miRNA. k indicates the number of complete iterations of the iterative equation. The best prediction performance is obtained when $k = 2$ and $d_1 = 128$, $d_2 = 64$.

2.5 Performance evaluation

In order to evaluate the performance of model in predicting the potential interaction between circRNA and miRNA, the 10-fold cross-validation experiment is used to evaluate the performance. In 10-fold cross-validation, the known circRNA-miRNA interactions are randomly divided into 10 subsets. Then, in each round of cross-validation experiments, nine subsets are taken from 10 subsets as the training set for model training and the remaining subset is used as the test set. The final interaction prediction score of circRNA and miRNA is obtained. The higher the score, the higher the probability that there is a biological interaction between circRNA and miRNA. Afterward, we ranked the interaction prediction scores between circRNA and miRNA in descending order. Then, the true positive rate (*TPR*) and false positive rate (*FPR*) are calculated by modifying the threshold. The calculation of *TPR* and *FPR* are defined as follows:

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

Finally, a receiver operating curve (ROC) based on the true positive rate and false positive rate is plotted, and the area under the ROC curve (AUROC value) is calculated to evaluate the predictive ability of the model. Similarly, the area of the curve (AUPR value) based on *precision* and *recall* is also used to evaluate the performance of the predictive model. The calculation of *precision* and *recall* is defined as follows:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

where TP means that the classifier predicts the number of positive samples in the actual positive samples. FP represents the number of positive samples is predicted in the actual negative samples. TN means that the classifier predicts the number of negative samples in the actual negative samples. FN indicates the number of actual positive samples that are predicted to be negative.

3. Results and discussion

3.1 Compare with other models

In order to further demonstrate the performance of IIMCCMA, we compare it with the other six prediction methods

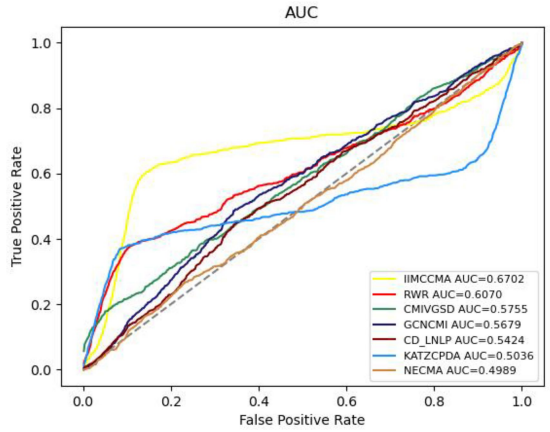

FIGURE 2
Performance comparison between IIMCCMA and other models on dataset 1 in term of AUROC.

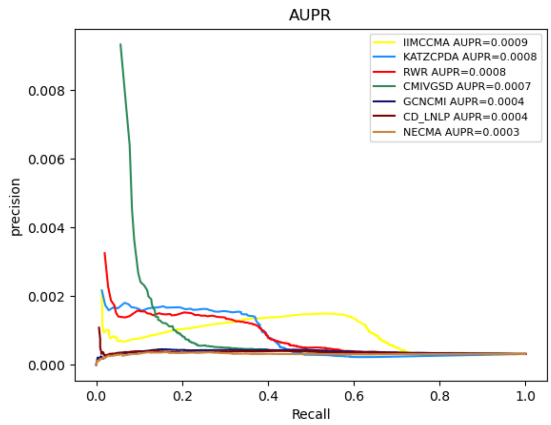

FIGURE 3
Performance comparison between IIMCCMA and other models on dataset 1 in term of AUPR.

(NECMA; Lan et al., 2021b, GCNCMI; He et al., 2022, CMIVGSD; Qian et al., 2021, CCD-LNLP; Zhang et al., 2019, RWR; Vural et al., 2019, and KATZCPDA; Fan et al., 2018). As shown in Figure 2, under the 10-fold cross-validation experiment on dataset 1, the AUROC value of NECMA is 0.4898, the AUROC value of CMIVGSD is 0.5755, the AUROC value of GCNCMI is 0.5679, the AUROC value of CD-LNLP is 0.5424, the AUROC value of RWR is 0.6070, the AUROC value of KATZCPDA is 0.5036, and the AUROC value of IIMCCMA is 0.6702. Therefore, from the experimental results, it can be found that the IIMCCMA model has a higher AUROC value than other interaction prediction models on dataset 1.

As shown in Figure 3, under the 10-fold cross-validation experiment on dataset 1, the AUROC value of NECMA is 0.0003, the AUPR value of CMIVGSD is 0.0007, the AUPR value of GCNCMI is 0.0004, the AUPR value of CD-LNLP is 0.0004, the AUPR value of RWR is 0.0008, the AUPR value of KATZCPDA is 0.0008, and the AUPR value of the IIMCCMA model is 0.0009. It can be found from the experimental results that the IIMCCMA model achieves a higher AUPR value than the other models on dataset 1.

The Figure 4 shows the performance comparison in term of AUROC on dataset 2. It can be found that the AUROC value of NECMA is 0.5021, the AUROC value of CMIVGSD is 0.7081, the AUROC value of GCNCMI is 0.4789, the AUROC value of CD-LNLP is 0.6751, the AUROC value of RWR is 0.6729, the AUROC value of KATZCPDA is 0.6292, and the AUROC value of IIMCCMA is 0.7333. It demonstrates that IIMCCMA outperforms than other prediction models on dataset 2.

As shown in Figure 5, the AUPR value of NECMA is 0.0002, the AUPR value of CMIVGSD is 0.0011, the AUPR value of GCNCMI is 0.0002, the AUPR value of CD-LNLP is 0.0008, the AUPR value of RWR is 0.0007, the AUPR value of KATZCPDA is 0.0006, and the AUPR value of the IIMCCMA model is 0.0011. It can be found that the IIMCCMA model achieves a higher AUPR value than the other models on dataset 2. In conclusion, under the 10-fold cross-validation experiment, we can find that theIIMCCMA has achieved higher AUROC and AUPR values than the other prediction models. Thus, it can be proved that theIIMCCMA performs better in the potential circRNA-miRNA interactions identification.

3.2 Ablation experiment

In order to verify the effectiveness of the improvements of IIMCCMA, we conduct ablation experiment on dataset 1: CircRNA-miRNA potential interaction prediction model based on multi-source similarity and inductive matrix completion (IIMCCMA without improved IMC and topological features). CircRNA-miRNA potential interaction prediction model based on fusion features and inductive matrix completion (IIMCCMA without improved IMC). We adopt the 10-fold cross-validation experiment and use the AUROC value as the evaluation metrics. As shown in Figure 6, the AUROC value of the circRNA-miRNA potential interaction prediction model based on multi-source similarity (IIMCCMA without improved IMC and topological features) is 0.6728. The AUROC value of the circRNA-miRNA potential interaction prediction model based on fusion features (IIMCCMA without improved IMC) is 0.6816. The AUROC value of IIMCCMA is 0.6938. In summary, based on the original inductive matrix completion algorithm, fusion of similarity features and topological features can improve the predictive ability of the model. Adding improved inductive matrix completion on the basis of fusion features can further improve the performance of the prediction model.

3.3 Case study

In order to prove the ability of the circRNA-miRNA potential interaction model (IIMCCMA) based on the multi-source

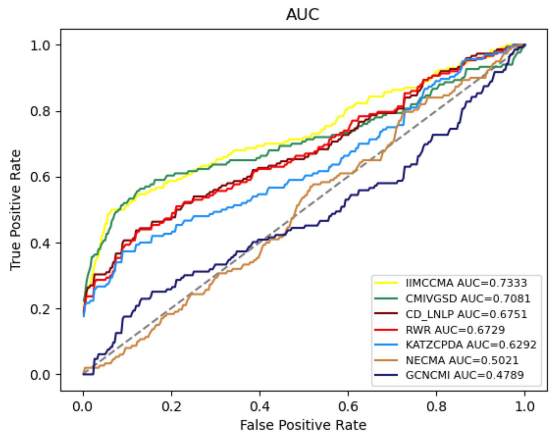

FIGURE 4
Performance comparison between IIMCCMA and other models on dataset 2 in term of AUROC.

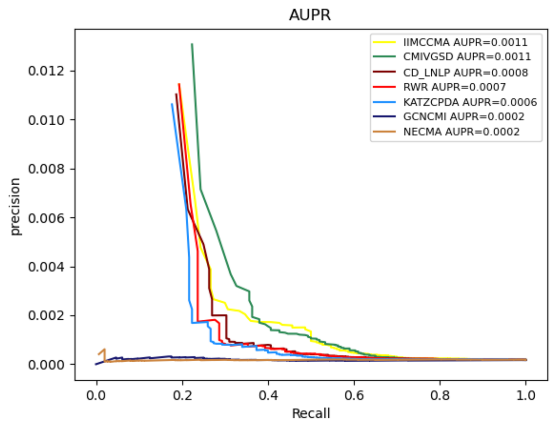

FIGURE 5
Performance comparison between IIMCCMA and other models on dataset 2 in term of AUPR.

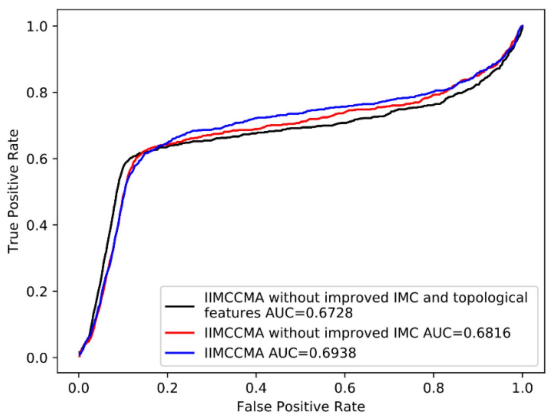

FIGURE 6
Performance comparison between IIMCCMA model and benchmark model.

TABLE 1 Case study based on microRNA miR-145-5p.

Rank	CircRNA	Evidence	Reference
1	hsa_circ_0058063	PMID: 30362519	Sun et al. (2019a)
2	hsa_circRNA_101981	PMID: 30136305	He et al. (2018)
3	hsa_circRNA_091420	PMID: 30136305	He et al. (2018)
4	hsa_circ_100242	PMID: 32218853	Zhu et al. (2020)
5	circPTN	PMID: 31511040	Chen et al. (2019)
6	circPVT1	PMID: 31986409	Zheng and Xu (2020)
7	hsa_circRNA_101996	PMID: 30136305	He et al. (2018)
8	circCEP128	PMID: 30939216	Sun et al. (2019b)
9	hsa_circ_0003855	PMID: 31776711	Zhang et al. (2020)
10	hsa_circ_0001955	PMID: 31822654	Yao et al. (2019)

biological interaction data to identify the potential interaction between circRNA and miRNA. This paper builds a case study based on miRNA miR-145-5p. Finally, this paper selects the top 10 circRNAs predicted by the IIMCCMA model that are related to miRNA miR-145-5p, and manually searches the existing literature to prove their relevance.

The top 10 circRNAs related to miRNA miR-145-5p predicted by the IIMCCMA model are shown in Table 1. From Table 1, 10 circRNAs related to miRNA miR-145-5p (hsa_circ_0058063, hsa_circRNA_101981, hsa_circRNA_091420, hsa_circ_100242, circPTN, circPVT1, hsa_circRNA_101996, circCEP128, hsa_circ_0003855, and hsa_circ_0001955) have been confirmed by existing literature. Specifically, the first circRNA hsa_circ_0058063 can be used as the sponge of miRNA miR-145-5p to regulate the expression of miRNA target gene CDK6 and promote the development of bladder cancer (Sun et al., 2019a). In prostate cancer cells, the expression pattern of the second-ranked circRNA hsa_circRNA_101981 was significantly down-regulated. Further experiments showed that miRNA miR-145-5p can regulate the expression of circRNA hsa_circRNA_101981 (He et al., 2018). The expression pattern of the third-ranked circRNA hsa_circRNA_091420 in prostate cancer cells was significantly upregulated. Overexpressed miRNA miR-145-5p can inhibit the expression of circRNA hsa_circRNA_091420 (He et al., 2018). Experimental results show that the fourth-ranked circRNA hsa_circ_100242 can interact with miRNA miR145-5p in bladder cancer cells (Zhu et al., 2020). Experiments show that the fifth-ranked circRNA circPTN is overexpressed in glioma cells and tissues. Further experiments showed that circRNA circPTN can spongy miRNA miR-145-5p and thus play a carcinogenic effect in glioma cells (Chen et al., 2019). CircRNA circPVT1, ranked sixth, was significantly up-regulated in lung adenocarcinoma cells. Experiments show that in lung adenocarcinoma cells, circRNA circPVT1 can be used as an ceRNA for miRNA miR145-5p (Zheng and Xu, 2020). Experiments show that the seventh-ranked circRNA hsa_circRNA_101996 can interact with miRNA miR-145-5p in prostate cancer cells. In addition, overexpressed

miRNA miR-1455p can inhibit the expression of circRNA hsa_circRNA_101996 (He et al., 2018). The eighth-ranked circRNA circCEP128 can promote the development of bladder cancer by regulating miRNA miR-145-5p and miRNA's target gene MYD88 (Sun et al., 2019b). The expression pattern of circRNA hsa_circ_0003855, ranked ninth, was significantly increased in gastric cancer cells. Experimental results show that circRNA hsa_circ_0003855 can take on the sponge effect of miRNA miR-145-5p to promote the proliferation and migration of gastric cancer cells (Zhang et al., 2020). The tenth-ranked circRNA hsa_circ_0001955 can assume the role of miRNA miR-145-5p sponge. Additionally, the downregulated circRNA hsa_circ_0001955 can inhibit the growth of hepatocellular carcinoma tumors (Yao et al., 2019). In summary, through the case study results based on miRNA miR-145-5p, it can be found that the IIMCCMA model can correctly identify the potential biological interaction between circRNA and miRNA.

4. Conclusion

Experiments show that circRNA can play an important role in cancer as a miRNA sponge. Therefore, correct identification of the interaction between circRNA and miRNA not only helps to understand the complex disease mechanism, but also contributes to the diagnosis, treatment and prognosis of the disease. Based on circRNA-miRNA interaction, circRNA-cancer interaction and miRNA-cancer interaction, this paper proposes a circRNA-miRNA potential interaction prediction model based on multi-source biological interaction data, IIMCCMA. This model first uses the Gaussian kernel function to calculate the GIP kernel similarity of circRNA and the GIP kernel of miRNA based on the circRNA-miRNA interaction network. Then, on the basis of the circRNA-cancer interaction network and the miRNA-cancer interaction network, the cosine function is used to calculate the functional similarity of circRNA and miRNA, respectively. Afterward, the different similarities of circRNAs and the different similarities of miRNAs were integrated separately. The known circRNA-miRNA interaction network is used to construct a heterogeneous network for extracting topological features of circRNA and miRNA, and the network embedding algorithm (NetMF) is used to obtain the low-dimensional space vectors of circRNA and miRNA, respectively. Finally, based on the fusion features, an improved inductive matrix completion algorithm is used to predict the potential interaction between circRNA and miRNA. In order to test the performance of the IIMCCMA, this paper selects four circRNA-disease potential interaction prediction models for comparison. The 10-fold cross-validation results show that compared with the other four models, the IIMCCMA achieved higher AUROC and AUPR values. Therefore, it is proved that IIMCCMA has better predictive ability. Moreover, the results of a case study based on miRNA miR-1455p show that the IIMCCMA model can correctly identify the potential interaction between circRNA and miRNA.

Although, the IIMCCMA model has shown excellent performance in predicting the potential interaction between circRNA and miRNA. However, there are still some shortcomings and limitations. (1) The imbalance of positive and negative samples in interaction data. Because the efficiency of identifying circRNA and miRNA interactions through biological experiments is low, in the existing circRNA-miRNA interaction network, the experimentally verified interactions are far less than the unknown interactions. The sparse circRNA-miRNA interaction network greatly affects the performance of the prediction model (Lan et al., 2016b, 2020b; Lei et al., 2020). Therefore, in the follow-up work, we will try to pre-fill the original interaction matrix to alleviate the sparsity of the known interaction network and enhance the performance of the model. (2) Parameter setting. There are a certain number of parameters in the IIMCCMA model that need to be set manually. The quality of the parameters needs to be confirmed through experimental verification. In addition, too many parameters will reduce the learning and generalization capabilities of the model. Therefore, no parameter or self-learning parameter model will be the main work in the future (Lan et al., 2016a, 2017, 2022b; Chen et al., 2021).

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

SY designed the work. DY and LN performed all the experiments. DY, LN, MQ, and SW wrote the manuscript. All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

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.

Publisher's note

References

Chen, J., Chen, T., Zhu, Y., Li, Y., Zhang, Y., Wang, Y., et al. (2019). circPTN sponges miR-145-5p/miR-330-5p to promote proliferation and stemness in glioma. *J. Exp. Clin. Cancer Res.* 38, 1–17. doi: 10.1186/s13046-019-1376-8

Chen, Q., Lai, D., Lan, W., Wu, X., Chen, B., Liu, J., et al. (2021). ILDMSF: inferring associations between long non-coding RNA and disease based on multi-similarity fusion. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 18, 1106–1112. doi: 10.1109/TCBB.2019.2936476

Fan, C., Lei, X., and Wu, F. X. (2018). Prediction of CircRNA-disease interactions using KATZ model based on heterogeneous networks. *Int. J. Biol. Sci.* 14, 1950–1959. doi: 10.7150/ijbs.28260

Fang, Z., and Lei, X. (2019). Prediction of miRNA-circRNA interactions based on k-NN multi-label with random walk restart on a heterogeneous network. *Big Data Min. Analy.* 2, 261–272. doi: 10.26599/BDMA.2019.9020010

Glažar, P., Papavasileiou, P., and Rajewsky, N. (2014). circBase: a database for circular RNAs. *RNA* 20, 1666–1670. doi: 10.1261/rna.043687.113

Guo, L. X., You, Z. H., Wang, L., Yu, C. Q., Zhao, B. W., Ren, Z. H., et al. (2022). A novel circRNA-miRNA association prediction model based on structural deep neural network embedding. *Brief. Bioinform.* 23:bbac391. doi: 10.1093/bib/bbac391

He, J. H., Han, Z. P., Zhou, J. B., Chen, W. M., Lv, Y. B., He, M. L., et al. (2018). MiR-145 affected the circular RNA expression in prostate cancer LNCaP cells. *J. Cell. Biochem.* 119, 9168–9177. doi: 10.1002/jcb.27181

He, J., Xiao, P., Chen, C., Zhu, Z., Zhang, J., and Deng, L. (2022). GCNCMI: a graph convolutional neural network approach for predicting circRNA-miRNA interactions. *Front. Genet.* 13:959701. doi: 10.3389/fgene.2022.959701

Jain, P., and Dhillon, I. S. (2013). Provable inductive matrix completion. arXiv [Preprint]. doi: 10.48550/arXiv.1306.0626

Lan, W., Dong, Y., Chen, Q., Liu, J., Wang, J., Chen, Y. P. P., et al (2021a). IGNSCDA: predicting CircRNA-disease associations based on improved graph convolutional network and negative sampling. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi: 10.1109/TCBB.2021.3111607 (Epub ahead of print).

Lan, W., Dong, Y., Chen, Q., Zheng, R., Liu, J., Pan, Y., et al. (2022a). KGANCDA: predicting circRNA-disease associations based on knowledge graph attention network. *Brief. Bioinform.* 23:bbab494. doi: 10.1093/bib/bbab494

Lan, W., Lai, D., Chen, Q., Wu, X., Chen, B., Liu, J., et al. (2020a). LDICDL: LncRNA-disease association identification based on collaborative deep learning. *IEEE/ACM Trans. Comput. Biol. Bioinform.*

Lan, W., Lai, D., Chen, Q., Wu, X., Chen, B., Liu, J., et al. (2020b). LDICDL: LncRNA-disease interaction identification based on collaborative deep learning. *IEEE/ACM Trans. Comput. Biol. Bioinform.*

Lan, W., Li, M., Zhao, K., Liu, J., Wu, F. X., Pan, Y., et al. (2017). LDAP: a web server for lncRNA-disease association prediction. *Bioinformatics* 33, 458–460. doi: 10.1093/bioinformatics/btw639

Lan, W., Wang, J., Li, M., Liu, J., Li, Y., Wu, F. X., et al. (2016a). Predicting drug–target interaction using positive-unlabeled learning. *Neurocomputing* 206, 50–57. doi: 10.1016/j.neucom.2016.03.080

Lan, W., Wang, J., Li, M., Liu, J., and Pan, Y. (2015a). "Predicting microRNA-disease associations by integrating multiple biological information." in *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 183–188.

Lan, W., Wang, J., Li, M., Liu, J., Wu, F. X., and Pan, Y. (2016b). Predicting microRNA-disease associations based on improved microRNA and disease similarities. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 15, 1774–1782. doi: 10.1109/TCBB.2016.2586190

Lan, W., Wang, J., Li, M., Peng, W., and Wu, F. (2015b). Computational approaches for prioritizing candidate disease genes based on PPI networks. *Tsinghua Sci. Technol.* 20, 500–512. doi: 10.1109/TST.2015.7297749

Lan, W., Wu, X., Chen, Q., Peng, W., Wang, J., and Chen, Y. P. (2022b). GANLDA: graph attention network for lncRNA-disease associations prediction. *Neurocomputing* 469, 384–393. doi: 10.1016/j.neucom.2020.09.094

Lan, W., Zhang, H., Dong, Y., Chen, Q., Cao, J., Peng, W., et al. (2022c). DRGCNCDA: predicting circRNA-disease interactions based on knowledge graph and disentangled relational graph convolutional network. *Methods* 208, 35–41. doi: 10.1016/j.ymeth.2022.10.002

Lan, W., Zhu, M., Chen, Q., Chen, B., Liu, J., Li, M., et al. (2020c). CircR2Cancer: A manually curated database of interactions between circRNAs and cancers. Database, 2020

Lan, W., Zhu, M., Chen, Q., Chen, J., Ye, J., Liu, J., et al. (2021b). Prediction of circRNA-miRNA associations based on network embedding. *Complexity* 2021, 1–10. doi: 10.1155/2021/6659695

Lei, X., Mudiyanselage, T. B., Zhang, Y., Bian, C., Lan, W., Yu, N., et al. (2020). A comprehensive survey on computational methods of non-coding RNA and disease interaction prediction. *Brief. Bioinform.*

Liu, M., Wang, Q., Shen, J., Yang, B. B., and Ding, X. (2019). Circbank: a comprehensive database for circRNA with standard nomenclature. *RNA Biol.* 16, 899–905. doi: 10.1080/15476286.2019.1600395

Nazarov, I., Shirokikh, B., Burkina, M., Fedonin, G., and Panov, M. (2018). Sparse group inductive matrix completion. arXiv [Preprint]. doi: 10.48550/arXiv.1804.10653

Qian, Y., Zheng, J., Jiang, Y., Li, S., and Deng, L. (2022). Prediction of circRNA-miRNA association using singular value decomposition and graph neural networks. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 1, 1–9. doi: 10.1109/TCBB.2022.3222777

Qian, Y., Zheng, J., Zhang, Z., Jiang, Y., Zhang, J., and Deng, L. (2021). "CMIVGSD: circRNA-miRNA interaction prediction based on Variational graph auto-encoder and singular value decomposition." in 2021 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE, 205–210.

Qiu, J., Dong, Y., Ma, H., Li, J., Wang, K., and Tang, J. (2018). "Network embedding as matrix factorization: unifying deepwalk, line, pte, and node2vec." in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 459–467.

Qu, S., Yang, X., Li, X., Wang, J., Gao, Y., Shang, R., et al. (2015). Circular RNA: a new star of noncoding RNAs. *Cancer Lett.* 365, 141–148. doi: 10.1016/j.canlet.2015.06.003

Rophina, M., Sharma, D., Poojary, M., and Scaria, V. (2020). Circad: A comprehensive manually curated resource of circular RNA associated with diseases. Database, 2020.

Rybak-Wolf, A., Stottmeister, C., Glažar, P., Jens, M., Pino, N., Giusti, S., et al. (2015). Circular RNAs in the mammalian brain are highly abundant, conserved, and dynamically expressed. *Mol. Cell* 58, 870–885. doi: 10.1016/j.molcel.2015.03.027

Si, S., Chiang, K. Y., Hsieh, C. J., Rao, N., and Dhillon, I. S. (2016). "Goal-directed inductive matrix completion." in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1165–1174.

Sun, M., Zhao, W., Chen, Z., Li, M., Li, S., Wu, B., et al. (2019a). Circ_0058063 regulates CDK6 to promote bladder cancer progression by sponging miR-145-5p. *J. Cell. Physiol.* 234, 4812–4824. doi: 10.1002/jcp.27280

Sun, M., Zhao, W., Chen, Z., Li, M., Li, S., Wu, B., et al. (2019b). Circular RNA CEP128 promotes bladder cancer progression by regulating Mir-145-5p/Myd88 via MAPK signaling pathway. *Int. J. Cancer* 145, 2170–2181. doi: 10.1002/ijc.32311

Vural, H., Kaya, M., and Alhajj, R. (2019). "A model based on random walk with restart to predict circRNA-disease interactions on heterogeneous network." in *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 929–932.

Wang, X. F., Yu, C. Q., Li, L. P., You, Z. H., Huang, W. Z., Li, Y. C., et al. (2022). KGDCMI: a new approach for predicting circRNA–miRNA interactions from multi-source information extraction and deep learning. *Front. Genet.* 13:958096. doi: 10.3389/fgene.2022.958096

Wilusz, J. E., and Sharp, P. A. (2013). A circuitous route to noncoding RNA. *Science* 340, 440–441. doi: 10.1126/science.1238522

Xie, F., Li, Y., Wang, M., Huang, C., Tao, D., Zheng, F., et al. (2018). Circular RNA BCRC-3 suppresses bladder cancer proliferation through miR-182-5p/p27 axis. *Mol. Cancer* 17, 1–12. doi: 10.1186/s12943-018-0892-z

Yao, Z., Xu, R., Yuan, L., Xu, M., Zhuang, H., Li, Y., et al. (2019). Circ_0001955 facilitates hepatocellular carcinoma (HCC) tumorigenesis by sponging miR-516a-5p to release TRAF6 and MAPK11. *Cell Death Dis.* 10, 1–12. doi: 10.1038/s41419-019-2176-y

Yu, C. Q., Wang, X. F., Li, L. P., You, Z. H., Huang, W. Z., Li, Y. C., et al. (2022). SGCNCMI: a new model combining multi-modal information to predict circRNA-related miRNAs, diseases and genes. *Biology* 11:1350. doi: 10.3390/biology11091350

Zhang, Z., Wang, C., Zhang, Y., Yu, S., Zhao, G., and Xu, J. (2020). CircDUSP16 promotes the tumorigenesis and invasion of gastric cancer by sponging miR-145-5p. *Gastric Cancer* 23, 437–448. doi: 10.1007/s10120-019-01018-7

Zhang, W., Yu, C., Wang, X., and Liu, F. (2019). Predicting CircRNA-disease interactions through linear neighborhood label propagation method. *IEEE Access* 7, 83474–83483. doi: 10.1109/ACCESS.2019.2920942

Zheng, Y., Chen, C. J., Lin, Z. Y., Li, J. X., Liu, J., Lin, F. J., et al. (2020). Circ_KATNAL1 regulates prostate cancer cell growth and invasiveness through the miR-145-3p/WISP1 pathway. *Biochem. Cell Biol.* 98, 396–404. doi: 10.1139/bcb-2019-0211

Zheng, F., and Xu, R. (2020). CircPVT1 contributes to chemotherapy resistance of lung adenocarcinoma through miR-145-5p/ABCC1 axis. *Biomed. Pharmacother.* 124:109828. doi: 10.1016/j.biopha.2020.109828

Zhu, Z., Chang, F., Liu, J., Wang, J., and Zhang, X. (2020). Comprehensive circular RNA profiling reveals the regulatory role of circ_100242/miR-145 pathway in bladder cancer. *Oncol. Lett.* 19, 2971–2978. doi: 10.3892/ol.2020.11380

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 22 December 2022
DOI 10.3389/fmicb.2022.1074398

OPEN ACCESS

EDITED BY
Qi Zhao,
University of Science and Technology Liaoning, China

REVIEWED BY
Luís Fernando Parizi,
Federal University of Rio Grande do Sul, Brazil
Chaoguang Tian,
Tianjin Institute of Industrial Biotechnology (CAS), China

*CORRESPONDENCE
Jianwei He
jwhe@lnu.edu.cn
Nan Ma
manan@motcats.ac.cn
Jingjing Jing
hellojjjing@163.com

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

RECEIVED 19 October 2022
ACCEPTED 28 November 2022
PUBLISHED 22 December 2022

CITATION
Niu T, Cui Y, Shan X, Qin S, Zhou X, Wang R, Chang A, Ma N, Jing J and He J (2022) Comparative transcriptomic analysis-based identification of the regulation of foreign proteins with different stabilities expressed in *Pichia pastoris*. *Front. Microbiol.* 13:1074398. doi: 10.3389/fmicb.2022.1074398

Comparative transcriptomic analysis-based identification of the regulation of foreign proteins with different stabilities expressed in *Pichia pastoris*

Tingting Niu[1†], Yi Cui[1†], Xu Shan[1], Shuzhen Qin[1], Xuejie Zhou[1], Rui Wang[1], Alan Chang[2], Nan Ma[3*], Jingjing Jing[4*] and Jianwei He[1*]

[1]School of Life Sciences, Liaoning University, Shenyang, China, [2]College of Life and Environmental Sciences, Wenzhou University, Wenzhou, China, [3]China Academy of Transportation Sciences, Beijing, China, [4]Tumor Etiology and Screening Department of Cancer Institute and General Surgery, The First Hospital of China Medical University, Shenyang, China

Introduction: The industrial yeast *Pichia pastoris* is widely used as a cell factory to produce proteins, chemicals and advanced biofuels. We have previously constructed *P. pastoris* strains that overexpress protein disulfide isomerase (PDI), which is a kind of molecular chaperone that can improve the expression of an exogenous protein when they are co-expressed. Chicken cystatin (cC) is a highly thermostable cysteine protease inhibitor and a homologous protein of human cystatin C (HCC). Wild-type cC and the two mutants, I66Q and ΔW (a truncated cC lacking the á-helix 2) represent proteins with different degrees of stability.

Methods: Wild-type cC, I66Q and ΔW were each overexpressed in *P. pastoris* without and with the coexpression of PDI and their extracellular levels were determined and compared. Transcriptomic profiling was performed to compare the changes in the main signaling pathways and cell components (other than endoplasmic reticulum quality control system represented by molecular chaperones) in *P. pastoris* in response to intracellular folding stress caused by the expression of exogenous proteins with different stabilities. Finally, hub genes hunting was also performed.

Results and discussion: The coexpression of PDI was able to increase the extracellular levels of both wild-type cC and the two mutants, indicating that overexpression of PDI could prevent the misfolding of unstable proteins or promote the degradation of the misfolded proteins to some extent. For *P. pastoris* cells that expressed the I66Q or ΔW mutant, GO (Gene Ontology) and KEGG (Kyoto Encyclopedia of Genes and Genomes) analyses of the common DEGs in these cells revealed a significant upregulation of the genes involved in protein processing, but a significant downregulation of the genes enriched in the Ribosome, TCA and Glycolysis/Gluconeogenesis pathways. Hub genes hunting indicated that the most downregulated ribosome protein, C4QXU7 in this case, might be an important target protein that could be

manipulated to increase the expression of foreign proteins, especially proteins with a certain degree of instability.

Conclusion: These findings should shed new light on our understanding of the regulatory mechanism in yeast cells that responds to intracellular folding stress, providing valuable information for the development of a convenient platform that could improve the efficiency of heterologous protein expression in *P. pastoris*.

KEYWORDS

cystatin, transcriptomic, amyloid disease, *pichia pastoris*, biofuel

1 Introduction

Human cystatin C (HCC) is a papain-like cysteine protease inhibitor that belongs to the cystatin superfamily, and it is also one of the most extensively studied endogenous inhibitors as well as an important biomarker of renal function (Dubin, 2005). Abnormal changes in the expression and secretion of HCC in the brain have been described for various neurological disorders such as amyotrophic lateral sclerosis (ALS), rare heritable neurodegenerative disorders, ischemia, some forms of epilepsy, Alzheimer's disease (AD) (Mathews and Levy, 2016) and recurrent hemorrhagic stroke (Merz et al., 1997; Zhou et al., 2015).

Previous studies have reported that the fatal amyloid disease, hereditary cystatin amyloid angiopathy (HCCAA), found in young Icelanders is mainly caused by the HCC hereditary amyloidogenic mutant L68Q, which has a high dimerization potential that can lead to self-aggregation and hyper-amyloidosis (Janowski et al., 2001; Palsdottir et al., 2006). The instability of the soluble HCC monomer has constrained any structural studies on its physicochemical properties. Meanwhile, chicken cystatin (cC) has a number of characteristics similar to HCC, and both proteins share about 44% sequence homology. Thus, cC is considered an ideal model for studying protein domain exchange and amyloid-related diseases (Bode et al., 1988; Yu et al., 2010). Residue 66 in cC corresponds to residue 68 in HCC, and the I66Q mutant of cC has similar amyloidogenic properties to L68Q of HCC under physiological conditions (Bjarnadottir et al., 2001).

The AS (appending structure) region of cC contains α-helix 2, which is crucial for the stability of cC and is considered to be the biggest difference between HCC and cC (Grubb et al., 1984). Therefore, the α-helix 2-truncated mutant (ΔW) with a deletion at residues 77–85 was constructed as the unstable cC model protein. Based on our previous results, the secreted amount of ΔW is much lower than that of WT cC or I66Q when expressed in *P. pastoris*, indicating that the absence of α-helix 2 in the AS region may be one of the factors contributing to the structural instability of HCC (Zhou et al., 2019).

Pichia pastoris (reclassified as *Komagataella phaffii/pastoris*) is a methylotrophic yeast and a highly successful system for producing recombinant proteins in the pharmaceutical and biofuel industries (Yu et al., 2017). The ability of *P. pastoris* to express recombinant proteins is facilitated by the strong promoter of its alcohol oxidase 1 (AOX1) gene. The activity of the AOX1 promoter is tightly regulated by the carbon source. Thus, recombinant proteins expressed from the AOX1 promoter in *P. pastoris* cells can be induced with methanol once cell growth has reached high densities to obtain a high level of expression for the proteins. For example, a His-Qtagged lipase A from *Beauveria bassiana* has been successfully produced in *P. pastoris* and shown to have potential use for biodiesel production *via* ethanolysis (Vici et al., 2015). Another example is the expression of α-L-arabinofuranosidase (ARA) in *P. pastoris*, which can be improved 5.5-fold by codon optimization. The recombinant ARA has significant potential in the catalytic conversion of corn stover to fermentable sugars during biofuel production (Vici et al., 2015). However, overexpression of recombinant proteins may lead to more misfolded proteins and trigger endoplasmic reticulum (ER) stress (Jolly and Morimoto, 2000; Oakes and Papa, 2015). The cells might then respond to ER stress by increasing the expression of some molecular chaperones, including PDI, HSP90, and HSP72 (Vogl and Glieder, 2013; Delic et al., 2014; Gu et al., 2015). To prevent protein misfolding and aggregation, the newly synthesized molecular chaperones would increase folding efficiency by capturing the folded intermediates and promoting refolding or degradation. Co-expression of PDI has been used to improve the expression of heterologous proteins in *P. pastoris* by overcoming the burden of protein folding and secretion (Inan et al., 2006; Navone et al., 2021a). In our previous study, the overexpression of PDI in *P. pastoris* GS115 strains was found to significantly increase the expression of cC (Zhou et al., 2019). On this basis, three representative proteins with different stabilities (WT cC and its two mutants I66Q and ΔW) (**Figure 1**) were used as model proteins to screen for factors other than the ER quality control system represented by the molecular chaperone PDI that could influence the expression of foreign proteins that may not be properly folded in *P. pastoris*. Subsequently, transcriptomic profiling was performed to identify the transcriptomic changes

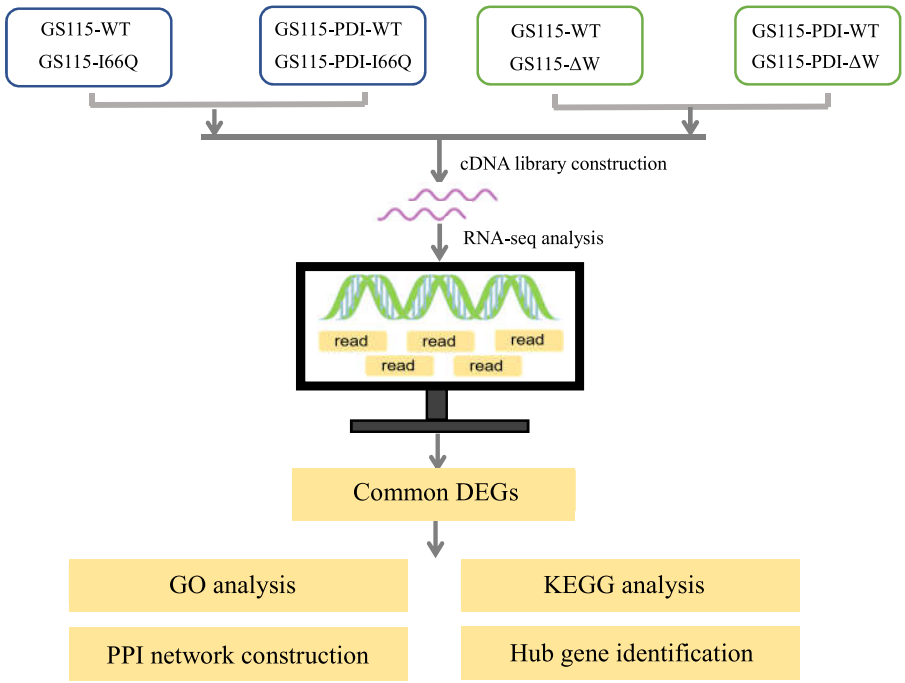

FIGURE 1
Schematic diagram of the workflow conducted in this study.

and pathways involved in the molecular network and changes in the dynamic mechanism of the foreign protein secretion pathway in *P. pastoris*.

2 Materials and methods

2.1 Strains, plasmids, and culture conditions

Pichia Pastoris GS115 strain was provided by Dr. Shutao Liu at Fuzhou University. The plasmids pPIC3.5K and pPICZαA were purchased from Invitrogen. GS115 strain and the previously constructed GS115 PDI-overexpressing strain were used as starting strains for the construction of the PDI and cC co-overexpressing strains. The yeast cells were first cultured at 30°C in Yeast Peptone Dextrose medium (YPD) (1% yeast extract, 2% peptone, and 2% glucose) to logarithmic growth phase (OD_{600} = 5.0) followed by methanol induction in Yeast Extract Peptone Medium (YPM) [1% yeast extract, 2% peptone, 0.5% Methanol (v/v)] for 72 h to induce the expression of PDI and cC.

2.2 Construction of recombinant strains

GS115 competent cells were transformed with the linearized plasmid pPICZαA-cC, pPICZαA-I66Q, and pPICZαA-ΔW by electroporation to generate GS115-cC, GS115-I66Q, GS115-ΔW recombinant strains, respectively. Similarly, GS115-PDI-cC, GS115-PDI-I66Q, GS115-PDI-ΔW strains were obtained by transforming GS115 PDI with pPICZαA-cC, pPICZαA-I66Q, and pPICZαA-ΔW, respectively.

2.3 Protein expression analysis

Extracellular protein samples were obtained as previously described (Zhou et al., 2019). Intracellular protein samples were extracted from yeast cells after treatment with Yeast Protein Extraction Reagent (Takara, Dalian, China). SDS-PAGE and western blotting were carried out following the procedure described previously (Zhou et al., 2019). The protein bands in one gel were visualized by staining the gel with PAGE Gel Silver Staining Kit (Takara, Beijing, China) whereas the protein bands in the other gel were transferred to a PVDF membrane (Millipore, MA, USA) for western blot analysis. After protein transfer, the PVDF membrane was incubated in a blocking buffer containing TBST plus 5% skimmed milk powder for 2 h. This was followed by three 10 min washes in TBST buffer, and 1 h of incubation in rabbit anti-cC antiserum (1:2000) at room temperature. After that, the membrane was again washed three times in TBST, with each wash lasting for 10 min. Finally, the blot was incubated with anti-rabbit peroxidase conjugate (1:10000) for 1 h at RT, and then subjected

to detection using the eECL Reagent (Beyotime, Shanghai, China).

2.4 RNA sequencing

Total RNA was extracted from the different GS115 strains using Yeast RNAiso Kit (Takara, Dalian, China). The extraction was performed according to the manufacturer's protocol. The mRNA fraction was purified from the total RNA using MicroPoly Purist kit (Takara, Dalian, China) according to the manufacturer's protocol. The concentration and integrity of the mRNA were measured using a NanoDrop 2000 (Thermo Fisher Scientific, MA, USA) and the Agilent 2100 LabChip system (Agilent Technologies, CA, USA). The RNA was sheared, and reverse transcribed using random primers to obtain the cDNA, which was then used for the construction of a cDNA library. Illumina RNA sequencing (RNA-Seq) libraries were subsequently performed using VAHTS Universal V6 RNA-seq Library P the SMARTer Stranded RNA Seq Kit (Vazyme Biotech Co. Ltd.) according to the manufacturer's instructions. Finally, RNA-Seq data were generated in Fastq format. The sequencing data have been submitted to National Center for Biotechnology Information (NCBI) under accession PRJNA892887[1].

2.5 Screening for differentially expressed genes

The differentially expressed genes (DEGs) for WT vs. I66Q, PDI-WT vs. PDI-I66Q, WT vs. ΔW and PDI-WT vs. PDI-ΔW were identified by the BioMarker cloud platform with adjusted p-value <0.01 and $\log_2$ fold change (FC) >2. Moreover, the common DEGs between different groups have been identified by the same method.

2.6 Functional enrichment analyses for common DEGs

Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses were performed on the BioMarker cloud platform with a p-value <0.05. ClueGo plug-in in Cytoscape software (3.8 version) was used for showing the ClueGO network diagram.

1 https://submit.ncbi.nlm.nih.gov/subs/bioproject/SUB12189176/overview

2.7 Protein–protein interaction network construction and hub gene identification

Common DEGs were used to construct the protein–protein interaction (PPI) network by using the SRTING online database with a confidence score of more than 0.7. Hub genes of the PPI network were identified using a degree algorithm from cytoHubba, a plugin in Cytoscape, and visualized using Cytoscape (v3.8.0).

3 Results

3.1 Effect of PDI-overexpression on the expression of WT cC and cC mutants

To investigate the intracellular distribution pattern and retention level of recombinant cC as well as its extracellular secretion in *P. pastoris*, WT cC, I66Q and ΔW were expressed in both *P. pastoris* GS115 strains without and with the overexpression of PDI. The extracellular secretion of wild-type cC and its two mutants was detected by silver staining (**Figure 2A**). Overexpression of PDI (57 kDa) can significantly enhance the expression of WT cC (14 kDa) and the I66Q mutant. In the case of ΔW, the protein was only secreted when it was co-expressed with PDI (**Figure 2B**). As shown in **Figure 2B**, almost no I66Q and ΔW were detected as an intracellular form in both PDI-overexpressing GS115 and wild-type GS115 strains. Interestingly, for WT cC, no significant difference in intracellular level was observed between the two yeast strains. This might suggest that when the yeast expressed WT cC, most of the proteins were capable of folding into the native form and were subsequently secreted out of the cell, with little misfolded protein being produced and residing in the ER despite an increase in the amount of the newly synthesized protein entering the ER. For the mutant I66Q, its intracellular level was proportional to its extracellular level in either yeast strain because of its amyloidogenic properties. Consequently, the three cC-overexpressing GS115 strains and the three cC-overexpressing GS115 strains that also co-expressed PDI were used for the following transcriptomic studies.

3.2 Transcriptomic analysis of different recombinant GS115 strains

It has been proven that genes are being expressed at different levels in different individual organisms as a result of biological variability (Robasky et al., 2014). Therefore, biological replicates were included to ensure the validity of the following experiments. The Pearson Correlation Coefficient r refers to

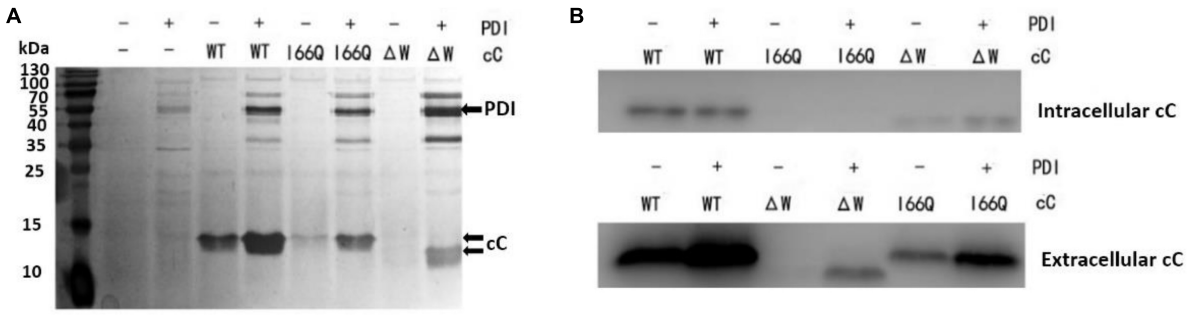

FIGURE 2
Analysis of the expression of the different versions of chicken cystatin (cC) (MW: 13 kDa) in both GS115 strains that did and did not overexpress PDI (MW: 57 kDa). **(A)** Secretion of the WT cC, I66Q and ΔW as detected in the culture supernatant after centrifugation as shown by SDS-PAGE. The gel was visualized by silver staining. **(B)** Expression of WT cC, I66Q and ΔW in GS115 as analyzed by western blot. Intracellular refers to the soluble fraction of cell lysate, and Extracellular refers to the culture supernatant. "+" and "−" indicates the cells were transfected with and without the corresponding cC-coding gene, respectively.

TABLE 1 Summary of the samples analyzed by RNA—sequencing (RNA-seq).

Samples for RNA-seq	Strain	Overexpressed protein
T1, T2, T3	GS115	/
T4, T5, T6	GS115-PDI	PDI
T7, T8, T9	GS115-cC	cC
T10, T11, T12	GS115-PDI-cC	PDI/cC
T13, T14, T15	GS115-I66Q	I66Q
T16, T17, T18	GS115-PDI-I66Q	PDI/I66Q
T19, T20, T21	GS115-ΔW	ΔW
T22, T23, T24	GS115-PDI-ΔW	PDI/ΔW

the biological assessment of repeated samples and it was used to analyze the correlation between every two samples (Schulze et al., 2012). The closer r^2 is to 1, the stronger the correlation between the two replicates (Figure 3). Subsequently, 24 samples were subjected to the transcriptomic analysis after RNA-sequencing was completed, with three replicates included for each strain and the data are summarized in Table 1. The clean reads of each sample were compared with the designated *P. pastoris* GS115 genome. The mapped data obtained after alignment were used to evaluate the quality of the library such as randcheck, insert size, and saturation test. Typically, FPKM (Fragments Per Kilobase of transcript per Million fragments mapped) was used as an indicator to measure the expression level of a gene (Zhao et al., 2021). Identification of DEGs was carried out according to the gene expression levels in different samples.

Different numbers of up-and down-regulated DEGs were obtained by comparing each of the two sample groups as shown in Figure 4. Interestingly, when the WT cC-overexpressing strain was compared with either the I66Q- or ΔW-overexpressing strain, the DEGs, especially the up-regulated genes, were significantly increased. Nevertheless, this trend was not observed in the comparison of PDI-WT vs. PDI-I66Q/ΔW, implying a healthy intracellular cell condition in GS115-PDI-I66Q and GS115-PDI-ΔW afforded by the overexpression of PDI. From this point, it became important to investigate the common DEGs identified from the comparison of both WT vs. I66Q/ΔW and PDI-WT vs. PDI-I66Q/ΔW, since these common DEGs are vital for the overexpression of foreign proteins that are less stable. The total number of up-regulated and down-regulated genes in each comparison group was depicted in a Venn diagram (Figure 4B). Among all the compared genes, 203 and 210 common DEGs indicated in Figure 4B were selected for subsequent KEGG analysis, GO annotation, and hub gene identification.

3.3 KEGG analysis of common DEGs

Kyoto encyclopedia of genes and genomes analysis was applied to explore the potential molecular functions and molecular mechanisms associated with the functions of the common DEGs. As shown in Figure 5A, for WT vs. I66Q and PDI-WT vs. PDI-I66Q comparisons, several signaling pathways were significantly enriched, including the pathways for the biosynthesis of amino acids, citrate cycle, and Tricarboxylic Acid cycle (TCA cycle), and the metabolic pathways. In the comparison of WT vs. ΔW and of PDI-WT vs. PDI-ΔW, in addition to the pathways that were enriched in both WT vs. I66Q and PDI-WT vs. PDI-I66Q, the DNA replication and ribosome pathways were also indicated, suggesting that the activation of genes in different KEGG pathways may be due to the different physicochemical properties of the I66Q and ΔW cC mutants (Figure 5B).

Enrichment analysis was performed on the 203 and 210 common DEGs using the ClueGO v2.5.4 plugin. After setting

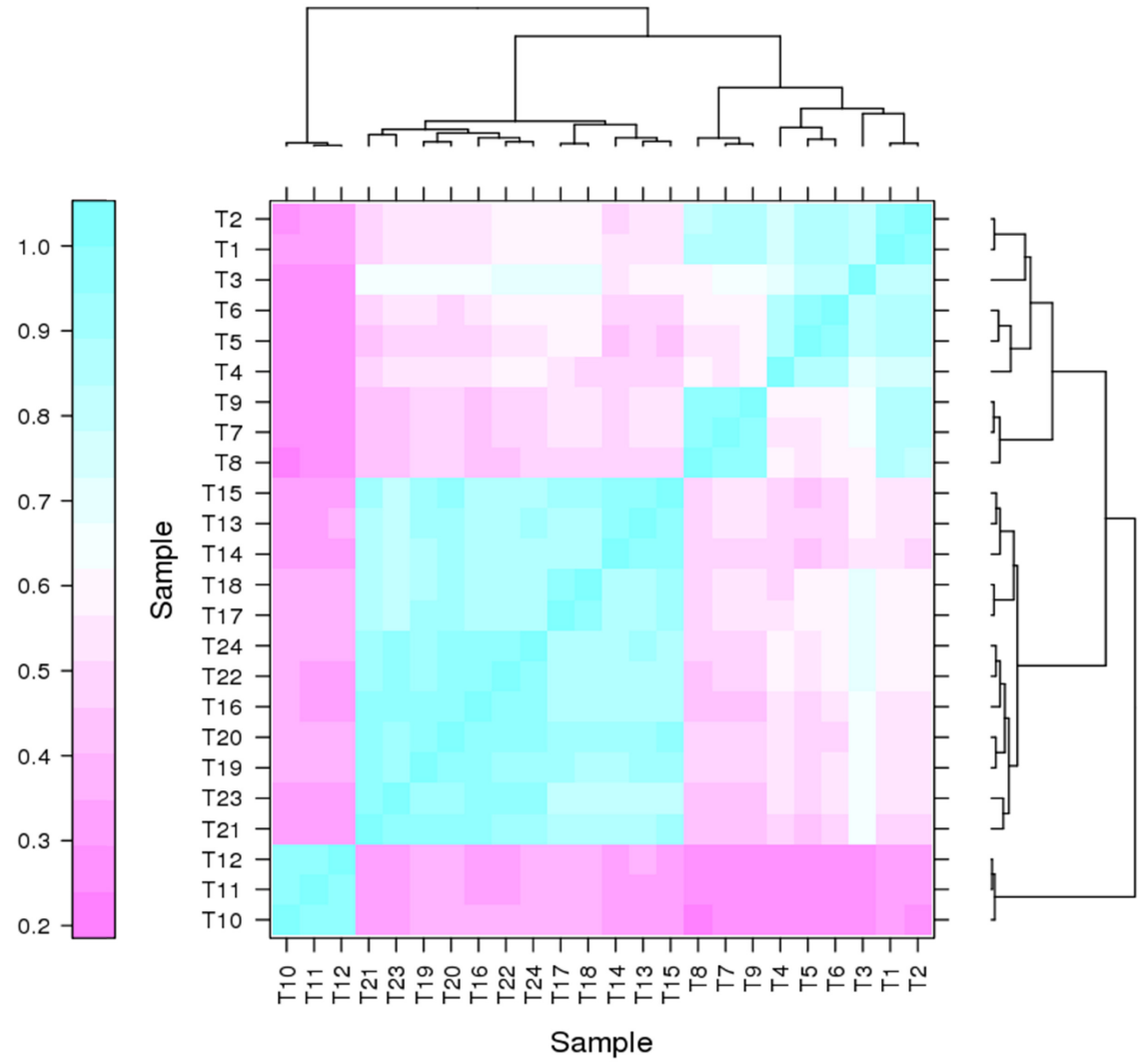

FIGURE 3
Correlation heat map for two pairs of samples.

the *P*-value as <0.05 and the Kappa Score Threshold as 0.4, 25, and 24 GO terms were identified from the group WT vs. I66Q and PDI-WT vs. PDI-I66Q, WT vs. ΔW and PDI-WT vs. PDI-ΔW, respectively. For the comparison of WT vs. I66Q and PDI-WT vs. PDI-I66Q, the enriched genes were mainly involved in protein folding, translation, and the acetyl-CoA metabolic and monocarboxylic acid biosynthetic processes (**Figure 6A**). For WT vs. ΔW and PDI-WT vs. PDI-ΔW, the enriched genes were not only distributed in the above pathways but also in the DNA metabolic process and porphyrin-containing compound biosynthetic process (**Figure 6B**). The result obtained from ClueGO enrichment analysis provided a global understanding of the scenario when proteins with different stabilities were expressed in *P. pastoris*. The common DEGs enriched in both the metabolic and protein processing pathways were quite noticeable in the comparison of WT vs. I66Q and that of WT vs. ΔW, suggesting that when amyloid mutants and unstable exogenous proteins are expressed in the *P. pastoris*, there might be a need to adjust the basic metabolic speed/efficiency and reproduction speed of the cells.

3.4 PPI network construction and hub genes identification

To further investigate the key cellular components and biological processes in the wild-type GS115 (I66Q/ΔW) and PDI-overexpressing GS115 strains (I66Q/ΔW), both of which were found to have common enriched DEGs, the STRING (Search Tool for the Retrieval of Interacting Genes) online tool was used to construct a PPI network of common DEGs. A combined score of >0.7 was set as the cut-off criterion for

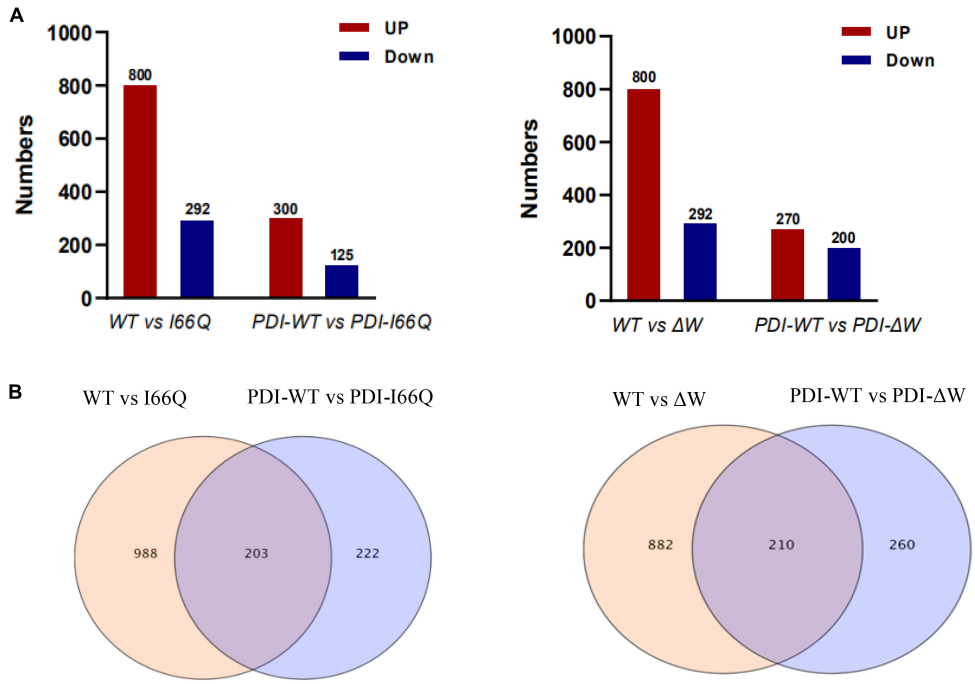

FIGURE 4
Count of differentially expressed genes (DEGs) in WT vs. I66Q/ΔW and PDI-WT vs. PDI-I66Q/ΔW comparisons. **(A)** Total number of up and downregulated DEGs in the four pairs: WT vs. I66Q, WT vs. ΔW, PDI-WT vs. PDI-I66Q and PDI-WT vs. PDI-ΔW. **(B)** Venn diagram depicting the total number of proteins, including the up and down-regulated genes in the four groups.

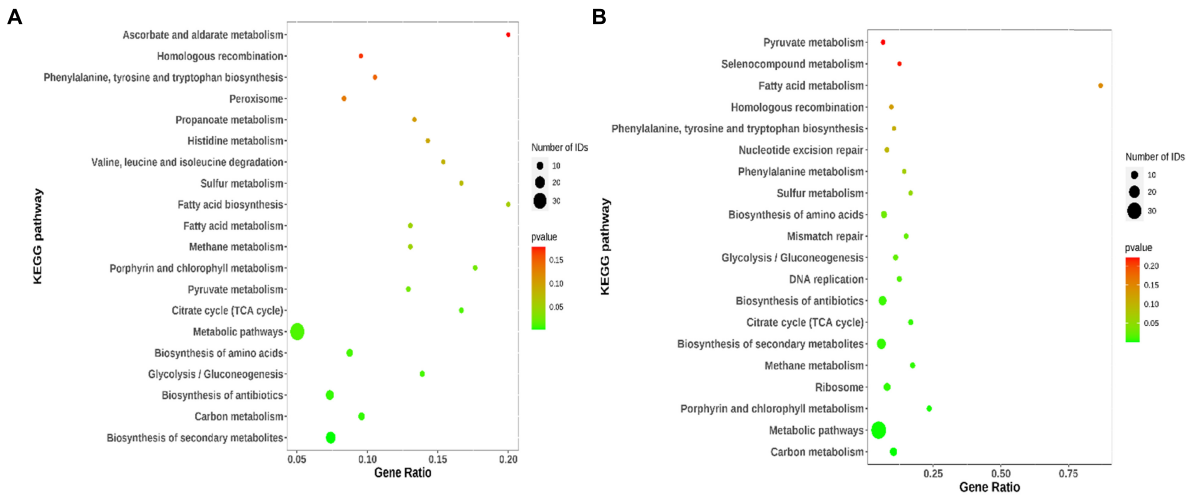

FIGURE 5
Kyoto encyclopedia of genes and genomes (KEGG) pathway enrichment analysis of common differentially expressed genes (DEGs). **(A)** KEGG pathway enrichment of 203 common DEGs identified from the WT vs. I66Q and PDI-WT vs. PDI-I66Q comparisons. **(B)** KEGG pathway enrichment of 210 common DEGs from the WT vs. ΔW and PDI-WT vs. PDI-ΔW comparisons.

statistical significance. Next, the PPI network was downloaded and visualized as shown in Figure 7. A total of 132 nodes and 101 edges were identified for the common genes belonging to the comparison of WT vs. I66Q and of PDI-WT vs. PDI-I66Q (Figure 7A). According to the Degree sores in the cytoHubba, the top ten highest-scored genes were selected as the hub genes, including the common DEGs that encode the proteins C4QXU7, C4QZL4, C4R7T6, C4R447, C4R196, C4R0F8, C4R7T7, C4R7D3, C4R596 (Figure 7B). As for the comparison of WT vs. ΔW and of PDI-WT vs. PDI-ΔW,

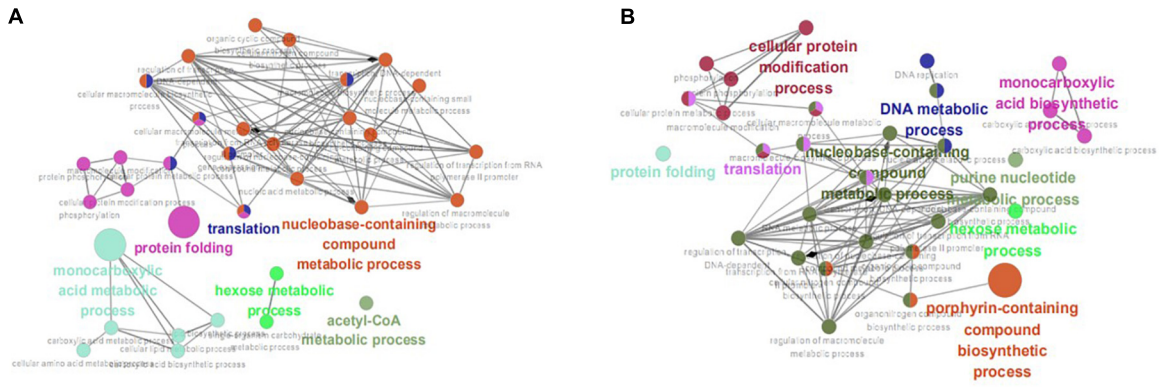

FIGURE 6
ClueGO enrichment analysis. **(A)** Significantly enriched gene ontology (GO) terms of the common differentially expressed genes (DEGs) identified from the WT vs. I66Q and PDI-WT vs. PDI-I66Q comparisons. **(B)** Significantly enriched GO terms of the common DEGs from the WT vs. ΔW and PDI-WT vs. PDI-ΔW comparisons.

140 nodes and 118 edges were found in the PPI network (Figure 7C), and the Hub genes included the common DEGs that encode C4QXU7, C4QVA2, C4R0T8, C4R447, C4QZE7, C4R7A1, C4R8M3, C4QWG3, C4R7T6, C4R7T6, C4QWY6 (Figure 7D).

C4QXU7, C4R7T6, and C4R0F8 were the top three key proteins identified in the comparison of WT vs. I66Q and of PDI-WT vs. PDI-I66Q. C4QXU7 is a small subunit of ribosomal protein S28e, whereas C4R7T6 is the large subunit of the ribosomal protein LP1, and C4R0F8 is a small subunit of the ribosomal protein S26e. In the case of WT vs. ΔW and PDI-WT vs. PDI-ΔW comparisons, the top three key proteins included C4QXU7, C4QVA2, and C4R0T8. C4QVA2 is a small subunit of ribosomal protein S3e, and C4R0T8 is a small subunit of ribosomal protein S15A. All of the above-mentioned proteins are responsible for the structural integrity of the ribosome and the protein translation rate. Interestingly, all the top three hub genes belong to the components of ribosomal subunits and they were found to be downregulated. Moreover, they were all found to be enriched in the translation pathway. This observation underlined the importance of the rate of protein synthesis by ribosomes and suggested that protein synthesis by ribosomes may be the ultimate goal that yeast cells need to adjust when an exogenous unstable protein is expressed in *P. pastoris*.

4 Discussion

It is generally believed that the amount of secreted amyloid mutant proteins and unstable mutant proteins is usually lower than that of their wild-type counterparts when eukaryotic expression systems are used (Whyteside et al., 2011; Bou Ali et al., 2013; He et al., 2019). When yeast cells express foreign proteins, the misfolded proteins may become toxic to the cells, and therefore, the co-overexpression of molecular chaperones may promote the folding and posttranslational modifications of the expressed proteins, thereby facilitating the secretion of the foreign proteins. Our results showed that co-expression of PDI with wild-type cC or mutated cC (I66Q or ΔW) could improve the secretion of cC to different extents. However, the distinct client-recruiting system of molecular chaperones may have a limit in improving the secretion of foreign proteins that may not be properly folded (Yan et al., 2021). This limitation could prevent *P. pastoris* from becoming an ideal protein production platform that can accommodate a variety of production requirements. Although modifications of specific transcription factors have been adopted to alter the regulation mode, it is still difficult to markedly increase the expression of foreign proteins in wild-type *P. pastoris* (Nusse and Lindau, 1988; Wang et al., 2017; Vogl et al., 2018). In this context, it is essential to pursue other regulatory genes and related cellular pathways that are capable of increasing the expression of foreign proteins in *P. pastoris*. Accordingly, WT, I66Q and ΔW were designed to be secreted by *P. pastoris* cells that did not overexpress PDI and those that did overexpress PDI, since both cell types could achieve different levels of protein expression in response to the various stabilities of WT cC and the two cC mutants. Our results, along with previously reported results, have shown that following their expression in *P. pastoris*, improperly folded foreign proteins may be found in lower levels compared with those that are properly folded, such as the improperly folded cC mutants versus their wide-type counterpart. Interestingly, the two cC mutants had different stabilities, which directly impacted their extracellular and intracellular levels. Consequently, the order of decreasing stability for WT cC, I66Q and ΔW turned out to be well-suited for establishing *P. pastoris* strains that could be used to investigate the key regulatory genes that the host cell would modulate when expressing foreign proteins that may not be properly folded.

10.3389/fmicb.2022.1074398

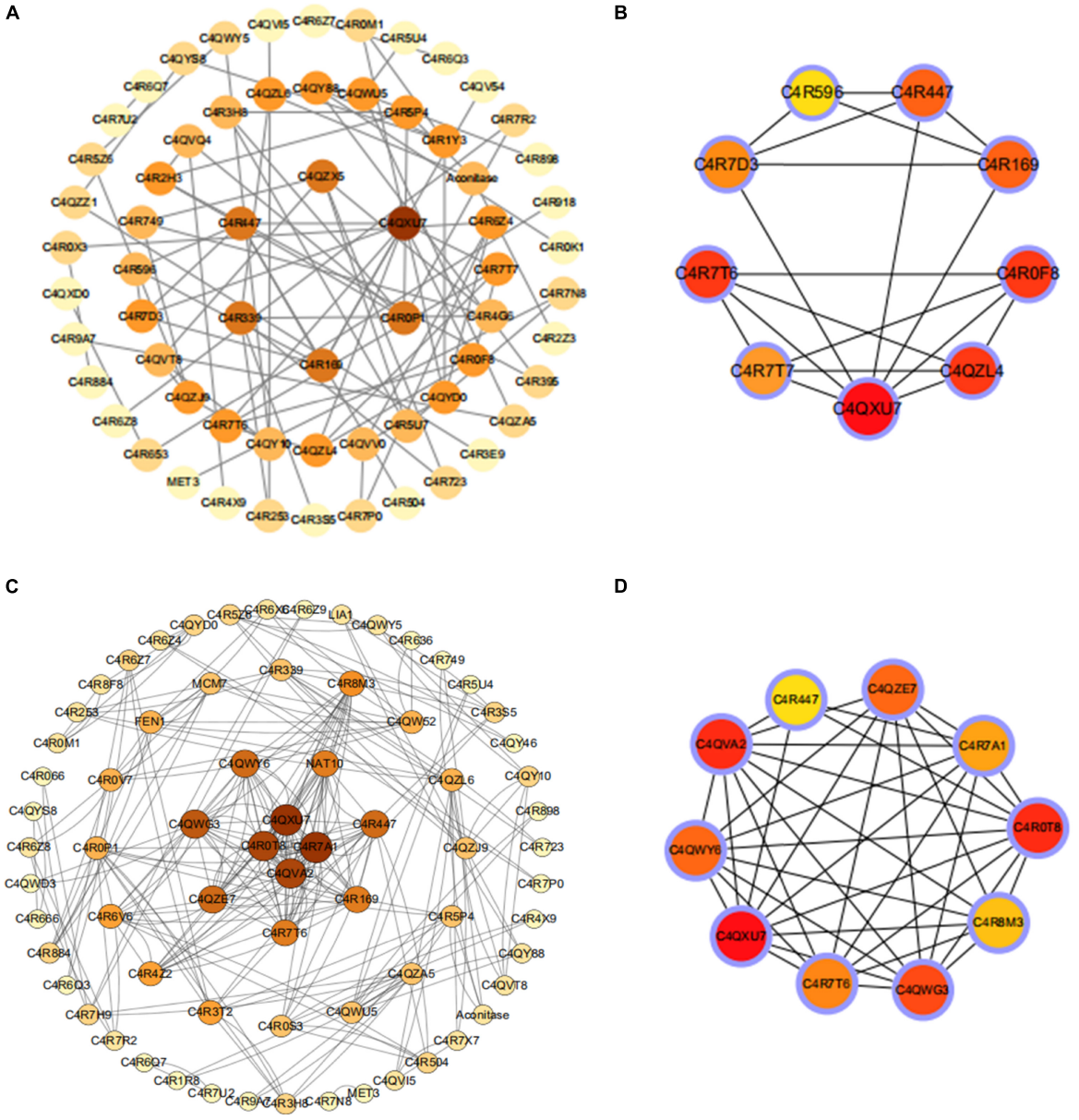

FIGURE 7
Protein–protein interaction (PPI) network of the proteins encoded by the common differentially expressed genes (DEGs) and selection of hub genes. **(A,B)** PPI network and the top 10 common DEGs between WT and I66Q and between PDI-WT and PDI-I66Q. **(C,D)** PPI network and the top 10 common DEGs between the group WT and ΔW and between PDI-WT and PDI-ΔW.

Both *P. pastoris* GS115 which did not overexpress PDI and *P. pastoris* GS115 overexpressed PDI were used as the starting strains to express different variants of cC having different stabilities (WT cC, I66Q and ΔW). RNA-seq analysis was then performed in order to screen for the common DEGs between the WT cC-expressing strain and I66Q-expressing strain or ΔW-expressing strain. For the comparison of WT vs. I66Q with PDI-WT vs. PDI-I66Q, GO analysis indicated the consumption of energy by the host cells as an important factor in the case of foreign protein expression since protein synthesis is an energy-costing process. Glycolysis and the TCA cycle are processes of energy acquisition, so reducing energy acquisition is a way to force the slow-down of protein synthesis because, without ATP and GTP, protein synthesis cannot occur. Glycolysis also leads to the pentose phosphate pathway, which provides nucleotides for nucleic acid synthesis. Thus, lowering glycolysis also affects RNA synthesis, and hence protein synthesis. The result of KEGG pathway analysis showed that in both groups, the common DEGs were enriched in the ribosome and TCA cycle, as well as in

carbon metabolism and glycolysis/gluconeogenesis, consistent with the results of GO analysis. The basal metabolic level of *P. pastoris* is closely related to the protein expression level (Liang et al., 2012; Renuse et al., 2014). When *P. pastoris* expressed I66Q and ΔW during the methanol induction stage, the common DEGs enriched in the TCA and glycolytic pathways were significantly decreased compared with those in the WT cC-expressing strain. At the same time, the common DEGs enriched in the protein translation and protein folding processes were up-regulated, indicating an initial increased expression of protein synthesis-related genes, and even glycolysis and the TCA cycle-related genes in the case of the I66Q- and ΔW-expressing strains. At a later time, the burden caused by these misfolded proteins started to affect the cells and resulted in an adjustment to reduce the expression of the foreign proteins. However, the selective pressure of methanol induction was still ongoing, so protein synthesis could only be slowed down by reducing the flow of energy to protein synthesis (Sola et al., 2007; Orman et al., 2009). These findings indicated that when foreign proteins that may not be properly folded are expressed in *P. pastoris*, the cells need to adjust their own metabolic states in order to maintain intracellular homeostasis, based on the degree of protein instability.

Hub genes are considered to be key genes that play vital roles in biological processes and can affect the regulation of other genes in a related pathway (Liu et al., 2022; Shu et al., 2022). It is of significance to note that the key gene *C4QXU7*, a small subunit of the ribosomal protein S28e, was identified by both the comparisons of WT vs. I66Q/ΔW and PDI-WT vs. PDI-I66Q/ΔW (**Figure 4B**). Yeast ribosomal proteins play important roles in the biogenesis and function of the ribosome (Aw et al., 2017). Deletion of a particular ribosome protein can delay or impair the subunit assembly, indicating that the decelerated elongation stage of translation might promote the co-translational folding rate of heterologous proteins (Liao et al., 2019). In *P. pastoris*, overexpression of xylanase A (a foreign protein) can lead to a significant down-regulation of numerous ribosomal proteins, resulting in decelerated translation elongation and enhanced folding efficiency for xylanase A (Navone et al., 2021b). Meanwhile, studies have shown that knocking out the *C4QXU7* gene in *P. pastoris* does not affect its growth, but can lead to a significant increase in the secretion levels of exogenous proteins (e.g., Pfu and Phytase), indicating that the decelerated elongation rate caused by the loss of C4QXU7 might promote the co-translational folding rate of heterologous proteins, increasing the expression of Pfu and Phytase (Liao et al., 2019). Together with our data on protein expression in *P. pastoris*, these observations could collectively indicate that knocking out the C4QXU7 gene may promote the expression of foreign proteins that are not easily folded in *P. pastoris*.

On the other hand, because *P. pastoris* is widely used for the expression of foreign proteins in industrial protein production, how to design and develop a new *P. pastoris* expression system capable of yielding a high expression level and flexible regulation characteristics is one of the key problems and important goals faced by bioengineering and synthetic biotechnology. In this study, the protein expression levels of three model proteins with the order of decreasing stability WT > I66Q > ΔW were significantly increased in *P. pastoris* GS115 that simultaneously overexpressed PDI. In addition to molecular chaperones, our data also revealed that some ribosomal proteins, e.g., C4QXU7, may also be important targets that can be modulated to increase the expression of foreign proteins. The modulation of key *P. pastoris* ribosomal protein genes will expand its application potential in a broader scenario. From this viewpoint, our research has provided valuable information for developing a convenient platform to improve the efficiency of heterologous protein expression in *P. pastoris*, which may also contribute to the application of synthetic biology in a special field, such as in the field of biofuel production.

Data availability statement

The data presented in this study are deposited in the BioSample database repository of National Center for Biotechnology Information (NCBI), accession number PRJNA892887.

Author contributions

JH and JJ: conceived and designed the research. TN, JH, and YC: wrote the manuscript. SQ and XZ: performed the experiments. TN, YC, and NM: performed the data analyses. XS and RW: provided experiment assistance, data curation, and validation. JH and NM: funding acquisition. AC: writing—review and editing. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by grants from National Natural Science Foundation of China (No. 31670103).

Acknowledgments

We thank Hailong Li and Defu Liu for their kind assistance in the transcriptomic analysis.

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.

Publisher's note

References

Aw, R., Barton, G. R., and Leak, D. J. (2017). Insights into the prevalence and underlying causes of clonal variation through transcriptomic analysis in *Pichia pastoris*. *Appl. Microbiol. Biotechnol.* 101, 5045–5058. doi: 10.1007/s00253-017-8317-2

Bjarnadottir, M., Nilsson, C., Lindstrom, V., Westman, A., Davidsson, P., Thormodsson, F., et al. (2001). The cerebral hemorrhage-producing cystatin C variant (L68Q) in extracellular fluids. *Amyloid* 8, 1–10. doi: 10.3109/13506120108993809

Bode, W., Engh, R., Musil, D., Thiele, U., Huber, R., Karshikov, A., et al. (1988). The 2.0 A X-ray crystal structure of chicken egg white cystatin and its possible mode of interaction with cysteine proteinases. *EMBO J.* 7, 2593–2599. doi: 10.1002/j.1460-2075.1988.tb03109.x

Bou Ali, M., Karray, A., Gargouri, Y., and Ben Ali, Y. (2013). N-terminal domain of turkey pancreatic lipase is active on long chain triacylglycerols and stabilized by colipase. *PLoS One* 8:e71605. doi: 10.1371/journal.pone.0071605

Delic, M., Gongrich, R., Mattanovich, D., and Gasser, B. (2014). Engineering of protein folding and secretion-strategies to overcome bottlenecks for efficient production of recombinant proteins. *Antioxid Redox Signal.* 21, 414–437. doi: 10.1089/ars.2014.5844

Dubin, G. (2005). Proteinaceous cysteine protease inhibitors. *Cell. Mol. Life Sci.* 62, 653–669. doi: 10.1007/s00018-004-4445-9

Grubb, A., Jensson, O., Gudmundsson, G., Arnason, A., Lofberg, H., and Malm, J. (1984). Abnormal metabolism of gamma-trace alkaline microprotein. The basic defect in hereditary cerebral hemorrhage with amyloidosis. *N. Engl. J. Med.* 311, 1547–1549. doi: 10.1056/NEJM198412133112406

Gu, L., Zhang, J., Du, G., and Chen, J. (2015). Multivariate modular engineering of the protein secretory pathway for production of heterologous glucose oxidase in *Pichia pastoris*. *Enzyme Microb. Technol.* 68, 33–42. doi: 10.1016/j.enzmictec.2014.10.006

He, J., Tang, F., Chen, D., Yu, B., Luo, Y., Zheng, P., et al. (2019). Design, expression and functional characterization of a thermostable xylanase from *Trichoderma reesei*. *PLoS One* 14:e0210548. doi: 10.1371/journal.pone.0210548

Inan, M., Aryasomayajula, D., Sinha, J., and Meagher, M. M. (2006). Enhancement of protein secretion in *Pichia pastoris* by overexpression of protein disulfide isomerase. *Biotechnol. Bioeng.* 93, 771–778. doi: 10.1002/bit.20762

Janowski, R., Kozak, M., Jankowska, E., Grzonka, Z., Grubb, A., Abrahamson, M., et al. (2001). Human cystatin C, an amyloidogenic protein, dimerizes through three-dimensional domain swapping. *Nat. Struct. Biol.* 8, 316–320. doi: 10.1038/86188

Jolly, C., and Morimoto, R. I. (2000). Role of the heat shock response and molecular chaperones in oncogenesis and cell death. *J. Natl. Cancer Inst.* 92, 1564–1572. doi: 10.1093/jnci/92.19.1564

Liang, S., Wang, B., Pan, L., Ye, Y., He, M., Han, S., et al. (2012). Comprehensive structural annotation of *Pichia pastoris* transcriptome and the response to various carbon sources using deep paired-end RNA sequencing. *BMC Genom.* 13:738. doi: 10.1186/1471-2164-13-738

Liao, X., Zhao, J., Liang, S., Jin, J., Li, C., Xiao, R., et al. (2019). Enhancing co-translational folding of heterologous protein by deleting non-essential ribosomal proteins in *Pichia pastoris*. *Biotechnol. Biofuels* 12:38. doi: 10.1186/s13068-019-1377-z

Liu, W., Jiang, Y., Peng, L., Sun, X., Gan, W., Zhao, Q., et al. (2022). Inferring gene regulatory networks using the improved markov blanket discovery algorithm. *Interdiscip. Sci.* 14, 168–181. doi: 10.1007/s12539-021-00478-9

Mathews, P. M., and Levy, E. (2016). Cystatin C in aging and in alzheimer's disease. *Ageing Res. Rev.* 32, 38–50. doi: 10.1016/j.arr.2016.06.003

Merz, G. S., Benedikz, E., Schwenk, V., Johansen, T. E., Vogel, L. K., Rushbrook, J. I., et al. (1997). Human cystatin C forms an inactive dimer during intracellular trafficking in transfected CHO cells. *J. Cell. Physiol.* 173, 423–432. doi: 10.1002/(SICI)1097-4652(199712)173:3<423::AID-JCP15>3.0.CO;2-C

Navone, L., Vogl, T., Luangthongkam, P., Blinco, J. A., Luna-Flores, C. H., Chen, X., et al. (2021a). Disulfide bond engineering of AppA phytase for increased thermostability requires co-expression of protein disulfide isomerase in *Pichia pastoris*. *Biotechnol. Biofuels* 14:80. doi: 10.1186/s13068-021-01936-8

Navone, L., Vogl, T., Luangthongkam, P., Blinco, J. A., Luna-Flores, C., Chen, X., et al. (2021b). Synergistic optimisation of expression, folding, and secretion improves E. coli AppA phytase production in *Pichia pastoris*. *Microb. Cell Fact.* 20:8. doi: 10.1186/s12934-020-01499-7

Nusse, O., and Lindau, M. (1988). The dynamics of exocytosis in human neutrophils. *J. Cell Biol.* 107, 2117–2123. doi: 10.1083/jcb.107.6.2117

Oakes, S. A., and Papa, F. R. (2015). The role of endoplasmic reticulum stress in human pathology. *Annu. Rev. Pathol.* 10, 173–194. doi: 10.1146/annurev-pathol-012513-104649

Orman, M. A., Calik, P., and Ozdamar, T. H. (2009). The influence of carbon sources on recombinant-human- growth-hormone production by *Pichia pastoris* is dependent on phenotype: A comparison of muts and mut+ strains. *Biotechnol. Appl. Biochem.* 52, 245–255. doi: 10.1042/BA20080057

Palsdottir, A., Snorradottir, A. O., and Thorsteinsson, L. (2006). Hereditary cystatin C amyloid angiopathy: Genetic, clinical, and pathological aspects. *Brain Pathol.* 16, 55–59. doi: 10.1111/j.1750-3639.2006.tb00561.x

Renuse, S., Madugundu, A. K., Kumar, P., Nair, B. G., Gowda, H., Prasad, T. S., et al. (2014). Proteomic analysis and genome annotation of *Pichia pastoris*, a recombinant protein expression host. *Proteomics* 14, 2769–2779. doi: 10.1002/pmic.201400267

Robasky, K., Lewis, N. E., and Church, G. M. (2014). The role of replicates for error mitigation in next-generation sequencing. *Nat. Rev. Genet.* 15, 56–62. doi: 10.1038/nrg3655

Schulze, S. K., Kanwar, R., Golzenleuchter, M., Therneau, T. M., and Beutler, A. S. (2012). SERE: Single-parameter quality control and sample comparison for RNA-Seq. *BMC Genom.* 13:524. doi: 10.1186/1471-2164-13-524

Shu, J., Wei, W., and Zhang, L. (2022). Identification of molecular signatures and candidate drugs in vascular dementia by bioinformatics analyses. *Front. Mol. Neurosci.* 15:751044. doi: 10.3389/fnmol.2022.751044

Sola, A., Jouhten, P., Maaheimo, H., Sanchez-Ferrando, F., Szyperski, T., and Ferrer, P. (2007). Metabolic flux profiling of *Pichia pastoris* grown on glycerol/methanol mixtures in chemostat cultures at low and high dilution rates. *Microbiology* 153, 281–290. doi: 10.1099/mic.0.29263-0

Vici, A. C., da Cruz, A. F., Facchini, F. D., de Carvalho, C. C., Pereira, M. G., Fonseca-Maldonado, R., et al. (2015). *Beauveria bassiana* lipase A expressed in *Komagataella* (*Pichia*) pastoris with potential for biodiesel catalysis. *Front. Microbiol.* 6:1083. doi: 10.3389/fmicb.2015.01083

Vogl, T., and Glieder, A. (2013). Regulation of *Pichia pastoris* promoters and its consequences for protein production. *N. Biotechnol.* 30, 385–404. doi: 10.1016/j.nbt.2012.11.010

Vogl, T., Sturmberger, L., Fauland, P. C., Hyden, P., Fischer, J. E., Schmid, C., et al. (2018). Methanol independent induction in *Pichia pastoris* by simple derepressed overexpression of single transcription factors. *Biotechnol. Bioeng.* 115, 1037–1050. doi: 10.1002/bit.26529

Wang, J., Wang, X., Shi, L., Qi, F., Zhang, P., Zhang, Y., et al. (2017). Methanol-independent protein expression by AOX1 promoter with trans-acting elements engineering and glucose-glycerol-shift induction in *Pichia pastoris*. *Sci. Rep.* 7:41850. doi: 10.1038/srep41850

Whyteside, G., Alcocer, M. J., Kumita, J. R., Dobson, C. M., Lazarou, M., Pleass, R. J., et al. (2011). Native-state stability determines the extent of degradation relative to secretion of protein variants from *Pichia pastoris*. *PLoS One* 6:e22692. doi: 10.1371/journal.pone.0022692

Yan, P., Zou, Z., Zhang, S., Wang, R., Niu, T., Zhang, X., et al. (2021). Defining the mechanism of PDI interaction with disulfide-free amyloidogenic proteins:

Implications for exogenous protein expression and neurodegenerative disease. *Int. J. Biol. Macromol.* 174, 175–184. doi: 10.1016/j.ijbiomac.2021.01.172

Yu, X. W., Sun, W. H., Wang, Y. Z., and Xu, Y. (2017). Identification of novel factors enhancing recombinant protein production in multi-copy *Komagataella* phaffii based on transcriptomic analysis of overexpression effects. *Sci. Rep.* 7:16249. doi: 10.1038/s41598-017-16577-x

Yu, Y., Wang, Y., He, J., Liu, Y., Li, H., Zhang, H., et al. (2010). Structural and dynamic properties of a new amyloidogenic chicken cystatin mutant I108T. *J. Biomol. Struct. Dyn.* 27, 641–649. doi: 10.1080/07391102.2010.10508578

Zhao, Y., Li, M. C., Konate, M. M., Chen, L., Das, B., Karlovich, C., et al. (2021). TPM, FPKM, or normalized counts? a comparative study of quantification measures for the analysis of RNA-seq data from the NCI patient-derived models repository. *J. Transl. Med.* 19:269. doi: 10.1186/s12967-021-02936-w

Zhou, X., Lu, X., Qin, S., Xu, L., Chong, X., Liu, J., et al. (2019). Is the absence of alpha-helix 2 in the appendant structure region the major contributor to structural instability of human cystatin C? *J. Biomol. Struct. Dyn.* 37, 4522–4527. doi: 10.1080/07391102.2018.1552625

Zhou, Y., Zhou, Y., Li, J., Chen, J., Yao, Y., Yu, L., et al. (2015). Efficient expression, purification and characterization of native human cystatin C in *Escherichia coli* periplasm. *Protein Expr. Purif.* 111, 18–22. doi: 10.1016/j.pep.2015.03.006

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 12 January 2023
DOI 10.3389/fmicb.2022.1090770

OPEN ACCESS

EDITED BY
Liang Wang,
Guangdong Provincial People's Hospital, China

REVIEWED BY
Nguyen Quoc Khanh Le,
Taipei Medical University, Taiwan
Yuan Zhu,
China University of Geosciences Wuhan, China
Yibao Zhang,
Peking University, China
Sai Ho Ling,
University of Technology Sydney, Australia

*CORRESPONDENCE
Xiance Jin
✉ jinxc1979@hotmail.com
Congying Xie
✉ wzxiecongying@163.com

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

RECEIVED 06 November 2022
ACCEPTED 28 December 2022
PUBLISHED 12 January 2023

CITATION
Wei C, Xiang X, Zhou X, Ren S, Zhou Q, Dong W, Lin H, Wang S, Zhang Y, Lin H, He Q, Lu Y, Jiang X, Shuai J, Jin X and Xie C (2023) Development and validation of an interpretable radiomic nomogram for severe radiation proctitis prediction in postoperative cervical cancer patients. *Front. Microbiol.* 13:1090770. doi: 10.3389/fmicb.2022.1090770

Development and validation of an interpretable radiomic nomogram for severe radiation proctitis prediction in postoperative cervical cancer patients

Chaoyi Wei[1†], Xinli Xiang[2†], Xiaobo Zhou[3], Siyan Ren[3], Qingyu Zhou[3], Wenjun Dong[3], Haizhen Lin[3], Saijun Wang[3], Yuyue Zhang[3], Hai Lin[1], Qingzu He[1], Yuer Lu[1], Xiaoming Jiang[1], Jianwei Shuai[1], Xiance Jin[4,5]* and Congying Xie[3]*

[1]Wenzhou Key Laboratory of Biophysics, Wenzhou Institute, University of Chinese Academy of Sciences, Wenzhou, Zhejiang Province, China, [2]The Second Affiliated Hospital of Wenzhou Medical University, Wenzhou, Zhejiang Province, China, [3]Medical and Radiation Oncology, The Second Affiliated Hospital of Wenzhou Medical University, Wenzhou, Zhejiang Province, China, [4]Radiotherapy Center, The First Affiliated Hospital of Wenzhou Medical University, Wenzhou, Zhejiang Province, China, [5]School of Basic Medical Sciences, Wenzhou Medical University, Wenzhou, Zhejiang Province, China

Background: Radiation proctitis is a common complication after radiotherapy for cervical cancer. Unlike simple radiation damage to other organs, radiation proctitis is a complex disease closely related to the microbiota. However, analysis of the gut microbiota is time-consuming and expensive. This study aims to mine rectal information using radiomics and incorporate it into a nomogram model for cheap and fast prediction of severe radiation proctitis prediction in postoperative cervical cancer patients.

Methods: The severity of the patient's radiation proctitis was graded according to the RTOG/EORTC criteria. The toxicity grade of radiation proctitis over or equal to grade 2 was set as the model's target. A total of 178 patients with cervical cancer were divided into a training set (n=124) and a validation set (n=54). Multivariate logistic regression was used to build the radiomic and non-raidomic models.

Results: The radiomics model [AUC=0.6855(0.5174-0.8535)] showed better performance and more net benefit in the validation set than the non-radiomic model [AUC=0.6641(0.4904-0.8378)]. In particular, we applied SHapley Additive exPlanation (SHAP) method for the first time to a radiomics-based logistic regression model to further interpret the radiomic features from case-based and feature-based perspectives. The integrated radiomic model enables the first accurate quantitative assessment of the probability of radiation proctitis in postoperative cervical cancer patients, addressing the limitations of the current qualitative assessment of the plan through dose-volume parameters only.

Conclusion: We successfully developed and validated an integrated radiomic model containing rectal information. SHAP analysis of the model suggests that radiomic features have a supporting role in the quantitative assessment of the probability of radiation proctitis in postoperative cervical cancer patients.

KEYWORDS

radiomics, nomogram, radiation proctitis, SHapley Additive exPlanation (SHAP), microbiota

1. Introduction

Cervical cancer is a malignant neoplasm at the junction of the squamous epithelial cells of the vaginal or transitional zone of the cervix and the endocervical canal columnar epithelial cells. Cervical cancer is the fourth most common cancer worldwide (Sung et al., 2021). Radiotherapy is one of the most effective methods for treating pelvic malignancies and has an irreplaceable role in treating cervical cancer at all stages. The main complication of radiotherapy for pelvic malignancies is radiation proctitis (Yeung et al., 2020). Fifty percent of patients with cervical cancer or endometrial cancer who received postoperative intensity-modulated radiotherapy developed acute rectal toxicity, and 5%–10% developed chronic rectal toxicity (Zelefsky et al., 2008; Yeung et al., 2020).

Unlike simple radiation damage to other organs, radiation proctitis is a complex disease closely related to the microbiota. A study using a rectal radiation mouse model showed that radiation affected both host and intestinal microbiota (Gerassy-Vainberg et al., 2018). Radiation therapy could induce local microbial ecological dysbiosis, and the dysbiosis microbiota could exert a direct pro-inflammatory effect on epithelial cells. In another study of 32 female patients with chronic radiation proctitis, differential patterns of dysbiosis were observed after analyzing the gut microbiota of patients with or without hematochezia (Liu et al., 2021). Gut microbiota could offer a set of biomarkers for radiation enteritis prediction, disease activity evaluation, and treatment selection (Wang et al., 2019).

However, the current prediction models of radiation proctitis were focused mainly on clinical and radiotherapy dose features. Several univariate and multivariate analyses showed that features, including tumor size, pathological characteristics, and radiological parameters, were significantly correlated with post-radiotherapy comorbidities in patients undergoing pelvic radiotherapy (Albert et al., 2008; Schmidt et al., 2022). A study by Fiorino et al. showed that rectal function was significantly correlated with treatment volume, PTV margins, radiation therapy dose, hemorrhoids presence, anticoagulant use during follow-up, and relative (%) and absolute (cm^3) values of rectal V38Gy and V40Gy correlated with rectal bleeding (Fiorino et al., 2008). A study by Mahal et al. also noted that total radiation dose, dose per fraction, radiotherapy techniques, and treatment volume affected the rectum of patients (Mahal et al., 2014).

Another review also suggested features associated with radiation proctitis, including vascular diseases such as smoking, diabetes, hypertensive diabetes and atherosclerosis, collagen vascular disease, comorbid inflammatory bowel disease, and human immunodeficiency virus infection. Also, the review noted that specific underlying genetic changes could affect patients' sensitivity to radiation. There was a correlation between genes and higher risks of gastrointestinal and genitourinary tract radiotoxicity after radiotherapy (Shadad et al., 2013).

The above studies have shown a strong correlation between patients' oncologic features, pathologic features, and radiologic dose and the appearance of radiation proctitis in postoperative radiotherapy for pelvic cancer. However, the conclusions of these studies are inconsistent, and the accuracy of the prediction of radiation proctitis is unsatisfactory.

In recent years, computer-aided diagnosis, especially machine learning methods, has also been used for postoperative radiotherapy side effects and comorbidities prediction in oncology patients. Lee et al. applied machine learning methods such as random forest and bioinformatics to genome-wide data to predict and interpret advanced urogenital toxicity (Lee et al., 2018). They designed more robust predictive models and identified plausible biomarkers and biological processes associated with late urogenital toxicity. A study by Lewis & Kemp et al. showed that the integration of machine learning and genome-scale metabolic modeling identified multi-group biomarkers of radiation resistance (Lewis and Kemp, 2021).

However, it should be noted that the machine learning models above were developed using clinical features only. It ignored the large number of features embedded in computed tomography images (CT) that are imperceptible to the naked eye. Moreover, in the process of treatment plan determination, physicians are more focused on extracting focal area information and lack awareness of pelvic rectal information. Therefore, a comprehensive model is urgently needed to deepen the understanding of patient rectal image information to accurately assess radiotherapy treatment plans and reduce severe complications of radiation proctitis.

Abbreviations: SHAP, SHapley Additive exPlanation; CT, computed tomography; ROI, Regions of Interest; LASSO, the least absolute shrinkage and selection operator regularization; DCA, decision curve analysis; AUC, area under curve; OR, odds ratio; ROC, receiver operating characteristic analysis; ECCR, Ethics Committee in Clinical Research; RTOG, Radiation Therapy Oncology Group; EORTC, European Organization for Research and Treatment of Cancer.

Deep learning, as well as dynamical modeling, is demonstrating powerful feature extraction and modeling capabilities in various medical fields (Li et al., 2021; Qian et al., 2021; Chen et al., 2022; Hu et al., 2022; Li Y. et al., 2022; Li X. et al., 2022). In data-driven disease research, a graph neuro network was used to predict the potential associations of disease-related metabolites (Sun et al., 2022). Deep learning can also be used to explore the identification of circRNA-disease associations (Wang et al., 2021) and predict the potential human lncRNA interactions (Zhang et al., 2021; Jingxuan et al., 2022; Wang et al., 2022). In drug metabolism research, deep learning can be used to predict the ability of a compound to permeate across the blood–brain barrier (Tang et al., 2022) and drug response (Kuenzi et al., 2020). At the same time, deep learning is also a popular tool for radiotherapy research. Zhong et al. developed a deep learning-based radiomic nomogram that could predict the prognosis of patients with different treatment regimens (Zhong et al., 2021). Qiang et al. established a prognosis prediction system based on deep learning for locoregionally advanced nasopharyngeal carcinoma (Qiang et al., 2021). Although deep learning has been widely applied in the analysis and prediction of various diseases, the poor interpretability of the deep learning model makes it difficult for clinicians to understand and trust these tools (Huff et al., 2021).

Radiomics can extract biomedical images containing information reflecting the underlying pathophysiology and reveal the relationships through quantitative image analysis (Le et al., 2021; Lam et al., 2022). In previous studies, radiomics has been used to predict postoperative radiotherapy-induced toxicity in prostate cancer patients. Mostafaei et al. showed that models based on CT radiomics, clinical features, and dose-volume parameters could predict radiation toxicity. The combination of imaging and clinical features could improve the performance of radiotoxicity prediction models (Mostafaei et al., 2020). However, no study has been performed to predict postoperative radiotherapy comorbidity in cervical cancer patients using radiomic features. Due to the limited resolution, the information on microbiota is almost impossible to extract directly by radiomics in theory, and no relevant studies have been reported. However, it is feasible that radiomics can indirectly reflect the effect of microbiota on the rectum.

Therefore, this study uses radiomics to extract the rectal information from medical images and improve the model performance and diagnostic accuracy through quantitative image analysis. Moreover, this study creatively introduces SHapley Additive exPlanation (SHAP) values to explore the interpretability of nomogram to improve the clinicians' understanding of the model and its radiomic features, which facilitates later clinical promotion.

2. Materials and methods

2.1. Patients

The study protocol was reviewed and approved by the Ethics Committee in Clinical Research (ECCR) of the First Affiliated Hospital of Wenzhou Medical University. It was conducted following the Declaration of Helsinki. The Transparent Reporting of Individual Prognosis or Diagnosis Multivariate Predictive Models (TRIPOD) guidelines and the Strengthening Reports of Observational Studies in Epidemiology (STROBE) statement were applied. As this study was a retrospective cohort study, informed consent was waived, and all patient data were anonymized and desensitized.

Patients with cervical cancer from 1st January 2015 to 31st December 2020 in the First Affiliated Hospital of Wenzhou Medical University were collected for this study. These patients received a cervical cancer diagnosis, oncological surgery, and postoperative radiotherapy at the First Affiliated Hospital of Wenzhou Medical University.

The inclusion criteria (Figure 1) include (a) no severe symptoms at the time of diagnosis and good general physical condition; (b) patients with relatively intact organ functions, basically standard hematological system, and no contraindications to treatment; (c) no previous history of other malignant tumor diseases and radiotherapy; (d) postoperative pathological examination results confirming the diagnosis of cervical cancer; (e) postoperative radiotherapy; (f) patients with postoperative treatment; (g) the patient had complete pathology, imaging, and radiation therapy dose information; (h) no intracavitary brachytherapy was performed.

The exclusion criteria include (a) no definite postoperative pathological findings; (b) no complete clinicopathological data; (c) no CT scan was performed before postoperative radiotherapy; (d) patient's pathology, imaging, and radiation treatment dose data are missing; (e) intolerance to radiotherapy or chemotherapy, and treatment plan was not completed due to severe acute toxic reactions during treatment; (f) Have received intracavitary brachytherapy.

The Mann–Whitney *U*-test and the Chi-square test were used to evaluate the performance of clinical and dose-volume features. Patients were randomly divided into a training set and a validation set.

2.2. Extraction of radiomic features

The entire rectal region on the patient's CT image was defined as the Regions of Interest (ROI). Using ITK-SNAP software (Yushkevich et al., 2006), a pelvic radiologist with 10 years of experience at the First Affiliated Hospital of Wenzhou Medical University outlined this target region manually. Another radiologist with 20 years of experience reviewed it. Extraneous components other than the rectum, such as peripheral vessels, peripheral tissues, and peripheral organs, were not outlined by radiologists to minimize interfering information. The two radiologists did not know the patient's information. If the two doctors had the same opinion, the ROI would be included in the imaging data set.

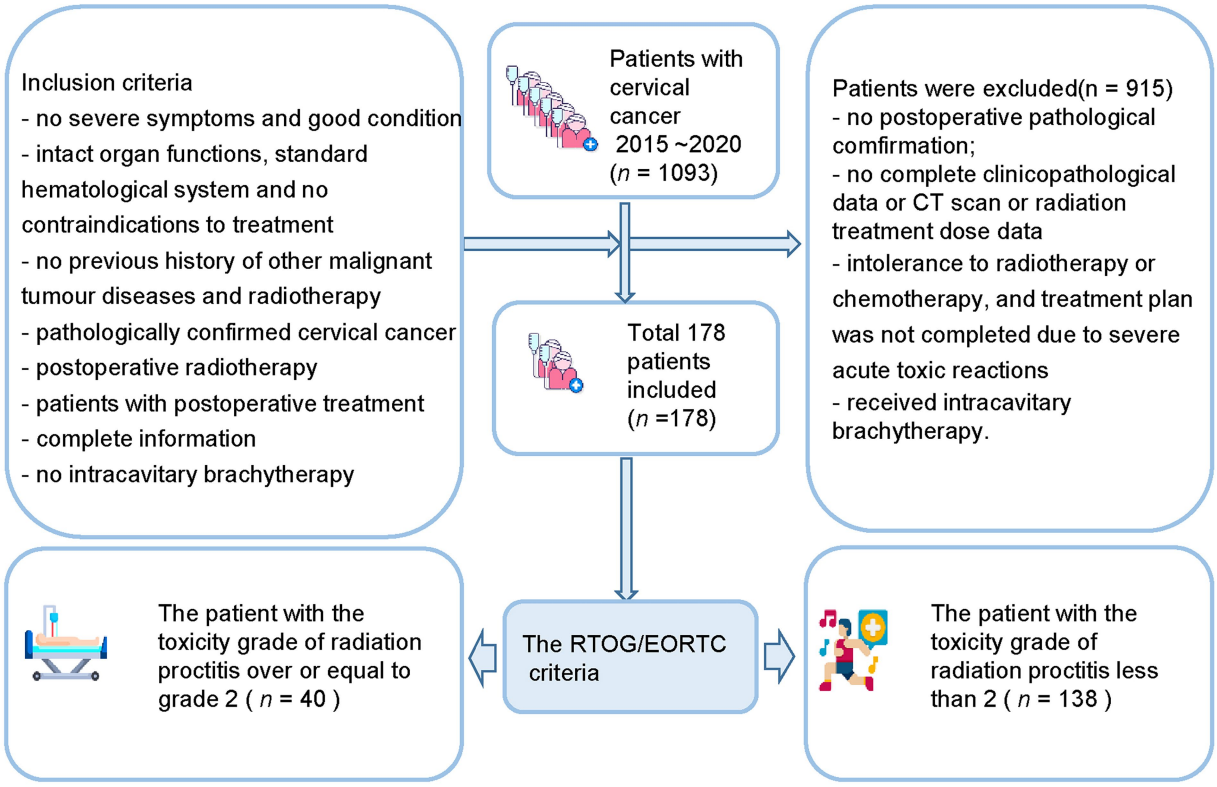

FIGURE 1
Flow diagram of the study enrolment patients.

Quantitative radiological features were automatically extracted using a feature extraction platform based on the Python package PyRadiomics (van Griethuysen et al., 2017). After segmentation and reconstruction of the patient CT, each patient extracted ROI was imported into Python in nrrd format. We extracted 1,409 radiomic features, including 8 feature classes used for further analysis and regression modeling. Radiomics features were dependent on the CT hardware, scanning parameters, and contrast agents. The process of generation and selection of radiomic features was illustrated in Figure 2.

2.3. Feature selection and model development

The variance equality of radiomics features was assessed by Levene's test. Independent t-test or Wilcoxon's test was used for feature selection. After standardizing the radiomic features using the z fraction transform, the high-dimensional imaging features extracted from the ROI were selected by the least absolute shrinkage and selection operator (LASSO) regularization algorithm. We performed univariate logistic regression on all features to screen out the key features significantly associated with the severity of radiation proctitis. The value of p was usually set at $p < 0.2$, but can also be set at $p < 0.05$ or $p < 0.1$. It requires the researcher to adjust the value of p according to the sample size. Due to the limited amount of data in this study, we set $p < 0.2$ as the threshold. Features with $p > 0.2$ in the univariate logistic regression were excluded, and features with $p < 0.2$ were included. Finally, the features left after multiple screenings were introduced into a stepwise logistic regression analysis to build a comprehensive model.

2.4. Model simplification and model evaluation

The critical features remained after multiple screenings were included in the multivariate logistic regression model generated by the stepwise forward and backward methods. Finally, to transform the complex regression equations into simple and visual graphs and make the prediction models' results more readable, a visual nomogram was constructed based on these features that can be stably present in the unified model. All model evaluations were performed on the unseen validation set. In addition, calibration curves were used to evaluate the model performance of the nomogram.

To further validate the performance of the radiomic features, we built a simplified non-radiomic model by removing the radiomic features. To evaluate the performance of these two

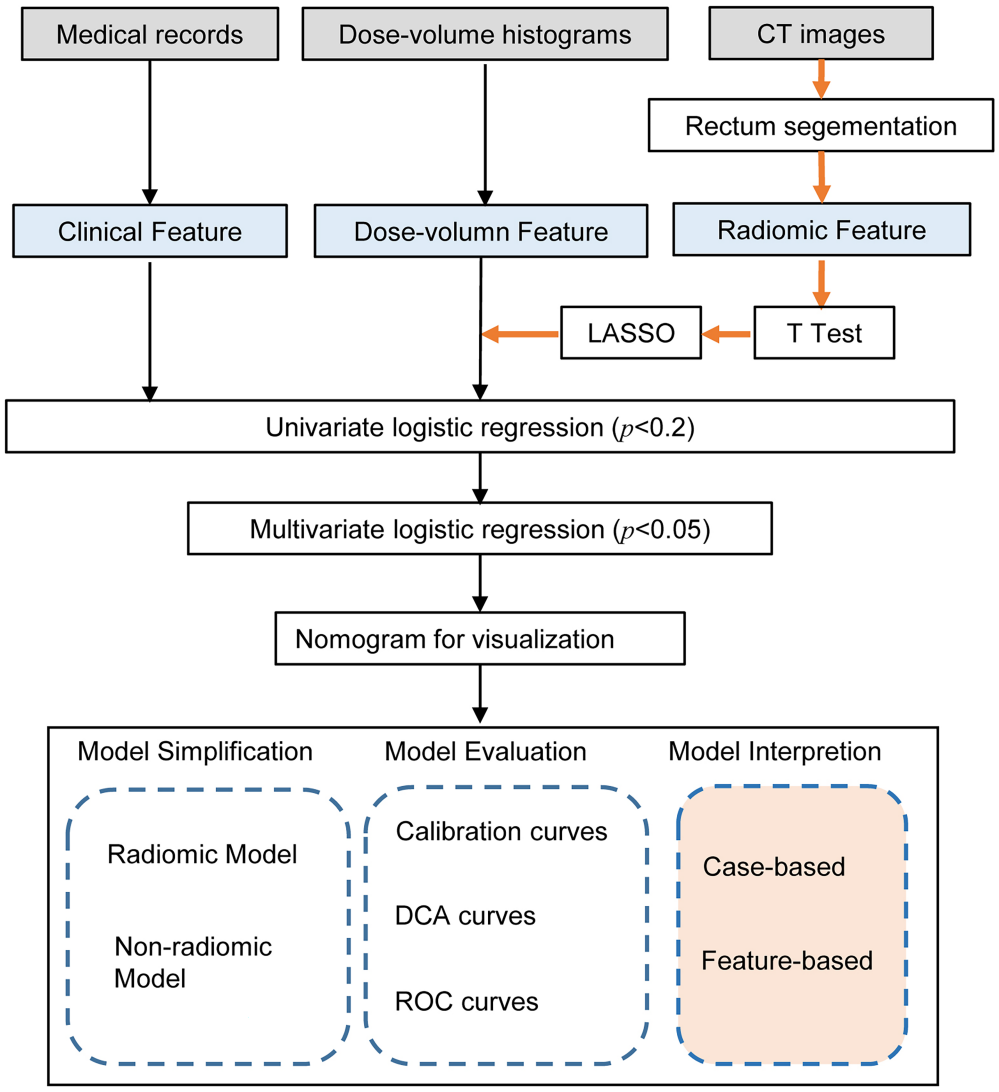

FIGURE 2
Workflow of the radiomic model development and model analysis process. The orange arrows in the flow chart represented the processing of the radiomic features. The radiomic features were generated by the PyRadiomics package after outlining the rectal region on the original image. After feature selection by *T*-test, LASSO and univariate logistic regression, multivariate logistic regression models were developed and visualized as nomograms. The model analysis consists of three parts: model simplification, model evaluation, and model interpretation. By comparing the performance change before and after model simplification, we could measure the importance of the radiomic features. In particular, we applied SHapley Additive exPlanation (SHAP) values for the first time to a radiomics-based logistic regression model to further interpret the radiomic features from case-based and feature-based perspectives.

models, we assessed the discrimination using the receiver operating characteristic (ROC) analysis. The area under the receiver operating characteristic curve (AUC) was used to assess the predictive discrimination of these two models. In addition, in order to verify the validity of the model from another perspective, a k-Nearest Neighbor (KNN) model was built using the same data as the radiomic nomogram. The root-mean-square error (RMSE) and 10-fold cross-validation were used to select the optimal hyperparameter of the KNN model. We used decision curve analysis (DCA) to assess clinical validity by quantifying the net benefit at each threshold probability. All statistical analyses were performed using R (version 4.2.2), Python (version 3.9.12), and SPSS (version 24.0, IBM). The workflow of the model analysis process after modeling was shown in Figure 2.

2.5. Model interpretation

SHapley Additive exPlanation (SHAP) method is a game-theoretic-based model interpretation method. From a game theory perspective, SHAP treats each feature variable as a player. The predicted outcome obtained by the model is considered as the gain from the cooperation of many players to complete a project. It connects optimal credit allocation with local explanations using

the classical Shapley values from game theory and their related extensions (Lundberg and Lee, 2017). We used scikit-learn (Pedregosa et al., 2011) to build the logistic regression model and used the SHAP package to calculate the SHAP values for the logistic regression model and further analyze the SHAP values with the SHAP plot module. The decision process of each patient could be presented by force plot. By overlaying the force plots and sorting the output values, we could see how all patients made their decisions. In addition to analyzing the model from the patient's perspective, we can also use SHAP to understand the model from the feature's perspective. SHAP provides bar plots and scatter plots of features to help us understand which feature was most important to the model.

3. Results

3.1. Baseline information of patients

This study included 1,093 patients with cervical cancer who needed to initiate radiotherapy at the First Affiliated Hospital of Wenzhou Medical University between 1st January 2015 and 31st December 2020. After screening and exclusion, a total of 178 patients were finally included in our study. The study included 40 patients (22.5%) with a toxicity grade greater than or equal to grade 2 after radiation therapy and 138 patients (77.5%) less than grade 2 after radiation therapy. The patients were divided into a training set ($n = 124$) and a validation set ($n = 54$). Table 1 shows the baseline information of the patients.

3.2. Radiomic features selection and multivariate analysis

We extracted a total of 1,409 radiomic features from the patients' CTs and selected them using the LASSO algorithm. Multivariate logistic regression analysis was performed on all features selected by LASSO and univariate logistic regression. The results of the multivariate logistic regression are shown in Table 2. Radiotherapy techniques [OR = 0.000 (0.000–0.086), $p = 0.005$], Maximum rectal dose [OR = 1.006 (1.001–1.011), $p = 0.020$], Contrast [OR = 0.000 (0.000–0.002), $p = 0.046$] were independent risk factors for severe radiation proctitis.

3.3. Establishment of nomogram and model evaluation

In order to develop a clinically applicable method to predict the occurrence of radiation proctitis, we constructed a radiomics nomogram. The results of the nomogram were shown in Figure 3A. All model evaluations were performed on the unseen validation set. The calibration curve of the combined radiomics nomogram was shown in Figure 3B. To further validate the performance of the radiomic features, we built a simplified non-radiomic model based only on the clinical feature and dose-volume feature by removing the radiomic feature and comparing its performance with the full radiomic model. The ROC curves for the two nomogram models (Figure 3C) showed that the prediction effect of the radiomic model [AUC = 0.6855 (0.5174–0.8535)] performed better than the non-radiomic model [AUC = 0.6641 (0.4904–0.8378)]. The AUC of radiomic nomogram [AUC = 0.6855 (0.5174–0.8535)] was close to that of the KNN model [AUC = 0.7051 (0.5602–0.85)]. It illustrated the validity of the model from another perspective. The decision curve analysis (DCA; Figure 3D) was used to assess the utility of both prediction models by calculating the net benefit at various probability thresholds. According to the decision curves, the radiomic model showed more benefit in predicting the risk of radiation proctitis than the non-radiomic model. It suggested that radiomic features were supporting features for severe radiation proctitis prediction.

3.4. Model interpretation

3.4.1. Case-based model interpretation

To further understand how decision-making occurred for individual and entire patient populations, we used SHAP to analyze from a case-based perspective. Figure 4A represented the decision process for SHAP values across all patients, with the vertical axis representing the magnitude of the SHAP values. As the graph was ordered by model output, we could clearly see the boundary line between red and blue. Features pushing the prediction higher were shown in red, and those pushing the prediction lower were in blue.

In addition to the model interpretation for all cases, we could also provide a clearer picture of the decision-making situation for individual patients through the waterfall or force plot. For example, by selecting the patient on the far right of Figure 4A, the decision-making process could be visualized in Figure 4B or Figure 4C. Although the presentation was different, the information in Figures 4B,C was consistent. These two plots indicated the proportion and absolute SHAP value of various features in the decision-making process for that patient. SHAP could provide a quantitative and visual representation of the decision mechanisms of the radiomics model for any patient.

3.4.2. Feature-based model interpretation

We calculated and visualized the SHAP values for each feature in the radiomics model. The beeswarm plot (Figure 5A) demonstrated an overview of the feature contribution of all patients. In the beeswarm plot, features were sorted by the sum of SHAP value magnitudes over all samples, and SHAP values were used to show the distribution of each feature's impacts. The bar plot shown in Figure 5B demonstrated the mean absolute value of the SHAP values for each feature. The plot showed that radiotherapy techniques and the maximum rectal dose have a high mean value. Since SHAP values represented a feature's responsibility for a

TABLE 1 Baseline information of all patients.

Variables		Primary queue (n=178)		
		<grade 2 (n=138)	≥grade 2 (n=40)	p-value
Age (years)		53.5 (46–61)	52(47.75–60.75)	0.957
Therapy	3D-CRT	53 (63.9%)	30 (36.1%)	<0.001
	VMAT	85 (89.5%)	10 (10.5%)	
Vascular invasion		48 (69.6%)	21 (30.4%)	0.043
FIGO Staging		2(1–2)	1(1–2.75)	0.800
Total rectal volume		65.937 (51.421–94.235)	69.422 (51.921–89.546)	0.875
Minimum rectal dose		1320.85 (417.6–2205.55)	583.35 (407.775–1817.025)	0.189
Maximum rectal dose		4907.35 (4145.175–5293.575)	4187.7 (4136.525–4819.25)	0.038
Average rectal dose		3958.55 (3738.35–4144.6)	3928.8 (3831.325–3996.2)	0.289
V5Gy(cm^3)		63.444 (47.286–89.492)	69.422 (51.921–89.546)	0.427
V5Gy(%)		100 (98.148–100)	100(99.555–100)	0.987
V10Gy(cm^3)		62.984 (47.127–90.466)	69.422 (53.625–92.057)	0.260
V10Gy(%)		99.95 (95.415–100)	99.245 (98.365–100)	0.976
V15Gy(cm^3)		62.511 (46.688–90.466)	69.422 (53.625–91.919)	0.283
V15Gy(%)		99.18 (93.693–100)	98.52 (97.268–100)	0.837
V20Gy(cm^3)		62.511 (46.234–89.878)	69.422 (53.625–90.757)	0.274
V20Gy(%)		97.755 (92.235–100)	97.855 (95.11–99.743)	0.705
V25Gy(cm^3)		61.865 (45.854–87.485)	68.366 (53.122–88.878)	0.240
V25Gy(%)		95.475 (89.273–99.065)	96.645 (92.355–98.093)	0.304
V30Gy(cm^3)		59.571 (42.243–80.618)	67.518 (52.789–87.533)	0.181
V30Gy(%)		90.145 (84.488–95.985)	95.06 (87.873–97.298)	0.028
V35Gy(cm^3)		56.05 (39.968–76.027)	61.987 (48.092–86.923)	0.203
V35Gy(%)		83.955 (75.073–91.188)	92.34 (84.098–95.993)	0.001
V40Gy(cm^3)		42.149 (27.754–60.107)	50.823 (33.909–65.466)	0.108
V40Gy(%)		59.66 (50.4–70.375)	65.595 (58.38–76.108)	0.035
V45Gy(cm^3)		16.175 (0–30.87)	0(0–15.492)	0.003
V45Gy(%)		26.465 (0–38.913)	0(0–16.76)	0.001

change in the model output, Figures 5A,B indicated that the radiotherapy technique and the maximum rectal dose were essential.

To understand how each feature affected the model's output, we plotted gray bar plots to show the SHAP values for each feature and scatter plots to show the SHAP values of the other features most relevant to that feature (Figures 5C,D).

TABLE 2 Result of multivariate logistic regression.

Features	B	P	OR (95% CI)
Therapy	−8.225	0.005	0.000(0.000–0.086)
Maximum rectal dose	0.006	0.020	1.006(1.001–1.011)
Contrast	−349.316	0.046	0.000(0.000–0.002)
Constant	−23.030	0.026	0.000

4. Discussion

With the advancement of radiotherapy techniques, the postoperative survival rates of cancers such as cervical cancer have increased dramatically (Citrin, 2017). However, complications and side effects caused by postoperative radiotherapy or chemotherapy are difficult to avoid. Radiation proctitis is one of the most common complications of postoperative radiotherapy in patients with pelvic tumors (Rustagi et al., 2015; Qian et al., 2021), with mild diarrhea or mild rectal exudate in mild cases and even intestinal necrosis or bleeding in severe cases, endangering patients' lives (Citrin, 2017). In clinical practice, doctors currently rely on the dose-volume features of radiotherapy plans to assess the risk of radiation proctitis. However, there is a lot of valuable information in pathology and clinical imaging that is not considered by

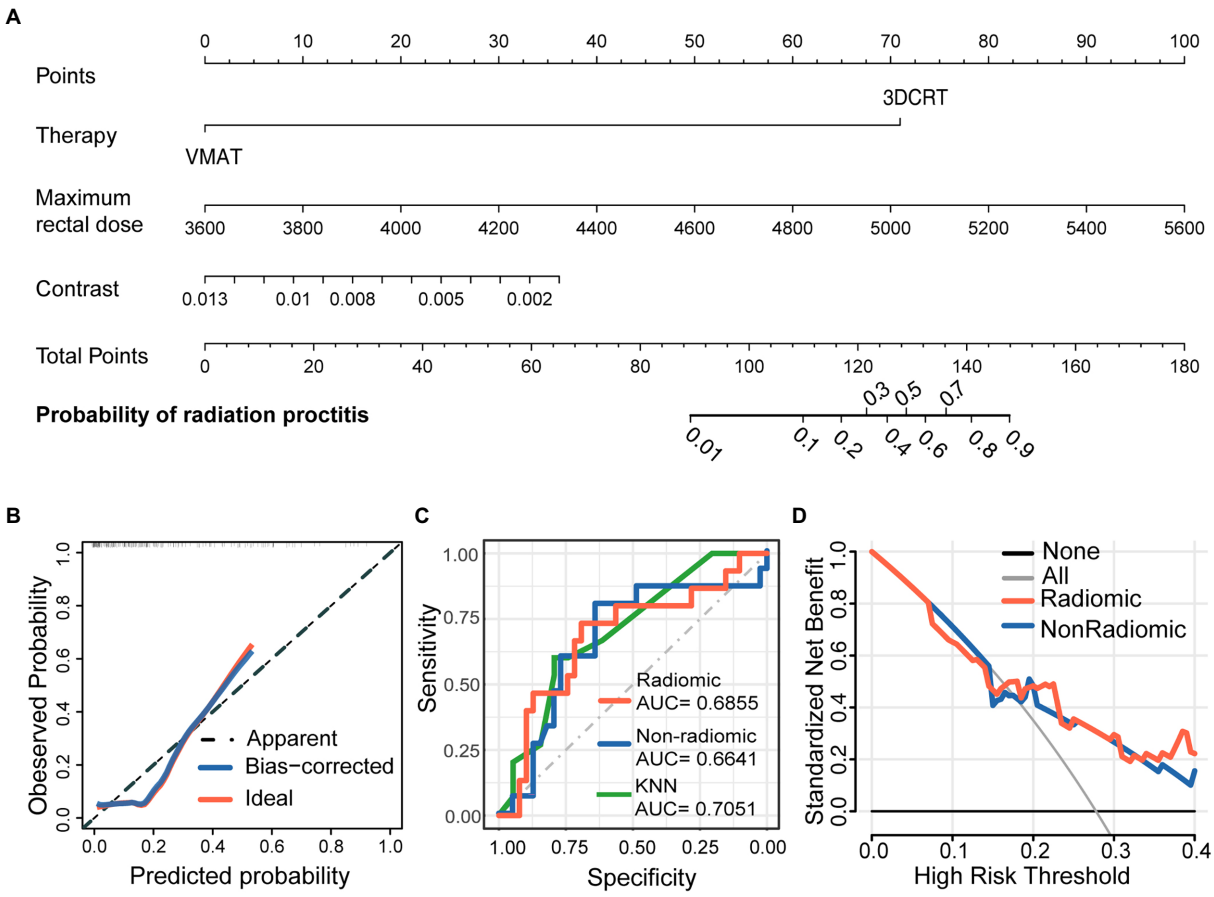

FIGURE 3
Nomogram for severe radiation proctitis prediction in postoperative cervical cancer patients, calibration of the nomogram, and decision curves in the overall patients. The combined nomogram **(A)** incorporated clinical, dose-volume, and radiomic features. By accumulating the points for each feature, we could quickly calculate the probability of radiation proctitis. All model evaluations were performed on the validation set. The Calibration curves of the combined radiomics nomogram **(B)** illustrated the relationship between the observed outcome frequencies and the predicted probabilities. The ROC curves **(C)** demonstrated the accuracy of the radiomic and non-radiomic models and KNN radiomic model. The DCA curves **(D)** demonstrated the net benefit of the radiomic and non-radiomic models.

clinicians. Moreover, the sensitivity of the rectum to radiotherapy radiation also varies significantly between individuals.

To further refine the assessment of radiation proctitis, we selected radiomic features associated with radiation proctitis by univariate regression and the LASSO algorithm. Radiotherapy techniques (OR = 0.000 (0.000–0.086), $p = 0.005$), Maximum rectal dose (OR = 1.006 (1.001–1.011), $p = 0.020$), Contrast (OR = 0.000 (0.000–0.002), $p = 0.046$) were independent risk factors for radiation proctitis. Finally, we developed an integrated prediction model based on clinical and radiomic features [AUC = 0.6855 (0.5174–0.8535)]. Current studies of radiation proctitis had mainly focused on local radiotherapy dose limits rather than comprehensive predictive models (Snyder et al., 2001; Huang et al., 2004). There was only one study using radiomics to build a predictive model for radiation proctitis (Mostafaei et al., 2020). In gastrointestinal toxicities modeling, the AUC of radiomic model of their study was 0.71, which was relatively higher compared with our study. However, the study was conducted based on data from only 64 patients and was only suitable for patients with prostate cancer.

The radiomic features of the model potentially incorporated the effect of microbiota on rectal radiosensitivity. The model without radiomic features showed lower validity, while the model containing both radiomic features and clinical features showed better performance on the ROC curve. The change of net benefit in Figure 3D suggested that radiomic features had played a supporting role in predictive models. And as a measure of the local intensity variation, a larger contrast correlated with a greater disparity in intensity values among neighboring voxels. In our study, the contrast suggested that a lower tissue density compared to the surrounding tissue was associated with higher radiosensitivity.

In most cases, PyRadiomics followed the image biomarker standardization initiative (IBSI)'s definition of features. PyRadiomics development was also involved in the standardization effort by the IBSI team. Still, there were some

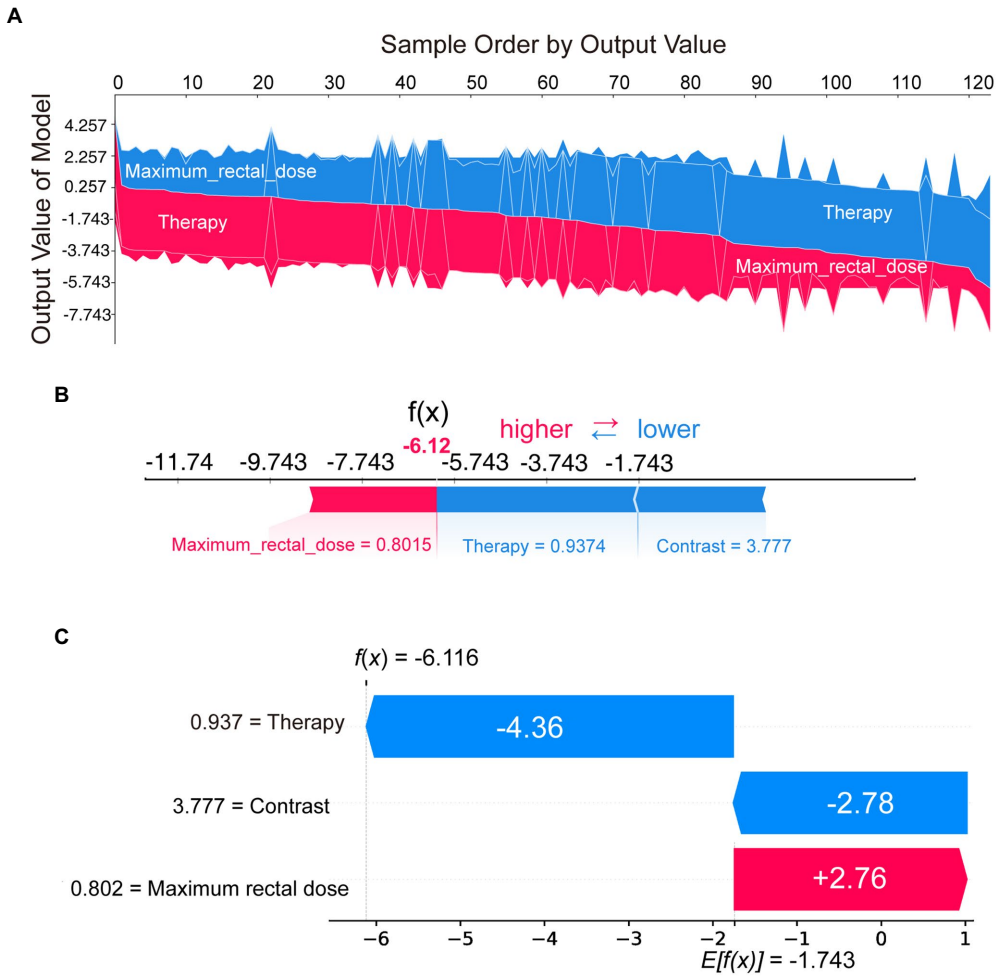

FIGURE 4
SHAP plots demonstrated SHAP values from a case-based perspective. Sampled by model output, the overall SHAP plot **(A)** showed the decision process of all patients. The force plot **(B)** and the waterfall plot **(C)** demonstrated the proportion and absolute SHAP value of various features in the decision-making process for a single patient.

differences between PyRadiomics and feature extraction as defined in the IBSI documents. Most notably were the differences in gray value discretization (just for the fixed bin size type) and resampling. In summary, the definitions of PyRadiomics and IBSI were slightly different, but did not represent one over the other. Moreover, IBSI was only an initiative, not a standard. For these reasons, IBSI would not significantly impact the reproducibility and validity of this study.

While SHAP was often used to explain features in machine learning algorithms and neural network models (Bang et al., 2021; Park et al., 2022; Shaji et al., 2022; Shi et al., 2022), SHAP analysis of logistic regression models had not yet been mentioned. Although logistic regression algorithms were simpler and more explicit than other machine learning algorithms and neural networks, logistic regression models were more challenging to understand than they may seem. Users could not directly measure the importance of features between continuous and categorical variables through odds ratio (OR) or coefficients (Table 2). In particular, for radiomic models, the significant variation in the magnitude of radiomic features made it more challenging to understand the actual decision-making process of the model through the coefficients and OR values of logistic regression. We wanted to help users better understand each feature's role in the model. In subsequent clinical treatment, model users can further quantify the contribution of radiomic features in each model output.

To address this issue, we introduced SHAP for the first time to a radiomics-based logistic regression model, which further revealed the model's decision-making mechanism (Figure 4A). The total contribution of SHAP for each feature included in the model was analyzed (Figure 5B). Radiotherapy techniques and the maximum rectal dose occupied vital positions in the model contribution. Notably, the SHAP value of the radiomic feature was the lowest. It suggested that the radiomic feature was

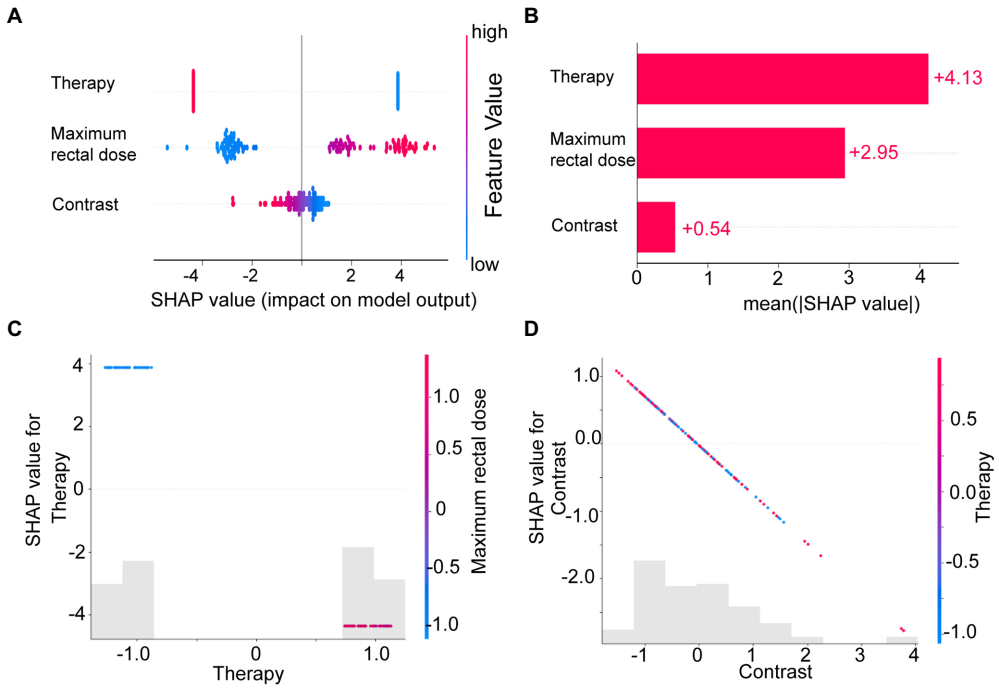

FIGURE 5
The SHAP plots illustrated the feature-based model interpretation process. **(A)** The beeswarm plot used SHAP values to show the distribution of each feature's impacts. **(B)** The standard bar plot demonstrated the mean absolute value of the SHAP values for each feature. These two plots **(C,D)** showed the SHAP values in different features. Gray bar plots showed the SHAP values for each feature. Scatter plots showed the SHAP values of the other features most relevant to that feature. Vertical dispersion represented interaction effects between the horizontal and vertical features.

weaker than the clinical feature and dose-volume feature. The SHAP could also analyze correlations between variables (Figures 5C,D). Correlations in SHAP values were observed between the three features. It may suggest an inter-collaborative relationship between variables in the model. However, this can only indicate a correlation between SHAP values, not between the values of the variables. In the subplot of therapy (Figure 5C), we can find that the most relevant variable was the maximum rectal dose. There was a harmful effect of maximum rectal dose in the decision-making process of these VMAT therapy samples. However, no fixed pattern was observed in the subplot of contrast (Figure 5D).

SHAP had a unique role in radiomics-based logistic models as a game-theoretic approach. SHAP helped us understand radiomic features that vary significantly in magnitude. Furthermore, SHAP provided a quantitative and visual representation of the decision mechanisms within the model for each patient.

We recommend that clinicians can reduce the value of the maximum rectal dose by modifying the plan when the model suggests that the current radiotherapy plan has a high probability of radiation proctitis. Clinicians can rely on interpretable models to precisely control the risk of the final plan to an acceptable level. Patients with cervical cancer can reduce unnecessary radiation doses and the incidence of radiation proctitis with the help of the comprehensive model.

5. Conclusion

We successfully developed and validated an integrated radiomic model containing rectal information in this study. The integrated radiomic model enables the accurate quantitative assessment of the probability of radiation proctitis in postoperative cervical cancer patients, addressing the limitations of the current qualitative assessment based on dose-volume parameters only. Based on the model output and SHAP values analysis, we suggest that clinicians can adjust the radiation dose to minimize the occurrence of severe radiation proctitis while not compromising the effectiveness of radiation therapy.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by the Ethics Committee in Clinical Research (ECCR) of the First Affiliated Hospital of Wenzhou Medical University.

Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

CW, XX, and CX conceived the project, developed the prediction method, designed and implemented the experiments, analyzed the result, and wrote the manuscript. XZ, SR, and XJin implemented the experiments and analyzed the result. QZ, WD, HaizL, SW, and YZ analyzed the result. HaiL, QH, YL, XJia, and JS contributed to the interpretation of the results. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by Zhejiang Engineering Research Center for Innovation and Application of Intelligent Radiotherapy Technology in the Second Affiliated Hospital of Wenzhou Medical University, Wenzhou key Laboratory of radiotherapy and Translational Research of Cancer (2021100848), and Wenzhou Science and Technology Bureau Y2020733. This work was also supported by the Ministry of Science and Technology of the People's Republic of China under grant 2021ZD0201900, the National Natural Science Foundation of China under grant numbers 12090052 and 11874310, and Major Projects in Fujian Province under grant 2020Y4001.

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.

Publisher's note

References

Albert, M., Song, J. S., Schultz, D., Cormack, R. A., Tempany, C. M., Haker, S., et al. (2008). Defining the rectal dose constraint for permanent radioactive seed implantation of the prostate. *Urol. Oncol.* 26, 147–152. doi: 10.1016/j.urolonc.2007.03.026

Bang, M., Eom, J., An, C., Kim, S., Park, Y. W., Ahn, S. S., et al. (2021). An interpretable multiparametric radiomics model for the diagnosis of schizophrenia using magnetic resonance imaging of the corpus callosum. *Transl. Psychiatry* 11:462. doi: 10.1038/s41398-021-01586-2

Chen, X., Zhu, R., Zhong, J., Ying, Y., Wang, W., Cao, Y., et al. (2022). Mosaic composition of RIP1–RIP3 signalling hub and its role in regulating cell death. *Nat. Cell Biol.* 24, 471–482. doi: 10.1038/s41556-022-00854-7

Citrin, D. E. (2017). Recent developments in radiotherapy. *N. Engl. J. Med.* 377, 1065–1075. doi: 10.1056/NEJMra1608986

Fiorino, C., Alongi, F., Broggi, S., Cattaneo, G. M., Cozzarini, C., Di Muzio, N., et al. (2008). Physics aspects of prostate tomotherapy: planning optimization and image-guidance issues. *Acta Oncol.* 47, 1309–1316. doi: 10.1080/02841860802266755

Gerassy-Vainberg, S., Blatt, A., Danin-Poleg, Y., Gershovich, K., Sabo, E., Nevelsky, A., et al. (2018). Radiation induces proinflammatory dysbiosis: transmission of inflammatory susceptibility by host cytokine induction. *Gut*, 67, 97–107. doi: 10.1136/gutjnl-2017-313789

Hu, H., Liu, R., Zhao, C., Lu, Y., Xiong, Y., Chen, L., et al. (2022). CITEMO(XMBD): a flexible single-cell multimodal omics analysis framework to reveal the heterogeneity of immune cells. *RNA Biol.* 19, 290–304. doi: 10.1080/15476286.2022.2027151

Huang, E. Y., Wang, C. J., Hsu, H. C., Hao, L., Chen, H. C., and Sun, L. M. (2004). Dosimetric factors predicting severe radiation-induced bowel complications in patients with cervical cancer: combined effect of external parametrial dose and cumulative rectal dose. *Gynecol. Oncol.* 95, 101–108. doi: 10.1016/j.ygyno.2004.06.043

Huff, D. T., Weisman, A. J., and Jeraj, R. (2021). Interpretation and visualization techniques for deep learning models in medical imaging. *Phys. Med. Biol.* 66:04TR01. doi: 10.1088/1361-6560/abcd17

Jingxuan, J. S., Shuai, S. C., Qi, Z., and Jianwei, S. (2022). Predicting potential interactions between lncRNAs and proteins via combined graph auto-encoder methods. *Brief. Bioinform.* bbac527. doi: 10.1093/bib/bbac527

Kuenzi, B. M., Park, J., Fong, S. H., Sanchez, K. S., Lee, J., Kreisberg, J. F., et al. (2020). Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell* 38, 672–684.e6. doi: 10.1016/j.ccell.2020.09.014

Lam, L. H. T., Do, D. T., Diep, D. T. N., Nguyet, D. L. N., Truong, Q. D., Tri, T. T., et al. (2022). Molecular subtype classification of low-grade gliomas using magnetic resonance imaging-based radiomics and machine learning. *NMR Biomed.* 35:e4792. doi: 10.1002/nbm.4792

Le, N. Q. K., Kha, Q. H., Nguyen, V. H., Chen, Y. C., Cheng, S. J., and Chen, C. Y. (2021). Machine learning-based Radiomics signatures for EGFR and KRAS mutations prediction in non-Small-cell lung cancer. *Int. J. Mol. Sci.* 22:9254. doi: 10.3390/ijms22179254

Lee, S., Kerns, S., Ostrer, H., Rosenstein, B., Deasy, J. O., and Oh, J. H. (2018). Machine learning on a genome-wide association study to predict late genitourinary toxicity after prostate radiation therapy. *Int. J. Radiat. Oncol. Biol. Phys.* 101, 128–135. doi: 10.1016/j.ijrobp.2018.01.054

Lewis, J. E., and Kemp, M. L. (2021). Integration of machine learning and genome-scale metabolic modeling identifies multi-omics biomarkers for radiation resistance. *Nat. Commun.* 12:2700. doi: 10.1038/s41467-021-22989-1

Li, Y., He, Q., Guo, H., Zhong, C. Q., Li, X., Li, Y., et al. (2022). MSSort-DIA(XMBD): a deep learning classification tool of the peptide precursors quantified by OpenSWATH. *J. Proteomics* 259:104542. doi: 10.1016/j.jprot.2022.104542

Li, X., Zhang, P., Yin, Z., Xu, F., Yang, Z. H., Jin, J., et al. (2022). Caspase-1 and Gasdermin D afford the optimal targets with distinct switching strategies in NLRP1b Inflammasome-induced cell death. *Research* 2022, 9838341–9838317. doi: 10.34133/2022/9838341

Li, X., Zhong, C. Q., Wu, R., Xu, X., Yang, Z. H., Cai, S., et al. (2021). RIP1-dependent linear and nonlinear recruitments of caspase-8 and RIP3 respectively to necrosome specify distinct cell death outcomes. *Protein Cell* 12, 858–876. doi: 10.1007/s13238-020-00810-x

Liu, L., Chen, C., Liu, X., Chen, B., Ding, C., and Liang, J. (2021). Altered Gut Microbiota Associated With Hemorrhage in Chronic Radiation Proctitis. *Front Oncol.* 11:637265. doi: 10.3389/fonc.2021.637265

Lundberg, S. M., and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Adv. Neural Inf. Proces. Syst.* 30, 4765–4774. doi: 10.48550/arXiv.1705.07874

Mahal, B. A., Ziehr, D. R., Hyatt, A. S., Neubauer-Sugar, E. H., O'Farrell, D. A., O'Leary, M. P., et al. (2014). Use of a rectal spacer with low-dose-rate brachytherapy for treatment of prostate cancer in previously irradiated patients: initial experience and short-term results. *Brachytherapy* 13, 442–449. doi: 10.1016/j.brachy.2014.05.001

Mostafaei, S., Abdollahi, H., Kazempour Dehkordi, S., Shiri, I., Razzaghdoust, A., Zoljalali Moghaddam, S. H., et al. (2020). CT imaging markers to improve radiation toxicity prediction in prostate cancer radiotherapy by stacking regression algorithm. *Radiol. Med.* 125, 87–97. doi: 10.1007/s11547-019-01082-0

Park, Y. W., Eom, J., Kim, D., Ahn, S. S., Kim, E. H., Kang, S. G., et al. (2022). A fully automatic multiparametric radiomics model for differentiation of adult pilocytic astrocytomas from high-grade gliomas. *Eur. Radiol.* 32, 4500–4509. doi: 10.1007/s00330-022-08575-z

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Machine Learn. Res.* 12, 2825–2830. doi: 10.48550/arXiv.1201.0490

Qian, X., Qiu, Y., He, Q., Lu, Y., Lin, H., Xu, F., et al. (2021). A review of methods for sleep arousal detection using Polysomnographic signals. *Brain Sci.* 11:1274. doi: 10.3390/brainsci11101274

Qiang, M., Li, C., Sun, Y., Sun, Y., Ke, L., Xie, C., et al. (2021). A prognostic predictive system based on deep learning for Locoregionally advanced nasopharyngeal carcinoma. *J. Natl. Cancer Inst.* 113, 606–615. doi: 10.1093/jnci/djaa149

Rustagi, T., Corbett, F. S., and Mashimo, H. (2015). Treatment of chronic radiation proctopathy with radiofrequency ablation (with video). *Gastrointest. Endosc.* 81, 428–436. doi: 10.1016/j.gie.2014.04.038

Schmidt, D. R., Bhagwat, M., Glazer, D. I., Chen, M. H., Moteabbed, M., McMahon, E., et al. (2022). MRI-based radiotherapy planning to reduce rectal dose in excess of tolerance. *Prostate Cancer* 2022, 7930744–7930749. doi: 10.1155/2022/7930744

Shadad, A. K., Sullivan, F. J., Martin, J. D., and Egan, L. J. (2013). Gastrointestinal radiation injury: symptoms, risk factors and mechanisms. *World J. Gastroenterol.* 19, 185–198. doi: 10.3748/wjg.v19.i2.185

Shaji, S., Palanisamy, R., and Swaminathan, R. (2022). Explainable optimized LightGBM based differentiation of mild cognitive impairment using MR Radiomic features. *Stud. Health Technol. Inform.* 295, 483–486. doi: 10.3233/SHTI220770

Shi, Y., Zou, Y., Liu, J., Wang, Y., Chen, Y., Sun, F., et al. (2022). Ultrasound-based radiomics XGBoost model to assess the risk of central cervical lymph node metastasis in patients with papillary thyroid carcinoma: individual application of SHAP. *Front. Oncol.* 12:897596. doi: 10.3389/fonc.2022.897596

Snyder, K. M., Stock, R. G., Hong, S. M., Lo, Y. C., and Stone, N. N. (2001). Defining the risk of developing grade 2 proctitis following 125I prostate brachytherapy using a rectal dose-volume histogram analysis. *Int. J. Radiat. Oncol. Biol. Phys.* 50, 335–341. doi: 10.1016/s0360-3016(01)01442-0

Sun, F., Sun, J., and Zhao, Q. (2022). A deep learning method for predicting metabolite–disease associations via graph neural network. *Brief. Bioinform.* 23:bbac266. doi: 10.1093/bib/bbac266

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 71, 209–249. doi: 10.3322/caac.21660

Tang, Q., Nie, F., Zhao, Q., and Chen, W. (2022). A merged molecular representation deep learning method for blood-brain barrier permeability prediction. *Brief. Bioinform.* 23:bbac357. doi: 10.1093/bib/bbac357

van Griethuysen, J. J. M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., et al. (2017). Computational Radiomics system to decode the radiographic phenotype. *Cancer Res.* 77, e104–e107. doi: 10.1158/0008-5472.CAN-17-0339

Wang, Z., Wang, Q., Wang, X., Zhu, L., Chen, J., Zhang, B., et al. (2019). Gut microbial dysbiosis is associated with development and progression of radiation enteritis during pelvic radiotherapy. *J. Cell. Mol. Med.* 23, 3747–3756. doi: 10.1111/jcmm.14289

Wang, C. C., Han, C. D., Zhao, Q., and Chen, X. (2021). Circular RNAs and complex diseases: from experimental results to computational models. *Brief. Bioinform.* 22:bbab286. doi: 10.1093/bib/bbab286

Wang, W., Zhang, L., Sun, J., Zhao, Q., and Shuai, J. (2022). Predicting the potential human lncRNA-miRNA interactions based on graph convolution network with conditional random field. *Brief. Bioinform.* 23:bbac463. doi: 10.1093/bib/bbac463

Yeung, A. R., Pugh, S. L., Klopp, A. H., Gil, K. M., Wenzel, L., Westin, S. N., et al. (2020). Improvement in patient-reported outcomes with intensity-modulated radiotherapy (RT) compared with standard RT: a report from the NRG oncology RTOG 1203 study. *J. Clin. Oncol.* 38, 1685–1692. doi: 10.1200/JCO.19.02381

Yushkevich, P. A., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., Gee, J. C., et al. (2006). User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage* 31, 1116–1128. doi: 10.1016/j.neuroimage.2006.01.015

Zelefsky, M. J., Levin, E. J., Hunt, M., Yamada, Y., Shippy, A. M., Jackson, A., et al. (2008). Incidence of late rectal and urinary toxicities after three-dimensional conformal radiotherapy and intensity-modulated radiotherapy for localized prostate cancer. *Int. J. Radiat. Oncol. Biol. Phys.* 70, 1124–1129. doi: 10.1016/j.ijrobp.2007.11.044

Zhang, L., Yang, P., Feng, H., Zhao, Q., and Liu, H. (2021). Using network distance analysis to predict lncRNA–miRNA interactions. *Interdisciplin. Sci. Comput. Life Sci.* 13, 535–545. doi: 10.1007/s12539-021-00458-z

Zhong, L., Dong, D., Fang, X., Zhang, F., Zhang, N., Zhang, L., et al. (2021). A deep learning-based radiomic nomogram for prognosis and treatment decision in advanced nasopharyngeal carcinoma: a multicentre study. *EBioMedicine* 70:103522. doi: 10.1016/j.ebiom.2021.103522

Frontiers in Microbiology

Explores the habitable world and the potential of microbial life

The largest and most cited microbiology journal which advances our understanding of the role microbes play in addressing global challenges such as healthcare, food security, and climate change.

Discover the latest Research Topics

See more →

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact