



ARTIFICIAL INTELLIGENCE IN ENVIRONMENTAL MICROBIOLOGY

EDITED BY: Mohammad-Hossein Sarrafzadeh, Seyed Soheil Mansouri,
Javad Zahiri and Solange I. Mussatto

PUBLISHED IN: *Frontiers in Microbiology*



frontiers

Frontiers eBook Copyright Statement

The copyright in the text of individual articles in this eBook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this eBook is the property of Frontiers.

Each article within this eBook, and the eBook itself, are published under the most recent version of the Creative Commons CC-BY licence.

The version current at the date of publication of this eBook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or eBook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714

ISBN 978-2-88976-511-9

DOI 10.3389/978-2-88976-511-9

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

ARTIFICIAL INTELLIGENCE IN ENVIRONMENTAL MICROBIOLOGY

Topic Editors:

Mohammad-Hossein Sarrafzadeh, University of Tehran, Iran

Seyed Soheil Mansouri, Technical University of Denmark, Denmark

Javad Zahiri, Tarbiat Modares University, Iran

Solange I. Mussatto, Technical University of Denmark, Denmark

Citation: Sarrafzadeh, M.-H., Mansouri, S. S., Zahiri, J., Mussatto, S. I., eds. (2022). Artificial Intelligence in Environmental Microbiology. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88976-511-9

Table of Contents

- 04 Editorial: Artificial Intelligence in Environmental Microbiology**
Mohammad-Hossein Sarrafzadeh, Seyed Soheil Mansouri, Javad Zahiri, Solange I. Mussatto and Hashem Asgharnejad
- 07 Association Between Oral Microbiota and Human Brain Glioma Grade: A Case-Control Study**
Yuqi Wen, Le Feng, Haorun Wang, Hu Zhou, Qianqian Li, Wenyan Zhang, Ming Wang, Yeming Li, Xingzhao Luan, Zengliang Jiang, Ligang Chen and Jie Zhou
- 18 Deep Learning Driven Drug Discovery: Tackling Severe Acute Respiratory Syndrome Coronavirus 2**
Yang Zhang, Taoyu Ye, Hui Xi, Mario Juhas and Junyi Li
- 26 T1SEstacker: A Tri-Layer Stacking Model Effectively Predicts Bacterial Type 1 Secreted Proteins Based on C-Terminal Non-repeats-in-Toxin-Motif Sequence Features**
Zewei Chen, Ziyi Zhao, Xinjie Hui, Junya Zhang, Yixue Hu, Runhong Chen, Xuxia Cai, Yueming Hu and Yejun Wang
- 40 A Comparative Study of Deep Learning Classification Methods on a Small Environmental Microorganism Image Dataset (EMDS-6): From Convolutional Neural Networks to Visual Transformers**
Peng Zhao, Chen Li, Md Mamunur Rahaman, Hao Xu, Hechen Yang, Hongzan Sun, Tao Jiang and Marcin Grzegorzec
- 61 Machine Learning and Deep Learning Applications in Metagenomic Taxonomy and Functional Annotation**
Alban Mathieu, Mickael Leclercq, Melissa Sanabria, Olivier Perin and Arnaud Droit
- 71 Study on the Bacterial Communities of the Biofilms on Titanium, Aluminum, and Copper Alloys at 5,772 m Undersea in Yap Trench**
Xiaofan Zhai, Wei Cao, Yimeng Zhang, Peng Ju, Juna Chen, Jizhou Duan and Chengjun Sun
- 85 MDGNN: Microbial Drug Prediction Based on Heterogeneous Multi-Attention Graph Neural Network**
Jiangsheng Pi, Peishun Jiao, Yang Zhang and Junyi Li
- 95 EMDS-6: Environmental Microorganism Image Dataset Sixth Version for Image Denoising, Segmentation, Feature Extraction, Classification, and Detection Method Evaluation**
Peng Zhao, Chen Li, Md Mamunur Rahaman, Hao Xu, Pingli Ma, Hechen Yang, Hongzan Sun, Tao Jiang, Ning Xu and Marcin Grzegorzec
- 107 Interfacing Machine Learning and Microbial Omics: A Promising Means to Address Environmental Challenges**
James M. W. R. McElhinney, Mary Krystelle Catacutan, Aurelie Mawart, Ayesha Hasan and Jorge Dias
- 118 Clean and Safe Drinking Water Systems via Metagenomics Data and Artificial Intelligence: State-of-the-Art and Future Perspective**
Asala Mahajna, Inez J. T. Dinkla, Gert Jan W. Euverink, Karel J. Keesman and Bayu Jayawardhana



Editorial: Artificial Intelligence in Environmental Microbiology

Mohammad-Hossein Sarrafzadeh^{1*}, Seyed Soheil Mansouri², Javad Zahiri^{3,4}, Solange I. Mussatto⁵ and Hashem Asgharnejad⁶

¹ School of Chemical Engineering, College of Engineering, University of Tehran, Tehran, Iran, ² Department of Chemical and Biochemical Engineering, Technical University of Denmark, Kongens Lyngby, Denmark, ³ Department of Neuroscience, University of California, San Diego, San Diego, CA, United States, ⁴ Department of Pediatrics, University of California, San Diego, San Diego, CA, United States, ⁵ Department of Biotechnology and Biomedicine, Technical University of Denmark, Kongens Lyngby, Denmark, ⁶ Department of Civil, Geological, and Mining Engineering, Polytechnique Montréal, Montreal, QC, Canada

Keywords: artificial intelligence, deep learning, machine learning, environmental microbiology, predictive modeling, image analysis, metagenomics

Editorial on the Research Topic

Artificial Intelligence in Environmental Microbiology

Perhaps twenty-first century is so called “Digital Era” since digitalization and artificial intelligence (AI) is finding its way into every aspect of human life. Nowadays, AI-based approaches are gaining a lot of traction as components of research and development in different scientific and technological fields. One of the areas that is experiencing a digital revolution is environmental microbiology, which is the science of studying the interactions between the microorganisms and the environment and their mutual impacts (Pepper et al., 2011). Approaches such as machine learning (ML), deep learning (DL), image processing, pattern recognition and internet of things (IoT) are being widely implemented in this field in all aspects from theoretical development and identification to process monitoring and optimization (Asgharnejad and Sarrafzadeh, 2020; Gargalo et al., 2020; Asgharnejad et al., 2021). Outbreak of the global challenge of COVID-19 pandemic during the last 2 years and the huge impacts of this virus on socioeconomic infrastructures has also highlighted the necessity of innovative approaches for controlling and monitoring microbial communities in the environment. This special issue provides a platform for gathering the most recent advances in the fields of environmental microbiology from the perspective of AI. It includes 10 scientific papers (six original research articles, two mini-reviews and two reviews) that cover a wide range of AI approaches including ML, DL, and image processing. Two of these papers are specifically focused on using AI for diagnosis and tackling the SARS-CoV-2 virus, which is the species causing COVID-19 and, in this regard, the current Research Topic can be a reference for ongoing research on the edge of science to overcome the pandemic and prevent future such catastrophic outbursts.

Moreover, AI can be used to diagnose and find effective treatments for microbial-risen diseases. Previous studies have shown that oral microbiota has a close relation with different types of cancer. Wen et al. have studied the possible relation between the oral microbiota and gliomas, which are the most prevalent form of primary malignant brain tumors. They conducted an association rule mining algorithm to find the relation between the microbiota existed in the saliva of a compound sample containing 35 patients diagnosed with high-grade and low-grade glioma and 24 control samples. The results of their study determined the oral microbiota features and gene functions that were associated with glioma malignancy, which is a great achievement in terms of cancer therapy.

Zhang et al., conducted a literature review on how DL can accelerate the procedure of drug discovery to tackle Severe Acute Respiratory Syndrome Coronavirus 2, which is globally known

OPEN ACCESS

Edited by:

George Tsiamis,
University of Patras, Greece

Reviewed by:

Vassiliki Karapapa,
Municipality of Agrinio, Greece

*Correspondence:

Mohammad-Hossein Sarrafzadeh
sarrafzdh@ut.ac.ir

Specialty section:

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

Received: 15 May 2022

Accepted: 30 May 2022

Published: 13 June 2022

Citation:

Sarrafzadeh M-H, Mansouri SS,
Zahiri J, Mussatto SI and
Asgharnejad H (2022) Editorial:
Artificial Intelligence in Environmental
Microbiology.
Front. Microbiol. 13:944242.
doi: 10.3389/fmicb.2022.944242

as COVID-19. Besides improving the efficacy of antimicrobial screening against a wide range of pathogens, DL has shown potential to reliably identify drug candidates against newly-emerged diseases such as COVID-19 and a number of drugs including Atazanavir, Remdesivir, Kaletra, Enalaprilat, Venetoclax, and Posaconazole have been proposed to be effective based on applying DL algorithm on the datasets of genetic and protein types of SARS-CoV-2. Pi et al., also applied Heterogeneous Multi-Attention Graph Neural Network with the objective of drug prediction for COVID-19 and two drugs were successfully predicted and verified by their model.

DL can also be used to train a model for identifying the substrate secretion mechanism in Gram-negative bacteria. Chen et al. followed this approach using the sequence-based non-RTX-motif features and combined it into a tri-layer stacking model, T1SEstacker, which predicted the RTX proteins accurately. This model can accurately estimate the various substrate proteins being secreted through the two bacterial cell membranes by one step (classical) or two steps (non-classical) into extracellular environment.

Image datasets of environmental microorganisms such as EMDS-6 are powerful tools for diagnosis and classifications of newly discovered microorganisms. In the research carried out by Zhao, Li, Rahaman, Xu, Yang, et al., a comparative study was conducted of DL methods of classification using the EMDS-6 image dataset. The authors compared 21 DL-based methods of image classification including direct classification, imbalanced training, and hyper-parameters tuning. Zhao, Li, Rahaman, Xu, Ma, et al. also used the classic algorithms of image processing such as image denoising, image segmentation, and object detection to analyze the EMDS-6 and its potential for being used to evaluate the performance of image processing algorithms.

Metagenomics is a revolutionary field that burst fundamental changes in environmental microbiology, which allows the characterization of all microorganisms in a sequencing experiment. To distinguish the microbes in terms of taxonomy and biological activity, the sequenced reads must essentially be associated with known genomes/genes. However, current association methods are inadequate in terms of rapidity and also accuracy, especially when detecting bacterial species or in specific cases such as virus, plasmids, and gene detection. Machine and deep learning methods can use the newly reconstructed genomes by metagenomics as models and be a platform for association. Mathieu et al. tried to assess the different machine learning based methods and their efficiency to enhance the annotation of metagenomic sequences.

The application of metagenomics data can be expanded to clean and safe drinking water systems. Mahajna et al. reviewed

the literature on what kind of occupancy-abundance patterns are exhibited in the drinking water microbiome, how the drinking water microbiome evolves both spatially and temporally, and how different microbial communities can co-exist in the drinking water environment. They also evaluated the potential role of AI in addressing the predictive and mechanistic questions in this field.

However, applications of AI are not limited to human health and genetics, but they can be expanded to environmental protection areas as well, especially in hardly controllable environments such as aquacultures. Zhai et al. used AI to study the bacterial biofilms on the metallic alloys at 5,700 m depth undersea. They derived the sequencing data of the microbial communities of the biofilms and applied big data analytics methods to study the dataset and compare the microbial composition of the biofilms on different alloy surfaces. McElhinney et al. also reviewed the capability of ML algorithms for analyzing the big microbiological datasets produced as a result to the advent of microbial omics. The authors provided a briefing for ML, highlighting the concept of retaining biological sample information for supervised ML, and reviewed the state-of-the-art of ML-driven microbial ecology.

In this Research Topic, cutting-edge scientific research works are gathered focused on applications of AI methods for identifying, monitoring, and analyzing environmental microorganisms and alleviating their hazardous impacts on human life. The gaps and challenges addressed in this Research Topic can be the hot topic of further studies in the future regarding the comprehension of what is going to be anticipated as a crucial concept for tackling the challenges in the area of environmental microbiology in the forthcoming years.

AUTHOR CONTRIBUTIONS

All authors have made a significant, equal, direct and intellectual contribution to the preparation of this editorial note, and approved it for publication.

ACKNOWLEDGMENTS

This Research Topic was focused on the most recent advances in the field of AI and environmental microbiology. There is a total of 10 accepted manuscripts by 61 authors from China, Germany, France, Canada, Netherlands and UAE. We would like to show our gratitude to the researchers and all the esteemed reviewers from all over the globe including Canada, India, US, UK, Australia, Iran, Taiwan, Spain and South Africa whose precious contributions significantly enhanced the quality of the topic.

REFERENCES

- Asgharnejad, H., and Sarrafzadeh, M.-H. (2020). Development of digital image processing as an innovative method for activated sludge biomass quantification. *Front. Microbiol.* 11, 574966. doi: 10.3389/fmicb.2020.574966
- Asgharnejad, H., Sarrafzadeh, M.-H., Abhar-Shegoftah, O., Nazloo, E. K., and Oh, H.-M. (2021). Biomass quantification and 3-D topography reconstruction of microalgal biofilms using digital image processing. *Algal Res.* 55, 102243. doi: 10.1016/j.algal.2021.102243
- Gargalo, C. L., Udugama, I., Pontius, K., Lopez, P. C., Nielsen, R. F., Hasanazadeh, A., et al. (2020). Towards smart biomanufacturing:

a perspective on recent developments in industrial measurement and monitoring technologies for bio-based production processes. *J. Indust. Microbiol. Biotechnol.* 47, 947–964. doi: 10.1007/s10295-020-02308-1

Pepper, I. L., Gerba, C. P., Gentry, T. J., and Maier, R. M. (2011). *Environmental Microbiology*. London: Academic Press.

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.

Copyright © 2022 Sarrafzadeh, Mansouri, Zahiri, Mussatto and Asgharnejad. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Association Between Oral Microbiota and Human Brain Glioma Grade: A Case-Control Study

Yuqi Wen^{1,2†}, Le Feng^{3†}, Haorun Wang^{1,2†}, Hu Zhou^{1,2}, Qianqian Li¹, Wenyan Zhang^{1,2}, Ming Wang^{1,2}, Yeming Li^{1,2}, Xingzhao Luan^{1,2}, Zengliang Jiang^{4,5*}, Ligang Chen^{1,2,6,7*} and Jie Zhou^{1,2,6,7*}

OPEN ACCESS

Edited by:

Solange I. Mussatto,
Technical University of Denmark,
Denmark

Reviewed by:

Zhiqiang Qin,
University of Arkansas for Medical
Sciences, United States
Dana Marshall,
Meharry Medical College,
United States

*Correspondence:

Zengliang Jiang
jiangzengliang@westlake.edu.cn
Ligang Chen
chengligang.cool@163.com
Jie Zhou
zj000718@yeah.net

[†]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: 24 July 2021

Accepted: 24 September 2021

Published: 18 October 2021

Citation:

Wen Y, Feng L, Wang H, Zhou H,
Li Q, Zhang W, Wang M, Li Y, Luan X,
Jiang Z, Chen L and Zhou J (2021)
Association Between Oral Microbiota
and Human Brain Glioma Grade: A
Case-Control Study.
Front. Microbiol. 12:746568.
doi: 10.3389/fmicb.2021.746568

¹Department of Neurosurgery, The Affiliated Hospital of Southwest Medical University, Luzhou, China, ²Sichuan Clinical Medical Research Center for Neurosurgery, Luzhou, China, ³Department of Prosthodontics, The Affiliated Stomatology Hospital of Southwest Medical University, Luzhou, China, ⁴School of Life Sciences, Westlake University, Hangzhou, China, ⁵Key Laboratory of Growth Regulation and Translational Research of Zhejiang Province, School of Life Sciences, Westlake University, Hangzhou, China, ⁶Neurological Diseases and Brain Function Laboratory, The Affiliated Hospital of Southwest Medical University, Luzhou, China, ⁷Academician (Expert) Workstation of Sichuan Province, The Affiliated Hospital of Southwest Medical University, Luzhou, China

Gliomas are the most prevalent form of primary malignant brain tumor, which currently have no effective treatments. Evidence from human studies has indicated that oral microbiota is closely related to cancers; however, whether oral microbiota plays a role in glioma malignancy remains unclear. The present study aimed to investigate the association between oral microbiota and grade of glioma and examine the relationship between malignancy-related oral microbial features and the isocitrate dehydrogenase 1 (IDH1) mutation in glioma. High-grade glioma (HGG; $n = 23$) patients, low-grade glioma (LGG; $n = 12$) patients, and healthy control (HCs; $n = 24$) participants were recruited for this case-control study. Saliva samples were collected and analyzed for 16S ribosomal RNA (rRNA) sequencing. We found that the shift in oral microbiota β -diversity was associated with high-grade glioma ($p = 0.01$). The phylum Patescibacteria was inversely associated with glioma grade (LGG and HC: $p = 0.035$; HGG and HC: $p < 0.01$). The genera *Capnocytophaga* (LGG and HC: $p = 0.043$; HGG and HC: $p < 0.01$) and *Leptotrichia* (LGG and HC: $p = 0.044$; HGG and HC: $p < 0.01$) were inversely associated with glioma grades. The genera *Bergeyella* and *Capnocytophaga* were significantly more positively correlated with the IDH1 mutation in gliomas when compared with the IDH1-wild-type group. We further identified five oral microbial features (*Capnocytophaga Porphyromonas*, *Haemophilus*, *Leptotrichia*, and *TM7x*) that accurately discriminated HGG from LGG (area under the curve [AUC]: 0.63, 95% confidence interval [CI]: 0.44–0.83) and HCs (AUC: 0.79, 95% CI: 0.68–0.92). The functional prediction analysis of oral bacterial communities showed that genes involved in cell adhesion molecules ($p < 0.001$), extracellular matrix molecule-receptor interaction ($p < 0.001$), focal adhesion ($p < 0.001$), and regulation of actin cytoskeleton ($p < 0.001$) were associated with glioma grades, and some microbial gene functions involving lipid metabolism and the adenosine 5'-monophosphate-activated protein kinase signaling pathway were significantly more enriched in IDH1 mutant gliomas

than compared with the IDH1-wild-type gliomas. In conclusion, our work revealed oral microbiota features and gene functions that were associated with glioma malignancy and the IDH1 mutation in glioma.

Keywords: glioma, malignant grade, oral microbiota, isocitrate dehydrogenase 1 mutation, human cohort

INTRODUCTION

Gliomas, the most common primary tumor of the central nervous system, are stratified into grades 1–4 based on the histological features defined by the World Health Organization (WHO; Louis et al., 2016). This classification system has been transformed into a molecular feature-based classification system and can be used for the formulation of targeted therapeutic methods, which have been shown to have a higher level of prognostic accuracy (Louis et al., 2016, 2020; Brat et al., 2020). WHO grades 1–2 gliomas (low-grade gliomas; LGG) exhibit low aggressive tendencies and have a better prognosis, whereas WHO grades 3–4 gliomas (high-grade gliomas; HGG) have a high rate of deterioration and a poor prognosis (Louis et al., 2016). Evidence from human studies has indicated that oral microbiota is closely related to cancers (Michaud et al., 2013; Fan et al., 2018); however, whether oral microbiota plays a role in glioma malignancy remains unclear.

The presence of mutant forms of isocitrate dehydrogenase 1 (IDH1) is a key factor in determining the prognosis of patients with gliomas. Generally speaking, glioma patients with the IDH1 mutation have a more favorable prognosis, and the mutation is frequently expressed in patients with an LGG but rarely detected in patients with a WHO grade 4 glioma (Ceccarelli et al., 2016). Oral microbial-produced substances may be carcinogenic (Kurkivuori et al., 2007; Meurman and Uitto, 2008). Therefore, the mechanism underlying the link between the IDH1 mutation and better prognosis may involve the variation in the composition and function of oral microbiota. Currently, the relationship between oral microbiota and the glioma IDH1 mutation is uncertain.

Therefore, we aimed to investigate the association between oral microbiota and glioma grade and examine the relationship between the composition and functional features of malignancy-related oral microbiota and the glioma IDH1 mutation.

MATERIALS AND METHODS

Study Subjects and Study Design

The cohort was a prospective cohort that included 59 participants of Han Chinese ethnicity. This study was a cross-sectional analysis of the retrospective cohort at baseline. Briefly, 59 participants aged 20–74 years who were living in Luzhou City, Southern China, were recruited by the Department of Neurosurgery, Affiliated Hospital of Southwest Medical University, between April 2019 and October 2020. We collected sociodemographic, lifestyle, and dietary factor information. Anthropometric parameters, which included weight and height, were measured by trained nurses. The glioma IDH1 mutation was assessed by postoperative pathological diagnosis. Clinical physiological variables (blood

glucose value) were measured using a blood glucose monitor. Oral saliva samples of participants were collected during participants' visits to the study site. We excluded participants who had been physiologically diagnosed with a non-glioma and had not received surgery. Finally, 59 participants were included in the present analysis (see inclusion and exclusion criteria for further details; **Supplementary Methods; Supplementary Figure 1**).

Glioma grades were assessed according to the 2016 WHO Classification of Tumors of the Central Nervous System. Participants were divided into three groups: (i) healthy controls (HCs; $n=24$), (ii) LGG group ($n=12$), and (iii) HGG group ($n=23$) according to the criteria for glioma grades. HCs were recruited from patients' families, such as their spouses and parents. The study protocol was approved by the ethics committee of the Affiliated Hospital of Southwest Medical University (No. KY2019030), and all participants provided written informed consent.

Sample Collection and DNA Extraction

Participants were asked to refrain from drinking, eating, brushing teeth, or smoking on the morning of the study visit and to rinse out impurities in the mouth with sterile saline. During the visit to the study center, participants were provided with a saliva sampler and detailed instructions for the saliva sample collection. Briefly, each participant collected their saliva sample by natural secretion, which was kept in the mouth for 3 min. They then spat 20 ml of saliva into a 50-ml sterile centrifugal tube. All saliva samples were immediately frozen and stored in a -80°C freezer.

All DNA extraction steps were performed in a biosafety cabinet. Oral saliva DNA was isolated using a QIAamp Fast DNA Stool Mini Kit (QIAGEN, Hilden, Germany) according to standard protocols. NanoDrop was used to quantitatively detect the concentration of DNA in each sample, and 1% agarose gel electrophoresis was used to evaluate the integrity of DNA. The DNA samples that met the quality requirements (i.e., $A_{260}/A_{280}=1.8\text{--}2.8$, total DNA $>500\text{ ng}$, the main strip of the gel was complete without an obvious tail) were frozen at -80°C for subsequent analyses, and non-conforming samples were discarded.

Oral Microbiota Profiling Using 16S rRNA Sequencing

The 16S ribosomal RNA (rRNA) gene amplification procedure was divided into two polymerase chain reaction (PCR) steps. For the first PCR reaction, the V3-V4 hypervariable region of the 16S rRNA gene was amplified from the genomic DNA using primers 338F (ACTCCTACGGGAGGCAGCAG) and 806R (GGACTACHVGGGTWTCTAAT). The amplification products were purified by gel extraction (AxyPrep DNA Gel Extraction Kit, Axygen Biosciences, Union City, CA, United States) according to manufacturer instructions. The concentration of the

pooled libraries was determined using the Qubit quantification system. Amplicon sequencing was performed on the MiSeq PE250 platform (Illumina, San Diego, California, United States). Automated cluster generation and 2×250bp paired-end sequencing with dual-index reads were performed.

16S rRNA Gene Sequencing Bioinformatics Analysis

Sequence analysis was performed using the Quantitative Insights into Microbial Ecology Pipeline (QIIME) software version 2-2020.2 (Bolyen et al., 2019). The divisive amplicon denoising algorithm (DADA2; Callahan et al., 2016) was used for amplicon sequence variant (ASV) clustering equaling 100%. A representative sequence was selected for each ASV, and the SILVA reference database was used to annotate taxonomic information. The absolute abundance table was extracted from the pipeline and converted into relative abundances by normalization for analyzing the composition of gut microbiota by QIIME2 for downstream analysis.

Statistical Analyses

For comparisons between the three groups, we used the chi-square test for categorical variables and analysis of variance (ANOVA) for continuous variables. We examined the associations between glioma grade and oral microbial α -diversity indices (Shannon, ACE, and Chao1 indices and Good's coverage), which were estimated based on species richness in the ASV subsample table. The association between glioma grade and β -diversity (between-subject diversity) dissimilarity, based on ASV-level Bray-Curtis distance, was analyzed using permutational ANOVA (999 permutations), adjusted for age, sex, and body mass index (BMI).

We used Wilcoxon rank-sum tests to determine oral microbiota and microbial function associations with glioma grade (a value of $p \leq 0.05$ was considered statistically significant). The Benjamini-Hochberg method was used to control for false discovery rate (FDR). Receiver operator characteristic curves based on the identified oral microbial features were used to discriminate different glioma grade patients from HCs. The true positive rate (sensitivity) was plotted against the false positive rate (100% – specificity), and the area under the curve (AUC) values were reported with 95% confidence intervals (CI) as an estimate of diagnostic utility.

We examined the associations between the composition and functional features of malignancy-related oral microbiota and the glioma IDH1 mutation using multivariable linear regression, adjusted for age, sex, and BMI. The Benjamini-Hochberg method was used to control for FDR. We also conducted stratified analyses for smoking and alcohol status. Analyses were carried out using R statistical software (version 3.3.1, R Foundation). A value of $p < 0.05$ was considered statistically significant.

RESULTS

Characteristics of Study Participants

The demographic characteristics of participants are shown in Table 1. The mean (standard deviation) age was 46.12 (12.69) years, and 47.46% were women (Table 1).

Association Between Glioma Grade and Oral Microbiota Diversity

We first investigated the associations between glioma grade and microbiome α -/ β -diversity. A significant difference in microbial β -diversity ($p = 0.01$) was found between the HGG and HC groups (Figure 1A), whereas no significant difference was found between the LGG and HC groups ($p = 0.51$; Figure 1B), and no significant difference was found between the LGG and HGG groups ($p = 0.89$; Figure 1C). There were no significant differences in measures of α -diversity between the glioma groups and the HCs [HGG and HC: Shannon index: $p = 0.34$, phylogenetic diversity (PD): $p = 0.86$; LGG and HC: Shannon index: $p = 0.91$, PD: $p = 0.21$].

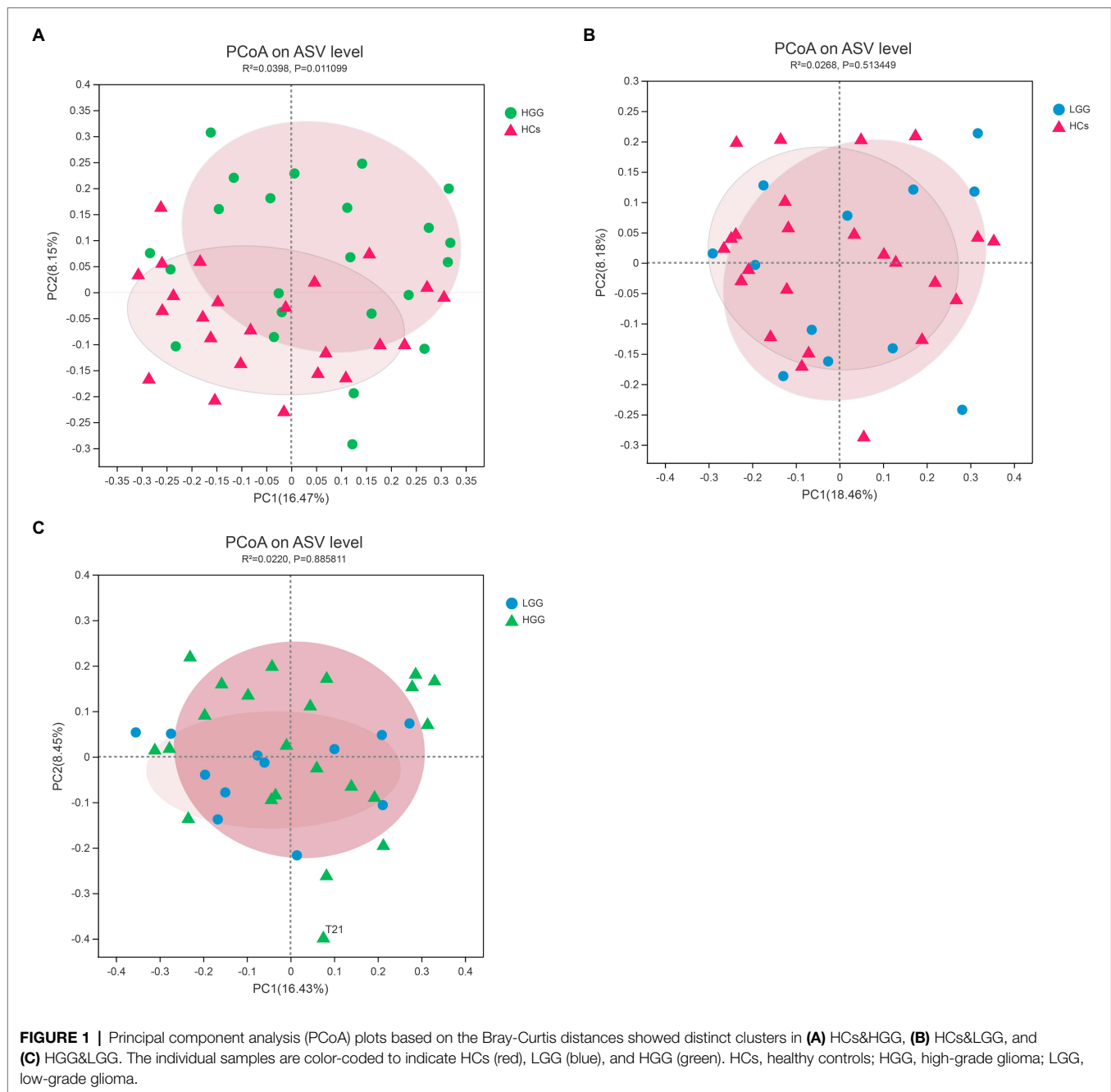
Glioma Grades Were Associated With Oral Microbiota Composition and Gene Function

Our results showed that 99.57% of the oral microbiota was aligned to seven phyla, which included *Firmicutes*, *Bacteroidetes*, *Proteobacteria*, *Actinobacteria*, *Fusobacteria*, *Patescibacteria*, and *Spirochaetota* (Figures 2A,B). The abundance of *Patescibacteria* decreased significantly with increasing malignancy of glioma from LGG ($p = 0.035$) to HGG ($p < 0.01$), compared with HCs

TABLE 1 | Clinical characteristics of the study population in this study.

Characteristics		Glioma grade		
		HGs (N = 24)	LGG (N = 12)	HGG (N = 23)
Age (years)				
Range		32–57	20–61	21–74
Mean \pm SD		45.29 \pm 6.72	37.67 \pm 13.49	51.39 \pm 14.75
Gender				
Male		10 (41.67%)	8 (66.67%)	13 (56.52%)
Female		14 (58.33%)	4 (33.33%)	10 (43.48%)
BMI (kg/m²)				
Range		19.59–37.78	18.51–25.77	18.07–28.69
Mean \pm SD		25.59 \pm 3.42	22.18 \pm 1.90	22.99 \pm 2.90
Blood glucose				
Range		—	4.4–14.9	4.77–22.5
Mean \pm SD		—	7.0 \pm 3.06	8.13 \pm 3.63
Excrement regularity				
Yes		17 (70.83%)	9 (75%)	16 (69.57%)
No		7 (29.17%)	3 (25%)	7 (30.43%)
Brushing habits				
Numbers	0 time	0 (0%)	3 (25%)	3 (13.05%)
	1 time	20 (83.33%)	7 (58.33%)	13 (56.52%)
	2 times	4 (16.67%)	2 (16.67%)	7 (30.43%)
Time	< 2 min	21 (87.5%)	6 (50%)	22 (95.65%)
	> 2 min	3 (12.5%)	6 (50%)	1 (4.35%)
Tooth missing				
Yes		7 (29.17%)	5 (41.67%)	9 (39.13%)
No		17 (70.83%)	7 (58.33%)	14 (60.87%)

In this study, according to the recruitment standard, 59 cases including 24 HCs and 35 glioma patients were included. Glioma patients were divided into LGG (N = 12), HGG (N = 23). HCs, healthy controls; LGG, low-grade glioma; HGG, high-grade glioma.

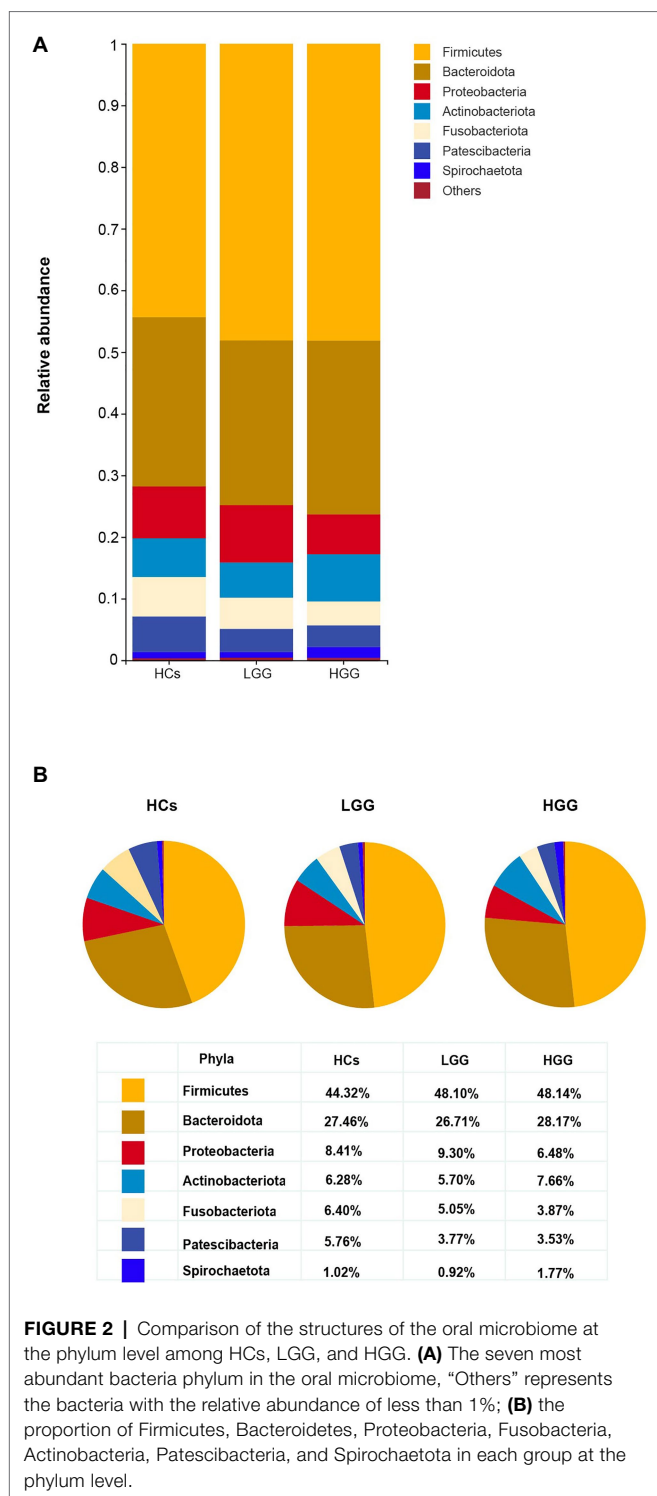


(Figure 3A). The proportions of the other six phyla (*Firmicutes*, *Bacteroidetes*, *Proteobacteria*, *Actinobacteria*, *Fusobacteria*, and *Spirochaetota*) were not associated with glioma grade (Supplementary Figure 2).

The ASVs were assigned to 181 individual genera of which 18 were present in all samples with a relative abundance of more than 1% in at least one sample (Supplementary Figures 3A,B). The genera *Capnocytophaga* (LGG and HC: $p=0.043$; HGG and HC: $p<0.01$) and *Leptotrichia* (LGG and HC: $p=0.044$; HGG and HC: $p<0.01$) were inversely associated with glioma grade (Figure 3B). Five oral microbial features [*Porphyromonas* ($p<0.05$), *Haemophilus*, *Leptotrichia* ($p<0.05$), *TM7x*

($p<0.05$), and *Capnocytophaga* ($p<0.05$)] were significantly lower in the HGG group compared with the HC group (Figure 3C). The five oral microbial features (*Porphyromonas*, *Haemophilus*, *Leptotrichia*, *TM7x*, and *Capnocytophaga*) accurately discriminated the HGG group from HCs (AUC: 0.79, 95% CI: 0.68–0.92; Figure 4 and Supplementary Table 1). We also use bacterial marker panels to discriminate the HGG group from the LGG group (AUC: 0.63, 95% CI: 0.44–0.83) and the LGG group from HCs (AUC: 0.57, 95% CI: 0.36–0.78; Figure 4).

We used the Phylogenetic Investigation of Communities by Reconstruction of Unobserved States (PICRUST) to predict



the oral microbiome functions *via* the saliva microbiome data sets. Signaling molecules and interactions, such as cell adhesion molecules (CAMs), extracellular matrix (ECM)-receptor interactions, cellular community-eukaryotes (focal adhesion), and actin cytoskeleton regulation, were positively associated with glioma grade, while the polycyclic aromatic

hydrocarbon degradation and the Bile secretion were inversely associated with HCs (Figure 5).

Glioma IDH1 Mutation Was Associated With Malignancy-Related Oral Microbiota and Gene Function

Results showed that the abundance of *Firmicutes* was significantly lower in the IDH-mutant samples compared with that of the IDH-wild-type samples at the phylum level (Figure 6A). The genus-level profiling showed that the abundance of *Bergeyella* and *Capnocytophaga* was significantly positively correlated with the IDH-mutant samples (Figure 6B). We also used PICRUSt to predict Kyoto Encyclopedia of Genes and Genomes (KEGG)-based functional orthologs between the IDH-mutant and IDH-wild-type groups to characterize the functional alterations, which were inferred from the 16S rRNA gene sequencing data. We found 82 pathways that were significantly greater in the IDH-mutant group compared with those of the IDH-wild-type group using the Mann-Whitney U test. Our result revealed that several metabolic-related pathways, such as lipid metabolism (linoleic acid, ether lipid metabolism, fatty acid biosynthesis, glycerophospholipid, and biosynthesis of unsaturated fatty acids), fatty acid metabolism (alpha-linolenic acid), amino acid metabolism (tyrosine metabolism, tryptophan metabolism, lysine degradation, and glycine, serine, and threonine metabolism), carbohydrate metabolism (inositol phosphate metabolism), and cofactors and vitamin metabolism (retinol metabolism), were significantly higher in the IDH-mutant group compared with those of the IDH-wild-type group. Moreover, signal transduction, such as the adenosine 5'-monophosphate-activated protein kinase (AMPK) signaling pathway, the phosphatidylinositol signaling system, the sphingolipid signaling pathway, and the phospholipase D signaling pathway, was more enriched in the IDH-mutant group than in the IDH-wild-type group (Supplementary Figure 4).

DISCUSSION

To our knowledge, this is the first comprehensive clinical 16S rRNA sequencing data set to characterize the community features of oral microbiota in different glioma grades. We found that the shift in oral microbiota β -diversity was associated with HGG. The phylum *Patescibacteria* was inversely associated with glioma grade, and the genera *Capnocytophaga* and *Leptotrichia* were inversely associated with glioma grade. We identified five oral microbial features (*Porphyromonas*, *Haemophilus*, *Leptotrichia*, *TM7x*, and *Capnocytophaga*) that accurately discriminated patients with HGG from those with LGG and HCs. The gene function of oral bacterial communities was associated with glioma grade. Moreover, the abundance of the phylum *Firmicutes* was significantly negatively correlated with IDH-mutant samples, whereas the genera *Bergeyella* and *Capnocytophaga* were significantly

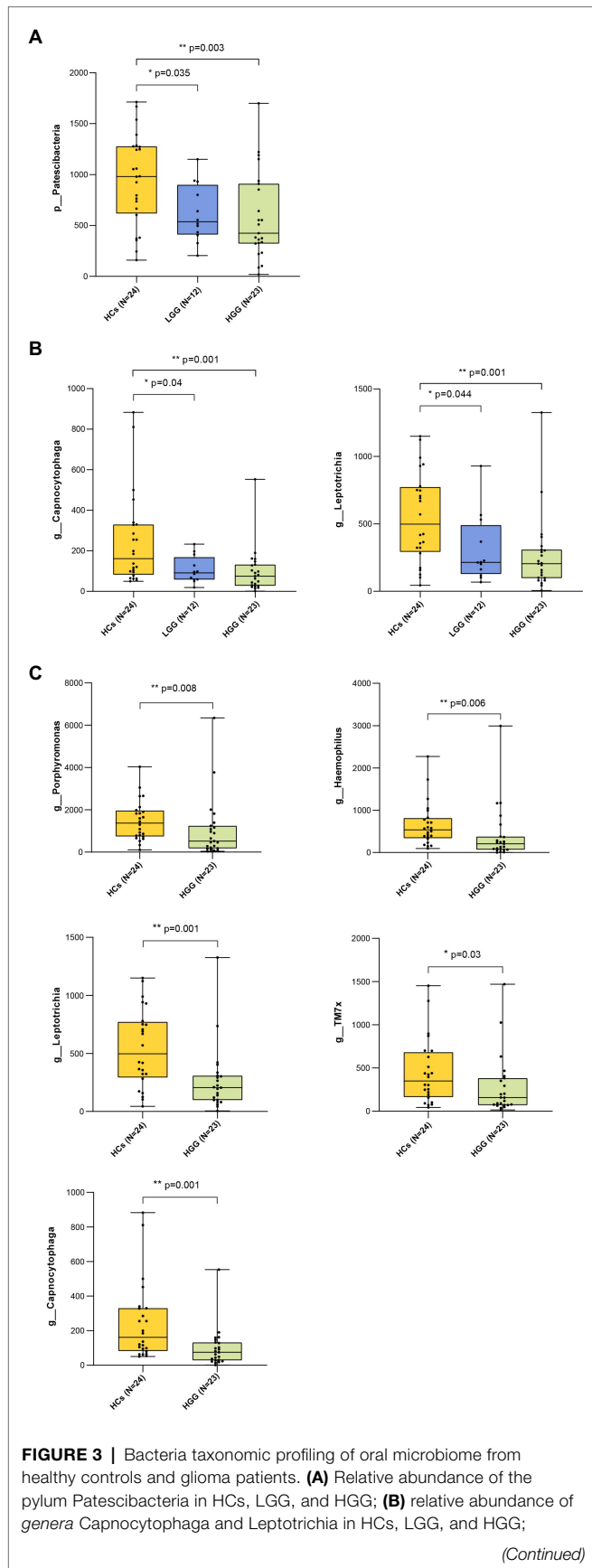
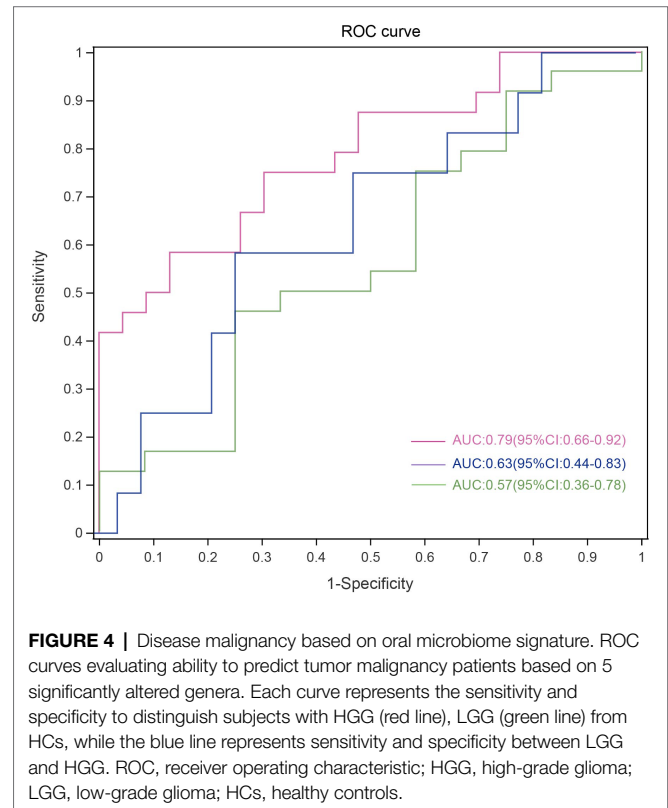


FIGURE 3 | **(C)** relative abundance of genera Porphyromonas, Haemophilus, Leptotrichia, TM7x, Capnocytophaga in HCs, and HGG. The box plot represented the relative abundance of bacteria genus in HCs, LGG, and HGG. The p value was calculated by non-parametric Mann-Whitney U test. Each box plot represents the median, interquartile range, minimum, and maximum values. p value <0.05 indicated the statistical significance. HCs, healthy controls; LGG, low-grade glioma; HGG, high-grade glioma.



positively correlated with IDH-mutant samples. Several microbial (lipid metabolism, amino acid metabolism, and energy metabolism) and signal transduction (AMPK signaling pathway) pathways were significantly higher in the glioma IDH-mutant group than those in the glioma IDH-wild-type group. Our findings revealed that oral microbiota features and gene functions are associated with glioma malignancy and the IDH1 mutation.

Previous studies have shown that oral microbiome significantly affected the composition of gut microbiome (Hou et al., 2021; Pandya et al., 2021; Park et al., 2021). The oral microbiome and gut microbiome can spread to the brain through cranial nerves or cellular infections or produced certain metabolites to affect the brain by both direct and indirect means (Dominy et al., 2019; Jing et al., 2021; Narengaowa et al., 2021; Vyhnałova et al., 2021; Wang et al., 2021). Our results showed significant differences in oral microflora between HGG patients and HCs; oral *Patescibacteria* was significantly decreasing during the progression of glioma malignancy. Few studies have reported



FIGURE 5 | Microbial functions altered in the HCs and HGG. Heat map showing the median abundance of all significant modules as determined by PICRUST analysis at HCs and HGG. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. PICRUST, Phylogenetic Investigation of Communities by Reconstruction of Unobserved States; HCs, healthy controls; HGG, high-grade glioma.

a correlation between oral *Patescibacteria* and human disease. This is the first study examining *Patescibacteria* in oral saliva samples of glioma patients and suggests *Patescibacteria* as a negatively associated risk factor for disease progression from LGG to HGG. Our results highlight the potential of *Patescibacteria* detection as a diagnostic and prognostic determinant for glioma malignancy, but further experimental research to establish the mechanistic basis of these relationships is needed.

In addition, the phylum *Fusobacteriota* was significantly lower in the HGG group than in the HCs. Furthermore, the family *Leptotrichiaceae* (**Supplementary Figure 3C**) and

the genus *Leptotrichia*, which belongs to *Fusobacteriota*, were inversely associated with glioma malignancy. This is consistent with the results of two large, nested, case-control studies in which a greater abundance of *Leptotrichia* was associated with a decreased risk of pancreatic cancer (Michaud et al., 2013; Fan et al., 2018). *Leptotrichia* is considered an opportunistic pathogen and can stimulate human immune system responses (Eribe and Olsen, 2017). Moreover, *Leptotrichia* may elicit the immune response and thus protect against pancreatic carcinogenesis (Inman et al., 2014).

Genus-level analysis showed that a bacterial marker panel with *Capnocytophaga*, *TM7x*, *Porphyromonas*, *Haemophilus*,

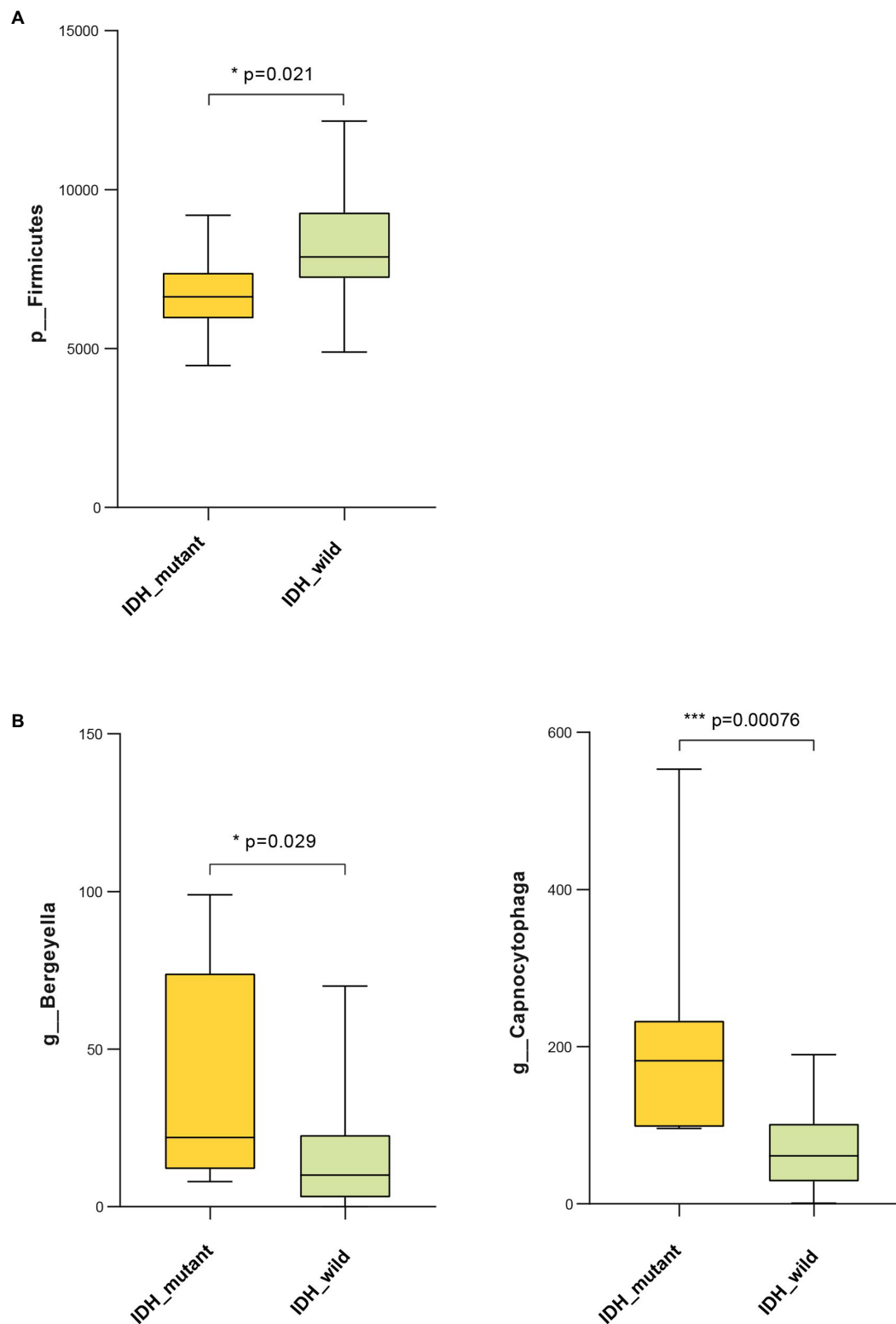


FIGURE 6 | Bacteria taxonomic profiling of oral microbiome from IDH1-mutant group and IDH1-wild group. **(A)** Relative abundance of phylum Proteobacteria in IDH1-mutant group and IDH1-wild group; **(B)** relative abundance of genera Bergeyella and Capnocytophaga in IDH1-mutant group and IDH1-wild group. Cross coordinates represent different group names and longitudinal represent the abundance of a species in different groups. The p value was calculated by non-parametric Mann-Whitney U test. Each box plot represents the median, interquartile range, minimum, and maximum values. IDH1, isocitrate dehydrogenase 1.

and *Leptotrichia* had an AUC of 0.79 for discriminating between HGG and HCs. We found that the relative abundance of genus *Capnocytophaga* was inversely associated with glioma malignancy. *Capnocytophaga* is a genus of Gram-negative anaerobes that inhabit the oral cavity (Lopez et al., 2010), which has been reported to be inversely associated with human diseases, such as chorioamnionitis, neonatal infection (Lopez et al., 2010), and lung cancer (Thirumala et al., 2012). Consistent with our results, Hayes et al. reported that a greater abundance of the genus *Capnocytophaga* was significantly associated with a reduced risk of larynx cancer (Hayes et al., 2018). However, the protective mechanism of *Capnocytophaga* for glioma malignancy remains unclear. Previous studies have consistently shown a decrease in *Haemophilus* and *Porphyromonas* in the saliva of patients with cancer compared with that of HCs (Mei et al., 2018; Yang et al., 2018; Lu et al., 2019; Li et al., 2020). In addition, the high abundance of *Haemophilus* and *Porphyromonas* is associated with anticancer-associated immunity (Lu et al., 2019). A high level of antibodies to *Porphyromonas gingivalis* in the serum correlates with a lower risk of pancreatic cancer (Michaud et al., 2013). This study is the first to report the inverse association between the abundance of *TM7x* and glioma malignancy and that *TM7x* is a useful bacteria marker for glioma malignancy diagnosis.

The functional prediction showed that environmental information processing, such as CAMs and ECM-receptor interactions, was significantly higher in HGG patients than in the HCs. Several CAMs, such as neural cell adhesion molecule L1, have been identified to underlie the occurrence of glioma malignancies (Senner et al., 2002; Jiang et al., 2019; Lyu et al., 2021). ECM plays an important role in gliomas, such as in the higher expression of laminin $\alpha 2$ in glioblastoma (Lathia et al., 2012). The key role of the focal adhesion pathway and the level of actin in the cytoskeleton during the migration and invasion of glioblastoma have also been reported (de Semir et al., 2020; Wu et al., 2020).

Our results demonstrated that the genera *Bergeyella* and *Capnocytophaga* were significantly positively correlated with the glioma IDH-mutant, which is consistent with the results of glioma malignancy. The IDH1-mutant plays an important role in glioma cell glucose induction, glutamine metabolism, lipid synthesis, and cell redox regulation (Hartmann et al., 2009; Maus and Peters, 2017; Stuni et al., 2018). Moreover, metabolism deregulation plays an important role in cell growth, proliferation, angiogenesis, and invasion; thus, it has been considered one of the emerging hallmarks of cancer cells (Hanahan and Weinberg, 2011). Recent studies have found that lipid metabolism reprogramming plays a crucial role in the progression of cancer cells, such as in membrane synthesis, energetic production, and signal transduction (Liu et al., 2017). The activation of the AMPK signaling pathway contributes to the anti-inflammatory microenvironment of IDH1-mutated gliomas and thus causes better prognoses in patients with an IDH1-mutated glioma (Han et al., 2019). Our results demonstrated that the malignancy inverse-related microbial gene functions involving lipid metabolism and

AMPK signaling pathway were significantly enriched in the IDH-mutant group, suggesting that changes in oral microbial gene functions may underlie the link between the positive association between IDH-mutant gliomas and better prognosis. Moreover, several studies reported that mutant IDH1 in gliomas regulated a number of physiological processes such as inflammatory pathways, metabolic metabolism, hypoxia sensing, histone demethylation, and changes in DNA methylation causing DNA strand breaks, apoptosis, autophagy, and tumor cell death (Gilbert et al., 2014; Viswanath et al., 2018; Zhang et al., 2019; Kadiyala et al., 2021; Pirozzi and Yan, 2021). Therefore, the possible mechanism underlying the association between oral microbiome and IDH1 mutation is that IDH1 mutation specifically selects some oral microbiota, which can produce specific metabolites involved in lipid metabolism and AMPK signaling pathway to regulate intracellular energy homeostasis, increase brain glioma cell apoptosis and autophagy, prevent brain glioma cell proliferation, and contribute to the formation of an anti-inflammatory tumor microenvironment in the brain, and further causes better prognoses in patients with an IDH1-mutated glioma. Certainly, animal and cell experiments are further needed to determine the causality of IDH1 mutation on the oral microbiome under glioma status.

The present study has several strengths. First, to the best of our knowledge, this is the first study examining the role of oral microbiota in glioma malignancy. Second, we developed a novel bacterial marker panel to discriminate HGG patients from LGG patients and HCs. Third, our study revealed that the composition and gene function of oral microbiota were significantly associated with the IDH1 mutation in glioma, which can be used to predict the prognosis of glioma patients. The present study also has several limitations. First, the sample size was relatively small. This study was a small-sample, single-center study because of the challenges in recruiting this type of cohort and the strict inclusion criteria. However, this also guaranteed the consistency of the sample. Second, although we demonstrated that oral microbiota was associated with glioma malignancy and the IDH mutation, the underlying causality remains unclear. Finally, no plaque or tongue-coating specimens were included because of the difficulty in collecting such samples.

In summary, the present study indicated that oral microbiota composition and gene functions are significantly associated with glioma malignancy and the IDH1 mutation. We also discovered a microbial biomarker panel to distinguish HGG patients from HCs. Our results suggest that oral microbiota may be an important preventive target to mitigate glioma malignancy and achieve better prognoses for glioma patients.

DATA AVAILABILITY STATEMENT

The data presented in the study are deposited in the National Library of Medicine repository (<https://submit.ncbi.nlm.nih.gov/>), BioProject accession number: PRJNA750937 (<https://www.ncbi.nlm.nih.gov/bioproject/PRJNA750937>).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics committee of the Affiliated Hospital of Southwest Medical University (No. KY2019030). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

YW, LC, and JZ conceived the study and designed the experiments. YW, LF, and HW analyzed and interpreted the data. YW, LF, HW, and ZJ generated the figure, drafted the manuscript, and contributed to critical revision of the manuscript. HZ, QL, WZ, and MW visited patients, collected the data, and critical revision of the manuscript. YL and XL collected the samples and critical revision of the manuscript. YW, LF, and HW revised the manuscript with input from JZ and ZJ. LC and JZ obtained funding and contributed to study supervision. All authors contributed to the article and approved the submitted version.

FUNDING

The research was supported by National Natural Science Foundation project, Grant No. 00022986, and technology projects of Sichuan Province and Grant No. 2018JY0403. The research was further supported by Medical Research Fund for Young Scholars of the Sichuan Medical Association, Grant No. Q16076 and Natural Science Foundation of Southwest Medical University, Grant Nos. 2016XNYD217, 2018-ZRQN-032, and 2016LZXNYD-G03.

ACKNOWLEDGMENTS

We acknowledge the Shanghai Majorbio Bio-Pharm Technology Co., Ltd. for assistance with MiSeq sequencing.

REFERENCES

- Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* 37, 852–857. doi: 10.1038/s41587-019-0209-9
- Brat, D. J., Aldape, K., Colman, H., Figarella-Branger, D., Fuller, G. N., Giannini, C., et al. (2020). cIMPACT-NOW update 5: recommended grading criteria and terminologies for IDH-mutant astrocytomas. *Acta Neuropathol.* 139, 603–608. doi: 10.1007/s00401-020-02127-9
- Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J., and Holmes, S. P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13, 581–583. doi: 10.1038/nmeth.3869
- Ceccarelli, M., Barthel, F. P., Malta, T. M., Sabedot, T. S., Salama, S. R., Murray, B. A., et al. (2016). Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse Glioma. *Cell* 164, 550–563. doi: 10.1016/j.cell.2015.12.028
- de Semir, D., Bezrookove, V., Nosrati, M., Scanlon, K. R., Singer, E., Judkins, J., et al. (2020). Phip drives glioblastoma motility and invasion by regulating

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmicb.2021.746568/full#supplementary-material>

Supplementary Figure 1 | Study flow-chart. A total of 59 participants from the cross-sectional study (HCs=24, LGG=12, HGG=24) with 16S oral microbiome profiling were included in the study. *1 patient with LGG was pathologically diagnosed with non-glioma. **1 patient with HGG was not received surgery. Abbreviations: HCs, healthy controls; LG, low glioma group; HGG, high-grade glioma.

Supplementary Figure 2 | Bacteria taxonomic profiling at the phylum level of oral microbiome from healthy controls and glioma patients. Box plots show the relative abundance of Firmicutes, Bacteroidetes, Proteobacteria, Actinobacteria, Fusobacteria, and Spirochaetota in HCs, LGG, and HGG. Each box plot represents the median, interquartile range, minimum, and maximum values. The *p* value was calculated by non-parametric Mann-Whitney U test. Value of *p*<0.05 indicated the statistical significance. HCs, healthy controls; HGG: high-grade glioma; LGG: low-grade glioma.

Supplementary Figure 3 | Relative abundance of different bacteria genus in healthy and glioma patients. Bacteria taxonomic profiling at the genus level of oral microbiome from healthy controls and glioma patients. (A) The 18 bacteria at the genus level in the oral microbiome, “Others” represents the bacteria with the relative abundance of less than 1%. (B) The proportion of Streptococcus, Prevotella, Porphyromonas, Veillonella, Haemophilus, Actinomyces, Neisseria, Porphyromonas, and Granulicatella among HGG and HC groups at the genus level; (C) the box plots show the relative abundance of Leptotrichiaceae in HCs, LGG, and HGG. Each box plot represents the median, interquartile range, minimum, and maximum values. The *p* value was calculated by non-parametric Mann-Whitney U test. *p* value <0.05 indicated the statistical significance. HCs, healthy controls; LGG: low-grade glioma; HGG: high-grade glioma.

Supplementary Figure 4 | Microbial functions altered in the IDH1-mutant group and IDH1-wild-type group. Heat map showing the medial abundance of all significant modules as determined by PICRUSt analysis at IDH1-mutant group and IDH1-wild-type group. Notes: * *p*<0.05, ** *p*<0.01, *** *p*<0.001. Abbreviations: PICRUSt, Phylogenetic Investigation of Communities by Reconstruction of Unobserved States; HCs, healthy controls; HGG, high-grade glioma. PICRUSt, Phylogenetic Investigation of Communities by Reconstruction of Unobserved States; IDH1, isocitrate dehydrogenase 1; HCs, healthy controls; HGG: high-grade glioma.

- the focal adhesion complex. *Proc. Natl. Acad. Sci. U. S. A.* 117, 9064–9073. doi: 10.1073/pnas.1914505117
- Dominy, S. S., Lynch, C., Ermini, F., Benedyk, M., Marczyk, A., Konradi, A., et al. (2019). Porphyromonas gingivalis in Alzheimer’s disease brains: evidence for disease causation and treatment with small-molecule inhibitors. *Sci. Adv.* 5:eaa03333. doi: 10.1126/sciadv.aau3333
- Eribe, E. R. K., and Olsen, I. (2017). Leptotrichia species in human infections II. *J. Oral Microbiol.* 9:1368848. doi: 10.1080/20002297.2017.1368848
- Fan, X., Alekseyenko, A. V., Wu, J., Peters, B. A., Jacobs, E. J., Gapstur, S. M., et al. (2018). Human oral microbiome and prospective risk for pancreatic cancer: a population-based nested case-control study. *Gut* 67, 120–127. doi: 10.1136/gutjnl-2016-312580
- Gilbert, M. R., Liu, Y., Neltner, J., Pu, H., Morris, A., Sunkara, M., et al. (2014). Autophagy and oxidative stress in gliomas with IDH1 mutations. *Acta Neuropathol.* 127, 221–233. doi: 10.1007/s00401-013-1194-6
- Han, C. J., Zheng, J. Y., Sun, L., Yang, H. C., Cao, Z. Q., Zhang, X. H., et al. (2019). The oncometabolite 2-hydroxyglutarate inhibits microglial activation via the AMPK/mTOR/NF-kappaB pathway. *Acta Pharmacol. Sin.* 40, 1292–1302. doi: 10.1038/s41401-019-0225-9

- Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell* 144, 646–674. doi: 10.1016/j.cell.2011.02.013
- Hartmann, C., Meyer, J., Balss, J., Capper, D., Mueller, W., Christians, A., et al. (2009). Type and frequency of IDH1 and IDH2 mutations are related to astrocytic and oligodendroglial differentiation and age: a study of 1,010 diffuse gliomas. *Acta Neuropathol.* 118, 469–474. doi: 10.1007/s00401-009-0561-9
- Hayes, R. B., Ahn, J., Fan, X., Peters, B. A., Ma, Y., Yang, L., et al. (2018). Association of oral microbiome with risk for incident head and neck squamous cell cancer. *JAMA Oncol.* 4, 358–365. doi: 10.1001/jamaoncol.2017.4777
- Hou, W., Li, J., Cao, Z., Lin, S., Pan, C., Pang, Y., et al. (2021). Decorating bacteria with a therapeutic nanocoating for synergistically enhanced biotherapy. *Small* 17:e2101810. doi: 10.1002/smll.202101810
- Inman, K. S., Francis, A. A., and Murray, N. R. (2014). Complex role for the immune system in initiation and progression of pancreatic cancer. *World J. Gastroenterol.* 20, 11160–11181. doi: 10.3748/wjg.v20.i32.11160
- Jiang, Q., Xie, Q., Hu, C., Yang, Z., Huang, P., Shen, H., et al. (2019). Glioma malignancy is linked to interdependent and inverse AMOG and L1 adhesion molecule expression. *BMC Cancer* 19:911. doi: 10.1186/s12885-019-6091-5
- Jing, Y., Yu, Y., Bai, F., Wang, L., Yang, D., Zhang, C., et al. (2021). Effect of fecal microbiota transplantation on neurological restoration in a spinal cord injury mouse model: involvement of brain-gut axis. *Microbiome* 9:59. doi: 10.1186/s40168-021-01007-y
- Kadiyala, P., Carney, S. V., Gauss, J. C., Garcia-Fabiani, M. B., Haase, S., Alghamri, M. S., et al. (2021). Inhibition of 2-hydroxyglutarate elicits metabolic reprogramming and mutant IDH1 glioma immunity in mice. *J. Clin. Invest.* 131:e139542. doi: 10.1172/JCI139542
- Kurkivuori, J., Salaspuro, V., Kaihovaara, P., Kari, K., Rautemaa, R., Gronroos, L., et al. (2007). Acetaldehyde production from ethanol by oral streptococci. *Oral Oncol.* 43, 181–186. doi: 10.1016/j.oraloncology.2006.02.005
- Lathia, J. D., Li, M., Hall, P. E., Gallagher, J., Hale, J. S., Wu, Q., et al. (2012). Laminin alpha 2 enables glioblastoma stem cell growth. *Ann. Neurol.* 72, 766–778. doi: 10.1002/ana.23674
- Li, Y., Tan, X., Zhao, X., Xu, Z., Dai, W., Duan, W., et al. (2020). Composition and function of oral microbiota between gingival squamous cell carcinoma and periodontitis. *Oral Oncol.* 107:104710. doi: 10.1016/j.oraloncology.2020.104710
- Liu, Q., Luo, Q., Halim, A., and Song, G. (2017). Targeting lipid metabolism of cancer cells: a promising therapeutic strategy for cancer. *Cancer Lett.* 401, 39–45. doi: 10.1016/j.canlet.2017.05.002
- Lopez, E., Raymond, J., Patkai, J., El Ayoubi, M., Schmitz, T., Moriette, G., et al. (2010). Capnocytophaga species and preterm birth: case series and review of the literature. *Clin. Microbiol. Infect.* 16, 1539–1543. doi: 10.1111/j.1469-0691.2010.03151.x
- Louis, D. N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W. K., et al. (2016). The 2016 World Health Organization classification of tumors of the central nervous system: a summary. *Acta Neuropathol.* 131, 803–820. doi: 10.1007/s00401-016-1545-1
- Louis, D. N., Wesseling, P., Aldape, K., Brat, D. J., Capper, D., Cree, I. A., et al. (2020). cIMPACT-NOW update 6: new entity and diagnostic principle recommendations of the cIMPACT-Utrecht meeting on future CNS tumor classification and grading. *Brain Pathol.* 30, 844–856. doi: 10.1111/bpa.12832
- Lu, H., Ren, Z., Li, A., Li, J., Xu, S., Zhang, H., et al. (2019). Tongue coating microbiome data distinguish patients with pancreatic head cancer from healthy controls. *J. Oral Microbiol.* 11:1563409. doi: 10.1080/20002297.2018.1563409
- Lyu, Y., Yang, H., and Chen, L. (2021). Metabolic regulation on the immune environment of glioma through gut microbiota. *Semin. Cancer Biol.* doi: 10.1016/j.semcancer.2021.05.005 [Epub ahead of print]
- Maus, A., and Peters, G. J. (2017). Glutamate and alpha-ketoglutarate: key players in glioma metabolism. *Amino Acids* 49, 21–32. doi: 10.1007/s00726-016-2342-9
- Mei, Q. X., Huang, C. L., Luo, S. Z., Zhang, X. M., Zeng, Y., and Lu, Y. Y. (2018). Characterization of the duodenal bacterial microbiota in patients with pancreatic head cancer vs. healthy controls. *Pancreatol.* 18, 438–445. doi: 10.1016/j.pan.2018.03.005
- Meurman, J. H., and Uttamo, J. (2008). Oral micro-organisms in the etiology of cancer. *Acta Odontol. Scand.* 66, 321–326. doi: 10.1080/00016350802446527
- Michaud, D. S., Izard, J., Wilhelm-Benartzi, C. S., You, D. H., Grote, V. A., Tjonneland, A., et al. (2013). Plasma antibodies to oral bacteria and risk of pancreatic cancer in a large European prospective cohort study. *Gut* 62, 1764–1770. doi: 10.1136/gutjnl-2012-303006
- Narengaowa, Kong, W., Lan, F., Awan, U. F., Qing, H., and Ni, J. (2021). The Oral-gut-brain AXIS: the influence of microbes in Alzheimer's disease. *Front. Cell. Neurosci.* 15:633735. doi: 10.3389/fncel.2021.633735
- Pandya, G., Kirtonia, A., Singh, A., Goel, A., Mohan, C. D., Rangappa, K. S., et al. (2021). A comprehensive review of the multifaceted role of the microbiota in human pancreatic carcinoma. *Semin. Cancer Biol.* doi: 10.1016/j.semcancer.2021.05.027 [Epub ahead of print]
- Park, S. Y., Hwang, B. O., Lim, M., Ok, S. H., Lee, S. K., Chun, K. S., et al. (2021). Oral-gut microbiome axis in gastrointestinal disease and cancer. *Cancers* 13:2124. doi: 10.3390/cancers13092124
- Pirozzi, C. J., and Yan, H. (2021). The implications of IDH mutations for cancer development and therapy. *Nat. Rev. Clin. Oncol.* 18, 645–661. doi: 10.1038/s41571-021-00521-0
- Senner, V., Kismann, E., Puttmann, S., Hoess, N., Baur, I., and Paulus, W. (2002). L1 expressed by glioma cells promotes adhesion but not migration. *Glia* 38, 146–154. doi: 10.1002/glia.10058
- Stuani, L., Riols, F., Millard, P., Sabatier, M., Batut, A., Saland, E., et al. (2018). Stable isotope labeling highlights enhanced fatty acid and lipid metabolism in human acute myeloid Leukemia. *Int. J. Mol. Sci.* 19:3325. doi: 10.3390/ijms19113325
- Thirumala, R., Rappo, U., Babady, N. E., Kamboj, M., and Chawla, M. (2012). Capnocytophaga lung abscess in a patient with metastatic neuroendocrine tumor. *J. Clin. Microbiol.* 50, 204–207. doi: 10.1128/JCM.05306-11
- Viswanath, P., Radoul, M., Izquierdo-Garcia, J. L., Ong, W. Q., Luchman, H. A., Cairncross, J. G., et al. (2018). 2-Hydroxyglutarate-mediated autophagy of the endoplasmic reticulum leads to an unusual downregulation of phospholipid biosynthesis in mutant IDH1 Gliomas. *Cancer Res.* 78, 2290–2304. doi: 10.1158/0008-5472.CAN-17-2926
- Vyhnalova, T., Danek, Z., Gachova, D., and Linhartova, P. B. (2021). The role of the oral microbiota in the etiopathogenesis of oral squamous cell carcinoma. *Microorganisms* 9:1549. doi: 10.3390/microorganisms9081549
- Wang, Y., Tong, Q., Ma, S. R., Zhao, Z. X., Pan, L. B., Cong, L., et al. (2021). Oral berberine improves brain dopa/dopamine levels to ameliorate Parkinson's disease by regulating gut microbiota. *Signal Transduct. Target. Ther.* 6:77. doi: 10.1038/s41392-020-00456-5
- Wu, S., Qiao, Q., and Li, G. (2020). A radiosensitivity gene signature and XPO1 predict clinical outcomes for Glioma patients. *Front. Oncol.* 10:871. doi: 10.3389/fonc.2020.00871
- Yang, C. Y., Yeh, Y. M., Yu, H. Y., Chin, C. Y., Hsu, C. W., Liu, H., et al. (2018). Oral microbiota community dynamics associated with oral squamous cell carcinoma staging. *Front. Microbiol.* 9:862. doi: 10.3389/fmicb.2018.00862
- Zhang, Y., Pusch, S., Innes, J., Sidlauskas, K., Ellis, M., Lau, J., et al. (2019). Mutant IDH sensitizes Gliomas to endoplasmic reticulum stress and triggers apoptosis via miR-183-mediated inhibition of Semaphorin 3E. *Cancer Res.* 79, 4994–5007. doi: 10.1158/0008-5472.CAN-19-0054

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.

Copyright © 2021 Wen, Feng, Wang, Zhou, Li, Zhang, Wang, Li, Luan, Jiang, Chen and Zhou. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Deep Learning Driven Drug Discovery: Tackling Severe Acute Respiratory Syndrome Coronavirus 2

Yang Zhang^{1*}, Taoyu Ye¹, Hui Xi¹, Mario Juhas² and Junyi Li^{3*}

¹ College of Science, Harbin Institute of Technology (Shenzhen), Shenzhen, China, ² Medical and Molecular Microbiology Unit, Department of Medicine, Faculty of Science and Medicine, University of Fribourg, Fribourg, Switzerland, ³ School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen, China

OPEN ACCESS

Edited by:

Mohammad-Hossein Sarrafzadeh,
University of Tehran, Iran

Reviewed by:

Hashem Asgharnejad,
Polytechnique Montréal, Canada
Tanmay Majumdar,
National Institute of Immunology (NII),
India

*Correspondence:

Yang Zhang
zhangyang07@hit.edu.cn
Junyi Li
lijunyi@hit.edu.cn

Specialty section:

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

Received: 11 July 2021

Accepted: 30 September 2021

Published: 28 October 2021

Citation:

Zhang Y, Ye T, Xi H, Juhas M and
Li J (2021) Deep Learning Driven Drug
Discovery: Tackling Severe Acute
Respiratory Syndrome Coronavirus 2.
Front. Microbiol. 12:739684.
doi: 10.3389/fmicb.2021.739684

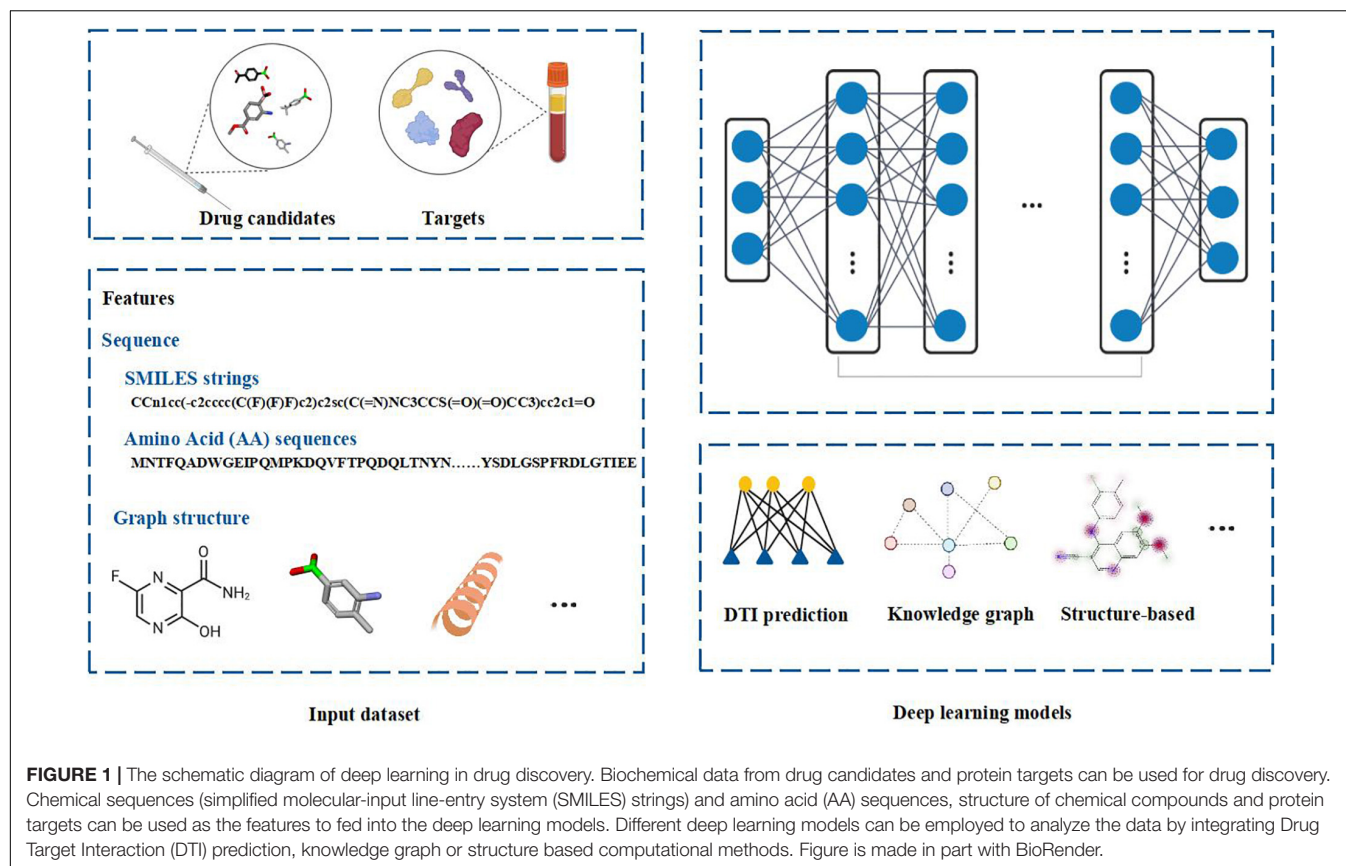
Deep learning significantly accelerates the drug discovery process, and contributes to global efforts to stop the spread of infectious diseases. Besides enhancing the efficiency of screening of antimicrobial compounds against a broad spectrum of pathogens, deep learning has also the potential to efficiently and reliably identify drug candidates against Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). Consequently, deep learning has been successfully used for the identification of a number of potential drugs against SARS-CoV-2, including Atazanavir, Remdesivir, Kaletra, Enalaprilat, Venetoclax, Posaconazole, Daclatasvir, Ombitasvir, Toremifene, Niclosamide, Dexamethasone, Indomethacin, Pralatrexate, Azithromycin, Palmatine, and Saquinavir. This mini-review discusses recent advances and future perspectives of deep learning-based SARS-CoV-2 drug discovery.

Keywords: deep learning, database, drug discovery, antibiotics, antimalarial drug, drug repurposing, SARS-CoV-2

INTRODUCTION

Deep learning is a branch of machine learning. It is an algorithm that abstracts data by using multiple processing layers composed of complex structures or multiple non-linear transformations. Compared with the shallow machine learning methods, deep learning algorithm is a process of automatic feature engineering. Deep learning frameworks, such as convolutional neural network and recursive neural network, have been applied in the fields of bioinformatics and biomedicine and achieved excellent results (Lipinski et al., 2019). Deep learning methods have good applications in microbiology including metagenomic data analysis, microbial-related drug discovery, disease microbial association and so on (Duch et al., 2007). When analyzing microbial related data, it shows high prediction accuracy in practice. Because deep learning algorithms are good at obtaining very complex underlying patterns in data, they are especially suitable for large and high-dimensional data sets. Moreover, it is easy to update the model with the new data. The hidden layer of the network obviously reduces the demand for Feature Engineering and is conducive to the completion of prediction tasks. The schematic diagram of the deep learning in drug discovery is shown in Figure 1.

Deep learning has revolutionized most areas of science and technology, including drug discovery. Traditional drug discovery methods are not time and cost efficient and therefore often unable to keep pace with the rapidly emerging and re-emerging pathogenic microorganisms. More recent drug discovery methods include Naive Bayesian, Support Vector Machines and Neural Networks (Bender et al., 2007; Stephenson et al., 2019). These alternative drug discovery



methods usually use bigger data sets generated from high throughput screenings and allow more accurate prediction of bioactivities and molecular properties of the targets (Stephenson et al., 2019). Compared to these alternative machine learning methods used for drug discovery, deep learning is characterized by the flexibility of the architecture of Neural Networks (Chen et al., 2018). Given the cost and time required for traditional drug discovery, deep learning has the potential to significantly accelerate the drug discovery process. By using information on the biological, chemical, and topological properties of compounds and their putative targets from the large-scale libraries, deep learning can be employed to identify the most promising drugs against specific diseases (Neves et al., 2020; Stokes et al., 2020). Various deep learning methods have been developed over the last few years, but their application in drug discovery has still not reached its full potential. One of the main hurdles for researchers planning to build their own deep learning model for drug identification is the amount of resources and time required to collect large amounts of data. A number of computational screen open databases have been made to prioritize drug candidates, recently. A representative set of open access datasets which can be used to train deep learning models for specific research projects is shown in Table 1.

The outbreak of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) causing coronavirus disease (COVID-19) has been declared a global pandemic. By September 2021, more than 220 million people have been infected with

SARS-CoV-2 and more than 4.5 million of those infected have died. In addition, several SARS-CoV-2 variants with mutations that increase their potential to contribute to the severity of the pandemic have emerged and are spreading around the globe (Zhang Y. et al., 2020).

Besides non-structural proteins, SARS-CoV-2 genome encodes four structural proteins: envelope (E), membrane (M), nucleocapsid (N), and spike (S) (Zhang Y. et al., 2020). S protein mediates entry of SARS-CoV-2 into the host cells by binding and fusing with the host's cellular receptor, the angiotensin-converting enzyme 2 (ACE2). Mutations in S protein, particularly in its receptor binding domain (RBD) were shown to play a role in the increased transmissibility and infectivity of the emerging SARS-CoV-2 variants (Zahradnik et al., 2021).

Although several SARS-CoV-2 vaccines have been developed over the last few months, there are not many efficient and reliable drugs available for the treatment of SARS-CoV-2 infections. This is caused partially by the fact that the traditional drug discovery process may be time-consuming and costly to keep pace with the rapid spread of SARS-CoV-2 and its variants with increased transmissibility and other enhanced features.

Deep learning has been previously applied for the identification of a number of antiviral compounds, including antiviral peptides (Timmons and Hewage, 2021) and small drug-like compounds with the potential to inhibit HIV-1 (Andrianov et al., 2021).

The computational approaches employing deep learning will aid also faster discovery of novel and active potential inhibition agents against SARS-CoV-2 and its emerging variants.

High-throughput technologies have generated an increasing amount of data in chemoinformatics. As a result, it is believed that the application of the recent deep learning advances into the drug discovery process will lead to novel, more reliable and efficient therapeutics against SARS-CoV-2.

ANTIMICROBIAL DRUGS IDENTIFIED BY DEEP LEARNING

Deep learning can reduce time and costs of the drug discovery process, particularly in its early stages. Consequently, deep learning-based approaches have been successfully used to identify novel antimicrobial compounds against a wide variety of pathogenic microorganisms, including bacteria, protozoan parasites and viruses.

Training of the deep learning model to identify molecules active against antibiotic-resistant bacteria led to the discovery of Halicin and eight additional potential antibiotics from the ZINC database (Table 1; Stokes et al., 2020). Interestingly, these compounds identified by deep learning are all structurally divergent from conventional antibiotics (Stokes et al., 2020). Subsequent tests revealed strong antibacterial activity of Halicin against a number of antibiotic-resistant bacteria, including Carbapenemase-producing Enterobacterales, *Mycobacterium tuberculosis*, *Acinetobacter baumannii*, and *Clostridioides difficile* (Stokes et al., 2020).

In parasite research, deep learning has been applied to predict new antimalarial drug candidates. Neves et al. (2020) employed deep learning to obtain binary, continuous Quantitative Structure-Activity Relationships (QSAR) models using datasets extracted from ChEMBL database (Table 1). QSAR mathematical models can predict the relationship between the structure of a molecule and biological activity or physicochemical property. This study led to the discovery of two new families of the potential next generation antimalarial drugs with activity against *Plasmodium* causing malaria at nanomolar concentrations and low cytotoxicity in mammalian cells (Neves et al., 2020).

Deep learning has been also applied for the identification of a number of antiviral compounds. Timmons and Hewage developed a novel method called ENNAVIA, which employs deep learning and chemoinformatics, to identify peptides with low toxicity and excellent biological activity. The peptides identified in this study represent promising candidates for antiviral drugs (Timmons and Hewage, 2021). Furthermore, deep learning in combination with molecular modeling has been applied for the identification of three small drug-like compounds from millions of molecules in the ZINC15 database (Andrianov et al., 2021). Based on machine learning, molecular docking, molecular dynamics and quantum chemical calculations, the compounds identified in this study are promising HIV-1 drugs with the potential to block CD4-binding site of the viral envelope protein,

TABLE 1 | Representative biochemical datasets used in deep learning studies.

Dataset	Description	URL	References
ZINC	ZINC database contains over 230 million compounds.	http://zinc.docking.org/	Bai et al., 2020; Choi et al., 2020; Stokes et al., 2020; Ton et al., 2020
ChEMBL	ChEMBL (version 27) chemical database contains over 1.9 million specific compounds.	https://www.ebi.ac.uk/chembl/	Stokes et al., 2020
Drug target commons (DTC)	DTC crowdsourcing database contains 204,901 annotated bioactivity data points among 4,276 compounds and 1,007 specific protein targets.	https://drugtargetcommons.fimm.fi/	Beck et al., 2020
BindingDB	BindingDB database of measured binding affinities contains 2,061,017 binding data for 8,160 protein targets and 907,259 small molecules.	http://www.bindingdb.org/bind/index.jsp	Beck et al., 2020
DrugBank	DrugBank pharmaceutical database contains detailed molecular information about drugs, their mechanisms, interactions and targets.	https://go.drugbank.com/releases/latest	Choi et al., 2020; Zeng et al., 2020
PDBbind	PDBbind database provides binding data of 21,382 biomolecular complexes, including protein-ligand (17,679), nucleic acid-ligand (136), protein-nucleic acid (973), and protein-protein complexes (2,594).	http://www.pdbbind.org.cn	Bai et al., 2020

thus inhibiting HIV-1 entry (Andrianov et al., 2021). Li et al. have developed a dual-channel deep neural network for identifying variable-length antiviral peptides (DeepAVP) which could block entry of the virus into the host cell (Li et al., 2020). Deep learning has been also used for the prediction of plant-exclusive virus-derived small interfering RNAs (PVsiRNAPred) (He et al., 2019).

DEEP LEARNING IN TACKLING SEVERE ACUTE RESPIRATORY SYNDROME CORONAVIRUS 2

Reliable and efficient computing methods employing deep learning are urgently needed for the discovery of drugs against SARS-CoV-2 and its emerging variants.

Drug repurposing is considered to be among the fastest and most promising methods for identification of effective SARS-CoV-2 treatments. A good example of the drug repurposing involving deep learning is a recent work by Zhang Y. et al. (2020). This study employed a deep learning-based drug-target interaction model called Molecule Transformer-Drug Target Interaction (MT-DTI) utilizing chemical sequences [simplified molecular-input line-entry system (SMILES) strings] and amino acid (AA) sequences as the input (**Figure 1**). MT-DTI model was trained with a combined and curated chemical-protein pairs from BindingDB and Drug Target Commons (DTC) databases (**Table 1**). This study led to identification of several commercially available antiviral drugs with the potential to interact also with the SARS-CoV-2 proteins (Beck et al., 2020). Subsequent experiments showed that several of the antiviral agents identified by MT-DTI model could be potentially used to treat SARS-CoV-2 (Beck et al., 2020). These include Atazanavir (Kd 94.94 nM), Remdesivir (Kd 113.13 nM), and Kaletra (Lopinavir/Ritonavir) (**Table 2**). Atazanavir, showing an inhibitory potency against SARS-CoV-2 3-C like proteinase is an antiviral drug used for the treatment of the human immunodeficiency virus (HIV) infections. Remdesivir has been previously predicted to act against SARS-CoV-2. Furthermore, Lopinavir and Ritonavir were shown to target viral proteinases (Beck et al., 2020). The BindingDB is a public database containing measured binding affinities for three types of coronaviruses, SARS-CoV-2, SARS-CoV and MERS-CoV.¹

MT-DTI model was also used to select compounds from 1,400 approved drugs in DrugBank and ZINC databases (**Table 1**) with strong affinity to the host cell targets crucial for viral infection (Zahradnik et al., 2021). This approach led to identification of drugs candidates with a strong binding affinity (Kd < 100 nM) against ACE2 receptor and transmembrane protease serine 2 (TMPRSS2) (Zahradnik et al., 2021). Drug candidates identified in this study include an ACE2 inhibitor Enalaprilat (Kd 1.46 nM) and several drugs with predicted strong affinity for TMPRSS2, namely Venetoclax (Kd 6.12 nM), Posaconazole (Kd 17.11 nM), Daclatasvir (Kd 6.65 nM), and Ombitasvir (Kd 5.91 nM) (**Table 2**). Strong affinity of Enalaprilat for ACE2 suggests that it might prevent the entry of SARS-CoV-2 to human cells. Notably, two of the drug candidates identified, namely Daclatasvir and Ombitasvir, are known Hepatitis C virus (HCV) inhibitors, thus suggesting that they may act against both HCV and SARS-CoV-2 (Zahradnik et al., 2021). The DrugBank has collected data for 65 drugs against 385 drug targets, which is web accessible at <https://go.drugbank.com/covid-19>.

Zeng et al. used deep learning-based knowledge graph to select promising SARS-CoV-2 drug candidates (Zeng et al., 2020). Knowledge graph in this study encompasses 15 million edges across 39 types of relationships connecting expression patterns, genes, pathways, drugs and diseases and incorporates data from over 20 million PubMed articles and the DrugBank database (**Table 1**). Deep learning employed to learn the representation of nodes and relationships in this knowledge graph led to identification of 41 promising drug candidates, including

Toremifene, Niclosamide, Dexamethasone and Indomethacin (**Table 2**; Beck et al., 2020). Toremifene is a selective estrogen receptor modulator, which has shown antiviral activity against a number of viruses, including SARS-CoV-2. Dexamethasone is an anti-inflammatory agent with the potential to treat SARS-CoV-2 infections (Beck et al., 2020). Niclosamide, a drug used to treat tapeworm and an anti-inflammatory drug Indomethacin were also shown to have antiviral activity *in vitro*. The 41 promising drug candidates identified in this study (including Toremifene, Niclosamide, Dexamethasone and Indomethacin) were also validated by gene expression and proteomics of cells infected with SARS-CoV-2 (Beck et al., 2020).

A hybrid deep learning and molecular simulation-based screening procedure was used to select drug candidates targeting RNA-dependent RNA polymerase (RdRp) from 1906 approved drugs, recently (Choi et al., 2020). Commercially available drug candidates, Pralatrexate and Azithromycin, (**Table 2**) identified in this study were confirmed to inhibit SARS-CoV-2 replication *in vitro* (Choi et al., 2020). While Pralatrexate was shown to act after entry of the virus into the cells, Azithromycin was active at both the entry and post-entry of SARS-CoV-2 into the host cells (Choi et al., 2020).

Bai et al. developed MolAICal software tool combining deep learning model and classical algorithm for identification of drugs interacting with 3D pocket of protein targets (Bai et al., 2020). Deep learning model of MolAICal software was trained using approved drug fragments in PDBbind database and drug-like molecules in the ZINC database (**Table 1**). Drug design functions of MolAICal software were demonstrated using the membrane protein glucagon receptor (GCGR) and SARS-CoV-2 main protease (Mpro) (Zeng et al., 2020).

Ton et al. (2020) developed a Deep Docking (DD) deep learning platform which uses QSAR models for screening of potential drug candidates in the ZINC database (**Table 1**). This approach led to the identification of 1,000 potential ligands for SARS-CoV-2 Mpro (Ton et al., 2020).

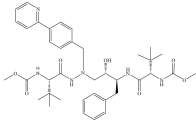
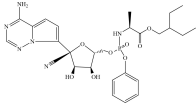
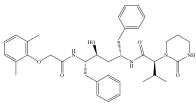
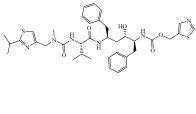
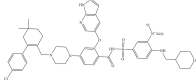
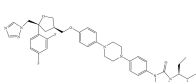
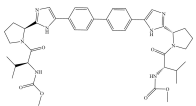
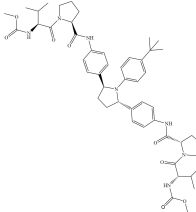
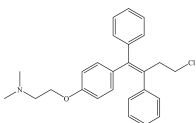
Deep learning and molecular docking methods were developed for screening of natural compounds against SARS-CoV-2 Mpro in the ChEMBL database (**Table 1**; Bai et al., 2020). ChEMBL database is an open large-scale chemical database of bioactive molecules, containing 8,200 potential anti-SARS-CoV-2 drug candidates. This study led to the identification of two natural compounds with potential as therapeutics against SARS-CoV-2, namely Palmatine (Kd 1096.4 nM) and Sauchinone (**Table 2**) (Kd 389.05 nM). Palmatine and Sauchinone are an alkaloid and a lignan, respectively, with previously shown pharmacological properties. Furthermore, both Palmatine and Sauchinone form a stable complex with SARS-CoV-2 Mpro and have been predicted to inhibit SARS-CoV-2 (Bai et al., 2020).

Deep learning combined with multiple sequence alignment drug-likeness screening, molecular docking, chemical space mapping and molecular dynamics simulation was also used to identify drug candidates by screening 1528 anti-HIV-1 compounds against 3-chymotrypsin-like cysteine protease (3CLpro) of SARS-CoV-2 (Nand et al., 2020).

Given the lack of therapeutics against SARS-CoV-2, deep learning approaches combined with other computational

¹<https://www.bindingdb.org/bind/Covid19.jsp>

TABLE 2 | Drug candidates against SARS-CoV-2.

Drug	Molecular formula	Structural formula	SMILES	Target	References
Atazanavir	C ₃₈ H ₅₂ N ₆ O ₇		<chem>COC([N]([C@@H])(C(C)(O)O)C(NN(C([C@H])(O)[C@H](CC1 = CC = CC = C1)NC([C@H](C(C)(C)C)NC(OC) = O) = O)CC(C = C2) = CC = C2C3 = NC = CC = C3) = O) = O</chem>	3C-like proteinase	Beck et al., 2020
Remdesivir	C ₂₇ H ₃₅ N ₆ O ₈ P		<chem>CP = O.COC1 = CC = CC = C1.O = C(OCC(C)C)[C@H](C)NC.NC2 = NC = NN3C2 = CC = C3[C@@@](C#N)[C@H](C)[C@H](O)[C@@H](C(O))O4.[OH].[OH]</chem>	3C-like proteinase	Beck et al., 2020
Kaletra (Lopinavir/Ritonavir)	C ₇₄ H ₉₆ N ₁₀ O ₁₀ S ₂ (C ₃₇ H ₄₈ N ₄ O ₅ /C ₃₇ H ₄₈ N ₆ O ₅ S ₂)		<chem>CC1 = CC = CC(C) = C1OCC([N]([C@@H])(CC2 = CC = CC = C2)[C@@H](O)C[C@H](CC3 = CC = CC = C3)NC([C@H](C(C)C)N4C(NCCC4) = O) = O) = O</chem>	Helicase	Beck et al., 2020
Enalaprilat	C ₁₈ H ₂₄ N ₂ O ₅		<chem>CC(C)C1 = NC(CN(C)C([N]([C@@H](C(C)O)C([N]([C@@H])(CC2 = CC = CC = C2)[C@H](O)[C@H](CC3 = CC = CC = C3)NC(OCC4 = CN = CS4) = O) = O) = O) = CS1</chem>	ACE2	Choi et al., 2020
Venetoclax	C ₄₅ H ₅₀ ClN ₇ O ₇ S		<chem>C1C1 = CC = C(C2 = C(CN3CCN(C4 = CC = C(C(=NS(C5 = CC = C(NCC6CCOCC6)C([N +]([O-]) = O) = C5)(= O) = O) = O)C(OC7 = CC(C = CN8) = C8N = C7) = C4)CC3)CCC(C)(C)C2)C = C1</chem>	TMPRSS2 ACE2	Choi et al., 2020
Posaconazole	C ₃₇ H ₄₂ F ₂ N ₈ O ₄		<chem>FC1 = CC(F) = CC = C1[C@@@]2(CN3N = CN = C3)[C@H](COC4 = CC = C(N5CCN(C6 = CC = C(N7C = NN([C@@H](CC)[C@H](O)C)C7 = O)C = C6)CC5)C = C4)CO2</chem>	TMPRSS2 ACE2	Choi et al., 2020
Daclatasvir	C ₄₀ H ₅₀ N ₈ O ₆		<chem>COC([N]([C@@H])(C(C)C)C(N1[C@H](C2 = NC = C(C3 = CC = C(C4 = C(C = C(C5 = CN = C([C@H]6N(C([C@H](NC(OC) = O)C(C)C) = O)CCC6)N5)C = C4)C = C3)N2)CCC1) = O) = O</chem>	TMPRSS2 ACE2	Choi et al., 2020
Ombitasvir	C ₅₀ H ₆₇ N ₇ O ₈		<chem>CC(O)(C)C(C = C1) = CC = C1N([C@H](C2 = CC = C(NC([C@H]3CCCCN3C([C@H](C(C)C)NC(OC) = O) = O) = O)C = C2)CC4)[C@@H]4C5 = CC = C(NC([C@H]6N(C([C@H]7NC(O)C) = O)C(C)C) = O)CCC6) = O)C = C5</chem>	TMPRSS2 ACE2	Choi et al., 2020
Toremifene	C ₂₆ H ₂₈ ClNO		<chem>CN(C)CCOC1 = CC = C(C(C2 = CC = CC = C2) = C(C3 = CC = CC = C3)/CCC)C = C1</chem>	—	Zeng et al., 2020

(Continued)

TABLE 2 | (Continued)

Drug	Molecular formula	Structural formula	SMILES	Target	References
Niclosamide	C ₁₃ H ₈ Cl ₂ N ₂ O ₄		<chem>ClC1 = CC = C(O)C(C(=O)NC2=CC=C([N+](=O)[O-])=O)C2=CC(Cl)=O = C1</chem>	–	Zeng et al., 2020
Dexamethasone	C ₂₂ H ₂₉ FO ₅		<chem>O = C1C = C[C@@]2(C)C(C)[C@H]3C([C@@H](C)[C@H]3O)C(C(=O)O)C([H])([C@H]2)C(=O)C1</chem>	–	Zeng et al., 2020
Indomethacin	C ₁₉ H ₁₆ ClNO ₄		<chem>COC1 = CC = C(NC(=O)C2=CC=C(Cl)C(=O)C2)C1=CC(Cl)=O</chem>	–	Zeng et al., 2020
Pralatrexate	C ₂₃ H ₂₃ N ₇ O ₅		<chem>NC1 = C2C(N = CC(C(=O)O)C(=O)C2)C(=O)C1=CC(Cl)=O</chem>	RdRp	Zhang H. et al., 2020
Azithromycin	C ₃₈ H ₇₂ N ₂ O ₁₂		<chem>CN1[C@H](C)[C@@H](O)C(C(=O)O)C1=CC(Cl)=O</chem>	RdRp	Zhang H. et al., 2020
Palmitate	C ₂₁ H ₄₂ N ₄ O ₄ ⁺		<chem>COC1 = CC(=O)C(C(=O)O)C1=CC(Cl)=O</chem>	Mpro enzyme of SARS-CoV-2	Joshi et al., 2020
Saquinone	C ₂₀ H ₂₀ O ₆		<chem>O = C1C = C2[C@H]3(OCO2)[C@@H]4([H])[C@@H]1[C@H]3C(=O)C1=CC(Cl)=O</chem>	Mpro enzyme of SARS-CoV-2	Joshi et al., 2020

Table shows molecular and structural formulas, simplified molecular-input line-entry system (SMILES) strings and corresponding targets of the potential drugs against SARS-CoV-2.

methods will play an important role in the identification of potential drugs targeting SARS-CoV-2. Compounds selected by deep learning will subsequently undergo standard clinical evaluation.

DISCUSSION

Deep learning has a number of advantages compared to more conventional methods, including its ability to learn complex features independently. Although deep learning has played an important role in the identification of novel drugs against

a wide range of pathogens, including SARS-CoV-2, many challenges still remain.

The connection between the data fed into the deep learning model and the delivered output is inscrutable, which hidden inside is a so-called black box. Deep neural network due to its black-box nature therefore often lacks interpretability. Therefore, the interpretability of the future neural networks on the output results will be a key factor in understanding the logic of machine. This will aid analysis of the chemical compounds identified by deep learning and better design of the drug discovery studies.

Furthermore, the input data affects the prediction performance of the deep learning model. Consequently, a large,

standardized and reliable biochemical dataset is necessary to achieve better learning of the deep learning model. Development of a large open dataset in the future will enable potential standardization of the deep learning-based drug discovery.

Antibody-based therapy represents an interesting SARS-CoV-2 treatment option. Deep learning models have been developed for the discovery and design of therapeutic antibodies (Mason et al., 2021; Saka et al., 2021). Thus, drug repositioning and screening from computational libraries containing a massively diverse antibody sequences could be used to engineer anti-viral SARS-CoV-2 treatment.

Furthermore, most recent studies describe methodologies separately and test them individually. Application of deep learning to combine chemoinformatics with other types of data, such as imaging, cellular and molecular biology data for integrative analysis would be an important direction for future research. To this end, it might be necessary to identify the best neural network architecture for handling those vast troves of data.

We believe that integrative and systematic analysis will be important for future deep learning-based drug discovery that

involves complicated large biological, chemical and clinical datasets. Using such large datasets to streamline and accelerate drug discovery, deep learning will be crucial not only for the identification of drug candidates against SARS-CoV-2 but also against a broad spectrum of other emerging and reemerging pathogens.

AUTHOR CONTRIBUTIONS

All authors contribute to the writing and reviewing the manuscript.

FUNDING

This work was supported by the Natural Science Foundation of Shenzhen City (Project No. JCYJ20180306172131515) and the Shenzhen Science and Technology Program the university stable support program (20200821222112001).

REFERENCES

- Andrianov, A. M., Nikolaev, G. I., Shuldov, N. A., Bosko, I. P., Anischenko, A. I., and Tuzikov, A. V. (2021). Application of deep learning and molecular modeling to identify small drug-like compounds as potential HIV-1 entry inhibitors. *J. Biomol. Struct. Dyn.* 15, 1–19. doi: 10.1080/07391102.2021.1905559
- Bai, Q., Tan, S., Xu, T., Liu, H., Huang, J., and Yao, X. (2020). MolAICal: a soft tool for 3D drug design of protein targets by artificial intelligence and classical algorithm. *Brief Bioinform.* 11:bbaa161. doi: 10.1093/bib/bb aa161
- Beck, B. R., Shin, B., Choi, Y., Park, S., and Kang, K. (2020). Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model. *Comput. Struct. Biotech.* 18, 784–790. doi: 10.1016/j.csbj.2020.03.025
- Bender, A., Scheiber, J., Glick, M., Davies, J. W., Azzaoui, K., Hamon, J., et al. (2007). Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. *ChemMedChem* 2, 861–873. doi: 10.1002/cmdc.200700026
- Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., and Blaschke, T. (2018). The rise of deep learning in drug discovery. *Drug Discov. Today* 23, 1241–1250. doi: 10.1016/j.drudis.2018.01.039
- Choi, Y., Shin, B., Kang, K., Park, S., and Beck, B. R. (2020). Target-centered drug repurposing predictions of Human Angiotensin-Converting Enzyme 2 (ACE2) and Transmembrane Protease Serine Subtype 2 (TMPRSS2) interacting approved drugs for Coronavirus disease 2019 (COVID-19) treatment through a drug-target interaction deep learning model. *Viruses* 12:1325. doi: 10.3390/v12111325
- Duch, W., Swaminathan, K., and Meller, M. (2007). Artificial intelligence approaches for rational drug design and discovery. *Curr. Pharm. Des.* 13, 1497–1508. doi: 10.2174/138161207780765954
- He, B., Huang, J., and Chen, H. (2019). PVsiRNAPred: Prediction of plant exclusive virus-derived small interfering RNAs by deep convolutional neural network. *J. Bioinform. Comput. Biol.* 17:1950039. doi: 10.1142/S0219720019500392
- Joshi, T., Joshi, T., Pundir, H., Sharma, P., Mathpal, S., and Chandra, S. (2020). Predictive modeling by deep learning, virtual screening and molecular dynamics study of natural compounds against SARS-CoV-2 main protease. *J. Biomol. Struct. Dyn.* 5, 1–19. doi: 10.1080/07391102.2020.1802341
- Li, J., Pu, Y., Tang, J., Zou, Q., and Guo, F. (2020). DeepAVP: A dual-channel deep neural network for identifying variable-length antiviral peptides. *IEEE J. Biomed. Health Inform.* 24, 3012–3019. doi: 10.1109/JBHI.2020.2977091
- Lipinski, C. F., Maltarollo, V. G., Oliveira, P. R., daSilva, A. B. F., and Honorio, K. M. (2019). Advances and Perspectives in applying deep learning for drug design and discovery. *Front. Robot.* 6:108. doi: 10.3389/frobt.2019.0108
- Mason, D. M., Friedensohn, S., Weber, C. R., Jordi, C., Wagner, B., Meng, S., et al. (2021). Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nat. Biomed. Eng.* 5, 600–612. doi: 10.1038/s41551-021-00699-9
- Nand, M., Maiti, P., Joshi, T., and Chandra, S. (2020). A Ramakrishnan Virtual screening of anti-HIV1 compounds against SARS-CoV-2: machine learning modeling, chemoinformatics and molecular dynamics simulation based analysis. *Sci. Rep.* 10:20397. doi: 10.1038/s41598-020-77524-x
- Neves, B. J., Braga, R. C., Alves, V. M., Lima, M. N. N., Cassiano, G. C., Muratov, E., et al. (2020). Deep Learning-driven research for drug discovery: Tackling Malaria. *PLoS Comput. Biol.* 16:e1007025. doi: 10.1371/journal.pcbi.1007025
- Saka, K., Kakuzaki, T., Metsugi, S., Kashiwagi, D., Yoshida, K., Wada, M., et al. (2021). Antibody design using LSTM based deep generative model from phage display library for affinity maturation. *Sci. Rep.* 11:5852. doi: 10.1038/s41598-021-85274-7
- Stephenson, N., Shane, E., Chase, J., Rowland, J., Ries, D., Justice, N., et al. (2019). Survey of machine learning techniques in drug discovery. *Curr. Drug Metab.* 20, 185–193. doi: 10.2174/1389200219666180820112457
- Stokes, J. M., Yang, K., Swanson, K., Jin, W. G., Cubillos-Ruiz, A., Donghia, N. M., et al. (2020). A deep learning approach to antibiotic discovery. *Cell* 180, 688–702. doi: 10.1016/j.cell.2020.01.021
- Timmons, P. B., and Hewage, C. M. (2021). ENNAVIA is a novel method which employs neural networks for antiviral and anti-coronavirus activity prediction for therapeutic peptides. *Brief Bioinform.* 2021:bbab258. doi: 10.1093/bib/bbab258
- Ton, A. T., Gentile, F., Hsing, M., Ban, F., and Cherkasov, A. (2020). Rapid identification of potential inhibitors of SARS-CoV-2 main protease by deep docking of 1.3 billion compounds. *Mol. Inform.* 8:e2000028. doi: 10.1002/minf.202000028

- Zahradnik, J., Marciano, S., Shemesh, M., Zoler, E., Harari, D., Chiaravalli, J., et al. (2021). SARS-CoV-2 variant prediction and antiviral drug design are enabled by RBD in vitro evolution. *Nat. Microbiol.* 6, 1188–1198. doi: 10.1038/s41564-021-00954-4
- Zeng, X., Song, X., Ma, T., Pan, X., Zhou, Y., Hou, Y., et al. (2020). Repurpose open data to discover therapeutics for COVID-19 using deep learning. *J. Proteome. Res.* 19, 4624–4636. doi: 10.1021/acs.jproteome.0c00316
- Zhang, H., Yang, Y., Li, J., Wang, M., Saravanan, K. M., Wei, J., et al. (2020). A novel virtual screening procedure identifies Pralatrexate as inhibitor of SARS-CoV-2 RdRp and it reduces viral replication in vitro. *PLoS Comput. Biol.* 16:e1008489. doi: 10.1371/journal.pcbi.1008489
- Zhang, Y., Xi, H., and Juhas, M. (2020). Biosensing detection of the SARS-CoV-2 D614G mutation. *Trends Genet.* 37, 299–302. doi: 10.1016/j.tig.2020.12.004

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.

Copyright © 2021 Zhang, Ye, Xi, Juhas and Li. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



T1SEstacker: A Tri-Layer Stacking Model Effectively Predicts Bacterial Type 1 Secreted Proteins Based on C-Terminal Non-repeats-in-Toxin-Motif Sequence Features

Zewei Chen^{1†}, Ziyi Zhao^{1†}, Xinjie Hui^{2†}, Junya Zhang¹, Yixue Hu¹, Runhong Chen¹, Xuxia Cai¹, Yueming Hu¹ and Yejun Wang^{1*}

OPEN ACCESS

Edited by:

Mohammad-Hossein Sarrafzadeh,
University of Tehran, Iran

Reviewed by:

Timothy James Wells,
The University of Queensland,
Australia
Jack Christopher Leo,
Nottingham Trent University,
United Kingdom

*Correspondence:

Yejun Wang
wangyj@szu.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: 11 November 2021

Accepted: 20 December 2021

Published: 08 February 2022

Citation:

Chen Z, Zhao Z, Hui X, Zhang J,
Hu Y, Chen R, Cai X, Hu Y and
Wang Y (2022) T1SEstacker:
A Tri-Layer Stacking Model Effectively
Predicts Bacterial Type 1 Secreted
Proteins Based on C-Terminal
Non-repeats-in-Toxin-Motif Sequence
Features.
Front. Microbiol. 12:813094.
doi: 10.3389/fmicb.2021.813094

¹ Youth Innovation Team of Medical Bioinformatics, Shenzhen University Health Science Center, Shenzhen, China,

² Department of Respiratory Medicine, Xuanwu Hospital, Capital Medical University, Beijing, China

Type 1 secretion systems play important roles in pathogenicity of Gram-negative bacteria. However, the substrate secretion mechanism remains largely unknown. In this research, we observed the sequence features of repeats-in-toxin (RTX) proteins, a major class of type 1 secreted effectors (T1SEs). We found striking non-RTX-motif amino acid composition patterns at the C termini, most typically exemplified by the enriched “[FLI][VAI]” at the most C-terminal two positions. Machine-learning models, including deep-learning ones, were trained using these sequence-based non-RTX-motif features and further combined into a tri-layer stacking model, T1SEstacker, which predicted the RTX proteins accurately, with a fivefold cross-validated sensitivity of ~0.89 at the specificity of ~0.94. Besides substrates with RTX motifs, T1SEstacker can also well distinguish non-RTX-motif T1SEs, further suggesting their potential existence of common secretion signals. T1SEstacker was applied to predict T1SEs from the genomes of representative *Salmonella* strains, and we found that both the number and composition of T1SEs varied among strains. The number of T1SEs is estimated to reach 100 or more in each strain, much larger than what we expected. In summary, we made comprehensive sequence analysis on the type 1 secreted RTX proteins, identified common sequence-based features at the C termini, and developed a stacking model that can predict type 1 secreted proteins accurately.

Keywords: T1SS, T1SE, RTX proteins, T1SEstacker, prediction, deep learning

INTRODUCTION

Type 1 secretion systems (T1SSs) are uniquely distributed in Gram-negative bacteria, which can secrete various substrate proteins through the two bacterial cell membranes by one step (classical) or two steps (non-classical) into extracellular milieu (Smith et al., 2018b; Spitz et al., 2019). A T1SS is composed by three elementary components—an ATP-binding cassette (ABC) transporter located

in the inner membrane, an outer membrane factor (OMF), and a membrane fusion protein (MFP) connecting the ABC transporter (Kanonenberg et al., 2018). A wide variety of proteins are secreted through this oligomeric secretion channel to play their biological roles. Due to the simple structure of the system, T1SSs have been widely applied in biomedical engineering applications (Schwarz et al., 2012; Ryu et al., 2015; Park et al., 2020).

The T1SS substrates, also called type 1 secreted effectors (T1SEs), have various biological functions, such as host invasion (virulence factors, e.g., HlyA) (Felmlee et al., 1985), enzymolysis (digestion enzymes, e.g., TliA and PrtA) (Son et al., 2012), nutrient acquisition (iron-scavenger proteins, e.g., HasA) (Kanonenberg et al., 2013), and biofilm formation (adhesins, e.g., LapA) (Guo et al., 2019). Since the first T1SS substrate, hemolysin A (HlyA), was discovered in 1979 and its nucleotide sequence was determined in 1985 (Noegel et al., 1979; Felmlee et al., 1985), the structural characteristics and function of T1SEs have been studied extensively. Typical T1SEs can be classified into three classes simply according to their T1SS ABC transporter types: C39-containing ABC transporters with hydrolase activity, C39-like domain (CLD)-containing ABC transporters without hydrolase activity, and a third type of ABC transporters without any additional N-terminal domain (Hui et al., 2021). Class 1 T1SEs, known as the smallest T1SS substrates, normally contain N-terminal leader peptides. The C termini of the leader peptides contain a canonical double glycine ("GG") motif, which can be recognized and cleaved by the C39 domains of corresponding ABC transporters before the mature proteins are secreted through T1SSs (Kanonenberg et al., 2013). Class 2 T1SEs have remarkable repeats-in-toxin (RTX) domains and are also known as RTX proteins. The glycine-rich nanopeptide repeats in RTX domains show a "GGxGxDxUx" consensus sequence motif where "x" is any amino acid and "U" represents a large or hydrophobic amino acid. Class 3 T1SEs may also contain RTX repeat sequences but not necessarily. The last two categories do not contain N-terminal leader peptides, but instead potentially have secretion signal sequences in the C termini. However, the C-terminal signal patterns and function mechanisms remain to be clarified (Kanonenberg et al., 2013). Recently, a group of non-classical T1SEs named RTX adhesins (class 4) have been reported, which are closely related to biofilm formation (Smith et al., 2018b). Different from class 1–3 T1SEs, the RTX adhesins are transported from cytoplasm to extracellular environment by a two-step secretion mechanism, which involves periplasmic intermediates. This subgroup of T1SS machinery is linked with a bacterial transglutaminase-like cysteine proteinase (BTLCP) (Smith et al., 2018b). The RTX adhesion proteins have dialanine BTLCP cleavage sites in the N-terminal retention module that can be recognized and cleaved by the machinery-coupled BTLCP in periplasm before the cross-outer-membrane transport (Boyd et al., 2014; Smith et al., 2018b). The currently known RTX adhesins also have RTX repeats and signal sequences in the C termini (Boyd et al., 2014; Smith et al., 2018b).

Both the function and sequences of T1SEs show large diversity, and until now only ~100 T1SEs have been validated, which are homology-not-filtered, i.e., being redundant with high sequence homology, and therefore could represent fewer independent

validated effectors¹. Bioinformatic strategies have also been tried to predict novel T1SEs, but mainly focused on the RTX proteins with the consensus RTX motifs (Linhartova et al., 2010; Luo et al., 2015). For instance, Linhartova et al. (2010) combined pattern searching, Hidden Markov Model profiles, and the RPS-BLAST tool finding conserved domains to predict 1,024 candidate RTX proteins from 840 bacterial genomes, as comprised the most comprehensive list of RTX T1SE candidates. Luo et al. (2015) made the first attempt to develop a machine-learning model to predict RTX proteins. The random forest-based model learned amino acid sequence-derived features extracted from the full-length and C-terminal sequences of T1SE candidates predicted by Luo et al. (2015). Regretfully, neither a software tool nor a web server was provided for users to implement the method. Besides, both the homology-based and machine-learning methods completely focused on the RTX proteins and the conserved RTX motif was placed with a large weight. The methods are hardly generalized to find more novel T1SEs without RTX motif features.

By careful sequence pattern analysis, previously, we identified the position-specific amino acid composition (Aac), secondary structure element (Sse), and solvent accessibility (Acc) features of type 3 secreted effectors within their N termini and the various Aac, Sse, and Acc profiles of type 4 secreted effectors within their C termini (Wang et al., 2011, 2014). Given the evidence about the potential C-terminal secretion signals of T1SEs (Koronakis et al., 1989; Masure et al., 1990; Zhang et al., 1995; Delepelaire, 2004; Holland et al., 2005; Thomas et al., 2014), in this research, we comprehensively observed the amino acid sequence patterns, especially non-RTX-motif features within the C termini of RTX proteins, and also the Sse and Acc property. Furthermore, we developed machine-learning models to learn the newly observed sequence-derived features and predicted T1SEs with or without typical RTX motifs. Deep learning models and ensemblers have recently been widely used to predict bacterial secretion signals and achieved good performance (Wang et al., 2018, 2019; Almagro Armenteros et al., 2019; Xue et al., 2019; Hui et al., 2020). We also tested Deep Neural Network models and integrated them and others within a stacked model to improve the prediction performance.

MATERIALS AND METHODS

Datasets

Bacterial RTX proteins were collected from Linhartova et al. (2010). In total, there were 1,024 RTX proteins predicted from 840 bacterial genomes (Linhartova et al., 2010). CD-HIT was used to detect homology among the RTX proteins, while 30% was considered as the similarity cutoff and only one representative was retained if there were multiple proteins showing sequence similarity above the cutoff (Li and Godzik, 2006). Proteins were also sampled randomly from the whole proteomes derived from various bacterial genome sequences. The known T1SEs, RTX proteins, and their homologs with >30% blastp similarity were

¹<http://61.160.194.165/TxSEdb>

removed, and a homology filtering strategy similar to that applied for RTX proteins were used to identify the non-redundant non-RTX proteins. In total, 512 non-redundant RTX proteins were retained, which were considered as the positive dataset (p). A total of 2,000 proteins were also randomly selected from the processed non-RTX proteins, and three groups, each with 512 proteins, were further picked out to match the number and general length distribution of the RTX proteins, forming the negative datasets ($n1 \sim n3$). The p and $n1$ were used as the main observation datasets. A fivefold cross-validation strategy was used for training the machine-learning prediction models, for which both the positive and negative datasets were split into five subsets of equal size of protein sequences, with four of them being served as training datasets and the remaining one as testing datasets in each round of model analysis. Experimentally validated T1SEs were also annotated manually from literature. These proteins could be RTX or other type of proteins with experimental evidence to be transported through T1SSs. All the datasets were publically available together with the standalone T1SEstacker package (see Section “Software Availability”; see Text Footnote 1).

Once the datasets were collected and annotated, the sequence-based features were analyzed with in-house scripts. The secondary structure and solvent accessibility were predicted with SSpro/ACCpro5, with three elements encoded for secondary structure (“H” for helix, “E” for strand, and “C” for coil) and two elements for accessibility (“B” for buried and “E” for exposed) (Magnan and Baldi, 2014).

Sequential and Position-Specific Amino Acid Composition Feature-Based Non-deep-Learning Models

The number and position distribution of RTX motifs featured as “GGxGxD” was observed within the RTX and non-RTX proteins. Sequential Aac, continuous and 1 or 2 amino acid interrupted bi-residue amino acid composition (bAac) features were extracted from the C-terminal 20- or 60-residue fragments of both the positive and negative datasets, respectively, observed, and compared. The features were used for training Random Forest (RF), Support Vector Machine (SVM), and Naive Bayesian (NB) models, with R packages of “randomForest,” “e1071,” and the “e1071” method “naiveBayes,” respectively². The neighbor-position Aac conditional constraint features in the C termini were learned in Markov models (Wang et al., 2013). Bi-profile Bayesian position-specific Aac features were extracted and trained with SVM models (Wang et al., 2011). For the SVM models, four kernels (“linear,” “polynomial,” “sigmoid,” and “radial”) were tested and the corresponding parameters, e.g., *gamma* and/or *cost*, were optimized using a 10-fold cross-validation grid search strategy within each training dataset. For the other models, the features were also extracted based on each training dataset. The details about the models and the optimized parameters refer to the website of T1SEstacker (see Section “Software Availability”).

²<https://www.r-project.org/>

Deep Learning Models

Deep learning models were trained with the Aac features of RTX proteins within the C-terminal 20 (C20) and 60 amino acid positions (C60). Each position was represented by a 20-element feature vector describing the composition of amino acids. An $m \times 20$ L matrix was built to represent the original Aac features of training datasets, where m is the number of training proteins and L is 20 or 60 for C20 or C60 models, respectively. Fully connected Deep Neural Network (DNN), Self Attention (SelfAttention), and models with Long-Short Term Memory (LSTM) cells (RNN) were trained and tested with a fivefold cross-validation strategy. The details about the models and the optimized parameters refer to the website of T1SEstacker (see Section “Software Availability”).

A Stacked Model Featured by the Prediction Results of Individual Models

To achieve better prediction performance, we proposed a new stacking scheme to integrate prediction results of individual models (Figure 1). A primary stacked model was built for each original fivefold training dataset and its based individual models. For each original fivefold testing dataset, an embedded fivefold cross-validation was adopted to evaluate the performance of stacked models. The prediction result of each-fold best-trained model of individual algorithms on each protein of the corresponding testing dataset was based, and encoded as 1 (RTX) or 0 (non-RTX) according to the model-specific optimized cutoff score. Each protein within an embedded fivefold training dataset was represented as a feature vector of “0” and “1,” and an $m' \times n$ matrix was generated for the whole training dataset, where m' is the protein number of the embedded training dataset and n is the number of individual machine-learning models. SVM models with “linear” kernels were trained and the parameters (costs) were optimized with a 10-fold cross-validation grid-searching strategy.

A voting strategy was used to integrate the five primary stacked models, with the same weight assigned for each model.

Performance Evaluation of the Individual and Stacked Models

Sensitivity (Sn), specificity (Sp), accuracy (ACC), the area under the curve of receiver operating characteristic (rocAUC), and Matthews correlation coefficient (MCC) were defined and used as measures to assess the performance of models based on a fivefold cross-validation strategy.

$$Sn = TP / (TP + FN)$$

$$Sp = TN / (TN + FP)$$

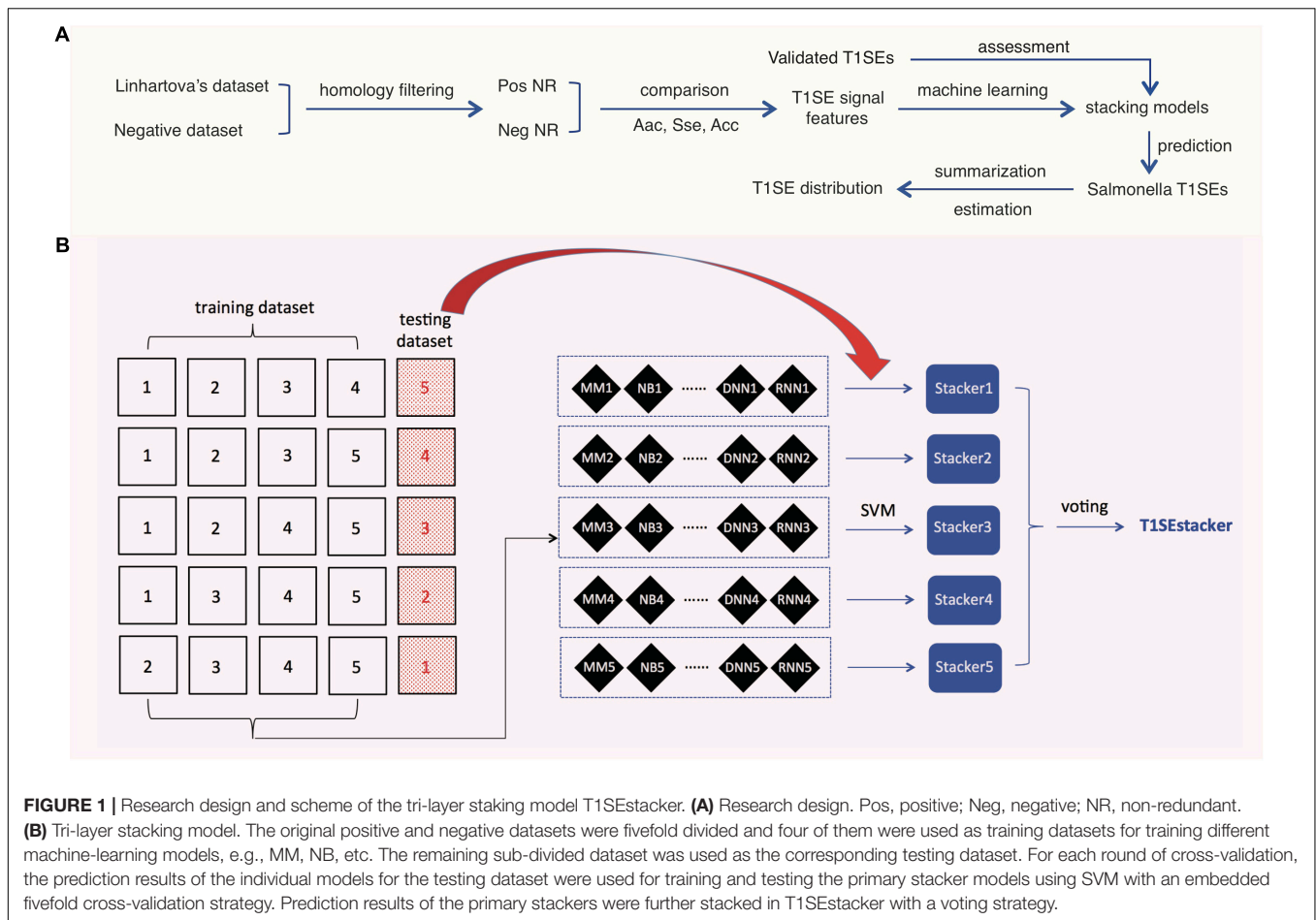
$$ACC = (TP + TN) / (TP + FN + TN + FP)$$

$$MCC = [(TP \times TN) - (FN \times FP)] / \sqrt{[(TP + FN) \times (TN + FP) \times (TP + FP) \times (FN + FN)]}$$

TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively.

Statistics

Individual amino acids were counted within C-terminal 20, 60, or 110-aa fragments, and Mann–Whitney tests were performed to compare their distribution between RTX and non-RTX proteins,



followed by Bonferroni corrections. For two continuous or non-continuous amino acids (bi-AAAs), the composition was also compared between the C termini of RTX and non-RTX proteins using the same statistical methods. Another balanced rate comparison method, EBT, was also adopted to compare the C-terminal occurrence of bi-AAAs between the two classes of proteins (Hui et al., 2017). The alpha levels for all tests were preset as 0.05.

Software Availability

T1SEstackers and its modules were developed with Python, Perl, and R. The packages and user manual can be downloaded freely via the link, <http://www.szu-bioinf.org/tools/T1SEstacker>. A web server was also initiated to make internet-based prediction service: <http://www.szu-bioinf.org/T1SEstacker>.

Salmonella Genomes

In total, 26 representative strains were included, which covered the known *Salmonella* phylogenetic groups. N268_08, NCTC12419, and RKS3044 belong to *Salmonella bongori*; RKS2983 and RSK2980 belong to *Salmonella enterica* subsp. *arizonae*; ATCC_BAA_1581 and RKS3027 belong to *Salmonella enterica* subsp. *houtenae*; 2439-64 and RKS3013 belong to *Salmonella enterica* subsp. *vii*; 11_01853, 11_01854, 11_01855,

and RKS2978 belong to *S. enterica* subsp. *diarizonae*; RKS2986 and ST114 belong to *Salmonella enterica* subsp. *salamae*; 1121 and RKS3057 belong to *Salmonella enterica* subsp. *indica*; while P12519, 287/91, ATCC9150, SPB7, RKS4594, ATCC9120, CT18, 14028S, and LT2 represent various serovars of *Salmonella enterica* subsp. *enterica*. The genome and genome-encoding proteome were downloaded from NCBI genome database³. T1SEstacker was applied to predict the T1SE candidates with default settings.

RESULTS

Research Design

The major obstacles for training machine-learning models in prediction of bacterial T1SEs include (1) the limited number of experimentally validated positive proteins and (2) the large sequence diversity of T1SE groups. Comprehensive literature searching and manual annotation only curated 99 validated T1SEs, and only 49 were retained after a strict homology-filtering process, which were distributed in all the four major T1SE groups (see Text Footnote 1). To better analyze the likely novel sequential features that could facilitate understanding the mechanisms of

³<https://www.ncbi.nlm.nih.gov/genome>

type 1 secretion and prediction of new T1SEs, and as performed by others previously (Luo et al., 2015), we took the larger-scale RTX T1SE candidates identified by Linhartova et al. (2010) as training data for analysis of features other than RTX motifs and building models to predict novel T1SEs.

After removing the homologs, the remaining non-redundant T1SEs and paired non-T1SEs were compared for their sequential and position-specific Aac, Sse, and Acc features, especially non-RTX motif features (Figure 1A). With the sequence-based features, a stacking model was developed to predict T1SEs (Figure 1B). Representative strains of *Salmonella* phylogenetic branches were predicted with the newly developed model, and the possible number and distribution of candidate T1SEs were evaluated (Figure 1A).

Distance Distribution of Repeats-in-Toxin Motifs to the C Termini in Repeats-in-Toxin Proteins

The 512 non-redundant RTX proteins show a length distribution from 70 to 36,805 amino acids, with a median of 1,112 residues and 7 super-long proteins with larger than 10,000 amino acids (Figure 2A). In addition, 494 from the 512 positive proteins could be found with at least one RTX motif within each protein sequence (Figure 2B). As a control, only 13 from the total 2,341 non-redundant negative proteins contained RTX motifs, which were filtered for further comparative or model-training analysis. The most C-terminal residue of each most C-terminal RTX motif shows a distance of 1–21,948 amino acids to the C terminus of the corresponding full-length protein, with a median of 110 amino acids (Figure 2C). Fewer than 9% of the C-terminal RTX motifs have a distance of smaller than 60 amino acids from the protein C termini, and only ~5% are shorter than 20 amino acids (Figure 2D).

Sequential Amino Acid Composition Features Buried in the C Termini of Repeats-in-Toxin Proteins

We compared the composition of individual amino acids (Aac) and two continuous or non-continuous amino acids (bAac) among the C termini of RTX proteins since there were possibly atypical secretion signals (Boyd et al., 2014; Smith et al., 2018a). To avoid the possible misinterpretation caused by RTX motifs, we mainly observed the Aac and bAac profiles within the C-terminal 20 (C20) and C-terminal 60 (C60) residues (Supplementary Dataset 1). Within C20, most individual amino acids show different compositions between the positive and negative proteins, with aspartic acid (D), leucine (L), threonine (T), valine (V), isoleucine (I), and phenylalanine (F) being most typically enriched and arginine (R), lysine (K), glutamic acid (E), and proline (P) being most strikingly depleted in RTX proteins (Figure 3A; Mann–Whitney *U*-tests with Bonferroni correction, $p < 0.001$). Glycine (G) was not different between the two types of proteins (Figure 3A; $p = 1$). When the observed length increases to C-terminal 60-aa, most of the featured residues identified from shorter fragments remain different between groups for the composition, whereas some others start to show difference or no

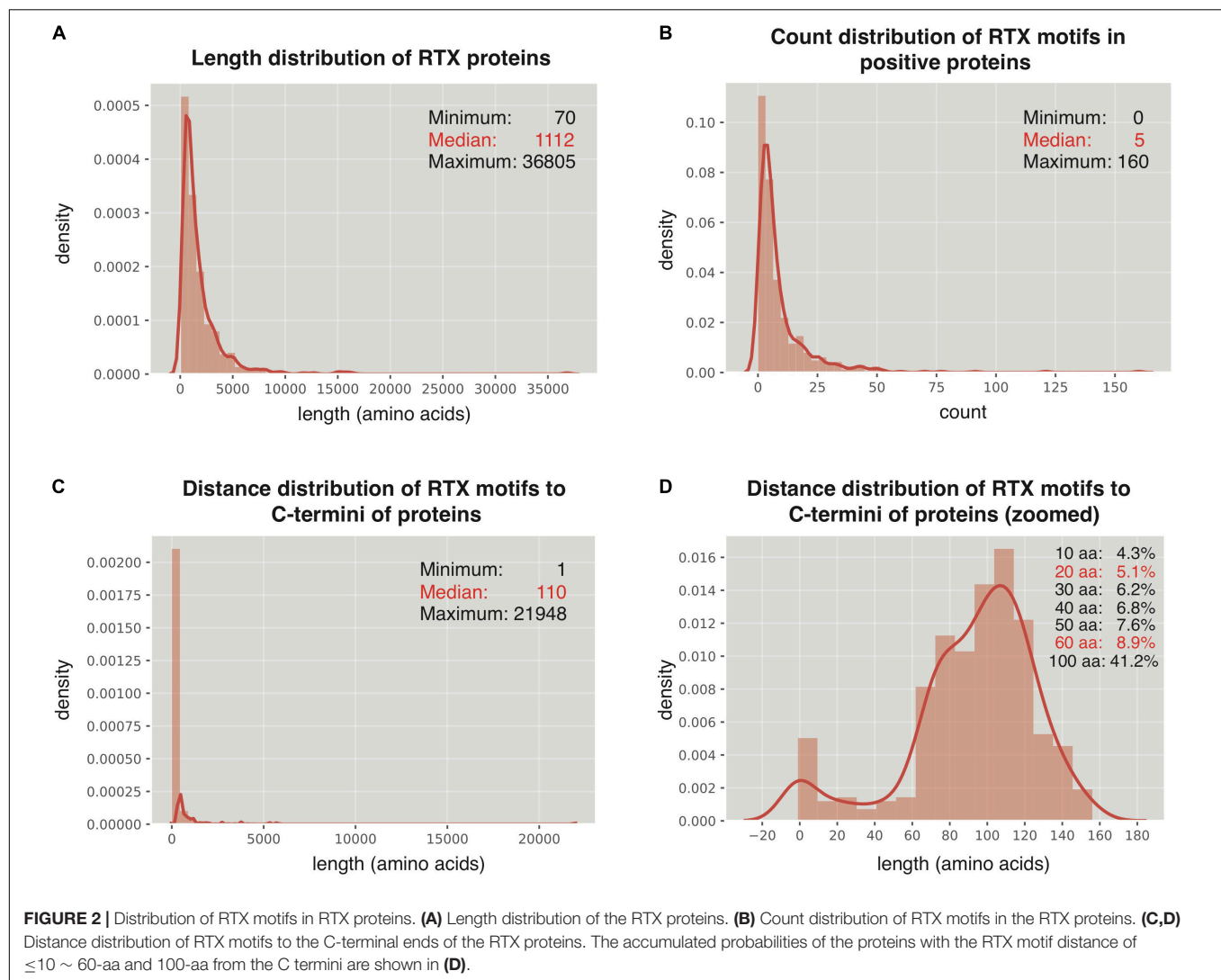
difference, e.g., “G” being enriched in RTX proteins and “L,” “V,” and “I” becoming no difference (Figure 3A). The enrichment of “G” in RTX C60 fragments is not likely due to the increasing occurrence of RTX motifs, which is enriched with “G,” since the RTX motifs are lowly represented and the RTX motif featured “GG” is either not strikingly higher in the C60 fragments of RTX proteins (Figures 2D, 3B). For the C-terminal 110-aa fragments, the amino acid species with significantly different composition and the amplitude of difference further increase (Figure 3A). It cannot be excluded that the increased number of RTX motifs leads to the most striking composition amplitude change of “D” and “G,” especially in C110, for which half of the sequences contained the RTX motifs. However, the “L” composition change is interesting, which shows higher composition in C20, no difference in C60, and lower composition in C110 of RTX proteins (Figure 3A).

The continuous and interrupted bAac profile also shows difference in C termini between RTX and non-RTX proteins. For example, “D[FL],” “TL/LT,” “AxD,” “Tx[LT],” “TxxD,” and “Dxx[FI]” most frequently occur, whereas “R[RK],” “K[KA],” “AxR,” “Rx[RL],” “AxxR,” “Kxx[KE],” and “Rxx[QR]” are most strikingly depleted in the C-terminal 20-aa fragments of RTX proteins in contrast to non-RTX proteins (Figures 3B–D; Mann–Whitney *U*-tests with Bonferroni correction, $p < 0.001$; EBT- $p < 0.001$). As the observed C-terminal length increases (to 60 aa), the general bAac profile difference between RTX and non-RTX proteins remains or becomes more typical, with only a few changes. The main changes involve the reduced “L” and increased “G” combinations in the RTX C60 enriched list (Figures 3B–D). It is noted that either “GG” or “GxG,” which is supposed to be highly represented by RTX motifs, does not show the most significant different composition or occurrence in C60 between RTX and non-RTX proteins, suggesting that the observed different “G”-combination compositions are not due to the increased percent of RTX motifs in C60 of RTX proteins. In C110, however, the composition shows striking difference for both “GG” and “GxG” between RTX and non-RTX proteins (Supplementary Dataset 1).

Other independent non-RTX proteins datasets are also paired and the profile difference for Aac and bAac in C termini between RTX and non-RTX proteins shows large consistence.

Position-Specific Amino Acid Composition Features Buried in the C Termini of Repeats-in-Toxin Proteins

The C-terminal position-specific amino acid composition (psAac) profiles were also compared between RTX and non-RTX proteins. Generally, RTX proteins show much larger amino acid composition preference (Figure 4A). C20 and C21–60 in RTX proteins also show different preference profiles. C20 shows apparent preference for non-polar “L” and “A” while C21–60 more prefers polar “G” (Figure 4A). “D,” “S,” and “T” are preferred in both C20 and C21–60 of RTX proteins. The results are consistent with and explain the observations on sequential Aac and bAac in C termini of RTX and non-RTX proteins. Remarkably, the C-terminal endmost two positions in RTX proteins show the



most typical psAac bias, with a pattern of non-polar hydrophobic “[FLI][VAI]” motif (**Figure 4A**).

The psAac profile of C termini of RTX proteins and the difference between them and non-RTX proteins were confirmed with other, paired, independent negative datasets (**Supplementary Figure 1**). We also compared the psAac profile of N termini of RTX and non-RTX proteins (**Supplementary Figure 2**). There was a difference, but not as typical as that observed within the C termini. Moreover, until now there is no evidence suggesting the existence of type 1 secretion signals within N termini of the substrate proteins. Therefore, the N termini were not further studied in this study.

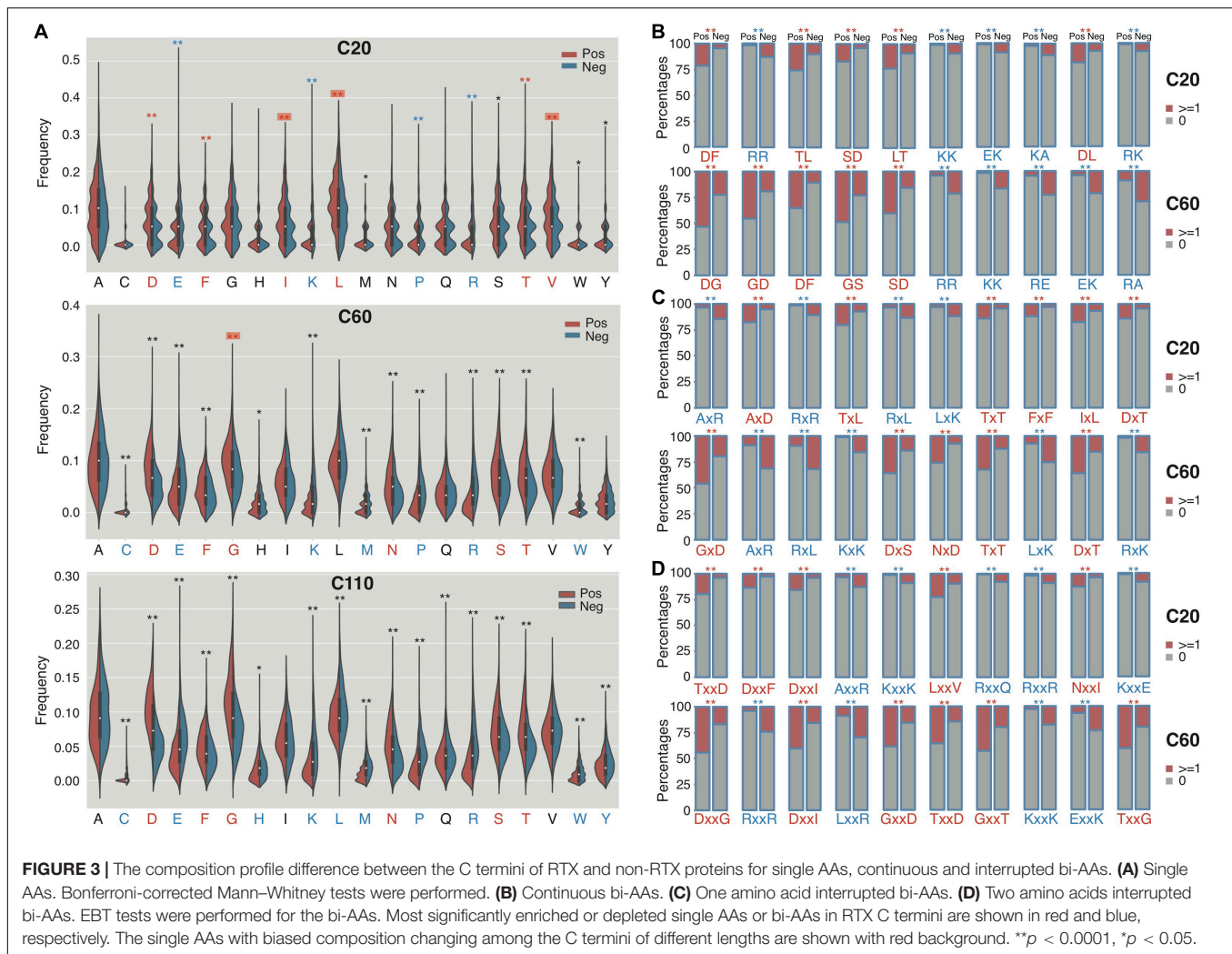
Enrichment of β -Strands and Depletion of α -Helices Within the C Termini of Repeats-in-Toxin Proteins

An apparent difference between the C termini of RTX T1SEs and non-T1SEs was the depletion of α -helices or enrichment of β -strands and coiled coils, no matter in C20 or C60 (**Figure 4B**).

The solvent accessibility was not different between the RTX and non-RTX proteins within the C termini (data not shown). The different forms of secondary structure are likely related with the composition preference of residues. For instance, both polar “G” and non-polar “A” are enriched in β -strands, while “F” and “I” are not for beneficial for maintenance of the stability of α -helices (**Figure 4A**). It remains to be clarified whether the residue composition and structure features are associated with specific recognition of the proteins for specific type 1 secretion.

C-Terminal Non-repeats-in-Toxin Motif Features Accurately Classify Repeats-in-Toxin From Non-repeats-in-Toxin Proteins

A list of machine-learning models were trained to learn the sequence-based non-RTX motif features buried within the C termini of RTX proteins, including NB, RF, and SVM models learning sequential Aac and bAac features, MM models using adjacent amino acid dependent Aac features, and SVM models



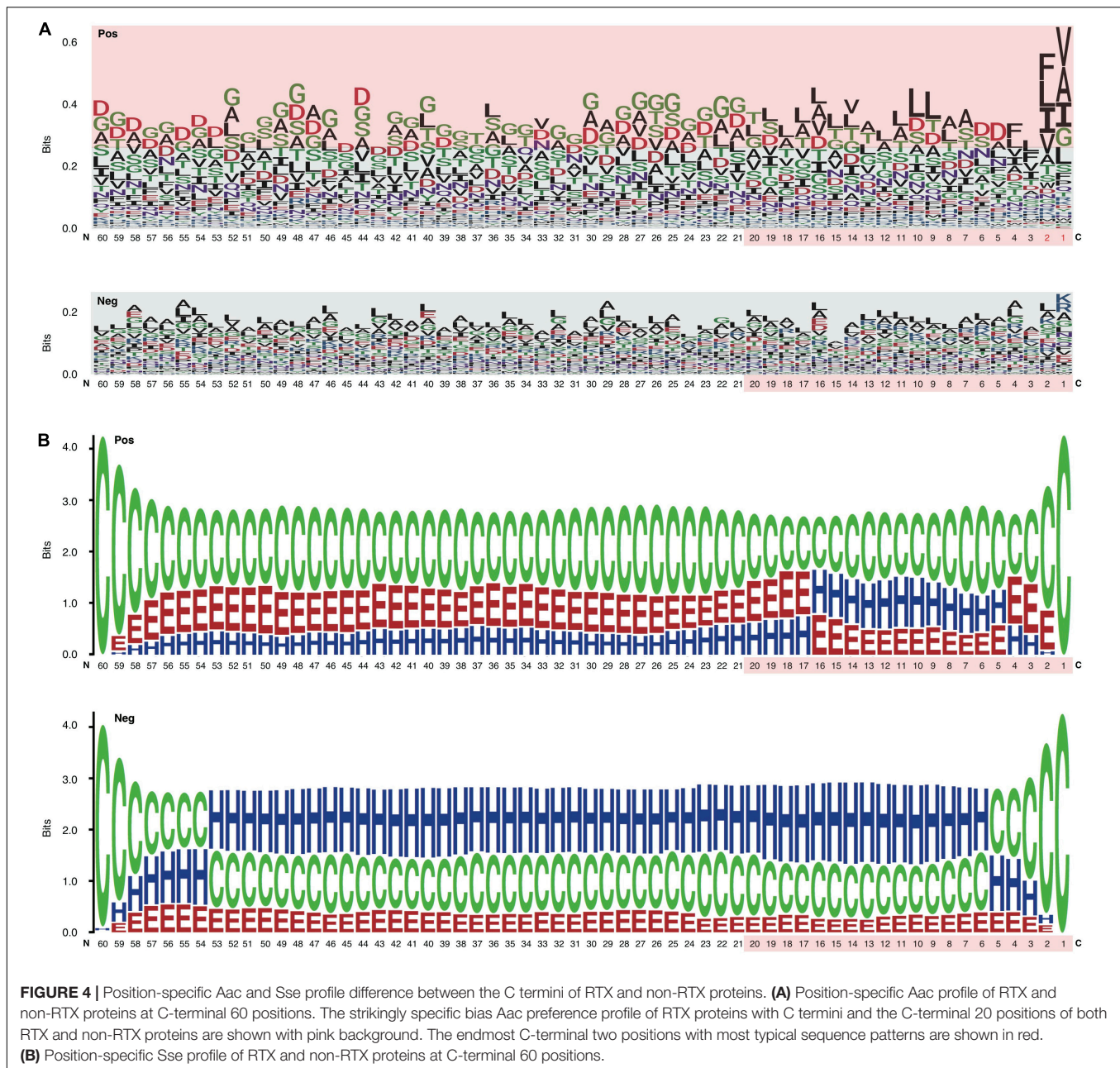
analyzing position-specific Aac features (Table 1). Moreover, five types of DL models were trained, with three among them of best performance retained (DNN, Attention, and RNN), which also learned the C-terminal Aac features of RTX proteins (Table 1). Secondary structure features were not learned in the models since they are not stable, which were predicted with varied accuracy using different software tools.

All the models showed certain ability to classify RTX proteins from the non-RTX ones correctly only based on the Aac features within C-terminal 20-aa peptide fragments of known RTX proteins (Table 2 and Figure 5A). RNN, MM, RF, and seqSVM showed best prediction performance with the same average rocAUC of 0.88, while BPBAac and DNN appeared poorest with a rocAUC of 0.85 (Table 2 and Figure 4A). C60 models outperformed C20 ones obviously, and MM, RF, and seqSVM remained the best-performed models, reaching a rocAUC of 0.94 (Table 2 and Figure 5B).

Taken together, the results demonstrate that the C termini of RTX proteins contain non-RTX Aac signals, which can be used to recognize RTX proteins accurately. The signals are likely distributed along the C-terminal 60-aa positions.

A Stacked Model Shows Striking Performance Improvement in Prediction of Repeats-in-Toxin Types of Type 1 Secreted Effectors

To achieve better performance, we designed a tri-layer stacking model, which integrates the prediction results of individual models learning sequence-based features, to classify RTX and non-RTX proteins (Figure 1). The primary SVM-based stacked models (pT1SEstacker) trained with the prediction results of original fivefold cross-validated testing datasets showed better performance than individual models for both C20 and especially C60, with average rocAUC of 0.85 and 0.95, respectively (Table 2 and Figure 5C). The prediction results of the primary stacked models based on cross-validated testing datasets were assembled in the final model (T1SEstacker) with a voting strategy. It is noted that, with an independent dataset, which will be explained in the next section, the voting-based tri-layer stacker T1SEstacker generally balanced the effect of individual pT1SEstacker models and always achieved slightly better performance when voting cutoff was set as 0.6 (Figure 6A and Supplementary Figure 3).



T1SEstacker Can Recognize the Common Secretion Signals Among Different Types of Type 1 Secreted Effectors

We curated experimentally validated T1SEs and applied the RTX protein prediction models to identify them. It should be noted that none of the C60 or C20 of the verified T1SEs contained any RTX motif. Both T1SEstacker_C20 and T1SEstacker_C60 could well predict the T1SEs (Figure 6A and Supplementary Figure 3). The recalling rate of T1SEstacker_C20 and T1SEstacker_C60 reached 77 and 81%, respectively (Figure 6B). As a control, we used an independent negative dataset, and the specificity

of T1SEstacker_C20 and T1SEstacker_C60 was 89 and 96%, respectively (Figure 6B).

Among the validated T1SEs, 25% (25/99) do not contain any putative RTX motif along the full-length protein sequences (Figure 6C and Supplementary Dataset 2). Interestingly, T1SEstacker_C60 correctly recalled 52% (13/25) of the non-RTX-motif T1SEs (Figure 6C). Another non-RTX-motif T1SE was predicted to be non-effector by the final T1SEstacker_C60 model, yet it was correctly recalled by two primary models. The recalling rates of non-RTX-motif T1SEs are much higher than the false-positive rates of the negative dataset for both C60 and C20 models (Figures 6B,C). Therefore, the results further suggested that C termini of T1SEs, with-RTX-motif or non-RTX-motif

TABLE 1 | Models and the optimized parameters.

Model	Algorithm Features
MM	Markov model Aac conditional on that of the preceding position.
RF	Random forest AAs, continuous and interrupted bi-AAAs with striking sequential composition difference between positive and negative sequences.
NB	Naïve Bayes same with RF.
seqSVM	Support vector machine same with RF.
BPBAac	Support vector machine bi-profile position-specific Aac profiles.
DNN	Simple full-connected deep neural network Aac profiles.
SelfAttention	Softmax deep neural network Aac profiles.
RNN	Deep neural network with LSTM cells Aac profiles.

type, potentially contained common signals, which can guide the accurate prediction of these proteins.

Most of the validated T1SEs were not well classified into one of the four T1SE classes, except for seven being clear class 4 effectors, including enterotoxigenic *Escherichia coli* CexE (accession: ABM92275.1), *Gallibacterium anatis* GtxA (OBW99045.1), *Pseudomonas fluorescens* LapA (ABA71877.1), *Legionella pneumophila* RtxA (CAH11847.1), *Bordetella bronchiseptica* BrtA (CAE31684.1), *Shewanella oneidensis* BpfA (Q8EIX3.1), and *Vibrio cholera* FrhA (AWB74152.1). Five could be predicted by T1SEstacker_C60 correctly and only two (BpfA and CexE) were not recalled (**Supplementary Dataset 2**). The well-known class 2 effector, *E. coli* HlyA (P08715.1), other two class 2 effectors, *Aggregatibacter actinomycetemcomitans* LtxA (WP_148335754.1) and *Neisseria meningitidis* FrpC (AAA99902.1), and one typical class 3 effector, *Serratia marcescens* LipA (Q59933), were all correctly predicted (**Supplementary Dataset 2**). Because the other effectors were not well classified, we did not further

compare the prediction performance of T1SEstacker on different T1SE classes. Interestingly, five validated T1SEs were annotated to be bacteriocins, including *Rhizobium leguminosarum* RzcA (AAF36415.1), *Bradyrhizobium elkanii* BAB55900.1, *Xylella fastidiosa* XF2407 (AAF85206.1) and XF2759 (AAF85544.1), *Xanthomonas oryzae* AAW74644.1, and *Agrobacterium tumefaciens* RzcA (AAK89027.2). Four of the bacteriocins were correctly predicted, except for RzcA (**Supplementary Dataset 2**).

Large Variation of Type 1 Secreted Effectors Composition in *Salmonella* Strains

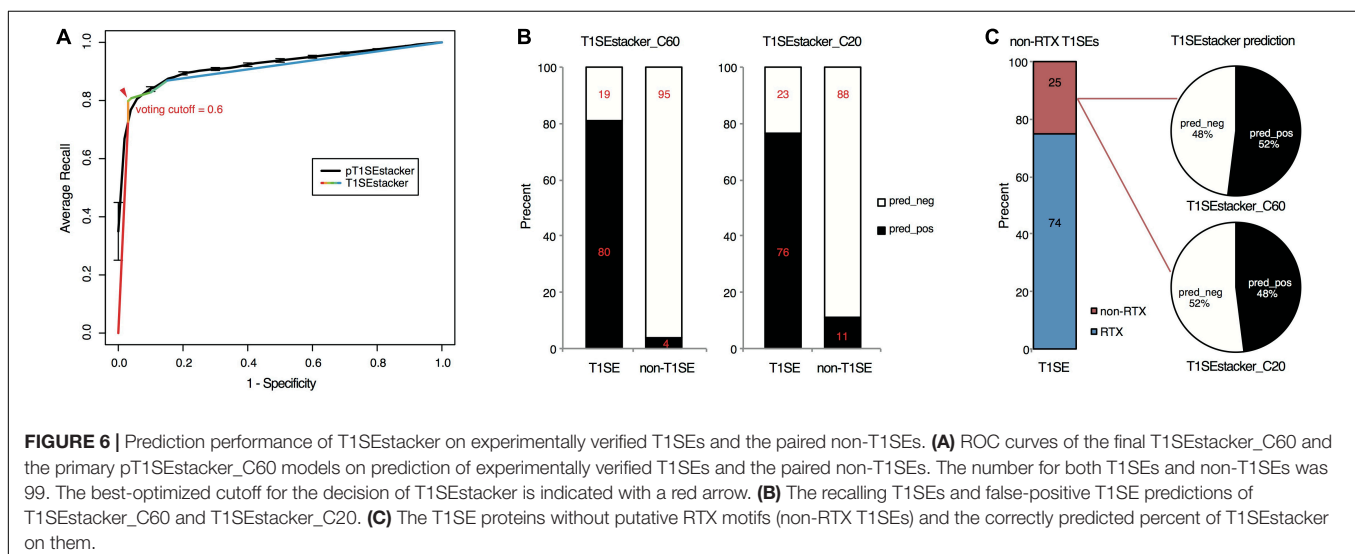
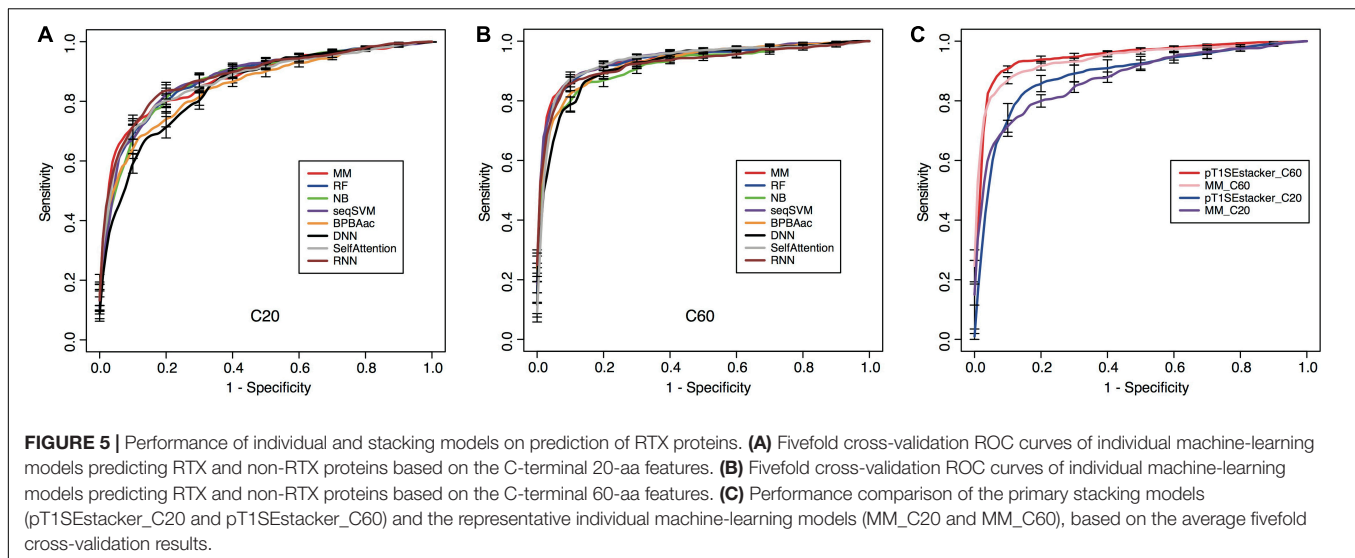
The chromosomes of 26 representative strains from all *Salmonella* major phylogenetic branches were scanned with T1SEstacker C60 model (**Supplementary Dataset 3**). In each strain, 269 ± 22 T1SE candidates were predicted (**Figure 7A**). With the recalling rate of 0.81 and false-positive rate of 0.04 evaluated previously on the validated T1SE dataset, the real number of T1SEs was estimated to reach 88 to 154, with an average of 123, in *Salmonella* strains (**Figure 7A**). The precision of predicted T1SE candidates was only ~ 0.37 ($123 \times 0.81/269$). However, it is difficult to improve the precision by shifting the decision cutoff values or to distinguish the true positives from the false ones. Moreover, most of the real T1SEs were included in the predictions. Therefore, we used the original T1SEstacker predictions to analyze the distribution of T1SE candidates among the *Salmonella* strains.

Despite a relatively stable number of T1SE candidates in different strains, the protein composition varied a lot. The candidates were clustered into 1,004 orthologous families, among which 240 (24%) were strain-specific proteins, 670 (67%) were

TABLE 2 | Performance of models.

Model	SN	SP	ACC	rocAUC	MCC
MM_C20	0.81 \pm 0.06	0.81 \pm 0.05	0.81 \pm 0.02	0.88 \pm 0.02	0.62 \pm 0.04
RF_C20	0.79 \pm 0.06	0.82 \pm 0.09	0.80 \pm 0.06	0.88 \pm 0.04	0.61 \pm 0.12
NB_C20	0.89 \pm 0.05	0.69 \pm 0.06	0.79 \pm 0.04	0.87 \pm 0.03	0.59 \pm 0.09
seqSVM_C20	0.79 \pm 0.05	0.82 \pm 0.06	0.81 \pm 0.04	0.88 \pm 0.04	0.61 \pm 0.09
BPBAac_C20	0.72 \pm 0.06	0.82 \pm 0.02	0.77 \pm 0.03	0.85 \pm 0.03	0.55 \pm 0.06
DNN_C20	0.77 \pm 0.05	0.75 \pm 0.05	0.76 \pm 0.04	0.85 \pm 0.03	0.53 \pm 0.07
SelfAttention_C20	0.80 \pm 0.03	0.80 \pm 0.05	0.80 \pm 0.04	0.87 \pm 0.02	0.60 \pm 0.07
RNN_C20	0.82 \pm 0.05	0.80 \pm 0.05	0.81 \pm 0.04	0.88 \pm 0.04	0.63 \pm 0.07
pT1SEstacker_C20	0.83 \pm 0.06	0.85 \pm 0.04	0.84 \pm 0.04	0.88 \pm 0.06	0.69 \pm 0.09
MM_C60	0.86 \pm 0.06	0.93 \pm 0.04	0.89 \pm 0.02	0.94 \pm 0.02	0.79 \pm 0.02
RF_C60	0.85 \pm 0.06	0.90 \pm 0.02	0.88 \pm 0.03	0.94 \pm 0.03	0.76 \pm 0.05
NB_C60	0.86 \pm 0.03	0.83 \pm 0.05	0.84 \pm 0.04	0.92 \pm 0.02	0.69 \pm 0.07
seqSVM_C60	0.84 \pm 0.08	0.92 \pm 0.02	0.88 \pm 0.04	0.94 \pm 0.02	0.77 \pm 0.07
BPBAac_C60	0.84 \pm 0.04	0.89 \pm 0.02	0.87 \pm 0.02	0.93 \pm 0.01	0.73 \pm 0.03
DNN_C60	0.87 \pm 0.06	0.85 \pm 0.04	0.86 \pm 0.02	0.92 \pm 0.02	0.72 \pm 0.04
SelfAttention_C60	0.89 \pm 0.02	0.89 \pm 0.03	0.89 \pm 0.02	0.93 \pm 0.01	0.78 \pm 0.03
RNN_C60	0.85 \pm 0.07	0.90 \pm 0.07	0.87 \pm 0.03	0.93 \pm 0.03	0.75 \pm 0.06
pT1SEstacker_C60	0.89 \pm 0.04	0.94 \pm 0.02	0.91 \pm 0.02	0.95 \pm 0.02	0.83 \pm 0.03

Sn, Sensitivity; Sp, specificity; ACC, accuracy; rocAUC, the area under the curve of receiver operating characteristic; MCC, Matthews correlation coefficient. The best performance was highlighted in bold font.



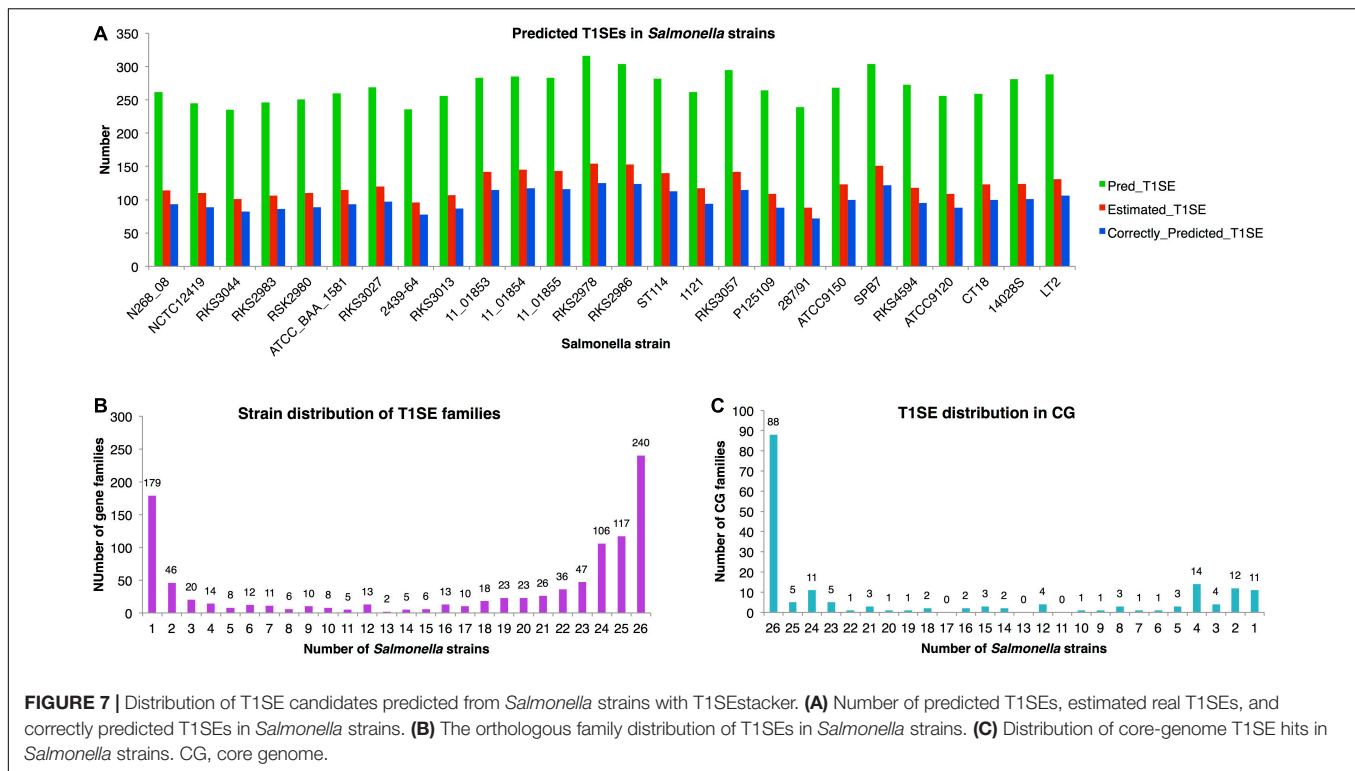
present in fewer than half of the strains, and only 179 (18%) were distributed in the core genome of the *Salmonella* strains (**Figure 7B** and **Supplementary Dataset 3**). For the core-genome hits, only 49% (88/179) were recognized as T1SEs in all the strains, and 31% (55/179) of the families were predicted as T1SEs only in fewer than half of the strains (**Figure 7C** and **Supplementary Dataset 3**). The results suggested that there is a large variety for the composition of T1SEs in different bacterial strains, and that a T1SE homolog does not necessarily remain a T1SE since mutations in the C terminus could frequently avoid the recognition of T1SS.

DISCUSSION

Like other secreted proteins, bacterial type 1 secreted proteins (T1SEs) also play important roles in various infection diseases. Some T1SEs, e.g., bacteriocins, show non-self bacteria-killing

activities and therefore have been used for anti-bacteria drug or probiotic development. How many T1SEs are there in each bacterial strain? How diverse is their function? The questions remain unanswered since we are still at the very beginning on understanding the mechanisms of type 1 secretion. Only around 100 T1SEs have been verified by experiments, and many of them contain RTX motifs nearby the C termini of protein sequences. However, not all T1SEs contain RTX motifs, while the proteins with RTX motifs, although more likely to be, are not necessarily T1SEs. Therefore, T1SEs could have other common targeted signals that mediate their specific type 1 secretion. More novel T1SEs could be identified based on these common signals.

Previous studies suggested possible signals within C termini of RTX and non-RTX T1SEs (Delepeleire, 2004; Huang et al., 2010; Wakeel et al., 2011). In this research, we focused on RTX T1SEs, observed the Aac features within their C termini comprehensively, and compared them with the C termini of non-RTX proteins or N termini of the RTX and non-RTX



proteins. It was interesting to identify specific Aac preference in C termini of RTX proteins (**Figure 4A**). As control, no apparent difference was found between the N termini of RTX and non-RTX proteins (**Supplementary Figures 1, 2**). The Aac preference profile was not biased by possibly included RTX motifs. On one hand, very few RTX motifs were retained in the observed length of C-terminal sequences (both C20 and C60) (**Figures 2C,D**). On the other hand, the motif-enriched bi-AAAs were not as strikingly different as other bi-AAAs (**Figure 2B**). Moreover, the real occurrence of some individual AAs or bi-AAAs within C termini of RTX proteins, especially C60, e.g., “G” and “D,” was much higher than the percentage of proteins with putative RTX motifs within the region. Therefore, such Aac preference could be independent of RTX motif. Alternatively, RTX motifs could also represent the preference, but a more specific and conserved pattern. Besides the enriched Aac, significantly depleted Aac should also be noted, e.g., “E,” “K,” “R,” and “P.” In the research, by observing the position-specific Aac profiles, we also identified a typical amino acid composition pattern at the C termini of RTX proteins, with a motif feature of “[FLI][VAI].” Previous studies on known T1SEs found the enrichment of “[LDAVTSIF]” residues in C-terminal signal regions (Delepleaire, 2004; Huang et al., 2010; Wakeel et al., 2011). The features were also evident in our C-terminal sequence-based or position-specific Aac analysis on the T1SEs. HlyA and its homologs in *E. coli*, *Proteus vulgaris*, and *Morganella morganii* were all shown with a preference of “[LS][AV]” at the C termini (Koronakis et al., 1989), consistent with our position-specific Aac observation. We also found that the C termini of RTX proteins preferred β -strands rather than α -helices as in non-RTX proteins (**Figure 4B**). It is intriguing to

further investigate whether the unique amino acid composition and secondary structure contribute to the specificity of signal recognition of type 1 secretion.

Machine-learning models based on the C-terminal non-RTX-motif Aac features well predicted RTX proteins from non-RTX proteins (**Figure 5** and **Table 2**). The features within C20 showed certain power, while those buried in C60 showed better distinguishing capability (**Figure 5** and **Table 2**). The C60 models could also accurately recall verified T1SEs at high prediction specificity (larger than 95%) (**Figure 6B**). It should be pointed out again that none of the verified T1SEs contained any RTX motif within C20 or C60 regions. More interestingly, 25 of the verified T1SEs do not contain RTX motif throughout their full-length sequences, and yet 12 and 13 were still predicted by C20 and C60 models, respectively, as positive results (**Figure 6C**). Among the correctly predicted T1SEs, some are bacteriocins and others are not putative RTX proteins. Therefore, the features identified in this study can be used for development of general T1SE prediction models. In future studies and as more non-RTX T1SEs have been identified, the common features can be reanalyzed, with a more balanced training dataset of different types of T1SEs.

We developed a tri-layer stacking model, T1SEstacker, and showed that the stackers generally outperformed the individual machine-learning models (**Table 2** and **Figure 5C**). However, some individual models also showed good performance, e.g., MM, RNN, SelfAttention, and RF, but generally not as good or stable as the stackers, pT1SEstacker (**Table 2** and **Figure 5C**). We made a second round of stacking for the pT1SEstackers trained with sub-divided cross-validated datasets because for pT1SEstackers, we adopted a SVM model to

integrate the prediction results of individual machine-learning models (**Figure 1**). Similar with T1SEstacker that integrates pT1SEstacker results, other ensemblers often use voting strategy (Wang et al., 2019) or linearly weight each individual model (Hui et al., 2020). The parameters, i.e., linear weights for individual models and decision cutoffs for those models, were generally stable and not very sensitive to the sub-divided or full training datasets. However, for pT1SEstacker models, we trained the prediction results of individual models using SVM, and the parameters were pretty sensitive to the training datasets. Therefore, the five pT1SEstackers were each with different optimized parameters. To integrate their respective prediction results, another round of stacking had to be performed. The final model T1SEstacker appeared not apparently better than the pT1SEstacker models. However, once the optimized voting cutoff was selected (≥ 0.6 , 3/5, consensus prediction), the prediction of T1SEstacker always showed best performance, with a compromise of sensitivity and specificity (**Figure 6A** and **Supplementary Figure 3**).

The false-positive rate (FPR) of T1SEstacker_C60 was low and close to 0.04. It is important since many tools predicting bacterial secreted proteins showed a high FPR and the experimental research seldom benefited from the tool (Hui et al., 2020). As an example, we showed the influence of FPR on the final prediction performance, by prediction and estimation of T1SE candidates in *Salmonella* with T1SEstacker (**Figure 7A**). Despite the high specificity (0.96), among the predicted T1SE candidates, majority were false positives, and the precision was only ~ 0.37 (**Figure 7A**). It is largely because for each genome, most genes are non-T1SEs, and even 1% FPR could generate 50–100 false-positive predictions, for which the number is close to that of true T1SEs. Therefore, it appears essential and urgent to further reduce FPR in predictor development, not merely for T1SE, but also for all types of secreted proteins.

Currently, there is still a lack of computational methods predicting T1SEs (Hui et al., 2021). Although Luo et al. (2015) developed a random forest predictor, the tool or codes were not publically available and therefore a direct comparison could not be performed. An important factor that impedes development of prediction tools for T1SEs is the very limited number of experimentally validated T1SE proteins. Linhartova et al. (2010) and Luo et al. (2015) we in this research used Linhartova's RTX proteins as the positive dataset. In fact, we also used the validated T1SEs to build a similar model, and the performance was only slightly inferior to T1SEstacker but the variance was much larger among the cross-validated replicates. Moreover, the T1SEstacker could accurately predict the novel ones in the validated effector dataset at a high specificity. Therefore, we presented the T1SEstacker based on Linhartova's RTX proteins finally. With T1SEstacker and *Salmonella* strains, we also made estimation on the distribution of T1SEs. Roughly, there could be ~ 100 T1SEs in each bacterial strain. Therefore, the current T1SEs and function of T1SSs could be largely underestimated and underinvestigated. We also found that the T1SE composition varied a lot among different bacterial strains, suggesting they could exert specific function for better fitting and bacterial survival. Therefore, it is of great significance to identify and

investigate the function of T1SEs for both microbiologists and computational biologists.

Very few T1SEs have been validated from *Salmonella* spp., and SiIE represents the most well-known one, a large non-fimbrial adhesin of 600 kDa consisting of 53 repeats of Ig domains, which is encoded in an T1SS operon within *Salmonella* Pathogenicity Island 4 (SPI-4) of *S. enterica* strains (Gerlach et al., 2007; Barlag and Hensel, 2015; Klingl et al., 2020). We found that it was conserved in 19 out of the total 26 *Salmonella* strains (ID: 19CG0093; **Supplementary Dataset 3**). Interestingly, the gene was also detected from *S. bongori* besides all the seven subspecies of *S. enterica*. However, for *S. bongori*, *S. enterica* subsp. *diarizonae*, *indica*, and *enterica*, there were always representative strains missing the gene (**Supplementary Dataset 3**). More efforts should be placed to check whether there is the gene but mis-annotated or the gene has been actually lost. If the gene is lost, it is also interesting to know how its function is complemented in the corresponding strains. In this research, we also provided a list of possible T1SE candidates and their distribution among *Salmonella* spp., which comprise a valuable resource for the research community to further investigate *Salmonella* T1SEs and their function in bacterial pathogenicity.

T1SEstacker is one of the earliest machine-learning models predicting T1SEs. The performance requires further assessment and improvement. In this study, only sequence-derived features of T1SEs were analyzed and learned. Integration of other features such as the genomic context, i.e., proximity of the candidate genes to those encoding secretion components (Glaser et al., 1988; Welch and Pellett, 1988; Welch, 1991), common motifs located in promoters for transcription co-regulation (Mukherjee et al., 2015), physiochemical properties of proteins (Welch et al., 1983), and so on, may be helpful in improving the prediction performance. In addition, T1SS type-specific or species-specific substrate feature analysis and model development could further improve the precision of prediction. Despite the functional relevance, what we have known on T1SSs and T1SEs remains much fewer than unknowns (Alav et al., 2021). It remains a big challenge for computational biologists to make thorough and systematic analysis of T1SE features and develop more effective prediction models.

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

YW conceived and designed the project. ZC, ZZ, and YW developed the models and evaluated the performance. XH developed the web server. JZ, YXH, RC, XC, and YMH annotated the data and performed the analysis. All authors participated in the manuscript preparation.

FUNDING

This project was supported by the Funds for Medical Bioinformatics Youth Innovation Team of Shenzhen University (406/0000080805) and Natural Science Funds of Shenzhen (JCYJ201607115221141 and JCYJ20190808165205582). ZC and XH were supported by Special Funds for the Cultivation of Guangdong College Students' Scientific and Technological Innovation, Climbing Program (pdjha0427). ZC was supported by a National Undergraduate Training Program of China for Innovation and Entrepreneurship (no. 201910590003).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmicb.2021.813094/full#supplementary-material>

REFERENCES

- Alav, I., Kobylka, J., Kuth, M. S., Pos, K. M., Picard, M., Blair, J. M. A., et al. (2021). Structure, assembly, and function of tripartite efflux and type 1 secretion systems in gram-negative bacteria. *Chem. Rev.* 121, 5479–5596. doi: 10.1021/acs.chemrev.1c00055
- Almagro Armenteros, J. J., Tsirigos, K. D., Sonderby, C. K., Petersen, T. N., Winther, O., Brunak, S., et al. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* 37, 420–423. doi: 10.1038/s41587-019-0036-z
- Barlag, B., and Hensel, M. (2015). The giant adhesin SiiE of *Salmonella enterica*. *Molecules* 20, 1134–1150. doi: 10.3390/molecules20011134
- Boyd, C. D., Smith, T. J., El-Kirat-Chatel, S., Newell, P. D., Dufrene, Y. F., and O'Toole, G. A. (2014). Structural features of the *Pseudomonas fluorescens* biofilm adhesin LapA required for LapG-dependent cleavage, biofilm formation, and cell surface localization. *J. Bacteriol.* 196, 2775–2788. doi: 10.1128/JB.01629-14
- Delepelaire, P. (2004). Type I secretion in gram-negative bacteria. *Biochim. Biophys. Acta* 1694, 149–161.
- Felmlee, T., Pellett, S., and Welch, R. A. (1985). Nucleotide sequence of an *Escherichia coli* chromosomal hemolysin. *J. Bacteriol.* 163, 94–105.
- Gerlach, R. G., Jäckel, D., Stecher, B., Wagner, C., Lupas, A., Hardt, W. D., et al. (2007). *Salmonella* Pathogenicity Island 4 encodes a giant non-fimbrial adhesin and the cognate type 1 secretion system. *Cell Microbiol.* 9, 1834–1850. doi: 10.1111/j.1462-5822.2007.00919.x
- Glaser, P., Sakamoto, H., Bellalou, J., Ullmann, A., and Danchin, A. (1988). Secretion of cycloolysin, the calmodulin-sensitive adenylate cyclase-haemolysin bifunctional protein of *Bordetella pertussis*. *EMBO J.* 7, 3997–4004.
- Guo, S., Vance, T. D. R., Stevens, C. A., Voets, I., and Davies, P. L. (2019). RTX Adhesins are key bacterial surface megaproteins in the formation of biofilms. *Trends Microbiol.* 27, 453–467.
- Holland, I. B., Schmitt, L., and Young, J. (2005). Type 1 protein secretion in bacteria, the ABC-transporter dependent pathway. *Mol. Membr. Biol.* 22, 29–39. doi: 10.1080/09687860500042013
- Huang, B., Troese, M. J., Howe, D., Ye, S., Sims, J. T., Heinzen, R. A., et al. (2010). Anaplasma phagocytophilum APH_0032 is expressed late during infection and localizes to the pathogen-occupied vacuolar membrane. *Microb. Pathog.* 49, 273–284. doi: 10.1016/j.micpath.2010.06.009
- Hui, X., Chen, Z., Lin, M., Zhang, J., Hu, Y., Zeng, Y., et al. (2020). T3SEpp: an integrated prediction pipeline for bacterial type III secreted effectors. *mSystems* 5, e00288–20. doi: 10.1128/mSystems.00288-20
- Hui, X., Chen, Z., Zhang, J., Lu, M., Cai, X., Deng, Y., et al. (2021). Computational prediction of secreted proteins in gram-negative bacteria. *Comput. Struct. Biotechnol. J.* 19, 1806–1828.
- Hui, X., Hu, Y., Sun, M. A., Shu, X., Han, R., Ge, Q., et al. (2017). EBT: a statistic test identifying moderate size of significant features with balanced power and precision for genome-wide rate comparisons. *Bioinformatics* 33, 2631–2641. doi: 10.1093/bioinformatics/btx294
- Kanonenberg, K., Schwarz, C. K., and Schmitt, L. (2013). Type I secretion systems - a story of appendices. *Res. Microbiol.* 164, 596–604. doi: 10.1016/j.resmic.2013.03.011
- Kanonenberg, K., Spitz, O. I., Erenburg, N., Beer, T., and Schmitt, L. (2018). Type I secretion system-it takes three and a substrate. *FEMS Microbiol. Lett.* 365:fny094. doi: 10.1093/femsle/fny094
- Klingl, S., Kordes, S., Schmid, B., Gerlach, R. G., Hensel, M., and Muller, Y. A. (2020). Recombinant protein production and purification of SiiD, SiiE and SiiF - Components of the SPI4-encoded type I secretion system from *Salmonella Typhimurium*. *Protein Expr. Purif.* 172:105632. doi: 10.1016/j.pep.2020.105632
- Koronakis, V., Koronakis, E., and Hughes, C. (1989). Isolation and analysis of the C-terminal signal directing export of *Escherichia coli* hemolysin protein across both bacterial membranes. *EMBO J.* 8, 595–605.
- Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158
- Linhartova, I., Bumba, L., Masin, J., Basler, M., Osicka, R., Kamanova, J., et al. (2010). RTX proteins: a highly diverse family secreted by a common mechanism. *FEMS Microbiol. Rev.* 34, 1076–1112. doi: 10.1111/j.1574-6976.2010.00231.x
- Luo, J., Li, W., Liu, Z., Guo, Y., Pu, X., and Li, M. (2015). A sequence-based two-level method for the prediction of type I secreted RTX proteins. *Analyst* 140, 3048–3056. doi: 10.1039/c5an00311c
- Magnan, C. N., and Baldi, P. (2014). SSpro/ACCpro5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. *Bioinformatics* 30, 2592–2597. doi: 10.1093/bioinformatics/btu352
- Masure, H. R., Au, D. C., Gross, M. K., Donovan, M. G., and Storm, D. R. (1990). Secretion of the *Bordetella pertussis* adenylate cyclase from *Escherichia coli* containing the hemolysin operon. *Biochemistry* 29, 140–145. doi: 10.1021/bi00453a017
- Mukherjee, D., Pal, A., Chakravarty, D., and Chakrabarti, P. (2015). Identification of the target DNA sequence and characterization of DNA binding features of HlyU, and suggestion of a redox switch for hlyA expression in the human pathogen *Vibrio cholerae* from in silico studies. *Nucleic Acids Res.* 43, 1407–1417. doi: 10.1093/nar/gku1319
- Noegel, A., Rdest, U., Springer, W., and Goebel, W. (1979). Plasmid cistrons controlling synthesis and excretion of the exotoxin alpha-haemolysin of *Escherichia coli*. *Mol. Gen. Genet.* 175, 343–350. doi: 10.1007/BF00397234

- Park, Y., Eom, G. T., Oh, J. Y., Park, J. H., Kim, S. C., Song, J. K., et al. (2020). High-level production of bacteriotoxic phospholipase A1 in bacterial host *Pseudomonas fluorescens* via ABC transporter-mediated secretion and inducible expression. *Microorganisms* 8:239. doi: 10.3390/microorganisms8020239
- Ryu, J., Lee, U., Park, J., Yoo, D. H., and Ahn, J. H. (2015). A vector system for ABC transporter-mediated secretion and purification of recombinant proteins in *Pseudomonas* species. *Appl. Environ. Microbiol.* 81, 1744–1753. doi: 10.1128/AEM.03514-14
- Schwarz, C. K. W., Landsberg, C. D., Lenders, M. H. H., Smits, S. H. J., and Schmitt, L. (2012). Using an *E. coli* Type 1 secretion system to secrete the mammalian, intracellular protein IFABP in its active form. *J. Biotechnol.* 159, 155–161. doi: 10.1016/j.jbiotec.2012.02.005
- Smith, T. J., Sondermann, H., and O'Toole, G. A. (2018b). Type 1 does the two-step: type 1 secretion substrates with a functional periplasmic intermediate. *J. Bacteriol.* 200, e00168–18. doi: 10.1128/JB.00168-18
- Smith, T. J., Font, M. E., Kelly, C. M., Sondermann, H., and O'Toole, G. A. (2018a). An N-Terminal retention module anchors the giant adhesin LapA of *Pseudomonas fluorescens* at the cell surface: a novel subfamily of type I secretion systems. *J. Bacteriol.* 200, e00734–17. doi: 10.1128/JB.00734-17
- Son, M., Moon, Y., Oh, M. J., Han, S. B., Park, K. H., Kim, J. G., et al. (2012). Lipase and protease double-deletion mutant of *Pseudomonas fluorescens* suitable for extracellular protein production. *Appl. Environ. Microbiol.* 78, 8454–8462. doi: 10.1128/AEM.02476-12
- Spitz, O., Erenburg, N. I., Beer, T., Kanonenberg, K. I., Holland, B., and Schmitt, L. (2019). Type I secretion systems—one mechanism for all? *Microbiol. Spectr.* 7:PSIB-0003-2018. doi: 10.1128/microbiolspec.PSIB-0003-2018
- Thomas, S. I., Holland, B., and Schmitt, L. (2014). The Type 1 secretion pathway – the hemolysin system and beyond. *Biochim. Biophys. Acta* 1843, 1629–1641. doi: 10.1016/j.bbamcr.2013.09.017
- Wakeel, A., den Dulk-Ras, A., Hooykaas, P. J. J., and McBride, J. W. (2011). *Escherichia chaffeensis* tandem repeat proteins and Ank200 are type 1 secretion system substrates related to the repeats-in-toxin exoprotein family. *Front. Cell. Infect. Microbiol.* 1:22. doi: 10.3389/fcimb.2011.00022
- Wang, J., Yang, B., An, Y., Marquez-Lago, T., Leier, A., Wilksch, J., et al. (2019). Systematic analysis and prediction of type IV secreted effector proteins by machine learning approaches. *Brief Bioinform.* 20, 931–951. doi: 10.1093/bib/bbx164
- Wang, J., Yang, B., Leier, A., Marquez-Lago, T. T., Hayashida, M., Rocker, A., et al. (2018). Bastion6: a bioinformatics approach for accurate prediction of type VI secreted effectors. *Bioinformatics* 34, 2546–2555. doi: 10.1093/bioinformatics/bty155
- Wang, Y., Sun, M., Bao, H., and White, A. P. (2013). T3_MM: a Markov model effectively classifies bacterial type III secretion signals. *PLoS One* 8:e58173. doi: 10.1371/journal.pone.0058173
- Wang, Y., Wei, X., Bao, H., and Liu, S. L. (2014). Prediction of bacterial type IV secreted effectors by C-terminal features. *BMC Genomics* 15:50. doi: 10.1186/1471-2164-15-50
- Wang, Y., Zhang, Q., Sun, M. A., and Guo, D. (2011). High-accuracy prediction of bacterial type III secreted effectors based on position-specific amino acid composition profiles. *Bioinformatics* 27, 777–784. doi: 10.1093/bioinformatics/btr021
- Welch, R. A. (1991). Pore-forming cytotoxins of gram-negative bacteria. *Mol. Microbiol.* 5, 521–528. doi: 10.1111/j.1365-2958.1991.tb00723.x
- Welch, R. A., Hull, R., and Falkow, S. (1983). Molecular cloning and physical characterization of a chromosomal hemolysin from *Escherichia coli*. *Infect. Immun.* 42, 178–186. doi: 10.1128/iai.42.1.178-186.1983
- Welch, R. A., and Pellett, S. (1988). Transcriptional organization of the *Escherichia coli* hemolysin genes. *J. Bacteriol.* 170, 1622–1630.
- Xue, L., Tang, B., Chen, W., and Luo, J. (2019). DeepT3: deep convolutional neural networks accurately identify Gram-negative bacterial type III secreted effectors using the N-terminal sequence. *Bioinformatics* 35, 2051–2057. doi: 10.1093/bioinformatics/bty931
- Zhang, F., Yin, Y., Arrowsmith, C. H., and Ling, V. (1995). Secretion and circular dichroism analysis of the C-terminal signal peptides of HlyA and LktA. *Biochemistry* 34, 4193–4201. doi: 10.1021/bi00013a007

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.

Copyright © 2022 Chen, Zhao, Hui, Zhang, Hu, Chen, Cai, Hu and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



A Comparative Study of Deep Learning Classification Methods on a Small Environmental Microorganism Image Dataset (EMDS-6): From Convolutional Neural Networks to Visual Transformers

Peng Zhao¹, Chen Li^{1*}, Md Mamunur Rahaman¹, Hao Xu¹, Hechen Yang¹, Hongzan Sun², Tao Jiang^{3*} and Marcin Grzegorzczek⁴

¹ Microscopic Image and Medical Image Analysis Group, College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China, ² Shengjing Hospital, China Medical University, Shenyang, China, ³ School of Control Engineering, Chengdu University of Information Technology, Chengdu, China, ⁴ Institute of Medical Informatics, University of Lübeck, Lübeck, Germany

OPEN ACCESS

Edited by:

Mohammad-Hossein Sarrafzadeh,
University of Tehran, Iran

Reviewed by:

Kaveh Kavousi,
University of Tehran, Iran
Hashem Asgharnejad,
Polytechnique Montréal, Canada

*Correspondence:

Chen Li
lichen201096@hotmail.com
Tao Jiang
jiang@cuit.edu.cn

Specialty section:

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

Received: 09 October 2021

Accepted: 02 February 2022

Published: 02 March 2022

Citation:

Zhao P, Li C, Rahaman MM, Xu H, Yang H, Sun H, Jiang T and Grzegorzczek M (2022) A Comparative Study of Deep Learning Classification Methods on a Small Environmental Microorganism Image Dataset (EMDS-6): From Convolutional Neural Networks to Visual Transformers. *Front. Microbiol.* 13:792166. doi: 10.3389/fmicb.2022.792166

In recent years, deep learning has made brilliant achievements in *Environmental Microorganism* (EM) image classification. However, image classification of small EM datasets has still not obtained good research results. Therefore, researchers need to spend a lot of time searching for models with good classification performance and suitable for the current equipment working environment. To provide reliable references for researchers, we conduct a series of comparison experiments on 21 deep learning models. The experiment includes direct classification, imbalanced training, and hyper-parameters tuning experiments. During the experiments, we find complementarities among the 21 models, which is the basis for feature fusion related experiments. We also find that the data augmentation method of geometric deformation is difficult to improve the performance of VTs (ViT, DeiT, BotNet, and T2T-ViT) series models. In terms of model performance, Xception has the best classification performance, the vision transformer (ViT) model consumes the least time for training, and the ShuffleNet-V2 model has the least number of parameters.

Keywords: deep learning, convolutional neural network, visual transformer, image classification, small dataset, environmental microorganism

1. INTRODUCTION

With the advancement of industrialization, industrial pollution becomes increasingly serious. Therefore, finding effective methods to control, reduce, or eliminate pollution is a top priority. Biological approaches have outstanding performance in solving environmental pollution problems. The Biological approaches have four main advantages in environmental treatment: no new pollution, no additional energy consumption, gentle process, decomposition products can feedback to nature, and make a virtuous cycle of material changes (McKinney, 2004). Microorganisms are all tiny creatures that are invisible to the naked eyes. They are tiny and simple in structure, and usually can only be seen with a microscope. *Environmental Microorganisms* (EMs) specifically refer to those species of microorganisms that live in natural environments (such as mountains,

streams, and oceans) and artificial environments (such orchards and fish ponds). EMs play a vital role in whole nature for better or worse. For example, lactic acid bacteria can decompose some organic matter in the natural environment to provide nutrients for plants; actinomycetes can digest organic waste in sludge and improve water quality; microalgae can fix carbon dioxide in the air and be used as a raw material for biodiesel (Zhao et al., 2021); activated sludge composed of microorganisms has a strong ability to adsorb and oxidize organic matter and purify water (Asgharnejad and Sarrafzadeh, 2020). Harmful rhizosphere bacteria can inhibit plant growth by producing phytotoxins (Fried et al., 2000). Sludge bulking is caused by bacterial proliferation and the accumulation of sticky material, which poses a fundamental challenge for wastewater treatment (Fan et al., 2017). Therefore, EMs research helps solve environmental pollution problems, and the classification of EMs is the cornerstone of related research (Kosov et al., 2018).

Generally, the size of EMs is between 0.1 and 100 μm , which is challenging to be identified and found. The traditional microbial classification method typically uses the “morphological method,” which requires a skilled operator to observe the EMs under a microscope. Then the results are given according to the shape characteristics. This is very time-consuming and financial (Pepper et al., 2011). In addition, if researchers do not refer to the literature, even very experienced researchers cannot guarantee the accuracy and objectivity of the analysis results. Therefore, using the computer-aided classification of EM images can enable researchers to use the slightest professional knowledge and the least time to make the most accurate judgments.

Currently, the analysis of EMs by computer vision is already achieved. For example, RGB (Red, Green, Blue) color analysis measures the number of microorganisms (Filzmoser and Todorov, 2011; Sarrafzadeh et al., 2015), and deep learning methods are used to achieve the classification and segmentation of EM images. Among them, the research of EM classification using deep learning methods obtains more and more attention. Deep learning is a new research direction in the field of machine learning, and it provides good performance for image classification (Zhang et al., 2020). Traditional machine learning-based EM classification methods rely on feature extraction, which requires many human resources (Çayır et al., 2018). In contrast, deep learning-based algorithms perform feature extraction in an automated manner, allowing researchers to use minimal domain knowledge and workforce to extract prominent features. Furthermore, the classification results of deep learning are better than that of traditional machine learning in the case of super-large training samples (Wang et al., 2021). However, for small datasets, the performance of deep learning is limited. Because the collection of EMs is usually carried out outdoors, for some sensitive EMs, transportation, storage, and observation during the period may affect the quality of the final images. Therefore, it is difficult to obtain enough high-quality images, and this case results in the problem of small datasets. Therefore, this paper compares the performance of various deep learning models on small data sets of EMs and aims to find models with better performance on small data sets.

This article compares a series of Convolutional Neural Networks (CNNs), such as ResNet-18, 34, 50, 101 (He et al., 2016), VGG11, 13, 16, 19 (Simonyan and Zisserman, 2014), DenseNet-121, 169 (Huang et al., 2017), Inception-V3 (Szegedy et al., 2016), Xception (Chollet, 2017), AlexNet (Krizhevsky et al., 2012), GoogleNet (Szegedy et al., 2015), MobileNet-V2 (Sandler et al., 2018), ShuffleNet-V2x0.5 (Ma et al., 2018), Inception-ResNet-V1 (Szegedy et al., 2017), and a series of visual transformers (VTs), such as vision transformer (ViT) (Dosovitskiy et al., 2020), BotNet (Srinivas et al., 2021), DeiT (Touvron et al., 2020), T2T-ViT (Yuan et al., 2021). The purpose is to find deep learning models that are suitable for EM small datasets. The workflow diagram of this study is shown in **Figure 1**. Step (b) is to rotate the training set and validation set images by 90°, 180°, 270°, and mirror images up and down, left and right, augment the dataset by six times. Step (c) is uniform image size to 224×224 to facilitate training and classification. Step (d) is to input the processed data into different network models for training. Step (e) is to input the test set into the trained network for classification, and step (f) is to calculate the *Average Precision* (AP), accuracy, precision, recall, and F1-score based on the classification results to evaluate the performance of the network model.

The structure of this paper is as follows. Section 2 introduces the related methods of deep learning in image classification, the impact of small datasets on image classification, and the related work of deep learning models. Section 3 introduces the dataset and experimental design in detail. Section 4 compares and summarizes the experimental results. Section 5 summarizes the whole paper and looks forward.

2. RELATED WORK

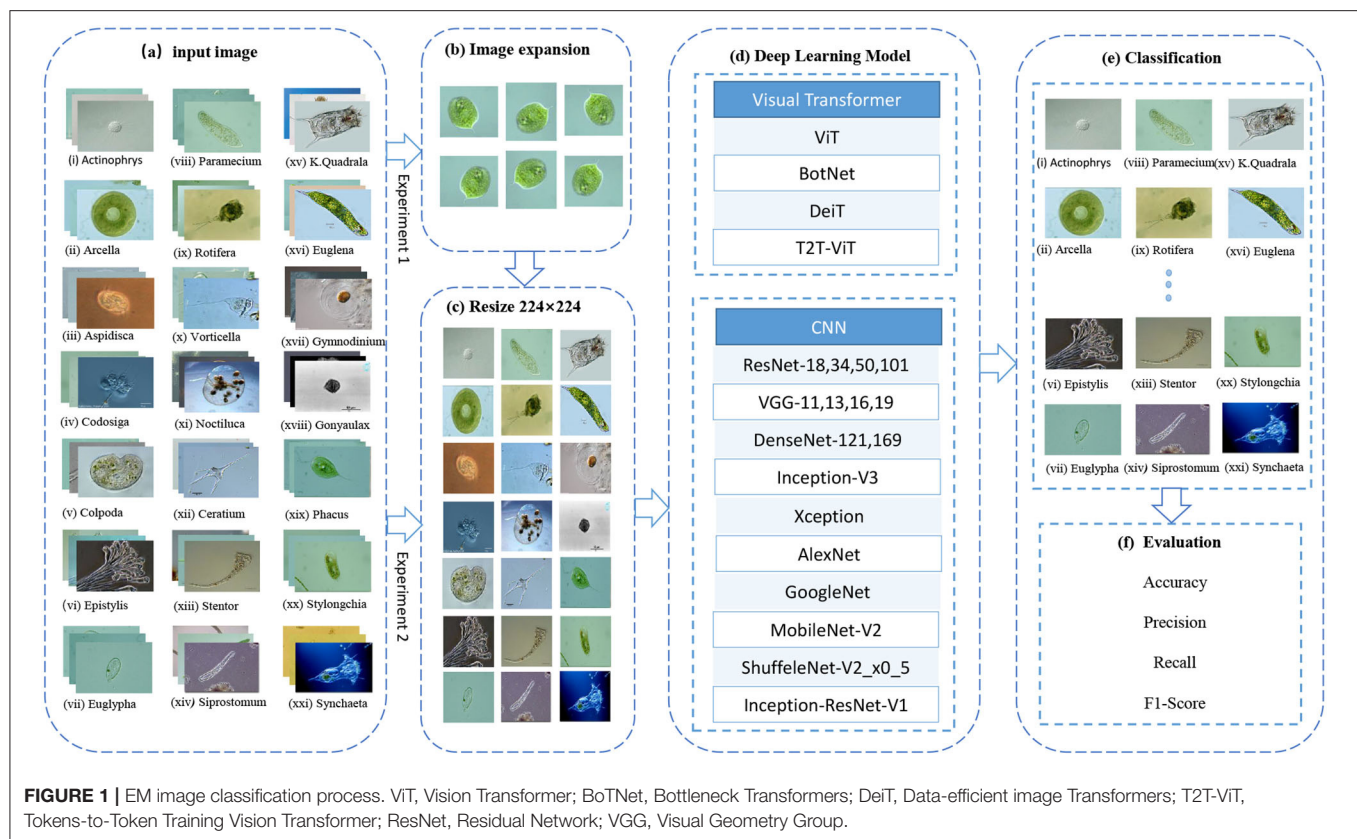
This section summarizes the impact of small datasets on classification, including basic deep learning image classification methods.

2.1. The Impact of Small Datasets on Image Classification

In rectal histopathology deep learning classification research, a large number of labeled pathological images are needed. However, the preparation of large datasets requires expensive labor costs and time costs, leading to the fact that existing studies are primarily based on small datasets. In addition, the lack of sufficient data leads to overfitting problems during the training process. A conditional sliding windows arithmetic is proposed in Haryanto et al. (2021) to solve this problem, which generates histopathological images. This arithmetic successfully solves the limitation of rectal histopathological data.

In climate research, the use of deep learning in cloud layer analysis often requires a lot of data. Therefore, classification in the case of a small dataset cannot achieve higher accuracy. In order to solve this problem, a classification model with high accuracy on small datasets is proposed. The method improves from three aspects:

1. A network model for a small dataset is designed.



2. A regularization technique to increase the generalization ability of the model is applied.

3. The average ensemble of models is used to improve the classification accuracy.

Therefore, the model not only has higher accuracy but also has better robustness (Phung and Rhee, 2019).

In deep learning research, small datasets often lead to classification over-fitting and low classification accuracy. According to this problem, a kind of deep CNN based transfer learning is designed to solve the problem of the small dataset. This method mainly improves data and models. In terms of data, the model transfers the feature layer of the CNN model pre-trained on big sample dataset to a small sample dataset. In terms of model, the whole series average pooling is used instead of the fully connected layer, and Softmax is used for classification. This method has a good classification performance on small sample datasets (Zhao, 2017).

Because of the limited training data, a two-phase classification method using migration learning and web data augmentation technology is proposed. This method increases the number of samples in the training set through network data augmentation. In addition, it reduces the requirements on the number of samples through transfer learning. This classifier reduces the over-fitting problem while improving the generalization ability of the network (Han et al., 2018).

In radar image recognition, due to the complex environment and particular imaging principles, Synthetic Aperture Radar

(SAR) images have the problem of sample scarcity. A target recognition method of SAR image based on Constrained Naive Generative Adversarial Networks and CNN is proposed to solve this problem. This method combines Least Squares Generative Adversarial Networks and designs a shallow network structure based on the traditional CNNs model. The problem of high model complexity and over-fitting caused by the deep network structure is avoided, to improve the recognition performance. This method can better solve the problems of few image samples and intense speckle noise (Mao et al., 2021).

Lack of sufficient training data can seriously deteriorate the performance of neural networks and other classifiers. Due to this problem, a self-aware multi-classifier system suitable for “small data” cases is proposed. The system uses Neural Network, Support Vector Machines (SVMs) and Naive Bayes models as component classifiers. In addition, this system uses the confidence level as a criterion for classifier selection. The system performs well in various test cases and is incredibly accurate on small datasets (Kholerdi et al., 2018).

Convolutional Neural Networks are very effective for face recognition problems, but training such a network requires a large number of labeled images. Such large datasets are usually not public and challenging to collect. According to this situation, a method based on authentic face images to synthesize a vast training set is proposed. This method swaps the facial components of different face images to generate a new face. This technology achieves the most advanced face recognition

performance on the Labeled Faces in the Wild (LFW) face database (Hu et al., 2017).

The effectiveness of tuning the number of convolutional layers to classify small datasets is proven in Chandrarathne et al. (2020). In addition, related experiments suggest that by employing a very low learning rate (LR), the accuracy of classification of small datasets can be greatly increased.

In medical signal processing, very small datasets often lead to the problems of model overfitting and low classification accuracy. According to this situation, a method combining deep learning and traditional machine learning is proposed. This method uses the first few layers of CNN for feature extraction. Then, the extracted features are fed back to traditional supervised learning algorithms for classification. This method can avoid the overfitting problem caused by small datasets. In addition, it has better performance than traditional machine learning methods (Alabandi, 2017).

2.2. Deep Learning Techniques

Due to the excellent performance of AlexNet in the image classification competition (Krizhevsky et al., 2012), improvements in the CNN architecture are very active. A series of CNN-based networks continue to appear, making CNN an irreplaceable mainstream method in the field of computer vision. In recent years, Transformer frequently appears in computer vision tasks and provides good performance, which is sufficient to attract the attention of researchers.

2.2.1. Convolutional Neural Networks

AlexNet is the first large-scale CNN architecture to perform well in ImageNet classification. The innovation of the network lies in the successful application of the Rectified Linear Unit (Relu) activation function and the use of the Dropout mechanism and data enhancement strategy to prevent overfitting. To improve the model generalization ability, the network uses a Local Response Normalization layer. In addition, the maximum pooling of overlap is used to avoid the blurring effect caused by average pooling (Krizhevsky et al., 2012).

The Visual Geometry Group of Oxford proposes the VGG network. The network uses a deeper network structure with depths of 11, 13, 16, and 19 layers. Meanwhile, VGG networks use a smaller convolution kernel (3×3 pixels) instead of the larger convolution kernel, which reduces the parameters and increases the expressive power of the networks (Simonyan and Zisserman, 2014).

GoogLeNet is a deep neural network model based on the Inception module launched by Google. The network introduces an initial structure to increase the width and depth of the network while removing the fully connected layer and using average pooling instead of the fully connected layer to avoid the disappearance of the gradient. The network adds two additional softmax to conduct the gradient forward (Szegedy et al., 2015).

ResNet solves the “degradation” problem of deep neural networks by introducing residual structure. ResNet networks use multiple parameter layers to learn the representation of residuals between input and output, rather than using parameter layers to directly try to learn the mapping between input and output as

VGGs networks do. Residual networks are characterized by ease of optimization and the ability to improve accuracy by adding considerable depth (He et al., 2016).

The DenseNet network is inspired by the ResNet network. DenseNet uses a dense connection mechanism to connect all layers. This connection method allows the feature map learned by each layer to be directly transmitted to all subsequent layers as input, so that the features and the transmission of the gradient is more effective, and the network is easier to train. The network has the following advantages: it reduces the disappearance of gradients, strengthens the transfer of features, makes more effective use of features, and reduces the number of parameters to a certain extent (Huang et al., 2017).

The inception-V3 network is mainly improved in two aspects. Firstly, branch structure is used to optimize the Inception Module; secondly, the larger two-dimensional convolution kernel is unpacked into two one-dimensional convolution kernels. This asymmetric structure can deal with more and richer spatial information and reduce the computation (Szegedy et al., 2016).

Xception is an improvement of Inception-V3. The network proposes a novel Depthwise Separable Convolution align them in column, the core idea of which lies in space transformation and channel transformation. Compared with Inception, Xception has fewer parameters and is faster (Chollet, 2017).

MobileNets and Xception have the same ideas but different pursuits. Xception pursues high precision, but MobileNets is a lightweight model, pursuing a balance between model compression and accuracy. A new unit Inverted residual with linear bottleneck is applied in MobileNet-V2. The inverse residual first increases the number of channels, then performs convolution and then increases the number of channels. This can reduce memory consumption (Sandler et al., 2018).

ShuffleNet makes some improvements based on MobileNet. The 1×1 convolution used by MobileNet is a traditional convolution method with a lot of redundancy. However, ShuffleNet performs shuffle and group operations on 1×1 convolution. This operation implements channel shuffle and pointwise group convolution. In addition, this operation dramatically reduces the number of model calculations while maintaining accuracy (Ma et al., 2018).

The Inception-ResNet network is inspired by ResNet, which introduces the residual structure of ResNet in the Inception module. Adding the residual structure does not significantly improve the model effect. But the residual structure helps to speed up the convergence and improve the calculation efficiency. The calculation amount of Inception-ResNet-v1 is the same as that of Inception-V3, but the convergence speed is faster (Szegedy et al., 2017).

2.2.2. Visual Transformers

The ViT model applies transformers in the field of natural language processing to the field of computer vision. The main contribution of this model is to prove that CNN is not the only choice for image classification tasks. Vision transformer divides the input image into fixed-size patches and then obtains patch embedding through a linear transformation. Finally, the

patch embeddings of the image are sent to the transformer to perform feature extraction to classification. The model is more effective than CNN on super-large-scale datasets and has high computational efficiency (Dosovitskiy et al., 2020).

The BoTNet is proposed by Srinivas. This network introduces self-attention into ResNet. Therefore, BoTNet has both the local perception ability of CNN and the global information acquisition ability of Transformer. The top-1 accuracy on ImageNet is as high as 84.7%, and the performance is better than models such as SENet and Efficient-Net (Srinivas et al., 2021).

T2T-ViT is an upgraded version of ViT. It proposes a novel Tokens-to-Token mechanism based on the characteristics and structure of ViT. This mechanism allows the deep learning model to model local and global information. The performance of this model is better than ResNet in the ImageNet data test, and the number of parameters and calculations are significantly reduced. In addition, the performance of its lightweight model is better than that of MobileNet (Yuan et al., 2021).

DeiT is proposed by Touvron et al. The innovation of DeiT is proposes a new distillation process based on a distillation token, which has the same function as a class token. It is a token added after the image block sequence. The output after the transformer encoder and the output of the teacher model calculates the loss together. The training of DeiT requires fewer data and fewer computing resources (Touvron et al., 2020).

2.3. EM Image Classification

With the development of technology, good results are achieved using computer-aided EM classification. In Kruk et al. (2015), a system for automatic identification of different species of microorganisms in soil is proposed. The system first separates microorganisms from the background using the Otsu. Then shape features, edge features, and color histogram features are extracted. Then the features are filtered using a fast correlation-based filter. Finally, the random forest (RF) classifier is used for classification. This system frees researchers from the tedious task of microbial observation.

In Amaral et al. (1999), a semi-automatic microbial identification system is proposed. The system can accurately identify seven species of protozoa commonly found in wastewater. The system first enhances the image to be processed and then undergoes data collection and complex morphological operations to generate a 3D model of EMs. The 3D model is used to determine the species of protozoa. In Amaral et al. (2008), a semi-automatic image analysis procedure is proposed. It is found that geometric features have good recognition ability. It is possible to detect the presence of two microorganisms, Opercularia and Vorticella, in wastewater plants. In Chen and Li (2008), an improved neural network classification method based on microscopic images of sewage bacteria is proposed. The method uses principal component analysis to reduce the extracted EM features. Also, the method applies the daptive accelerated back propagation (BP) algorithm to learn image classification.

An automatic classification method with high robustness of EMs is suggested in Li et al. (2013), which describes the shape of EMs in microscopic images by Edge Histograms, Extended

Geometrical Features, etc. The support vector machine classifier is used to achieve the best classification result of 89.7%. A shape-based method for EM classification is suggested in Yang et al. (2014), which introduces very robust two-dimensional feature descriptors for EM shapes. The main process of this method is to separate EMs from the background. Then a new EM feature descriptor is used and finally a SVM is used for classification.

A new method for automatic classification of bacterial colony images is proposed in Nie et al. (2015), which enables the classification of colonies in different growth stages and contexts. In addition, the method mainly uses a multilayer middle layer CNN model for classification and uses the patches segmented from the CDBN model as input. Finally, a voting scheme is used for prediction. The results show that the method achieves results that exceed the classical model.

3. MATERIALS AND METHODS

This section explains the EMDS-6 dataset, data augmentation methods, the distribution of the dataset, and the evaluation metrics for classification.

3.1. Dataset

3.1.1. Data Description

This experiment uses Environmental Microorganism Dataset 6th Version (EMDS-6) to compare model performance. The dataset contains a total of 840 EM images of different sizes. These images contain a total of 21 types of EMs, each with 40 images, namely: *Actinophrys*, *Arcella*, *Aspidisca*, *Codosiga*, *Colpoda*, *Epistylis*, *Euglypha*, *Paramecium*, *Rotifera*, *Vorticella*, *Noctiluca*, *Ceratium*, *Stentor*, *Siprostomum*, *K. Quadrata*, *Euglena*, *Gymnodinium*, *Gymlyano*, *Phacus*, *Stylongchia*, *Synchaeta*. Some examples are shown in **Figure 2** (Zhao et al., 2021).

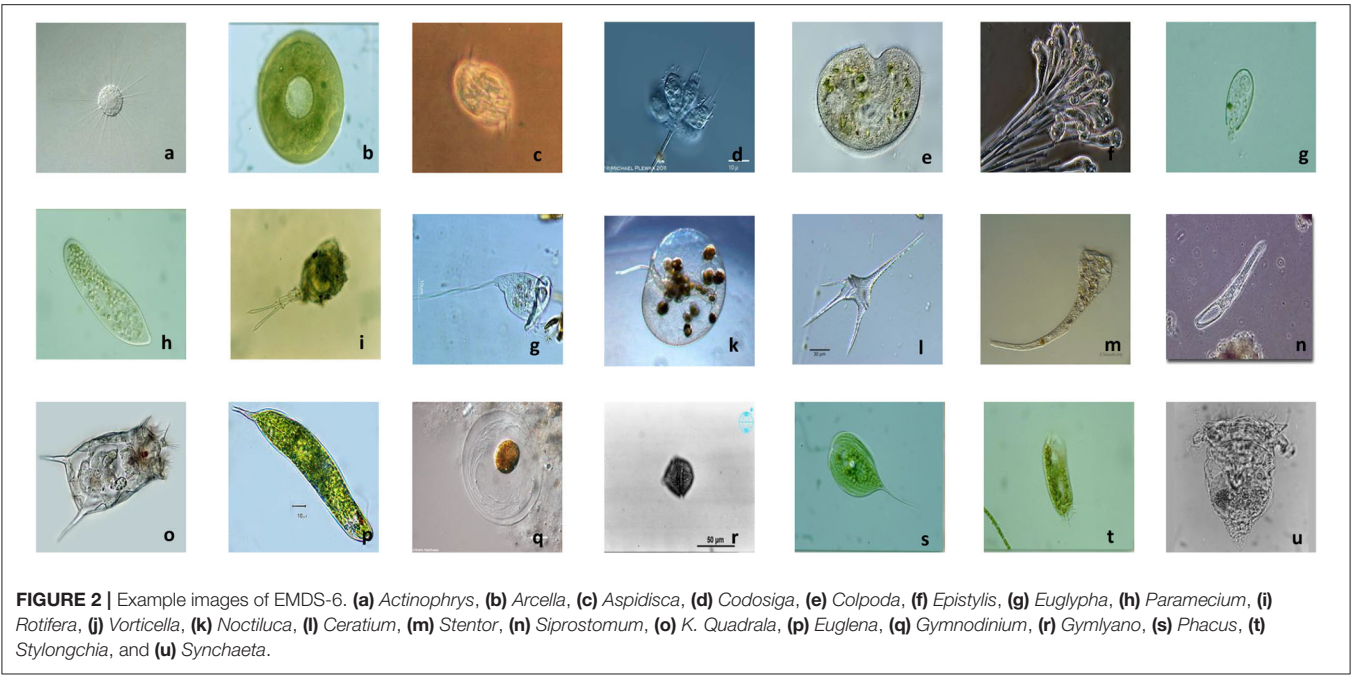
3.1.2. Data Preprocessing

In order to improve the accuracy of the model and reduce the degree of model overfitting, the images in EMDS-6 are augmented. Due to the security problem of data augmentation, the only geometric transformation of the data is performed here. The geometric transformation includes rotation 90°, 180°, and 270°, up and down mirroring, and left and right mirroring. These transformations do not break the EM label and ensure data security. In addition, the image sizes in EMDS-6 is inconsistent, but the input required by the deep learning models is the same. Therefore, all images in EMDS-6 are standardized to 224 × 224 pixels.

3.1.3. Data Settings

Experiment A: Randomly select 37.5% of the dataset as the training set, 25% as the validation set, and 37.5% as the test set. Experiment A is to directly perform classification tasks on 21 types of microorganisms through the deep learning model. The details of the training set, validation set, and test set are shown in **Table 1**.

Experiment B: Randomly select 37.5% of the dataset as the training set, 25% as the validation set, and 37.5% as the test set. Specifically, 21 types of microorganisms



are sequentially regarded as positive samples and the remaining 20 types of samples are regarded as negative samples. In this way, 21 new datasets are generated. For example, if *Actinophrys* images are used as positive samples, the remaining 20 types of EMs such as *Arcella* and *Aspidisca* are used as negative samples. Experiment B is imbalanced training to assist in verifying the performance of the model.

Because EMDS-6 is a very small dataset, the experimental results are quite contingent. Therefore, 37.5% of the data is used to test the performance of the model to increase the reliability of the experiment. This also expresses our sincerity to the experimental results.

3.2. Evaluation Methods

To scientifically evaluate the classification performance of deep learning models, choosing appropriate indicators is a crucial factor. Recall, Precision, Accuracy, F1-score, AP, and *mean Average Precision* (mAP) are commonly used evaluation indicators (Xie et al., 2015). The effectiveness of these indicators is proven. The Recall is the probability of being predicted to be positive in actual positive samples. Precision is the probability of being actual positive in all predicted positive samples. Average Precision refers to the average value of recall rate from 0 to 1. The mAP is the arithmetic average of all AP. F1-score is the harmonic value of precision rate and recall rate. Accuracy refers to the percentage of correct results predicted in the total sample (Powers, 2020). The specific calculation methods of these indicators are shown in Table 2.

In Table 2, TN is the number of negative classes predicted as negative classes, FP represents the number of negative classes predicted as positive classes, FN refers to the

TABLE 1 | Dataset details of EMDS-6.

Class\Dataset	Train	Val	Text	Total
<i>Actinophrys</i>	15	10	15	40
<i>Arcella</i>	15	10	15	40
<i>Aspidisca</i>	15	10	15	40
<i>Codosiga</i>	15	10	15	40
<i>Colpoda</i>	15	10	15	40
<i>Epistylis</i>	15	10	15	40
<i>Euglypha</i>	15	10	15	40
<i>Paramecium</i>	15	10	15	40
<i>Rotifera</i>	15	10	15	40
<i>Vorticella</i>	15	10	15	40
<i>Noctiluca</i>	15	10	15	40
<i>Ceratium</i>	15	10	15	40
<i>Stentor</i>	15	10	15	40
<i>Siprostomum</i>	15	10	15	40
<i>K. Quadrala</i>	15	10	15	40
<i>Euglena</i>	15	10	15	40
<i>Gymnodinium</i>	15	10	15	40
<i>Gonyaulax</i>	15	10	15	40
<i>Phacus</i>	15	10	15	40
<i>Stylongchia</i>	15	10	15	40
<i>Synchaeta</i>	15	10	15	40
Total	315	210	315	840

number of positive classes predicted as negative classes, and TP is the number of positive classes predicted as positive classes.

TABLE 2 | Evaluation metrics for image classification. Sample classification (K), number of positive samples (M).

Assessments	Formula
Precision (<i>P</i>)	$\frac{TP}{TP+FP}$
Recall (<i>R</i>)	$\frac{TP}{TP+FN}$
F1-score	$2 \times \frac{P \times R}{P+R}$
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
AP	$\frac{1}{M} \sum_{i=1}^M \text{Precision}_{\max}(i)$
mAP	$\frac{1}{K} \sum_{j=1}^K \text{AP}(j)$

TABLE 3 | Computer hardware configuration.

Hardware	Product number
CPU	Intel Core i7-10700
GPU	NVIDIA Quadro RTX 4000
Motherboard	HP 8750 (LPC Controller-0697)
RAM	SAMSUNG DDR4 3200MHz
SSD	HP SSD S750 256GB

TABLE 4 | Deep learning hyper-parameters.

Parameter	Value
Batch Size	32
Epoch	100
Learning	0.002
Optimizer	Adam

4. COMPARISON OF CLASSIFICATION EXPERIMENTS

4.1. Experimental Environment

This comparative experiment is performed on the local computer. The computer hardware configuration is shown in **Table 3**. The computer software configuration is as follows: Win10 Professional operating system, Python 3.6, and Pytorch 1.7.1. In addition, the code runs in the integrated development environment Pycharm 2020 Community Edition.

This experiment mainly uses some classic deep learning models and some relatively novel deep learning models. The hyper-parameters uniformly set by these models are shown in **Table 4**.

4.2. Experimental Results and Analysis

4.2.1. The Classification Performance of Each Model on the Training and Validation Sets

Figure 3 shows the accuracy and loss curves of the CNNs and VT series models. **Table 5** shows the performance indicators of different deep learning models on the validation set. According to **Figure 3** and **Table 5**, the performance of different deep learning models using small EM dataset cases is briefly evaluated.

As shown in **Figure 3**, the accuracy rate of the training set is much higher than that of the validation set of each

model. Densenet169, Googlenet, Mobilenet-V2, ResNet50, ViT, and Xception network models are particularly over-fitted. In addition, AlexNet, InceptionResnetV1, ShuffleNet-V2, and VGG11 network models do not show serious overfitting. Among 21 models in **Table 5**, the accuracy rates of the Deit, ViT, and T2T-ViT models are at the 10th, 12th, and 14th. The VT models are in the middle and downstream position among the 21 models.

The Xception network model has the highest accuracy, precision, and recall rates in the test set results, which are 40.32, 49.71, and 40.33%. The AlexNet, ViT, and ShuffleNet-V2 network models require the shortest training time, which are 711.64, 714.56, and 712.95 s. In addition, the ShuffleNet-V2 network model has the smallest parameter amount, which is 1.52 MB.

VGG16 and VGG19 networks cannot converge in EMDS-6 classification task. The VGG13 network model has the lowest accuracy, precision, and recall rates in the validation set results, which are 20.95, 19.23, and 20.95%. The VGG19 network model requires the longest training time, which is 1036.68 s. In addition, the VGG19 network model has the largest amount of parameters, which is 521 MB.

Xception is a network with excellent performance in EMDS-6 classification. In the Xception network accuracy curve, the accuracy of the Xception network training set is rising rapidly, approaching the highest point of 90% after 80 epochs. Meanwhile, the accuracy of the validation set is close to the highest point 45%, after 30 epochs. In addition, the Xception network training set loss curve declines steadily and approaches its lowest point after 80 epochs. But the validation set loss begins to approach the lowest point after 20 epochs and stops falling. VGG13 is a network that performs poorly on EMDS-6 classification. In the VGG13 network, the accuracy curve of the training set and the accuracy curve of the validation set have similar trends, and there are obvious differences after 80 epochs. Meanwhile, the loss of the training set and the loss of the validation set are also relatively close, and there are obvious differences after 60 epochs. Networks such as Xception, ResNet34, and Googlenet are relatively high-performance networks. The training accuracy of these networks is much higher than the validation accuracy. Furthermore, the validation accuracy is close to the highest point in a few epochs. In addition, the training set loss of these networks is usually lower than 0.3 at 100 epochs. VGG11 and AlexNet are poorly performing networks. These network training accuracy curves are relatively close to the validation accuracy curves. Disagreements usually occur after many epochs. In addition, the training set loss of these networks is usually higher than 0.3 at 100 epochs.

4.2.2. The Classification Performance of Each Model on Test Set

Table 6, shows the performance indicators of each model on the test set, including precision, recall, F1-score, and accuracy. Moreover, the confusion matrix of the CNNs and VTs models are shown in **Figure 4**.

It is observed from the test set results that the accuracy ranking of each model remains unchanged. The accuracy rate of the Xception network on the test set is still ranked first, at

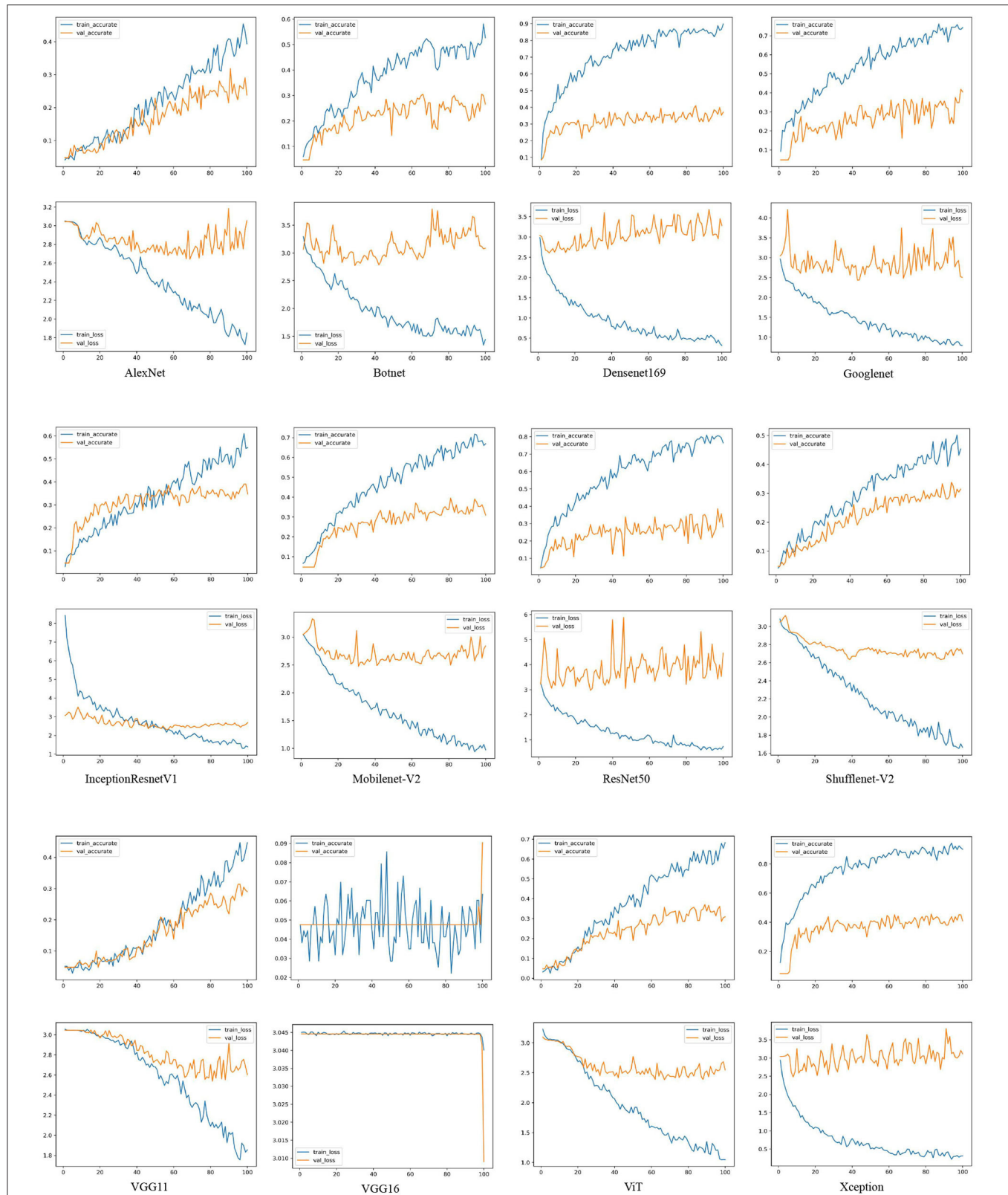


FIGURE 3 | The loss and accuracy curves of different deep learning networks on the training and validation sets. For example, AlexNet, Botnet, Densenet169, Googlenet, InceptionResnet-V1, Mobilenet-V2, ResNet50, ShuffleNet-V2, VGG11, VGG16, ViT, and Xception. train-accurate is the accuracy curve of the training set, train-accurate is the accuracy curve of the validation set, train-loss is the loss curve of the training set, and val-loss is the loss curve of the validation set.

TABLE 5 | Comparison of classification results of different deep learning models on the validation set.

Model	Avg. R(%)	Avg. P(%)	Avg. F1_score(%)	Accuracy(%)	Params Size (MB)	Time (s)
Xception	45.71	52.48	44.95	45.71	79.8	996
ResNet34	42.86	45.33	42.31	42.86	81.3	780
Googlenet	41.90	42.83	40.49	41.91	21.6	772
Densenet121	40.95	43.61	40.09	40.95	27.1	922
Densenet169	40.95	43.62	39.89	40.95	48.7	988
ResNet18	40.95	45.55	41.05	40.95	42.7	739
Inception-V3	40.00	45.01	39.70	40.00	83.5	892
MobileNet-V2	39.52	39.57	37.01	39.52	8.82	767
InceptionResnetV1	39.05	41.54	37.96	39.05	30.9	800
Deit	39.05	39.37	37.70	39.05	21.1	817.27
ResNet50	38.57	43.84	38.02	38.57	90.1	885
ViT	37.14	41.02	35.95	37.14	31.2	715
ResNet101	34.76	36.52	32.99	34.76	162	1021
T2T-ViT	34.29	38.17	34.54	34.28	15.5	825.3
ShuffleNet-V2	33.81	33.90	31.68	33.81	1.52	713
AlexNet	31.90	32.53	29.32	31.91	217	712
VGG11	31.43	41.20	29.97	31.43	491	864
BotNet	30.48	32.61	30.06	30.48	72.2	894
VGG13	20.95	19.23	18.37	20.95	492	957
VGG16	9.05	1.31	2.10	9.05	512	990
VGG19	4.76	0.23	0.44	4.76	532	1036

P denotes Precision, and *R* represents Recall. (Sort in descending order of classification accuracy).

40.32%, and is 3.81% higher than the second. Meanwhile, the average accuracy, average recall rate, and average F1-score of the Xception network also remain in the first place, at 40.32, 40.33, and 41.41%. Excluding the non-convergent VGG16 and VGG19 networks, the accuracy of the VGG13 validation set is still ranked at the bottom, at 15.55%. However, the ranking of the T2T-ViT network on the validation set accuracy rate changes dramatically. The accuracy rate of the T2T-ViT network is 34.28%, and the ranking rose from 12th to 5th. In addition, the AP, average recall and average F1-score of the T2T-ViT network are 38.17, 34.29, and 34.54%. Judging from the time consumed for the models, the ViT model consumes the least time at 3.77 s. On the other hand, the Densenet169 model consumes the most time at 11.13 s.

Figure 4 depicts the confusion matrix generated by part of the test dataset to more intuitively show the classification performance of the CNNs and VTs models on small EM datasets. In **Table 6**, Xception is the network with the best overall performance, and VGG13 is the network with the worst overall performance. In the confusion matrix of the Xception network, 127 EM images out of 315 EM images are classified into the correct category. In addition, the 11th type of EM classification performs the best, with 12 EM images are correctly classified and three EM images are misclassified into other categories. Meanwhile, the Xception network performs the worst in the 13th category of EM classification results. Three EM images are correctly classified and 14 EM images are misclassified into other categories. For the VGG13 network, 49 of the 315 EM images are classified into the correct category. Among them, the 16th EM classification performs best. Six EM images are

correctly classified, and 9 EM images are mistakenly classified into other categories. Comparing the CNNs and VTs models, all of the models perform well on the 11th EM classification and perform poorly on the 13th EM classification. For example, the ViT model correctly classifies 9 EM images and 0 EM images in the classification of the 11th and 13th class EMs, respectively.

Figure 4 shows that Xception better classifies the 11th and 16th types of EM images. ResNet is better at classifying tasks of the 11th and 16th types of images. Googlenet is better at classifying the 9th, 17th, and 21st EMs. The overall classification performance of T2T-ViT is poor. However, there are still outstanding performances in the 16th EM classification. The BotNet hybrid model is good at the 11th type of EM classification. However, the classification performance on the 12th and 13th images is abysmal. ResNet is good at image classification in the 9th, 11th, and 17th categories. The ViT model is good at the 11th, 12th, and 17th EM image classification. It is found from **Figure 4** that the images that each model is good at classifying are not the same. Therefore, there is a certain degree of complementarity among different deep learning models.

From **Figure 4**, Xception and Googlenet are highly complementary. For example, Googlenet has a good performance in the classification of EMs in classes 17 and 21, but Xception has a poor performance in the classification of EMs in classes 17 and 21. In addition, Xception is better at classifying the 11th class of EM images than Googlenet. This result shows that the features extracted by the two models are quite different. Two networks can extract features that each other network cannot extract. Therefore, there is a strong complementarity between

TABLE 6 | Comparison of classification results of different deep learning models on the test set.

Model	Avg. R(%)	Avg. P(%)	Avg. F1_score(%)	Accuracy(%)	Params Size (MB)	Time (s)
Xception	40.33	49.71	41.41	40.32	79.8	5.63
ResNet34	36.51	42.92	36.22	36.51	81.3	6.14
Googlenet	35.23	37.70	34.21	35.24	21.6	5.97
Mobilenet-V2	34.29	38.21	33.07	34.29	8.82	5.13
T2T-ViT	34.29	38.17	34.54	34.28	15.5	4.44
Densenet169	33.65	36.55	33.79	33.65	48.7	11.13
InceptionResnetV1	33.64	35.71	32.90	33.65	30.9	5.11
ResNet18	33.33	38.10	32.36	33.33	42.7	4.92
ResNet50	33.33	40.98	33.44	33.33	90.1	6.23
Densenet121	33.01	39.20	33.79	33.02	27.1	9.27
Deit	32.39	34.40	32.74	32.38	21.1	5.43
ViT	31.75	33.84	31.47	31.74	31.2	3.77
Inception-V3	31.11	34.84	31.32	31.11	83.5	7.49
ResNet101	27.94	34.59	28.31	27.94	162	8.83
VGG11	27.61	29.64	26.00	27.62	491	4.98
ShuffleNet-V2	27.30	25.02	24.98	27.30	1.52	5.42
BotNet	25.40	29.65	26.04	25.39	72.2	6.5
AlexNet	24.44	23.98	22.65	24.44	217	3.9
VGG13	15.55	15.18	14.38	15.55	492	5.28
VGG16	8.26	1.28	1.93	8.25	512	5.79
VGG19	4.76	0.23	0.44	4.76	532	6.42

P denotes Precision, and *R* represents Recall. (Sort in descending order of classification accuracy).

the two features. In addition, although VGG11 performs poorly in the classification of EMs. However, VGG11 is better at class 1 and class 19 classification tasks than Resnet34. Therefore, there is still a certain complementarity between the features extracted by the two models. This complementarity makes it possible to improve model performance through feature fusion.

In the study, we combine 18 models in pairs. Regardless of the specific feature fusion method or the possibility of a particular implementation, we calculate the ideal performance of the two models after fusion based on the current results. Part of the results is shown in **Table 7**. All results of the table are in the appendix. In **Table 7**, the ideal accuracy rate of each combination is calculated by the following steps. For each combination, the best results of every model are firstly accumulated. Then, the accumulated results are divided by the total number of images in the test set, and the result is the ideal accuracy rate. For example, the combination of Xception and Googlenet. In class 1 EM classification, Xception correctly classifies four images, and Googlenet correctly classifies five images. Here, 5 are the best results. The other categories can be deduced by analogy. The calculation method of model performance improvement is as follows: Use the ideal accuracy to subtract the highest accuracy of the two models to obtain the performance that can be improved in the ideal state after the fusion. In **Table 7**, the fusion of Xception and Googlenet performs best on the EMDS-6, with a classification accuracy of 46.03%. However, ResNet101 and VGG11 are improved the most after the fusion, and the two models have the strongest complementarity. On the left side of **Table 7**, we can clearly see the ideal effect of improving

the accuracy after the fusion of the two features. The improved accuracy after fusion reflects the complementarity of the two models to some extent. This complementarity can provide some help to researchers who are engaged in feature fusion.

4.3. Extended Experiments

4.3.1. After Data Augmentation, the Classification Performance of Each Model on the Validation Set

In this section, we augment the dataset, and the performance indicators of the models are calculated and exhibited in **Table 8**, including precision, recall, F1-score, and accuracy. In addition, we compare the accuracy changes before and after data augmentation, as shown in **Figure 5**.

After data augmentation, the time required for model training also increases significantly. The training time of the ViT models is the least, which is 902.27 s. Although the training set is augmented to six times, the training time of the ViT models is increased by 187.27 s compared with the 715 s. The classification accuracy of the Xception network ranks first at 52.62%. The T2T-ViT network has the lowest classification rate of 35.56%.

After data augmentation, the classification performance of each model is improved. **Figure 5** shows the changes in the accuracy of each model after data augmentation. The validation set accuracy of the VGG16 network is increased the most, at 28.41%. This is because the VGG16 network can converge on the augmentation dataset. In addition, the validation set accuracy of VGG13 and VGG11 are improved significantly, increasing by 21.59 and 16.67%, respectively. The accuracy of the VGG11 validation set rose from 17th to 3th. The accuracy of the VGG13

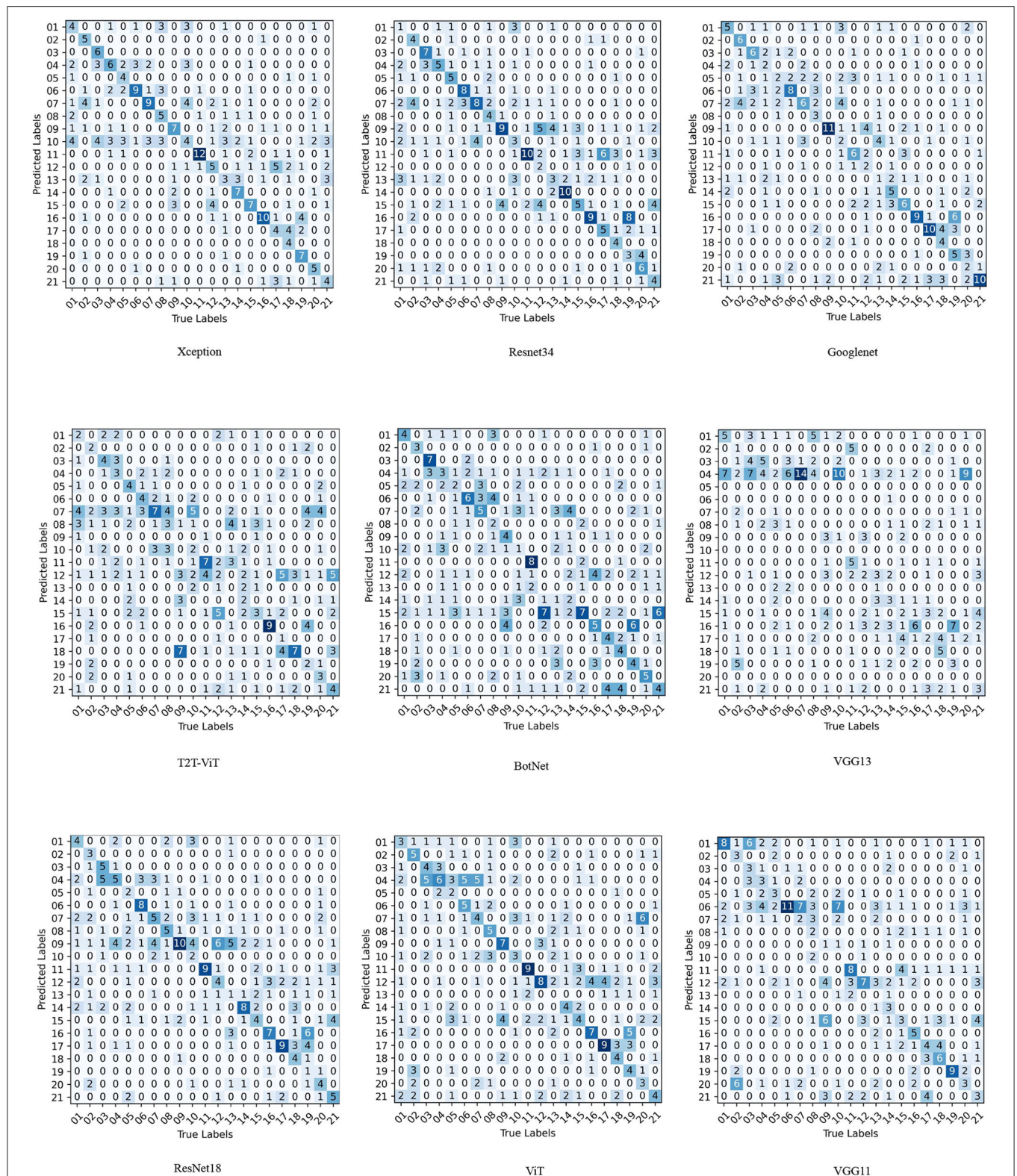


FIGURE 4 | Confusion matrix comparison of different networks on test set, Xception, Resnet34, Googlenet, T2T-ViT, BotNet, VGG13, ResNet18, ViT, and VGG11. (In the confusion matrix, 01, 02, 03, 04, 05, 06, 07, 08, 09, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21 represent Actinophrys, Arcella, Aspidisca, Codosiga, Colpoda, Epistylis, Euglypha, Paramecium, Rotifera, Vorticella, Noctiluca, Ceratium, Stentor, Siprostomum, K. Quadrala, Euglena, Gymnodinium, Gymmyano, Phacus, Stylongchia, and Synchaeta, respectively).

TABLE 7 | After fusing the two features, it has ideal precision and ideal performance improvement.

Model			Model		
Change (up) (%)			Accuracy		
ResNet101	VGG11	9.52	Googlenet	Xception	46.03%
InceptionResnetV1	ResNet18	7.94	Inception-V3	Xception	44.76%
Inception-V3	Shufflenet-V2	7.62	ResNet50	Xception	44.76%
Shufflenet-V2	VGG11	7.62	Deit	Xception	44.44%
Deit	VGG11	7.30	Densenet161	Xception	44.13%
Inception-V3	VGG11	7.30	VGG11	Xception	44.13%
ResNet18	ResNet50	7.30	Densenet121	Xception	43.81%
ResNet34	ResNet50	7.30	Mobilenet-V2	Xception	43.81%
ResNet34	VGG11	7.30	ResNet34	ResNet50	43.81%
ResNet101	Shufflenet-V2	7.30	ResNet34	VGG11	43.81%
Googlenet	Mobilenet-V2	7.30	Densenet121	ResNet34	43.49%
Alexnet	T2T-ViT	6.98	Googlenet	ResNet34	43.49%
Deit	Mobilenet-V2	6.98	InceptionResnetV1	Xception	43.49%
Deit	ViT-5	6.98	Mobilenet-V2	ResNet34	43.49%
Densenet121	Googlenet	6.98	ResNet18	Xception	43.49%

The left side of the table shows the improved accuracy of feature fusion under ideal conditions, and the right side of the table shows the accuracy of feature fusion under ideal conditions.

validation set rose from 19th to 11th. After data augmentation, the validation set accuracy of T2T-ViT, Densenet169, and ViT are not improved significantly, increasing by 1.28, 1.19, and 1.91%.

From a specific series of models, the performance of VGG series models is improved significantly after data augmentation. The performance improvement of the Densenet series models is not apparent. The accuracy of the Densenet121 and the Densenet169 validation sets are increased by 1.43 and 1.19%, respectively. Meanwhile, the performance improvement of the ViT series models is not apparent. The classification accuracy of the T2T-ViT validation set is increased by 1.28%, ViT is increased by 1.91%, and Diet is increased by 4.28%. In the ResNet series models, ResNet18, ResNet34, and ResNet50 are increased by 3.49, 3.25, and 3.65%, and the improvement is not obvious. However, the classification accuracy of the ResNet101 validation set is increased by 8.65%, which is obvious.

4.3.2. After Data Augmentation, the Classification Performance of Each Model on the Test Set

After data augmentation, the performance of each model on the test set is shown in **Table 9**. In **Table 9**, the Xception network has the highest accuracy of 45.71%. Meanwhile, the Xception network has an excellent recall index of 50.43%. Excluding the non-convergent VGG19, the VGG16 model has the worst performance, with an accuracy of 24.76%. The ViT model consumes the least time, which is 3.72 s. The Densenet169 model consumes the most time, which is 11.04 s.

Figure 6 shows the change of accuracy on the test set before and after the augmentation. In **Figure 6**, we can see that the accuracy of each deep learning model on the test set is generally increased. Among them, the accuracy of the VGG series models is improved the most. VGG11 is increased by 9.25%, VGG13 is increased by 21.28%, and VGG16 is increased by 16.51%. However, the accuracy of the ViT series models test set is not

significantly improved. The accuracy of some model test sets even drops. After data augmentation, the accuracy of the Diet network validation set is not changed. The accuracy of the T2T-ViT network is dropped by 3.80%. The accuracy of the ViT model is dropped by 3.17%. However, the accuracy of BotNet, a mixed model of CNN and ViT, is improved significantly, reaching 11.12%.

4.3.3. In Imbalanced Training, After Data Augmentation, the Classification Performance of Each Model on the Validation Set

In this section, we re-split and combine the data. Take each of the 21 types of EMs as positive samples in turn and the remaining 20 types of microorganisms as negative samples. In this way, we repeat this process 21 times in our paper. The specific splitting method is shown in Section 3.1.3. The deep learning model can calculate an AP after training each piece of data. **Table 10** shows the AP and mAP of each model validation set. We select the classical VGG16, ResNet50, and Inception-V3 networks for experiments. Furthermore, a relatively novel ViT model is also selected. In addition, the Xception network, which has always performed well above, is selected for experiments. Since the VGG16 network cannot converge at a LR of 0.0001, this part of the experiment adjusts the LR of the VGG16 network to 0.00001.

It can be seen in **Table 10** that the mAP of the Xception network is the highest, which is 56.61%. The Xception network has the highest AP on the 10th data, and the AP is 82.97%. The Xception network has the worst AP on the 3rd data, with an AP of 29.72%. As shown in **Figure 7**, the confusion matrix (d) is drawn by the 10th data. In (d), 46 of the 60 positive samples are classified correctly, and 14 are mistakenly classified as negative samples. In the confusion matrix drawn by the third data, 8 of the 60 positive samples are classified correctly, and 52 are incorrectly classified as negative samples.

TABLE 8 | Comparison of classification results of different deep learning models on the validation set.

Model	Avg. R(%)	Avg. P(%)	Avg. F1_score(%)	Accuracy(%)	Params Size (MB)	Time (s)
Xception	52.62	52.05	50.63	52.62	79.80	2636.08
Mobilenet-V2	49.67	51.91	48.82	49.68	8.82	1237.49
VGG11	48.10	52.40	48.44	48.10	491.00	1745.73
ResNet34	46.10	47.85	44.68	46.11	81.30	1335.87
ResNet18	44.44	51.87	43.03	44.44	42.70	1090.39
Googlenet	44.29	47.16	43.50	44.29	21.60	1257.33
Inception-V3	43.97	50.78	43.41	43.97	83.50	2004.08
AlexNet	43.58	45.02	43.05	43.57	217.00	951.27
ResNet101	43.41	46.08	43.33	43.41	162.00	2786.95
Deit	43.34	46.62	43.29	43.33	21.10	1306.99
VGG13	42.54	41.38	41.21	42.54	492.00	2307.04
Densenet121	42.38	46.91	42.39	42.38	27.10	2169.11
ResNet50	42.22	47.76	42.10	42.22	90.10	1968.28
Densenet169	42.14	48.04	42.79	42.14	48.70	2526.61
InceptionResnetV1	41.66	47.83	41.68	41.67	30.90	1451.76
ViT	39.05	43.50	38.52	39.05	31.20	902.27
ShuffleNet-V2	37.62	39.37	36.84	37.62	1.52	965.81
VGG16	37.47	38.21	36.80	37.46	512.00	2589.15
BotNet	36.59	36.38	35.59	36.59	72.20	2000.17
T2T-ViT	35.56	38.43	36.19	35.56	15.50	1385.62
VGG19	4.76	0.23	0.44	4.76	532.00	1022.57

P denotes Precision, and *R* represents Recall. The training set is augmented. (Sort in descending order of classification accuracy).

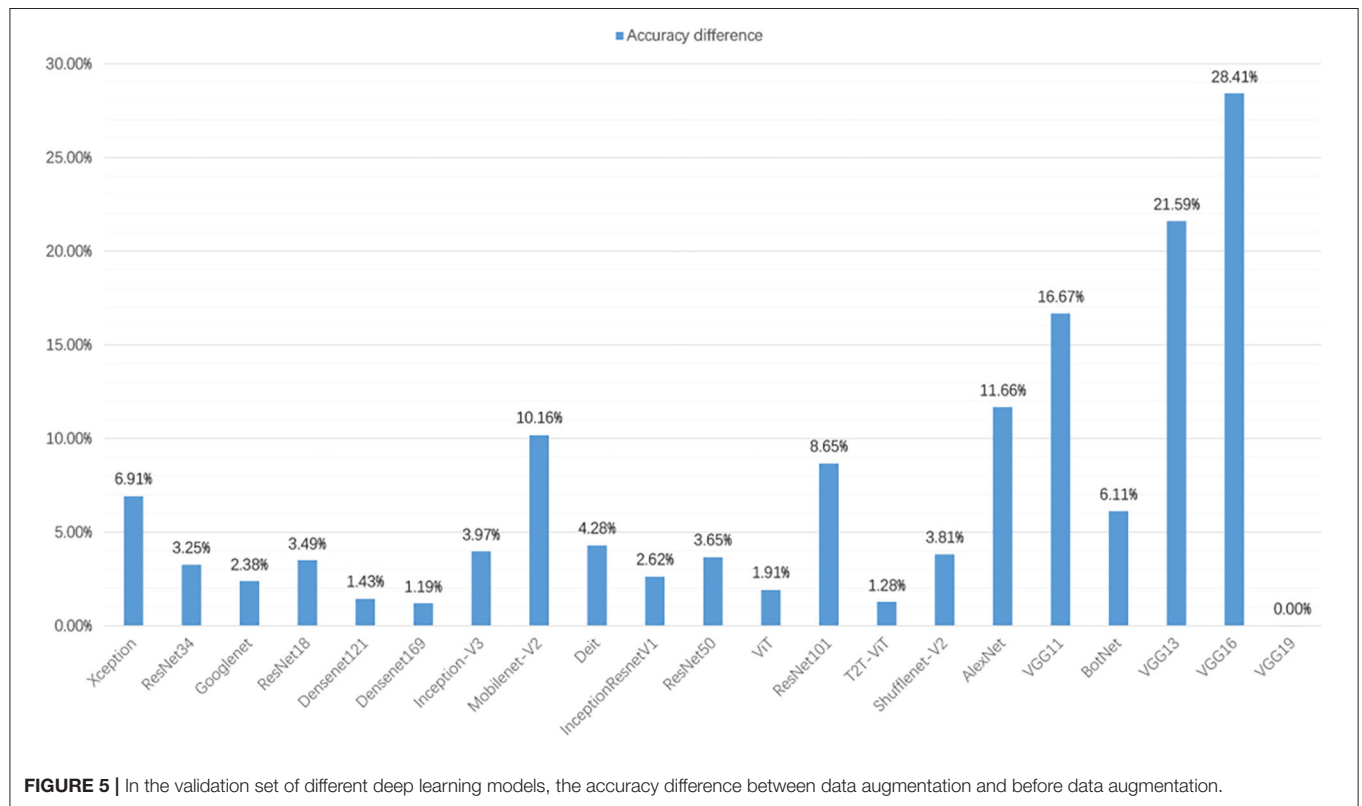
**FIGURE 5** | In the validation set of different deep learning models, the accuracy difference between data augmentation and before data augmentation.

TABLE 9 | Comparison of classification results of different deep learning models on the test set.

Model	Avg. R(%)	Avg. P(%)	Avg. F1_score(%)	Accuracy(%)	Params Size (MB)	Time (s)
Xception	45.71	50.43	46.15	45.71	79.8	5.49
Mobilenet-V2	42.54	47.56	43.07	42.54	8.22	5.04
ResNet18	39.05	44.82	39.22	39.05	42.7	4.90
Densenet121	38.73	40.28	38.20	38.73	27.1	8.98
ResNet34	38.73	42.25	37.84	38.73	81.3	6.07
ResNet50	38.10	41.56	36.97	38.10	90.1	6.20
Inception-V3	37.78	44.32	38.00	37.78	83.5	7.47
Googlenet	37.46	43.55	37.92	37.46	21.6	6.03
Densenet169	37.14	41.51	37.37	37.14	48.7	11.04
VGG11	37.14	38.81	36.70	37.14	491	4.96
InceptionResnetV1	36.82	41.47	36.75	36.83	30.9	5.11
VGG13	36.82	38.46	36.25	36.83	492	5.28
BotNet	36.50	39.12	36.35	36.51	72.2	6.44
ResNet101	35.23	38.01	35.44	35.24	162	8.85
AlexNet	34.92	39.10	34.97	34.92	217	5.25
Deit	32.39	34.40	32.74	32.38	21.1	4.41
T2T-ViT	30.48	35.88	30.85	30.48	15.50	5.41
ShuffleNet-V2	28.57	35.64	29.41	28.57	1.52	5.42
ViT	28.58	29.63	27.86	28.57	31.2	3.72
VGG16	24.77	25.53	24.11	24.76	512	5.79
VGG19	4.76	0.23	0.44	4.76	532	6.36

P denotes Precision, and *R* represents Recall. The training set is augmented. (Sort in descending order of classification accuracy).

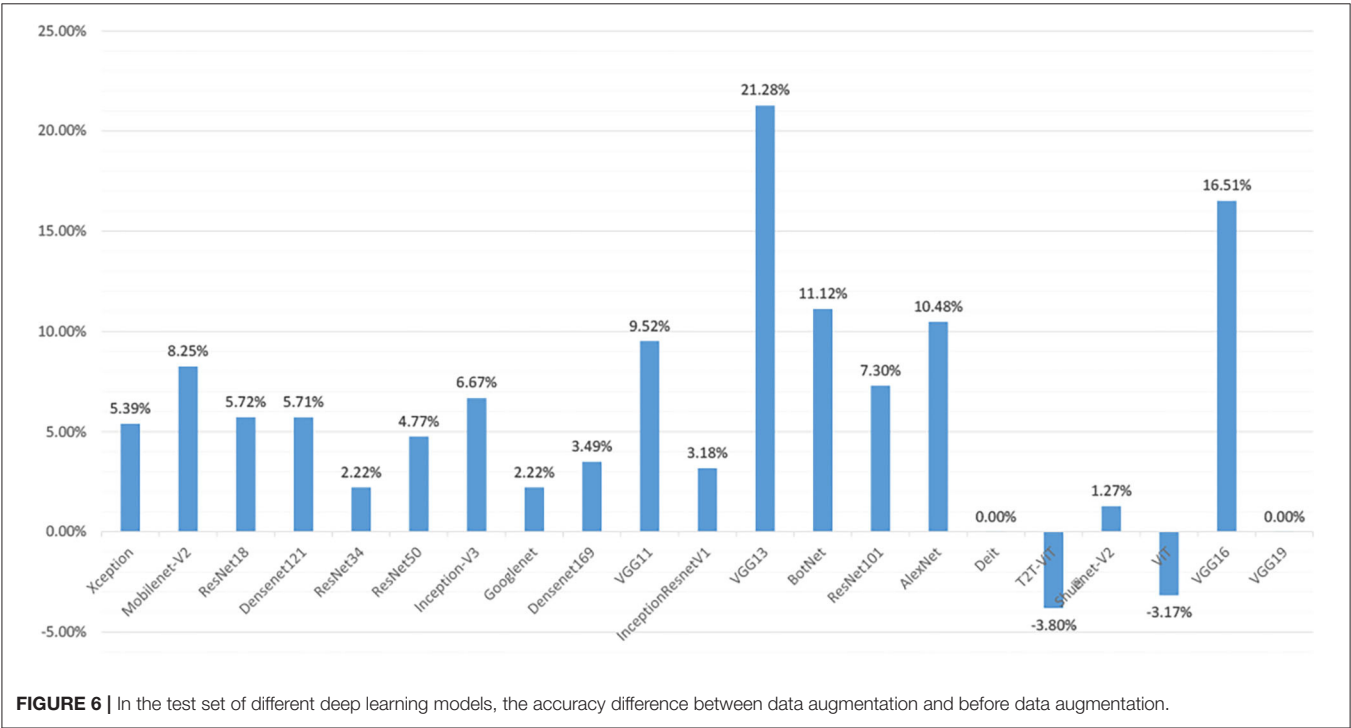


FIGURE 6 | In the test set of different deep learning models, the accuracy difference between data augmentation and before data augmentation.

The mAp of the VGG16 network is the lowest at 34.69%. The VGG16 network performs best on the 10th data AP, with an AP of 76.12%. The VGG16 network performs the worst on the 21st data AP, with an AP of 5.47%. Despite tuning the LR, the VGG16 network still fails to converge on the 3rd, 8th, 13th, 15th, and 21st data.

TABLE 10 | AP and MAP of different deep learning models in imbalanced training.

Model/Sample	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)	6 (%)	7 (%)	8 (%)	9 (%)	10 (%)	11 (%)
ViT	30.77	44.99	18.43	48.51	74.47	76.17	50.98	15.32	31.12	60.74	54.02
Xception	37.66	51.16	29.72	68.32	73.66	67.96	79.19	65.41	55.84	82.97	55.91
VGG16	48.38	41.43	9.63	51.05	52.61	42.23	76.92	5.97	27.57	76.12	34.77
ResNet50	30.58	45.96	14.24	68.19	66.15	43.10	71.24	46.51	31.87	62.19	36.79
Inception-V3	37.75	36.79	33.41	56.37	55.77	43.51	59.52	41.18	38.40	75.03	69.26
Model/Sample	12 (%)	13 (%)	14 (%)	15 (%)	16 (%)	17 (%)	18 (%)	19 (%)	20 (%)	21 (%)	mAPA
ViT	15.24	17.84	25.46	6.74	13.95	48.61	7.26	60.33	23.07	9.53	34.93
Xception	54.16	52.28	65.06	46.36	30.61	60.41	31.21	61.14	45.50	74.36	56.61
VGG16	24.06	16.22	63.90	5.80	10.49	33.87	24.77	44.00	33.14	5.47	34.69
ResNet50	15.59	42.12	68.57	24.94	17.49	47.52	6.64	49.04	16.73	56.10	41.03
Inception-V3	15.09	49.09	64.11	37.91	15.00	43.98	15.84	54.40	10.78	60.38	43.50

(In [%]).

The mAp of the ViT network and the VGG16 network are relatively close. The ViT network performs best on the 6th data AP, with an AP of 76.17%. Among the 60 positive samples, 35 are classified correctly, and 25 are classified as negative samples. The ViT network performs the worst on the 15th data AP, with an AP of 6.74%. Among the 60 positive samples, 0 are classified correctly and 60 are classified as negative samples.

In addition, Resnet50 performs the best on seven data AP and the worst on the 18th data AP. The Inception-V3 network performs best on 10 data AP and the worst on the 16th data AP.

4.3.4. Mis-classification Analysis

In the extended experiments, we randomly divide EMDS-6 three times and train the data for each division. The results and accuracy errors of the three experiments are shown in **Table 11** and **Figure 8**.

In **Table 11**, under the original dataset, Xception has the best classification performance on 21 deep learning models. After data augmentation, Xception still has the highest classification performance. In **Table 11**, the performance of the VGG series network has major changes compared to **Table 9**. In **Figure 9**, we can clearly understand that VGG11, VGG13, VGG16, and VGG19 failed to converge at least once in the three experiments. This phenomenon causes the VGG series models to fall behind in average performance. Except for the VGG series models, the performance of other models tends to be stable on the whole, and the errors are kept within $\pm 5\%$ of the average of the three experiments. Xception and Densenet169 networks show good robustness in the classification results before and after data augmentation. However, the classification performance of the AlexNet network fluctuates greatly in the three experiments, and the robustness is poor.

In **Figure 9**, after data augmentation, the performance of VGG13 improves the most, but this is mainly caused by the failure of some experiments on the original dataset to converge. In addition to the VGG13 network, the Mobilenet-V2, ShuffleNet-V2, and Densenet121 models improve the most, with

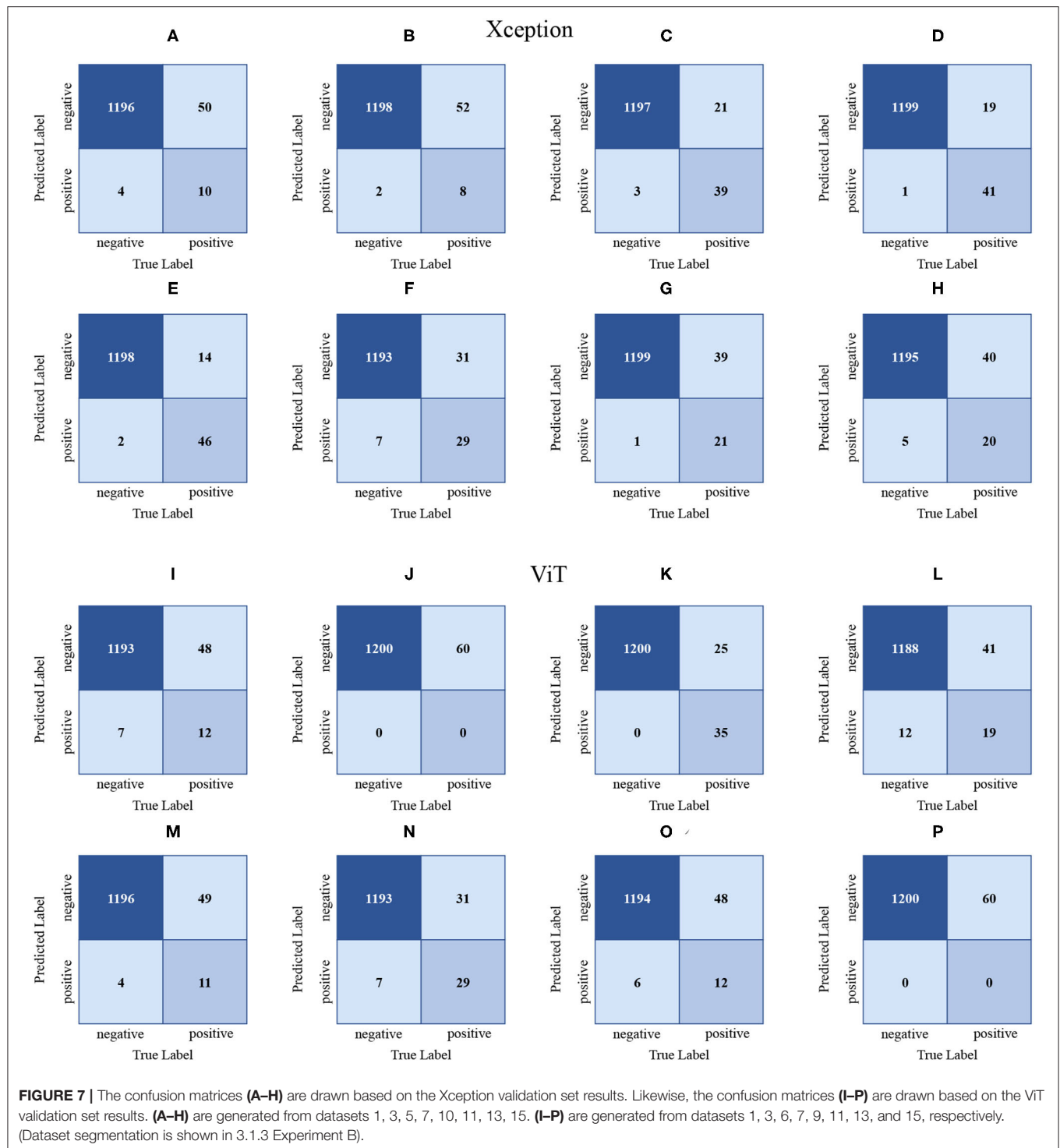
accuracy rates increase by 10.25, 9.52, and 8.89%. In addition, the performance improvement of ResNet34, ResNet18, and InceptionResnetV1 models is relatively small, and the accuracy are increase by 2.54, 2.96, and 3.5%. Generally speaking, after data augmentation, the CNN series models have a very obvious improvement in the precision, recall, F1-Score, and accuracy of the test set. However, the opposite situation appeared in the VTs after data augmentation. Taking the Accuracy index as an example, the accuracy of the ViT model in the test set has dropped by -2.5%, the Accuracy of the T2T-ViT model is equal to that before the augmentation, and the Accuracy of the Deit model has only increased by 1.16%.

In general, augmenting the dataset through geometric transformation can effectively improve the classification performance of the CNN series models. Nevertheless, for the VTs, the method of geometric transformation to augment the dataset is difficult to improve the classification performance of the VTs and even leads to a decrease in model performance.

4.3.5. Comparison of Experimental Results After Tuning Model Parameters

In this section, our extended experiments select representative models, namely the CNN-based Xception, the Transformer-based ViT, and the BotNet hybrid model based on CNN and VT. This section of the experiment trains 100 epochs. The purpose of the study is to observe the effect of changing two hyper-parameters, LR, and batch size (BS), on the experimental results. The experimental results are shown in **Table 12**.

Under the same BS and different LR conditions, the maximum fluctuation of ViT training time is only 4.6 s, the maximum fluctuation of BotNet training time is 74.6 s, and the maximum fluctuation of Xception training time is 80.6 s. Experiments indicate that tuning LR has little effect on the time required for training. However, the change of LR greatly influences the accuracy of experimental results. Taking the ViT as an example, the accuracy of the model is 16.83% under the conditions of BS = 16 and LR = 2×10^{-5} . Under the condition



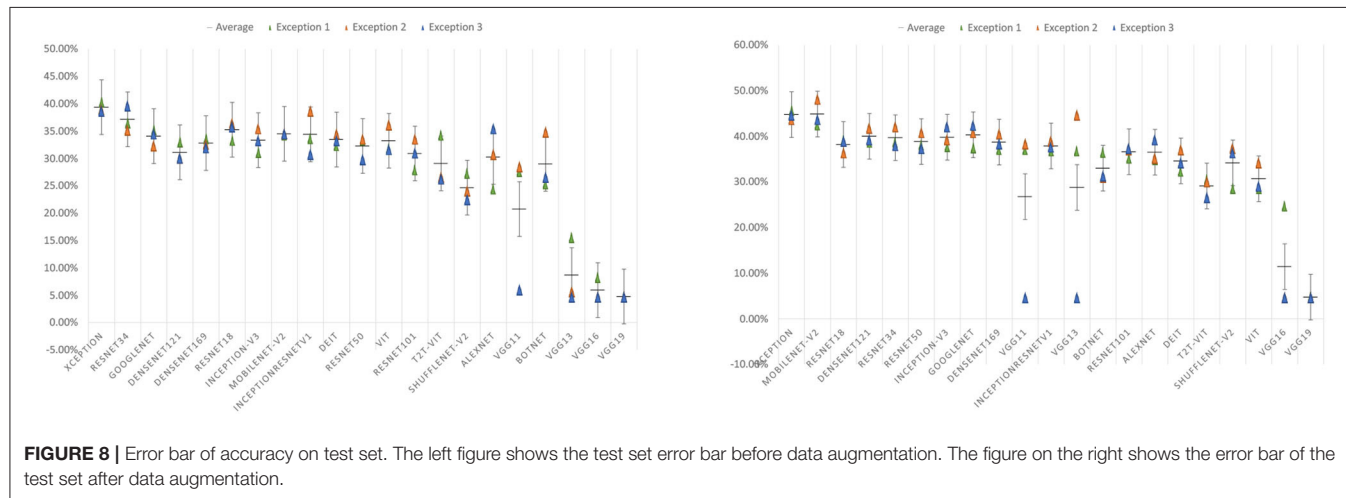
of $LR = 2 \times 10^{-4}$, the highest accuracy of the ViT can reach 31.11%. In addition, the accuracy of the model is only 3.17% under the condition of $LR = 2 \times 10^{-2}$. Experiments indicate that the performance of the model decreases when using an oversized LR ($LR = 2 \times 10^{-2}$) and an extremely small LR ($LR = 2 \times 10^{-5}$). An oversized LR may cause the network to fail

to converge, which means the model lingered near the optimal value and could not reach the optimal solution. This leads to performance degradation. The following two reasons explain the performance degradation when applying extremely small LRs. On the one hand, an extremely small LR makes the network hard to converge fastly. The related experiments show that the model

TABLE 11 | Comparison of different deep learning models on test set.

Model	Original data				Augmented data			
	Recall (%)	Precision (%)	F1-score (%)	Accuracy (%)	Recall (%)	Precision (%)	F1-score (%)	Accuracy (%)
Xception	39.37	44.25	39.07	39.37	44.76	47.97	44.53	44.76
ResNet34	37.14	41.96	36.93	37.14	39.68	43.15	39.54	39.68
ResNet18	35.24	40.53	34.33	35.24	38.20	42.64	38.38	38.20
Mobilenet-V2	34.50	37.24	33.86	34.50	44.75	48.31	44.82	44.75
InceptionResnetV1	34.39	36.46	33.92	34.39	37.88	41.09	37.53	37.89
Googlenet	34.07	36.89	33.48	34.07	40.32	44.59	40.37	40.32
Deit	32.27	34.08	31.92	33.44	34.60	37.01	34.76	34.60
Inception-V3	33.33	33.78	32.26	33.33	39.79	43.17	39.69	39.79
ViT	33.24	34.92	32.63	33.23	30.69	32.49	30.08	30.69
Densenet169	32.80	35.38	32.49	32.80	38.73	43.52	38.79	38.73
ResNet50	32.28	36.41	31.79	32.27	38.84	41.69	38.37	38.84
Densenet121	31.11	35.66	31.25	31.11	40.00	43.02	39.75	40.00
ResNet101	30.90	35.29	30.97	30.90	36.61	38.34	36.01	36.61
AlexNet	30.26	31.08	28.70	30.26	36.51	39.62	36.41	36.51
T2T-ViT	29.10	32.84	29.17	29.10	29.10	32.19	29.13	29.10
BotNet	29.00	31.11	28.46	28.99	33.02	34.29	32.45	33.02
ShuffleNet-V2	24.66	23.71	22.86	24.66	34.18	37.09	34.19	34.18
VGG11	20.74	19.99	18.31	20.74	26.77	26.98	25.39	26.77
VGG13	8.68	5.66	5.47	8.68	28.78	29.52	27.12	28.78
VGG16	5.93	0.58	0.94	5.92	11.43	8.66	8.33	11.43
VGG19	4.76	0.23	0.44	4.76	4.76	0.23	0.44	4.76

[In (%)].



is difficult to reach the optimal value within 100 epochs with an extremely small LR (2×10^{-5}). On the other hand, an extremely small LR may cause the network to fall into an optimal local solution, which leads to performance degradation.

In addition to the LR, the change of BS also dramatically affects the performance of the model. Different models show different patterns at different BS values. For example, the accuracy of the ViT model decreases rapidly with increasing BS at LR = 2×10^{-5} . The accuracy of the BotNet increases

sharply with increasing BS at LR = 2×10^{-5} . However, the relevant experiments show that BS does not seriously affect the performance of the model under large datasets (Radiuk, 2017). Nevertheless, with small datasets, only a slight change in the BS value can dramatically change the performance of the model.

Compared to a large dataset, tuning the BS and LR on a small dataset can significantly change model performance. Therefore, finding the optimal parameters to improve the model performance on small datasets is necessary.

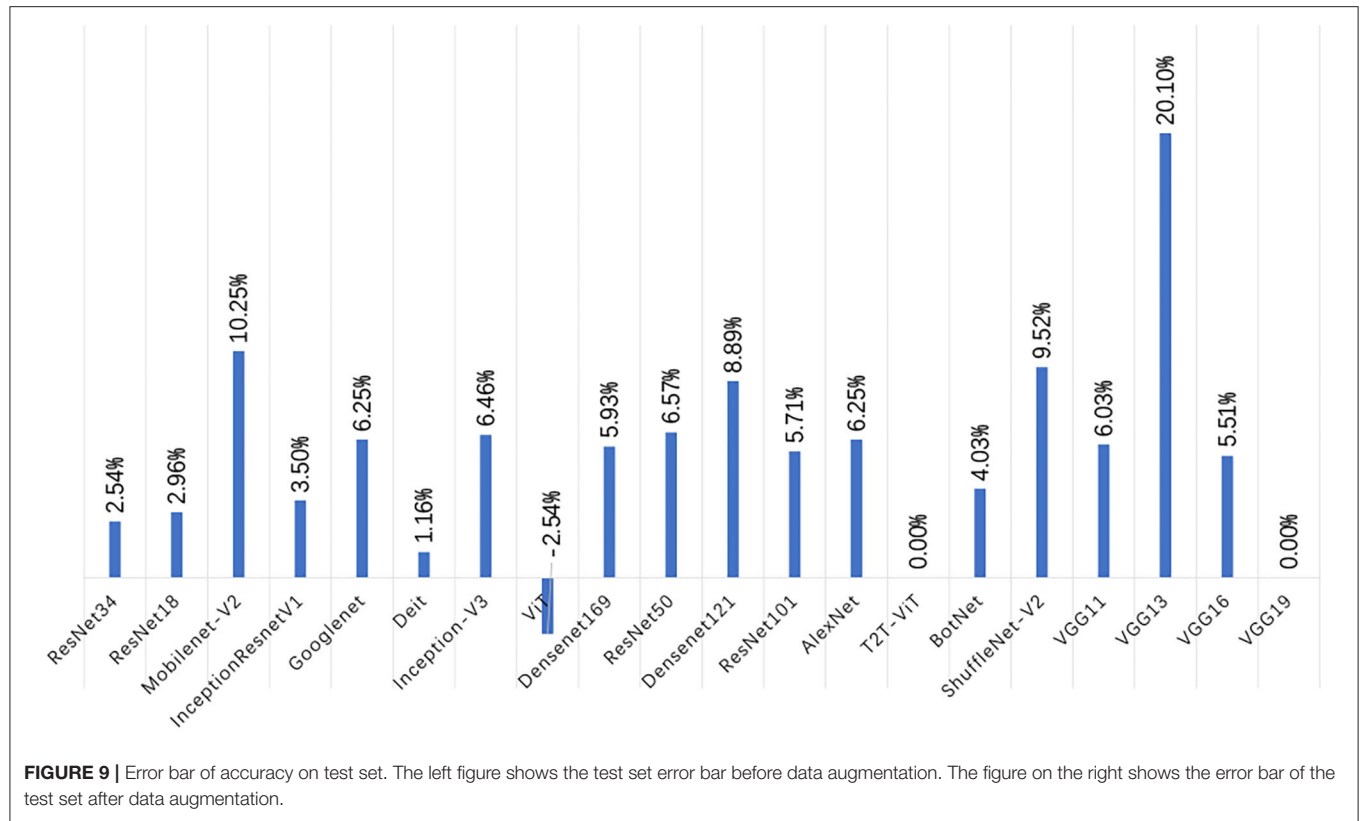


TABLE 12 | Comparison of training time consumption and test set accuracy of different networks.

LR	ViT (Times)				ViT (Accuracy)			
	BS 4	BS 8	BS 16	BS 32	BS 4 (%)	BS 8 (%)	BS 16 (%)	BS 32 (%)
2×10^{-5}	530.70	793.56	760.76	760.99	28.25	21.59	16.83	14.92
2×10^{-4}	530.32	793.12	761.17	762.88	30.16	27.94	31.11	30.16
2×10^{-3}	530.30	792.43	760.87	761.88	11.43	15.87	20.63	17.14
2×10^{-2}	535.30	794.01	760.67	760.99	4.76	4.76	3.17	7.62

LR	Xception (Times)				Xception (Accuracy)			
	BS 4	BS 8	BS 16	BS 32	BS 4 (%)	BS 8 (%)	BS 16 (%)	BS 32 (%)
2×10^{-5}	840.31	1106.94	1119.99	1074.65	38.10	37.46	37.14	37.78
2×10^{-4}	834.88	1107.95	1081.05	1088.86	51.43	50.48	41.90	38.73
2×10^{-3}	837.11	1113.56	1042.63	1042.75	23.49	34.29	28.25	30.48
2×10^{-2}	808.83	1086.24	1037.36	1073.62	14.29	16.83	20.00	17.46

LR	BotNet (Times)				BotNet (Accuracy)			
	BS 4	BS 8	BS 16	BS 32	BS 4 (%)	BS 8 (%)	BS 16 (%)	BS 32 (%)
2×10^{-5}	806.80	1006.82	977.10	950.71	16.51	17.14	20.32	21.27
2×10^{-4}	778.31	990.82	1022.45	1011.02	12.70	24.76	26.67	27.94
2×10^{-3}	772.63	984.33	967.09	937.65	9.21	7.94	14.29	10.79
2×10^{-2}	774.33	985.43	968.46	936.39	7.94	10.79	7.62	16.83

The left side of the table shows the training time consumption, while the right side of the table shows the accuracy of the test set. learning rate (LR), Batch Size (BS).

5. DISCUSSION

This experiment studies the classification performance of 21 deep learning models on small EM dataset (EMDS-6). The comparison results are obtained according to the evaluation indicators, as shown in **Tables 5, 6, 8, 9**. Meanwhile, some models are selected for imbalanced experiments to investigate the performance of the models further. The results are shown in **Table 10**. In order to increase the reliability of the conclusions, this paper repeats the main experiment three times. The average value is shown in **Table 11**, and the errors of the three experiments are shown in **Figure 8**. In addition, this paper explores the impact of hyper-parameters on small dataset classification, and the results are shown in **Table 12**.

The performance of the VGG network gradually decreases as the number of network layers increases. Especially the VGG16 and VGG19 networks cannot converge on EMDS-6. This may be because the dataset is too small, and the gradient disappears in the process of a continuous deepening of the network layer, which affects the convergence.

The training time of the ViT network on EMDS-6 is very short, but it does not make a significant difference with other models. After the data augmentation of EMDS-6, the ViT network has apparent advantages in the time of training the model, and the time consumption is much less than other models. We can speculate that the ViT model may further expand its advantage when trained on more training data.

In the experiments where the model parameters are tuned, slight changes in both the LR and BS parameters lead to drastic changes in model performance. This does not happen if the experiment is based on a large-scale dataset. However, in small datasets, each class of EMs only accounts for a portion of the image, and most of the others are noise. Moreover, some models that include batch normalization normalize the environmental noise at different BS leading to fluctuations in classification accuracy.

After data augmentation, the accuracy of CNN series models improves significantly. However, the increase of VT series model accuracy is slight, and some of them even decrease. The results are shown in **Figure 6**. To further prove the above experimental results, this paper re-divides the dataset and conducts three experiments, and the results are shown in **Figure 9**. Experiments once again prove that the geometric deformation augmented data method is difficult to improve the performance of the VT series models. This may be because our data augmentation method only makes geometric changes to the data. The geometric transformation is only changed the spatial position of the feature. However, the VT series models use attention to capture the global context information, and it pays more attention to global information. Operations such as rotation and mirroring have little effect on global information, and it is impossible to learn more global features. This makes the performance of the VT series models unable to improve after data augmentation significantly. However, the performance of BotNet, a hybrid model of CNN and VT, is significantly improved after data augmentation. This is because the BotNet network

only replaced three Bottlenecks with MHSA. The BotNet network is essentially more inclined to the feature extraction method of CNN.

6. CONCLUSION AND FUTURE WORK

The classification of small EM datasets are very challenging in computer vision tasks, which has attracted the attention of many researchers. Due to the development of deep learning, image classification of small datasets is developing rapidly. This article uses 17 CNN models, three VT models, and a hybrid CNN and VT model to test model performance. We have performed several experiments, including direct classification of each model, classification tasks after data augmentation, and imbalanced training tasks on some representative models. The experimental results prove that the Xception network is suitable for this kind of task. The ViT models take the least time for training. Therefore, the ViT model is suitable for large-scale data training. The ShuffleNet-V2 network has the least number of parameters, although its classification performance is average. Therefore, ShuffleNet-V2 is more suitable for occasions where high classification performance is not necessary and limited storage space.

This study provides an analysis table of the differences between the 18 models. This result can help related research on feature fusion quickly find models with significant differences and improve model performance. In addition, this study finds for the first time that the data augmentation method of geometric deformation is extremely limited or even ineffective in improving the performance of VT series models. This study and conclusion can provide relevant researchers with a conclusion with sufficient experimental support. Our research and conclusions reduce their workload in selecting experimental augmentation methods to a certain extent. This has a significant reference value.

Although the augmentation method of geometric deformation is effective for the performance improvement of CNNs, it does not help much for the performance improvement of VTs. We can improve the VT networks performance by studying new data augmentation methods in future work.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: <https://figshare.com/articles/dataset/EMDS-6/17125025/1>.

AUTHOR CONTRIBUTIONS

PZ: experiment, result analysis, and paper writing. CL: data preparation, method, result analysis, paper writing, proofreading, and funding support. MR: proofreading. HX and HY: experiment. HS: environmental microorganism knowledge support. TJ: result analysis and funding support. MG: method and result analysis. All authors contributed to the article and approved the submitted version.

FUNDING

This work is supported by the National Natural Science Foundation of China (No. 61806047), the Scientific Research

Fund of Sichuan Provincial Science and Technology Department under Grant (No. 2021YFH0069), and Project supported by the State Key Laboratory of Robotics (No. 2019-O13).

REFERENCES

- Alabandi, G. A. (2017). *Combining Deep Learning With Traditional Machine Learning to Improve Classification Accuracy on Small Datasets*. Thesis, Texas State University, San Marcos, TX.
- Amaral, A., Baptiste, C., Pons, M.-N., Nicolau, A., Lima, N., Ferreira, E., et al. (1999). Semi-automated recognition of protozoa by image analysis. *Biotechnol. Techniq.* 13, 111–118. doi: 10.1023/A:1008850701796
- Amaral, A., Ginoris, Y. P., Nicolau, A., Coelho, M., and Ferreira, E. (2008). Stalked protozoa identification by image analysis and multivariable statistical techniques. *Anal. Bioanal. Chem.* 391, 1321–1325. doi: 10.1007/s00216-008-1845-y
- Asgharnejad, H., and Sarrafzadeh, M.-H. (2020). Development of digital image processing as an innovative method for activated sludge biomass quantification. *Front. Microbiol.* 11, 2334. doi: 10.3389/fmicb.2020.574966
- Çayır, A., Yenidoğan, I., and Dağ, H. (2018). “Feature extraction based on deep learning for some traditional machine learning methods,” in *2018 3rd International Conference on Computer Science and Engineering (UBMK)*. IEEE (Sarajevo, Bosnia and Herzegovina), 494–497. doi: 10.1109/UBMK.2018.8566383
- Chandrarathne, G., Thanikasalam, K., and Pinidiyaarachchi, A. (2020). “A comprehensive study on deep image classification with small datasets,” in *Advances in Electronics Engineering, Lecture Notes in Electrical Engineering*, Vol. 619 (Singapore: Springer), 93–106. doi: 10.1007/978-981-15-1289-6_9
- Chen, C., and Li, X. (2008). “A new wastewater bacteria classification with microscopic image analysis,” in *Proceedings of the 12th WSEAS International Conference on Computers* (Heraklion), 915–921.
- Chollet, F. (2017). “Xception: deep learning with depthwise separable convolutions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 1251–1258. doi: 10.1109/CVPR.2017.195
- 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, N., Qi, R., Rossetti, S., Tandoi, V., Gao, Y., and Yang, M. (2017). Factors affecting the growth of microthrix parvicella: batch tests using bulking sludge as seed sludge. *Sci. Total Environ.* 609, 1192–1199. doi: 10.1016/j.scitotenv.2017.07.261
- Filzmoser, P., and Todorov, V. (2011). Review of robust multivariate statistical methods in high dimension. *Anal. Chim. Acta* 705, 2–14. doi: 10.1016/j.aca.2011.03.055
- Fried, J., Mayr, G., Berger, H., Traunsperger, W., Psenner, R., and Lemmer, H. (2000). Monitoring protozoa and metazoa biofilm communities for assessing wastewater quality impact and reactor up-scaling effects. *Water Sci. Technol.* 41, 309–316. doi: 10.2166/wst.2000.0460
- Han, D., Liu, Q., and Fan, W. (2018). A new image classification method using cnn transfer learning and web data augmentation. *Expert Syst. Appl.* 95, 43–56. doi: 10.1016/j.eswa.2017.11.028
- Haryanto, T., Suhartanto, H., Arymurthy, A. M., and Kusmardi, K. (2021). Conditional sliding windows: an approach for handling data limitation in colorectal histopathology image classification. *Inform. Med. Unlock.* 23, 100565. doi: 10.1016/j.imu.2021.100565
- 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, NV) 770–778. doi: 10.1109/CVPR.2016.90
- Hu, G., Peng, X., Yang, Y., Hospedales, T. M., and Verbeek, J. (2017). Frankenstein: learning deep face representations using small data. *IEEE Trans. Image Process.* 27, 293–303. doi: 10.1109/TIP.2017.2756450
- 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, HI), 4700–4708. doi: 10.1109/CVPR.2017.243
- Kholerdi, H. A., TaheriNejad, N., and Jantsch, A. (2018). “Enhancement of classification of small data sets using self-awareness—an iris flower case-study,” in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, IEEE (Florence), 1–5. doi: 10.1109/ISCAS.2018.8350992
- Kosov, S., Shirahama, K., Li, C., and Grzegorzec, M. (2018). Environmental microorganism classification using conditional random fields and deep convolutional neural networks. *Pattern Recogn.* 77, 248–261. doi: 10.1016/j.patcog.2017.12.021
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Adv. Neural Inform. Process. Syst.* 25:1097–1105. doi: 10.1145/3065386
- Kruk, M., Kozera, R., Osowski, S., Trzciński, P., Paszt, L. S., Sumorok, B., et al. (2015). “Computerized classification system for the identification of soil microorganisms,” in *AIP Conference Proceedings*, Vol. 1648 (Melville, NY: AIP Publishing LLC), 660018. doi: 10.1063/1.4912894
- Li, C., Shirahama, K., Grzegorzec, M., Ma, F., and Zhou, B. (2013). “Classification of environmental microorganisms in microscopic images using shape features and support vector machines,” in *2013 IEEE International Conference on Image Processing*, IEEE (Melbourne, VIC), 2435–2439. doi: 10.1109/ICIP.2013.6738502
- Ma, N., Zhang, X., Zheng, H.-T., and Sun, J. (2018). “Shufflenet v2: practical guidelines for efficient cnn architecture design,” in *Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science*, Vol. 11218, eds V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss (Springer, Cham), 116–131. doi: 10.1007/978-3-030-01264-9_8
- Mao, C., Huang, L., Xiao, Y., He, F., and Liu, Y. (2021). Target recognition of SAR image based on CN-GAN and CNN in complex environment. *IEEE Access* 9, 39608–39617. doi: 10.1109/ACCESS.2021.3064362
- McKinney, R. E. (2004). *Environmental Pollution Control Microbiology: A Fifty-Year Perspective*. Boca Raton, FL: CRC Press. doi: 10.1201/9780203025697
- Nie, D., Shank, E. A., and Jovic, V. (2015). “A deep framework for bacterial image segmentation and classification,” in *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics* (Atlanta, GA), 306–314. doi: 10.1145/2808719.2808751
- Pepper, I. L., Gerba, C. P., Gentry, T. J., and Maier, R. M. (2011). *Environmental Microbiology*. San Diego, CA: Academic Press.
- Phung, V. H., and Rhee, E. J. (2019). A high-accuracy model average ensemble of convolutional neural networks for classification of cloud image patches on small datasets. *Appl. Sci.* 9, 4500. doi: 10.3390/app9214500
- Powers, D. M. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv Preprint arXiv:2010.16061*.
- Radiuk, P. M. (2017). Impact of training set batch size on the performance of convolutional neural networks for diverse datasets. *Inform. Technol. Manage. Sci.* 20, 20–24. doi: 10.1515/itms-2017-0003
- Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). “Mobilenetv2: inverted residuals and linear bottlenecks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT), 4510–4520. doi: 10.1109/CVPR.2018.00474
- Sarrafzadeh, M. H., La, H.-J., Lee, J.-Y., Cho, D.-H., Shin, S.-Y., Kim, W.-J., et al. (2015). Microalgae biomass quantification by digital image processing and rgb color analysis. *J. Appl. Phycol.* 27, 205–209. doi: 10.1007/s10811-014-0285-7
- Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Srinivas, A., Lin, T.-Y., Parmar, N., Shlens, J., Abbeel, P., and Vaswani, A. (2021). Bottleneck transformers for visual recognition. *arXiv preprint arXiv:2101.11605*. doi: 10.1109/CVPR46437.2021.01625

- Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. (2017). "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31 (San Francisco, CA).
- 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. doi: 10.1109/CVPR.2015.7298594
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 2818–2826. doi: 10.1109/CVPR.2016.308
- Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. (2020). Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*.
- Wang, P., Fan, E., and Wang, P. (2021). Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. *Pattern Recogn. Lett.* 141, 61–67. doi: 10.1016/j.patrec.2020.07.042
- Xie, Y., Xing, F., Kong, X., Su, H., and Yang, L. (2015). Beyond classification: structured regression for robust cell detection using convolutional neural network. *Med. Image Comput. Comput. Assist. Interv.* 9351, 358–365. doi: 10.1007/978-3-319-24574-4_43
- Yang, C., Li, C., Tiebe, O., Shirahama, K., and Grzegorzec, M. (2014). "Shape-based classification of environmental microorganisms," in *2014 22nd International Conference on Pattern Recognition*, IEEE (Stockholm), 3374–3379. doi: 10.1109/ICPR.2014.581
- Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Tay, F. E., et al. (2021). Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986*.
- Zhang, Z., Cui, P., and Zhu, W. (2020). Deep learning on graphs: a survey. *IEEE Trans. Knowl. Data Eng.* 34, 249–270. doi: 10.1109/TKDE.2020.2981333
- Zhao, P., Li, C., Rahaman, M., Xu, H., Ma, P., Yang, H., et al. (2021). EMDS-6: Environmental microorganism image dataset sixth version for image denoising, segmentation, feature extraction, classification and detection methods evaluation. *arXiv Preprint arXiv: 2112.07111*. 1–11.
- Zhao, T., Liu, M., Zhao, T., Chen, A., Zhang, L., Liu, H., et al. (2021). Enhancement of lipid productivity in *Chlorella pyrenoidosa* by collecting cells at the maximum cell number in a two-stage culture strategy. *Algal Res.* 55, 102278. doi: 10.1016/j.algal.2021.102278
- Zhao, W. (2017). Research on the deep learning of the small sample data based on transfer learning. *AIP Confer. Proc.* 1864, 020018. doi: 10.1063/1.4992835

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.

Copyright © 2022 Zhao, Li, Rahaman, Xu, Yang, Sun, Jiang and Grzegorzec. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Machine Learning and Deep Learning Applications in Metagenomic Taxonomy and Functional Annotation

Alban Mathieu¹, Mickael Leclercq¹, Melissa Sanabria², Olivier Perin³ and Arnaud Droit^{1*}

¹ Computational Biology Laboratory, CHU de Québec - Université Laval Research Centre, Québec City, QC, Canada,

² Université Côte d'Azur, CNRS, INRIA, I3S, Nice, France, ³ Digital Sciences Department, L'Oréal Advanced Research, Aulnay-sous-Bois, France

OPEN ACCESS

Edited by:

Sayed Soheil Mansouri,
Technical University of Denmark,
Denmark

Reviewed by:

Jan Zrimec,
National Institute of Biology (NIB),
Slovenia
Asaf Levy,
Hebrew University of Jerusalem, Israel

*Correspondence:

Arnaud Droit
arnaud.droit@crchuq.ulaval.ca

Specialty section:

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

Received: 08 November 2021

Accepted: 02 February 2022

Published: 14 March 2022

Citation:

Mathieu A, Leclercq M,
Sanabria M, Perin O and Droit A
(2022) Machine Learning and Deep
Learning Applications in Metagenomic
Taxonomy and Functional Annotation.
Front. Microbiol. 13:811495.
doi: 10.3389/fmicb.2022.811495

Shotgun sequencing of environmental DNA (i.e., metagenomics) has revolutionized the field of environmental microbiology, allowing the characterization of all microorganisms in a sequencing experiment. To identify the microbes in terms of taxonomy and biological activity, the sequenced reads must necessarily be aligned on known microbial genomes/genes. However, current alignment methods are limited in terms of speed and can produce a significant number of false positives when detecting bacterial species or false negatives in specific cases (virus, plasmids, and gene detection). Moreover, recent advances in metagenomics have enabled the reconstruction of new genomes using *de novo* binning strategies, but these genomes, not yet fully characterized, are not used in classic approaches, whereas machine and deep learning methods can use them as models. In this article, we attempted to review the different methods and their efficiency to improve the annotation of metagenomic sequences. Deep learning models have reached the performance of the widely used k-mer alignment-based tools, with better accuracy in certain cases; however, they still must demonstrate their robustness across the variety of environmental samples and across the rapid expansion of accessible genomes in databases.

Keywords: machine learning, deep learning, metagenomic, whole genome shotgun, classification, taxonomic annotation, functional annotation

INTRODUCTION

The study of the microbial environments has benefited from the sequencing revolution, where technology improvement decreased the DNA sequencing cost and increased the number of sequenced nucleic bases. For approximately 20 years (depending on how we define the term metagenomics), it has allowed the decryption of the microbial composition of a huge variety of environments (Bahram et al., 2021). In the present publication, we use the term metagenome to refer to the directly sequenced DNA of one environment, without any prior amplification. This implies that a metagenome is a sample extracted from the total DNA of genomes, cut into fragments of hundreds to thousands base pair (bp) lengths. The fragments can be paired-end or not, depending on the technology used (Escobar-Zepeda et al., 2015). The DNA sample is then analyzed to answer the ambitious questions: “who is here?” and “what are they doing?” A variety of bioinformatic tools and software have been developed to annotate the sequences into

taxonomic and functional categories. They can be grouped into two categories: (i) alignment-based methods that infer taxonomy/functions based on similarity of sequences along reference databases such as BLAST (Altschul et al., 1990) and DIAMOND (Buchfink et al., 2014), (ii) k-mer-based approaches such as kraken2 (Wood et al., 2019) and CENTRIFUGE (Kim et al., 2016). However, these technologies suffer from dependence on prior knowledge and are not able to annotate sequences absent from the databases (at least with no resemblance). A good criterion to compare results between software will be to evaluate the capacity to annotate the sequenced reads to a taxonomy/functional entity, but because the annotation depends on the technology used, the choice of the associated parameters, or the intrinsic factors of the studied environment (**Figure 1**), the comparison is not feasible without a unique benchmark. Moreover, another aspect that affects the rate of annotation (e.g., the capacity to annotate the sequenced reads) is the level of analysis, which might be in terms of taxonomy, rank (species to domain), functions, and gene/pathway. The more the annotation is specific (threshold of similarity and level of analysis), the more the rate of annotation will be low. Nevertheless, to obtain interpretable information, having a detailed annotation in terms of taxonomy and functions will help to interpret the generated data. The trade-off has to be set according to the studies and their scientific questions (Inkpen et al., 2017).

In the last few years, new methods have emerged to analyze metagenomics data based on machine and deep learning approaches. These methods attempt to acquire the capacity to distinguish complex patterns among large datasets to make accurate predictions on future datasets that will be analyzed using the trained models (Greener et al., 2021). In metagenomic experiments, unsupervised or supervised models are widely used to make classification or clustering of samples based on annotation matrices. Current common approaches in the field are General Linearized Models to differentiate the microbial composition of samples, Principal Components Analysis to reduce data dimension and visualize data in an unsupervised way (Calle, 2019), and feature selection methods to define microbial signatures (Erickson et al., 2012; Loomba et al., 2017; Zhong et al., 2019). Learning and prediction of disease status of patient-related metagenomic samples have been rarely explored, but a successful application has been proposed using more than 2,400 metagenomic samples from clinical metagenomic studies (Pasolli et al., 2016). In this review, we do not attempt to expose all the machine learning methods and use cases existing in the literature, but we will try to unravel the issue of annotation that meets the machine and deep learning model requirements, exploring how it was applied in metagenomics annotation. **Table 1** summarizes models and tools reviewed in the following article.

CHALLENGES IN METAGENOMIC ANNOTATION

Taxonomic annotation of bacteria is complex and, because the microorganisms do not possess sexual reproduction, the definition of bacterial species is based on a laboratory experiment

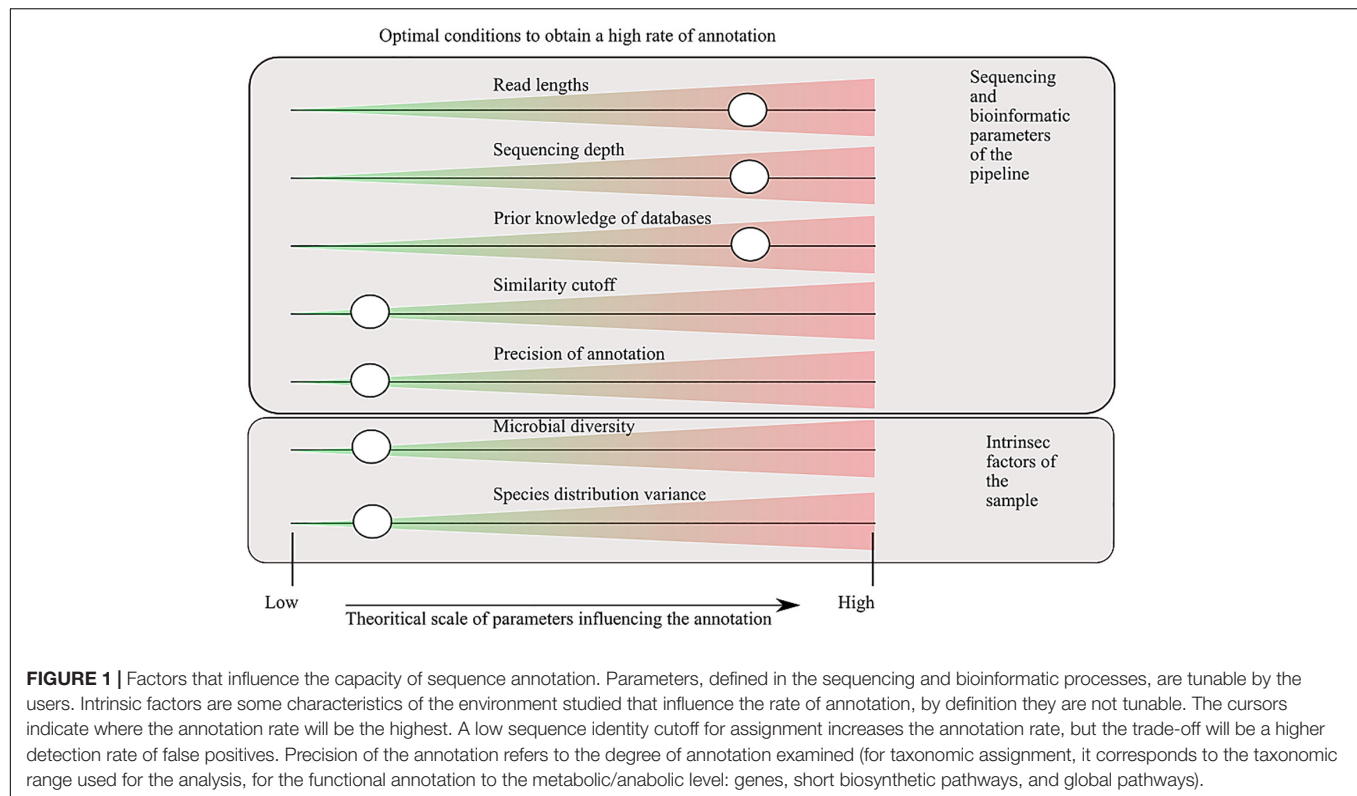
result to define a species, e.g., DNA–DNA hybridization of two bacterial genomes must be greater than or equal to 70% to be grouped in the same species (Wayne et al., 1987). However, this leads to high DNA heterogeneity functions in the species group. Breitwieser et al. (2019) collected information that demonstrated the difference in average nucleotide identity between different species, revealing the difficulty to classify them using DNA genomic sequences. Moreover, the microbial diversity is very large and not yet recovered. It has been estimated that we only accessed, using culture-based approaches, 0.001–1% of the total bacterial diversity present on earth (O’Leary et al., 2016). This emphasizes that genomes in reference databases do not cover the total diversity in the metagenomic samples. Finally, the emergence of assembly and *de novo* metagenomic reconstruction of genomes from metagenomic data, also called metagenomic assembled genomes (MAGs), has unveiled the numerous uncultured microorganisms in multiple environments (Qin et al., 2010; Lee et al., 2017; Delmont et al., 2018; Kroeger et al., 2018; Pedron et al., 2019). Because the genomes are not yet cultured, they can represent multiple genomes, and their taxonomy affiliation cannot be connected to known species. They are generally named with an identifier, e.g., (Genus) sp. (identifier) (for instance, *Bacillus* sp. M35), or are proposed with the species name preceded by the term “*Candidatus*.” MAG permitted the acquisition of yet uncultured genomes, but integrating MAG into metagenomic classifiers is complex because they may not be regular genomes and they are not fully integrated into the taxonomy.

Metagenomic data are also a source of functional information and, using reference databases, can be annotated to understand what the potential functions are that could reflect their ecological role in the studied environment. As there is a link between DNA sequences and functions, we can be more confident of the annotation process based on alignment, but the sequence similarity threshold to be confident is always a questionable point that will impact the annotation process (Treiber et al., 2020). Depending on the function, a different similarity percentage will be required to identify reads as a function. Moreover, the choice of the database can change the level of analysis and the interpretation. In addition, there are no official functional categories and a variety of databases has emerged, each with its own specificities (**Table 2**) (Caspi et al., 2014; Lombard et al., 2014; Kanehisa et al., 2016; Huerta-Cepas et al., 2019; Gene Ontology Consortium [GOC], 2021; Mistry et al., 2021; The UniProt Consortium et al., 2021). These non-standardized annotations alter our capacity to compare tool accuracies.

APPLICATION OF MACHINE LEARNING TO METAGENOMIC CLASSIFICATION

Characterization of the Metagenomic Annotation Process in Terms of Machine and Deep Learning

We represent here the machine learning process by characterizing the input, output, and the type of classification and model used



in the framework. The theoretical process of annotation can be analyzed as a multiclass classification problem, where a huge number of reads (input) must be uniquely classified into a wide variety of taxonomic ranks (output), meaning that it cannot be labeled in two different classes for one read. This is a supervised problem where the models are trained using a ground truth reference, i.e., true values are used to compare with the output of a machine learning model (Greener et al., 2021). In metagenomics in general, this ground truth is difficult to obtain because the metagenomic data are the fruit of complex and unresolved microbial phylogeny, as explained previously.

The machine learning model input *sensu stricto* will be millions of metagenomic reads and the output will be the category (or categories) to which the read belongs. To evaluate the classification performance, all the publications presented in this review used the same or equivalent metrics. These metrics are precision, which is the capacity of good assignment when there is an assignment, and recall (sensitivity), which is the number of reads correctly classified compared to the total number of reads to classify. All other possible metrics used in the literature are variants of these two metrics or their concatenation (F1-score, accuracy, use of taxonomic rank instead of reads count as unit of measure, etc.).

Precision

$$\frac{\# \text{ correct reads classified}}{\# \text{ reads classified}}$$

Recall

$$\frac{\# \text{ correct read classified}}{\# \text{ reads in datasets}}$$

Machine learning and deep learning models finally produce the output, which is the final classification of reads into categories and the associated probabilistic/confidence value. Applied to metagenomics, the confidence value is used to define a threshold of assignment of reads. These thresholds are defined by the authors and impact the model accuracy. A parallel with alignment-based method is the percentage of similarity required to annotate a read to its hit in the database.

Naive Bayes Classification Model

One of the first approaches of machine learning classifiers on nucleotide signatures was the application of naive Bayes (NB) models on 28 genomic data present in the genomic databases in 2001 (Sandberg et al., 2001). This work was the foundation for the development of the application of NB classification on shotgun metagenomic data by Rosen et al. (2008), who trained their classifier using 635 microbial genomes to construct k-mer frequency profiles of the genomes, then tested the classification of simulated fragments and metagenomic reads into different classes (strain, species, and genus).

The term “naive” in NB refers to the fact that the Bayes theorem assumes that the values of a particular feature are independent of the value of any other feature, which simplifies the problem and gives a starting point to estimate the degree of complexity of the problem, here, the metagenomic classification. The genomes were divided into 25-, 100-, and 500-bp length, and 3-, 6-, and 9- to 15-mer fragments were used to train the model. A total of 63,500 fragments were isolated to test the accuracy of the models. The log-likelihood score for each sequence was

TABLE 1 | Summary of the articles and models reviewed.

Publication	Machine/deep learning category	Models tested	Training input	Tested input	Real applications input	Output	Encoding scheme	Parameters	Hyper-parameters	Best model selected
NBC: the naive Bayes classification tool web server for taxonomic classification of metagenomic reads (Rosen et al., 2011)	Machine learning Supervised classification	Naive Bayes	Genome sequence from DB (25, 100, and 500 bp)	Genome sequence from DB (25, 100, and 500 bp) Metagenomic data	Metagenomic reads	Strain-species-genus classification	Compositional vectors ("Target encoding" like)	NA	k-mer size (3, 6, and 9–15)	Naive Bayes
Accurate phylogenetic classification of variable-length DNA fragments (McHardy et al., 2007)	Support vector machine	Linear or Gaussian SVM	Genome sequence from DB (1, 5, 10, and 15 kb)	Genome sequence from DB (25, 100, and 500 bp) Metagenomic assembled data	Contigs (assembled metagenomic data)	Genus to domain classification	Compositional vectors ("Target encoding" like)	Misclassification cost Gaussian/linear kernel	k-mer size (2–6) Input length (5, 10, 15, and 50 kb)	5–6-mer-size Gaussian SVM
Large-scale machine learning for metagenomics sequence classification. (Vervier et al., 2016)	Support vector machine	Linear SVM	Genome sequence	Genome sequence affiliated to the same species as trained. Simulated reads with sequencing error model introduction	Metagenomic reads	Rank flexible classification of metagenomic reads	Compositional vectors ("Target encoding" like)	Squared loss function Stochastic gradient descent	k-mer size (4, 5, and 6) Quantity of input data	Linear SVM classifier with rank-flexible classification
Deep learning models for bacteria taxonomic classification of metagenomic data (Fiannaca et al., 2018)	Deep neural network (DNN)	Convolutional neural network (CNN) Deep belief network (DBN)	Simulated reads of 16S RNA sequences	Simulated reads of 16S RNA sequences	16S amplicon reads or metagenomic reads	Domain to genus classification	One hot encoding	# hidden unit # hidden layers # kernel # kernel size # Pooling size	k-mer size (3–7)	CNN
DeepMicrobes: taxonomic classification for metagenomics with deep learning (Liang et al., 2020)	Deep neural network (DNN)	ResNet-like CNN, CNN + LSTM, Pool, CNN, LSTM, LSTM + ATTENTION	Simulated reads from MAGs sequence	Simulated reads from MAGs sequence (training excluded) Simulated mock communities of isolates Simulated reads from absent species	Metagenomic reads	Genus/species reads classification	One hot encoding K-mer embedding	# size of CNN filters # residual block # LSTM dimension # FC layers # FC units Type of pooling # window size of pooling Pooling stride # attention rows Penalization coefficient Batch size Learning rate and decay L2 regularization Activation function Optimizer	k-mer length and redundancy	k-mer embedding + LSTM + ATTENTION

(Continued)

TABLE 1 | (Continued)

Publication	Machine/deep learning category	Models tested	Training input	Tested input	Real applications input	Output	Encoding scheme	Parameters	Hyper-parameters	Best model selected
A fast and accurate functional annotator and classifier of genomic and metagenomic sequences (Sharma et al., 2015)	Machine learning supervised classification coupled to alignment method	Naive Bayesian classifier, Random Forest (RF), AdaBoost, Multiclass classifier and Lib-SVM	Peptides from eggNOG databases	Genomes Simulated metagenomic reads Real metagenomic reads	Genomic/metagenomic reads	Functional annotation of predicted genes	Compositional vectors of amino acid composition	# features # tree	NA	Random forest + RAPsearch2
DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data (Arango-Argoty et al., 2018)	Deep neural network (DNN)	Deep neural network (DNN)	UniProt genes with similarity against ARDB genes	Short gene fragments Novel AR genes	Genes Metagenomic reads	Antibiotic resistance genes prediction	Matrix of dissimilarity against AR genes	NA	NA	Deep neural network (DNN)

obtained, and the class with the highest score was attributed to the sequence. They compared their results at the strain level using BLAST as a gold standard procedure and, as a result, found similar results to BLAST in terms of accuracy (i.e., capacity of correct assignment). Their optimal k-mer length was between 9- and 15-bp lengths, depending on the length of the genomes to be detected. On the basis of these promising results, the researchers implemented a web service of their tool (Rosen et al., 2011) and added viral and fungal models (Rosen and Lim, 2012). However, in a 2017 benchmark study of 11 classifiers, the NB classifier was evaluated using simulated metagenomic data and experimental metagenomic mock communities, obtaining one of the lowest precision and recall in the benchmark (using three precision levels: strain, species, and genus) (McIntyre et al., 2017), hence showing the limitations of NB models. The low accuracy can be explained by the simplicity of the model itself or by the fact that the model did not integrate new genomes present in the tested datasets.

Support Vector Machine Models

Support Vector Machine (SVM) models are another supervised learning methodology applied to metagenomic read classification. SVMs compute the distance between the points of the datasets and try to find the hyperplane that represents the largest separation between two classes, generally using maximum margin as loss function (Han et al., 2017). Such hyperplane is determined by a kernel function (e.g., linear and Gaussian) (Steinwart and Christmann, 2008). In comparison to NB models, SVM models can handle the non-linearities of the data and take into account the interactions between data inputs. To our knowledge, the first use in metagenomics was in 2007, when McHardy et al. (2007) developed a multiclass SVM model to analyze the sequence composition of assembled metagenomic contigs to classify them into taxonomic ranges. As input training, they used the complete genome sequences of 340 organisms. In case of incomplete genomes, they arbitrary joined the contigs to obtain one sequence per genome. Different parameters were tuned to optimize the model. First, the DNA sequences were transformed into compositional vectors, as a target-encoding-like method, and then they counted the occurrence of k-mer patterns and chose the most appropriate k-mer size for a specific output class (e.g., 5 mers for genus to class levels and 6 mers for phylum and domains levels). The Gaussian and linear kernels were compared with benchmark approaches. The Gaussian kernel gave better results using cross validation. Then, the binary function for class determination of SVM was turned into multiclass using an “all vs. all” technique (i.e., performing each pair of comparison one vs. all), and the contigs are assigned to a class using a voting mechanism. To train and test the model, the genome sequences were divided into training and test data of a defined length (1, 5, 10, and 15 kb, according to the mean length of contigs retrieved in metagenomic assembling). Training data and test data came from the same genomes, but sequences sampled in training datasets were excluded from the testing datasets. Using these data, they defined the capacity of prediction using the class outputs from genus to domain taxonomic ranges, according to the length of the contig. In terms

TABLE 2 | Functional databases and their characteristics.

Functional databases	CAZy	Pfam	KEGG	eggNOG	GO Terms	MetaCyc	UniProt
Base unit	Carbohydrate-Active Enzymes	Protein domain	Ortholog gene	Ortholog gene	Vocabulary	Small-molecule metabolism	Protein
Grouping family	Protein family and sub-families	Family	Module pathway disease	Pathway	Ontology GO: Biological process Molecular function	Metabolic pathway	NA

of gene level, the sensitivity of the classifier was close to 90% for the long fragments, whereas the 1-kb fragments had a very low sensitivity percentage, close to 0. The authors tested their tools on real assembled data from different metagenomes and used as ground truth the taxonomic annotation made by state-of-the-art alignment-based tools from 2007 in the corresponding studies, making it difficult to consider as exact reality. The model was then implemented into a dedicated web server to annotate metagenomes (Patil et al., 2012).

Another SVM-based approach was recently developed in 2016 by Vervier et al. (2016), where an SVM model supports the expansion of genome sequences data availability. The authors highlighted the limit of compositional vectors approach (k-mer profile of 4, 5, and 6) for SVM model training, because the size of genome sequences is in millions of bases and the genomes available in databases increase exponentially. To overcome the problem, they optimized their model using a stochastic gradient descent, which allowed the optimization of the gradient using only one term at each step. To construct the training and test datasets, they selected three different quantities of genomes to evaluate the impact of genome numbers on the model prediction accuracy. Considering that certain alignment classifiers develop a lower common ancestor approach that allows classification of reads at different taxonomic levels, the authors built a rank-flexible approach that chooses the most adapted level to classify the reads based on the maximum score obtained with each of different rank-specific models. If a read is rejected at a specific taxonomic level, then it can be classified in upper levels if the score achieved the required threshold for the upper level. These thresholds are tunable parameters that can be optimized by taxon or set globally. The models were then tested on the remaining genomes available affiliated to the same species as the genome sequences in the training set. Moreover, they developed simulated reads that contained errors in sequencing bases, which was the first publication for machine learning classification to take into account this bias. In summary, it appeared that despite promising results on the tested genomes, especially in comparison to the NBC classifier methods, the evaluation on simulated data turned in favor of a better alignment method like kraken (Wood and Salzberg, 2014), which was less sensitive to sequencing errors and produced less false positive results.

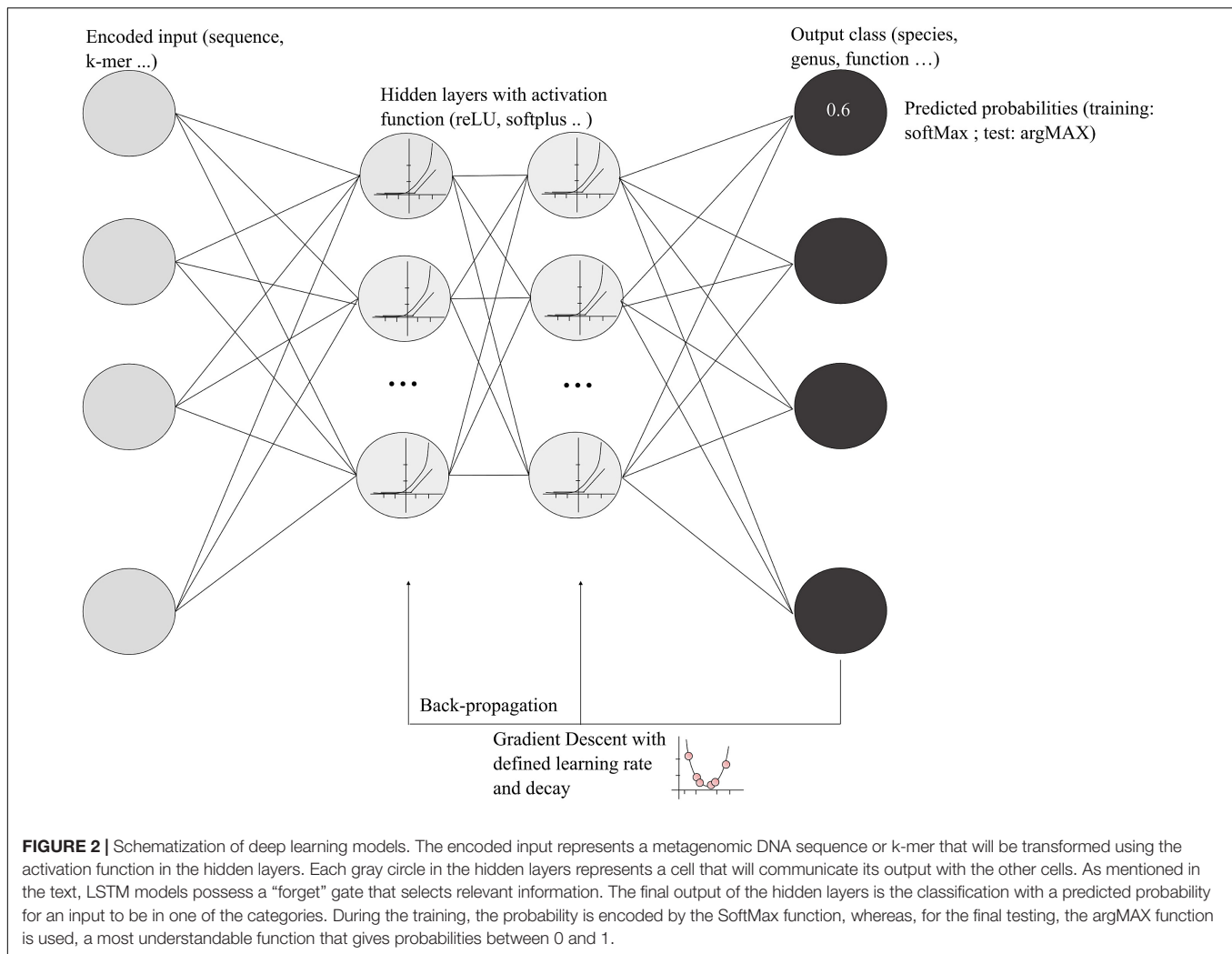
Deep Neural Network Models

Deep learning approaches are more sophisticated than classic machine learning. They may facilitate the use of large amount

of microbial genomic data available in 2022 and can take into consideration the interdependencies of input data. Deep learning refers to a category of machine learning based on artificial neural networks that generally adds more layers (hidden layers) and more units in a layer to extract more complex features from the raw input (Goodfellow et al., 2016). However, deep learning encompasses a large variety of networks, and, due to the complexity of deep learning algorithms, each model has a high number of tunable parameters. A schematization of deep neural networks (DNNs) and functions are presented in **Figure 2**, showing the main steps and associated vocabulary. An important concept in the deep learning process is the backpropagation, which allows the model to correct parameters based on the error of the network's output. Using a gradient descent algorithm with a defined learning rate and decay, the process finds the optimal weight for each neuron in each layer that minimizes the error of classification (**Figure 2**). Learning rate and decay are empiric hyper-parameters that must be defined during the model optimization.

One of the first applications of deep learning models on metagenomic classification was the use of convolutional neural networks (CNNs) and deep belief networks (DBNs) to annotate 16S fragments (Fiannaca et al., 2018). Amplicon-based metagenomics was out of scope of our review but, because they applied their model to whole genome metagenomic shotgun, we review here their model and performance. The two types of networks were compared to the 16S ribosome database project (RDP) classifier, a NB classifier for 16S data (Wang et al., 2007), which demonstrated that the CNN model had the best accuracy. The benchmark has its limitations because it has not been compared to alignment classifiers specialized in shotgun metagenomic data.

In 2020, Liang et al. (2020) analyzed different deep learning architectures for metagenomic taxonomic classification and developed a model to classify metagenomic reads based on bidirectional long short-term memory (LSTM) plus a self-attention mechanism, called DeepMicrobes. The input data used for training were taxonomically characterized MAGs from human gut metagenomes that were transformed into simulated metagenomic reads with the HiSeq 2500 Illumina sequencing error model. To evaluate the optimal parameters, the evaluation test inputs were other simulated reads from same MAGs, using another random seed of the ART simulator. Different parameters were tested such as the encoding of the input data, different DNN models, and the addition of a self-attention mechanism (a full



list of parameters is listed in **Table 1**). In total, approximately 30 parameters and hyperparameters were tested for each model, depending on if it can be tunable for the model or not (**Table 1**). As mentioned, the selected model was embedding-based recurrent self-attention model, with a batch size of 2,048, a training learning rate of 0.001 with a decay rate of 0.05, and Adam as stochastic gradient descent optimizer. During the comparison of DNN models with benchmark approaches, the authors emphasized that the one-hot encoding may be the reason why some of the models tested, ResNet like CNN, hybrid DNN, and seq2species (a deep learning model for 16S metagenomic annotation in preprint since 2019), have a low accuracy and low confidence in prediction. In contrast, k-mer embedding encoding gave better results, and an explanation made by the authors was that it considered that reverse-complementary DNA strands were the same. The results showed that the bidirectional LSTM model performed better. LSTM are recurrent neural networks, developed to process sequential data. In recurrent neural networks, the information generated by the treatment of the input goes sequentially into different cells, but this design suffers from short term memory. Therefore, LSTM models have

been developed to overcome this limitation. They possess internal states that learn to keep the relevant information and forget non-relevant data from one step to the next. This facilitates the use of long sequences as input. Finally, a self-attention mechanism was added to enable the model, to keep information at the sequence level. It enables the model to analyze the dependency of k-mers, by calculating a coefficient of relation between k-mers of a same sequence. It allows the model to focus on specific regions of the DNA sequence and the comparison of sequences with different read lengths. In fine, it increased the precision/recall score of the tested input. The best model was then compared to 2020 state-of-the-art classifiers: kraken2 (Wood et al., 2019), Centrifuge (Kim et al., 2016), Kaiju (Menzel et al., 2016), and CLARK-S (Ounit and Lonardi, 2016). As they mentioned, because there is no real metagenomic dataset that can serve as ground truth, one possibility is to simulate metagenomic samples by taking isolated sequencing reads. They thus created mock community and compared their results at the genus level because some classifiers did not contain related genomes in their native database. Globally, the DeepMicrobes model performed better than the different tools in terms of precision, recall, and

estimation of abundance of the genus level. One limitation of this article is the lack of comparison on the species level, as this information provides key insight into biological interactions. However, with the two best competitors, kraken2 and Kaiju, they obtained good results for abundance estimation at the species level even if less reads were classified. Kraken2 accuracy might have been improved by the fact that it can support larger databases than the native small database, allowing the detection of more species. To justify this key point of database dependency, they analyzed the detection of 121 genomes where species are absent from all databases of all tools and demonstrated that their model proposed less false positive results. Going through the literature highlighted that k-mer embedding encoding was already proposed for metagenomic classification (Menegaux and Vert, 2019), in a study that compared their model to the already described SVM model (Vervier et al., 2016) and burrows wheeler alignment (BWA) alignment tool (Li and Durbin, 2009). The model was not explicitly detailed, and it was based on a one-layer neural network and implemented in FastText software.¹ Because the benchmark was not compared to most efficient tools, the obtained results were difficult to evaluate.

APPLICATION OF MACHINE LEARNING TO FUNCTIONAL ANNOTATION AND OTHER SPECIFIC CASES

Machine learning models have not been yet fully applied on metagenomic functions. The only article that mentioned the utilization of machine learning models was the WOODS program, which developed a two-step pipeline, a first step of machine learning classification, and a second step of alignment annotation (Sharma et al., 2015). The machine learning step acts as pre-filter to align the reads against a specific functional category of genes. The alignment tool selected was RAPsearch2 (Zhao et al., 2012), and the functional database was eggNOG3 (Huerta-Cepas et al., 2019). The genes were regrouped into 22 functional categories, and different machine learning models were evaluated. Random forest was the best model to classify the test data, and the global pipeline achieved good results compared to the BLAST reference. However, the model was developed on complete ORF or amino acid sequences with a length larger than 500, and this makes the software useful only for assembled metagenomic data. For non-assembled metagenomic reads, this leads to the direct use of RAPsearch2.

Machine learning has also been applied to a specific functional case, the detection of antibiotic resistance gene (ARG). The screening of antibiotic resistance (AR) determinants in microbiome is a hot topic of research, as the increase of microbial resistance is a worldwide concern (Zaman et al., 2017; Nathan, 2020). To retrieve ARGs in microbiome, the analysis of shotgun metagenomic data is one of the most exhaustive ways, bypassing the culturing step. However, to retrieve these genes researchers are dependent on alignment tools and related databases. As alignment-based methods are not perfect and can produce false positive results (AR can be derived from non-ARG such as

efflux pump) and false negative results (no detection of genes variants from databases), applying learning models can be an efficient way to detect these genes. This was tested by Arango-Argoty et al. (2018) they proposed a new tool named DeepARG that contains two deep learning models to retrieve 30 classes of antibiotic determinants in metagenomic reads or full gene sequences, respectively. The model took as input a dissimilarity matrix based on the alignment bitscore of reads/genes mapped to an ARG database. Although the tool was also based on alignment score, the accuracy of ARG classes prediction compared to the alignment-based method was improved. This was explained by the fact that the deep learning model application did not require a set general threshold of similarity (i.e., percentage of similarity), instead allowing adaptation of the threshold function to the ARG classes (done in the training part). The proposed model may have been tuned with different combinations of parameters, but the article did not mention the different tests performed.

To analyze the pertinence of machine learning applications in functional metagenomic screening, the development of a methodology that analyzes the sequences by itself (with k-mer embedding for instance) and not a global score of dissimilarity matrix remains to be evaluated. Sequence identity threshold in functional screening is not extensively documented, although it is a critical key point of the functional annotation. A common threshold to assess the function is 30% of sequence similarity, even if a common value for different functions is highly critical (Pearson, 2013). HUMAnN3 (Beghini et al., 2021), a recent functional annotation pipeline, sets the identity threshold to 50% yet advises the user to configure the settings. LSTM models developed for taxonomic annotation, which allow the models to focus on specific parts of the sequences, may be a promising candidate to identify and annotate functional data.

CONCLUSION

Machine learning has been applied because the beginning of metagenomic annotation, but the increase of available microbial genomic data in databases leads to the obsolescence of the first models, too simple to accommodate the size and complexity of the data. Their accuracy was reduced in comparison to k-mer-based tools in the reviewed benchmark. Because the integration of genomic data is feasible in deep learning models, two models have been published for taxonomic annotation. The first one was not compared with enough benchmarks to conclude on their progress, and the second named DeepMicrobes demonstrated good performance, even compared to state-of-the-art alignment-based classifiers. The tool highlighted the benefit of k-mer embedding for the input treatment and the use of networks such as LSTM that learns important long-range interactions and “forgets” information not discriminant to build the model. The comparisons to the other tools were mostly achieved at the genus level, but a benchmark to the species level would have been of interest in terms of interpretation. In functional annotation, deep learning technologies have been applied to specific questions, or to build a model for pre-classification, but remain to be studied for a full functional annotation. Because no real microbiomes are

¹<https://fasttext.cc/>

known without the prism of metagenomic tools, the benchmarks in metagenomic annotation are based on simulated data or mock communities that present a reduced diversity. This leads the benchmarks to be case specific, and the tools developed to be overfitted to the generated data. Moreover, the data available in databases represent a low percentage of the overall microbial diversity, leading to the construction of models specific to what is known in databases. Specific machine learning algorithms have been proposed to answer these specific cases, as active learning that allows the selection of relevant data from the training set to improve the models and not overfit to the data. Active learning may be a framework that facilitates the building of models with high accuracy, by selecting certain data of the input to train the models (Settles, 2009). Despite the possibility of biases due to the targeted sampling, it may overcome the pitfalls of metagenomics (i.e., database orientation to certain bacterial species, the use of reconstructed genomes with no taxonomic annotation, and the deletion of non-informative sequences). Finally, as bacterial genome sequences in databases are still in expansion, current developed models have to be regularly tested/updated to remain up to date. All programs developed and commented in this article provide useful information to build the most adapted annotation framework. Because in the field of metagenomics

data availability and computational resource accessibility increase at a relatively high rate, current models may become obsolete and new models will be constructed. This must be done based on already developed algorithms and the use of successful tested parameters.

AUTHOR CONTRIBUTIONS

AM elaborated the review plan, conducted literature searches, researched data, selected relevant articles, created the figures and tables, wrote, formatted, and finalized the article for submission. ML and MS helped to elaborate the review plan, proposed corrections, and validated the approaches linked to machine learning. OP and AD supervised the work and helped to improve the manuscript. All authors contributed to writing and reviewing the manuscript.

FUNDING

AM, ML, and AD were supported by Research and Innovation Chair L'Oreal in Digital Biology.

REFERENCES

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.
- Arango-Argoty, G., Garner, E., Pruden, A., Heath, L. S., Vikesland, P., and Zhang, L. (2018). DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome* 6, 1–15. doi: 10.1186/s40168-018-0401-z
- Bahram, M., Netherway, T., Frioux, C., Ferretti, P., Coelho, L. P., Geisen, S., et al. (2021). Metagenomic assessment of the global diversity and distribution of bacteria and fungi. *Environ. Microbiol.* 23, 316–326. doi: 10.1111/1462-2920.15314
- Beghini, F., McIver, L. J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., et al. (2021). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with biobakery 3. *Elife* 10:e65088. doi: 10.7554/eLife.65088
- Breitwieser, F. P., Lu, J., and Salzberg, S. L. (2019). A review of methods and databases for metagenomic classification and assembly. *Brief Bioinform.* 20, 1125–1136. doi: 10.1093/bib/bbx120
- Buchfink, B., Xie, C., and Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60. doi: 10.1038/nmeth.3176
- Calle, M. L. (2019). Statistical analysis of metagenomics data. *Genomics Inform.* 17:e6. doi: 10.5808/GI.2019.17.1.e6
- Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C. A., et al. (2014). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.* 42, D459–D471.
- Delmont, T. O., Quince, C., Shaiber, A., Esen, ÖC., Lee, S. T., Rappé, M. S., et al. (2018). Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nat. Microbiol.* 3, 804–813. doi: 10.1038/s41564-018-0176-9
- Erickson, A. R., Cantarel, B. L., Lamendella, R., Darzi, Y., Mongodin, E. F., Pan, C., et al. (2012). Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of crohn's disease. *PLoS One* 7:e49138. doi: 10.1371/journal.pone.0049138
- Escobar-Zepeda, A., de León, A. V.-P., and Sanchez-Flores, A. (2015). The road to metagenomics: from microbiology to DNA sequencing technologies and bioinformatics. *Front. Genet.* 6:348. doi: 10.3389/fgene.2015.00348
- Fiannaca, A., La Paglia, L., La Rosa, M., Lo Bosco, G., Renda, G., Rizzo, R., et al. (2018). Deep learning models for bacteria taxonomic classification of metagenomic data. *BMC Bioinformatics* 19:198. doi: 10.1186/s12859-018-2182-6
- Gene Ontology Consortium [GOC] (2021). The Gene Ontology resource: enriching a Gold mine. *Nucleic Acids Res.* 49, D325–D334. doi: 10.1093/nar/gkaa1113
- Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Cambridge: MIT Press.
- Greener, J. G., Kandathil, S. M., Moffat, L., and Jones, D. T. (2021). A guide to machine learning for biologists. *Nat. Rev. Mol. Cell Biol.* 2021, 40–55. doi: 10.1038/s41580-021-00407-0
- Han, W., Wang, M., and Ye, Y. A. (2017). Concurrent subtractive assembly approach for identification of disease associated sub-metagenomes. *Res. Comput. Mol. Biol.* 2017, 18–33. doi: 10.1007/978-3-319-56970-3_2
- Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., et al. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 47, D309–D314. doi: 10.1093/nar/gky1085
- Inkpen, S. A., Douglas, G. M., Brunet, T. D. P., Leuschen, K., Doolittle, W. F., and Langille, M. G. I. (2017). The coupling of taxonomy and function in microbiomes. *Biol. Philos.* 32, 1225–1243.
- Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 44, D457–D462. doi: 10.1093/nar/gkv1070
- Kim, D., Song, L., Breitwieser, F. P., and Salzberg, S. L. (2016). Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.* 26, 1721–1729. doi: 10.1101/gr.210641.116
- Kroeger, M. E., Delmont, T. O., Eren, A. M., Meyer, K. M., Guo, J., Khan, K., et al. (2018). New biological insights into how deforestation in amazonia affects soil microbial communities using metagenomics and metagenome-assembled genomes. *Front. Microbiol.* 9:1635. doi: 10.3389/fmicb.2018.01635
- Lee, S. T. M., Kahn, S. A., Delmont, T. O., Shaiber, A., Esen, ÖC., Hubert, N. A., et al. (2017). Tracking microbial colonization in fecal microbiota transplantation experiments via genome-resolved metagenomics. *Microbiome* 5:50. doi: 10.1186/s40168-017-0270-x

- Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324
- Liang, Q., Bible, P. W., Liu, Y., Zou, B., and Wei, L. (2020). DeepMicrobes: taxonomic classification for metagenomics with deep learning. *NAR Genom. Bioinform.* 2:lqaa009. doi: 10.1093/nargab/lqaa009
- Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M., and Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* 42, D490–D495. doi: 10.1093/nar/gkt1178
- Loomba, R., Seguritan, V., Li, W., Long, T., Klitgord, N., Bhatt, A., et al. (2017). Gut microbiome based metagenomic signature for non-invasive detection of advanced fibrosis in human nonalcoholic fatty liver disease. *Cell Metab.* 25, 1054–1062.e5.
- McHardy, A. C., Martín, H. G., Tsirigos, A., Hugenholtz, P., and Rigoutsos, I. (2007). Accurate phylogenetic classification of variable-length DNA fragments. *Nat. Methods* 4, 63–72. doi: 10.1038/nmeth976
- McIntyre, A. B. R., Ounit, R., Afshinnikoo, E., Prill, R. J., Hénaff, E., Alexander, N., et al. (2017). Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome Biol.* 18, 1–19.
- Menegaux, R., and Vert, J. P. (2019). Continuous embeddings of DNA sequencing reads and application to metagenomics. *J. Comput. Biol.* 26, 509–518. doi: 10.1089/cmb.2018.0174
- Menzel, P., Ng, K. L., and Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.* 7:11257. doi: 10.1038/ncomms11257
- Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., et al. (2021). Pfam: the protein families database in 2021. *Nucleic Acids Res.* 49, D412–D419. doi: 10.1093/nar/gkaa913
- Nathan, C. (2020). Resisting antimicrobial resistance. *Nat. Rev. Microbiol.* 18, 259–260. doi: 10.1038/s41579-020-0348-5
- O’Leary, N. A., Wright, M. W., Brister, J. R., Ciufu, S., Haddad, D., McVeigh, R., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44, D733–D745. doi: 10.1093/nar/gkv1189
- Ounit, R., and Lonardi, S. (2016). Higher classification sensitivity of short metagenomic reads with CLARK-S. *Bioinformatics* 32, 3823–3825. doi: 10.1093/bioinformatics/btw542
- Pasolli, E., Truong, D. T., Malik, F., Waldron, L., and Segata, N. (2016). Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. *PLoS Comput. Biol.* 12:e1004977. doi: 10.1371/journal.pcbi.1004977
- Patil, K. R., Roun, L., and McHardy, A. C. (2012). The phylopythias web server for taxonomic assignment of metagenome sequences. *PLoS One* 7:e38581. doi: 10.1371/journal.pone.0038581
- Pearson, W. R. (2013). An introduction to sequence similarity (“homology”) searching. *Curr. Protoc. Bioinform.* 3:10.1002/0471250953.bi0301s42. doi: 10.1002/0471250953.bi0301s42
- Pedron, R., Esposito, A., Bianconi, I., Pasolli, E., Tett, A., Asnicar, F., et al. (2019). Genomic and metagenomic insights into the microbial community of a thermal spring. *Microbiome* 7:8. doi: 10.1186/s40168-019-0625-6
- 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
- Rosen, G. L., and Lim, T. Y. (2012). NBC update: the addition of viral and fungal databases to the Naïve Bayes classification tool. *BMC Res. Notes* 5:81. doi: 10.1186/1756-0500-5-81
- Rosen, G. L., Reichenberger, E. R., and Rosenfeld, A. M. (2011). NBC: the Naïve Bayes Classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics* 27, 127–129. doi: 10.1093/bioinformatics/btq619
- Rosen, G., Garbarine, E., Caseiro, D., Polikar, R., and Sokhansanj, B. (2008). Metagenome fragment classification using N-Mer frequency profiles. *Adv Bioinform.* 2008:205969. doi: 10.1155/2008/205969
- Sandberg, R., Winberg, G., Bränden, C.-I., Kaske, A., Ernberg, I., and Cöster, J. (2001). Capturing whole-genome characteristics in short sequences using a naïve bayesian classifier. *Genome Res.* 11, 1404–9. doi: 10.1101/gr.186401
- Settles, B. (2009). *Active Learning Literature Survey*. Madison: Univ Wisconsin–Madison.
- Sharma, A. K., Gupta, A., Kumar, S., Dhakan, D. B., and Sharma, V. K. (2015). Woods: a fast and accurate functional annotator and classifier of genomic and metagenomic sequences. *Genomics* 106, 1–6. doi: 10.1016/j.ygeno.2015.04.001
- Steinwart, I., and Christmann, A. (2008). *Support Vector Machines*. Berlin: Springer Science & Business Media.
- The UniProt Consortium, Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Agivetova, R., et al. (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* 49, D480–D489. doi: 10.1093/nar/gkaa1100
- Treiber, M. L., Taft, D. H., Korf, I., Mills, D. A., and Lemay, D. G. (2020). Pre- and post-sequencing recommendations for functional annotation of human fecal metagenomes. *BMC Bioinformatics* 21:74. doi: 10.21203/rs.2.16066/v3
- Vervier, K., Mahé, P., Tournoud, M., Veyrieras, J.-B., and Vert, J.-P. (2016). Large-scale machine learning for metagenomics sequence classification. *Bioinformatics* 32, 1023–1032.
- Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naive bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* 73, 5261–5267. doi: 10.1128/AEM.00062-07
- Wayne, L. G., Brenner, D. J., Colwell, R. R., Grimont, P. A. D., Kandler, O., Krichevsky, M. I., et al. (1987). Report of the Ad Hoc committee on reconciliation of approaches to bacterial systematics. *Int. J. Syst. Evol. Microbiol.* 37, 463–464.
- Wood, D. E., and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 15:R46.
- Wood, D. E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biol.* 20:257. doi: 10.1186/s13059-019-1891-0
- Zaman, S. B., Hussain, M. A., Nye, R., Mehta, V., Mamun, K. T., and Hossain, N. (2017). A review on antibiotic resistance: alarm bells are ringing. *Cureus* 9:e1403. doi: 10.7759/cureus.1403
- Zhao, Y., Tang, H., and Ye, Y. (2012). RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics* 28, 125–126. doi: 10.1093/bioinformatics/btr595
- Zhong, H., Ren, H., Lu, Y., Fang, C., Hou, G., Yang, Z., et al. (2019). Distinct gut metagenomics and metaproteomics signatures in prediabetics and treatment-naïve type 2 diabetics. *EBioMedicine* 47, 373–383. doi: 10.1016/j.ebiom.2019.08.048

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.

Copyright © 2022 Mathieu, Leclercq, Sanabria, Perin and Droit. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Study on the Bacterial Communities of the Biofilms on Titanium, Aluminum, and Copper Alloys at 5,772 m Undersea in Yap Trench

Xiaofan Zhai^{1,2,3,4†}, Wei Cao^{1†}, Yimeng Zhang^{2,3,4}, Peng Ju^{1*}, Juna Chen^{5*}, Jizhou Duan^{2,3,4} and Chengjun Sun^{1*}

OPEN ACCESS

Edited by:

Mohammad-Hossein Sarrafzadeh,
University of Tehran, Iran

Reviewed by:

Fariba Rezvani,
Iranian Research Organization
for Science and Technology, Iran
Neda Fakhimi,
Carnegie Institution for Science (CIS),
United States

*Correspondence:

Peng Ju
jupeng@fio.org.cn
Juna Chen
jys130@163.com
Chengjun Sun
csun@fio.org.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: 09 December 2021

Accepted: 25 January 2022

Published: 18 March 2022

Citation:

Zhai X, Cao W, Zhang Y, Ju P,
Chen J, Duan J and Sun C (2022)
Study on the Bacterial Communities
of the Biofilms on Titanium,
Aluminum, and Copper Alloys
at 5,772 m Undersea in Yap Trench.
Front. Microbiol. 13:831984.
doi: 10.3389/fmicb.2022.831984

¹ Key Laboratory of Marine Eco-Environmental Science and Technology, Marine Bioresource and Environment Research Center, First Institute of Oceanography, Ministry of Natural Resources, Qingdao, China, ² CAS Key Laboratory of Marine Environmental Corrosion and Bio-Fouling, Institute of Oceanology, Chinese Academy of Sciences, Qingdao, China, ³ Open Studio for Marine Corrosion and Protection, Pilot National Laboratory for Marine Science and Technology (Qingdao), Qingdao, China, ⁴ Center for Ocean Mega-Science, Chinese Academy of Sciences, Qingdao, China, ⁵ Navy Submarine Academy, Qingdao, China

Biofilms formed on metal surfaces strongly affect metallic instruments serving in marine environments. However, due to sampling difficulty, less has been known about the bacterial communities of the biofilm on metallic surfaces in hadal environments, so the failure process of these deep-sea metallic instruments influenced by microbial communities could be hardly predicted. In this research, seven alloys, including titanium, aluminum, and copper alloys, were exposed in Yap Trench hadal environment for 1 year. Thus, the communities of the biofilms formed on metallic surfaces at 5,772 m undersea in Yap Trench were initially reported in previous studies. Then, 16S rRNA gene sequencing was performed to visualize the *in situ* bacterial communities of the biofilms formed on titanium, aluminum, and copper alloys at 5,772 m undersea in Yap Trench. It was found that Proteobacteria was the dominant phylum in all samples, but distinct genera were discovered on various alloys. The titanium alloy provided a suitable substrate for a mutualistic symbiotic biofilm with abundant bacterial richness. Aluminum alloys without copper components showed the least bacterial richness and formed a cold-adapted and oligotrophic-adapted biofilm containing the genera *Sulfurimonas* and *PS1 Clade*, while copper-present alloys showed relatively high bacterial richness with copper-resistant or even copper-utilizing biofilms constituting the genera *Stenotrophomonas*, *Burkholderia-Caballeronia-Paraburkholderia*, and *Achromobacter* on the surfaces. Furthermore, among all the element components contained in alloys investigated in this research, copper element showed the strongest influences on the composition and function of microbial communities in the biofilms formed on various metallic surfaces.

Keywords: microbial communities, biofilm, metal alloys, hadal environment, Yap Trench

INTRODUCTION

The hadal biosphere at the deep-sea environment in Yap Trench has been less studied and poorly explored. Until the recent years, microorganisms started to come to light owing to technical development (Li L. et al., 2020; Zhang et al., 2021). Microbial diversity in seawater and sediments was initially reported by the researchers. The microbial community composition in seawater was found to be dominated by Gammaproteobacteria with heterotrophic processes as the most common metabolisms (Zhang et al., 2018), whereas in the sediments, the microbial populations had fluctuating distributions and chemolithoautotrophic metabolic processes dominated by Proteobacteria and Thaumarchaeota (Fu et al., 2020). The studies showed that, in this dark realm, unique and highly adapted microbial communities have formed. Especially the detection/enrichment of genes involved in stress response and metal resistance in the seawater and sediment of the Yap Trench suggested special adaptation strategies of the hadal microorganisms toward high pressure and/or nutrient availability, while the enrichment of metal resistance genes might be a hypothesized characteristic of the hadal seawater microbial communities (Zhang et al., 2018). Besides that, a typical “V-shape” topography, and frequent sediment collapses on trench walls, high total organic carbon (TOC%) and total nitrogen (TN%) were found in this environment, especially in the core sediments with distinct microbial populations of *Proteobacteria* and *Thaumarchaeota* (Fu et al., 2020). However, the details of the species and functions are still unknown.

At present, further studies on hadal environments are highly dependent on the advanced and expensive metallic instruments that are capable of serving in these extremely low-temperature and high-pressure environments. However, according to reports on metallic instruments serving in the offshore area, microorganisms play an important role on metal failure, which is called microbiologically influenced corrosion (Remazeilles et al., 2010; Zhao et al., 2018; Zhou et al., 2018; Ma et al., 2020). It was found that various bacteria showed different effects on metal failure process, i.e., corrosion acceleration, corrosion inhibition, or irrelevance—for example, sulfate-reducing bacteria are recognized as the corrosion-accelerating bacteria (Enning and Garrelfs, 2014; Guan et al., 2020), while some metal-reducing bacteria have been proven to successfully inhibit metal corrosion (Zuo, 2007). However, in a real marine environment, especially in the little-known Yap Trench environment, the biofilms formed on metallic surfaces are complex, heterogeneous, and far more than the reported sulfate-reducing bacteria and metal-reducing bacteria (Li et al., 2017; Zhang et al., 2019a).

Previous studies revealed that dramatic differences showed up between the communities in surrounding seawater and the biofilms on various metallic surfaces. Thus, the microbial diversity in seawater and sediments provide us limited knowledge on analyzing the feasibility and predicting the failure of metallic instruments. Until now, nothing about the influence of hadal communities on these metallic materials is known. Hence, clarifying the microbial compositions on metallic surfaces makes

a significant sense to predict the safety and service life of metallic instruments applied in Yap Trench.

What is more, the biofilms formed on metallic surfaces are not only highly dependent on the environment but also closely related to metal-inherent qualities, such as element components, alloy phases, and so on (Dang and Lovell, 2016). These inherent qualities make metal alloys display various surface status, including surface free energy, roughness, hydrophilic/lipophilic property, and electrostatic charge, which attract certain microorganisms to adhere—for example, no electronically charged surface was more attractive to marine *Pseudomonas* sp. rather than the hydrophilic and negatively charged surface (Fletcher and Loeb, 1979). It was also proved that diverse microbial communities develop on the surfaces of metallic plates, which differed from the surrounding oligotrophic bacteria in seawater (Li et al., 2017; Zhang et al., 2019a). Furthermore, marine surface-associated biofilms formed on the copper alloys possess distinct microbial compositions compared with those formed on aluminum alloys (Zhang et al., 2019b). As a result, figuring out the bacterial communities of the biofilm on metallic surfaces in hadal environments would greatly help researchers to evaluate the microbial influence on metallic instruments, which might be favorable to predict the failure process of these metallic instruments employed in deep-sea environments.

In this work, several typical metal alloys, including titanium alloy, aluminum alloy, and copper alloy, which are commonly used for deep-sea instruments were employed as testing substrates. These alloys were exposed in Yap Trench for 1 year to observe the bacterial communities of the formed biofilm. According to this research, an initial attempt is made to know more about hadal environments and concerned more on metallic instruments serving in Yap Trench.

EXPERIMENTAL

Sample Collection

Titanium alloy TA2, aluminum alloy ZAL, aluminum alloy 5A06, aluminum alloy 1060, copper alloy T2, copper alloy B10, and copper alloy B30 were employed in this study. The composition of the alloys is shown in **Tables 1–3**. Coupons of 40 mm × 120 mm × 5 mm, made of alloy TA2, alloy ZAL, alloy 5A06, alloy 1060, alloy T2, alloy B10, and alloy B30, were prepared separately for seawater immersion tests. The coupons were fixed in an insulated frame cage which was fastened on a subsurface buoy. This buoy was exposed in Yap Trench (138°43′434″ E, 9°51′0215″ N) at 5,772 m undersea. The salinity of the seawater was detected as 3.47%, and the temperature was 1.58°C. The pressure at the exposure location was determined to be 5,881 MPa.

The immersion started from 23 May 2016 to 2 June 2017, which lasted for 375 days. After exposure, coupons were stored at −20°C until they were taken back to the laboratory.

A sterilized soft brush was used to scrape the biofilm from each coupon surface. Then, the biofilm was transferred into a sterilized beaker with phosphate-buffered saline. The biomass in solution

TABLE 1 | Chemical composition of the alloys.

	Alloy TA2	Alloy ZAL	Alloy 5A06	Alloy 1060	Alloy T2	Alloy B10	Alloy B30
Ti (%)	Residual	0.15–0.35	0.02–0.10	≤0.03	/	/	/
Al (%)	/	Residual	Residual	Residual	/	/	/
Cu (%)	/	4.50–5.30	≤0.10	≤0.05	Residual	Residual	Residual
Fe (%)	≤0.30	≤1.00	0.00–0.40	/	≤0.005	≤0.02	≤0.90
C (%)	≤0.15	/	/	/	≤0.03	≤0.03	≤0.05
N (%)	≤0.05	/	/	/	/	/	/
O (%)	≤0.20	/	/	/	/	/	/
Mn (%)	/	0.60–1.00	0.50–0.8	≤0.03	/	/	≤1.20
Mg (%)	/	≤0.05	5.8–6.8	≤0.03	/	/	/
Si (%)	/	≤0.30	≤0.40	≤0.25	/	/	≤0.15
Zn (%)	/	≤0.20	≤0.20	≤0.05	/	/	/
V (%)	/	/	/	≤0.05	/	/	/
Ni (%)	/	≤0.10	/	/	/	/	/
Zr (%)	/	≤0.20	/	/	/	/	/
Ni–Co (%)	/	/	/	/	/	9.50–10.50	29.00–33.00
Pb (%)	/	/	/	/	≤0.005	≤0.01	≤0.05
S (%)	/	/	/	/	≤0.01	≤0.01	≤0.01
Bi (%)	/	/	/	/	≤0.02	≤0.02	/
Sb (%)	/	/	/	/	≤0.005	≤0.005	/
P (%)	/	/	/	/	≤0.01	≤0.01	≤0.006
As (%)	/	/	/	/	≤0.01	≤0.01	/

The chemical element compositions (mass fraction%) of alloy TA2, alloy ZAL, alloy 5A06, alloy 1060, alloy T2, alloy B10, and alloy B30 employed in this research were according to national standards GB/T 3620.1-2016, GB/T 3190-2008, and GB/T 5231-2001.

TABLE 2 | Diversity estimators for bacteria from seven metallic surface samples exposed in Yap Trench using 16S rRNA gene sequencing.

Sample	Observed species	Good coverage	Chao1	Faith's phylogenetic diversity	ACE	Shannon	Simpson
TA2	556	0.9987	569	57.10	573	4.33	0.78
ZAL	383	0.9987	407	34.65	407	5.25	0.92
Al5A06	241	0.9994	248	28.42	258	4.34	0.88
Al1060	197	0.9997	205	19.15	205	4.82	0.92
T2	555	0.9988	573	45.52	565	5.67	0.91
B10	366	0.9989	380	38.15	379	5.01	0.92
B30	374	0.9990	389	34.56	375	4.26	0.84

TA2, ZAL, Al5A06, Al1060, T2, B10, and B30 represented the biofilms collected from the corresponding alloy surfaces. The observed species, Chao1 and ACE, represented the species richness of each sample. The Shannon and Simpson indices of these samples revealed community diversities, and Faith's phylogenetic diversity evaluated the evolution differences.

TABLE 3 | Relative abundance of 16S rRNA gene sequences of the seven samples exposed in Yap Trench for 1 year at the bacterial class level.

Sample	TA2	ZAL	Al5A06	Al1060	T2	B10	B30
Alphaproteobacteria	58%	22%	19%	21%	29%	9%	40%
Gammaproteobacteria	28%	65%	41%	41%	47%	65%	48%
Actinobacteria	3%	3%	0%	0%	4%	1%	5%
Bacteroidia	2%	1%	6%	10%	3%	1%	1%
Bacilli	2%	5%	0%	1%	6%	20%	1%
Campylobacteria	1%	0%	29%	26%	1%	–	0%
Deltaproteobacteria	1%	0%	5%	1%	1%	0%	0%
Clostridia	1%	1%	0%	–	55%	2%	1%

TA2, ZAL, Al5A06, Al1060, T2, B10, and B30 represented the biofilms collected from the corresponding alloy surfaces. The top five abundant classes of each sample are shown in this table.

was further filtrated through 0.22- μ m filter membranes to obtain a concentrate of the microorganisms (Mittelman et al., 1997). These biofilm samples collected from alloy TA2, alloy ZAL, alloy 5A06, alloy 1060, alloy T2, alloy B10, and alloy B30 were named as TA2, ZAL, Al5A06, Al1060, T2, B10, and B30, respectively. Based on the differences of the major component in each metallic alloy, TA2, as a representative titanium alloy, was named as group A1; ZAL, Al5A06, and Al1060, as representatives of aluminum alloy, were named as group A2; and T2, B10, and B30, as representatives of copper alloy, were named as group A3.

DNA Extraction and Sequencing

The DNA of these microorganisms was extracted using a previously reported method. DNA concentration and purity were determined with a spectrophotometer (Lambda 1A; Perkin-Elmer). A A260/A280 ratio of 1.8–2.1 was considered acceptable for PCR-based procedures (Zhang et al., 2019a). The extracted DNA of the biofilms was used as a template to amplify the 16S rRNA genes by PCR with the universal forward primer 338F (5'-ACTCCTACGGGAGGCAGCA-3') and reverse primer 806R (5'-GGACTACHVGGGTWTCTAAT-3'). PCR purification kit (QIAGEN, Hilden, NRW, Germany) was used to purify the PCR products. The PCR libraries were conducted using TruSeq DNA PCRFree Sample Preparation Kit (Illumina, San Diego, CA, United States). After quantification with Qubit, the PCR libraries were sequenced on the Illumina HiSeq PE250 platform.

Sequence Data Analysis

Based on the unique barcodes of each sample, raw paired-end reads were assigned. Subsequently, FLASH (V1.2.7) was used to merge these reads according to their overlap after the barcodes and primer cuts (Magoc and Salzberg, 2011). Then, based on the process for quality control in QIIME, these sequences were filtered, followed by detecting and removing the chimera sequences by UCHIME algorithm (Caporaso et al., 2010; Edgar et al., 2011). Operational taxonomic units (OTUs) were clustered with 97% similarity using UPARSE software, version 7.11 (Edgar, 2013). The taxonomy of each 16S rRNA gene sequence was analyzed by RDP Classifier 2.2 against the GreenGene database (DeSantis et al., 2006; Wang et al., 2007).

Then, the sequence data were normalized after unique tags dislodge to analyze OTU abundance and diversity index. Good coverage was calculated by QIIME to represent sequencing depth. Alpha diversity indices were employed to indicate the bacterial diversity of each sample, including Chao and ACE for species richness, Simpson and Shannon for community diversity evaluating both the species richness and evenness, and Faith's phylogenetic diversity (Faith pd) for phylogenetic diversity. Besides these, beta diversity using clustering analysis and principal coordinate analysis based on unweighted unifrac distances were employed for the comparison of the community differences between groups. Furthermore, Venn diagrams were used to show the unique and shared OTUs of the three groups, and PICRUSt2 (Langille et al., 2013) was employed to predict the functional genes based on the 16S rRNA sequencing data, which were annotated against the Kyoto Encyclopedia of Genes and Genomes (KEGG) database V2018-01 (Kanehisa et al., 2017).

Statistical tests based on analysis of variance were used to determine the difference in functional gene abundance, and factors with *p*-values less than 0.05 were considered to have a significant difference.

Data Availability

The raw sequences obtained were deposited in the NCBI Short Read Archive database under Bioproject accession number PRJNA438021, with Biosample numbers SAMN23711893-23711899.

RESULTS

Microbial Richness and Diversity of Biofilm on the Alloys

A total of 331,905 high-quality bacterial sequences, ranging from 38,102 to 58,460, were obtained for further analysis.

As shown in **Table 2**, the bacterial coverage ranged from 99.87 to 99.97%, indicating that the sequences obtained by V3–V4 Illumina sequencing captured their core microbial communities. In addition, all rarefaction curves of bacteria reached saturation, revealing that the amount of sequencing data was enough to capture the great majority of bacterial communities (**Figure 1**). The observed species and Chao1, which represented species richness of each sample, were quite different in these samples. In total, 556 species (Chao1 index 569 and ACE index 573) were found on titanium alloy TA2. On average, 274 observed species (Chao1 index 287 and ACE index 290) were found on aluminum alloys, and 432 observed species (Chao1 index 447 and ACE index 440) were on copper alloys. The biofilm on ZAL alloy showed the highest species richness among the three aluminum alloys, and the biofilm on T2 alloy showed the highest value among copper alloys, although the average values of the observed species, Chao1 index, and ACE index showed the following trend: titanium alloy > copper alloy > aluminum alloy. The Shannon and Simpson indices of these samples, revealing community diversities, showed similar results. The Shannon and Simpson indices of TA2 were 4.33 and 0.78, respectively. The average Shannon and Simpson indices of aluminum alloys were 4.80 and 0.91, while those of copper alloys were calculated as 4.98 and 0.89. Faith pd (shown in **Table 2**) was used to evaluate the evolution differences. The Faith pd indices of TA2 (57.10) and T2 (45.52) were obviously higher than the other samples, showing relatively high phylogenetic diversities.

Comparison of the Microbial Composition of Biofilms on the Alloys

Differences in the composition of the bacterial community were detected for the seven alloys, i.e., A1 group (TA2), A2 group (ZAL, Al5A06, and Al1060), and A3 group (T2, B10, and B30).

As shown in **Figure 2A**, in the Venn diagram, there were 556 OTUs shown in group A1 (titanium alloy), 610 OTUs in group A2 (aluminum alloys), and 1,043 OTUs in group A3 (copper alloys). They shared only 94 OTUs at group level. The A2 and A3 groups shared more OTUs (116 OTUs) than those they shared with the

A1 group (i.e., 40 OTUs for A1 and A2 groups and 50 OTUs for A1 and A3 groups). As to the sample level (**Figure 2B**), all seven samples only shared 17 OTUs, revealing dramatic differences in species composition. TA2 showed the highest unique OTUs of up to 67%, while Al1060 showed the lowest at 28%. Even in each group, the shared OTUs were found to be as low as 55 OTUs for the aluminum alloy samples (**Figure 2C**) and 63 OTUs for the copper alloy samples (**Figure 2D**).

The average pair-group method with an arithmetic mean based on unweighted unifracs distances was performed to determine the differences between these samples, as shown in **Figure 3**. The clustering analysis showed that these samples could be divided into three groups: titanium alloy (TA2); copper alloys and aluminum alloy with copper element (T2, B10, B30, and ZAL); and aluminum alloys without copper (Al1060 and Al5A06). Furthermore, principal coordinate analysis (PCoA) plots of unweighted unifracs distances based on OTUs are shown in **Figure 4**. The results showed that different brands of each alloy were similar, such as the aluminum alloys without the copper group and the copper alloy group. An exception was found on ZAL, a kind of aluminum alloy containing copper, which was similar to copper alloys instead of aluminum alloys.

Bacterial Community Compositions of Biofilms on the Alloys

In total, more than 21 bacterial phyla were found in these samples. In terms of the average abundance of seven samples, Proteobacteria was found to be the dominant phylum, accounting for 77% of the total sequences (**Figure 5**). Then, it was followed by Epsilonbacteraeota accounting for 8%, Firmicutes accounting for 7%, Bacteroidetes accounting for 3%, and Actinobacteria accounting for 3%. Cyanobacteria, Acidobacteria, Planctomycetes, and Patescibacteria were also found in the samples with a relatively low proportion.

Corresponding to the hierarchical cluster tree and PCoA plot results, Al5A06 and Al1060 showed similar community compositions, while T2, B10, B30, and ZAL clustered more closely with each other in general. At the phylum level (**Figure 5**), Proteobacteria was the dominant phylum in all samples, ranging from 62 to 89%. Firmicutes were found as the second represented phylum on titanium alloy TA2 and copper-present alloys B10, B30, T2, and ZAL. However, Epsilonbacteraeota was the secondary represented phylum on non-copper aluminum alloys Al1060 and Al5A06. On the class level, Alphaproteobacteria (58%) was the dominant class on titanium alloy TA2, while Gammaproteobacteria (40–65%) was the dominant class on the copper alloys and aluminum alloys. Gammaproteobacteria (28%) took the second place, followed by Actinobacteria (3%), Bacteroidia (2%), Bacilli (2%), Campylobacteria (1%), and Deltaproteobacteria (1%) on TA2. Furthermore, Alphaproteobacteria (19–40%) took the second place on copper alloys and aluminum alloys except for B10, on which Bacilli was found to be the secondary (**Table 3**).

However, distinct dominant bacteria were found on different metals at the genus level (**Figure 6**). *PS1 Clade* (45%), *Stenotrophomonas* (13%), and *Acinetobacter* (3%) made up the dominant genus on TA2. *Sulfurimonas* (27% on average) and

PS1 Clade (15% on average) were dominant on non-copper aluminum alloys Al1060 and Al5A06, while *Stenotrophomonas* (15%), *Cobetia* (14%), and *Vibrio* (14%) were dominant on the copper-present aluminum alloy ZAL. *Stenotrophomonas* (18% on average), *Burkholderia–Caballeronia–Paraburkholderia* (6% on average) and *Achromobacter* (3% on average) formed the communities on copper alloys B10, B30, and T2. These results illustrated that the composition of microbial communities attached on the metal surfaces highly depended on the metal composition.

Key Functional Gene Prediction

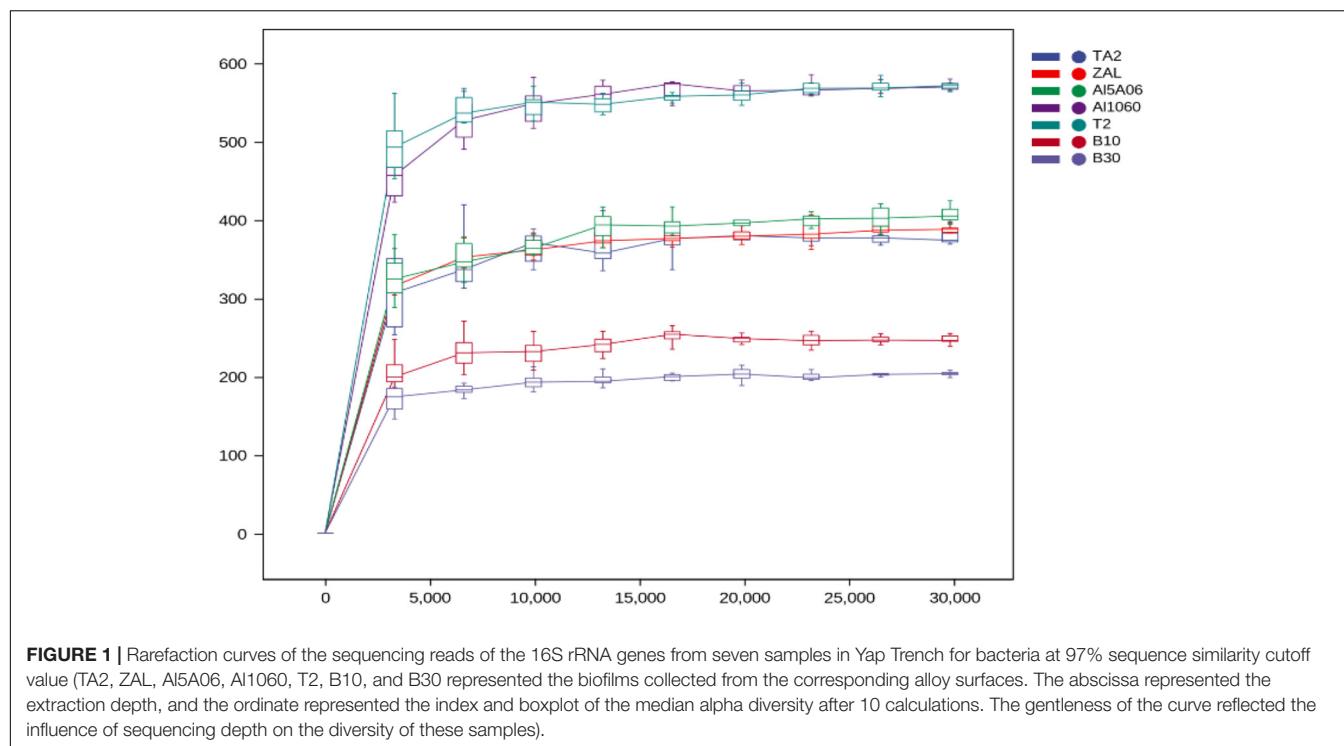
The functional gene profile of the microbial communities on the seven samples was analyzed by PICRUSt2 based on KEGG database. **Figure 7** shows the relative abundance of the top 15 identified genes in each sample. Distinctive functional genes were found in the biofilms on different metal alloys. The most abundant functional gene sets were RNA polymerase sigma-70 factor-encoding gene (*rpoE*) on TA2 (0.44%), Al5A06 (0.31%), Al1060 (0.33%), ZAL (0.31%), T2 (0.34%), B10 (0.18%), and B30 (0.35%). Besides this, glutathione S-transferase-encoding gene (*GST*), which played an important role in biodefense system, was shown to be relatively high in these samples. Another 3-oxoacyl-[acyl-carrier protein] reductase-encoding gene (*fabG*) showed a high abundance, which might be related to fatty acid synthesis and environmental tolerance. What is more, methyl-accepting chemotaxis protein-encoding gene (*mcp*) was found to be dramatically abundant on non-copper aluminum alloys Al1060 and Al5A06. The ABC-2-type transport system permease protein-encoding gene (*ABC-2.P*), LacI family transcriptional regulator-encoding gene (*lacI* and *galR*), and ATP-binding cassette-encoding gene (*ABCB-BAC*) were found to be relatively rich on copper-containing alloys ZAL, T2, B30 and B10.

Copper Resistance Genes

Since copper is considered toxic to microorganisms, various genes related to copper resistance, such as copper tolerance two-component regulatory system *cusSR*, Cu⁺ transporting ATPase-encoding genes *copAB* (Silver and Phung, 2005), and copper resistance protein-encoding genes *pcoBCD*, were identified in these samples and shown in **Table 4**. According to the copper contents in the alloys, these samples could be divided into two groups, i.e., copper-free alloys (TA2, Al1060, and Al5A06) and copper-present alloys (ZAL, T2, B10, and B30). It was found that the abundance of *cusRS* genes of copper-free alloys was significantly lower than that of copper-present alloys ($P < 0.01$, Student's *t*-test), while the abundance of *copAB* genes of copper-free alloys was higher than that of copper-present alloys, which showed a significant difference ($P < 0.01$, Student's *t*-test). Besides these, most of the *pcoBCD* genes were shown to be more abundant in copper-present alloys than in copper-free alloys with $P < 0.01$ (for *pcoBC* genes).

DISCUSSION

Yap Trench has attracted much attention due to its specific physical and geochemical characteristics as well as its hadal



biosphere. Several studies have reported the microbial diversity and metabolic potentials of seawater and surface sediment (Zhang et al., 2018; Fu et al., 2020), but seldom focused on the microbial composition on the serving metals in Yap Trench. This research provided a brief glimpse of the biofilm on several metal alloys at 5,772 m undersea in Yap Trench. The biofilms formed on these metals with distinct composition not only reflected the deep-sea environment to some degree but also provoked new thoughts of the interaction of microorganisms with metals in deep-sea conditions.

Microbial Richness and Diversity Analysis

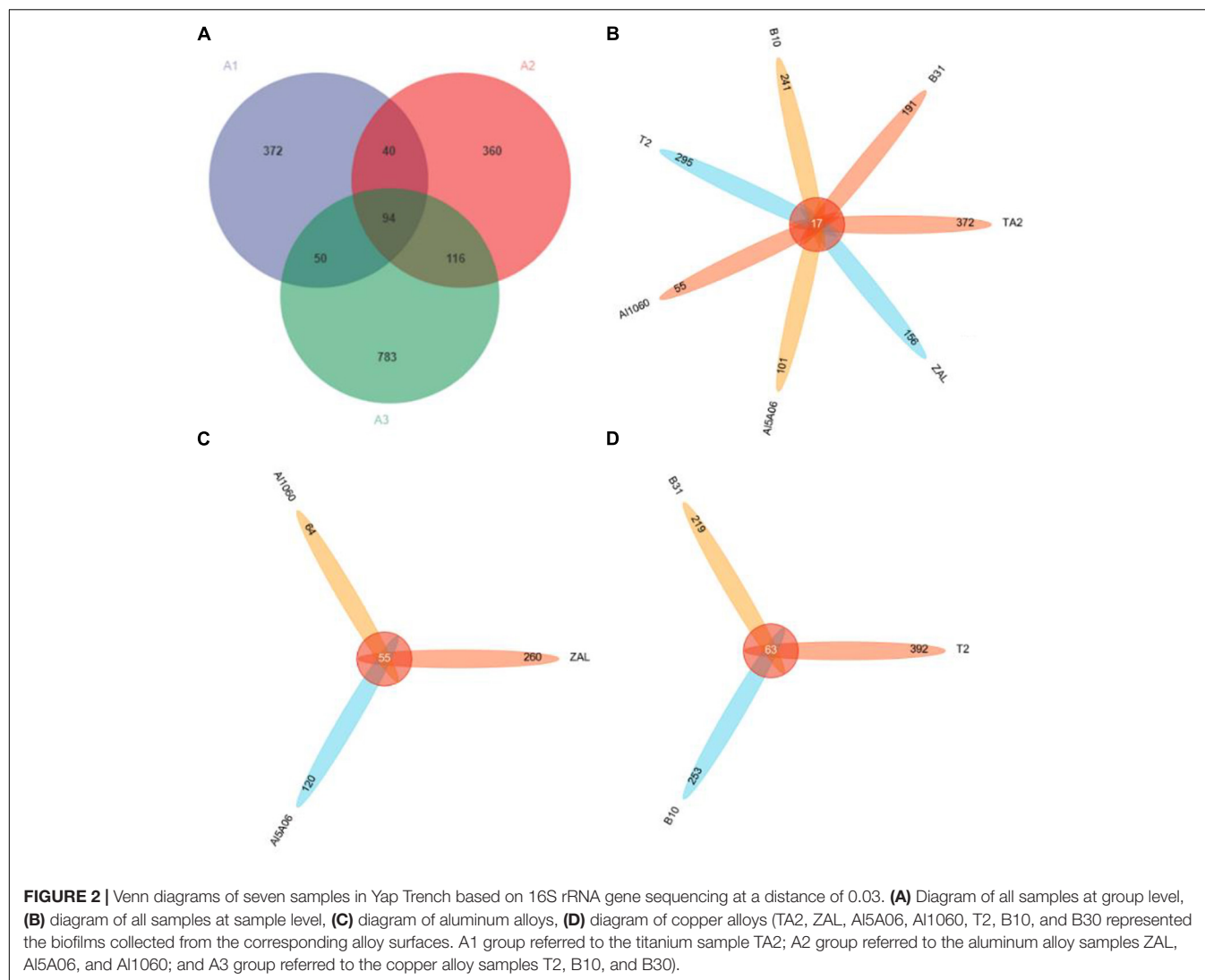
Metal alloy surfaces are ideal sites for biofilm formation and allowing biofilm-associated microorganisms to improve their growth (Beveridge et al., 1997; Loto, 2017). However, microbial richness and diversity showed obvious differences between the various alloy surfaces studied in this research. The highest richness was found on titanium alloy TA2 due to its best biocompatibility (Long and Rack, 1998; Niinomi, 2008; Geetha et al., 2009), followed by copper-present alloys (ZAL, T2, B10, and B30), and the least was aluminum alloys Al5A06 and Al1060. However, these results differed from previous shallow-sea results. Samples immersed at a depth of 1–1.5 m below sea level for 30 months in the coastal zone of Hongtang Bay showed that higher richness was found on aluminum alloy, while lower richness was found on copper alloy (Zhang et al., 2019a), which could be attributed to the oligotrophic environment, leading to the different planktonic microorganisms in Yap Trench. Gammaproteobacteria constituted up to 92.2%

of the total microbial community in the hadal seawater of Yap Trench (Zhang et al., 2018), while Alphaproteobacteria (35.3%) made up the major bacterial groups in the shallow surface seawater (1–1.5 m below sea level) of Hongtang Bay (Zhang et al., 2019a). As a result, on the surface of copper alloys in the hadal environment, biofilms tended to be formed, which provided suitable living environments for microbial organisms (Zhang et al., 2019b). Besides this, an oxidation passivation film composed of Al_2O_3 usually formed on aluminum alloys (Wolowik et al., 1998). The super-hydrophobic property and oligotrophy might also influence the bacterial attachment (Yu et al., 2014; Xiao et al., 2015). Furthermore, the detection of various metal resistance genes, including Cu resistance, in the Yap Trench metagenomes was reported (Zhang et al., 2018), illustrating that hadal microorganisms would be more adapted to Cu-rich environments than to shallow seawater. The environmental copper-resistant microbiological composition as well as the surface conditions of the alloys both contributed to the high bacterial richness on copper-present alloys.

Among these seven alloys, even on the same type of alloys, such as copper alloys T2, B10, and B30, the composition of the bacterial communities showed great distinctions. The proportions of the unique OTUs of each copper alloy sample ranged from 78 to 86%, revealing that even alloying elements with low concentrations would greatly influence the bacterial communities.

Microbial Community Analysis

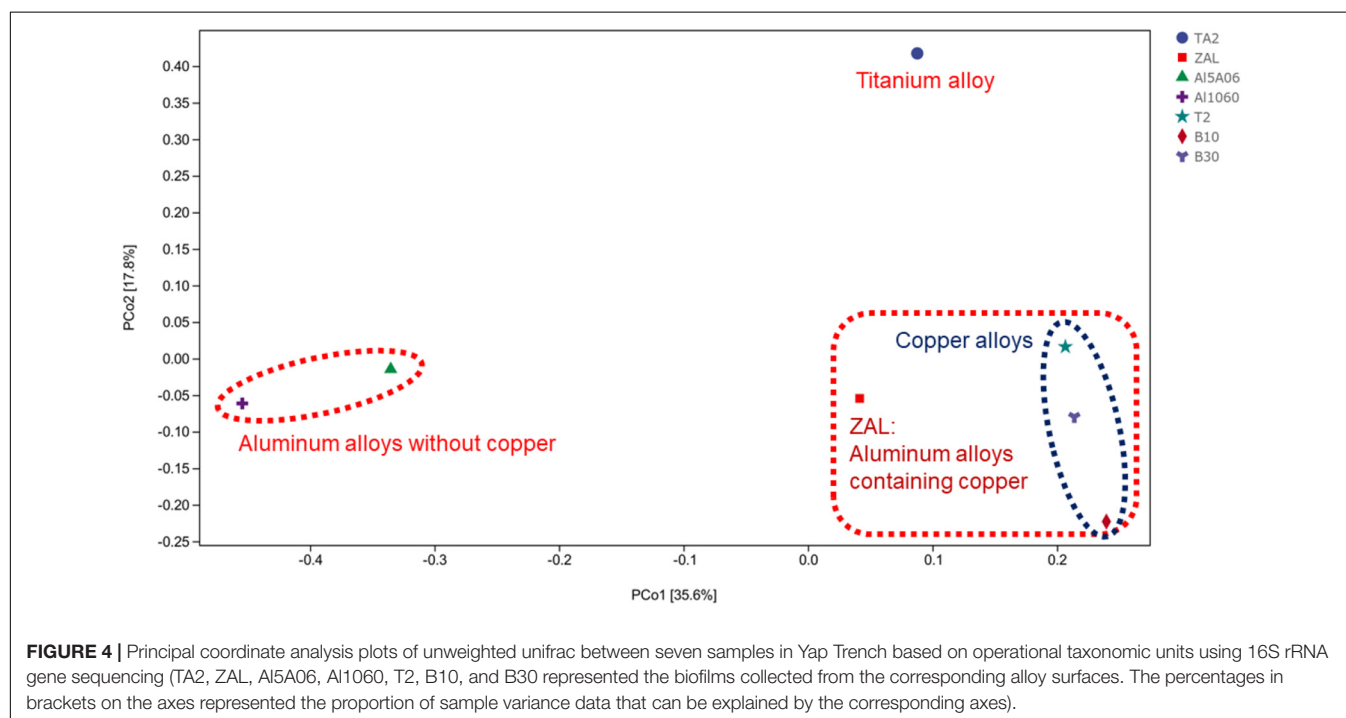
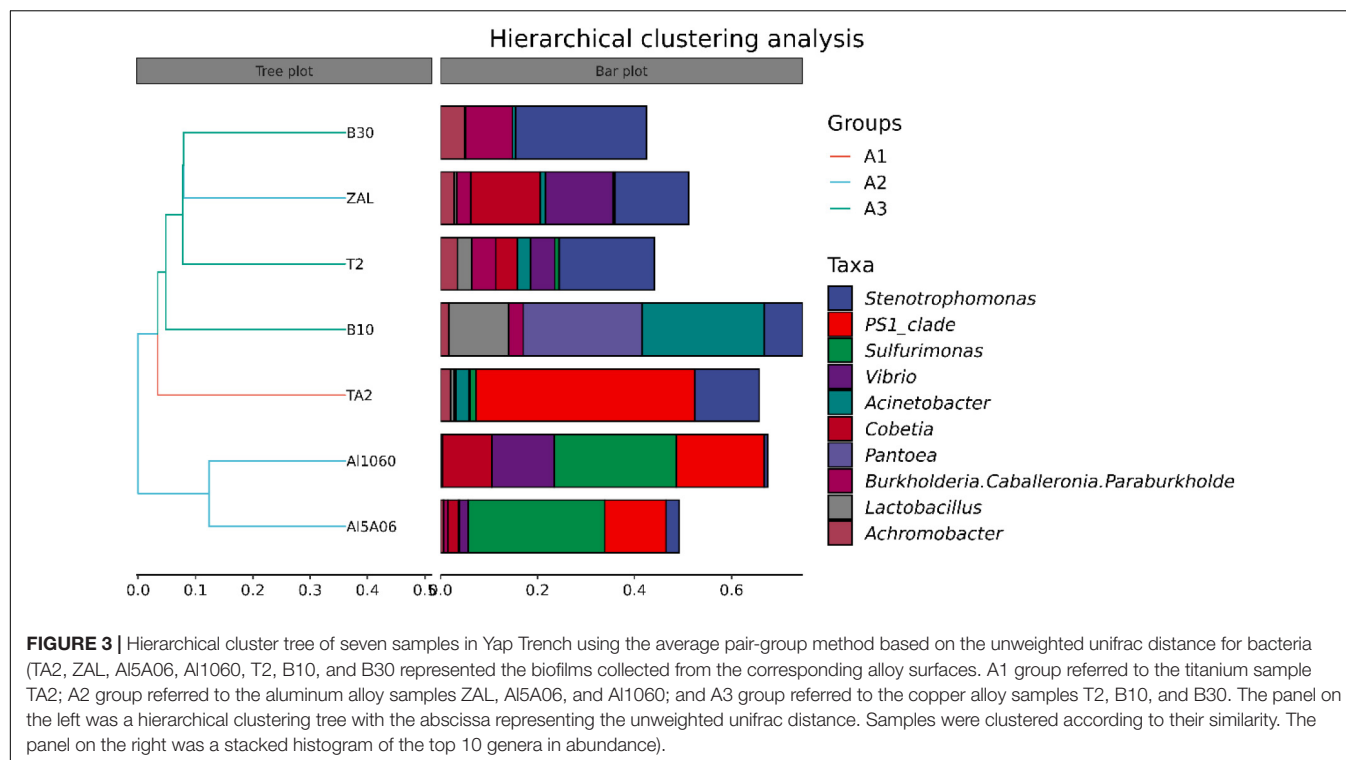
Metallic surfaces might promote bacterial attachment and biofilm formation by enriching nutrients or acting as electron donors for microorganisms (Beveridge et al., 1997; Yu et al., 2013;



Guan et al., 2016, 2021). Diverse and distinct bacterial communities developed on the surfaces of different alloys, which highly depended on the composition of the substrate.

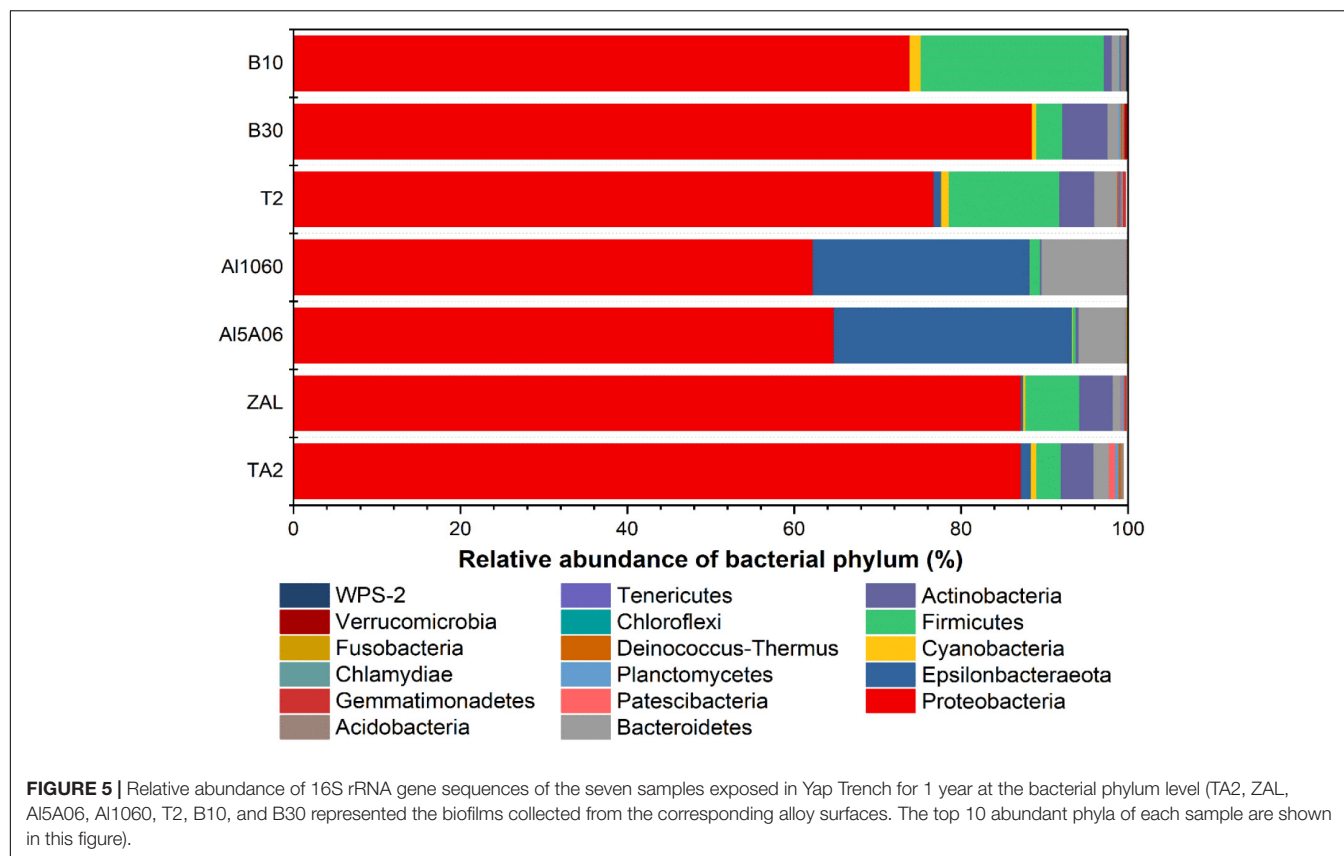
On titanium alloy TA2, *PS1 Clade* played the leading role in the biofilm. *PS1 Clade* of Alphaproteobacteria was firstly isolated from a coastal station in the East Sea, Western Pacific Ocean, and reported by SJ Yang (Yang et al., 2012). *PS1 Clade* was a member of a putatively novel order closely related to Rhizobiales. The *PS1* lineage stem would adapt to various marine habitats, including the oligotrophic Sargasso Sea as well as tropical and temperate environments (Jimenez-Infante et al., 2014). The core genome of the *PS1 Clade* suggested an aerobic, heterotrophic lifestyle with genes encoding for gluconeogenesis, citric acid cycle, and the Entner–Doudoroff pathway, implying that the *PS1 Clade* might not be primary cellulose degraders but opportunists utilizing cellobiose and small oligosaccharides (Jimenez-Infante et al., 2014; Daniel and Ana, 2020). What is more important is that the genome of *PS1 Clade* strains represented numerous high-affinity transporter-encoding genes,

which were genomic hallmarks for cells proliferating in low-nutrient environments (Lauro et al., 2009). *Stenotrophomonas* took the second place in TA2 biofilm. Although it was well-known as a nosocomial and human infection pathogen (Coenye et al., 2004), *Stenotrophomonas* strains dwelling in marine environments remained unclear. Many *Stenotrophomonas* strains showed high resistance to high-level intrinsic resistance to heavy metals. It was also proved that they could degrade a wide range of organic compounds, including pollutants, which would potentially be used in bioremediation (Ryan et al., 2009). The metabolites of *Stenotrophomonas* strains always showed antifungal or antibacterial activities (Romanenko et al., 2008), which led to a biofilm with a relatively simple constitution on the TA2 surface. Another dominating genus, *Acinetobacter* on TA2, is one of the commonly found Gram-negative bacteria in marine environments. Some *Acinetobacter* species isolated from deep-sea sediments were found to be cold-adapted (Xue et al., 2019). *Acinetobacter* played an important role in hydrocarbon degradation and has a key role in bioremediation processes



(MacCormack and Fraile, 1997). Various *Acinetobacter* strains were reported to be oil-, sulfonamide-, and phenol-degrading (Kobayashi et al., 2012; Zhang et al., 2012; Luo et al., 2013), indicating that they could make use of multiple organic carbon sources. Thus, *Stenotrophomonas* and *Acinetobacter* might act as the primary degraders for *PS1 Clade*, forming mutualistic

symbiosis in the biofilm. It is worth mentioning that although the hadal environment was considered oligotrophic, organic matters were discovered in this area, and heterotrophic processes were found as the most common microbial metabolisms in the Yap Trench seawater (Zhang et al., 2018). Researchers took the view that the typical “V-shape” topography of the trenches would



accumulate organic matters by a funneling effect. These organic matters could come from sinking particulates from the upper ocean, terrestrial inputs, chemosynthesis from the dark ocean, or even cell lysates at the trench axis (Jover et al., 2014). Then, under gravity, these organic materials would slowly migrate to the deepest trench axis (Ichino et al., 2015). The funneling effect due to the “V-shape” of the trench flanks also played an important role in the formation of increasing surface sedimentary organic carbon content which might come from the upper seawater layer (Li D. et al., 2020). Besides these, abundant genes involved in the degradation of various types of carbohydrates, hydrocarbons, and aromatics were reported by previous studies (Zhang et al., 2018), showing their potentials to be used organic carbon sources in the Yap Trench environment indicating organic matter-enriched environments.

On non-copper aluminum alloys Al1060 and Al5A06, *Sulfurimonas* and *PS1 Clade* dominated the bacterial groups in the biofilms. The genus *Sulfurimonas* belonged to the class Campylobacteria within the phylum of Epsilonbacteraeota (Inagaki et al., 2003). *Sulfurimonas* strains were discovered in various habitats, including marine sediments, deep-sea hydrothermal vents, and pelagic water column redoxclines (Han and Perner, 2015). Although the lineage *Sulfurimonas* was well-known as small sulfur-oxidizing bacteria utilizing reduced sulfur compounds such as sulfide, thiosulfate, and elemental sulfur as an electron donor for growth, organic compounds including formate, fumarate, amino acid, and alcohol mix could work as

a preferred electron donor and contribute to bacterial growth (Labrenz et al., 2013). The versatile metabolic strategies of *Sulfurimonas* species helped them adapt to a broad type of environments (Han and Perner, 2015), including the deep-sea environment in Yap Trench. The versatile metabolic strategies might also cooperate with *PS1 Clade* to form mutualistic symbiosis in the biofilm.

On copper alloys T2, B10, 30, and copper-present aluminum alloy ZAL, *Stenotrophomonas* was found as a major constitution in these biofilms. As mentioned above, *Stenotrophomonas*, which phylogenetically belonged to Gammaproteobacteria, showed high resistance to heavy metals, including Cu (Ye et al., 2013; Chen et al., 2016; Hou et al., 2020). As a result, the *Stenotrophomonas* strains were successfully isolated from various copper-rich environments, such as copper-polluted agricultural soils and well-adapted Cu(II)-reduced biocathodes of microbial fuel cells (Altimira et al., 2012; Tao et al., 2017). *Stenotrophomonas* strains could convert Cu(II) into Cu(0) on the cell surface in the absence of cathodic electrons (Shen et al., 2017). On copper surface, *Stenotrophomonas* tended to release more amounts of extracellular polymeric substances (EPS) to form biofilms with a strong Cu(II) complexation effect (Hou et al., 2020). Therefore, *Stenotrophomonas* showed high tolerance or might get use of Cu(II) on copper alloy surfaces by forming a biofilm with high EPS. At the same time, this Cu(II) reduction process would inhibit the corrosion of the copper alloys serving in this environment because the

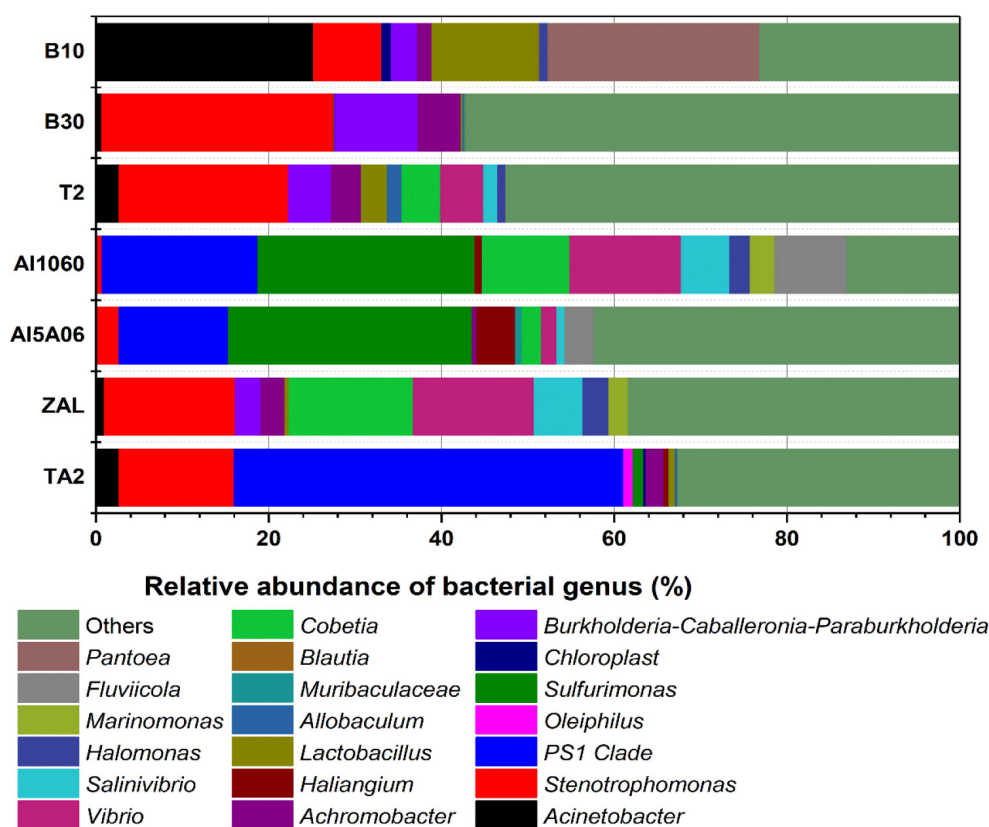


FIGURE 6 | Relative abundance of 16S rRNA gene sequences of the samples exposed in Yap Trench for 1 year at the bacterial genus level (TA2, ZAL, AI5A06, AI1060, T2, B10, and B30 represented the biofilms collected from the corresponding alloy surfaces. The top 10 abundant genera of each sample are shown in this figure).

essence of corrosion was defined as the oxidation process of metals. Due to the existence of a biofilm composed of *Stenotrophomonas* which would reduce Cu(II) (Shen et al., 2017), the oxidation process was restrained, leading to an inhibited corrosion process. *Burkholderia-Caballeronia-Paraburkholderia* showed high proportions on these copper-present alloys, too. They were discovered on copper-rich microbial fuel cells, revealing a high tolerance to copper (Wu et al., 2019; Ai et al., 2020). Another dominating genus in the biofilm on B10 was *Acinetobacter*. Besides the cold-adapting and hydrocarbon-degrading characteristics mentioned above, biosorption of Cu(II) was found on *Acinetobacter* in 2017 (Zhang et al., 2017). In the biofilm containing *Acinetobacter*, Cu(II) would promote protein secretion and bound with EPS, thus leading to more compact granules with better ability to settle (Jiang et al., 2020). Above all, the biofilm formed on copper-present alloys all showed high tolerance to copper. The biofilm would use, reduce, or biosort Cu(II) to form a stable and functional mutualistic symbiosis.

Besides that, previous reports showed that the abundant genes of bacterial communities in Yap Trench seawater were involved in stress response and metal resistance (Zhang et al., 2018). The attached bacteria on these alloys were considered to be derived from the planktonic communities and enriched on metal

surfaces. Thus, high bacterial diversity and stable functional mutualistic symbiosis could form on metallic surfaces, even on toxic copper alloys.

Key Functional Gene Analysis

The functional gene *rpoE* was found to be relatively abundant in all seven samples. *RpoE* gene was known as an important stress response gene. *RpoE* could encode key RNA polymerase component, which contributed to the protein expression for periplasmic and outer membrane component integrity (Helmman, 2002; Woods and McBride, 2017). The relatively high abundance of *rpoE* revealed a high resistance of these biofilms to the extreme deep-sea environment with low temperature, high pressure, and oligotrophic features. *FabG* gene encoding 3-oxoacyl-[ACP] reductase and *GST* gene encoding glutathione S-transferase were also commonly found in these samples. *FabG* gene was a key enzyme in the type II fatty acid synthase system in bacteria and catalyzes beta-ketoacyl-ACP reduction, while A played key roles on fatty acid biosynthesis (Li et al., 2006; Huang et al., 2008). The bacterial *GSTs* were reported to be active in catalyzing specific reactions in the degradation pathways of recalcitrant chemicals for growth (Vuilleumier, 1997). Thus, the obvious presence of *fabG* and *GST* genes indicated harsh living conditions in Yap

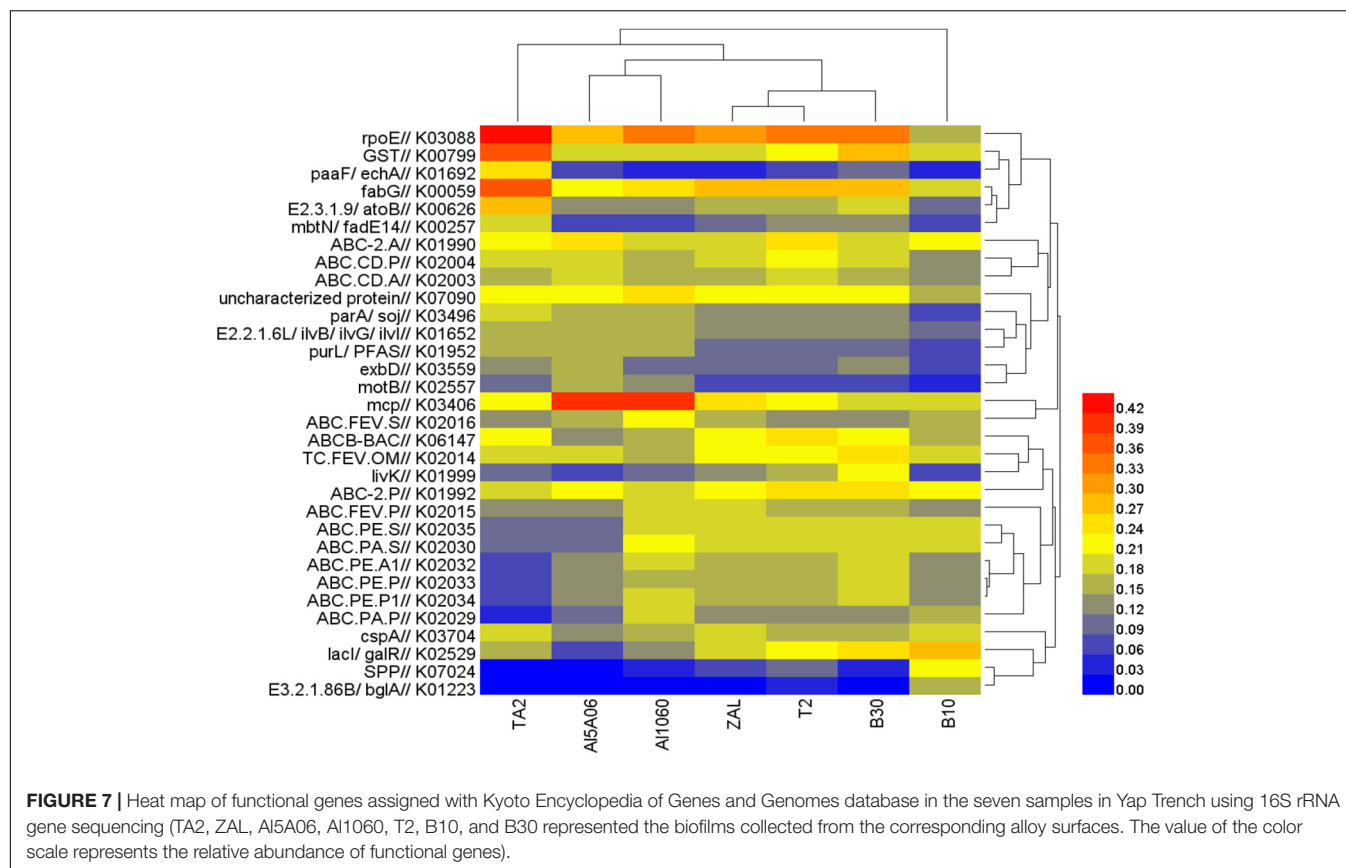


TABLE 4 | Relative abundance of key functional genes related to copper resistance in the seven samples in Yap Trench using 16S rRNA gene sequencing.

Gene	Description	TA2	ZAL	AI5A06	AI1060	T2	B10	B30
<i>cusR</i>	Copper resistance phosphate regulon response regulator	0.009%	0.021%	0.004%	0.005%	0.022%	0.027%	0.029%
<i>cusS</i>	Heavy metal sensor histidine kinase	0.010%	0.021%	0.004%	0.006%	0.023%	0.027%	0.030%
<i>copA</i>	Cu ⁺ -transporting ATPase	0.066%	0.051%	0.091%	0.096%	0.053%	0.051%	0.044%
<i>copB</i>	Cu ⁺ -transporting ATPase	0.054%	0.033%	0.047%	0.044%	0.031%	0.019%	0.029%
<i>pcoB</i>	Copper resistance protein B	0.014%	0.022%	0.004%	0.005%	0.026%	0.021%	0.030%
<i>pcoC</i>	Copper resistance protein C	0.012%	0.023%	0.004%	0.006%	0.024%	0.026%	0.029%
<i>pcoD</i>	Copper resistance protein D	0.013%	0.023%	0.004%	0.006%	0.022%	0.025%	0.026%

TA2, ZAL, AI5A06, AI1060, T2, B10, and B30 represented the biofilms collected from the corresponding alloys surfaces.

Trench (Vuilleumier and Pagni, 2002). Besides that, several genes connected to transport systems, such as *ABC.CD.P*, *ABC.CD.A*, *ABC-2.A*, *ABC-2.P*, and so on, were detected. ATP-binding cassette (ABC) transporter-encoding genes might mainly come from *PS 1 Clade*. ABC transporters were known to transport a wide variety of substrates, such as amino acids, oligopeptides, and sugars (Davidson et al., 2008). These ABC transporters contributed significantly to the uptake of extensive substrates for growth in relatively oligotrophic and pelagic environments.

It is worth mentioning that the existence of Cu element seemed to make a great influence on the functional genes of the biofilms than the other alloying elements. The abundance of several genes related to copper resistance, such as the copper tolerance two-component regulatory system *cusSR*,

Cu⁺-transporting ATPase-encoding gene *copAB* (Silver and Phung, 2005), and copper resistance protein-encoding gene *pcoBCD*, was found to be distinctly differentiated. Based on copper contents, the alloys could be divided into two groups: copper-free alloys (TA2, AI1060, and AI5A06) and copper-present alloys (ZAL, T2, B10, and B30). The copper-free alloys showed low *cusSR* and *pcoBCD* but high *copAB* abundance, while the copper-present alloys revealed a contrast. The *cusS*–*cusR* two-component systems were significant for bacteria in sensing, responding, and adapting to the changing environments, such as the elevation of Cu(I) ions in the periplasm (Affandi and McEvoy, 2019). The plasmid-encoded gene *pcoBCD* would detoxify copper in the periplasm and further strengthen the copper resistance ability (Rensing and Grass, 2003), while *copA* and *copB* genes could

encode Cu⁺-transporting ATPase, which would act as ATPase membrane pump to transport copper ions (Mana-Capelli et al., 2003; Silver and Phung, 2005). The high abundance of high *copAB* on copper-free alloys was an interesting phenomenon because of the opposite results compared to that obtained in shallow surface seawater (Zhang et al., 2019a,b). This might be attributed to the fact that, in the hadal environment, the environment might be quite oligotrophic with low copper contents (less than 0.000002% in surface sediments; Huang et al., 2020). No results were found in hadal seawater in Yap Trench; it might be even lower. However, copper was considered as one of the essential elements for living cells (Favre et al., 2019), so they needed to transport copper into the cell to enrich copper for growth and metabolism, which could lead to high *copAB*. That might also be the reason why the abundance of bacteria that survived on copper-free alloys was lower than that on copper alloys. So, based on these above-mentioned results, a hypothesis could be proposed. On copper-free alloys, the biofilms showed low copper-resistant gene abundance but high copper-sensitive gene abundance, which might be used for transporting copper ions according to the growth and metabolism requirements. On the contrary, the biofilms on copper-present alloys showed a relatively high copper resistance. They might not consume ATP to export copper ions. They might either make use of copper ions in the cell or export copper ions in other ways. The diverse response to copper led to totally different bacterial communities and functions of the biofilm on these alloys.

CONCLUSION

This study caught a brief glimpse of biofilms formed on metal alloys in Yap Trench. Although it was known that planktonic bacterial communities showed a great difference with the biofilm communities, this research found out that the bacterial communities on the biofilms at various substrates revealed obvious differences. Among the alloys studied in this research, copper element showed strong influences on microbial communities and key functional genes even at a relatively low content in the alloy, such as ZAL. Titanium alloy provided a suitable substrate for a mutualistic symbiotic biofilm. Aluminum alloys without copper components showed the least bacterial

richness and formed cold-adapted and oligotrophic-adapted biofilms. Copper-present alloys showed a relatively high bacterial richness with copper-resistant or even copper-utilizing biofilms on the surfaces. Besides that, the copper-related biofilm would participate in Cu(II) reduction, which could effectively inhibit copper corrosion. Furthermore, the bacterial communities of biofilms on these alloys were found to be highly different from those in shallow sea, and many bacterial genera remained unclear based on our existing database. Thus, future research on extreme environments, such as deep-sea environments, are critically needed and of great significance.

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/>, PRJNA438021 SAMN23711893-23711899.

AUTHOR CONTRIBUTIONS

XZ and WC were responsible for designing and conducting the experiments, analyzing the data, and drafting the manuscript. XZ and YZ revised the manuscript and provided much needed insight into the result interpretation. WC and PJ took part in plate placement and sample collection. JD have overseen all aspects of this project in terms of scientific significance. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Basic Scientific Fund for National Public Research Institutes of China (2020S02 and 2019Y03), National Natural Science Foundation of China (51702328 and 41706080), National Basic Research Program of China (973 Program) (No. 2015CB755904), Natural Science Foundation of Shandong Province (No. ZR2021QD099), Key Research and Development Program of Shandong Province (Major Scientific and Technological Innovation Project, 2019JZZY020711), and the Young Elite Scientists Sponsorship Program by CAST.

REFERENCES

- Affandi, T., and McEvoy, M. M. (2019). Mechanism of metal ion-induced activation of a two-component sensor kinase. *Biochem. J.* 476, 115–135. doi: 10.1042/bcj20180577
- Ai, C., Yan, Z., Hou, S., Zheng, X., Zeng, Z., Amanze, C., et al. (2020). Effective treatment of acid mine drainage with microbial fuel cells: an emphasis on typical energy substrates. *Minerals* 10:443. doi: 10.3390/min10050443
- Altimira, F., Yanez, C., Bravo, G., Gonzalez, M., Rojas, L. A., and Seeger, M. (2012). Characterization of copper-resistant bacteria and bacterial communities from copper-polluted agricultural soils of central Chile. *BMC Microbiol.* 12:193. doi: 10.1186/1471-2180-12-193
- Beveridge, T. J., Makin, S. A., Kadurugamuwa, J. L., and Li, Z. S. (1997). Interactions between biofilms and the environment. *FEMS Microbiol. Rev.* 20, 291–303. doi: 10.1016/s0168-6445(97)00012-0
- Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336. doi: 10.1038/nmeth.f.303
- Chen, S., Yin, H., Tang, S., Peng, H., Liu, Z., and Dang, Z. (2016). Metabolic biotransformation of copper-benzo a pyrene combined pollutant on the cellular interface of *Stenotrophomonas maltophilia*. *Bioresour. Technol.* 204, 26–31. doi: 10.1016/j.biortech.2015.12.068
- Coenye, T., Vanlaere, E., Falsen, E., and Vandamme, P. (2004). *Stenotrophomonas africana* Drancourt et al. 1997 is a later synonym of *Stenotrophomonas maltophilia* (Hugh 1981) Palleroni and Bradbury 1993. *Int. J. Syst. Evol. Microbiol.* 54, 1235–1237. doi: 10.1099/ijs.0.63093-0
- Dang, H., and Lovell, C. R. (2016). Microbial surface colonization and biofilm development in marine environments. *Microbiol. Mol. Biol. Rev.* 80, 91–138. doi: 10.1128/mmbr.00037-15

- Daniel, F. R. C., and Ana, R. M. P. (2020). Marine lake populations of jellyfish, mussels and sponges host compositionally distinct prokaryotic communities. *Hydrobiologia* 847, 3409–3425. doi: 10.1007/s10750-020-04346-3
- Davidson, A. L., Dassa, E., Orelle, C., and Chen, J. (2008). Structure, function, and evolution of bacterial ATP-binding cassette systems. *Microbiol. Mol. Biol. Rev.* 72, 317–364. doi: 10.1128/mmbr.00031-07
- DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., et al. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* 72, 5069–5072. doi: 10.1128/aem.03006-05
- Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods* 10, 996–998. doi: 10.1038/nmeth.2604
- Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., and Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27, 2194–2200. doi: 10.1093/bioinformatics/btr381
- Enning, D., and Garrelfs, J. (2014). Corrosion of iron by sulfate-reducing bacteria: new views of an old problem. *Appl. Environ. Microbiol.* 80, 1226–1236. doi: 10.1128/AEM.02848-13
- Favre, L., Ortalo-Magne, A., Kerloch, L., Pichereaux, C., Misson, B., Briand, J. F., et al. (2019). Metabolomic and proteomic changes induced by growth inhibitory concentrations of copper in the biofilm-forming marine bacterium *Pseudoalteromonas lipolytica*. *Metallomics* 11, 1887–1899. doi: 10.1039/c9mt00184k
- Fletcher, M., and Loeb, G. I. (1979). Influence of substratum characteristics on the attachment of a marine *Pseudomonad* to solid-surfaces. *Appl. Environ. Microbiol.* 37, 67–72. doi: 10.1128/aem.37.1.67-72.1979
- Fu, L. L., Li, D., Mi, T. Z., Zhao, J., Liu, C. G., Sun, C. J., et al. (2020). Characteristics of the archaeal and bacterial communities in core sediments from Southern Yap Trench via *in situ* sampling by the manned submersible Jiaolong. *Sci. Total Environ.* 703:134884. doi: 10.1016/j.scitotenv.2019.134884
- Geetha, M., Singh, A. K., Asokamani, R., and Gogia, A. K. (2009). Ti based biomaterials, the ultimate choice for orthopaedic implants - A review. *Prog. Mater. Sci.* 54, 397–425. doi: 10.1016/j.pmatsci.2008.06.004
- Guan, F., Duan, J., Zhai, X., Wang, N., Zhang, J., Lu, D., et al. (2020). Interaction between sulfate-reducing bacteria and aluminum alloys—Corrosion mechanisms of 5052 and Al-Zn-In-Cd aluminum alloys. *J. Mater. Sci. Technol.* 36, 55–64. doi: 10.1016/j.jmst.2019.07.009
- Guan, F., Liu, Z., Dong, X., Zhai, X., Zhang, B., Duan, J., et al. (2021). Synergistic effect of carbon starvation and exogenous redox mediators on corrosion of X70 pipeline steel induced by *Desulfovibrio singaporensis*. *Sci. Total Environ.* 788:147573. doi: 10.1016/j.scitotenv.2021.147573
- Guan, F., Zhai, X., Duan, J., Zhang, M., and Hou, B. (2016). Influence of sulfate-reducing bacteria on the corrosion behavior of high strength steel EQ70 under cathodic polarization. *PLoS One* 11:e0162315. doi: 10.1371/journal.pone.0162315
- Han, Y., and Perner, M. (2015). The globally widespread genus *Sulfurimonas*: versatile energy metabolisms and adaptations to redox clines. *Front. Microbiol.* 6:989. doi: 10.3389/fmicb.2015.00989
- Helmann, J. D. (2002). The extracytoplasmic function (ECF) sigma factors. *Adv. Microb. Physiol.* 46, 47–110. doi: 10.1016/s0065-2911(02)46002-x
- Hou, J., Huang, L., Zhou, P., Qian, Y., and Li, N. (2020). Understanding the interdependence of strain of electrothroph, cathode potential and initial Cu(II) concentration for simultaneous Cu(II) removal and acetate production in microbial electrosynthesis systems. *Chemosphere* 243:125317. doi: 10.1016/j.chemosphere.2019.125317
- Huang, H., Wu, D., Tian, W. X., Ma, X. F., and Wu, X. D. (2008). Antimicrobial effect by extracts of rhizome of *Alpinia officinarum* Hance may relate to its inhibition of beta-ketoacyl-ACP reductase. *J. Enzyme Inhib. Med. Chem.* 23, 362–368. doi: 10.1080/14756360701622099
- Huang, Y. H., Sun, C. J., Yang, G. P., Yue, X. N., Jiang, F. H., Cao, W., et al. (2020). Geochemical characteristics of hadal sediment in the northern Yap Trench. *J. Oceanol. Limnol.* 38, 650–664. doi: 10.1007/s00343-019-9010-3
- Ichino, M. C., Clark, M. R., Drazen, J. C., Jamieson, A., Jones, D. O. B., Martin, A. P., et al. (2015). The distribution of benthic biomass in hadal trenches: a modelling approach to investigate the effect of vertical and lateral organic matter transport to the seafloor. *Deep Sea Res. I Oceanogr. Res. Pap.* 100, 21–33. doi: 10.1016/j.dsr.2015.01.010
- Inagaki, F., Takai, K., Hideki, K. I., Neelson, K. H., and Horikishi, K. (2003). *Sulfurimonas autotrophica* gen. nov., sp nov., a novel sulfur-oxidizing epsilon-proteobacterium isolated from hydrothermal sediments in the Mid-Okinawa Trough. *Int. J. Syst. Evol. Microbiol.* 53, 1801–1805. doi: 10.1099/ijs.0.02682-0
- Jiang, Y., Liu, Y., Zhang, H., Yang, K., Li, J., and Shao, S. (2020). Aerobic granular sludge shows enhanced resistances to the long-term toxicity of Cu(II). *Chemosphere* 253:126664. doi: 10.1016/j.chemosphere.2020.126664
- Jimenez-Infante, F., Ngugi, D. K., Alam, I., Rashid, M., Baalawi, W., Kamau, A. A., et al. (2014). Genomic differentiation among two strains of the *PS1 clade* isolated from geographically separated marine habitats. *FEMS Microbiol. Ecol.* 89, 181–197. doi: 10.1111/1574-6941.12348
- Jover, L. F., Effler, T. C., Buchan, A., Wilhelm, S. W., and Weitz, J. S. (2014). The elemental composition of virus particles: implications for marine biogeochemical cycles. *Nat. Rev. Microbiol.* 12, 519–528. doi: 10.1038/nrmicro3289
- Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D361. doi: 10.1093/nar/gkw1092
- Kobayashi, F., Maki, T., and Nakamura, Y. (2012). Biodegradation of phenol in seawater using bacteria isolated from the intestinal contents of marine creatures. *Int. Biodeter. Biodegr.* 69, 113–118. doi: 10.1016/j.ibiod.2011.06.008
- Labrenz, M., Grote, J., Mammitzsch, K., Boschker, H. T. S., Laue, M., Jost, G., et al. (2013). *Sulfurimonas gotlandica* sp nov., a chemoautotrophic and psychrotolerant epsilonproteobacterium isolated from a pelagic redoxcline, and an emended description of the genus *Sulfurimonas*. *Int. J. Syst. Evol. Microbiol.* 63, 4141–4148. doi: 10.1099/ijs.0.048827-0
- Langille, M. G. I., 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
- Lauro, F. M., McDougald, D., Thomas, T., Williams, T. J., Egan, S., Rice, S., et al. (2009). The genomic basis of trophic strategy in marine bacteria. *Proc. Natl. Acad. Sci. U. S. A.* 106, 15527–15533. doi: 10.1073/pnas.0903507106
- Li, B. H., Zhang, R., Du, Y. T., Sun, Y. H., and Tian, W. X. (2006). Inactivation mechanism of the beta-ketoacyl-acyl carrier protein reductase of bacterial type-II fatty acid synthase by epigallocatechin gallate. *Biochem. Cell Biol.* 84, 755–762. doi: 10.1139/o06-047
- Li, D., Zhao, J., Yao, P., Liu, C. G., Sun, C. J., Chen, J. F., et al. (2020). Spatial heterogeneity of organic carbon cycling in sediments of the northern Yap Trench: implications for organic carbon burial. *Mar. Chem.* 223:103813. doi: 10.1016/j.marchem.2020.103813
- Li, L., Bai, S., Li, J., Wang, S., Tang, L., Dasgupta, S., et al. (2020). Volcanic ash inputs enhance the deep-sea seabed metal-biogeochemical cycle: a case study in the Yap Trench, western Pacific Ocean. *Mar. Geol.* 430:106340. doi: 10.1016/j.margeo.2020.106340
- Li, X., Duan, J., Xiao, H., Li, Y., Liu, H., Guan, F., et al. (2017). Analysis of Bacterial Community Composition of Corroded Steel Immersed in Sanya and Xiamen Seawaters in China via Method of Illumina MiSeq Sequencing. *Front. Microbiol.* 8:01737. doi: 10.3389/fmicb.2017.01737
- Long, M., and Rack, H. J. (1998). Titanium alloys in total joint replacement—a materials science perspective. *Biomaterials* 19, 1621–1639. doi: 10.1016/s0142-9612(97)00146-4
- Loto, C. A. (2017). Microbiological corrosion: mechanism, control and impact—a review. *Int. J. Adv. Manuf. Tech.* 92, 4241–4252. doi: 10.1007/s00170-017-0494-8
- Luo, Q., Zhang, J. G., Shen, X. R., Fan, Z. Q., He, Y., and Hou, D. Y. (2013). Isolation and characterization of marine diesel oil-degrading *Acinetobacter* sp strain Y2. *Ann. Microbiol.* 63, 633–640. doi: 10.1007/s13213-012-0513-9
- Ma, Y., Zhang, Y., Zhang, R., Guan, F., Hou, B., and Duan, J. (2020). Microbiologically influenced corrosion of marine steels within the interaction between steel and biofilms: a brief view. *Appl. Microbiol. Biotechnol.* 104, 515–525. doi: 10.1007/s00253-019-10184-8
- MacCormack, W. P., and Fraile, E. R. (1997). Characterization of a hydrocarbon degrading psychrotrophic Antarctic bacterium. *Antarct. Sci.* 9, 150–155. doi: 10.1017/s0954102097000199
- Magoc, 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

- Mana-Capelli, S., Mandal, A. K., and Arguello, J. M. (2003). *Archaeoglobus fulgidus* CopB is a thermophilic Cu²⁺-ATPase-Functional role of its histidine-rich N-terminal metal binding domain. *J. Biol. Chem.* 278, 40534–40541. doi: 10.1074/jbc.M306907200
- Mittelman, M. W., Habash, M., Lacroix, J. M., Khoury, A. E., and Krajden, M. (1997). Rapid detection of *Enterobacteriaceae* in urine by fluorescent 16S rRNA *in situ* hybridization on membrane filters. *J. Microbiol. Methods* 30, 153–160. doi: 10.1016/s0167-7012(97)00061-4
- Niinomi, M. (2008). Titanium alloys with high biological and mechanical biocompatibility. *J. Jpn Soc. Powder Powder Metall.* 55, 303–311. doi: 10.2497/jjspm.55.303
- Remazeilles, C., Saheb, M., Neff, D., Guilminot, E., Tran, K., Bourdoiseau, J. A., et al. (2010). Microbiologically influenced corrosion of archaeological artefacts: characterisation of iron(II) sulfides by Raman spectroscopy. *J. Raman Spectrosc.* 41, 1425–1433. doi: 10.1002/jrs.2717
- Rensing, C., and Grass, G. (2003). *Escherichia coli* mechanisms of copper homeostasis in a changing environment. *FEMS Microbiol. Rev.* 27, 197–213. doi: 10.1016/s0168-6445(03)00049-4
- Romanenko, L. A., Uchino, M., Tanaka, N., Frolova, G. M., Slinkina, N. N., and Mikhailov, V. V. (2008). Occurrence and antagonistic potential of *Stenotrophomonas* strains isolated from deep-sea invertebrates. *Arch. Microbiol.* 189, 337–344. doi: 10.1007/s00203-007-0324-8
- Ryan, R. P., Monchy, S., Cardinale, M., Taghavi, S., Crossman, L., Avison, M. B., et al. (2009). The versatility and adaptation of bacteria from the genus *Stenotrophomonas*. *Nat. Rev. Microbiol.* 7, 514–525. doi: 10.1038/nrmicro2163
- Shen, J., Huang, L., Zhou, P., Quan, X., and Li Puma, G. (2017). Correlation between circuit current. Cu(II) reduction and cellular electron transfer in EAB isolated from Cu(II)-reduced biocathodes of microbial fuel cells. *Bioelectrochemistry* 114, 1–7. doi: 10.1016/j.bioelechem.2016.11.002
- Silver, S., and Phung, L. T. (2005). A bacterial view of the periodic table: genes and proteins for toxic inorganic ions. *J. Ind. Microbiol. Biotechnol.* 32, 587–605. doi: 10.1007/s10295-005-0019-6
- Tao, Y., Xue, H., Huang, L., Zhou, P., Yang, W., Quan, X., et al. (2017). Fluorescent probe based subcellular distribution of Cu(II) ions in living electrotrophs isolated from Cu(II)-reduced biocathodes of microbial fuel cells. *Bioresour. Technol.* 225, 316–325. doi: 10.1016/j.biortech.2016.11.084
- Vuilleumier, S. (1997). Bacterial glutathione S-transferases: what are they good for? *J. Bacteriol.* 179, 1431–1441. doi: 10.1128/jb.179.5.1431-1441.1997
- Vuilleumier, S., and Pagni, M. (2002). The elusive roles of bacterial glutathione S-transferases: new lessons from genomes. *Appl. Microbiol. Biotechnol.* 58, 138–146. doi: 10.1007/s00253-001-0836-0
- Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* 73, 5261–5267. doi: 10.1128/aem.00062-07
- Wolowik, A., Janik-Czachor, M., Werner, Z., Wood, G. C., Skeldon, P., Thompson, G. E., et al. (1998). Inter-relationships between alloy composition, passive film composition and pitting behaviour of Al-Mo and Al-Mo-Si metastable alloys. *Corros. Sci.* 40, 731–740. doi: 10.1016/s0010-938x(97)00175-3
- Woods, E. C., and McBride, S. M. (2017). Regulation of antimicrobial resistance by extracytoplasmic function (ECF) sigma factors. *Microbes Infect.* 19, 238–248. doi: 10.1016/j.micinf.2017.01.007
- Wu, Z., Kong, Z., Lu, S., Huang, C., Huang, S., He, Y., et al. (2019). Isolation, characterization and the effect of indigenous heavy metal-resistant plant growth-promoting bacteria on sorghum grown in acid mine drainage polluted soils. *J. Gen. Appl. Microbiol.* 65, 254–264. doi: 10.2323/jgam.2018.11.004
- Xiao, X., Cao, G., Chen, F., Tang, Y., Liu, X., and Xu, W. (2015). Durable superhydrophobic wool fabrics coating with nanoscale Al₂O₃ layer by atomic layer deposition. *Appl. Surf. Sci.* 349, 876–879. doi: 10.1016/j.apsusc.2015.05.061
- Xue, D., Zeng, X., Lin, D., and Yao, S. (2019). Thermostable ethanol tolerant xylanase from a cold-adapted marine species *Acinetobacter johnsonii*. *Chin. J. Chem. Eng.* 27, 1166–1170. doi: 10.1016/j.cjche.2018.06.019
- Yang, S. J., Kang, I., and Cho, J. C. (2012). Genome sequence of strain IMCC14465, isolated from the East Sea, belonging to the *PS1 Clade* of *Alphaproteobacteria*. *J. Bacteriol.* 194, 6952–6953. doi: 10.1128/jb.01888-12
- Ye, J., Yin, H., Xie, D., Peng, H., Huang, J., and Liang, W. (2013). Copper biosorption and ions release by *Stenotrophomonas maltophilia* in the presence of benzo a pyrene. *Chem. Eng. J.* 219, 1–9. doi: 10.1016/j.cej.2012.12.093
- Yu, L., Duan, J., Du, X., Huang, Y., and Hou, B. (2013). Accelerated anaerobic corrosion of electroactive sulfate-reducing bacteria by electrochemical impedance spectroscopy and chronoamperometry. *Electrochem. Commun.* 26, 101–104. doi: 10.1016/j.elecom.2012.10.022
- Yu, Y., Hou, W., Hu, X., Yu, Y., Mi, L., and Song, L. (2014). Superhydrophobic modification of an Al₂O₃ microfiltration membrane with TiO₂ coating and PFDS grafting. *RSC Adv.* 4, 48317–48321. doi: 10.1039/c4ra07485h
- Zhang, C., Liu, Q., Li, X., Wang, M., Liu, X., Yang, J., et al. (2021). Spatial patterns and co-occurrence networks of microbial communities related to environmental heterogeneity in deep-sea surface sediments around Yap Trench, Western Pacific Ocean. *Sci. Total Environ.* 759:143799. doi: 10.1016/j.scitotenv.2020.143799
- Zhang, H., Hu, X., and Lu, H. (2017). Ni(II) and Cu(II) removal from aqueous solution by a heavy metal-resistance bacterium: kinetic, isotherm and mechanism studies. *Water Sci. Technol.* 76, 859–868. doi: 10.2166/wst.2017.275
- Zhang, W. W., Wen, Y. Y., Niu, Z. L., Yin, K., Xu, D. X., and Chen, L. X. (2012). Isolation and characterization of sulfonamide-degrading bacteria *Escherichia sp* HS21 and *Acinetobacter sp* HS51. *World J. Microbiol. Biotechnol.* 28, 447–452. doi: 10.1007/s11274-011-0834-z
- Zhang, X., Xu, W., Liu, Y., Cai, M., Luo, Z., and Li, M. (2018). Metagenomics reveals microbial diversity and metabolic potentials of seawater and surface sediment from a hadal biosphere at the Yap Trench. *Front. Microbiol.* 9:02402. doi: 10.3389/fmicb.2018.02402
- Zhang, Y., Ma, Y., Duan, J., Li, X., Wang, J., and Hou, B. (2019a). Analysis of marine microbial communities colonizing various metallic materials and rust layers. *Biofouling* 35, 429–442. doi: 10.1080/08927014.2019.1610881
- Zhang, Y., Ma, Y., Zhang, R., Zhang, B., Zhai, X., Li, W., et al. (2019b). Metagenomic resolution of functional diversity in copper surface-associated marine biofilms. *Front. Microbiol.* 10:02863. doi: 10.3389/fmicb.2019.02863
- Zhao, Y., Zhou, E., Xu, D., Yang, Y., Zhao, Y., Zhang, T., et al. (2018). Laboratory investigation of microbiologically influenced corrosion of 2205 duplex stainless steel by marine *Pseudomonas aeruginosa* biofilm using electrochemical noise. *Corros. Sci.* 143, 281–291. doi: 10.1016/j.corsci.2018.08.018
- Zhou, E., Li, H., Yang, C., Wang, J., Xu, D., Zhang, D., et al. (2018). Accelerated corrosion of 2304 duplex stainless steel by marine *Pseudomonas aeruginosa* biofilm. *Int. Biodeter. Biodegr.* 127, 1–9. doi: 10.1016/j.ibiod.2017.11.003
- Zuo, R. (2007). Biofilms: strategies for metal corrosion inhibition employing microorganisms. *Appl. Microbiol. Biotechnol.* 76, 1245–1253. doi: 10.1007/s00253-007-1130-6

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.

Copyright © 2022 Zhai, Cao, Zhang, Ju, Chen, Duan and Sun. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



MDGNN: Microbial Drug Prediction Based on Heterogeneous Multi-Attention Graph Neural Network

Jiangsheng Pi¹, Peishun Jiao¹, Yang Zhang^{2*} and Junyi Li^{1*}

¹ School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen, China, ² College of Science, Harbin Institute of Technology (Shenzhen), Shenzhen, China

OPEN ACCESS

Edited by:

George Tsiamis,
University of Patras, Greece

Reviewed by:

Yi Xiong,
Shanghai Jiao Tong University, China
Shahab S. Band,
National Yunlin University of Science
and Technology, Taiwan

*Correspondence:

Yang Zhang
zhangyang07@hit.edu.cn
Junyi Li
lijunyi@hit.edu.cn

Specialty section:

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

Received: 20 November 2021

Accepted: 07 March 2022

Published: 07 April 2022

Citation:

Pi J, Jiao P, Zhang Y and Li J
(2022) MDGNN: Microbial Drug
Prediction Based on Heterogeneous
Multi-Attention Graph Neural
Network.
Front. Microbiol. 13:819046.
doi: 10.3389/fmicb.2022.819046

Human beings are now facing one of the largest public health crises in history with the outbreak of COVID-19. Traditional drug discovery could not keep pace with newly discovered infectious diseases. The prediction of drug-virus associations not only provides insights into the mechanism of drug-virus interactions, but also guides the screening of potential antiviral drugs. We develop a deep learning algorithm based on the graph convolutional networks (MDGNN) to predict potential antiviral drugs. MDGNN is consisted of new node-level attention and feature-level attention mechanism and shows its effectiveness compared with other comparative algorithms. MDGNN integrates the global information of the graph in the process of information aggregation by introducing the attention at node and feature level to graph convolution. Comparative experiments show that MDGNN achieves state-of-the-art performance with an area under the curve (AUC) of 0.9726 and an area under the PR curve (AUPR) of 0.9112. In this case study, two drugs related to SARS-CoV-2 were successfully predicted and verified by the relevant literature. The data and code are open source and can be accessed from <https://github.com/Pijiangsheng/MDGNN>.

Keywords: antimicrobial drug prediction, graph convolution networks (GCN), heterogeneous network (Het-Net), representation learning, SARS-CoV-2

INTRODUCTION

Microorganisms are the unicellular or multicellular organisms, which include bacteria, archaea, viruses, protists, and fungi (Human Microbiome Project Consortium, 2012; Sommer and Bäckhed, 2013). Microbes sometimes can protect the human body from lethal pathogens, improve metabolism, and strengthen the immune system of the host (Ventura et al., 2009). On the other hand, the imbalance of the microbial community may cause a wide range of human diseases, such as obesity (Zhang et al., 2009), diabetes (Wen et al., 2008), rheumatoid arthritis (Lynch and Pedersen, 2016), and even cancer (Schwabe and Jobin, 2013).

As a novel coronavirus, SARS-CoV-2 has caused an unprecedented public health crisis recently. New variants of SARS-CoV-2 with the enhanced transmissibility are emerging globally. Traditional

drug development could not keep pace with threats from the fast-spreading SARS-CoV-2 and its variants, because of the complexity, high cost, and long experiment period of the traditional drug discovery process. The world needs to speed up the drug discovery process for COVID-19.

With the recent development of deep learning, especially the graph neural networks, more and more researchers have begun to try to find solutions based on the deep learning for their biological problems (Shamshirband et al., 2021; Zhang et al., 2021), such as drug interaction identification (Deng et al., 2020; Lin et al., 2020), protein function prediction (Gligorijević et al., 2021), virus classification (Deif et al., 2021), and disease-genes association prediction (Shu et al., 2021), etc. These studies show the potential of graph representation learning in biological questions.

In the research on microorganisms, there is a large amount of known information about the action of microorganisms and drugs, the genetic information of microorganisms, and the molecular formula information of small molecule drugs. We can use calculation-based methods to process these data to predict the possibility of interaction between microorganisms and drugs. This prediction allows us to initially screen out related therapeutic drugs for microorganisms that cause diseases, thereby speeding up the development of specific drugs for related diseases.

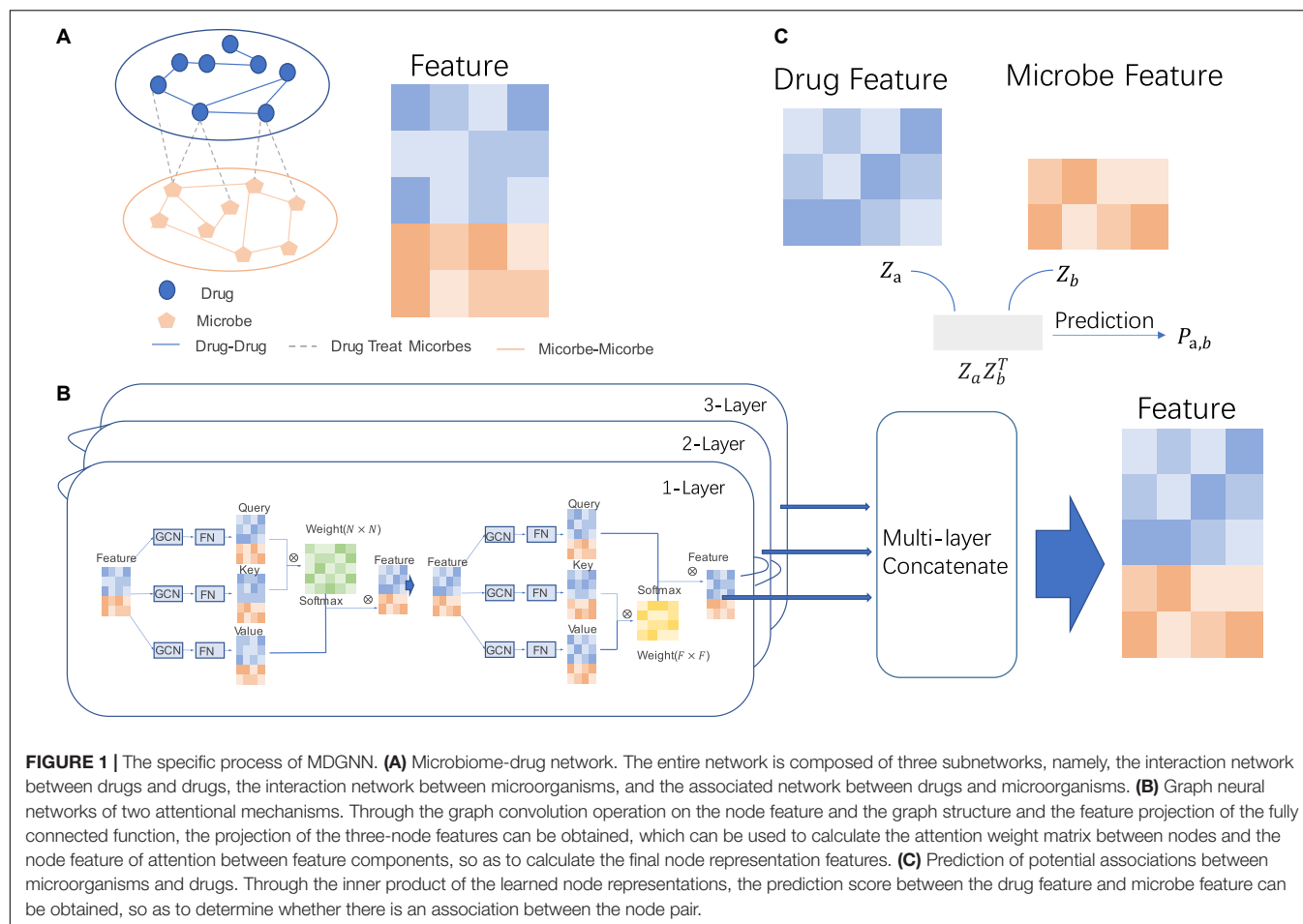
For microbial-drug association prediction, there are also several reported methods based on the graph representation algorithms. For example, Zhu et al. (2019) propose a method to predict human microbe-drug association, which is named Human Microbe-Drug Association by KATZ measure (HMDAKATZ). HMDAKATZ predicts possible drug-microbe associations using chemical similarity of drugs based on the Gaussian kernel similarity. Long et al. propose a Heterogeneous Network Embedding Representation framework for Microbe-Drugs Association prediction (HNERMDA) (Long and Luo, 2020). HNERMDA predicts drug-microbe association by heterogeneous graph neural network. Long et al. proposed a graph convolutional network (GCN)-based framework for predicting human microbe-drug associations, named GCNMDA (Long et al., 2020a). GCNMDA predicts drug-microbe association by introducing microbial protein interaction and chemical similarity of drugs. Long et al. propose a framework of heterogeneous graph attention networks to predict the association between drug and microbe (HGATDVA) (Long et al., 2020b). HGATDVA predicts drug-microbe associations by introducing a network of protein interactions between drug targets and microbial hosts. All of these previously reported methods first construct a heterogeneous network with microorganisms and drugs as nodes and then use some network representation methods to get the feature vectors of nodes in the heterogeneous network. For the prediction of the potential association between microorganisms and drugs, a common approach is to first build an action network with a variety of biological information, such as the interaction network between microorganisms and drugs. Then, the graph representation learning algorithm is used to learn node vector representation from the biological interaction network. Finally, the node representation vector obtained by the algorithm is

used to predict the probability of potential association between microorganisms and drugs.

Human Microbe-Drug Association by KATZ measure is the first algorithm used to predict potential links between microbes and drugs. In this method, the graph kernel similarity of microorganisms was calculated based on the known conditions to construct the microbial similarity network. Then, the drug similarity network was constructed according to the chemical structure similarity of drugs. By integrating the existing data of microbiota and drug association, a biological network with microbe and drug can be obtained. The KATZ algorithm was then used on this biological network to predict potential associations between microbes and drugs. HNERMDA is a method based on metapath2vec (Dong et al., 2017) to learn the node representation vector of microorganisms and drugs. By constructing an interaction network between microorganisms and drugs, it utilizes metapath2vec to learn their node representation vectors. Then, in the downstream prediction task, the bipartite graph recommendation algorithm with bias is used to predict the potential association between microorganisms and drugs. GCNMDA is a method that uses graph convolutional networks (GCNs) (Kipf and Welling, 2017) to learn node representation in heterogeneous biological networks composed of microorganisms and drugs, obtaining node representation vectors of microorganisms and drugs for the downstream prediction of potential drugs. Host protein information was introduced into the HGATDVA to construct two heterogeneous biological networks: One is a biological network composed of two isomeric nodes of microorganism and drug, and the other is a biological network composed of three isomeric nodes of microorganism, host protein, and drug. During node representation learning, graph attention networks (GAT) (Velickovic et al., 2018) were used to learn network representation of two biological networks, respectively, and two sets of node representation vectors were obtained. The node representation vectors of the two groups of microorganisms and drugs were added to predict the potential association between microorganisms and drugs. In the follow-up prediction of the association between microorganism and drug, the operation is carried out on these node feature vectors. Therefore, a good node feature can make our prediction result more accurate.

In the previously reported studies on the prediction of microbe-drug association, different graph representation learning algorithms are mainly used to improve the prediction performance. With the development of graph neural networks, there are more and more graph representation algorithms with better performance, such as GCN, GAT, heterogeneous graph attention networks (HANs) (Wang et al., 2019), and heterogeneous graph transformer (HGT) (Hu et al., 2020), etc.

In this article, we propose a model incorporating two attention mechanisms into a GCN to enhance the performance of graph characterization algorithms, thereby improving the performance of microbial-drug association prediction. In terms of relevant evaluation indicators, our model is better than the relevant benchmarks. In this case study, we predicted four SARS-CoV-2-related drugs on the relevant dataset through the trained model and verified the effectiveness of two of them in the latest database.



MATERIALS AND METHODS

We divide the information in the network into three levels of information. The first is the information inside the node. Generally speaking, entities such as drugs and microbes are abstracted into nodes in the network. To distinguish different nodes, unique features are assigned to nodes, namely, feature vectors of nodes in the network structure. The second is the information between nodes, the edges in the network. Finally, there is edge-to-edge information, the meta-path in the network.

The attention mechanism was first proposed in the field of natural language processing; that is, we can assign different weights to different word vectors. GAT is the earliest method to introduce an attention mechanism in the field of graph neural networks. It assigns different weights to different adjacency features in the stage of information aggregation. This kind of attention is not global attention, but only the attention between first-order neighbor nodes. HAN is an attention-based model for heterogeneous networks, which proposes two attention mechanisms: one is node-level attention, the other is attention between different meta-paths. First, feature vectors from different adjacent points are aggregated on each meta-path by the attention mechanism, and then, feature vectors from each source path

are aggregated by assigning different weights to each meta-path. Node-level attention in HAN is still the attention of local nodes, whereas attention between meta-paths is indirect global attention. But this approach relies heavily on setting up the meta-path. HGT is an improved approach to GCN that brings attention to message aggregation by introducing query vectors and key vectors. There are also two attention mechanisms in HGT, namely, attention between local nodes and attention between meta-paths.

However, the existing graph neural networks with attention mechanisms are all based on the local nodes; that is, the attention weight is only allocated between the source node and its neighbors. Due to the limitation of the network structure, the attention information between the source node and its higher-order neighbors is not calculated.

In view of the problems in the above methods, we propose two attention mechanisms, namely, the attention mechanism between all nodes and the attention mechanism between feature components within nodes. Through this new attention-based graph neural network, better node feature vectors for predicting microbial-drug association can be obtained. The whole prediction process is shown in **Figure 1**. Through node-attention, we can get the attention of one node in the graph to the

nodes of the whole graph. Through feature-attention, we can get the weight of each dimension between feature vectors of a node.

The prediction process is to first build a heterogeneous network with drug nodes and microbial nodes. In this network, there are microbial-microbial, microbial-drug, and drug-drug interactions. We mainly predict the potential association between microbial-drug interaction. Then, node-attention and feature-attention mechanisms are used to learn node representation on the network. Finally, after the representation vectors of the two heterogeneous nodes were obtained, they were directly used to predict the link between drugs and microorganisms.

The network representation algorithm is divided into three parts: node-attention, feature-attention, multi-layer feature fusion.

Node-Attention

Considering the sequence data, in which a single word is used as a data unit and connected together, we can think of sequence data as a special kind of graph structure, which can be regarded as a graph structure in which the in and out degrees of all nodes are 1. Different from GAT's node-attention mechanism, we also have a weight for higher-order neighbor nodes. Therefore, the advantage of using such global node-attention is that we can aggregate the node information of higher-order neighbors by calculating self-attention, instead of being limited to the structure of the graph to capture the information of other nodes.

Suppose there exists graph G , which can be represented by its adjacency matrix A and node feature matrix X , namely, $G = (A, X)$. For the nodes in graph G , we can calculate their weights and then aggregate the information based on the weights. Different from graph convolution operation, graph convolution operation aggregates information according to the graph structure. When aggregating information according to weight, it can break through the limitation of graph structure and aggregate corresponding information even when there is no edge connection between nodes (refer to **Figure 1B**).

In this paragraph, we will introduce some commonly used formulas in the following text, such as *GCN*, *Linear*. *GCN* is a neural network layer that can learn the structure information of graph structure data. The calculation method is shown in Formula (1)

$$Z = GCN(A, X) = ReLU(\tilde{A}XW), \quad (1)$$

which X is the original feature, $ReLU$ is activation function, W is the learnable parameter matrix, and \tilde{A} is the adjacency matrix with self-loop of the graph. *Linear* is a fully connected function, and its formula is shown in (2)

$$X' = Linear(X) = WX + b, \quad (2)$$

which X is the original feature, and W and b are learnable parameter matrix.

Our method is mainly based on the idea that *GCN* learns the structural information of the network and triplet attention learns the disconnect node interaction information. First, we aggregate node features in *GCN*, and after learning the structural information of the network, treat all nodes as sequence data

and temporarily ignore their structural information, as shown in Formula (3) (σ is non-linear activation function, like $ReLU$).

$$GCNLinear = \sigma(Linear(GCN(G, X))) \quad (3)$$

By using Formula (1), we can obtain the features of three groups of nodes needed to calculate triplet attention. Q_{Node} , K_{Node} , and V_{Node} . As shown in Formulas (4–6).

$$Q_{Node} = GCNLinear(G, X) \quad (4)$$

$$K_{Node} = GCNLinear(G, X) \quad (5)$$

$$V_{Node} = GCNLinear(G, X) \quad (6)$$

Then, the node weight matrix W_{Node} ($N \times N$) is obtained from its inner product, and its row direction is normalized, as shown in Formulas (7, 8).

$$W_{Node} = Q_{Node} \otimes K_{Node}^T \quad (7)$$

$$w_i = Softmax(w_i), i \in (0, 1, 2, \dots, n) \quad (8)$$

Finally, the inner product of weight matrix W_{Node} ($N \times N$) and V_{Node} ($N \times F$) is integrated to obtain the node feature matrix X_{Node} ($N \times F$), as shown in Formula (9).

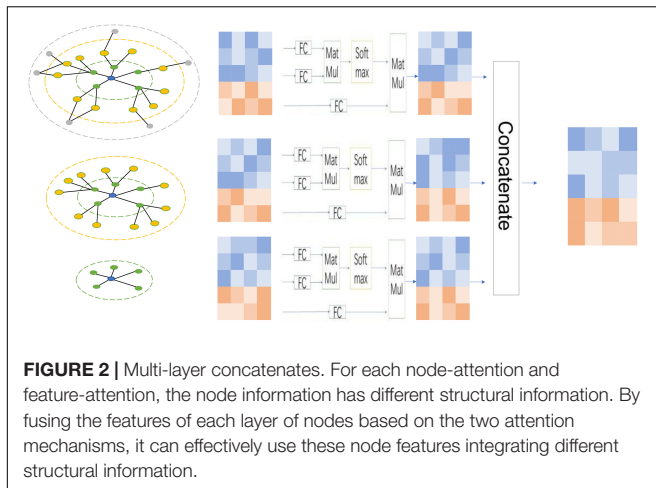
$$X_{Node} = X + W_{Node} \otimes V_{Node} \quad (9)$$

In this process, we model the information of interaction between nodes in the whole network by calculating a node weight matrix W_{Node} ($N \times N$). The node weight matrix W_{Node} ($N \times N$) is different from the adjacency matrix of the network A ($N \times N$), which can be regarded as the n power of the adjacency matrix A ($N \times N$), namely, W_{Node} ($N \times N$) = A^n ($N \times N$) and the n varies according to the size of the structure of the network.

Feature-Attention

The graph can be represented by the node set V and the edge set E , as well as the node eigenmatrix X ($N \times F$). For any node $N_i \in V$, node N_i can be represented by a node feature vector $(f_1, f_2, f_3, \dots, f_n)$. For a certain node N_i , we can express the importance of different features by feature weight vectors $(w_1, w_2, w_3, \dots, w_n)$, and distribute feature weights by inner product. In other words, for different nodes, there is always some feature components f_i , where $i \in 1, 2, 3, \dots, n$. In the dimension of f_i , this node is significantly different from other nodes. For some other feature components f_j , the values of all nodes are almost the same, so we need to give different weight values to these two different feature components. We use a feature component attention weight matrix to model the relationship between feature components within such nodes, as shown in the **Figure 1B**.

Just as in the calculation of node-attention, three feature vector matrices corresponding to node-features, query, key, and value, are first calculated. The difference lies in that we calculate the weight between node feature components through query



and key feature vector matrix, that is, attention weight matrix belonging to feature components, as shown in Formula (10)

$$W_{Feature} = Q_{Feature}^T \otimes K_{Feature} \quad (10)$$

It should be noted that the node weight matrix is $W_{Node} (N \times N)$ and the feature component weight matrix is $W_{Feature} (F \times F)$. After the matrix $W_{Feature} (F \times F)$ is obtained, the final $W_{Feature} (F \times F)$ is obtained through the normalization of the column direction, as shown in Formula (10). Then, the final feature vector of nodes is obtained by Formula (11, 12)

$$w_{\cdot j} = \text{Softmax}(w_{\cdot j}), j \in (0, 1, 2, \dots, n) \quad (11)$$

$$X_{Feature} = X + V_{Feature} \otimes W_{Feature} \quad (12)$$

Multi-Layer Feature Concatenates

Generally speaking, GCN can only aggregate information to first-order neighbors of the source node, whereas aggregation to higher-order information requires the number of layers of stacked GCN. For GCN with different layers, the node information represented by GCN is obtained by aggregating the node information within the scope of different graph structures, and these node feature vectors have different structural semantic information. By integrating the node information obtained from these different GCN layers, better results can be obtained for link prediction. For example, in jump-knowledge networks (Xu et al., 2018), node features from different GCN layers are added or spliced as the final node features. It is worth noting that jump-knowledge networks simply add up the node information learned from GCN of different layers and serve as the final node information.

Suppose that for graph $G(A, X)$, after n layer message aggregation, a list of node features $(X_1, X_2, X_3, \dots, X_n)$ will be obtained. The feature vector matrix in this list represents the node features obtained by integrating the substructure information of different graphs. We use triple-based attention to assign different weights to these node features and then fuse them for downstream tasks. We give a schematic diagram of node features

obtained by three-layer GNN, as shown in **Figure 2**. Specifically, for each set of node features, we use the following formulas to calculate,

$$Q_i = \sigma(\text{Linear}(X_i)), i \in (1, 2, \dots, n) \quad (13)$$

$$K_i = \sigma(\text{Linear}(X_i)), i \in (1, 2, \dots, n) \quad (14)$$

$$V_i = \sigma(\text{Linear}(X_i)), i \in (1, 2, \dots, n) \quad (15)$$

After calculating the Q_i , K_i , and V_i corresponding to each group of node features, the final feature vector X'_i of the group of nodes can be calculated by Formula (16)

$$X'_i = \text{Softmax}(Q_i \otimes K_i^T) \otimes V_i \quad (16)$$

Finally, by concatenating multiple sets of node features, the final node feature X' can be gain by Formula (17), which can be used to predict the score.

$$X' = (X'_1 \parallel X'_2 \dots \parallel X'_n) \quad (17)$$

Microbial Drug Association Prediction

After getting the final feature vector X of the microbe node and the drug node, the prediction score between a certain microbe and the drug node pair can be calculated, that is, the probability of the correlation between the microbe and the drug, as shown in Formula (18)

$$S_{(u,v)} = \text{Sigmoid}(X_u \otimes X_v) \quad (18)$$

where $X_u \in \mathcal{R}^{1 \times n}$, $X_v \in \mathcal{R}^{1 \times n}$, and Sigmoid is an activation function.

During the training process, we use binary cross-entropy as our loss function for training, as shown in Formula (19)

$$\text{loss} = \sum_{(u,v) \in \text{pos, neg}} \text{BCE}(S_{(u,v)}, A_{(u,v)}), \quad (19)$$

while A is the adjacency matrix, and $(u, v) \in \text{pos}$ means $A_{(u,v)} = 1$, and $(u, v) \in \text{neg}$ means $A_{(u,v)} = 0$.

RESULTS

Dataset

In the experiment, we used data coming from three datasets: DrugVirus (Andersen et al., 2020), MDAD (Sun et al., 2018), and aBiofilm (Rajput et al., 2018). We integrated the data of these three datasets after removing duplicate microorganisms and drugs. By calculating the similarity of drug structure, and taking the drug interaction with similarity greater than 0.5 as the relationship between drugs, the drug interaction network is obtained. Similarly, the microbial similarity is calculated through the microbial gene sequence, and the microbial similarity greater than 0.5 is taken as the microbial association to obtain the microbial interaction network. The data used in our experiment are shown in **Tables 1, 2**.

TABLE 1 | Data used in this study were obtained by integrating three datasets: DrugVirus, MDAD, and aBiofilm.

Name	Number
Drugs	3,091
Microbes	328
Drug–drug interaction	270,877
Microbe–microbe interaction	467
Drug–microbe interaction	3,900

TABLE 2 | The statistics for each microbe–drug association dataset.

Datasets	Microbes	Drugs	Associations
MDAD	173	1,373	2,470
aBiofilm	140	1,720	2,884
DrugVirus	95	175	933

TABLE 3 | Comparative experiment of different benchmarks and MDGNN.

Model	AUC	AUPR
GCN	0.9439 (0.0038)	0.8721 (± 0.0102)
GAT	0.9385 (0.0057)	0.8479 (0.0139)
HAN	0.9443 (0.0041)	0.8086 (0.0118)
HGT	0.9251 (0.0073)	0.8275 (0.0126)
GCNMDA	0.9541 (0.0036)	0.8796 (0.0103)
GraphSAINT	<u>0.9653</u> (0.0081)	<u>0.8938</u> (0.0135)
MDGNN	0.9721 (0.0053)	0.9102 (0.0118)

MDGNN outperforms all baselines including GCN, GAT, HAN, HGT, GCNMDA, and GraphSAINT.

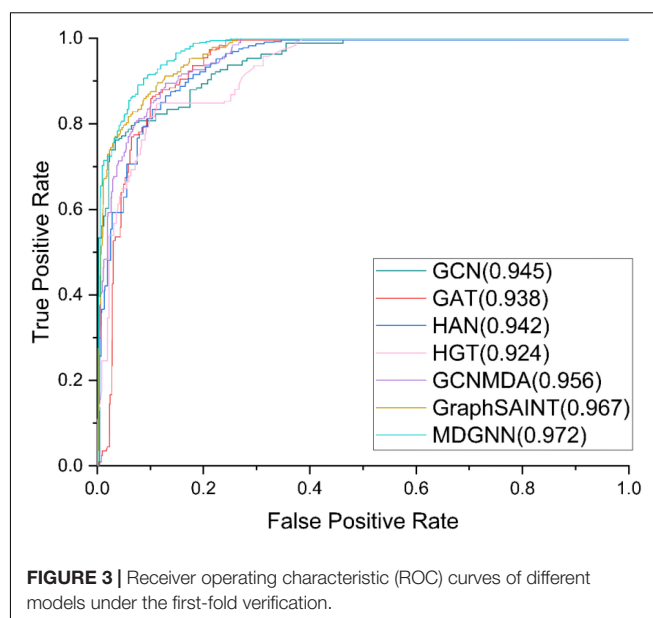
The bold value means the best model, and the underlined value means the second-best model.

Experiment Result

To verify the effectiveness of our method, we divided the dataset by 5-fold cross-validation of the data related to the known microorganisms a drug and randomly divided the data related to the known microorganisms and drugs into five groups. In each fold experiment, we take turns to select a group of related data as the test set, and the remaining four groups as the training set for training. In addition, because in the real world, it is more common that there is no interaction between microorganisms and drugs. At the same time, to compare the performance of each model in the case of unbalanced positive and negative samples, we set the number of negative samples in the experiment set to four.

In our model, we set that the learning rate in optimization algorithm was 0.001 with Adam optimizer, and other related hyperparameters, such as the number of model layers, feature dimensions, and training times, are described in the ablation experiments. The equipment used in the experiment is Intel(R) Xeon(R) Silver 4114 CPU @ 2.20 GHz, running memory is 128 GB, hard disk storage space is 10TB, and it is equipped with two Tesla P40 GPU with a total memory capacity of 48 GB.

The comparative models we used are GCN, GAT, HAN, HGT, GCNMDA, and GraphSAINT (Zeng et al., 2020). The

**FIGURE 3** | Receiver operating characteristic (ROC) curves of different models under the first-fold verification.

hyperparameters of the benchmark model are set according to their papers. The experimental results are shown in **Table 3**. ROC curves of the models are shown in **Figure 3**.

Area under the curve (AUC) is an index to measure the sorting performance. It is not sensitive to the balance of positive and negative samples. When the samples are unbalanced, it can also make a reasonable evaluation, which is suitable for measuring the sorting task. The closer of the result is to 1, the better performance it is.

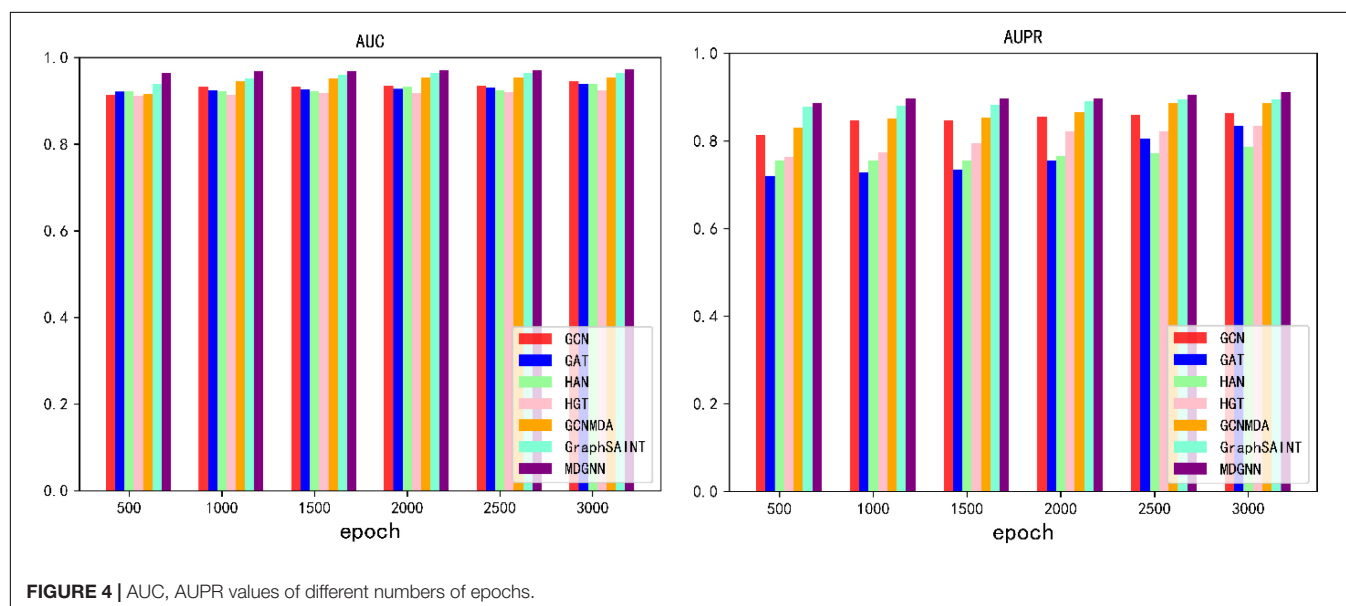
Area under the PR curve (AUPR) is the area value under the curve composed of recall rate and accuracy rate in the prediction results. It is generally used to measure the performance of correct prediction results in the dataset with unbalanced positive and negative samples.

Under a single index, the bold one is the best model, and the underlined one is the second-best model. It can be seen that the performance of our model under AUC evaluation index is ahead of state-of-the-art baseline GraphSAINT. Our model achieves an AUC of 0.9721, better than GraphSAINT, which is 0.9653. Under the evaluation index of AUPR, the performance of our model is significantly ahead of other models. Compared with state-of-the-art baseline GraphSAINT (0.8938), our model (0.9102) has increased by about 1.74%, which is better than GCN (0.8721), GAT (0.8479), HAN (0.8086), HGT (0.8275), and GCNMDA (0.8796).

Through comparative experiments with baseline, it can be seen that our model has achieved a great improvement in performance after calculating the attention between all nodes based on the entire graph. Compared with the model that calculates the attention between 1-hop neighbor nodes, our model is more able to mine the relationship between high-order neighbor nodes. In the association of microbial and drugs, an intuitive idea is if drug A interacts with drug B, and drug A interacts with microorganisms C, then we are likely to be inclined to speculate that drug B and microorganisms

TABLE 4 | Compare the 5-fold crossover experimental results of MDGNN and GCNMDA on three small datasets (MDAD, aBiofilm, and DrugVirus).

Methods	MDAD		aBiofilm		DrugVirus	
	AUC	AUPR	AUC	AUPR	AUC	AUPR
GCNMDA	0.9423 (0.0105)	0.9376 (0.0115)	0.9517 (0.0035)	0.9488 (0.0031)	0.8986 (0.0305)	0.9038 (0.0372)
MDGNN	0.9457 (0.0083)	0.9431 (0.0102)	0.9608 (0.0054)	0.9566 (0.0084)	0.8737 (0.0167)	0.8904 (0.0212)



C have an interaction. When calculating 1-hop-based attention (such as GAT, HAN, HGT, and GCNMDA), this indirect correlation between drug B and microorganism C is ignored. However, in MDGNN, this indirect correlation will be taken into consideration, and the message will be passed between the nodes B and C through our proposed method, thus improving the prediction performance of the model.

By comparing MDGNN and GCNMDA on three small datasets, we can further confirm our inference. On large dataset (MDAD, aBiofilm), our method is better than GCNMDA, and especially on aBiofilm dataset, our method can be nearly a percentage point higher than GCNMDA. This dataset is the most data in these three datasets. On the smallest dataset (DrugVirus), our method (AUC:0.8737, AUPR:0.8943) is inferior to GCNMDA (AUC:0.8986, AUPR:0.9038).

According to the results in **Table 4**, it can be seen that when the size of dataset grows, the number of indirect associations (like the relationship between B and C mentioned above) in the dataset will increase accordingly. This means that on large dataset, our method can learn more information about potential associations, and many of our final predictions of the association between microorganisms and drugs are inferred based on this potential association information.

It can be seen that the calculation of the two kinds of attention brings stronger fitting ability to the model. Moreover, this powerful fitting ability allows our model to learn more structural information every time it performs

TABLE 5 | Ablation experiments on modules of feature-attention and multi-layer feature.

Multi-layer	Feature-attention	Layer	AUC	AUPR
w/o Multi-layer	w/o Feature-attention	3 layers	0.9621	0.8891
		4 layers	0.9614	0.8816
		5 layers	0.9652	0.8909
	Feature-attention	3 layers	0.9637	0.8896
		4 layers	0.9640	0.8909
Multi-layer	w/o Feature-attention	5 layers	0.9694	0.9034
		3 layers	0.8696	0.7362
		4 layers	0.8760	0.7264
	Feature-attention	5 layers	0.8756	0.7410
		3 layers	0.9706	0.8982
		4 layers	0.9711	0.9097
		5 layers	0.9726	0.9112

The bold value means the best model.

gradient descent, so as to converge more quickly during the training process.

During the training process, we found that the optimization of these baseline training is extremely slow, and our model converges fast, so we train the model under different epochs settings to compare the effect of model training. When different numbers of epochs were set, the results obtained by each model are shown in **Figure 4**.

As can be seen from the experimental results, our model can converge to the optimal value within a very short training period.

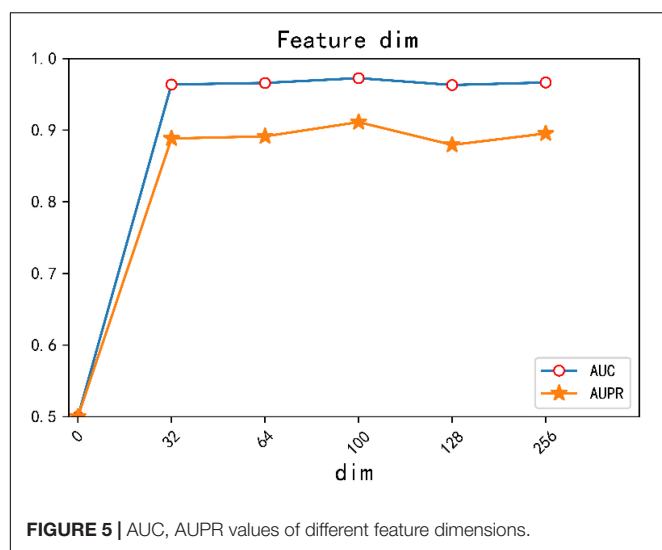


TABLE 6 | Predicted drugs that can treat SARS-CoV-2 (negative means that the drug is not associated with SARS in our dataset).

Predicted drugs	Prediction score
Mefloquine	0.9361
Darunavir (negative)	0.9177
Nelfinavir	0.9096
Azithromycin(negative)	0.8904
Vancomycin	0.8731
Dicinnamyl (negative)	0.8685
Nicosamide	0.8663
Chitosan (negative)	0.8529
Chlorpromazine	0.8467
Ribavirin	0.8406

Under the same epoch value, our model has greatly improved compared with other models. MDGNN requires less than 500 epochs to make the AUC converge to above 0.96, while other comparison models fail to exceed 0.95 after 3,000 epochs.

Through ablation experiments, we can analyze the role of each module. In the experiment, we analyze the function of each module by setting a model with different number of blocks. The specific ablation experimental results are shown in **Table 5**.

As can be seen from **Table 4**, when both modules are used, the performance is the best. Specifically, when the node information integrated with node-attention is directly aggregated through multi-layer module, the model will produce negative optimization. The reason may be that after removing the feature-attention, the calculation in the multi-layer module is performed directly on the node vector that incorporates the node-attention, which will cause the decoupling of the attention calculated in the node-attention, which results in a decrease in the result.

In addition, we conducted comparative experiments on the dimensions of different feature vectors and verified that the best experimental results were obtained when the dimension of feature vector was set to 100. The result is shown in **Figure 5**.

Case Study

In case study, we use the deduplicated datasets, which contains the SARS-CoV-2-related data from the DrugVirus dataset. We save the trained model parameters and use them to make predictions on the entire dataset. The parameters of the model are trained on the randomly divided training set, selected, and saved according to the results on the test set.

We load the trained model and then input the entire dataset into the model to obtain the feature vector of microorganisms and the feature vector of drugs. The corresponding microbial drug association score is obtained by inner product of the feature vector of microbe and the feature vector of drug.

Taking SARS-CoV-2 as an example, we predicted the drugs that may treat the virus and took out the 10 drugs with the greatest possibility. The results are shown in **Table 6**.

Among the ten drugs that we predicted to treat SARS-CoV-2, four drugs were not associated with SARS-CoV-2 in our dataset, but our model predicted that these four drugs had a high potential to treat SARS-CoV-2. Through searching PubChem database, we found that two of the four drugs can indeed treat SARS-CoV-2. Darunavir is an antiretroviral protease inhibitor that is used in the therapy and prevention of human immunodeficiency virus (HIV) infection and the acquired immunodeficiency syndrome (AIDS) (Deeks, 2018). In our dataset, there is indeed an association between Darunavir and HIV, but there is no association between Darunavir and SARS-CoV-2 (Costanzo et al., 2020). This real association does not exist in our dataset, and we can predict this association through the dataset. Similarly, Azithromycin is a drug that can treat SARS-CoV-2 (Rosenberg et al., 2020). However, there is no association between Azithromycin and SARS-CoV-2 in our dataset where Azithromycin is only associated with Hepatitis C virus and HIV. In addition, our model successfully predicts the potential association between Azithromycin and SARS-CoV-2.

CONCLUSION

With the rapid development of deep learning, there are many deep learning methods reported for drug development. For example, Beck et al. identified commercially available drugs to treat viral proteins using a pretrained deep learning-based drug target interaction model. Their results showed that drugs used to treat HIV might be effective against SARS-CoV-2 (Beck et al., 2020). Joshi et al. (2020) used deep learning methods to predict the structural formula of chemical molecules and predict potential drugs for SARS-CoV-2. A total of 39 potential drugs for SARS-CoV-2 were predicted based on the ChEMBL dataset.

The rapid spread of SARS-COV-2 and its variants have resulted a serious public health crisis. How to develop a specific drug quickly to tackle SARS-CoV-2 and its variants is an urgent problem. We propose a novel attentional mechanism-based graph neural network framework for learning network node representation and prove that our framework is superior to other state-of-the-art methods, which includes GCN, GAT, HAN, and HGT, etc. In addition, through a large number of

drug and microbial data, we have screened potential drugs for the treatment of SARS-CoV-2, most of which are known to treat SARS-CoV-2.

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.

AUTHOR CONTRIBUTIONS

JP and JL designed the study, performed bioinformatics analysis, and drafted the manuscript. JL conceived the study

and coordination and drafted the manuscript. All authors participated in the revision of the manuscript and read and approved the final manuscript.

FUNDING

This work was supported by the grants from the National Key R&D Program of China (2021YFA0910700), Shenzhen Science and Technology University Stable Support Program (GXWD20201230155427003-20200821222112001), Guangdong Key Area Research Program (2020B0101380001), Shenzhen Science and Technology Program (JCYJ20200109113201726), and Guangdong Basic and Applied Basic Research Foundation (2021A1515220115).

REFERENCES

- Andersen, P. I., Ianevski, A., and Lysvand, H. (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
- Beck, B. R., Shin, B., Choi, Y., Park, S., and Kang, K. (2020). Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model. *Comput. Struct. Biotechnol. J.* 18, 784–790. doi: 10.1016/j.csbj.2020.03.025
- Costanzo, M., De Giglio, M. A. R., and Roviello, G. N. S. A. R. S. (2020). CoV-2: recent reports on antiviral therapies based on lopinavir/ritonavir, darunavir/umifenovir, hydroxychloroquine, remdesivir, favipiravir and other drugs for the treatment of the new coronavirus. *Curr. Med. Chem.* 27, 4536–4541. doi: 10.2174/0929867327666200416131117
- Deeks, E. D. (2018). Darunavir/cobicistat/emtricitabine/tenofovir alafenamide: a review in HIV-1 infection. *Drugs* 78, 1013–1024. doi: 10.1007/s40265-018-0934-2
- Deif, M. A., Solyman, A. A. A., Kamarposhti, M. A., Band, S. S., and Hammam, R. E. (2021). A deep bidirectional recurrent neural network for identification of SARS-CoV-2 from viral genome sequences. *Math. Biosci. Eng.* 18, 8933–8950. doi: 10.3934/mbe.2021440
- Deng, Y., Xu, X., Qiu, Y., Xia, J., Zhang, W., and Liu, S. (2020). A multimodal deep learning framework for predicting drug–drug interaction events. *Bioinformatics* 36, 4316–4322. doi: 10.1093/bioinformatics/btaa501
- Dong, Y., Chawla, N. V., and Swami, A. (2017). “metapath2vec: scalable representation learning for heterogeneous networks,” in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, (New York, NY: ACM), 135144.
- Gligorijević, V., Renfrew, P. D., Kosciulek, T., Leman, J. K., Berenberg, D., Vatanen, T., et al. (2021). Structure-based protein function prediction using graph convolutional networks. *Nat. Commun.* 12, 1–14. doi: 10.1038/s41467-021-23303-9
- Hu, Z., Dong, Y., and Wang, K. (2020). “Heterogeneous graph transformer,” in *Proceedings of the Web Conference 2020*, (Taipei: Web Conference 2020).
- Human Microbiome Project Consortium. (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486:207. doi: 10.1038/nature11234
- Joshi, T., Joshi, T., Pundir, H., Sharma, P., Mathpal, S., and Chandra, S. (2020). Predictive modeling by deep learning, virtual screening and molecular dynamics study of natural compounds against SARS-CoV-2 main protease. *J. Biomol. Struct. Dyn.* 39, 6728–6746. doi: 10.1080/07391102.2020.1802341
- Kipf, T., and Welling, M. (2017). “Semi-supervised classification with graph convolutional networks,” in *Proceedings of the International Conference on Learning Representations*, (Vancouver, BC).
- Lin, X., Quan, Z., Wang, Z. J., Ma, T., and Zeng, X. (2020). “KGNN: knowledge graph neural network for drug-drug interaction prediction,” in *Proceedings of the 29th International Joint Conferences on Artificial Intelligence Virtual*, Vol. 380, (Saxony: IJCAI), 2739–2745.
- Long, Y., and Luo, J. (2020). Association mining to identify microbe drug interactions based on heterogeneous network embedding representation. *IEEE J. Biomed. Health Inform.* 25, 266–275. doi: 10.1109/JBHI.2020.2998906
- Long, Y., Wu, M., and Kwok, C. K. (2020a). Predicting human microbe–drug associations via graph convolutional network with conditional random field. *Bioinformatics* 36, 4918–4927. doi: 10.1093/bioinformatics/btaa598
- Long, Y., Zhang, Y., and Wu, M. (2020b). “Predicting drugs for COVID-19/SARS-CoV-2 via heterogeneous graph attention networks,” in *Proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, (New Jersey, NJ: IEEE).
- Lynch, S. V., and Pedersen, O. (2016). The human intestinal microbiome in health and disease. *New Engl. J. Med.* 375, 2369–2379.
- Rajput, A., Thakur, A., Sharma, S., and Kumar, M. (2018). aBiofilm: 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/gkx1157
- Rosenberg, E. S., Dufort, E. M., Udo, T., Wilberschied, L. A., Kumar, J., Tesoriero, J., et al. (2020). Association of treatment with hydroxychloroquine or azithromycin with in-hospital mortality in patients With COVID-19 in New York State. *JAMA* 323, 2493–2502. doi: 10.1001/jama.2020.8630
- Schwabe, R. F., and Jobin, C. (2013). The microbiome and cancer. *Nat. Rev. Cancer* 13, 800–812.
- Shamshirband, S., Fathi, M., Dehngani, A., Chronopoulos, A. T., Alinejad-Rokny, H., et al. (2021). A review on deep learning approaches in healthcare systems: taxonomies, challenges, and open issues. *J. Biomed. Inform.* 113:103627. doi: 10.1016/j.jbi.2020.103627
- Shu, J., Li, Y., Wang, S., Xi, B., and Ma, J. (2021). Disease gene prediction with privileged information and heteroscedastic dropout. *Bioinformatics* 37, i410–i417. doi: 10.1093/bioinformatics/btab310
- Sommer, F., and Bäckhed, F. (2013). The gut microbiota—masters of host development and physiology. *Nat. Rev. Microbiol.* 11, 227–238. doi: 10.1038/nrmicro2974
- Sun, Y. Z., Zhang, D. H., Cai, S. B., Ming, Z., Li, J.-Q., and Chen, X. (2018). MDAD: a special resource for microbe–drug associations. *Front. Cell. Infect. Microbiol.* 8:424. doi: 10.3389/fcimb.2018.00424
- Velickovic, P., Cucurull, G., and Casanova, A. (2018). “Graph attention networks,” in *Proceedings of the International Conference on Learning Representations*, (Vancouver, BC: ICLR).
- Ventura, M., O’flaherty, S., Claesson, M. J., Turrioni, F., Klaenhammer, T. R., van Sinderen, D., et al. (2009). Genome-scale analyses of health-promoting bacteria: probiogenomics. *Nat. Rev. Microbiol.* 7, 61–71. doi: 10.1038/nrmicro2047
- Wang, X., Ji, H., and Shi, C. (2019). “Heterogeneous graph attention network,” in *Proceedings of the World Wide Web Conference*, (Washington, DC: WWW).
- Wen, L., Ley, R. E., Volchkov, P. Y., Stranges, P. B., Avanesyan, L., Stonebraker, A. C., et al. (2008). Innate immunity and intestinal microbiota in the development of type 1 diabetes. *Nature* 455, 1109–1113. doi: 10.1038/nature07336
- Xu, K., Li, C., Tian, Y., Sonobe, T., Kawarabayashi, K., Jegelka, S. et al. (2018). “Representation learning on graphs with jumping knowledge networks,” in

- Proceedings of 35th International Conference on Machine Learning*, (Stockholm), 5453–5462.
- Zeng, H., Zhou, H., Srivastava, A., Kannan, R., and Prasanna, V. (2020). “GraphSAINT: graph sampling based inductive learning method,” in *Proceedings of the International Conference on Learning Representations*, (Addis Ababa).
- Zhang, H., DiBaise, J. K., Zuccolo, A., Kudrna, D., Braidotti, M., Yu, Y., et al. (2009). Human gut microbiota in obesity and after gastric bypass. *Proc. Natl. Acad. Sci. U.S.A* 106, 2365–2370. doi: 10.1073/pnas.0812600106
- Zhang, Y., Ye, T., Xi, H., Juhas, M., and Li, J. (2021). Deep learning driven drug discovery: tackling severe acute respiratory syndrome coronavirus 2. *Front. Microbiol.* 12:739684. doi: 10.3389/fmicb.2021.739684
- Zhu, L., Duan, G., and Yan, C. (2019). “Prediction of microbe-drug associations based on Katz measure,” in *Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, (New Jersey, NJ: IEEE).

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.

Copyright © 2022 Pi, Jiao, Zhang and Li. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



EMDS-6: Environmental Microorganism Image Dataset Sixth Version for Image Denoising, Segmentation, Feature Extraction, Classification, and Detection Method Evaluation

Peng Zhao¹, Chen Li^{1*}, Md Mamunur Rahaman^{1,2}, Hao Xu¹, Pingli Ma¹, Hechen Yang¹, Hongzan Sun³, Tao Jiang^{4*}, Ning Xu⁵ and Marcin Grzegorzczek⁶

¹ Microscopic Image and Medical Image Analysis Group, College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China, ² School of Computer Science and Engineering, University of New South Wales, Sydney, NSW, Australia, ³ Department of Radiology, Shengjing Hospital, China Medical University, Shenyang, China, ⁴ School of Control Engineering, Chengdu University of Information Technology, Chengdu, China, ⁵ School of Arts and Design, Liaoning Petrochemical University, Fushun, China, ⁶ Institute of Medical Informatics, University of Lübeck, Lübeck, Germany

OPEN ACCESS

Edited by:

George Tsiamis,
University of Patras, Greece

Reviewed by:

Muhammad Hassan Khan,
University of the Punjab, Pakistan
Elias Asimakis,
University of Patras, Greece

*Correspondence:

Chen Li
lichen201096@hotmail.com
Tao Jiang
jiang@cuit.edu.cn

Specialty section:

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

Received: 04 December 2021

Accepted: 28 March 2022

Published: 25 April 2022

Citation:

Zhao P, Li C, Rahaman MM, Xu H, Ma P, Yang H, Sun H, Jiang T, Xu N and Grzegorzczek M (2022) EMDS-6: Environmental Microorganism Image Dataset Sixth Version for Image Denoising, Segmentation, Feature Extraction, Classification, and Detection Method Evaluation. *Front. Microbiol.* 13:829027. doi: 10.3389/fmicb.2022.829027

Environmental microorganisms (EMs) are ubiquitous around us and have an important impact on the survival and development of human society. However, the high standards and strict requirements for the preparation of environmental microorganism (EM) data have led to the insufficient of existing related datasets, not to mention the datasets with ground truth (GT) images. This problem seriously affects the progress of related experiments. Therefore, This study develops the *Environmental Microorganism Dataset Sixth Version* (EMDS-6), which contains 21 types of EMs. Each type of EM contains 40 original and 40 GT images, in total 1680 EM images. In this study, in order to test the effectiveness of EMDS-6. We choose the classic algorithms of image processing methods such as image denoising, image segmentation and object detection. The experimental result shows that EMDS-6 can be used to evaluate the performance of image denoising, image segmentation, image feature extraction, image classification, and object detection methods. EMDS-6 is available at the <https://figshare.com/articles/dataset/EMDS6/17125025/1>.

Keywords: environmental microorganism, image denoising, image segmentation, feature extraction, image classification, object detection

1. INTRODUCTION

1.1. Environmental Microorganisms

Environmental Microorganisms (EMs) usually refer to tiny living that exists in nature and are invisible to the naked eye and can only be seen with the help of a microscope. Although EMs are tiny, they significantly impacts human survival (Madigan et al., 1997; Rahaman et al., 2020). Some beneficial bacteria can be used to produce fermented foods such as cheese and bread from a beneficial perspective. Meanwhile, Some beneficial EMs can degrade plastics, treat sulfur-containing waste gas in industrial, and improve the soil. From a harmful point of view,

EMs cause food spoilage, reduce crop production and are also one of the chief culprits leading to the epidemic of infectious diseases. To make better use of the advantages of environmental microorganisms and prevent their harm, a large number of scientific researchers have joined the research of EMs. The image analysis of EM is the foundation of all this.

EMs are tiny in size, usually between 0.1 and 100 microns. This poses certain difficulties for the detection and identification of EMs. Traditional “morphological methods” require researchers to look directly under a microscope (Madsen, 2008). Then, the results are presented according to the shape characteristics. This traditional method requires more labor costs and time costs. Therefore, using computer-assisted feature extraction and analysis of EM images can enable researchers to use their least professional knowledge with minimum time to make the most accurate decisions.

1.2. EM Image Processing and Analysis

Image analysis is a combination of mathematical models and image processing technology to analyze and extract certain intelligence information. Image processing refers to the use of computers to analyze images. Common image processing includes image denoising, image segmentation and feature extraction. Image noise refers to various factors in the image that hinder people from accepting its information. Image noise is generally generated during image acquisition, transmission and compression (Pitas, 2000). The aim of image denoising is to recover the original image from the noisy image (Buades et al., 2005). Image segmentation is a critical step of image processing to analyze an image. In the segmentation, we divide an image into several regions with unique properties and extract regions of interest (Kulwa et al., 2019). Feature extraction refers to obtaining important information from images such as values or vectors (Zebari et al., 2020). Moreover, these characteristics can be distinguished from other types of objects. Using these features, we can classify images. Meanwhile, the features of an image are the basis of object detection. Object detection uses algorithms to generate object candidate frames, that is, object positions. Then, classify and regress the candidate frames.

1.3. The Contribution of Environmental Microorganism Image Dataset Sixth Version (EMDS-6)

Sample collections of the EMs are usually performed outdoors. When transporting or moving samples to the laboratory for observation, drastic changes in the environment and temperature affect the quality of EM samples. At the same time, if the researcher observes EMs under a traditional optical microscope, it is very prone to subjective errors due to continuous and long-term visual processing. Therefore, the collection of environmental microorganism image datasets is challenging (Kosov et al., 2018). Most of the existing environmental microorganism image datasets are not publicly available. This has a great impact on the progress of related scientific research. For this reason, we have created the *Environmental Microorganism Image Dataset Sixth Version*

TABLE 1 | Basic information of EMDS-6 dataset, including Number of original images (NoOI), Number of GT images (NoGT).

Class	NoOI	NoGT	Class	NoOI	NoGT
Actinophrys	40	40	Ceratium	40	40
Arcella	40	40	Stentor	40	40
Aspidisca	40	40	Siprostomum	40	40
Codosiga	40	40	K. Quadrata	40	40
Colpoda	40	40	Euglena	40	40
Epistylis	40	40	Gymnodinium	40	40
Euglypha	40	40	Gonyaulax	40	40
Paramecium	40	40	Phacus	40	40
Rotifera	40	40	Stylonychia	40	40
Vorticella	40	40	Synchaeta	40	40
Noctiluca	40	40	-	-	-
Total	840	840	Total	840	840

(EMDS-6) and made it publicly available to assist related scientific researchers. Compared with other environmental microorganism image datasets, EMDS-6 has many advantages. The dataset contains a variety of microorganisms and provides possibilities for multi-classification of EM images. In addition, each image of EMDS-6 has a corresponding ground truth (GT) image. GT images can be used for performance evaluation of image segmentation and object detection. However, the GT image production process is extremely complicated and consumes enormous time and human resources. Therefore, many environmental microorganism image dataset does not have GT images. However, our proposed dataset has GT images. In our experiments, EMDS-6 can provide robust data support in tasks such as denoising, image segmentation, feature extraction, image classification and object detection. Therefore, the main contribution of the EMDS-6 dataset is to provide data support for image analysis and image processing related research and promote the development of EMs related experiments and research.

2. MATERIALS AND METHODS

2.1. EMDS-6 Dataset

There are 1680 images in the EMDS-6 dataset, including 21 classes of original EM images with 40 images per class, resulting in a total of 840 original images, and each original image is followed by a GT image for a total of 840. **Table 1** shows the details of the EMDS-6 dataset. **Figure 1** shows some examples of the original images and GT images in EMDS-6. EMDS-6 is freely published for non-commercial purpose at: <https://figshare.com/articles/dataset/EMDS6/17125025/1>.

The collection process of EMDS-6 images starts from 2012 till 2020. The following people have made a significant contribution in producing the EMDS-6 dataset: Prof. Beihai Zhou and Dr Fangshu Ma from the University of Science and Technology Beijing, China; Prof. Dr.-Ing. Chen Li and M.E. HaoXu from Northeastern University, China; Prof. Yanling Zou from Heidelberg University, Germany. The GT images of the EMDS-6

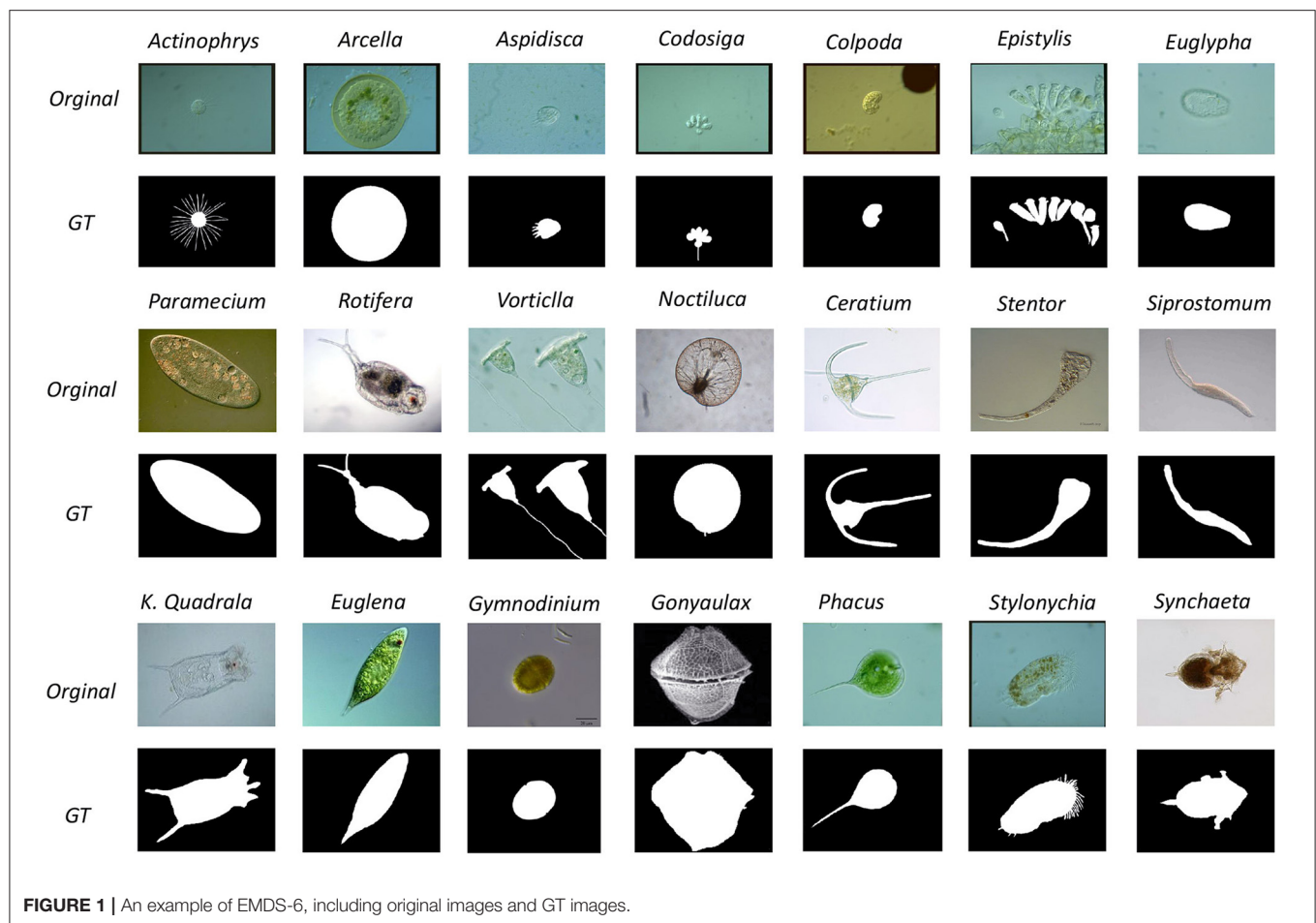


FIGURE 1 | An example of EMDS-6, including original images and GT images.

dataset are produced by Prof. Dr.-Ing Chen Li, M.E. Bolin Lu, M.E. Xuemin Zhu and B.E. Huaqian Yuan from Northeastern University, China. The GT image labeling rules are as follows: the area where the microorganism is located is marked as white as foreground, and the rest is marked as black as the background.

2.2. Experimental Method and Setup

To better demonstrate the functions of EMDS-6, we carry out noise addition and denoising experiments, image segmentation experiments, image feature extraction experiments, image classification experiments and object detection experiments. The experimental methods and data settings are shown below. Moreover, we select different critical indexes to evaluate each experimental result in this section.

2.2.1. Noise Addition and Denoising Method

In digital image processing, the quality of an image to be recognized is often affected by external conditions, such as input equipment and the environment. Noise generated by external environmental influences largely affects image processing and analysis (e.g., image edge detection, classification, and segmentation). Therefore, image denoising is the key step of image preprocessing (Zhang et al., 2022).

In this study, we have used four types of noise, Poisson noise, multiplicative noise, Gaussian noise and pretzel noise. By adjusting the mean, variance and density of different kinds of noise, a total of 13 specific noises are generated. They are multiplicative noise with a variance of 0.2 and 0.04 (marked as MN:0.2 and MN: 0.04 in the table), salt and pepper noise with a density of 0.01 and 0.03 (SPN:0.01, SPN:0.03), pepper noise (PpN), salt noise (SN), Brightness Gaussian noise (BGN), Positional Gaussian noise (PGN), Gaussian noise with a variance of 0.01 and a mean of 0 (GN 0.01–0), Gaussian noise with a variance of 0.01 and a mean of 0.5 (GN 0.01–0.5), Gaussian noise with a variance of 0.03 and a mean of 0 (GN 0.03–0), Gaussian noise with a variance of 0.03 and a mean of 0.5 (GN 0.03–0.5), and Poisson noise (PN). There are 9 kinds of filters at the same time, namely Two-Dimensional Rank Order Filter (TROF), 3×3 Wiener Filter [WF (3×3)], 5×5 Wiener Filter [WF (5×5)], 3×3 Window Mean Filter [MF (3×3)], Mean Filter with 5×5 Window [MF (5×5)], Minimum Filtering (MinF), Maximum Filtering (MaxF), Geometric Mean Filtering (GMF), Arithmetic Mean Filtering (AMF). In the experiment, 13 kinds of noise are added to the EMDS-6 dataset image, and then 9 kinds of filters are used for filtering. The result of adding noise into the image and filtering is shown in **Figure 2**.

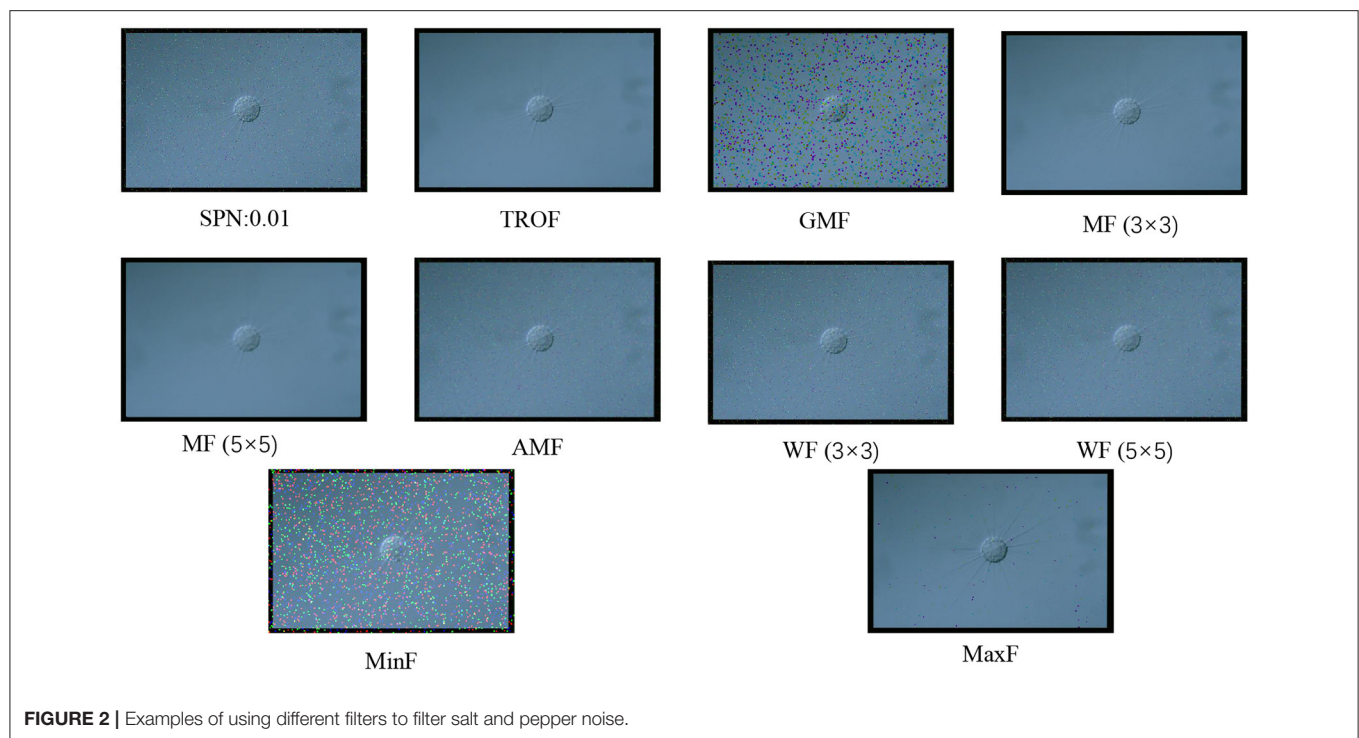


FIGURE 2 | Examples of using different filters to filter salt and pepper noise.

2.2.2. Image Segmentation Methods

This article designs the following experiment to prove that EMDS-6 can be used to test different image segmentation methods (Zhang et al., 2021). Six classic segmentation methods are used in the experiment: *k*-means (Burney and Tariq, 2014), Markov Random Field (MRF) (Kato and Zerubia, 2012), Otsu Thresholding (Otsu, 1979), Region Growing (REG) (Adams and Bischof, 1994), Region Split and Merge Algorithm (RSMA) (Chen et al., 1991) and Watershed Segmentation (Levner and Zhang, 2007) and one deep learning-based segmentation method, Recurrent Residual CNN-based U-Net (U-Net) (Alom et al., 2019) are used in this experiment. While using U-Net for segmentation, the learning rate of the network is 0.001 and the batch size is 1. In the *k*-means algorithm, the value of *k* is set to 3, the initial center is chosen randomly, and the iterations are stopped when the number of iterations exceeds the maximum number of iterations. In the MRF algorithm, the number of classifications is set to 2 and the maximum number of iterations is 60. In the Otsu algorithm, the BlockSize is set to 3, and the average value is obtained by averaging. In the region growth algorithm, we use a 8-neighborhood growth setting.

Among the seven classical segmentation methods, *k*-means is based on clustering, which is a region-based technology. Watershed algorithm is based on geomorphological analysis such as mountains and basins to implement different object segmentation algorithms. MRF is an image segmentation algorithm based on statistics. Its main features are fewer model parameters and strong spatial constraints. Otsu Thresholding is an algorithm based on global binarization, which can realize adaptive thresholds. The REG segmentation algorithm starts from a certain pixel and gradually adds neighboring pixels

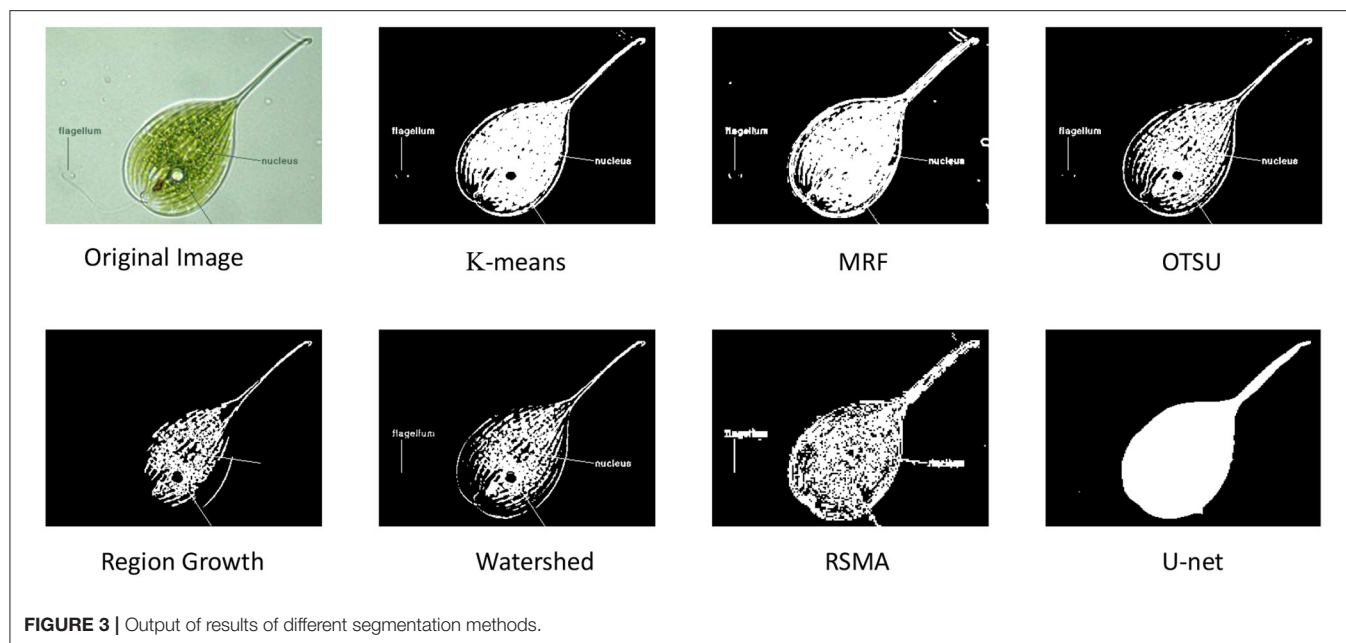
according to certain criteria. When certain conditions are met, the regional growth is terminated, and object extraction is achieved. The RSMA is first to determine a split and merge criterion. When splitting to the point of no further division, the areas with similar characteristics are integrated. **Figure 3** shows a sample of the results of different segmentation methods on EMDS-6.

2.2.3. Image Feature Extraction Methods

This article uses 10 methods for feature extraction (Li et al., 2015), including two-color features, One is HSV (Hue, Saturation, and Value) feature (Junhua and Jing, 2012), and the other is RGB (Red, Green, and Blue) color histogram feature (Kavitha and Suruliandi, 2016). The three texture features include the Local Binary Pattern (LBP) (Ojala et al., 2002), the Histogram of Oriented Gradient (HOG) (Dalal and Triggs, 2005) and the Gray-level Co-occurrence Matrix (GLCM) (Qunqun et al., 2013) formed by the recurrence of pixel gray Matrix. The four geometric features (Geo) (Mingqiang et al., 2008) include perimeter, area, long-axis and short-axis and seven invariant moment features (Hu) (Hu, 1962). The perimeter, area, long-axis and short-axis features are extracted from the GT image, while the rest are extracted from the original image. Finally, we use a support vector machine (SVM) to classify the extracted features. The classifier parameters are shown in **Table 2**.

2.2.4. Image Classification Methods

In this article, we design the following two experiments to test whether the EMDS-6 dataset can compare the performance of different classifiers (Li et al., 2019; Zhao et al., 2022). Experiment 1: use traditional machine learning methods to

**TABLE 2 |** Parameter setting of EMDS-6 feature classification using SVM.

Feature	Kernel	C	DFS	Tol	Max iter
LBP	rbf	50,000	ovr	1e-3	-1
GLCM	rbf	10,000	ovr	1e-3	-1
HOG	rbf	1,000	ovr	1e-3	-1
HSV	rbf	100	ovr	1e-3	-1
Geo	rbf	2,000,000	ovr	1e-3	-1
Hu	rbf	100,000	ovr	1e-3	-1
RGB	rbf	20	ovr	1e-3	-1

C, penalty coefficient; DFS, decision function shape; tol, the error value of stopping training; Geo, geometric features.

TABLE 3 | Deep learning model parameters.

Parameter	Parameter
Batch size, 32	Epoch, 100
Learning, 0.002	Optimizer, Adam

classify images. This chapter uses Geo features to verify the classifier's performance. Moreover, traditional classifiers used for testing includes, three k -Nearest Neighbor (k NN) classifiers ($k = 1, 5, 10$) (Abeywickrama et al., 2016)], three Random Forests (RF) (tree = 10, 20, 30) (Ho, 1995) and four SVMs (kernel function = rbf, polynomial, sigmoid, linear) (Chandra and Bedi, 2021). The SVM parameters are set as follows: penalty parameter $C = 1.0$, the maximum number of iterations is unlimited, the size of the error value for stopping training is 0.001, and the rest of the parameters are default values.

In Experiment 2, we use deep learning-based methods to classify images. Meanwhile, 21 classifiers are used to evaluate

TABLE 4 | Evaluation metrics of segmentation method.

Indicators	Formula
Dice	$\frac{2 \times V_{pred} \cap V_{gt} }{ V_{pred} + V_{gt} }$
Jaccard	$\frac{ V_{pred} \cap V_{gt} }{ V_{pred} \cup V_{gt} }$
Recall	$\frac{TP}{TP + FN}$

TP, True Positive; FN, False Negative; V_{pred} , the foreground predicted by the model; V_{gt} , the foreground in a GT image.

TABLE 5 | Classifier classification performance evaluation index.

Evaluation indicators	Formula
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	$\frac{TP}{TP+FP}$
F1-score	$2 \times \frac{P \times R}{P+R}$
Recall	$\frac{TP}{TP+FN}$

the performance, including, ResNet-18, ResNet-34, ResNet-50, ResNet-101 (He et al., 2016), VGG-11, VGG-13, VGG-16, VGG-19 (Simonyan and Zisserman, 2014), DenseNet-121, DenseNet-169 (Huang et al., 2017), Inception-V3 (Szegedy et al., 2016), Xception (Chollet, 2017), AlexNet (Krizhevsky et al., 2012), GoogleNet (Szegedy et al., 2015), MobileNet-V2 (Sandler et al., 2018), ShuffleNetV2 (Ma et al., 2018), Inception-ResNet - V1 (Szegedy et al., 2017), and a series of ViTs, such as ViT (Dosovitskiy et al., 2020), BotNet (Srinivas et al., 2021), DeiT (Touvron et al., 2020), T2T-ViT (Yuan et al., 2021). The above models are set with uniform hyperparameters, as detailed in **Table 3**.

TABLE 6 | Similarity comparison between denoised image and original image.

ToN / DM	TROF	MF: (3 × 3)	MF: (5 × 5)	WF: (3 × 3)	WF: (5 × 5)	MaxF	MinF	GMF	AMF
PN	98.36	98.24	98.00	98.32	98.15	91.97	99.73	99.21	98.11
MN:0.2	99.02	90.29	89.45	91.98	91.08	71.15	99.02	98.89	90.65
MN:0.04	99.51	99.51	99.51	95.57	95.06	82.35	99.51	98.78	94.92
GN 0.01-0	96.79	96.45	96.13	96.75	96.40	85.01	99.44	98.93	96.28
GN 0.01-0.5	98.60	98.52	98.35	98.97	98.81	96.32	99.67	64.35	98.73
GN 0.03-0	94.64	93.99	93.56	94.71	94.71	76.46	99.05	98.74	93.82
GN 0.03-0.5	97.11	96.95	96.66	98.09	97.79	94.04	99.24	66.15	97.54
SPN:0.01	99.28	99.38	99.14	99.60	99.37	95.66	99.71	99.44	99.16
SPN:0.03	98.71	98.57	98.57	99.29	98.87	92.28	99.24	99.26	98.80
PpN	98.45	98.53	98.30	99.46	99.02	96.30	99.04	99.61	98.61
BGN	97.93	97.74	97.74	97.91	97.69	90.00	99.66	99.16	97.60
PGN	96.97	96.63	96.33	97.16	96.85	85.82	99.47	98.98	96.47
SN	97.90	97.97	97.75	99.27	98.63	99.27	98.63	99.64	98.15

ToN, types of noise; DM, denoising method. (In [%]).

TABLE 7 | Comparison of variance between denoised image and original image.

ToN / DM	TROF	MF: (3 × 3)	MF: (5 × 5)	WF: (3 × 3)	WF: (5 × 5)	MaxF	MinF	GMF	AMF
PN	1.49	0.77	1.05	0.52	0.66	3.68	2.99	0.41	0.88
MN,v: 0.2	32.49	14.94	15.65	9.33	11.36	39.22	32.49	4.32	13.35
MN,v: 0.04	10.89	10.89	10.89	2.99	3.71	14.41	10.89	0.98	4.28
GN,m: 0,v: 0.01	3.81	3.06	3.44	2.06	2.62	11.68	7.36	1.16	3.00
GN,m: 0.5,v: 0.01	0.89	0.36	0.41	0.21	0.28	0.99	1.74	61.93	0.43
GN,m: 0,v: 0.03	8.60	7.78	8.34	5.04	5.04	27.23	16.55	4.24	7.33
GN,m: 0.5,v: 0.03	1.60	1.08	1.18	0.55	0.73	2.39	3.06	56.17	1.05
SPN,d: 0.01	1.92	1.21	1.46	0.10	0.30	6.37	2.90	4.73	1.25
SPN,d: 0.03	3.84	3.39	3.39	0.33	1.09	14.64	5.18	13.02	3.15
PpN	2.88	2.18	2.44	0.17	0.72	3.72	4.48	16.84	2.09
BGN	2.35	1.63	1.94	1.09	1.38	6.67	4.57	0.84	1.66
PGN	3.79	3.04	3.42	1.67	2.13	11.56	7.33	1.23	2.98
SN	3.86	3.17	3.44	0.31	1.35	4.82	6.25	5.58	2.94

(In [%]).

2.2.5. Object Detection Method

In this article, we use Faster RCNN (Ren et al., 2015) and Mask RCNN (He et al., 2017) to test the feasibility of the EMDS-6 dataset for object detection (Li C. et al., 2021). Faster RCNN provide excellent performance in many areas of object detection. The Mask RCNN is optimized on the original framework of Faster RCNN. By using a better skeleton (ResNet combined with FPN) and the AlignPooling algorithm, Mask RCNN achieves better detection results than Faster RCNN.

In this experiment, the learning rate is 0.0001, the model Backbone is ResNet50, and the batch size is 2. In addition, we used 25% of the EMDS-6 data as training, 25% is for validation, and the rest is for testing.

2.3. Evaluation Methods

2.3.1. Evaluation Method for Image Denoising

This article uses mean-variance and similarity indicators to evaluate filter performance. The similarity evaluation index can

be expressed as 1, where i represents the original image, i_1 represents the denoised image, N represents the number of pixels, and A represents the similarity between the denoised image and the original image. When the value of A is closer to 1, the similarity between the original image and the denoised image is higher, and the denoising effect is significant.

$$A = 1 - \frac{\sum_{i=1}^N |i_1 - i|}{N \times 255} \quad (1)$$

The variance evaluation index can be expressed as Equation (2), where S denotes the mean-variance, $L_{(i,j)}$ represents the value corresponding to the coordinates of the original image (i, j) , and $B_{(i,j)}$ the value associated with the coordinates of the denoised image (i, j) . When the value of S is closer to 0, the higher the similarity between the original and denoised images, the better the denoising stability.

TABLE 8 | Evaluation of feature extraction methods using EMDS-6 dataset.

Method/Index	Dice	Jaccard	Recal
k-means	47.78	31.38	32.11
MRF	56.23	44.43	69.94
Otsu	45.23	33.82	40.60
REG	29.72	21.17	26.94
RSMA	37.35	26.38	30.18
Watershed	44.21	32.44	40.75
U-Net	88.35	81.09	89.67

(In [%]).

TABLE 9 | Different results obtained by applying different features in the EMDS-6 classification experiments using SVM.

FT	LBP	GLCM	HOG
Acc	32.38	10.24	22.98
HSV	Geo	Hu	RGB
29.52	50.0	7.86	28.81

FT, Feature type; Acc, Accuracy. (In [%]).

$$S = 1 - \frac{\sum_{i=1}^n (L_{(ij)} - B_{(ij)})^2}{\sum_{i=1}^n L_{(ij)}^2} \quad (2)$$

2.3.2. Evaluation Method for Image Segmentation

We use segmented images and GT images to calculate Dice, Jaccard and Recall evaluation indexes. Among the three evaluation metrics, the Dice coefficient is pixel-level, and the Dice coefficient takes a range of 0-1. The more close to 1, the better the structure of the model. The Jaccard coefficient is often used to compare the similarity between two samples. When the Jaccard coefficient is larger, the similarity between the samples is higher. The recall is a measure of coverage, mainly for the accuracy of positive sample prediction. The computational expressions of Dice, Jaccard, and Recall are shown in **Table 4**.

2.3.3. Evaluation Index of Image Feature Extraction

Image features can be used to distinguish image classes. However, the performance of features is limited by the feature extraction method. In this article, we select ten classical feature extraction methods. Meanwhile, the classification accuracy of SVM is used to evaluate the feature performance. The higher the classification accuracy of SVM, the better the feature performance.

2.3.4. Evaluation Method for Image Classification

In Experiment 1 of Section 2.2.4, we use only the accuracy index to judge the performance of traditional machine learning classifiers. The higher the number of EMs that can be correctly classified, the better the performance of this classifier. In Experiment 2, the performance of deep learning models needs to be considered in several dimensions. In order to more accurately evaluate the performance of different deep learning models, we introduce new evaluation indicators. The evaluation indexes and the calculation method of the indexes are shown in **Table 5**. In

Table 5, TP means the number of EMs classified as positive and also labeled as positive. TN means the number of EMs classified as negative and also labeled as negative. FP means the number of EMs classified as positive but labeled as negative. FN means the number of EMs classified as negative but labeled as positive.

2.3.5. Evaluation Method for Object Detection

In this article, Average Precision (AP) and Mean Average Precision (mAP) are used to evaluate the object detection results. AP is a model evaluation index widely used in object detection. The higher the AP, the fewer detection errors. AP calculation method is shown in Equations 3 and 4.

$$AP = \sum_{n=1}^N (r_{n+1} - r_n) Pinter(r_{n+1}) \quad (3)$$

$$Pinter(r_{n+1}) = \max_{\hat{r}=r_{n+1}} P(\hat{r}) \quad (4)$$

Among them, r_n represents the value of the nth recall, and $p(\hat{r})$ represents the value of precision when the recall is \hat{r} .

3. EXPERIMENTAL RESULTS AND ANALYSIS

3.1. Experimental Results Analysis of Image Denoising

We calculate the filtering effect of different filters for different noises. Their similarity evaluation indexes are shown in **Table 6**. From **Table 6**, it is easy to see that the GMF has a poor filtering effect for GN 0.01-0.5. The TROF and the MF have better filtering effects for MN:0.04.

In addition, the mean-variance is a common index to evaluate the stability of the denoising method. In this article, the variance of the EMDS-6 denoised EM images and the original EM images are calculated as shown in **Table 7**. As the noise density increases, the variance significantly increases among the denoised and the original images. For example, by increasing the SPN density from 0.01 to 0.03, the variance increases significantly under different filters. This indicates that the result after denoising is not very stable.

From the above experiments, EMDS-6 can test and evaluate the performance of image denoising methods well. Therefore, EMDS-6 can provide strong data support for EM image denoising research.

3.2. Experimental Result Analysis of Image Segmentation

The experimental results of the seven different image segmentation methods are shown in **Table 8**. In **Table 8**, the REG and RSMA have poor segmentation performance, and their Dice, Jaccard, and Recall indexes are much lower than other segmentation methods. However, the deep learning-based, U-Net, has provided superior performance. By comparing these image segmentation methods, it can be concluded that EMDS-6 can provide strong data support for testing and assessing image segmentation methods.

TABLE 10 | Results of experiments to classify Geo features using traditional classifiers.

Classifier type	SVM: linear	SVM: polynomial	SVM: RBF	SVM: sigmoid	RF,nT: 30
Accuracy	51.67	27.86	28.81	14.29	98.33
kNN,k: 1	kNN,k: 5	kNN,k: 10	RF,nT: 10	RF,nT: 20	–
23.1	17.86	17.38	96.19	97.86	–

(In [%]).

TABLE 11 | Classification results of different deep learning models.

Model	Precision (%)	Recall (%)	F1-score (%)	Acc (%)	PS (MB)	Time (S)
Xception	44.29	45.36	42.40	44.29	79.8	1,079
ResNet34	40.00	43.29	39.43	40.00	81.3	862
Googlenet	37.62	40.93	35.49	37.62	21.6	845
Densenet121	35.71	46.09	36.22	35.71	27.1	1,002
Densenet169	40.00	40.04	39.16	40.00	48.7	1,060
ResNet18	39.05	44.71	39.94	39.05	42.7	822
Inception-V3	35.24	37.41	34.14	35.24	83.5	973
Mobilenet-V2	33.33	38.43	33.97	33.33	8.82	848
InceptionResnetV1	35.71	38.75	35.32	35.71	30.9	878
Deit	36.19	41.36	36.23	36.19	21.1	847
ResNet50	35.71	38.58	35.80	35.71	90.1	967
ViT	32.86	37.66	32.47	32.86	31.2	788
ResNet101	35.71	38.98	35.52	35.71	162	1,101
T2T-ViT	30.48	32.22	29.57	30.48	15.5	863
ShuffleNet-V2	23.33	24.65	22.80	23.33	1.52	790
AlexNet	32.86	34.72	31.17	32.86	217	789
VGG11	30.00	31.46	29.18	30.00	491	958
BotNet	28.57	31.23	28.08	28.57	72.2	971
VGG13	5.24	1.82	1.63	5.24	492	1,023
VGG16	4.76	0.23	0.44	4.76	512	1,074
VGG19	4.76	0.23	0.44	4.76	532	1,119

Acc, Accuracy; PS, Params size.

TABLE 12 | AP and mAP based on EMDS-6 object detection of different types of EMs.

Model\sample (AP)	Actinophrys	Arcella	Aspidisca	Codosiga	Colpoda	Epistylis	Euglypha	Paramecium
Faster RCNN	0.95	0.75	0.39	0.13	0.52	0.24	0.68	0.70
Mask RCNN	0.70	0.85	0.40	0.18	0.35	0.53	0.25	0.70
Model\sample	Rotifera	Vorticella	Noctiluca	Ceratium	Stentor	Siprostomum	K.Quadrula	Euglena
Faster RCNN	0.69	0.30	0.56	0.61	0.47	0.60	0.22	0.37
Mask RCNN	0.40	0.15	0.90	0.70	0.65	0.7	0.45	0.25
Model\sample	Gymnodinium	Gonyaulax	Phacus	Stylongchia	Synchaeta	mAP	–	–
Faster RCNN	0.53	0.25	0.43	0.42	0.61	0.50	–	–
Mask RCNN	0.60	0.28	0.50	0.68	0.48	0.51	–	–

3.3. Experimental Result Analysis of Feature Extraction

In this article, we use the SVM to classify different features. The classification results are shown in **Table 9**. The Hu features performed poorly, while the Geo features performed the best. In addition, the classification accuracy of FT, LBP, GLCM, HOG, HSV and RGB features are also very different. By comparing these classification results, we can conclude that EMDS-6 can be used to evaluate image features.

3.4. Experimental Result Analysis of Image Classification

This article shows the traditional machine learning classification results in **Table 10**, and the deep learning classification results are shown in **Table 11**. In **Table 10**, the RF classifier performs the best. However, the performance of the SVM classifier using the sigmoid kernel function is relatively poor. In addition, there is a big difference in Accuracy between other classical classifiers. From the computational results, the EMDS-6 dataset is able

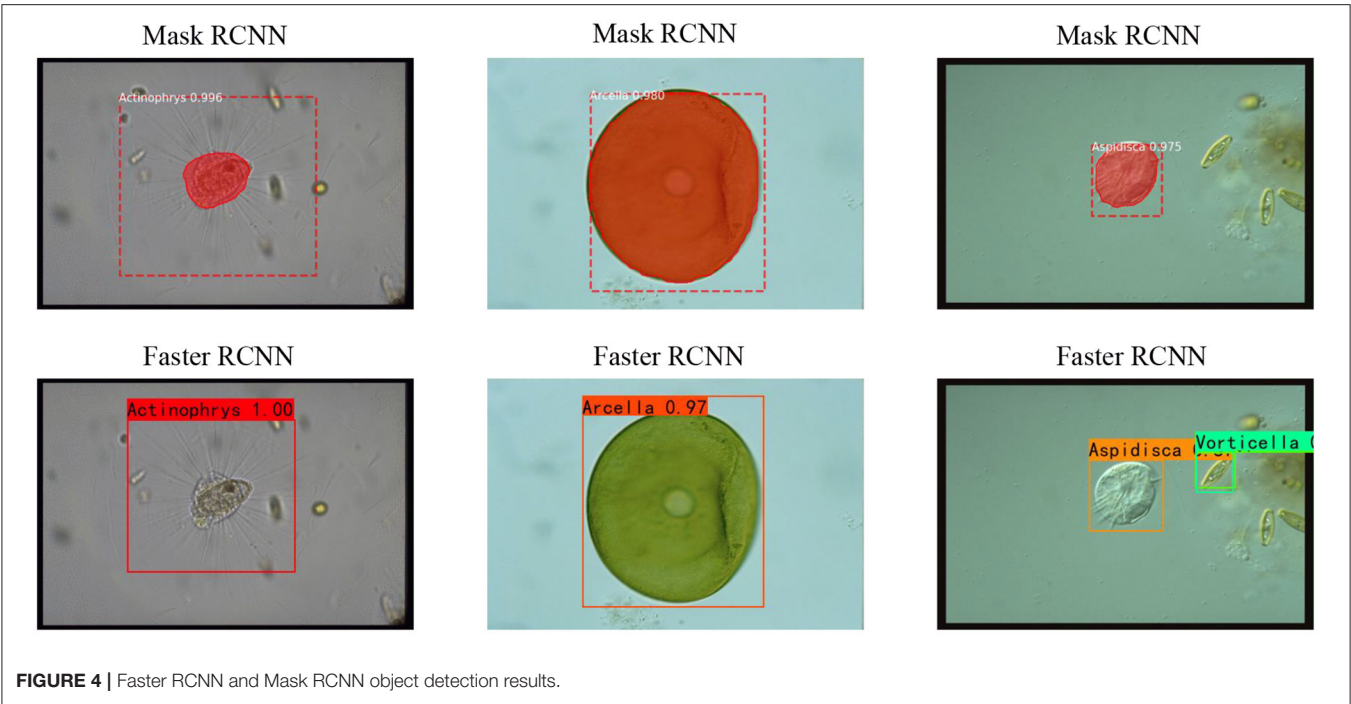


FIGURE 4 | Faster RCNN and Mask RCNN object detection results.

TABLE 13 | EMDS history versions and latest versions.

Dataset	ECN	OIN	GTIN	Dataset link	Functions
EMDS-1 (Li et al., 2013)	10	200	200	- -	IC, IS
EMDS-2 (Li et al., 2013)	10	200	200	- -	IC ,IS
EMDS-3 (Li et al., 2016)	15	300	300	- -	IC, IS
EMDS-4 (Zou et al., 2016)	21	420	420	https://research.project-10.de/em-classification/	IC, IS, IR
EMDS-5 (Li Z. et al., 2021)	21	420	840 (S 420, M 420)	https://github.com/NEUZihan/EMDS-5	ID, IED, SoIS, MoIS, SoFE, MoFE, IR
EMDS-6 [In this article]	21	840	840	https://figshare.com/articles/dataset/EMDS6/17125025/1	ID, IC, IS, IFE, IOD

IC, Image Classification; IS, Image Segmentation; SoIS, Single-object Image Segmentation; MoIS, Multi-object Image Segmentation; SoFE, Single-object Feature Extraction; MoFE, Multi-object Feature Extraction; IR, Image Retrieval; IFE, Image Feature Extraction; IOD, Image Object Detection; IED, Image Edge Detection; ID, Image denoising; ECN, EM Class Number; OIN, Original Image Number; GTIN, Ground Truth Image Number; S, Single Object; M, Multiple object.

to provide data support for classifier performance evaluation. According to **Table 11**, the classification accuracy of Xception is 44.29%, which is the highest among all models. The training of deep learning models usually consumes much time, but some models have a significant advantage in training time. Among the selected models, ViT consumes the shortest time in training samples. The training time of the ViT model is the least. The classification performance of the ShuffleNet-V2 network is average, but the number of parameters is the least. Therefore, experiments prove that EMDS-6 can be used for the performance evaluation of deep learning classifiers.

3.5. Experimental Result Analysis of Image Object Detection

The AP and mAP indicators for Faster CNN and Mast CNN are shown in **Table 12**. We can see from **Table 12** that Faster

RCNN and Mask RCNN have very different object detection effects based on their AP value. Among them, the Faster RCNN model has the best effect on Actinophrys object detection. The Mask RCNN model has the best effect on Arcella object detection. Based on the mAP value, it is seen that Faster RCNN is better than Mask RCNN for object detection. The result of object detection is shown in **Figure 4**. Most of the EMs in the picture can be accurately marked. Therefore it is demonstrated that the EMDS-6 dataset can be effectively applied to image object detection.

3.6. Discussion

As shown in **Table 13**, six versions of the EMs dataset are published. In the iteration of versions, different EMSs assume different functions. Both EMDS-1 and EMDS-2 have similar functions and can perform image classification and segmentation. In addition, both EMDS-1 and EMDS-2 contain

ten classes of EMs, 20 images of each class, with GT images. Compared with the previous version, EMDS-3 does not add new functions. However, we expand five categories of EMs.

We open-source EMDSs from EMDS-4 to the latest version of EMDS-6. Compared to EMDS-3, EMDS-4 expands six additional classes of EMs and adds a new image retrieval function. In EMDS-5, 420 single object GT images and 420 multiple object GT images are prepared, respectively. Therefore EMDS-5 supports more functions as shown in **Table 13**. The dataset in this article is EMDS-6, which is the latest version in this series. EMDS-6 has a larger data volume compared to EMDS-5. EMDS-6 adds 420 original images and 420 multiple object GT images, which doubles the number of images in the dataset. With the support of more data volume, EMDS-6 can achieve more functions in a better and more stable way. For example, image classification, image segmentation, object and object detection.

4. CONCLUSION AND FUTURE WORK

This article develops an EM image dataset, namely EMDS-6. EMDS-6 contains 21 types of EMs and a total of 1680 images. Including 840 original images and 840 GT images of the same size. Each type of EMs has 40 original images and 40 GT images. In the test, 13 kinds of noises such as multiplicative noise and salt and pepper noise are used, and nine kinds of filters such as Wiener filter and geometric mean filter are used to test the denoising effect of various noises. The experimental results prove that EMDS-6 has the function of testing the filter denoising effect. In addition, this article uses 6 traditional segmentation algorithms such as *k*-means and MRF and one deep learning algorithm to compare the performance of the segmentation algorithm. The experimental results prove that EMDS-6 can effectively test the image segmentation effect. At the same time, in the image feature extraction and evaluation experiment, this article uses 10 features such as HSV and RGB extracted from EMDS-6. Meanwhile, the SVM classifier is used to test the features. It is found that the classification results of different features are significantly different, and EMDS-6 has the function of testing the pros and cons of features. In terms of image classification, this article designs two experiments. The

first experiment uses three classic machine learning methods to test the classification performance. The second experiment uses 21 deep learning models. At the same time, indicators such as accuracy and training time are calculated to verify the performance of the model from multiple dimensions. The results show that EMDS-6 can effectively test the image classification performance. In terms of object detection, this article tests Faster RCNN and Mask RCNN, respectively. Most of the EMs in the experiment can be accurately marked. Therefore, EMDS-6 can be effectively applied to image object detection.

In the future, we will further expand the number of EM images of EMDS-6. At the same time, we will try to apply EMDS-6 to more computer vision processing fields to further promote microbial research development.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

PZ: experiment, result analysis, and article writing. CL: data preparation, method, result analysis, article writing, proofreading, and funding support. MR and NX: proofreading. HX and HY: experiment. PM: data treatment. HS: environmental microorganism knowledge support. TJ: result analysis and funding support. MG: method and result analysis. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (No.61806047).

ACKNOWLEDGMENTS

We thank Miss Zixian Li and Mr. Guoxian Li for their important discussion.

REFERENCES

- Abeywickrama, T., Cheema, M. A., and Taniar, D. (2016). K-nearest neighbors on road networks: a journey in experimentation and in-memory implementation. *arXiv preprint arXiv:1601.01549*. doi: 10.14778/2904121.2904125
- Adams, R., and Bischof, L. (1994). Seeded region growing. *IEEE Trans Pattern Anal. Mach. Intell.* 16, 641–647.
- Alom, M. Z., Yakopcic, C., Hasan, M., Taha, T. M., and Asari, V. K. (2019). Recurrent residual U-Net for medical image segmentation. *J. Med. Imaging* 6, 014006. doi: 10.1117/1.JMI.6.1.014006
- Buades, A., Coll, B., and Morel, J.-M. (2005). A review of image denoising algorithms, with a new one. *Multiscale Model. Simul.* 4, 490–530. doi: 10.1137/040616024
- Burney, S. M. A., and Tariq, H. (2014). K-means cluster analysis for image segmentation. *Int. J. Comput. App.* 96, 1–8.
- Chandra, M. A., and Bedi, S. S. (2021). Survey on svm and their application in image classification. *Int. J. Inf. Technol.* 13, 1–11. doi: 10.1007/s41870-017-0080-1
- Chen, S.-Y., Lin, W.-C., and Chen, C.-T. (1991). Split-and-merge image segmentation based on localized feature analysis and statistical tests. *CVGIP Graph. Models Image Process.* 53, 457–475.
- Chollet, F. (2017). “Xception: deep learning with depthwise separable convolutions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 1251–1258.
- Dalal, N., and Triggs, B. (2005). “Histograms of oriented gradients for human detection,” in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)* (San Diego, CA: IEEE), 886–893.
- 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.53yu.com/abs/2010.11929>

- He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). "Mask r-CNN," in *Proceedings of the IEEE International Conference on Computer Vision* (Honolulu, HI), 2961–2969.
- 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, NV), 770–778.
- Ho, T. K. (1995). "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition* (Montreal, QC: IEEE), 278–282.
- Hu, M.-K. (1962). Visual pattern recognition by moment invariants. *IRE Trans. Inform. Theory* 8, 179–187.
- 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, HI), 4700–4708.
- Junhua, C., and Jing, L. (2012). "Research on color image classification based on HSV color space," in *2012 Second International Conference on Instrumentation, Measurement, Computer, Communication and Control* (Harbin: IEEE), 944–947.
- Kato, Z., and Zerubia, J. (2012). *Markov Random Fields in Image Segmentation*. Hanover, MA: NOW Publishers.
- Kavitha, J., and Suruliandi, A. (2016). "Texture and color feature extraction for classification of melanoma using SVM," in *2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16)* (Kovilpatti: IEEE), 1–6.
- Kosov, S., Shirahama, K., Li, C., and Grzegorzec, M. (2018). Environmental microorganism classification using conditional random fields and deep convolutional neural networks. *Pattern Recogn.* 77, 248–261. doi: 10.1016/j.patcog.2017.12.021
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Adv. Neural Inform. Process. Syst.* 25, 1097–1105.
- Kulwa, F., Li, C., Zhao, X., Cai, B., Xu, N., Qi, S., et al. (2019). A state-of-the-art survey for microorganism image segmentation methods and future potential. *IEEE Access*. 7, 100243–100269.
- Levner, I., and Zhang, H. (2007). Classification-driven watershed segmentation. *IEEE Trans. Image Process.* 16, 1437–1445. doi: 10.1109/TIP.2007.894239
- Li, C., Ma, P., Rahaman, M. M., Yao, Y., Zhang, J., Zou, S., et al. (2021). A state-of-the-art survey of object detection techniques in microorganism image analysis: from traditional image processing and classical machine learning to current deep convolutional neural networks and potential visual transformers. *arXiv [Preprint]*. arXiv: 2105.03148. Available online at: <https://arxiv.org/abs/2105.03148>
- Li, C., Shirahama, K., and Grzegorzec, M. (2015). Application of content-based image analysis to environmental microorganism classification. *Biocybern. Biomed. Eng.* 35, 10–21. doi: 10.1016/j.bbe.2014.07.003
- Li, C., Shirahama, K., and Grzegorzec, M. (2016). Environmental microbiology aided by content-based image analysis. *Pattern Anal. Appl.* 19, 531–547. doi: 10.1007/s10044-015-0498-7
- Li, C., Shirahama, K., Grzegorzec, M., Ma, F., and Zhou, B. (2013). "Classification of environmental microorganisms in microscopic images using shape features and support vector machines," in *2013 IEEE International Conference on Image Processing* (Melbourne, VIC: IEEE), 2435–2439.
- Li, C., Wang, K., and Xu, N. (2019). A survey for the applications of content-based microscopic image analysis in microorganism classification domains. *Artif. Intell. Rev.* 51, 577–646.
- Li, Z., Li, C., Yao, Y., Zhang, J., Rahaman, M. M., Xu, H., et al. (2021). EMDS-5: Environmental microorganism image dataset fifth version for multiple image analysis tasks. *PLoS ONE* 16, e0250631. doi: 10.1371/journal.pone.0250631
- Ma, N., Zhang, X., Zheng, H.-T., and Sun, J. (2018). "Shufflenet v2: practical guidelines for efficient cnn architecture design," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Salt Lake City, UT), 116–131.
- Madigan, M. T., Martinko, J. M., Parker, J., et al. (1997). *Brock Biology of Microorganisms*, Vol. 11. Upper Saddle River, NJ: Prentice Hall.
- Madsen, E. L. (2008). *Environmental Microbiology: From Genomes to Biogeochemistry*. Oxford: Wiley-Blackwell.
- Mingqiang, Y., Kidiyo, K., and Joseph, R. (2008). A survey of shape feature extraction techniques. *Pattern Recognit.* 15, 43–90. doi: 10.5772/6237
- Ojala, T., Pietikainen, M., and Maenpää, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* 24, 971–987. doi: 10.1109/TPAMI.2002.1017623
- Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybernet.* 9, 62–66.
- Pitas, I. (2000). *Digital Image Processing Algorithms and Applications*. Hoboken, NJ: Wiley.
- Qunqun, H., Fei, W., and Li, Y. (2013). Extraction of color image texture feature based on gray-level co-occurrence matrix. *Remote Sens. Land Resour.* 25, 26–32. doi: 10.6046/gtzyyq.2013.04.05
- Rahaman, M. M., Li, C., Yao, Y., Kulwa, F., Rahman, M. A., Wang, Q., et al. (2020). Identification of covid-19 samples from chest x-ray images using deep learning: A comparison of transfer learning approaches. *J. Xray Sci. Technol.* 28, 821–839. doi: 10.3233/XST-200715
- Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-CNN: towards real-time object detection with region proposal networks. *Adv. Neural Inform. Process. Syst.* 28, 91–99. doi: 10.1109/TPAMI.2016.2577031
- Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). "MobileNetV2: inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT), 4510–4520.
- 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.53yu.com/abs/1409.1556>
- Srinivas, A., Lin, T.-Y., Parmar, N., Shlens, J., Abbeel, P., and Vaswani, A. (2021). Bottleneck transformers for visual recognition. *arXiv preprint arXiv:2101.11605*. Available online at: <https://arxiv.org/abs/2101.11605>
- Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. (2017). "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI Conference on Artificial Intelligence* (San Francisco, CA).
- 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.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 2818–2826.
- Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. (2020). Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*. Available online at: <https://arxiv.org/abs/2012.12877>
- Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Tay, F. E., et al. (2021). Tokens-to-token vit: training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986*. Available online at: <https://arxiv.53yu.com/abs/2101.11986>
- Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., and Saeed, J. (2020). A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. *J. Appl. Sci. Technol. Trends* 1, 56–70. doi: 10.38094/jastt1224
- Zhang, J., Li, C., Kosov, S., Grzegorzec, M., Shirahama, K., Jiang, T., et al. (2021). Lcunet: A novel low-cost u-net for environmental microorganism image segmentation. *Pattern Recognit.* 115, 107885. doi: 10.1016/j.patcog.2021.107885
- Zhang, J., Li, C., Rahaman, M., Yao, Y., Ma, P., Zhang, J., et al. (2022). A comprehensive review of image analysis methods for microorganism counting: from classical image processing to deep learning approaches. *Artif. Intell. Rev.* 55, 2875–2944. doi: 10.1007/s10462-021-10082-4
- Zhao, P., Li, C., Rahaman, M., Xu, H., Yang, H., Sun, H., et al. (2022). A comparative study of deep learning classification methods on a small environmental microorganism image dataset (emds-6): From convolutional neural networks to visual transformers. *arXiv [Preprint]*. arXiv: 2107.07699. Available online at: <https://arxiv.org/pdf/2107.07699.pdf>
- Zou, Y. L., Li, C., Boukhers, Z., Shirahama, K., Jiang, T., and Grzegorzec, M. (2016). "Environmental microbiological content-based image retrieval system using

internal structure histogram,” in *Proceedings of the 9th International Conference on Computer Recognition Systems*, 543–552.

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.

Copyright © 2022 Zhao, Li, Rahaman, Xu, Ma, Yang, Sun, Jiang, Xu and Grzegorzek. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Interfacing Machine Learning and Microbial Omics: A Promising Means to Address Environmental Challenges

James M. W. R. McElhinney^{1*}, Mary Krystelle Catacutan², Aurelie Mawart¹, Ayesha Hasan^{1,2} and Jorge Dias³

¹ Applied Genomics Laboratory, Center for Membranes and Advanced Water Technology, Khalifa University, Abu Dhabi, United Arab Emirates, ² Department of Biomedical Engineering, Khalifa University, Abu Dhabi, United Arab Emirates, ³ EECS, Center for Autonomous Robotic Systems, Khalifa University, Abu Dhabi, United Arab Emirates

OPEN ACCESS

Edited by:

George Tsiamis,
University of Patras, Greece

Reviewed by:

Felipe Hernandes Coutinho,
Institute of Marine Sciences (CSIC),
Spain

*Correspondence:

James M. W. R. McElhinney
james.mcelhinney@ku.ac.ae

Specialty section:

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

Received: 09 January 2022

Accepted: 14 March 2022

Published: 25 April 2022

Citation:

McElhinney JMW, Catacutan MK, Mawart A, Hasan A and Dias J (2022) Interfacing Machine Learning and Microbial Omics: A Promising Means to Address Environmental Challenges. *Front. Microbiol.* 13:851450. doi: 10.3389/fmicb.2022.851450

Microbial communities are ubiquitous and carry an exceptionally broad metabolic capability. Upon environmental perturbation, microbes are also amongst the first natural responsive elements with perturbation-specific cues and markers. These communities are thereby uniquely positioned to inform on the status of environmental conditions. The advent of microbial omics has led to an unprecedented volume of complex microbiological data sets. Importantly, these data sets are rich in biological information with potential for predictive environmental classification and forecasting. However, the patterns in this information are often hidden amongst the inherent complexity of the data. There has been a continued rise in the development and adoption of machine learning (ML) and deep learning architectures for solving research challenges of this sort. Indeed, the interface between molecular microbial ecology and artificial intelligence (AI) appears to show considerable potential for significantly advancing environmental monitoring and management practices through their application. Here, we provide a primer for ML, highlight the notion of retaining biological sample information for supervised ML, discuss workflow considerations, and review the state of the art of the exciting, yet nascent, interdisciplinary field of ML-driven microbial ecology. Current limitations in this sphere of research are also addressed to frame a forward-looking perspective toward the realization of what we anticipate will become a pivotal toolkit for addressing environmental monitoring and management challenges in the years ahead.

Keywords: machine learning, microbial ecology, metagenomics, environmental monitoring, microbiology, artificial intelligence, microbial omics, predictive modeling

INTRODUCTION

Expansion of the human population is increasing resource consumption and discharge of waste products, placing significant burdens on the biosphere (Burrell et al., 2020; Grantham et al., 2020; Lv et al., 2020; Albert et al., 2021; Lu et al., 2021; Naumann et al., 2021; Ortiz-Bobea et al., 2021). These activities are contributing to the multifaceted pollution of the global ecological systems

(Julinová et al., 2018; Santos et al., 2019; Turan et al., 2019; Vardhan et al., 2019; Briffa et al., 2020; Pulster et al., 2020; Simul Bhuyan et al., 2021; Sohrabi et al., 2021; Li and Fantke, 2022). Consequently, we are witnessing an accelerating loss of biodiversity, habitats, and climate change (Sintayehu, 2018; Brühl and Zaller, 2019). Gauging and forecasting such anthropogenic environmental impacts is often limited in scope due to scale-up challenges. At large scale, this endeavor remains an inordinately complex and resource-intensive task and therefore represents a major scientific goal.

At 93 gigatons carbon (Gt C), microbial communities comprise approximately 20% of the total estimated global biomass and exclusively form the deep subsurface biome (estimated at 70 Gt C) (Bar-On et al., 2018). These communities are ubiquitously distributed across the biosphere where their activities are central in shaping the environments of our planet (Gibbons and Gilbert, 2015); microbial communities possess exceptionally broad metabolic capabilities, enabling their utilization of many xenobiotics (Katsuyama et al., 2009; Junghare et al., 2019). Microbes can have short generation times and are amongst the first responders with perturbation-specific cues and markers (De Anda et al., 2018; Astudillo-García et al., 2019) these can therefore serve as a valuable source of biological information for establishing the status of their respective environmental niches and can serve as dynamic biosensors for monitoring and tracing environmental changes (Cesare et al., 2020; Morimura et al., 2020).

Omics methodologies enable rapid community-wide profiling of microbial populations across environmental perturbations. Omics data are information-rich, leading to an unprecedented volume of large multidimensional data sets with potential for predictive environmental classification and forecasting. However, the inherent complexity in these data conceals the patterns underlying the biological information, challenging manual curation and interpretation. Machine learning (ML) is well suited to address such challenges and there has been a sharp rise in their application in health-oriented microbiomics (Zeller et al., 2014; Szafranski et al., 2015; Knight et al., 2018). ML-driven omics is now being applied to address environmental challenges (Figure 1). Here, we will discuss the state of the art in this interdisciplinary field and highlight considerations, ongoing limitations, and challenges for future work. The interface between ML and molecular microbial ecology (MME) holds great promise for significantly advancing environmental monitoring and management practices. Indeed, ML will likely become a routine toolkit for the molecular microbiologist and will be essential to manage large multidimensional environmental omics data.

MAIN BODY

A Primer on Machine Learning

Machine learning approaches can be supervised (SML) or unsupervised (USML). In SML methods, data sets are reduced/converted into the sets of features which serve as the input and form a variable for the SML model. Features are measurable and informative properties of the data, e.g., taxa

abundances, annotated with metadata of interest (labels) which define the desired output (the target). Feature sets are subset into groups for model training and model testing/validation for SML learning. The SML architecture then attempts to derive a model that can predict the label for new input data. SML can be carried out to address regression or classification challenges. For regression, the SML tool predicts values for a continuous series (such as levels of environmental pollutants). For classification, the SML will predict the conditional label pertaining to the sample (such as contamination status). Deep learning (DL) is a subset of SML, which employs neural networks with multiple (>3) processing layers and has the highest capacity for learning. For USML, no label or target output is defined; instead the USML architecture establishes patterns in the data naively, usually by clustering or ordination projections. USML is particularly useful for exploratory analysis of microbial omics data and includes ordination methods that are commonly applied in microbiology. Here we focus primarily on SML applications for environmentally centered microbial omics research. For more details on the underlying principles of ML for microbial ecology, readers are encouraged to see reviews (Ghannam and Techtman, 2021; Goodswen et al., 2021).

Omics Data Sets Are Rich in Learnable Biological Information

Anthropogenic perturbations give rise to spatiotemporal patterns in microbial communities by influencing the following: abundances, interactions between, and dispersal of community members (Blaser et al., 2016; Liao et al., 2018). Community dynamics are perturbation-specific, reproducible, and predictable, affecting taxonomic diversity, differential abundances in taxa, functional gene clusters, and shifts in metabolic circuits which influence microbial interactions (Figure 1). Microbial omics approaches are rapidly advancing our views of these complex shifts and have opened myriad avenues for the utilization of microbial data to address environmental challenges. Often these omics approaches scrutinize a single systems level (e.g., DNA or RNA), but can synergistically provide more information when integrated with supporting omics data from other systems layers (Franzosa et al., 2015). Such integrative omics represents a powerful means to understand communities through cross-systems-level descriptions but is in its infancy and yet to be much applied in this area. A central challenge for any ML-led omics analyses is the preservation of the biological information hidden within the microbial community, throughout the workflow (Figure 1), to allow for effective learning. There are numerous ways *via* which the biological information in omics samples can be compromised. These pitfalls occur at virtually all decision points in the omics workflow and begin with the experimental design phase. The significance of a given pitfall is highly dependent on the phenomena under investigation and aims of the study but common pitfalls include inadequate sampling, improper preservation, sample transport conditions or subcommunity sampling (e.g., planktonic/sessile), biases arising from sample handling (e.g., during extraction and amplification), the choice of

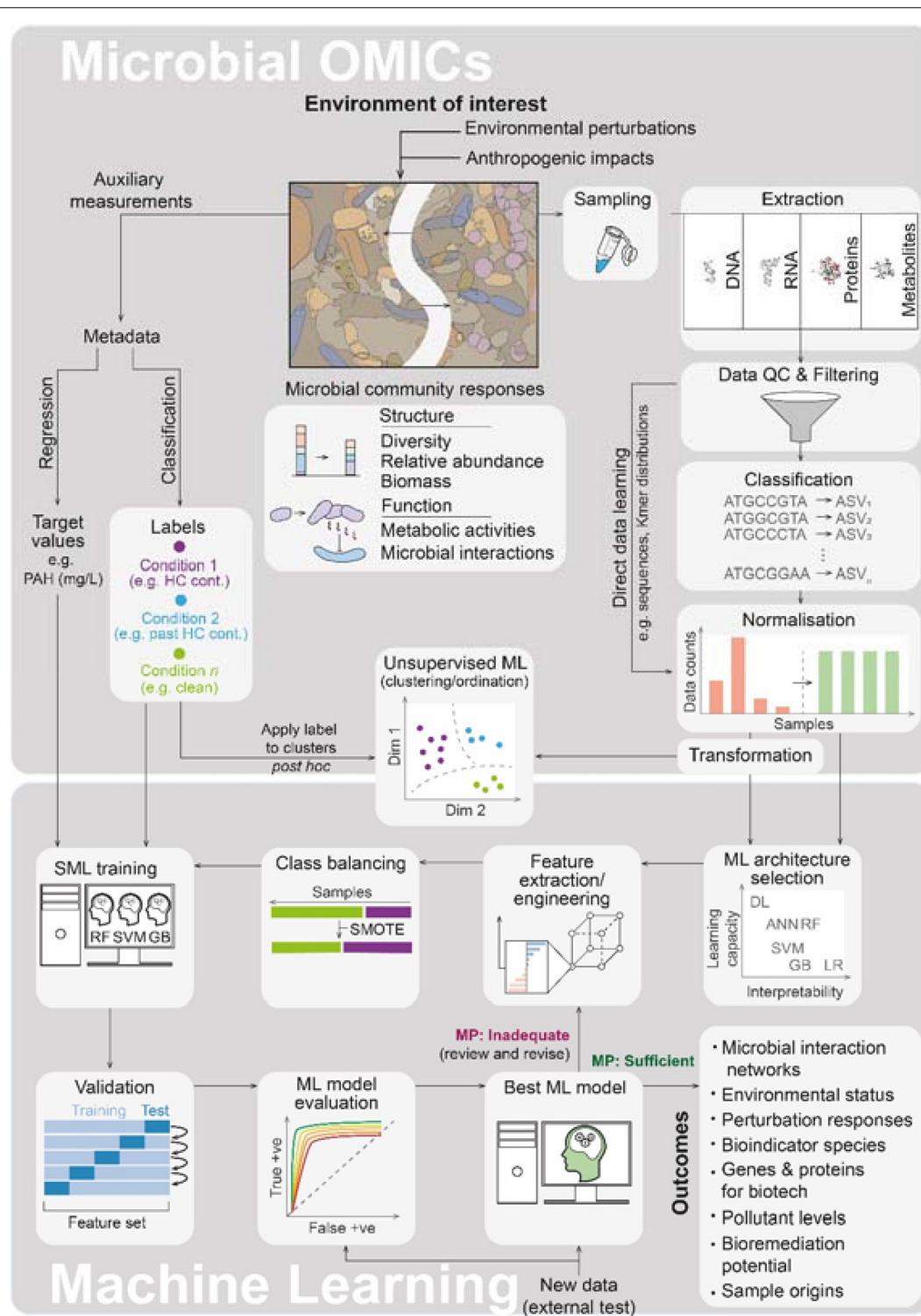


FIGURE 1 | The interface of microbial omics and machine learning (ML). A generalized and simplified overview of the workflows is presented highlighting the major steps in the microbial omics and ML workflows as they relate to one another along with key outcomes obtainable from the application of ML to omics data. Microbial community responses (biological information on which learning is aimed) are summarized below the cartoon snapshot of a contaminated environment of interest. Here, HC cont., hydrocarbon contamination; PAH, polycyclic aromatic hydrocarbons (as examples of targets in petroleum hydrocarbon scenarios); QC, quality control; ASV, amplicon sequence variant (ASVs are given here as an example of an omics classification, other examples include the often used OTU, genes, mRNA transcripts, protein categories or metabolite IDs); DL, deep learning; ANN, artificial neural networks (shallow); RF, random forest; SVM, support vector machine; GB, gradient boost; LR, logistic regression; SMOTE, synthetic minority oversampling technique; SML, supervised machine learning; and MP, model performance.

sequencing/liquid chromatography-mass spectrometry (LC-MS) platform and analytical methodology, classification and filtering of omics data (which can remove rare but important taxa, transcripts, or proteins), artifacts from data transformation and normalization approaches (correcting for library size is especially essential for meta-analyses), and the choice and engineering of features. A number of considerations can help in preserving the biological information for omics-led SML, and many are discussed in the following.

Workflow Considerations

Microbial Omics Input

Microbial omics pitfalls, from sampling to the bioinformatics pipeline, can reduce or bias the information yielded (Gutleben et al., 2018; Kaster and Sobol, 2020). Typically, some trade-off must be made in the experimental design, for which options have been suggested (Franzosa et al., 2015). In metataxonomics, resolution is usually limited to the genus level, though it is the most commonly used omics input for SML (Table 1), wherein relative operational taxonomic unit (OTU) abundances form the feature set (Miao et al., 2020; Janßen et al., 2021; Kim and Oh, 2021). However, the use of OTUs is inherently limiting for retaining community information and can miss important taxonomic groups. Indeed, since the development of the more biologically meaningful amplicon sequence variants (ASVs; Callahan et al., 2017), the absence of ASVs in most metataxonomic studies is striking. As ASVs represent a more accurate basis for taxa assignment, it will be interesting to see how their application influences ML performances in future.

Metagenomics is highly sensitive for low-abundance taxa, but is rarely applied for SML and carries additional costs which may limit sampling and options for ML (Chen and Tyler, 2020). Importantly, metagenomic approaches do not always convey a clear advantage over the more cost-effective metataxonomic approach (Xu et al., 2014). The choice between metataxonomics and metagenomics is evidently not clear-cut and should be considered in light of the expected community under study, choice of sequencing platform, and research goals. Microbial omics inputs are most often derived from closed-reference databases, leading to inevitable loss of learnable biological information in environmental samples due to unclassified/misclassified data (Chen and Tyler, 2020). However, the development of ML and DL tools (Liang et al., 2020) for enhancing taxonomic classification in metagenomic data sets could prove helpful. Alternatively, the direct use of biological sequences (from microbial omics surveys) circumvents this issue (by forgoing categorical assignment), thereby permitting the inclusion of more comprehensive feature spaces, at the cost of reducing the immediate interpretability for the user. Informative abstractions of omics data, such as the use of K-mer distributions as a feature set, have shown success in both taxonomic (Fiannaca et al., 2018) subtyping (Solis-Reyes et al., 2018) and phenotypic (Aun et al., 2018) classification, and are applicable to environmental applications. Indeed, K-mer abstractions have shown predictive potential for classifying sample environment and host-phenotype (an

environmental status) that excels over OTU features (Asgari et al., 2018). Environmental metatranscriptomics-led SML is currently limited. However, the approach has been shown to uncover the mixotrophic processes of protists in response to nutrient gradients in the Pacific Ocean (Lambert et al., 2021), thereby demonstrating that trophic modes can be readily predicted from metatranscriptomic data.

Choice of Machine Learning Architecture

There is a broad selection of the SML tools to select from and each carries its own advantages and limitations (Goodswen et al., 2021). Not a single architecture performs best in all environmental application cases and users must make a trade-off in terms of interpretability, learning performance, computational costs, data requirements, and ease of implementation (Ghannam and Techtmann, 2021). At the outset, selecting a set of architectures can help to ensure the delivery of research goals. Random forest (RF) is a popular choice for microbial omics-driven SML for its learning capacity, straightforward implementation, and high degree of interpretability (Ghannam and Techtmann, 2021). For especially complex tasks, or where knowledge is limited, DL approaches (multi-layered architectures) have the highest performance, as they can self-learn (i.e., do not require user extraction of) the feature set (Christin et al., 2019). However, DL comes with elevated computational costs and low interpretability of the underlying model (“black box” effect) and requires large volumes of data (thousands of samples). Consequently, though very promising, DL approaches for environmental omics are currently limited.

Feature Engineering

Feature selection and engineering are crucial for generating meaningful SML-based ecological models. Reducing the feature space can help to limit overfitting, reduce computational costs, improve cross-study comparison, and improve generalized prediction performance across data sets (Ghannam and Techtmann, 2021). However, care is needed when reducing features for training as biologically meaningful features can be missed if feature selection is based on abundance. This is especially so when assessing anthropogenic perturbations of pollutants in the environment, wherein the rare microbiome (taxa representing <0.1% of the total community) comprise a significant reservoir of gene clusters that enable the utilization and degradation of xenobiotic organic compounds (Wang et al., 2017). Taking embedded approaches for feature selection (that can evaluate across the full feature space) (Wang et al., 2017) or a biologically driven feature selection method (such as taxonomically aware hierarchical feature engineering) (Oudah and Henschel, 2018) may help in optimizing feature selection in metataxonomics-driven ML applications. Feature selection methods designed for functional feature sets are still notably lacking in this space.

Conventional statistics require assumptions on the underlying data and care is needed, given the compositional nature of microbial omics data sets (Gloor et al., 2017). For example, conventional ecological models often assume monotonicity in relationships, which can hinder ecological explanations

TABLE 1 | Example applications of the SML of microbial Omics data for addressing environmental challenges.

Environment	Niche	Application	Omics	Input data	Feature	Target(s)	SML architectures	Software	References
Aquatic	Marine (Coral Reef)	Prediction of environmental status	metataxonomics	16S rRNA OTUs	OTU abundance	Eutrophication indicators and temperature	RF	Caret and RF R packages	Glasl et al., 2019
Industrial	WWTP	Prediction of environmental variable to identify key subpopulations	metataxonomics	16S rRNA OTUs	OTU abundance, PCA coordinates	WWTP water temperature	LR, RF, SVML, DT, KNN, SVMRBF	Scikit-Learn	Kim and Oh, 2021
Terrestrial	Soil ¹	Prediction of carbon cycling	metataxonomics	16S rRNA OTUs	OTU abundance	[DOC]	RF, ANN	THEANO, Scikit-Learn	Thompson et al., 2019
Terrestrial	Compost	Classification of microbial biomarkers	metataxonomics	16S rRNA OTUs	OTU abundance	Compost cycle	RF	RF R package	Zhang et al., 2020
Terrestrial	Ground water + Soil ¹	Prediction of environmental contaminants	metataxonomics	16S rRNA OTUs	OTU abundance	[dioxane] and [CVOCs]	RF		Miao et al., 2020
Terrestrial	Soil	Prediction of environmental quality	metataxonomics	16S rRNA OTUs	OTU abundance	Soil physicochemical features	RF	RF R package	Hermans et al., 2020
Aquatic	Marine (coastal waters) ¹	Prediction of environmental contaminants	metataxonomics	16S rRNA OTUs	OTU abundance, 16S rRNA gene sequences	Glyphosate	RF, ANN	RF R package and DL4J	Janßen et al., 2019
Aquatic	Freshwater (river)	Classification of anthropogenic pathogen loads	metataxonomics ²	16S rRNA OTUs	OTU abundance	Fecal source	RF, MCMC	RF R package and SourceTracker	Dubinsky et al., 2016
Aquatic	Marine and Freshwater	Classification of microbial biomarkers	metataxonomics	16S rRNA and ITS OTUs	OTU abundance	Plastisphere communities	RF	RF R package	Li et al., 2021
Aquatic	Marine sediment (munitions dumpsite)	Prediction of environmental contaminants	metataxonomics	16S rRNA OTUs	OTU abundance	TNT	RF, ANN	Ranger R package ANN R keras framework + TensorFlow back end	Janßen et al., 2021
Aquatic	Freshwater (river)	Classification of sample origin	metataxonomics	16S rRNA OTUs	OTU abundance (top taxa)	Sample origin	RF	RF R package	Wang et al., 2021
Aquatic	Marine (oceanic waters)	Classification of trophic modes	Metatranscriptomics	Gene expression levels	expression levels of selected Pfam entries	Trophic mode (photo/hetero/mixo)	RF, DT, ANN	NR and XGBoost	Lambert et al., 2021
Terrestrial	Soil	Prediction of crop productivity	metagenomics	Shotgun sequencing	OTU abundance	Crop productivity	RF	Ranger R package	Chang et al., 2017
Terrestrial	Soil	Prediction of soil phylogroups from environmental metadata	metagenomics	NR	NR	<i>Listeria</i> species	RF	RF R package	Liao et al., 2021

¹Indirectly studied in microcosms.²Using PhyloChip array.

Here, ANN, Artificial Neural Network; CVOCs, Chlorinated Volatile Organic Compounds; DOC, Dissolved Organic Carbon; DT, Decision Tree; KNN, K-Nearest Neighbors; LR, Logistic Regression; MCMC, Markov Chain Monte Carlo; NR, Not reported; RF, Random Forest; SVML, Support Vector Machine (SVM) with a linear kernel; SVMRBF, SVM with a radial basis function kernel; TNT, trinitrotoluene; WWTP, Wastewater Treatment Plant.

of community variance across study sites. By applying SML (allowing for non-monotonic feature capture), the ability to capture this variance can increase nine-fold (Fontaine et al., 2021). It is important to note that the goal of SML should not be to replace classical statistical modeling, but rather to complement it. Integrating these two approaches presents an promising opportunity to leverage their advantages for predictive environmental microbiology (Lopatkin and Collins, 2020) and monitoring. For multi-omics studies, feature selection and engineering becomes increasingly complex with the successive systems levels, and there is much to be done in this area. In such studies, functional data across systems levels will likely need to be empirically assessed prior to SML to identify the most informative biomarkers for learning (Xu et al., 2014).

Evaluating Data Leakage

Data leakage is a subtle but important aspect of ML, referring to the unintended use or influence of data (that should not be available at the time of prediction) during the training process. This often occurs when the features used for training hide within themselves the result of the prediction, resulting in an overestimation of performance of the model during validation (Chiavegatto Filho et al., 2021). Due to the subtleties with which this can occur, avoiding data leakage is challenging and should be evaluated on a case by case basis. Important aspects for consideration here have been discussed previously (Wirbel et al., 2021) and include (1) data filtering that is influenced by the target label and (2) the splitting of dependent data (e.g., replicates and time-series data points) across training and validation sets. The use of an externally generated test data set (handled separately from the training set) for additional validation checks can help (Oyetunde et al., 2019; Wirbel et al., 2021), though data leakage is seldom discussed in microbial omics papers that use SML. We urge future authors in this space to consider including at least a statement on leakage assessment in studies based on SML.

Applications of Molecular Microbial Ecology–Machine Learning for Environmental Challenges

Microbes as Environmental Biosensors

Anthropogenic impacts are motivating the development of cost-effective and scalable environmental bioassessment methodologies (Fruehe et al., 2021). Microbes have long been recognized as potential *in situ* biosensors for following human impacts (Su et al., 2011), allowing for highly accurate quantitative SML predictions of the perturbation. Indeed, metataxonomic data can be valuable for the prediction of a variety of environmental contaminants (Table 1), spanning from relatively inert plastics (Li et al., 2021) to petroleum hydrocarbons [which illicit strong responses with detectable influences even after the pollutant is degraded and undetectable by conventional measures (Smith et al., 2015)]. Hydrocarbonoclastic indicator species have also been identified as key biosensors in ML-based bioprospecting of hydrocarbon seepage from subsurface reservoirs and can improve the likelihood of success in drilling for new assets (de Dios Miranda et al., 2019; Chitu et al., 2022). The same approach is also being explored as the potential

early-warning indicators of leakage from hydrocarbon transport lines (Shaheen et al., 2011). Indeed, the SML of microbial fingerprints has even demonstrated reasonable predictions (accuracies of 72–85%) of the future production of hydrocarbon reservoirs (using metataxonomic input) (Zijp et al., 2021) which can facilitate decision-making for enhanced asset management. These approaches thereby have real potential for reducing the carbon footprint and ecological impact of upstream oil and gas activities.

Microbes as Predictors of Environmental Status

Microbes have proved valuable as ecological assessment indicators in multiple diverse environments (Astudillo-García et al., 2019; Glasl et al., 2019; Hermans et al., 2020; Chen et al., 2021). Moreover, improvements in sequencing technologies are facilitating the upscaling and deployment of omics-based ML for more ambitious environmental monitoring and mitigation applications (Wang et al., 2021). These indicators can reveal important relationships for land management, when conventional field measurements are unhelpful (Chang et al., 2017). Indeed, the SML of microbial 16S rRNA abundances can directly predict soil productivity in arable land and risks posed for agriculture (Yuan et al., 2020). USML is routinely applied *via* ordination techniques to establish the organization of microbiome data in relation to their environmental parameters. However, in instances where conventional ordinations fail to determine clear relationships, SML may still yield community subpopulations that can serve as predictors for environmental parameters and processes of interest. For example, the influence between temperature and key phosphate and glycogen-accumulating organisms involved in the enhanced biological phosphorous removal processes of a set of wastewater treatment plants (WWTPs) in South Korea was identified using an SML approach, resulting in findings with clear implications for WWTP design and operation (Oh and Kim, 2021). Additionally, the SML of metabarcoded environmental DNA (eDNA) can provide superior performance for environmental quality monitoring over conventional bioindicator values for marine aquaculture monitoring (Fruehe et al., 2021). Furthermore, RF learning of eDNA has been shown to outperform conventional taxonomy-based biotic indices assessments (Cordier et al., 2018). Biodiversity in microbial communities can also be a useful proxy to assess the environmental impact of anthropogenic perturbations through changes in biotic indices (Aylagas et al., 2017). In these ways, SML is a useful means to improve environmental monitoring programs.

Predicting Sample Origin With Microbiological Data

The predictive power of ML for monitoring environmental status also enables sample origin to be established (Raza et al., 2021). Microbial metrics have proved to be exceptionally sensitive indicators of human impacts on freshwater environments (Liao et al., 2018). Indeed, *via* ML modeling, the partitioning of microbes along complex anthropogenic xenobiotic gradients from urban and agricultural runoffs is sufficient to identify the origin of water samples from the 30 most abundant taxa (Wang et al., 2021) and is able to resolve sample origin depth and local salinity in the Baltic Sea (Alneberg et al., 2020).

Such origin tracing carries the potential to inform for public health by accurately predicting the origins of fecal contaminants in public waters (Chen et al., 2021; Raza et al., 2021) and the source of food-borne pathogen outbreaks (Wheeler, 2019). The ability to identify sample origin sources is likely to be of critical importance moving forward for tracing runoffs from agricultural and industrial entities to ensure compliance with environmentally mindful legislation. It will be interesting to see whether this sort of tracing application will lend itself to following waterbodies in other settings, or indeed, other mobile elements within the environment (forensic analysis of migratory animals under conservation management, for example). Given the perceived stability in the gut microbiome, it is possible that this approach could also be extended as a biological tagging approach for following animal populations at the center of conservation efforts.

Supporting Environmental Meta-Analyses and Data Mining

The high volumes of omics data are enabling large-scale meta-analyses (Zeller et al., 2014) that can provide a global view of microbial roles within major environments (Ramirez et al., 2018; Wu et al., 2019; Yuan et al., 2020). However, several challenges arise in such studies owing to non-standardized sample collection, extraction methods, and primer choice (Ramirez et al., 2018). Additionally, technicalities of sequencing platforms, variable library sizes, and environmental confounders can reduce concordance across omics studies (though SML is alleviating this issue) (York, 2021). ML tools are well suited for uncovering patterns within these challenging data collections. For example, a meta-analysis of soil microbiomes with SML was able to reveal microbiological indicators for predicting propensity for *Fusarium* wilt (Yuan et al., 2020), an agriculturally important pest. Additionally, a meta-analysis of global soil (Ramirez et al., 2018) and WWTP (Wu et al., 2019) communities provided macroecological insights into the microbial biogeography communities and confirmed the importance of the rare microbiome members as bioindicators. There remains significant scope for standardizing the workflows in both omics and SML. Such standardizations are crucial to mitigating common pitfalls; these enhance reproducibility and promote meta-analyses and data mining. An important limiting factor here is that many data sets are unavailable, uploaded to repositories without raw data or lacking metadata descriptions. This issue has been raised before (Ramirez et al., 2018) and impedes otherwise valuable work. For instance, bioprospecting of biosynthetic gene clusters with SML-based omics data mining can yield proteins with biotechnological potential (Correia and Weimann, 2021) for bioremediation, biodegradable plastic production, and sustainable biofuels (Haque et al., 2020; Keasling et al., 2021). We therefore urge that omics data sets be uploaded in their raw form with metadata made available.

Supervised Machine Learning of Microbial Omics Data to Address Climate Change

The collective effects of anthropogenic perturbations are driving the consequences of climate change (notably, losses

of ecosystem function, services, biodiversity, and habitat) at unprecedented rates (Giuliani et al., 2017). The actions of microbial communities are implicitly tied to geochemical cycling, global water chemistries, nutrient availabilities, and soil/plant health (Gorbushina and Krumbein, 2000; Falkowski et al., 2008; Lian et al., 2008; Dong, 2010; Panke-Buisse et al., 2014). Microbes are thereby drivers of numerous ecosystem services on which the global population relies (Marco and Abram, 2019). Understanding microbe–ecosystem interactions and functions is therefore central to their utilization in ecological models and biotechnologies for intervening on climate change. The generation of high-resolution spatiotemporal dynamics data and incorporation of different omics data sets can provide important insights into the molecular mechanisms behind climate changes responses and improve the accuracy of forecasting models (Herold et al., 2020; Layton and Bradbury, 2021). Together with their ubiquitous nature, the core roles of microbial communities afford us with a broad framework for potential microbiological tools with which the fundamental impacts of global climate change can be understood, monitored, predicted, and conceivably, mitigated. The short generation times of microbial community members and their predictable changes following changing environmental parameters (Larsen et al., 2012) open the possibility for their use as early-warning indicators of climate change-led impacts on macroecological networks (Shah et al., 2022) before further biodiversity loss is observable on the macroscale. Conversely, microbial contributions to climate change *via* carbon cycle–climate feedback and N₂O production (Bardgett et al., 2008) are an ideal candidate for predictive SML modeling and intervention. Indeed, predictive models from microbial omics data have also shown utility across a range of climate change-linked phenomena, including browning (Fontaine et al., 2021), eutrophication (Glasl et al., 2019), harmful algal blooms (Hennon and Dyhrman, 2020), and arability of soils (Chang et al., 2017; Hennon and Dyhrman, 2020; Yuan et al., 2020). omics in soil–plant, subsurface, and aquatic microbiomes is also central to making inroads in the development of carbon capture and sequestration (CCS) biotechnologies (Schweitzer et al., 2021). It will be interesting to see whether such developments benefit from SML-based modeling, which could prove useful for establishing taxa and metabolisms that predict stability and sequestration rates in CCS systems. Therefore, SML modeling can facilitate the establishment and optimization of carbon fluxes in microbial communities (particularly for the poorly characterized deep subsurface microbiome) and may also help to bridge bioenergy production to CCS, which is considered essential for many climate change mitigation plans (Hanssen et al., 2020). At present, the ability of microbes to inform on, and forecast, climate change impacts *via* ecological monitoring programs is perhaps the most immediately applicable area for the SML of microbial omics in climate change research. In this way, microbes can assist decision-makers for sustainable policies and intervention measures to ensure food security and maintain ecosystem services before further ecological detriment occurs (Cordier et al., 2021; Shah et al., 2022). The potential future applications in this space, however, are vast and may be key for realizing goals in

global-scale climate management and engineering against climate change.

CONCLUDING REMARKS AND FUTURE PERSPECTIVES

Machine learning is a powerful toolbox for drawing meaningful biological insights from large multidimensional microbial data. Here, we discussed how SML can contribute to environmental challenges by valorizing microbial community data sets. The predictive potential of interfacing omics and SML has opened exciting new avenues for managing environmental pollution and status. The ability to identify key species and functional elements can be expected to accelerate biotechnological developments with implications for environmental intervention (such as bioremediation). Through the interface of these important disciplines, we are rapidly advancing our view of global microbiome and the ecological impacts from human activities.

This nascent, but fast-evolving, application area for ML has several notable opportunities which are yet to be exploited. Metataxonomics-centric ML efforts have dominated this space, but has yet to apply long-read and metagenome-assembled genomic data for feature set development in this research area. Additionally, several advanced systems-level techniques (metaproteomics, metabolomics, and in particular, integrative omics) remain at much earlier stages of development compared with DNA sequencing-based approaches and are consequently lagging in this arena. ML tools will likely become integral to pipelines for these advanced omics methodologies. We foresee SML becoming a routine complement to conventional statistics and expect that this will key for revealing the often-overlooked

rare microbiome. As omics approaches continue to advance, and sample costs reduce, we can expect to see a rise in the application of promising DL architectures at this interdisciplinary interface. DL tools will no doubt prove indispensable in data mining the ever-increasing public omics repositories and represent an exciting means to address feature engineering challenges *via* unsupervised feature extractions.

AUTHOR CONTRIBUTIONS

JM: structure of manuscript, figure design and production, literature review, manuscript writing, population of table, and revisions. MC: initial draft of manuscript, figure design, and literature review. AM: literature review, population of table, figure design, and development of content. AH: structure of the manuscript, secured funding, manuscript review, and development of content. JD: conceptualize the manuscript, manuscript review, and development of content. All authors contributed to the article and approved the submitted version.

FUNDING

This work was funded by the Competitive Internal Research Award (CIRA2019-019) of Khalifa University.

ACKNOWLEDGMENTS

We would like to acknowledge valuable discussions on this topic with Olivier Monga and Andreas Henschel.

REFERENCES

- Albert, J. S., Destouni, G., Duke-Sylvester, S. M., Magurran, A. E., Oberdorff, T., Reis, R. E., et al. (2021). Scientists' warning to humanity on the freshwater biodiversity crisis. *Ambio* 50, 85–94. doi: 10.1007/s13280-020-01318-8
- Alneberg, J., Bennke, C., Beier, S., Bunse, C., Quince, C., Ininbergs, K., et al. (2020). Ecosystem-wide metagenomic binning enables prediction of ecological niches from genomes. *Comm. Biol.* 3:119. doi: 10.1038/s42003-020-0856-x
- Asgari, E., Garakani, K., McHardy, A. C., and Mofrad, M. R. K. (2018). MicroPheno: predicting environments and host phenotypes from 16S rRNA gene sequencing using a k-mer based representation of shallow sub-samples. *Bioinformatics* 34, i32–i42. doi: 10.1093/bioinformatics/bty296
- Astudillo-García, C., Hermans, S. M., Stevenson, B., Buckley, H. L., and Lear, G. (2019). Microbial assemblages and bioindicators as proxies for ecosystem health status: potential and limitations. *Appl. Microbiol. Biotechnol.* 103, 6407–6421. doi: 10.1007/s00253-019-09963-0
- Aun, E., Brauer, A., Kisand, V., Tenson, T., and Remm, M. (2018). A k-mer-based method for the identification of phenotype-associated genomic biomarkers and predicting phenotypes of sequenced bacteria. *PLoS Comput. Biol.* 14:e1006434. doi: 10.1371/journal.pcbi.1006434
- Aylagas, E., Borja, Á., Tangherlini, M., Dell'Anno, A., Corinaldesi, C., Michell, C. T., et al. (2017). A bacterial community-based index to assess the ecological status of estuarine and coastal environments. *Mar. Poll. Bull.* 114, 679–688. doi: 10.1016/j.marpolbul.2016.10.050
- Bardgett, R. D., Freeman, C., and Ostle, N. J. (2008). Microbial contributions to climate change through carbon cycle feedbacks. *ISME J.* 2, 805–814. doi: 10.1038/ismej.2008.58
- Bar-On, Y. M., Phillips, R., and Milo, R. (2018). The biomass distribution on Earth. *Proc. Natl. Acad. Sci.* 115, 6506–6511. doi: 10.1073/pnas.1711842115
- Blaser, M. J., Cardon, Z. G., Cho, M. K., Dangl, J. L., Donohue, T. J., Green, J. L., et al. (2016). Toward a Predictive Understanding of Earth's Microbiomes to Address 21st Century Challenges. *mBio* 7:e00714-16. doi: 10.1128/mBio.00714-16
- Briffa, J., Sinagra, E., and Blundell, R. (2020). Heavy metal pollution in the environment and their toxicological effects on humans. *Heliyon* 6:e04691. doi: 10.1016/j.heliyon.2020.e04691
- Brühl, C. A., and Zaller, J. G. (2019). Biodiversity Decline as a Consequence of an Inappropriate Environmental Risk Assessment of Pesticides. *Front. Environ. Sci.* 7:177. doi: 10.3389/fenvs.2019.00177
- Burrell, A. L., Evans, J. P., and De Kauwe, M. G. (2020). Anthropogenic climate change has driven over 5 million km² of drylands towards desertification. *Nat. Commun.* 11:3853. doi: 10.1038/s41467-020-17710-7
- Callahan, B. J., McMurdie, P. J., and Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* 11, 2639–2643. doi: 10.1038/ismej.2017.119
- Cesare, A. Di, Pjevac, P., Eckert, E., Curkov, N., Miko Šparica, M., Corno, G., et al. (2020). The role of metal contamination in shaping microbial communities in heavily polluted marine sediments. *Environ. Poll.* 265:114823. doi: 10.1016/j.envpol.2020.114823

- Chang, H.-X., Haudenschild, J. S., Bowen, C. R., and Hartman, G. L. (2017). Metagenome-Wide Association Study and Machine Learning Prediction of Bulk Soil Microbiome and Crop Productivity. *Front. Microbiol.* 8:519. doi: 10.3389/fmicb.2017.00519
- Chen, F., Koh, X. P., Tang, M. L. Y., Gan, J., and Lau, S. C. K. (2021). Microbiological assessment of ecological status in the Pearl River Estuary. *Chin. Ecol. Indic.* 130:108084. doi: 10.1016/j.ecolind.2021.108084
- Chen, J.-C.-y., and Tyler, A. D. (2020). Systematic evaluation of supervised machine learning for sample origin prediction using metagenomic sequencing data. *Biol. Dir.* 15:29. doi: 10.1186/s13062-020-00287-y
- Chiavegatto Filho, A., Batista, A. F. D. M., and dos Santos, H. G. (2021). Data Leakage in Health Outcomes Prediction With Machine Learning. Comment on "Prediction of Incident Hypertension Within the Next Year: Prospective Study Using Statewide Electronic Health Records and Machine Learning". *J. Med. Internet Res.* 23:e10969. doi: 10.2196/10969
- Chitu, A. G., Zipp, M. H. A. A., and Zwaan, J. (2022). A novel exploration technique using the microbial fingerprint of shallow sediment to detect hydrocarbon microseepage and predict hydrocarbon charge — An Argentinian case study. *Interpretation* 10, 1F-T211.
- Christin, S., Hervet, É., and Lecomte, N. (2019). Applications for deep learning in ecology. *Methods Ecol. Evol.* 10, 1632–1644. doi: 10.1111/2041-210x.13256
- Cordier, T., Alonso-Sáez, L., Apothéloz-Perret-Gentil, L., Aylagas, E., Bohan, D. A., Bouchez, A., et al. (2021). Ecosystems monitoring powered by environmental genomics: A review of current strategies with an implementation roadmap. *Mol. Ecol.* 30, 2937–2958. doi: 10.1111/mec.15472
- Cordier, T., Forster, D., Dufresne, Y., Martins, C. I. M., Stoeck, T., and Pawlowski, J. (2018). Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring. *Mol. Ecol. Res.* 18, 1381–1391. doi: 10.1111/1755-0998.12926
- Correia, A., and Weimann, A. (2021). Protein antibiotics: mind your language. *Nat. Rev. Microbiol.* 19:7. doi: 10.1038/s41579-020-00485-5
- De Anda, V., Zapata-Peñasco, I., Blaz, J., Poot-Hernández, A. C., Contreras-Moreira, B., González-Laffitte, M., et al. (2018). Understanding the Mechanisms Behind the Response to Environmental Perturbation in Microbial Mats: A Metagenomic-Network Based Approach. *Front. Microbiol.* 9:2606. doi: 10.3389/fmicb.2018.02606
- de Dios Miranda, J., Seoane, J. M., Esteban, Á., and Espí, E. (2019). *Microbial Exploration Techniques: An Offshore Case Study, Oilfield Microbiology*. Florida: CRC Press, 271–298.
- Dong, H. (2010). Mineral-microbe interactions: a review. *Front. Earth Sci. Chin.* 4:127–147. doi: 10.1007/s11707-010-0022-8
- Dubinsky, E. A., Butkus, S. R., and Andersen, G. L. (2016). Microbial source tracking in impaired watersheds using PhyloChip and machine-learning classification. *Water Res.* 105, 56–64. doi: 10.1016/j.watres.2016.08.035
- Falkowski, P. G., Fenchel, T., and DeLong, E. F. (2008). The Microbial Engines That Drive Earth's Biogeochemical Cycles. *Science* 320, 1034–1039. doi: 10.1126/science.1153213
- Fiannaca, A., La Paglia, L., La Rosa, M., Lo Bosco, G., Renda, G., Rizzo, R., et al. (2018). Deep learning models for bacteria taxonomic classification of metagenomic data. *BMC Bioinform.* 19:198. doi: 10.1186/s12859-018-2182-6
- Fontaine, L., Khomich, M., Andersen, T., Hessen, D. O., Rasconi, S., Davey, M. L., et al. (2021). Multiple thresholds and trajectories of microbial biodiversity predicted across browning gradients by neural networks and decision tree learning. *ISME Commun.* 1:37.
- Franzosa, E. A., Hsu, T., Sirota-Madi, A., Shafquat, A., Abu-Ali, G., Morgan, X. C., et al. (2015). Sequencing and beyond: integrating molecular 'omics' for microbial community profiling. *Nat. Rev. Microbiol.* 13, 360–372. doi: 10.1038/nrmicro3451
- Fruehe, L., Cordier, T., Dully, V., Breiner, H. W., Lentendu, G., Pawlowski, J., et al. (2021). Supervised machine learning is superior to indicator value inference in monitoring the environmental impacts of salmon aquaculture using eDNA metabarcodes. *Mol. Ecol.* 30, 2988–3006. doi: 10.1111/mec.15434
- Ghannam, R. B., and Techtman, S. M. (2021). Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring. *Comput. Struct. Biotechnol. J.* 19, 1092–1107. doi: 10.1016/j.csbj.2021.01.028
- Gibbons, S. M., and Gilbert, J. A. (2015). Microbial diversity—exploration of natural ecosystems and microbiomes. *Curr. Opin. Genet. Dev.* 35, 66–72. doi: 10.1016/j.gde.2015.10.003
- Giuliani, G., Dao, H., De Bono, A., Chatenoux, B., Allenbach, K., De Laborie, P., et al. (2017). Live Monitoring of Earth Surface (LIMES): A framework for monitoring environmental changes from Earth Observations. *Rem. Sensing Environ.* 202, 222–233. doi: 10.1016/j.rse.2017.05.040
- Glasl, B., Bourne, D. G., Frade, P. R., Thomas, T., Schaffelke, B., and Webster, N. S. (2019). Microbial indicators of environmental perturbations in coral reef ecosystems. *Microbiome* 7:94. doi: 10.1186/s40168-019-0705-7
- Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome Datasets Are Compositional: And This Is Not Optional. *Front. Microbiol.* 8:2224. doi: 10.3389/fmicb.2017.02224
- Goodswen, S. J., Barratt, J. L. N., Kennedy, P. J., Kaufer, A., Calarco, L., and Ellis, J. T. (2021). Machine learning and applications in microbiology. *FEMS Microbiol. Rev.* 45:fua015.
- Gorbushina, A. A., and Krumbein, W. E. (2000). "Subaerial Microbial Mats and Their Effects on Soil and Rock," in *Microbial Sediments*, eds R. E. Riding and S. M. Awramik (Berlin, Heidelberg: Springer), 161–170. doi: 10.1007/978-3-662-04036-2_18
- Grantham, H. S., Duncan, A., Evans, T. D., Jones, K. R., and Beyer, H. L. (2020). Anthropogenic modification of forests means only 40% of remaining forests have high ecosystem integrity. *Nat. Comm.* 11:5978.
- Gutleben, J., De Mares, M., Chaib, van Elsas, J. D., Smidt, H., Overmann, J., and Sipkema, D. (2018). The multi-omics promise in context: from sequence to microbial isolate. *Crit. Rev. Microbiol.* 44, 212–229. doi: 10.1080/1040841X.2017.1332003
- Hanssen, S. V., Daioglou, V., Steinmann, Z. J. N., Doelman, J. C., Van Vuuren, D. P., and Huijbregts, M. A. J. (2020). The climate change mitigation potential of bioenergy with carbon capture and storage. *Nat. Clim. Change* 10, 1023–1029. doi: 10.1038/s41558-020-0885-y
- Haque, R., Paradisi, F., and Allers, T. (2020). *Haloferax volcanii* for biotechnology applications: challenges, current state and perspectives. *Appl. Microbiol. Biotechnol.* 104, 1371–1382. doi: 10.1007/s00253-019-10314-2
- Hennon, G. M. M., and Dyhrman, S. T. (2020). Progress and promise of omics for predicting the impacts of climate change on harmful algal blooms. *Harmful Algae* 91:101587. doi: 10.1016/j.hal.2019.03.005
- Hermans, S. M., Buckley, H. L., Case, B. S., Curran-Cournane, F., Taylor, M., and Lear, G. (2020). Using soil bacterial communities to predict physico-chemical variables and soil quality. *Microbiome* 8:79. doi: 10.1186/s40168-020-00858-1
- Herold, M., Martínez Arbas, S., Narayanasamy, S., Sheik, A. R., Kleine-Borgmann, L. A. K., Lebrun, L. A., et al. (2020). Integration of time-series meta-omics data reveals how microbial ecosystems respond to disturbance. *Nat. Comm.* 11:5281. doi: 10.1038/s41467-020-19006-2
- Janßen, R., Beck, A. J., Werner, J., Dellwig, O., Alneberg, J., Kreikemeyer, B., et al. (2021). Machine Learning Predicts the Presence of 2,4,6-Trinitrotoluene in Sediments of a Baltic Sea Munitions Dumpsite Using Microbial Community Compositions. *Front. Microbiol.* 12:626048. doi: 10.3389/fmicb.2021.626048
- Janßen, R., Zabel, J., von Lukas, U., and Labrenz, M. (2019). An artificial neural network and Random Forest identify glyphosate-impacted brackish communities based on 16S rRNA amplicon MiSeq read counts. *Mar. Poll. Bull.* 149:110530. doi: 10.1016/j.marpolbul.2019.110530
- Julínová, M., Vaňharová, L., and Jurča, M. (2018). Water-soluble polymeric xenobiotics – Polyvinyl alcohol and polyvinylpyrrolidone – And potential solutions to environmental issues: A brief review. *J. Environ. Manage.* 228, 213–222. doi: 10.1016/j.jenvman.2018.09.010
- Junghare, B., Spittler, D., and Schink, B. (2019). Anaerobic degradation of xenobiotic isophthalate by the fermenting bacterium *Syntrophorhabdus aromaticivorans*. *ISME J.* 13, 1252–1268. doi: 10.1038/s41396-019-0348-5
- Kaster, A.-K., and Sobol, M. S. (2020). Microbial single-cell omics: the crux of the matter. *Appl. Microbiol. Biotechnol.* 104, 8209–8220. doi: 10.1007/s00253-020-10844-0
- Katsuyama, C., Nakaoka, S., Takeuchi, Y., Tago, K., Hayatsu, M., and Kato, K. (2009). Complementary cooperation between two syntrophic bacteria in pesticide degradation. *J. Theor. Biol.* 256, 644–654. doi: 10.1016/j.jtbi.2008.10.024
- Keasling, J., Garcia Martin, H., Lee, T. S., Mukhopadhyay, A., Singer, S. W., and Sundstrom, E. (2021). Microbial production of advanced biofuels. *Nat. Rev. Microbiol.* 19, 701–715.

- Kim, Y., and Oh, S. (2021). Machine-learning insights into nitrate-reducing communities in a full-scale municipal wastewater treatment plant. *J. Environ. Manage.* 300:113795. doi: 10.1016/j.jenvman.2021.113795
- Knight, R., Vrbanc, A., Taylor, B. C., Akse, A., Callewaert, C., Debelius, J., et al. (2018). Best practices for analysing microbiomes. *Nat. Rev. Microbiol.* 16, 410–422. doi: 10.1038/s41579-018-0029-9
- Lambert, B. S., Groussman, R. D., Schatz, M. J., Coesel, S. N., Durham, B. P., Alverson, A. J., et al. (2021). The dynamic trophic architecture of open-ocean protist communities revealed through machine-guided metatranscriptomics. *Proc. Natl. Acad. Sci. U S A* 119:e2100916119. doi: 10.1073/pnas.2100916119
- Larsen, P. E., Field, D., and Gilbert, J. A. (2012). Predicting bacterial community assemblages using an artificial neural network approach. *Nat. Methods* 9, 621–625. doi: 10.1038/nmeth.1975
- Layton, K. K. S., and Bradbury, I. R. (2021). Harnessing the power of multi-omics data for predicting climate change response. *J. Anim. Ecol.* [Epub online ahead of print]. doi: 10.1111/1365-2656.13619
- Li, C., Wang, L., Ji, S., Chang, M., Wang, L., Gan, Y., et al. (2021). The ecology of the plastisphere: Microbial composition, function, assembly, and network in the freshwater and seawater ecosystems. *Water Res.* 2021:117428. doi: 10.1016/j.watres.2021.117428
- Li, Z., and Fantke, P. (2022). Toward harmonizing global pesticide regulations for surface freshwaters in support of protecting human health. *J. Environ. Manage.* 301:113909. doi: 10.1016/j.jenvman.2021.113909
- Lian, B., Chen, Y., Zhu, L., and Yang, R. (2008). Effect of Microbial Weathering on Carbonate Rocks. *Earth Sci. Front.* 15, 90–99. doi: 10.1016/s1872-5791(09)60009-9
- Liang, Q., Bible, P. W., Liu, Y., Zou, B., and Wei, L. (2020). DeepMicrobes: taxonomic classification for metagenomics with deep learning. *NAR Genom. Bioinform.* 2:lqaa009. doi: 10.1093/nargab/lqaa009
- Liao, J., Guo, X., Weller, D. L., Pollak, S., Buckley, D. H., Wiedmann, M., et al. (2021). Nationwide genomic atlas of soil-dwelling *Listeria* reveals effects of selection and population ecology on pangenome evolution. *Nat. Microbiol.* 6, 1021–1030. doi: 10.1038/s41564-021-00935-7
- Liao, K., Bai, Y., Huo, Y., Jian, Z., Hu, W., Zhao, C., et al. (2018). Integrating microbial biomass, composition and function to discern the level of anthropogenic activity in a river ecosystem. *Environ. Int.* 116, 147–155. doi: 10.1016/j.envint.2018.04.003
- Lopatkin, A. J., and Collins, J. J. (2020). Predictive biology: modelling, understanding and harnessing microbial complexity. *Nat. Rev. Microbiol.* 18, 507–520. doi: 10.1038/s41579-020-0372-5
- Lu, X., Ye, X., Zhou, M., Zhao, Y., Weng, H., Kong, H., et al. (2021). The underappreciated role of agricultural soil nitrogen oxide emissions in ozone pollution regulation in North China. *Nat. Comm.* 12:5021. doi: 10.1038/s41467-021-25147-9
- Lv, M., Luan, X., Liao, C., Wang, D., Liu, D., Zhang, G., et al. (2020). Human impacts on polycyclic aromatic hydrocarbon distribution in Chinese intertidal zones. *Nat. Sustain.* 3, 878–884. doi: 10.1038/s41893-020-0565-y
- Marco, D. E., and Abram, F. (2019). Editorial: Using Genomics, Metagenomics and Other “Omics” to Assess Valuable Microbial Ecosystem Services and Novel Biotechnological Applications. *Front. Microbiol.* 10:151. doi: 10.3389/fmicb.2019.00151
- Miao, Y., Johnson, N. W., Phan, T., Heck, K., Gedalanga, P. B., Zheng, X., et al. (2020). Monitoring, assessment, and prediction of microbial shifts in coupled catalysis and biodegradation of 1,4-dioxane and co-contaminants. *Water Res.* 173:115540. doi: 10.1016/j.watres.2020.115540
- Morimura, S., Zeng, X., Noboru, N., and Hosono, T. (2020). Changes to the microbial communities within groundwater in response to a large crustal earthquake in Kumamoto, southern Japan. *J. Hydrol.* 581:124341. doi: 10.1016/j.jhydrol.2019.124341
- Naumann, G., Cammalleri, C., Mentaschi, L., and Feyen, L. (2021). Increased economic drought impacts in Europe with anthropogenic warming. *Nat. Clim. Change* 11, 485–491. doi: 10.1038/s41558-021-01044-3
- Oh, S., and Kim, Y. (2021). Machine learning application reveal dynamic interaction of polyphosphate-accumulating organism in full-scale wastewater treatment plant. *J. Water Proc. Eng.* 44:102417. doi: 10.1016/j.jwpe.2021.102417
- Ortiz-Bobea, A., Ault, T. R., Carrillo, C. M., Chambers, R. G., and Lobell, D. B. (2021). Anthropogenic climate change has slowed global agricultural productivity growth. *Nat. Clim. Change* 11, 306–312. doi: 10.1038/s41558-021-01000-1
- Oudah, M., and Henschel, A. (2018). Taxonomy-aware feature engineering for microbiome classification. *BMC Bioinform.* 19:227. doi: 10.1186/s12859-018-2205-3
- Oyetunde, T., Liu, D., Martin, H. G., and Tang, Y. J. (2019). Machine learning framework for assessment of microbial factory performance. *PLoS One* 14:e0210558. doi: 10.1371/journal.pone.0210558
- Panke-Buisse, K., Poole, A. C., Goodrich, J. K., Ley, R. E., and Kao-Kniffin, J. (2014). Selection on soil microbiomes reveals reproducible impacts on plant function. *Isme J.* 9:980. doi: 10.1038/ismej.2014.196
- Pulster, E. L., Gracia, A., Armenteros, M., Toro-Farmer, G., Snyder, S. M., Carr, B. E., et al. (2020). A First Comprehensive Baseline of Hydrocarbon Pollution in Gulf of Mexico Fishes. *Sci. Rep.* 10:6437. doi: 10.1038/s41598-020-62944-6
- Ramirez, K. S., Knight, C. G., de Hollander, M., Brearley, F. Q., Constantinides, B., Cotton, A., et al. (2018). Detecting macroecological patterns in bacterial communities across independent studies of global soils. *Nat. Microbiol.* 3, 189–196. doi: 10.1038/s41564-017-0062-x
- Raza, S., Kim, J., Sadowsky, M. J., and Unno, T. (2021). Microbial source tracking using metagenomics and other new technologies. *J. Microbiol.* 59, 259–269. doi: 10.1007/s12275-021-0668-9
- Santos, A., Barbosa-Póvoa, A., and Carvalho, A. (2019). Life cycle assessment in chemical industry – a review. *Curr. Opin. Chem. Eng.* 26, 139–147. doi: 10.1016/j.coche.2019.09.009
- Schweitzer, H., Aalto, N. J., Busch, W., Chan, D. T. Chat, Chiesa, M., Elvevoll, E. O., et al. (2021). Innovating carbon-capture biotechnologies through ecosystem-inspired solutions. *One Earth* 4, 49–59. doi: 10.1016/j.oneear.2020.12.006
- Shah, R. M., Stephenson, S., Crosswell, J., Gorman, D., Hillyer, K. E., and Palombo, E. A. (2022). Omics-based ecosurveillance uncovers the influence of estuarine macrophytes on sediment microbial function and metabolic redundancy in a tropical ecosystem. *Sci. Total Environ.* 809:151175. doi: 10.1016/j.scitotenv.2021.151175
- Shaheen, M., Shahbaz, M., ur Rehman, Z., and Guergachi, A. (2011). Data mining applications in hydrocarbon exploration. *Artif. Intell. Rev.* 35, 1–18. doi: 10.1007/s10462-010-9180-z
- Simul Bhuyan, M., Venkatramanan, S., Selvam, S., Szabo, S., Hossain, M., Rashed-Un-Nabi, M., et al. (2021). Plastics in marine ecosystem: A review of their sources and pollution conduits. *Reg. Stud. Mar. Sci.* 41:101539. doi: 10.1111/j.gcb.14572
- Sintayehu, D. W. (2018). Impact of climate change on biodiversity and associated key ecosystem services in Africa: a systematic review. *Ecosyst. Health Sustain.* 4, 225–239. doi: 10.1080/20964129.2018.1530054
- Smith, M. B., Rocha, A. M., Smillie, C. S., Olesen, S. W., Paradis, C., Wu, L., et al. (2015). Natural Bacterial Communities Serve as Quantitative Geochemical Biosensors. *mbio* 6, e326–e315. doi: 10.1128/mbio.00326-15
- Sohrabi, H., Hemmati, A., Majidi, M. R., Eyvazi, S., Jahanban-Esfahlan, A., Baradaran, B., et al. (2021). Recent advances on portable sensing and biosensing assays applied for detection of main chemical and biological pollutant agents in water samples: A critical review. *Trends Anal. Chem.* 143:116344. doi: 10.1016/j.trac.2021.116344
- Solis-Reyes, S., Avino, M., Poon, A., and Kari, L. (2018). An open-source k-mer based machine learning tool for fast and accurate subtyping of HIV-1 genomes. *PLoS One* 13:e0206409. doi: 10.1371/journal.pone.0206409
- Su, L., Jia, W., Hou, C., and Lei, Y. (2011). Microbial biosensors: a review. *Biosens. Bioelectr.* 26, 1788–1799. doi: 10.1016/j.bios.2010.09.005
- Szafrański, S. P., Deng, Z.-L., Tomasch, J., Jarek, M., Bhuj, S., Meisinger, C., et al. (2015). Functional biomarkers for chronic periodontitis and insights into the roles of *Prevotella nigrescens* and *Fusobacterium nucleatum*; a metatranscriptome analysis. *Npj Biofilms and Microbiom.* 1:15017. doi: 10.1038/npjbiofilms.2015.17
- Thompson, J., Johansen, R., Dunbar, J., and Munsy, B. (2019). Machine learning to predict microbial community functions: An analysis of dissolved organic carbon from litter decomposition. *PLoS One* 14:e0215502. doi: 10.1371/journal.pone.0215502
- Turan, N. B., Erkan, H. S., Engin, G. O., and Bilgili, M. S. (2019). Nanoparticles in the aquatic environment: Usage, properties, transformation and toxicity—A review. *Proc. Safety Environ. Protect.* 130, 238–249. doi: 10.1016/j.psep.2019.08.014

- Vardhan, K. H., Kumar, P. S., and Panda, R. C. (2019). A review on heavy metal pollution, toxicity and remedial measures: Current trends and future perspectives. *J. Mol. Liquids* 290:111197. doi: 10.1016/j.molliq.2019.111197
- Wang, C., Mao, G., Liao, K., Ben, W., Qiao, M., Bai, Y., et al. (2021). Machine learning approach identifies water sample source based on microbial abundance. *Water Res.* 199:117185. doi: 10.1016/j.watres.2021.117185
- Wang, Y., Hatt, J. K., Tsementzi, D., Rodriguez, R. L., Ruiz-Pérez, C. A., Weigand, M. R., et al. (2017). Quantifying the Importance of the Rare Biosphere for Microbial Community Response to Organic Pollutants in a Freshwater Ecosystem. *Appl. Environ. Microbiol.* 83, e3321–e3316. doi: 10.1128/AEM.03321-16
- Wheeler, N. E. (2019). Tracing outbreaks with machine learning. *Nat. Rev. Microbiol.* 17, 269–269. doi: 10.1038/s41579-019-0153-1
- Wirbel, J., Zych, K., Essex, M., Karcher, N., Kartal, E., Salazar, G., et al. (2021). Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox. *Genom. Biol.* 22:93. doi: 10.1186/s13059-021-02306-1
- Wu, L., Ning, D., Zhang, B., Li, Y., Zhang, P., Shan, X., et al. (2019). Global diversity and biogeography of bacterial communities in wastewater treatment plants. *Nat. Microbiol.* 4, 1183–1195.
- Xu, Z., Malmer, D., Langille, M. G. I., Way, S. F., and Knight, R. (2014). Which is more important for classifying microbial communities: who's there or what they can do? *ISME J.* 8, 2357–2359. doi: 10.1038/ismej.2014.157
- York, A. (2021). Avoiding the pitfalls in microbiota studies. *Nat. Rev. Microbiol.* 19:2. doi: 10.1038/s41579-020-00480-w
- Yuan, J., Wen, T., Zhang, H., Zhao, M., Penton, C. R., Thomashow, L. S., et al. (2020). Predicting disease occurrence with high accuracy based on soil macroecological patterns of Fusarium wilt. *ISME J.* 14, 2936–2950. doi: 10.1038/s41396-020-0720-5
- Zeller, G., Tap, J., Voigt, A. Y., Sunagawa, S., Kultima, J. R., Costea, P. I., et al. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* 10, 766–766. doi: 10.15252/msb.20145645
- Zhang, C., Gao, Z., Shi, W., Li, L., Tian, R., Huang, J., et al. (2020). Material conversion, microbial community composition and metabolic functional succession during green soybean hull composting. *Biores. Technol.* 316:123823. doi: 10.1016/j.biortech.2020.123823
- Zijp, M., Mallinson, T., Zwaan, J., Chitu, A., and David, P. (2021). “Eagle Ford and Bakken Productivity Prediction Using Soil Microbial Fingerprinting and Machine Learning,” in *Paper Presented at the SPE/AAPG/SEG Unconventional Resources Technology Conference*, (Houston, Texas, USA).

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.

Copyright © 2022 McElhinney, Catacutan, Mawart, Hasan and Dias. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Clean and Safe Drinking Water Systems *via* Metagenomics Data and Artificial Intelligence: State-of-the-Art and Future Perspective

Asala Mahajna^{1,2*}, Inez J. T. Dinkla¹, Gert Jan W. Euverink², Karel J. Keesman³ and Bayu Jayawardhana²

¹Wetsus – European Centre of Excellence for Sustainable Water Technology, Leeuwarden, Netherlands, ²Engineering and Technology Institute Groningen, University of Groningen, Groningen, Netherlands, ³Mathematical and Statistical Methods – Biometris, Wageningen University, Wageningen, Netherlands

OPEN ACCESS

Edited by:

Mohammad-Hossein Sarrafzadeh,
University of Tehran, Iran

Reviewed by:

Ajaya Kumar Rout,
Central Inland Fisheries Research
Institute (ICAR), India
Elvis Fosso Kankeu,
University of South Africa,
South Africa

*Correspondence:

Asala Mahajna
a.mahajna@rug.nl

Specialty section:

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

Received: 09 December 2021

Accepted: 04 April 2022

Published: 05 May 2022

Citation:

Mahajna A, Dinkla IJT,
Euverink GJW, Keesman KJ and
Jayawardhana B (2022) Clean and
Safe Drinking Water Systems *via*
Metagenomics Data and Artificial
Intelligence: State-of-the-Art and
Future Perspective.
Front. Microbiol. 13:832452.
doi: 10.3389/fmicb.2022.832452

The use of next-generation sequencing technologies in drinking water distribution systems (DWDS) has shed insight into the microbial communities' composition, and interaction in the drinking water microbiome. For the past two decades, various studies have been conducted in which metagenomics data have been collected over extended periods and analyzed spatially and temporally to understand the dynamics of microbial communities in DWDS. In this literature review, we outline the findings which were reported in the literature on what kind of occupancy-abundance patterns are exhibited in the drinking water microbiome, how the drinking water microbiome dynamically evolves spatially and temporally in the distribution networks, how different microbial communities co-exist, and what kind of clusters exist in the drinking water ecosystem. While data analysis in the current literature concerns mainly with confirmatory and exploratory questions pertaining to the use of metagenomics data for the analysis of DWDS microbiome, we present also future perspectives and the potential role of artificial intelligence (AI) and mechanistic models to address the predictive and mechanistic questions. The integration of metagenomics, AI, and mechanistic models transcends metagenomics into functional metagenomics, enabling deterministic understanding and control of DWDS for clean and safe drinking water systems of the future.

Keywords: drinking water production, drinking water monitoring, high-throughput sequencing technology, metagenomics, machine learning, water distribution

INTRODUCTION

The importance of access to clean water and sanitation has been recognized worldwide as one of the main themes in the UN Sustainable Development Goals. While developed nations have connected their population to the water network, access to safe and clean water poses a challenge to the water management authorities. The rapid depletion of groundwater and the

contamination of surface water by industrial, agricultural, and urban waste streams have contributed to this problem. Sanitation and hygiene also rely heavily on adequate access to clean water for preventing and containing diseases to reduce the spread of pathogens and viruses (WHO, 2020). While the majority of drinking water bacteria is not dangerous for human health and is actually useful for the production of drinking water at the treatment plant, these organisms can cause unpleasant taste, odor, and turbidity of drinking water when present in excess (van Lieverloo et al., 2002; Vreeburg et al., 2004). Around 80% of customers' complaints to the water utilities are about unwanted aesthetic aspects of drinking water that are generated during its distribution. These impaired aesthetics, which are a result of the uncontrolled growth of indigenous bacteria in particles, sediments, and biofilms in distribution pipelines might even include the presence of invertebrates in the water (Polychronopolous et al., 2003; Vreeburg and Boxall, 2007).

Uncontrolled growth of indigenous bacteria in water distribution systems results in microbially induced operational problems in distribution pipes which introduce significant investment and maintenance costs for water utilities (Allion et al., 2011). In the Netherlands alone, investment costs on distribution pipelines require approximately 50% of water utility investments (de Moel et al., 2006). For example, sulfate-reducers and iron-oxidizers cause bio-corrosion of cast-iron pipes (Sun et al., 2014), and the growth of bacteria to high numbers in the form of a biofilm cause fouling of concrete pipes. In addition, the suspension of some of the bacteria which are attached to particles, sediments, or biofilms in distribution pipes can result in turbid or discolored water (Vreeburg et al., 2004). These bacteria are non-pathogenic and their excessive growth makes the water yellowish (Vreeburg and Boxall, 2007). Iron particles and manganese precipitates in water which are partially produced by bio-corrosion of iron pipes (Sun et al., 2014) or manganese-oxidizing or reducing organisms (Cerrato et al., 2010) cause water to be red or black colored (Seth et al., 2004). Other bacteria produce molecules affecting the taste and odor of water. For example, Actinomycetes produce Geosmin which is responsible for an earthy-muddy water taste (Srinivasan and Sorial, 2011), and sulfate-reducing or sulfur-oxidizing bacteria can enhance a sulfur-based odor (Scott and Pepper, 2010). On top of that, fungi, and yeast induce other aesthetic problem that has been recorded in drinking water systems. They negatively alter water odor and taste. Protozoa and invertebrates such as worms (e.g., Annelida), crustaceans (e.g., Asellidae), or snails (e.g., Mollusca) have also been found in distribution systems (Christensen et al., 2011). As protozoa and invertebrates are at the top of the trophic chain, they indicate the presence of a high number of bacteria in water.

This uncontrolled growth of indigenous bacteria during water distribution can result in the exceedance of water quality regulatory guidelines (Sartory, 2004). The current regulation dictates that water treatment processes should yield drinking water that causes less than 1 infection per 10,000 people per year. However, continuous threats from newly emerging micro-pollutants and the risk of recontamination due to the growth of environmental pathogens in drinking

water sources are still a concern. For instance, numerous pathogens which are opportunistic and hygienically threatening such as *Legionella pneumophila*, *Aeromonas hydrophila*, *Pseudomonas aeruginosa*, *Klebsiella pneumoniae*, *Mycobacteria*, and *Campylobacter* are able to grow at low nutrient concentrations in drinking water distribution systems and/or in household pipelines.

Limiting changes in the bacterial community during drinking water distribution and the prevention of uncontrolled growth up to high bacterial cell numbers and to the occurrence of unwanted microorganisms is done through removing carbon sources and nutrients, inactivating pathogenic organisms, removing chemical toxic compounds, and improving the transparency, taste, odor, and color of the water at the water treatment plant. Achieving high-quality drinking water that is biologically stable during transportation is done through physical, chemical, and biological processes such as dosing chlorine, aeration, ozonation, UV irradiation, active carbon filtrations, coagulation, flocculation, sedimentation steps, and/or rapid or slow sand filtration. The choice of which steps to apply to treat the water will depend on the source of the water and the initial water quality. After treatment, the water is transported *via* a pipeline system to the point of use or discharge. In this transportation process, residual organic material and microorganisms in the water may alter the quality of the water in this distribution system. The microbiological activity influences the chemical composition of the water and vice versa. The presence of organic material in water sustains the growth of microorganisms that form undesired biofilms and/or turbidity in the distribution system. The current removal of the organic material in the upstream purification steps aims to minimize regrowth but does not always result in biologically stable water. A balance between the efforts put in the removal and the risks for regrowth may be found in the specific quality of the organic material (Hijnen et al., 2014). However, detailed characteristics of the organic material are largely unknown, hampering the design of more effective treatment steps to produce biological stable water, i.e., water that does not support the growth of bacteria and other organisms in the distribution system.

While many countries around the world add disinfectant (such as chlorine, mono-chloramine, or chlorine dioxide) to drinking water as a secondary disinfection step, some European countries such as the Netherlands, Germany, Austria, and Switzerland use extensive treatment strategies which eliminate the bacterial growth supporting compounds (nutrients) in the water supplied to limit the potential regrowth in the distribution system. One disadvantage of using disinfectants in drinking water is that disinfectants react with organic compounds which results in the potential formation of carcinogenic by-products. Therefore, the concentrations of added disinfectants are kept to a minimum, with a higher risk of regrowth. Both methods are very effective at limiting bacterial growth in drinking water distribution systems. Yet, microbial changes in drinking water during distribution have been recorded in many countries. A more comprehensive overview of the drinking water distribution system microbiome is provided by Gomez and Aggarwal (2019).

This paper presents a review of recent advances in the monitoring, production, and distribution of drinking water using various -omics technologies. Firstly, the literature on microbial ecology in drinking water systems is revisited and various standard practices by water management authorities to monitor their activities are presented in “Microbiome in Water Systems” section. In “NGS Technology for Drinking Water Distribution Systems” section, the emergence of genetic sequencing technology as a new key-enabling water technology is discussed. This high throughput technology can shed light on microbial activities in much finer detail and allows us to understand the dynamics and various roles of microbial communities. This knowledge, through the employment of artificial intelligence and mechanistic models, can in turn be used to monitor and control the biological processes in drinking water systems as illustrated in “Artificial Intelligence Methods in DWDS” section.

MICROBIOME IN WATER SYSTEMS

Factors Affecting Drinking Water Microbial Ecology

The complexity of water in a DWDS, as a living aquatic ecosystem, is further enhanced by numerous aspects which are influencing the network of microbial interactions that exist in it during its distribution. Some of the aspects that influence bacterial growth during water distribution are: (1) the existence of the food chain, (2) concentration and type of nutrients, (3) type and concentration of residual disinfectant (if any), (4) microcosmic environmental conditions found in bulk water, sediment and/or biofilm, (5) system-wide environmental conditions (temperature, pH, etc.), (6) prevailing hydraulic condition and pipe materials, (7) and water residence time/water age (Prest et al., 2016).

Assessment of Drinking Water Microbial Quality

Characterizing organic material in water and quantifying its growth-promoting properties for micro-organisms has been previously done using different methods. The assimilable organic carbon (AOC) method is based on the measurement of the growth of two pure bacterial strains in a pasteurized water sample. The biodegradable dissolved organic carbon (BDOC) method measures the uptake of dissolved organic carbon (DOC) by the autochthonous bacteria in a water sample, the liquid chromatography–organic carbon detection technique (LC-OCD) identifies and quantifies natural organic matter constituents in aquatic environments, and the biofilm formation rate (BFR) method quantifies the ability of water to promote the growth of bacteria into a biofilm. However, these methods are indicative tools and do not provide detailed characteristics of the organic material which subsequently hampers real-time monitoring of treatment processes and their optimization.

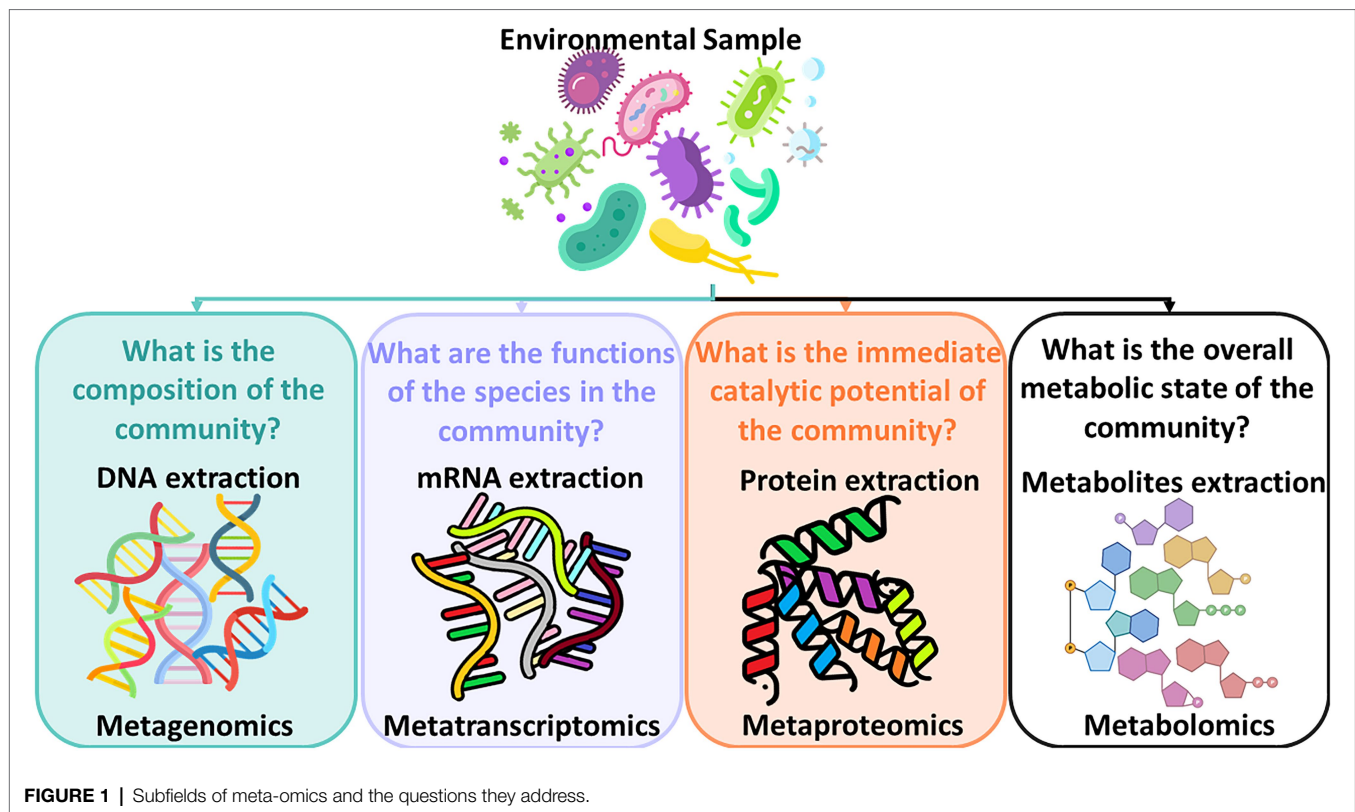
In addition, understanding microbial dynamics in drinking water distribution systems has been limited because of drawbacks

of available methods for characterizing drinking water bacterial communities which rely heavily on culture-based techniques. Assessing water microbial quality has been traditionally done using heterotrophic plate counts (HPC) which is a method for bacterial enumeration. Alternatively, bioassays which are analytical methods for determining the concentration or potency of a substance by its effect on living cells or tissues can be applied. When microorganisms grow on organic substrates, specific degradation pathways are induced to enzymatically metabolize these organic compounds. Specific assays that can detect these enzymes require time-consuming, lengthy laboratory work. These methods are hypothesis-driven whose goal is to detect a targeted suspected compound and a selection of enzyme assays needs to be determined upfront. As these methods generate an assessment of the water quality with a time lag, detect only a minute fraction of the bacteria found in water in reality, and are limited when it comes to identifying all characteristics of the bacterial community found in the water, Next Generation Sequencing (NGS) technologies have been introduced in order to better assess the microbial drinking water quality. Initially, NGS technologies were utilized by the medical field for studying the gut microbiome (Malla et al., 2019) and by pharmaceutical industries for drug discoveries and personalized medicine (Vandeputte, 2021). Progressively, this technology has been introduced into the field of environmental microbiology to study soil microbiome (Nesme et al., 2016), and aquatic systems (Behera et al., 2021), and subsequently into the fields of wastewater treatment and drinking water quality and their respective processes (Tan et al., 2015; Zhang and Liu, 2019). While the development of the NGS technologies is a process of continuous enhancements (Slatko et al., 2018), the greatest advantage of NGS technologies is that they can provide a comprehensive assessment of the abundance, viability, and community composition of the microorganisms found in the water sample. The new field of meta-omics enables scientists to study mixtures of genetic material from all organisms in a sample. **Figure 1** shows the subfields of meta-omics and what kind of questions these fields attempt to address.

Mechanistic Models for Simulating Drinking Water Quality in Distribution Networks

Water utilities have been using mechanistic hydraulic models to simulate drinking water quality in drinking water distribution systems. These simulation tools are used for the purpose of optimizing the design of the water infrastructure and its facilities, the real-time hydraulic operation and monitoring of the network, simulation of events of contamination and tracing the source of such an event, and establishing guidelines for the operation and maintenance (O&M) of the supply system.

In 1990, the United States Environmental Protection Agency (USEPA) developed the Environmental Protection Agency Network (EPANET) which is the first computational software package for modeling the hydraulics of drinking water distribution systems (Rossman, 2000). Since then several commercially



available spin-offs of EPANET were released. EPANET model start from a link-node structure where pipes are modelled as links, and junctions, hydraulic control elements, consumers, and sources are modelled as nodes. Drinking water quality is modeled in EPANET as an “additional simulation layer” on top of the hydraulic simulations which provide the core functionality of EPANET. Water age and source-tracing are two functionalities in EPANET which can provide an overarching assessment of the overall drinking water quality in distribution systems. Water age provides a proportional indicator of the decay of the residual disinfectant in the system and the formation of the respective disinfection by-products (DBP). On the other hand, source-tracing, which simulates the flow-path of water from the point of supply up to the point of consumption, has an added value when modeling drinking water quality in multi-quality water distribution systems where water comes from different sources. Source-tracing provides insight into a source of a contaminate in case of a contamination event, indicates potential mixing areas in the water supply network and provides knowledge about source influence areas in the system. Water age and source-tracing are mere high-level indicators of drinking water quality and in actuality drinking water quality may differ remarkably (Chenevey, 2022).

In EPANET, the Dynamic Water Quality Model (DWQM) serves as the basis for water quality modelling. For this, EPANET uses continuity equations for energy, mass, flow at nodes, flow for each storage component, mass for each storage component and each quality parameter, and equations for dilution requirements for modelling water quality under unsteady state

flow conditions (Todini and Rossman, 2012). DWQM models single species concentration in the distribution system under first-order kinetics and plug-flow advection assumptions. However, the single species models do not account for microbial growth in the drinking water system and are merely limited to modeling process parameters throughout the distribution network (Woolschlager et al., 2005).

Recently, the National Health Systems Resource Centre (NHSRC) released a Multi-Species eXtension to EPANET called EPANET-MSX that enables modelling of numerous interacting species in the bulk flow and on the pipe walls, while modelling microbial growth, as well. This extension models heterotrophic microbial growth in both their fixed and suspended forms through solving a set of interdependent, multispecies, mass balance equations which is an expansion of the fundamental equations provided in the DWQM (Shang et al., 2011). Other multi-species models which are empirical, semi-mechanistic, and mechanistic were developed for research purposes to simulate microbial drinking water quality are not commercially available (). However, the modeling of microbial growth in the multi-species models is limited to two species/values (i.e., mass of free bacteria in bulk water, and mass of attached bacteria on pipe wall), and does not account for the rich microbial diversity which exist in the drinking water. In addition, the computation nature of EPANET-MSX, which solves a set of differential-algebraic equations (DAEs) in semi-explicit form, renders this model computationally inefficient for modeling the concentration of each bacterial species in a system that contains bacterial

diversity in the magnitude of thousands. Hence, incorporating machine learning algorithms, which are good at handling data that are multi-dimensional and multi-variety, with metagenomics dataset can potentially present a computationally more efficient approach for simulating microbial drinking water quality (Rackauckas et al., 2020).

NGS TECHNOLOGY FOR DRINKING WATER DISTRIBUTION SYSTEMS

Metagenomics Analysis for Microbial Communities

The emergence of new genetic sequencing technologies has enabled the gathering of crucial *in-situ* information related to microbial communities and occupancy-abundance dynamics in drinking water. In the pioneering work of Santo Domingo et al. in 2003 at the US Environmental Protection Agency Test and Evaluation (T&E) facility, metagenomics was applied to investigate the role of heterotrophic bacteria and ammonia-oxidizing bacteria in drinking water. They used a Distribution System Simulator (DSS) to assess the biofilm microbial composition in drinking water distribution systems (DWDS) due to the role of biofilms, which can contain human microbial pathogens, on public health. The researchers conducted 16S rDNA sequence analysis on both biofilm and bulk water samples from the DSS which revealed that *α-Proteobacteria* and *β-Proteobacteria* were the predominant bacteria in the feed water, discharge water, and biofilm samples. This early metagenomics application has been used to determine the effectiveness of disinfectant treatment to control microbial communities in DWDS. In 2005, Tokajian et al., conducted a phylogenetic assessment of heterotrophic bacteria using 16S rDNA sequencing from an operational water distribution system in Lebanon. Water samples were taken from raw unchlorinated aquifer water and from different sites in the distribution network on a bimonthly basis over a period of 1 year. The analysis confirmed the aforementioned observations (Santo Domingo et al., 2003; Williams et al., 2004) that the majority of bacteria in drinking water were *α*-, *β*-, and *γ*-Proteobacteria. In addition, the study also revealed a higher presence of *sphingomonads* in drinking water samples than reported elsewhere in literature which can be attributed to the specific operational conditions in Lebanon.

Once microbial communities are identified using metagenomics data, the next step is to establish their specific role, function, and interaction with the environment. In 2006, Eichler et al. used RNA- and DNA-based 16S rRNA gene fingerprinting further to gain a comprehensive understanding of how different factors (i.e., different raw water sources, different treatment processes, and distribution) influence the microbial communities in tap water designated for human consumption. Based on the DWDS of the city Braunschweig in Germany involving two water reservoirs with two different surface water types: oligotrophic water and dystrophic water, Eichler et al. (2006) observed that that major taxonomic groups typical of freshwaters such as *α-Proteobacteria*, *β-Proteobacteria*, and

Bacteroidetes dominated the system. Comparative cluster analysis to the data revealed that there are three major types/clusters of communities in the system, each associated with the two types of surface water and to the chlorinated water, which is found to promote the growth of nitrifying bacteria. This work demonstrated the role of metagenomics analysis in revealing the importance of source water microflora to the drinking water microflora, in monitoring water quality, and in assessing the performance of different treatment processes. Further studies on the microbial diversity and composition in DWDS which support the metagenomic analysis in Eichler et al. (2006) were presented in Santo Domingo et al. (2003); Tokajian et al. (2005); Berney et al. (2009); Revetta et al. (2010), and Vital et al. (2012). The results of these studies are summarized in **Supplementary Table 1**.

Metagenomics Analysis for Temporal and Spatial Distributions and Intra-community Dynamics

The first study to investigate spatial and temporal dynamics of drinking water microbiota using metagenomics was presented in Rudi et al. (2009). The authors used 16S rRNA sequencing analysis to assess temporal and spatial diversity of tap water (namely, kitchen tap and toilet tap) microbiota in a Norwegian hospital between January and July 2006 (for temporal analysis). In their study, the researchers used density distribution analyses to investigate tap-specific distributions of the bacterial groups. Based on the hierarchical clustering analysis, they concluded that the microbiota clustered according to the location (spatial) and not to the season (temporal). Related to a potential public health issue, metagenomics analysis in their study provided additional insights. It is shown in Rudi et al. (2009) that *Legionella* had the highest relative abundance for the pathogen-related bacteria in the dataset, especially in the low-usage tap, which can be investigated further for controlling local *Legionella* or other pathogens colonization. Such spatial metagenomics analysis can prevent pathogenic outbreaks from reoccurring, such as the well-known *Pseudomonas aeruginosa* outbreak in an intensive care unit at Akershus university hospital which could be traced back to a single tap.

In 2014, Pinto et al. (2014) used a spatially distributed and temporally varying sampling approach to conduct spatial-temporal surveying and occupancy-abundance modelling techniques using metagenomics analysis in a chlorinated drinking water distribution system in the USA. They sampled and analyzed the bacterial communities in water leaving the treatment plant from June 2010 to August 2011 at the clean water reservoir of a wastewater treatment plant and at three locations from three different sectors in the drinking water distribution system (resulting in nine locations in total). The analysis, which was based on total DNA extracts, resulted in the identification of 4,369 Operational Taxonomic Units (OTUs) at a 97% similarity cut-off, across 20 different phyla in the 138 water samples over the 15-month sampling period. In spite of the high diversity of the bacterial community found in the water, the *Proteobacteria* phylum is again the dominant DW bacterial community

representing 60%–70% of the bacterial community for any given sample. Using Mantel's test, changes in the microbial community can be explained by around 5% of the highly diverse OTUs which indicates that this subset of OTUs can be used to track changes in the community. For instance, it was observed that β - and δ -*Proteobacteria* dominated the DWDS during the summer months while α - and γ -*Proteobacteria* were dominant in the winter. β -*Proteobacterium Hydrogenophaga* (a genus of *comamonas* bacteria) in contrast displayed peak relative abundance in the colder months. Pinto et al. (2014) concluded also that biofilms in the neighborhood of each sampling location or possibly even microbial ingress into the DWDS led to the observed location-specific OTUs in the system.

Prest (2015) studied temporal dynamics in bacterial community characteristics during a 2-year drinking water monitoring campaign in a full-scale distribution system operating without detectable disinfectant residual. The data collected came from a total of 360 water samples which were sampled on a biweekly basis from Kralingen water treatment plant effluent and at one fixed location in the DWDS. The samples were analyzed for heterotrophic plate counts (HPC), *Aeromonas* plate counts, adenosine-tri-phosphate (ATP) concentrations, flow cytometric (FCM) total and intact cell counts (TCC, ICC), water temperature, pH, conductivity, total organic carbon (TOC) and assimilable organic carbon (AOC). Computational multivariate analyses showed that the change in microbial parameters between the water treatment plant and DWDS had a predictable annual trend comparable to the water seasonal temperature fluctuations and was negatively correlated to the AOC concentration in the water treatment plant effluent. Prest (2015) concluded that microbial growth in DWDS was not attributed to a single parameter only in the treated effluent. Roeselers et al. (2015) conducted a similar study in which spatial and temporal patterns in phylogenetic diversity were investigated using high-throughput sequencing technology in 32 DWDS networks in the Netherlands where residual disinfectant is not used. They observed that the microbial community compositions from water samples can be differentiated based on the source of the water sample, e.g., raw water and processed water in different locations. In addition, the researchers observed that community structures of processed water did not differ substantially from end-point tap water which indicates that network-specific communities are stable in time. The analysis on microbial community clusters showed that the treatment plant rather than the sampling time points differentiates drinking water microbial communities.

All of the above-mentioned findings were consistent with the conclusions made by Blokker et al. (2016) who used self-organizing maps for relating water quality and water age in DWDS from a multi-year Dutch and United Kingdom dataset. Their analysis showed that water age and temperature may be treated as independent parameters influencing microbial water quality. In addition, they concluded that there is a clear influence of temperature, which is dictated by seasonal change, on *Aeromonas* and the HPC at 22°C. They also showed that while water age has been traditionally used as a mathematical modelling tool to give an indication for all system-specific

degradation of water quality, it appears to be of little value as an indicator for specific microbial water quality compared to water temperature. Their study recommends that specific DWDS conditions such as temperature, substrate concentration, and local shear stresses be incorporated in water quality models to better understand the risk of developing vulnerable water quality locations in drinking water distribution systems.

To assess the origin of bacteria in tap water and distribution system in an unchlorinated drinking water system, Liu et al. (2018) looked into the bacterial communities associated with biofilms, suspended particles, and loose deposits which are released in the distribution system as they are considered the major potential risk for drinking water bio-safety. They quantified the proportional contribution of the source water, treated water, and distribution system in determining the tap water bacterial community and concluded that the water purification process shaped the community of planktonic and suspended particle-associated bacteria in treated water. Correspondingly, Liu et al. (2018) recommended that tap water quality can be improved by both improving the purification steps and by cleaning the DWDS regularly.

In a recent study, Douerelo et al. (2018) used shotgun metagenomic sequencing to evaluate the taxonomic associations and functional aptitude of microbial communities found in chlorinated DWDS from two operational DWDS in the Southwest of the United Kingdom, where one DWDS is fed by surface water and the other one by groundwater. They isolated DNA from 24 samples which were taken from six bulk water and six biofilm samples at each sampling site. The shotgun metagenomic analysis showed that all domains of life (i.e., prokaryotes, eukaryotes, archaea, and viruses) are diversely present in the DWDS which is consistent with all previous metagenomics studies in DWDS. The researchers noted that the identification of metazoan DNA does not imply that the actual organisms are in the samples, but it can be used to indicate an ingress, e.g., free DNA released from animals or plants into the original source water or hydraulically introduced ingress. They concluded that limiting the entry of organic matter in the system can be an approach to inhibit the growth of biofilms in the system. Additionally, the researchers suggested that understanding the mechanism of biofilm formation can bring about the capacity to create the environmental conditions which favor the growth of infrastructure-protective extracellular polymeric substances (EPS) or exterminate pathogens. While the genus *Pseudomonas* has been used to indicate biofilm formation, they recommended the use of alternative bio-indicators of corrosion or biofilm formation in DWDS such as *Bacteroidetes*. Further studies on the microbial dynamics in DWDS which support the findings in the abovementioned studies are presented in Bae et al. (2019); Dai et al. (2019); Dias et al. (2019); Erdogan et al. (2019); Kori et al. (2019); Perrin et al. (2019); Brumfield et al. (2020); Maguvu et al. (2020); Siedlecka et al. (2020); Vavourakis et al. (2020); Atnafu et al. (2021); Bian et al. (2021); Kennedy et al. (2021), and Sevillano et al. (2021). A summary of the results of these studies is provided in **Supplementary Table 1**.

The aforementioned literature review has shown the applicability of metagenomics analysis to understand the role of spatial and temporal distribution and to study the dynamics

of microbial communities in DWDS. A number of genetic markers can be identified for monitoring the variation in the communities that in turn provide the health status of DWDS. There are many ongoing research projects that are built on these findings allowing the development of monitoring systems using predictive models based on the variation in the relative abundance of genetic markers and on recent advances in data science, statistical learning, and artificial intelligence.

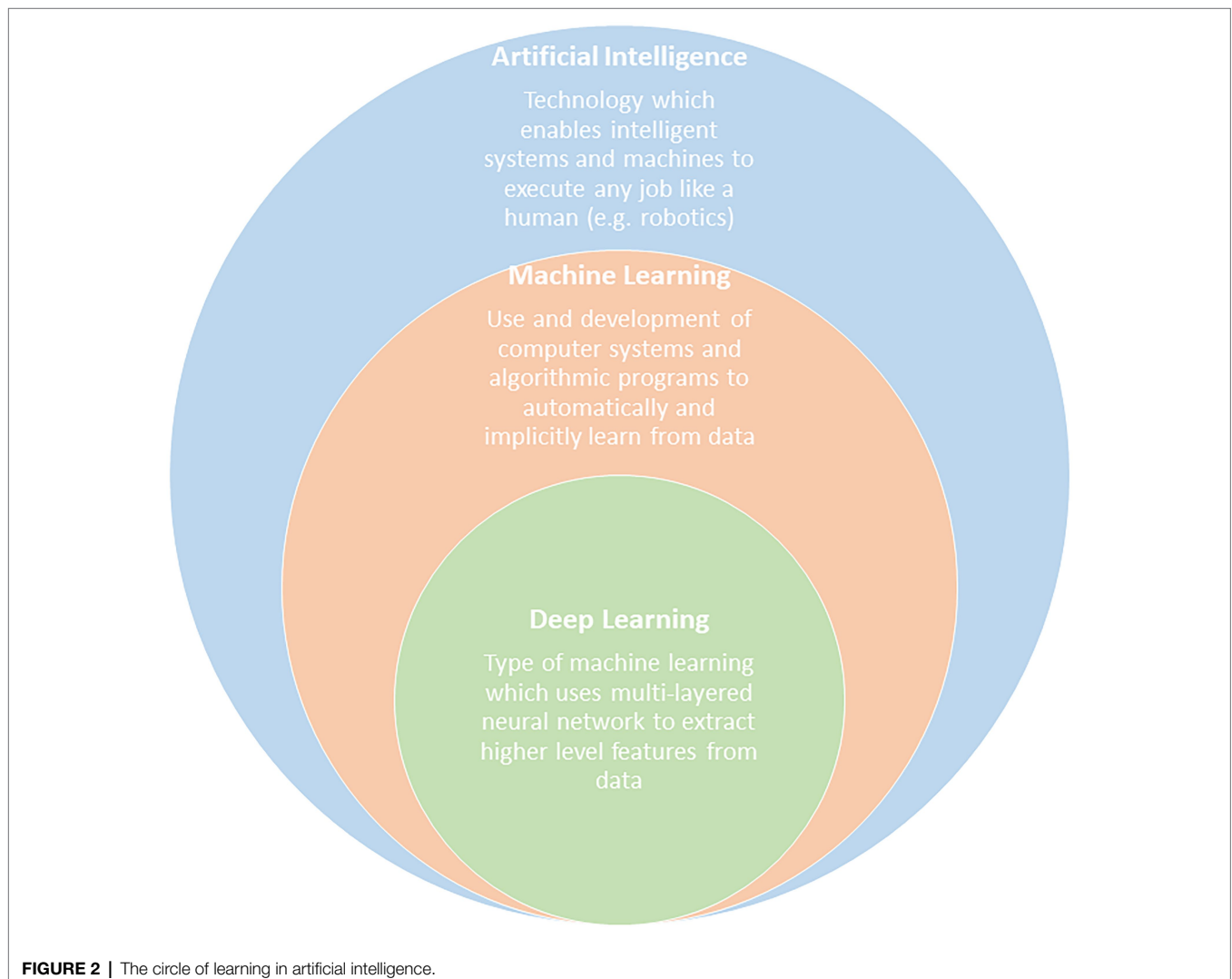
ARTIFICIAL INTELLIGENCE METHODS IN DWDS

Current Lines of Enquiry on Microbial Dynamics

In previous sections, a literature overview has been presented on the use of metagenomics data which have been collected over extended periods and analyzed temporally and spatially to understand the dynamics of microbial communities in DWDS. These works addressed mostly confirmatory and

exploratory questions corresponding to the use of metagenomics data for the analysis of DWDS. From a confirmatory angle, the results so far have addressed the questions on the associations between seasonality/location/type of source water/kind of disinfectant/treatment processes and different environmental parameters on the microbial community composition and structure found in DWDS. From an exploratory angle, research works hitherto have addressed the question of which factors influence most prevalently the microbial dynamics observed in DWDS. As the next step in data science, where data are used to answer predictive questions, there are currently many ongoing research activities where metagenomics data are analyzed for decision-making processes, such as process control and risk mitigation. These works involve the development of predictive models of DWDS that are enriched by real-time information of microbial communities' activities from metagenomics data.

The field of machine learning, which is encompassed by the field of artificial intelligence (AI), can be used to process metagenomics data into meaningful information that can enrich predictive models of DWDS. **Figure 2** shows the circles of



learning methods within the AI field that incorporate recent advances in machine learning and deep learning. Based on the data structure, problem formulation, and the machine learning algorithm used, data science can address different aspects of control and optimization of DWDS and the quality monitoring thereof. In this regard, machine learning can be deployed for four categories of application in DWDS: modelling microbial network interactions, prediction and forecasting of microbial and chemical water quality, decision support for maintenance and operation, and system optimization.

Addressing any type of question using data for different applications can be done through the use of three main types of machine learning: unsupervised learning, supervised learning, and reinforced learning. Unsupervised machine learning algorithms aim to identify meaningful patterns in the data by looking for hidden features in the unlabeled dataset and inferring clusters, accordingly. The use of such algorithms to answer questions regarding prevalence clusters within the microbial communities of drinking water has been illustrated by Pinto et al. (2014) as mentioned above. K-means clustering, Neural Networks (NN), and Principal Component Analysis (PCA) are some of the unsupervised machine learning approaches which are used for solving clustering problems. Supervised machine learning algorithms are deployed on labeled training data sets to make predictions. Classification problems are problems where supervised machine learning algorithms can be used to predict which category something falls into. Naive Bayes Classifier, Support Vector Machines (SVM), Logistic Regression, and Neural Networks are some of the approaches that can be deployed to solve classification problems. In the DWDS case, Liu et al. (2018) used the Bayesian “SourceTracker” method to assess the origin of bacteria in tap water and distribution system. Supervised machine learning algorithms can also be used to solve regression problems, for instance, in making predictions on a continuous scale. Various regression methods (linear, nonlinear, or Bayesian) using nonlinear static, dynamic, or spatially distributed models, can be used in these cases. Negara et al. (2019) has used SVM to solve a nonlinear regression problem that maps metagenomics data from a wastewater treatment plant into the process parameters. Finally, reinforcement learning algorithms use feedback-based learning algorithms where actions and rewards are defined, involving the decision-making agent and environment, in order to maximize a given utility/value function.

Knowledge Gaps and Latent Potential for the Discovery of Novel Lineages

In a study conducted by the United States Environmental Protection Agency (US EPA) which aimed at identifying microbial communities in drinking water by analyzing 16S rRNA-based clone libraries, the researcher found a majority mounting to 57.6% of the sequences belongs to the category of difficult-to-classify bacteria. The researchers observed that 44% of these difficult-to-classify sequences were closely related to sequences retrieved from preceding genomics-based drinking water studies. Thus, these hard-to-classify sequences are most likely indicative of novel lineages which are characteristic of the drinking water

microbiome and may play vital roles in drinking water biogeochemical processes (Revetta et al., 2010). As a consequence of this knowledge gap, light must be shed on the limitations of any artificial-intelligence-based models that use metagenomics data because the performance of any data-driven mathematical model depends on the quality of data it is fed (Sessions and Valtorta, 2006; Alves et al., 2021; Sambasivan et al., 2021).

In their opinion paper Hull et al. (2019), highlight that research in the field of drinking water (DW) microbiome is lagging behind compared to research advancements in the fields of the human microbiome, and environmental microbiomes. Thus, they suggest that the field of DW microbiome can benefit greatly from combining efforts for building a DW microbe project (DWMP). By going in the footprints of other genome databases, the field of DW microbiome can benefit from enriching a central database to include within-species resolution data. In addition, further whole-genome sequencing of DW samples can tackle the issue of unclassified/unknown/sequences (Hull et al., 2019).

Future Technology in DWDS: Meta-Transcriptomics

Meta-transcriptomics (RNA) data introduces additional dimensionality into the mathematical problem formulation that machine learning algorithms can accommodate to address questions regarding functionality. Meta-transcriptomics transcends metagenomics data analysis, where in addition to identifying the microbial communities in DWDS, it can provide information on the functions of each organism (functional metagenomics). One of the advantages of meta-transcriptomics is its ability to differentiate between the active part of a microbial community from the total community which can be quite distinct from one another. The extra knowledge on the functions of species in the microbial community in drinking water can provide valuable information for better understanding the metabolic pathways that are expressed in the bacteria that are present in the aquatic environment of drinking water. The information can be used by operators to deploy appropriate control actions that inhibit undesired metabolism and promote favorable ones (e.g., the metabolic pathways to convert major and minor carbon sources or specific compounds like pollutant degradation). Researchers in the medical field previously showed that meta-transcriptomics can provide a high-resolution picture of the microbiome's functional dynamics (Lavelle and Sokol, 2018). From a meta-omics point of view, it is envisioned that meta-transcriptomics will be crucial for the next step in an obtaining accurate understanding of microbial communities' activities in DWDS.

DISCUSSION

Metagenomics analysis of DWDS has revealed that high-resolution spatial and long-term temporal metagenomics data of DWDS provide insights on the variation of microbial communities under different environmental conditions. A group of genetic markers can subsequently be identified to monitor the dynamic changes in the drinking water microbiome. The ability to forecast the

spatial distribution and temporal dynamics of a drinking water bacterial community can make water quality monitoring more cost-effective, contribute to public health safety by ensuring a safe water supply and increase the performance of process control strategies. Knowing the normal conditions for the operation of the system in its steady-state allows for finding anomalies and invasive pathogens faster. While in all the aforementioned literature (**Supplementary Table 1**), metagenomics data has been effectively collected over extended periods and analyzed to understand the dynamics of microbial water quality in both wastewater treatment plants and water distribution systems, the data analysis has been limited to correlation analysis of available process data. An integrated approach that combines the meta-genomic data with predictive kinetic-mechanistic modelling, potentially combined with machine learning techniques, is still lacking. Consequently, current and future research directions should aim towards the development of a new approach using machine learning techniques to interpret DNA and RNA Next Generation Sequencing (NGS) data in combination with chemical and physical process knowledge to form the basis of a deeper understanding and prediction of the biological and chemical processes in the DWDS. It will transcend metagenomics into functional metagenomics in the drinking water management systems.

AUTHOR CONTRIBUTIONS

ID, G-JE, KK, and BJ contributed to the ideation, design, review, and editing of the literature review paper. AM wrote the manuscript and contributed to the ideation and the design

of the paper. All authors contributed to the article and approved the submitted version.

FUNDING

The work is co-funded by the Dutch Ministry of Economic Affairs and Climate Policy, the European Union Regional Development Fund, the Province of Fryslân, and the Northern Netherlands Provinces.

ACKNOWLEDGMENTS

This work was performed in the cooperation framework of Wetsus, European Centre of excellence for sustainable water technology (www.wetsus.nl). Wetsus is co-funded by the Dutch Ministry of Economic Affairs and Climate Policy, the European Union Regional Development Fund, the Province of Fryslân and the Northern Netherlands Provinces. The authors would like to thank the participants of the research theme “Genomics Based Water Quality Monitoring” for the fruitful discussions and their financial support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmicb.2022.832452/full#supplementary-material>

REFERENCES

- Allion, A., Lassiaz, S., Peguet, L., Boillot, P., Jacques, S., Peultier, J., et al. (2011). A long term study on biofilm development in drinking water distribution system: comparison of stainless steel grades with commonly used materials. *Revue de Métallurgie* 108, 259–268. doi: 10.1051/metal/2011063
- Alves, V. M., Auerbach, S. S., Kleinstreuer, N., Rooney, J. P., Muratov, E. N., Rusyn, I., et al. (2021). Curated data in - trustworthy in silico models out: the impact of data quality on the reliability of artificial intelligence models as alternatives to animal testing. *Altern. Lab. Anim* 49, 73–82. doi: 10.1177/02611929211029635
- Atnafu, B., Desta, A., and Assefa, F. (2021). Microbial community structure and diversity in drinking water supply, distribution systems as well as household point of use sites in Addis Ababa City, Ethiopia. *Microb. Ecol.* doi: 10.1007/s00248-021-01819-3 [Epub ahead of print].
- Bae, S., Lyons, C., and Onstad, N. (2019). A culture-dependent and metagenomic approach of household drinking water from the source to point of use in a developing country. *Water Res. X* 2:100026. doi: 10.1016/j.wroa.2019.100026
- Behera, B. K., Dehury, B., Rout, A. K., Patra, B., Mantri, N., Chakraborty, H. J., et al. (2021). Metagenomics study in aquatic resource management: recent trends, applied methodologies and future needs. *Gene Rep.* 25:101372. doi: 10.1016/j.genrep.2021.101372
- Berney, M., Vital, M., Hülshoff, I., Weilenmann, H. U., Egli, T., and Hammes, F. (2009). Rapid, cultivation-independent assessment of microbial viability in drinking water. *Water Res.* 43:2567. doi: 10.1016/j.watres.2009.03.032
- Bian, K., Wang, C., Jia, S., Shi, P., Zhang, H., Ye, L., et al. (2021). Spatial dynamics of bacterial community in chlorinated drinking water distribution systems supplied with two treatment plants: an integral study of free-living and particle-associated bacteria. *Environ. Int.* 154:106552. doi: 10.1016/j.envint.2021.106552
- Blokke, E. J. M., Furnass, W. R., Machell, J., Mounce, S. R., Schaap, P. G., and Boxall, J. B. (2016). Relating water quality and age in drinking water distribution systems using self-organising maps. *Environment* 3, 1–17. doi: 10.3390/environments3020010
- Brumfield, K. D., Hasan, N. A., Leddy, M. B., Cotruvo, J. A., Rashed, S. M., Colwell, R. R., et al. (2020). A comparative analysis of drinking water employing metagenomics. *PLoS One* 15, 1–27. doi: 10.1371/journal.pone.0231210
- Cerrato, J. M., Falkinham, J. O., Dietrich, A. M., Knocke, W. R., McKinney, C. W., and Pruden, A. (2010). Manganese-oxidizing and -reducing microorganisms isolated from biofilms in chlorinated drinking water systems. *Water Res.* 44, 3935–3945. doi: 10.1016/j.watres.2010.04.037
- Chenevey, B. (2022). Water quality modeling in distribution systems. *J. AWWA* 114, 26–33. doi: 10.1002/awwa.1864
- Christensen, S. C., Nissen, E., Arvin, E., and Albrechtsen, H. J. (2011). Distribution of *Asellus aquaticus* and microinvertebrates in a non-chlorinated drinking water supply system – effects of pipe material and sedimentation. *Water Res.* 45, 3215–3224. doi: 10.1016/j.watres.2011.03.039
- Dai, Z., Sevillano-Rivera, M. C., Calus, S. T., Bautista-de los Santos, Q. M., Murat Eren, A., van der Wielen, P. W. J. J., et al. (2019). Disinfection exhibits systematic impacts on the drinking water microbiome. *Microbiome* 8:42. doi: 10.1186/s40168-020-00813-0
- de Moel, P. J., Verberk, J. Q. J. C., and van Dijk, J. C. (2006). *Drinking Water: Principles and Practices*. Singapore: World Scientific Publishing.
- Dias, V. C. F., Durand, A.-A., Constant, P., Prévost, M., and Bédard, E. (2019). Identification of factors affecting bacterial abundance and community structures in a full-scale chlorinated drinking water distribution system. *Water* 11:627. doi: 10.3390/w11030627
- Douterelo, I., Calero-Preciado, C., Soria-Carrasco, V., and Boxall, J. B. (2018). Whole metagenome sequencing of chlorinated drinking water distribution

- systems. *Environ. Sci. Water Res. Technol.* 4, 2080–2091. doi: 10.1039/C8EW00395E
- Eichler, S., Christen, R., Hölting, C., Westphal, P., Bötel, J., Brettar, I., et al. (2006). Composition and dynamics of bacterial communities of a drinking water supply system as assessed by RNA- and DNA-based 16S rRNA gene fingerprinting. *Appl. Environ. Microbiol.* 72, 1858–1872. doi: 10.1128/AEM.72.3.1858-1872.2006
- Erdogan, I. G., Mekuto, L., Ntwampe, S. K. O., Fosso-Kankeu, E., and Waanders, F. B. (2019). Metagenomic profiling dataset of bacterial communities of a drinking water supply system (DWSS) in the arid Namaqualand region, South Africa: source (lower Orange River) to point-of-use (O'Kiep). *Data Br.* 25:104135. doi: 10.1016/j.dib.2019.104135
- Gomez, C. K., and Aggarwal, S. (2019). "Overview of drinking water distribution system microbiome and water quality." *Encyclopedia of Water*, 1–17.
- Hijnen, W., Schurer, R., Martijn, B., Bahlman, J., Hoogenboezem, W., and van der Wielen, P. (2014). Removal of easily and more complex biodegradable NOM by full-scale BAC filters to produce biological stable drinking water. *Water Res.* 48, 104–114. doi: 10.1016/j.watres.2013.10.011
- Hull, N. M., Ling, F., Pinto, A. J., Albertsen, M., Jang, H. G., Hong, P. Y., et al. (2019). Drinking water microbiome project: is it time? *Trends Microbiol.* 27, 670–677. doi: 10.1016/j.tim.2019.03.011
- Kennedy, L. C., Miller, S. E., Kantor, R. S., and Nelson, K. L. (2021). Effect of disinfectant residual, pH, and temperature on microbial abundance in disinfected drinking water distribution systems. *Environ. Sci. Water Res. Technol.* 7, 78–92. doi: 10.1039/D0EW00809E
- Kori, J. A., Mahar, R. B., Vistro, M. R., Tariq, H., Khan, I. A., and Goel, R. (2019). Metagenomic analysis of drinking water samples collected from treatment plants of Hyderabad City and Mehran University employees cooperative housing society. *Environ. Sci. Pollut. Res.* 26, 29052–29064. doi: 10.1007/s11356-019-05859-8
- Lavelle, A., and Sokol, H. (2018). Gut microbiota: beyond metagenomics, metatranscriptomics illuminates microbiome functionality in IBD. *Nat. Rev. Gastroenterol. Hepatol.* 15, 193–194. doi: 10.1038/nrgastro.2018.15
- Liu, G., Zhang, Y., Van Der Mark, E., Magic-knezev, A., and Pinto, A. (2018). Assessing the origin of bacteria in tap water and distribution system in an unchlorinated drinking water system by SourceTracker using microbial community fingerprints. *Water Res.* 138, 86–96. doi: 10.1016/j.watres.2018.03.043
- Maguvu, T. E., Bezuidenhout, C. C., Kritzing, R., Tsholo, K., Plaatjie, M., Molale-Tom, L. G., et al. (2020). Combining physicochemical properties and microbiome data to evaluate the water quality of south African drinking water production plants. *PLoS One* 15, 1–21. doi: 10.1371/journal.pone.0237335
- Malla, M. A., Dubey, A., Kumar, A., Yadav, S., Hashem, A., and Allah, E. F. A. (2019). Exploring the human microbiome: the potential future role of next-generation sequencing in disease diagnosis and treatment. *Front. Immunol.* 9:2868. doi: 10.3389/fimmu.2018.02868
- Negara, M. A. P., Cornelissen, E., Geurkink, A. K., Euverink, G. J. W., and Jayawardhana, B. (2019). "Next generation sequencing analysis of wastewater treatment plant process via support vector regression." in *Proceedings of the 1st IFAC Workshop on Control Methods for Water Resource Systems. Vol 52*. (IFAC-PapersOnLine), 37–42.
- Nesme, J., Achouak, W., Agathos, S. N., Bailey, M., Baldrian, P., Brunel, D., et al. (2016). Back to the future of soil metagenomics. *Front. Microbiol.* 7:73. doi: 10.3389/fmicb.2016.00073
- Perrin, Y., Bouchon, D., Delafont, V., Moulin, L., and Héchard, Y. (2019). Microbiome of drinking water: a full-scale spatio-temporal study to monitor water quality in the Paris distribution system. *Water Res.* 149, 375–385. doi: 10.1016/j.watres.2018.11.013
- Pinto, A. J., Schroeder, J., Lunn, M., Sloan, W., and Raskin, L. (2014). Spatial-temporal survey and occupancy-abundance modeling to predict bacterial community dynamics in the drinking water microbiome. *MBio* 5, 1–13. doi: 10.1128/mBio.01135-14
- Polychronopolous, M., Dudley, K., Ryan, G., and Hearn, J. (2003). Investigation of factors contributing to dirty water events in reticulation systems and evaluation of flushing methods to remove deposited particles. *Water Sci. Technol.* 3, 295–306. doi: 10.2166/ws.2003.0117
- Prest, E. (2015). Biological Stability in Drinking Water Distribution Systems. Dissertation. Netherlands: Ipskamp Drukkers.
- Prest, E. I., Hammes, F., van Loosdrecht, M. C. M., and Vrouwenvelder, J. S. (2016). Biological stability of drinking water: controlling factors, methods, and challenges. *Front. Microbiol.* 7:45. doi: 10.3389/fmicb.2016.00045
- Rackauckas, C., Ma, Y., Martensen, J., Warner, C., Zubov, K., Supekar, R., et al. (2020). Universal differential equations for scientific machine learning. *Proc. Natl. Acad. Sci. U. S. A.* 1–55. doi: 10.21203/rs.3.rs-55125/v1
- Revetta, R. P., Pemberton, A., Lamendella, R., Iker, B., and Santo Domingo, J. W. (2010). Identification of bacterial populations in drinking water using 16S rRNA-based sequence analyses. *Water Res.* 44, 1353–1360. doi: 10.1016/j.watres.2009.11.008
- Roeselers, G., Coolen, J., van der Wielen, P. W. J. J., Jaspers, M. C., Atsma, A., de Graaf, B., et al. (2015). Microbial biogeography of drinking water: patterns in phylogenetic diversity across space and time. *Environ. Microbiol.* 17, 2505–2514. doi: 10.1111/1462-2920.12739
- Rossman, L. A. (2000). *Epanet 2 Users Manual*. Cincinnati, Washington, D.C.: U.S. Environmental Protection Agency.
- Rudi, K., Tannæs, T., and Vatn, M. (2009). Temporal and spatial diversity of the tap water microbiota in a norwegian hospital. *Appl. Environ. Microbiol.* 75, 7855–7857. doi: 10.1128/AEM.01174-09
- Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., and Aroyo, L. (2021). "Everyone wants to do the model work, not the data work": data cascades in high-stakes AI." in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, May 8–13, 2021; 1–15.
- Santo Domingo, J. W., Meckes, M. C., Simpson, J. M., Sloss, B., and Reasoner, D. J. (2003). Molecular characterization of bacteria inhabiting a water distribution system simulator. *Water Sci. Technol.* 47, 149–154. doi: 10.2166/wst.2003.0305
- Sartory, D. P. (2004). Heterotrophic plate count monitoring of treated drinking water in the UK: a useful operational tool. *Int. J. Food Microbiol.* 92, 297–306. doi: 10.1016/j.ijfoodmicro.2003.08.006
- Scott, B. A., and Pepper, I. L. (2010). Water distribution systems as living ecosystems: impact on taste and odor. *Environ. Sci. Technol.* 45, 890–900. doi: 10.1080/10934521003709115
- Sessions, V., and Valtorta, M. (2006). "The effects of data quality on machine learning algorithms." in *Proceedings of the 11th International Conference on Information Quality*, November 10–12, 2006; (MIT, Cambridge, MA, USA), 485–498.
- Seth, A., Bachmann, R., Boxall, J., Saul, A., and Edyvean, R. (2004). Characterization of materials causing discoloration in potable water systems. *Water Sci. Technol.* 49, 27–32. doi: 10.2166/wst.2004.0080
- Sevillano, M., Vosloo, S., Cotto, I., Dai, Z., Jiang, T., Santiago Santana, J. M., et al. (2021). Spatial-temporal targeted and non-targeted surveys to assess microbiological composition of drinking water in Puerto Rico following hurricane Maria. *Water Res.* 193:117012. doi: 10.1016/j.watres.2021.117012
- Shang, F., Uber, J. G., and Rossman, L. (2011). *EPANET Multi-Species Extension Software and User's Manual*. Washington, DC: U.S. Environmental Protection Agency.
- Siedlecka, A., Wolf-Baca, M., and Piekarska, K. (2020). Spatiotemporal changes of antibiotic resistance and bacterial communities in drinking water distribution system in Wrocław, Poland. *Water* 12:2601. doi: 10.3390/w12092601
- Slatko, B. E., Gardner, A. F., and Ausubel, F. M. (2018). Overview of next generation sequencing technologies (and bioinformatics) in cancer. *Mol. Biol.* 122:e59. doi: 10.1002/cpm.59
- Srinivasan, R., and Sorial, G. A. (2011). Treatment of taste and odor causing compounds 2-methyl isoborneol and geosmin in drinking water: a critical review. *Environ. Sci. Technol.* 23, 1–13. doi: 10.1016/S1001-0742(10)60367-1
- Sun, H., Shi, B., Lytle, D. A., Bai, Y., and Wang, D. (2014). Formation and release behavior of iron corrosion products under the influence of bacterial communities in a simulated water distribution system. *Environ. Sci. Process Impacts* 16, 576–585. doi: 10.1039/c3em00544e
- Tan, B. F., Ng, C., Nshimiyimana, J. P., Loh, L. L., Gin, K. Y. H., and Thompson, J. R. (2015). Next-generation sequencing (NGS) for assessment of microbial water quality: current progress, challenges, and future opportunities. *Front. Microbiol.* 6:1027. doi: 10.3389/fmicb.2015.01027
- Todini, E., and Rossman, L. A. (2012). Unified framework for deriving simultaneous equation algorithms for water distribution networks. *J. Hydraul. Eng.* 139, 511–526.
- Tokajian, S. T., Hashwa, F. A., Hancock, I. C., and Zalloua, P. A. (2005). Phylogenetic assessment of heterotrophic bacteria from a water distribution system using 16S rDNA sequencing. *Can. J. Microbiol.* 51, 325–335. doi: 10.1139/w05-007

- Vandeputte, M. (2021). What has biochemistry done for us? The journey from next-generation sequencing to personalized medicine? *Biochemist* 43, 4–8. doi: 10.1042/bio_2021_192
- van Lieverloo, J. H. M., van der Kooij, D., and Hoogenboezem, W. (2002). *Encyclopedia of Environmental Microbiology*. New York: John Wiley & Sons.
- Vavourakis, C. D., Heijnen, L., Peters, M. C. F. M., Marang, L., Ketelaars, H. A. M., and Hijnen, W. A. M. (2020). Spatial and temporal dynamics in attached and suspended bacterial communities in three drinking water distribution systems with variable biological stability. *Environ. Sci. Technol.* 54, 14535–14546. doi: 10.1021/acs.est.0c04532
- Vital, M., Dignum, M., Magic-Knezev, A., Ross, P., Rietveld, L., and Hammes, F. (2012). Flow cytometry and adenosine tri-phosphate analysis: alternative possibilities to evaluate major bacteriological changes in drinking water treatment and distribution systems. *Water Res.* 46, 4665–4676. doi: 10.1016/j.watres.2012.06.010
- Vreeburg, J. H., and Boxall, J. B. (2007). Discolouration in potable water distribution systems: a review. *Water Res.* 41, 519–529. doi: 10.1016/j.watres.2006.09.028
- Vreeburg, J. H. G., Schaap, P. G., and van Dijk, J. C. (2004). Particles in the drinking water system: from source to discolouration. *Water Sci. Technol. Water Supply* 4, 431–438. doi: 10.2166/ws.2004.0135
- WHO (2020). Hand Hygiene Day. Available at: <https://www.who.int/news-room/events/detail/2020/05/05/default-calendar/hand-hygiene-day> (Accessed November 15, 2021).
- Williams, M. M., Domingo, J. W. S., Meckes, M. C., Kelty, C. A., and Rochon, H. S. (2004). Phylogenetic diversity of drinking water bacteria in a distribution system simulator. *J. Appl. Microbiol.* 96, 954–964. doi: 10.1111/j.1365-2672.2004.02229.x
- Wooschlager, J. E., Rittmann, B. E., and Piriou, P. (2005). Water quality decay in distribution systems – problems, causes, and new modeling tools. *Urban Water J.* 2, 69–79. doi: 10.1080/15730620500144027
- Zhang, Y., and Liu, W. T. (2019). The application of molecular tools to study the drinking water microbiome—current understanding and future needs. *Crit. Rev. Environ. Sci. Technol.* 49, 1188–1235. doi: 10.1080/10643389.2019.1571351

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.

Copyright © 2022 Mahajna, Dinkla, Euverink, Keesman and Jayawardhana. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Advantages of publishing in Frontiers



OPEN ACCESS

Articles are free to read
for greatest visibility
and readership



FAST PUBLICATION

Around 90 days
from submission
to decision



HIGH QUALITY PEER-REVIEW

Rigorous, collaborative,
and constructive
peer-review



TRANSPARENT PEER-REVIEW

Editors and reviewers
acknowledged by name
on published articles

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org

Contact us: frontiersin.org/about/contact



REPRODUCIBILITY OF RESEARCH

Support open data
and methods to enhance
research reproducibility



DIGITAL PUBLISHING

Articles designed
for optimal readership
across devices



FOLLOW US

@frontiersin



IMPACT METRICS

Advanced article metrics
track visibility across
digital media



EXTENSIVE PROMOTION

Marketing
and promotion
of impactful research



LOOP RESEARCH NETWORK

Our network
increases your
article's readership