frontiers | Frontiers in Microbiology

MINI REVIEW
published: 25 April 2022
doi: 10.3389/fmicb.2022.851450

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

James M. W. R. McElhinney[1], Mary Krystelle Catacutan[2], Aurelie Mawart[1], Ayesha Hasan[1,2] and Jorge Dias[3]*

[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

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

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 JMWR, 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

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; Szafrański 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 Techtmann, 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

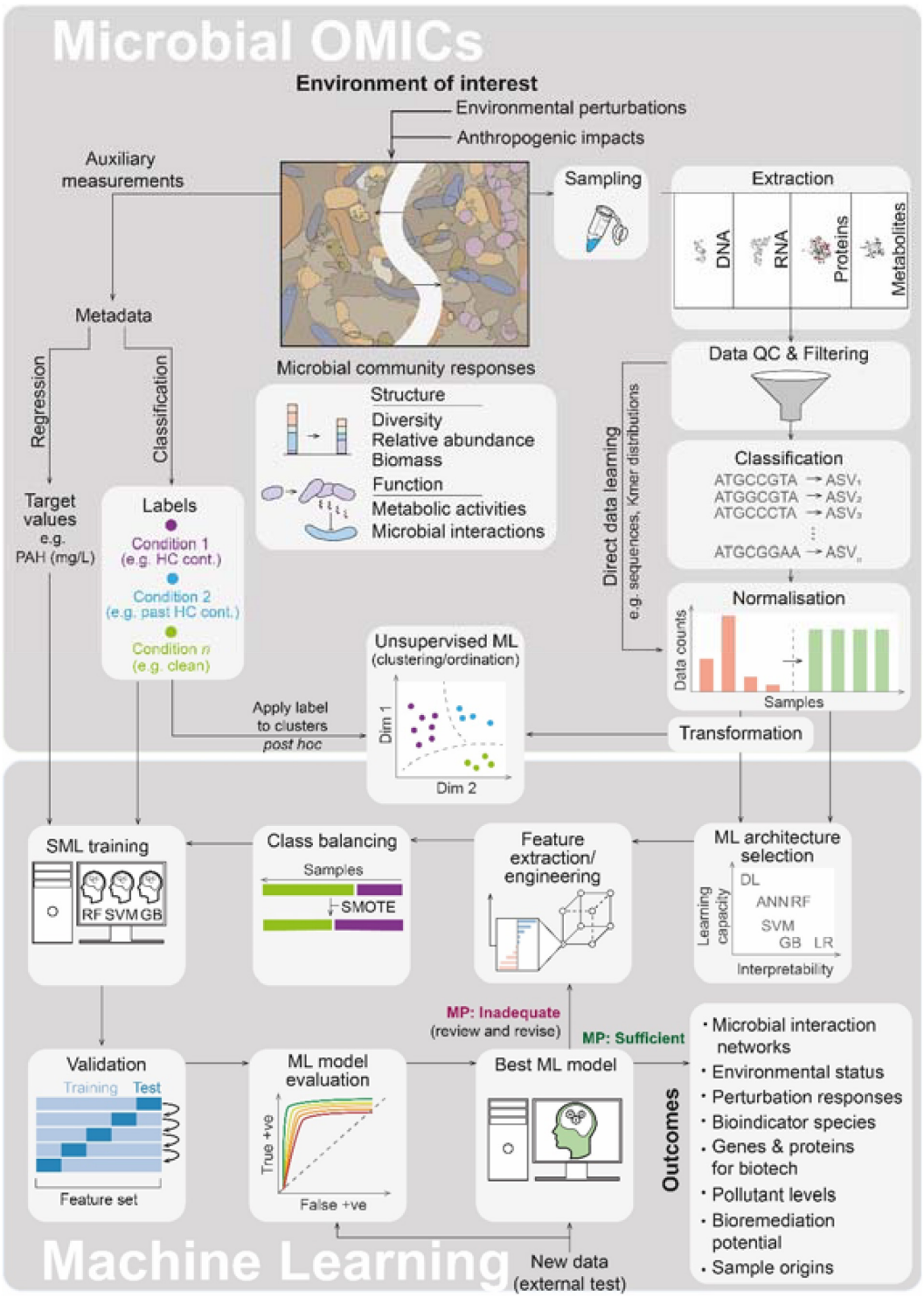

FIGURE 1 | The interface of microbial omics and machine learning (ML). A generalized and simplified overview of the workflows is presented highlighting the major steps in the microbial omics and ML workflows as they relate to one another along with key outcomes obtainable from the application of ML to 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, polyaromatic 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	*Listeria* species	RF	RF R package	Liao et al., 2021

[1] *Indirectly studied in microcosms.*

[2] *Using 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 N_2O 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.2019.0017

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., Haudenshield, 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. Indicat.* 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., Zijp, 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 Techtmann, 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:fuab015.

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

Julinová, M., Vaňharová, L., and Jurča, M. (2018). Water-soluble polymeric xenobiotics – Polyvinyl alcohol and polyvinylpyrrolidon – And potential solutions to environmental issues: A brief review. *J. Environ. Manage.* 228, 213–222. doi: 10.1016/j.jenvman.2018.09.010

Junghare, M., Spiteller, 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., Vrbanac, A., Taylor, B. C., Aksenov, 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/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., Bhuju, 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 Munsky, 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.