CellPress
REVIEWS

Opinion

Embracing Environmental Genomics and Machine Learning for Routine Biomonitoring

Tristan Cordier,[1,*] Anders Lanzén,[2,3] Laure Apothéloz-Perret-Gentil,[1] Thorsten Stoeck,[4] and Jan Pawlowski[1,5]

Genomics is fast becoming a routine tool in medical diagnostics and cutting-edge biotechnologies. Yet, its use for environmental biomonitoring is still considered a futuristic ideal. Until now, environmental genomics was mainly used as a replacement of the burdensome morphological identification, to screen known morphologically distinguishable bioindicator taxa. While prokaryotic and eukaryotic microbial diversity is of key importance in ecosystem functioning, its implementation in biomonitoring programs is still largely unappreciated, mainly because of difficulties in identifying microbes and limited knowledge of their ecological functions. Here, we argue that the combination of massive environmental genomics microbial data with machine learning algorithms can be extremely powerful for biomonitoring programs and pave the way to fill important gaps in our understanding of microbial ecology.

Biomonitoring Aquatic Environments in a Changing World

Aquatic ecosystems, including coastal waters, rivers, and lakes, provide numerous ecosystem services. These include, for example, aquaculture, fisheries, transportation, power generation, water supply for agriculture and human consumption, pollution dilution, exploitation of natural resources such as oil and gas, and numerous cultural and recreational services [1,2]. Due to the ever-growing anthropogenic pressures on aquatic ecosystems, their ability to sustain ecological communities and ecosystem services can be severely impacted [1,3]. Because there is a trade-off between acceptable environmental impact of ecosystem exploitation and socioeconomic benefits, international regulatory systems for sustainable industrial development with minimal environmental impacts are in place worldwide [2,4]. These regulatory systems are laid down in a variety of national and international directives governed by different national and international authorities. The backbone of environmental monitoring programs for aquatic habitats is the biological component (Box 1) [5,6]. Biological indicators are affected by the total range of environmental parameters that they are exposed to. As opposed to chemical and hydrological monitoring techniques [7], which provide an environmental quality snapshot, biological indicators provide a cumulative measure of ecosystem health resulting from the combined responses of these communities to all of the disturbances that they have been experiencing [3,8].

The Conventional Morphology-Based Biomonitoring Tools

Based on the well-studied predictable response of **biological quality elements (BQEs)** (see Glossary) to environmental changes, several **biotic indices (BIs)** for the assessment of **ecological quality status (EQS)** were developed and are applied worldwide to monitor aquatic habitats according to national and international regulations (Figure 1, Key Figure; Box 1). In marine coastal environments, the BIs – such as AMBI, NSI, BENTIX [9–11] – are based on benthic macrofauna composition. In freshwater ecosystems, the commonly targeted groups

Highlights

Environmental genomics targeting microbial communities offers an accurate and cost-effective alternative to conventional biomonitoring based on large-size morphologically identified bioindicators.

Machine learning algorithms are the most promising approach to establish a new routine biomonitoring framework because they allow bypassing the current biological and technical limits.

Microbial metabarcoding combined with machine-learning approaches will allow scaling-up both spatial and temporal resolution for larger and more ambitious biomonitoring programs.

[1]University of Geneva, Department of Genetics and Evolution, 1211 Geneva, Switzerland
[2]AZTI, Marine Research Division, Herrera Kaia, Portualdea z/g, 20110 Pasaia, Basque Country, Spain
[3]IKERBASQUE, Basque Foundation for Science, Bilbao, Spain
[4]University of Kaiserslautern, Ecology Group, D-67663 Kaiserslautern, Germany
[5]ID-Gene ecodiagnostics, Campus Biotech, Avenue Sécheron 15, 1202 Geneva, Switzerland

*Correspondence:
tristan.cordier@gmail.com,
tristan.cordier@unige.ch (T. Cordier).

https://doi.org/10.1016/j.tim.2018.10.012

Box 1. Biomonitoring and Biotic Indices

The backbone of environmental monitoring programs is the biological component [5,6]. Accordingly, the ecological status of surface waters is classified by using biological quality elements (BQEs) as key factors by the European Union (Water Framework Directive, WFD, Directive 2000/60/EC[i] and Marine Strategy Framework Directive, MSFD, Directive 2008/56/EC[ii]), by the USA (the Clean Water Act of the US Environmental Protection Agency[iii]) and by the United Nations Convention on the Law of the Sea (UNCLOS)[iv]. Biological indicators are affected by the total range of environmental parameters to which they are exposed. In comparison with chemical and hydrological monitoring techniques [7], they provide a cumulative measure of ecosystem health resulting from the combined responses of these communities to all of the stresses that they are exposed to [8]. Therefore, the use of biological communities to measure ecological quality is recognized as being a more effective and integrative way than the use of physicochemical variables alone [91,92]. Bioindicator systems are based on the predictable response of specific non-opportunistic organisms to physicochemical changes of the environment. In aquatic ecosystems, traditionally, fish and amphibians, macroinvertebrates, macrophytes, and diatoms are used as bioindicators. Typically, the identification of a bioindicator is based on significant correlations between a specific organism and a set of environmental variables. Based on the direction and strength of this relation and the characteristics of the environmental parameters, the bioindicator is usually assigned an **indicator value (IndVal)** [93] or an ecological weight or category [9]. The IndVal and the (relative) abundance of bioindicators present in a biological community are the basis for the calculation of a biotic index (BI) for ecological quality status (EQS) assessments of a specific habitat. A multitude of BIs has been developed and applied in recent decades, which are highly specialized for a particular type of environment, environmental stress, geographic region, and national legislative requirements [3].

are fish, macroinvertebrates, and diatoms [3,12]. The calculation of BIs relies mainly on morphotaxonomic identification of collected BQEs and their abundances [13]. As summarized previously [13–16], this traditional approach has a number of substantial shortcomings: (i) the biodiversity changes are monitored based on a limited number of morphologically identifiable taxa; (ii) accurate diagnostics of species and their abundances is hampered by a combination of variable taxonomic skills and numerous cryptic species, that is, undifferentiated morphologies, damaged specimens, and unidentifiable early-life stages [17–19]; and (iii) morphological identification is labor intensive and requires specialized experts in taxonomy [17]. This leads to long time lags between sampling and reporting of results, high costs, and a slow, nonautomatable sample throughput. Moreover, taxonomic expertise for species identification is declining worldwide, and so is the ability to train new experts in this field. Because relatively small numbers of samples can be processed due to these constraints, the traditional monitoring approach has low upscaling potential to detect significant biodiversity changes and to mirror the quality of the whole ecosystem under study.

As a result of the growing human pressure on aquatic ecosystems, an increasing demand for monitoring campaigns, and the shortcomings of morphology-based species identification, regulating agencies have highlighted the need to develop more cost-efficient, faster, and reliable tools for EQS assessments of aquatic habitats [20]. Hence, environmental genomics tools represent a promising alternative to morphology-based methods to screen bioindicators and meet these requirements.

The Challenges of Inferring Biotic Indices from Metabarcoding Data

The development of **DNA barcoding** and **environmental DNA (eDNA) metabarcoding** prompted the use of DNA sequences rather than morphology to identify species in biomonitoring surveys [21,22]. This led to a series of studies that explored the potential of metabarcoding for the detection of pollution gradients and for the inference of conventional BIs. The results of these studies showed that metabarcoding data provide a discriminatory power comparable or superior to morphology in terms of correlating community structure with environmental parameters [23–27]. Furthermore, several studies inferring BI from taxonomically assigned metabarcoding data showed that the obtained EQS assessments significantly correlate with the ones obtained by the traditional morphology-based approach. This is exemplified in both

Glossary

Biological quality elements (BQEs): a selection of taxonomic groups which are used for the assessment of ecological quality status (EQS). Depending on the water body type, the BQEs include ‘phytoplankton’, ‘diatoms’, ‘aquatic flora’, ‘macroalgae and angiosperms’, ‘benthic invertebrates’, and ‘fish fauna’.

Biotic indices (BIs): continuous biological metrics that classify an environment based on taxonomic richness, composition, abundance, and ecological functions, in comparison to reference conditions. Ecological quality status (EQS) assessment relies on those BIs and includes several ordered classes, usually from ‘poor’ to ‘high’.

DNA barcoding: a molecular method that uses a short genetic marker, that is taxonomically informative, to identify a particular species.

Ecological quality status (EQS): discrete variables, usually ordered from ‘poor’ to ‘high’, that refer to the ‘ecological quality’ of a water body.

Environmental DNA (eDNA) metabarcoding: diversity survey of PCR-selected taxa present in environmental DNA samples, through high-throughput amplicon sequencing.

Environmental RNA metatranscriptomics: high-throughput sequencing of retrotranscribed RNA. Unlike metabarcoding, there is no taxonomic selection, and it provides a snapshot of the functional activities at the time of sampling.

Indicator value (IndVal): quantify the fidelity and specificity of a species in relation to an individual or a set of environmental parameters.

Morphospecies: a pragmatic and operational definition of closely related individuals sharing similar morphological traits.

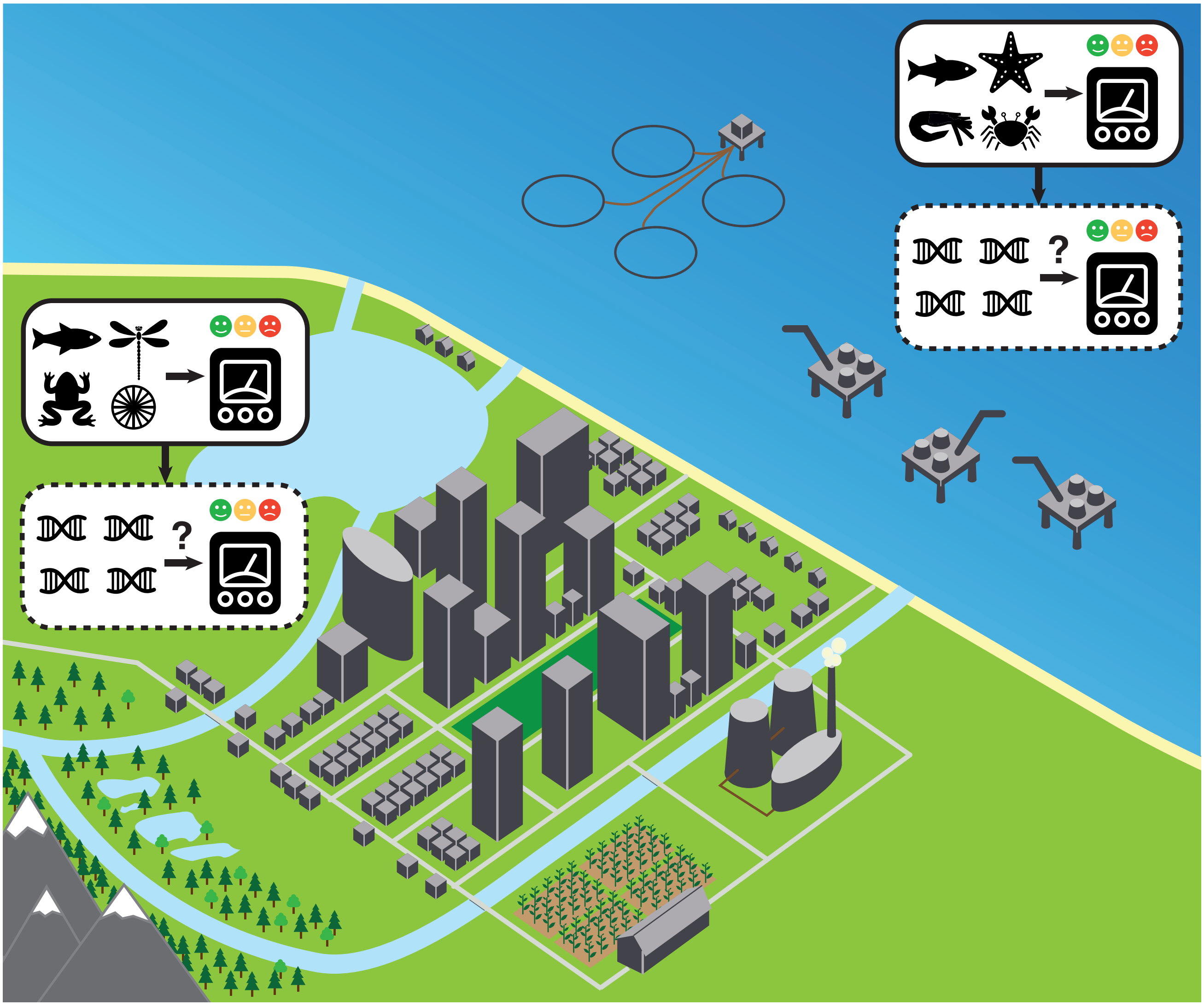
Operational taxonomic unit (OTU): a cluster of similar DNA sequences, obtained from metabarcoding data, that are considered as a proxy for molecular species.

Polymerase chain reaction (PCR): a molecular biology technique used to exponentially amplify a single copy or a few copies of a specific marine [28–30] and freshwater [31–33] environments. These pilot studies concluded that metabarcoding offers a relatively accurate and cost-effective alternative to morphology-based biomonitoring.

segment of DNA to generate millions of copies of a particular DNA sequence.

Key Figure

Routine Biomonitoring of Freshwater and Marine Ecosystems as Currently Implemented in Regulations and Ongoing Transition towards a DNA-Based Approach



Trends in Microbiology

Figure 1. In both environments, biodiversity monitoring currently relies on the morphological identification of known bioindicator taxa. The question marks in the boxes illustrate the difficulties encountered in using environmental genomics to screen these taxa.

However, these studies also raised several issues that impede a perfect match between BIs inferred from morphology-based surveys and metabarcoding data. Both methods are subject to different biases affecting richness, abundance, and taxonomic composition, which are the principal components of conventional BIs (reviewed in [13]). The richness of **operational taxonomic units (OTUs)** is not analogous to **morphospecies** richness, due to cryptic diversity, intragenomic variation, or the presence of eDNA from dead or inactive organisms [25,34]. The abundance derived from sequence data is influenced not only by the number of individuals but also by the number of cells for multicellular organisms and the copy number of the marker gene in the genomes, which is known to vary widely among eukaryotes [35,36] and to a lesser extent for prokaryotes [37,38]. The taxonomic composition retrieved from metabarcoding data can be biased by the specificity of **PCR** primers and may suffer from a strong sampling effect due to the fact that DNA extractions are typically performed on small amounts of material, making the large-size organisms not well represented in eDNA extracts [39].

Another challenging issue is related to the incompleteness of the reference database, which impedes the assignment of a large part of sequences. Despite sustained efforts to complete DNA barcode databases [40–43], there are still many bioindicator species that have not been sequenced yet. This results in large number of OTUs that remain unclassified or that cannot be taxonomically classified to the depth required for EQS assessments, which usually corresponds to the species or genus ranks. The importance of this issue varies greatly between geographical and ecological regions, taxonomic groups, and molecular markers [44–46]. The proportion of unassigned sequences may reach 90%, depending on the specificity of the selected marker [28,47]. Usually, it is lower if the molecular marker targets specifically one taxonomic group, such as diatoms and ciliates, but even in this case, often less than 50% of sequences can be taxonomically assigned due to the gaps in reference databases [31,48]. Altogether, such results may challenge the practical usefulness of metabarcoding as they do for routine biomonitoring because the majority of generated data remain of limited use for EQS assessment, although the development of OTU-based BIs could partially alleviate this issue (see below).

A Plea for Using Microbial Communities in Biomonitoring

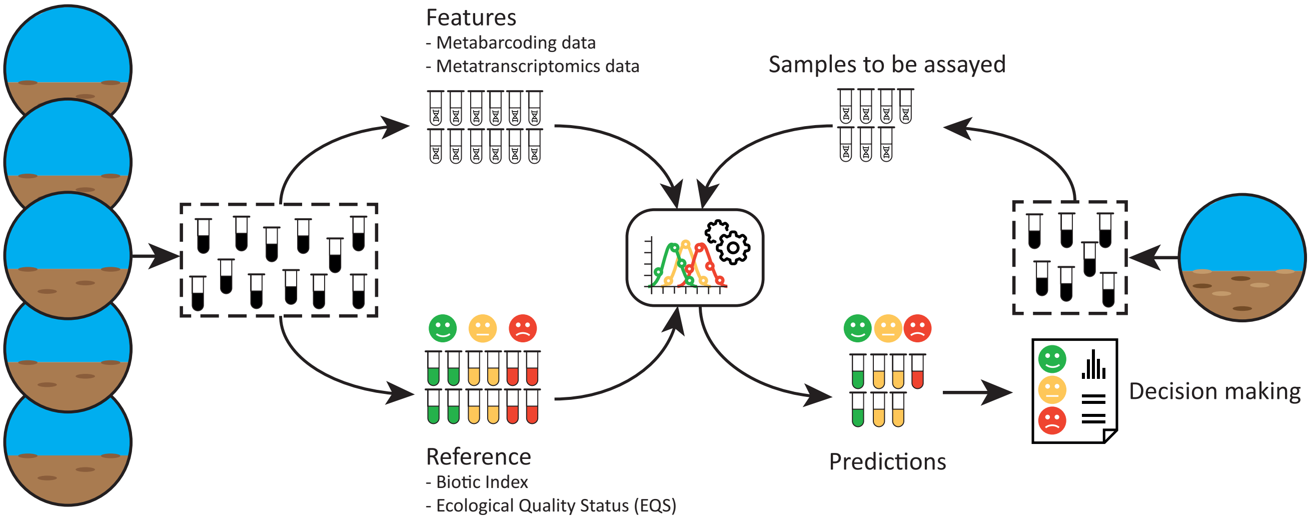
The potential of microorganisms to be used as bioindicators has long been recognized, albeit with different applications for prokaryotic and eukaryotic biomes. High environmental sensitivity and short generation times make them particularly responsive to environmental disturbances, while their functional importance, ubiquity, small size, and high abundance contribute to the technical feasibility and ecological relevance of using them for routine biomonitoring [49], which indeed has already been suggested [13,16]. However, only microbial taxa with species that can be distinguished morphologically are currently used as bioindicators, including diatoms [50,51], foraminifera [52,53], ciliates [54,55], and some bacterial groups [56,57]. Other microbial taxa are largely ignored in conventional biomonitoring surveys because it is too challenging to identify them morphologically and because of the lack of knowledge about their ecological function.

Pilot metabarcoding studies have confirmed the utility of microbial communities for the assessment of ecological impacts. Numerous studies demonstrated the accuracy of the metabarcoding approach to screen conventional microbial bioindicator taxa, such as diatoms [31,32,58–60], foraminifera [24,61–63], and ciliates [48]. Other studies reported changes associated with an anthropogenic impact in global eukaryotic [23,27,64,65] and prokaryotic [27,66–69] diversity surveys. Those molecular studies that compared the response of different taxonomic groups showed that protist or bacterial communities are often as efficient as large-size organisms when used as proxies to detect environmental impacts [25,27].

In view of these studies, we argue that a focus on microbial diversity that largely dominates most ecosystems, and is readily captured by molecular protocols, is likely the best option at hand to perform efficient next-generation biomonitoring. This avenue would indeed overcome both the mismatches of morphological and molecular approaches and the issue of low representativeness of large-size organisms in metabarcoding data. However, it raises another issue related to the interpretation of microbial community data from the EQS assessment perspective, namely that such interpretations require a certain knowledge of the ecological function of identified microbial taxa, which is generally unavailable for most microorganisms. It has been proposed that this issue could be solved by using published reports of mainly higher ranking microbial taxa (e.g., microgAMBI, [70,71]). However, such approaches ignore that distinct ecological adaptations often exist at lower taxonomic ranks, even at the resolution of a single-nucleotide variation [72]. To amend our knowledge regarding ecological signals of microbial OTUs, and to overcome the incompleteness of sequence reference databases, alternative ‘taxonomy-free’ sequence-centered approaches have been proposed recently. Apothéloz-Perret-Gentil *et al*. [60] proposed to give autoecological values to benthic diatom OTUs based on their occurrence in samples of known ecological status. Keeley *et al*. [68] pioneered a similar approach, deriving a ‘multitrophic metabarcoding biotic index’ combining the use of prokaryotes and total eukaryotes. Keck *et al*. [73] envisioned using phylogenetic analyses to impute ecological profiles to diatom OTUs, assuming a niche conservatism. Finally, Cordier *et al*. [74,75] suggested that BI values be predicted by using supervised machine-learning (SML) algorithms (Figure 2).

Combining Environmental Genomics with Machine Learning

In our opinion, machine-learning algorithms (Box 2) are the most promising approach to establish a new routine biomonitoring framework with metabarcoding datasets. First, based



Trends in Microbiology

Figure 2. Proposed Workflow Using a Predictive Model to Make Ecological Quality Status (EQS) Assessment from Environmental Genomics Data. The predictive model is built from training data that include both reference diagnostics and metabarcoding data using a supervised machine learning (SML) algorithm.

Box 2. The Different Types of Problem in Machine Learning (ML)

The field of machine learning aims to develop computer algorithms that can 'learn' from a set of data and improve their performance with experience, to assist humans in big data and complex classification tasks. Such algorithms have proven to be useful in a wide range of problems, such as natural language processing [94], physics [95], game theory [96], computational biology [97,98], biomedical research [87,99], and ecological research [100,101]. Because ML algorithms are best fitted for large and noisy datasets, they are also particularly suited for the analysis of highly dimensional microbial genomics datasets [67,87,89]. ML problems can be broadly divided into two fundamentally different categories, the supervised and the unsupervised ones.

In a supervised problem, the aim is to train a predictive model using a labelled dataset, that is, the solution is known for each provided sample, in order to classify upcoming unlabeled samples. The training of such a model consists in identifying, among a usually large number of features (i.e., the variables that we assume to be important for the problem at hand), the ones (or a combination of them) that correlate or explain the known solutions of the training observations. This extracted knowledge is used by the predictive model, trained on only a subset of the possible real-world situations, to make predictions on upcoming samples.

In an unsupervised problem, nothing is known about the solution, and the aim is to identify a relevant pattern or rules of associations in the data. An intermediate category is the semisupervised machine learning, in which only a subset of samples is labelled within the full dataset, that actually corresponds to most real-world situations, because obtaining a fully labelled dataset is time-consuming and often practically out of reach.

on recent work, we argue that SML combined with metabarcoding would allow us to complement and expand the currently implemented morphology-based methods to produce the same EQS assessment, but in a faster and more cost-effective way. Second, we propose a framework that would incorporate routinely assayed samples into a semisupervised ML approach, to refine the decision boundaries of the current decision-making standards. Finally, we propose to gradually move towards a fully data-driven approach to iteratively update our classification system, which will truly unlock the disruptive potential of metabarcoding for biomonitoring.

The routine biomonitoring of an ecosystem can intuitively be transposed into a supervised classification problem (SML, Figure 2). Indeed, EQS is a discrete variable – easily readable by regulators – that is derived from BI continuous values. Such continuous BI values can serve as a reference for the metabarcoding samples, provided that both molecular and morphotaxonomic data are obtained in parallel on the same samples. Hence, this labelled dataset constitutes a training set for the building of a predictive biomonitoring model (Figure 2). Ideally, the training data covers a wide range of environmental conditions of a given ecosystem in a balanced effort. The predictive model performance has to be tested, using cross-validation, in order to measure the generalization error [76]. A low error usually means that the model can be reliably used for making predictions on upcoming samples. This has recently been tested in marine environments [67,74] and proved to outperform taxonomy-based metabarcoding approach in EQS assessments [75], demonstrating the applicability of SML algorithms for biomonitoring purposes.

Using SML algorithms to build predictive models from labelled metabarcoding data provides several advantages. First, SML is *de facto* eliminating the problems arising from the lack of a taxonomic framework and the dependency on reference databases, because the ecological signal (linear or not) of individual OTUs and association rules within the full community are automatically disentangled from the background noise during the training of the predictive model. Second, as opposed to morphological identification, genomic data are unambiguous and can be easily stored and compared across time and practitioners. Third, the data collection can be standardized and the analysis automatized. Finally, thanks to its cost-effectiveness, such an approach is allowing for scaling-up both spatial and temporal resolution for more ambitious and extensive biomonitoring programs.

These advantages have led us to think that environmental genomics is a mature enough technology to be implemented in routine biomonitoring, provided that we consider it for its own value, instead of trying to make it fit into the current morphotaxonomic standards. Environmental genomics is best fitted to unravel microbial diversity and generate meaningful ecological data for these organisms. Adopting SML means moving from an established, taxonomy-driven paradigm to a sequence-centered, taxonomy-free one. This contrasts with one of the main goals of the metabarcoding approach, namely identifying the pool of species present in a sample from a short taxonomically informative DNA marker [22]. However, the above-mentioned biological and technical limits keep preventing us from using environmental genomics according to the same current biomonitoring system. Therefore, we think that it is more reasonable to completely bypass the issues related to the taxonomic assignment of the sequences by combining the metabarcoding data generated for microbial communities with ML technologies.

This approach could also be extended to microbial **environmental RNA metatranscriptomics** datasets. Collecting functional data from microbial communities provides an additional, potentially more mechanistic, layer of information to taxonomic profiles. In a biomonitoring context, we could expect expressed microbial functions to be even better proxies than taxa to detect a given environmental disturbance, especially because functional redundancy is likely widespread among prokaryotes [77]. However, collecting RNA samples for metatranscriptomics is far more constraining in practice (and less cost-effective) than collecting DNA samples for metabarcoding, due to the instability and conservation issues of RNA molecules. This could make metatranscriptomics very challenging to implement in routine biomonitoring programs.

The ‘Black Box’ Issue in Machine Learning

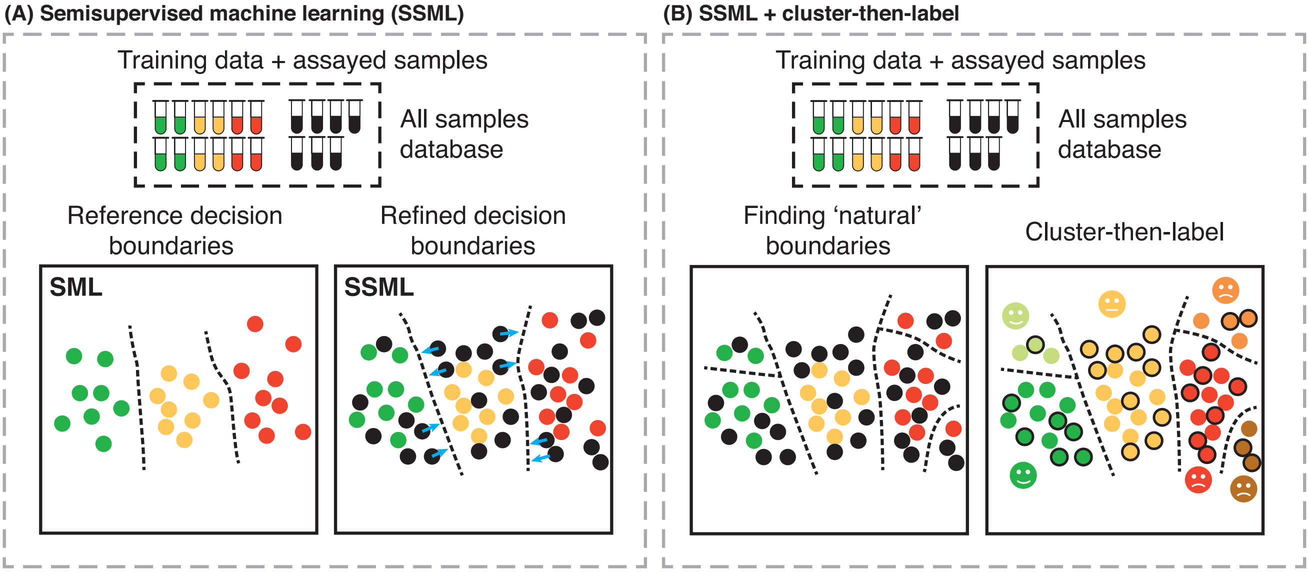
A legitimate concern that may arise from the perspective of applying ML to taxonomy-free metabarcoding data for routine biomonitoring has to do with the perceived opacity of ML algorithms and the trained model they produce [78]. The so-called ‘black box’ problem comes from the difficulty in keeping meaningful human oversight over the algorithm decision-making process [79]. Indeed, the high dimensionality of the data is out of reach for human readability, while easily accomplishable by computers. However, human interpretability, and to some extent understanding of the trained model, appears crucial to operate a smooth transition and routinely adopt ML approaches for complex decision-making problems [80]. This need for ‘trust’ in the model prompted the development of tools to explain the predictions, by identifying the features of the data that contribute the most to a given prediction (see Outstanding Questions) [81]. Such feature importance measurement can also be built into some ML algorithms, such as random forest [82]. With such information alongside the predictions, domain experts would be able to judge the prediction trustworthy or not, and by extension, the trained model [80]. For instance, recent studies testing the usefulness of ML confirmed the strength of established biomarkers, but also identified new powerful ones by isolating sequences with a high predictive power [67,75]. Hence, adopting ML is not only solving biological and technical issues for routine biomonitoring, it is also ecologically meaningful because the trained model can be interpreted in terms of features that matter the most. The model can therefore be controlled by biologists and ecologists, without falling into the ‘black box’ issue.

Concluding Remarks: Toward Environmental Genomics Data-Driven Biomonitoring

In most ML problems, the dataset is not fully labelled because a reference diagnostic for all samples is not always available. The analysis of such datasets falls into the semisupervised category of ML problems (SSML). Integrating unlabeled samples within the training set can improve the performance of a predictive model [83,84]. The rationale for this improvement is that unlabeled samples allow coverage of wider conditions within the studied system (sampled space) and hence increase the ability of ML algorithms to identify meaningful patterns. This reshapes the existing classification boundaries into more data-driven clusters (Figure 3A) [85].

In the case of biomonitoring, the routinely assayed samples could complement the training data as unlabeled samples and may increase the accuracy of a trained model to classify upcoming samples. Such strategies are based on the assumption that samples close to each other in the data space are more likely to share the same label (but see [86]), even though other processes, such as biotic interactions, random demographic drift, or dispersal limitation may interact with the anthropogenic pressures in the assembly of communities. Nevertheless, predictive models remain accurate over regional geographic distance [75], meaning that the data contain at least some dimensions that can be exploited by a trained model to make accurate predictions. We believe that incorporating unlabeled data can improve ML-based monitoring, and suggest 'blind tests' to measure the effect of incorporating unlabeled samples on the accuracy of predictions as a priority for future studies.

With time, a dataset including both reference and assayed samples would cover most of the possible environmental conditions along a gradient from reference sites to impacted ones and the numerous OTUs encountered in the studied ecosystem. At this point, it would be possible to use pattern-recognition ML algorithms to find the optimal boundaries between redefined environmental states. Hence, we could explore the extent to which the current biomonitoring classification matches these new classes (Figure 3B) and possibly update our management decision-making criteria.


(A) Semisupervised machine learning (SSML)
Training data + assayed samples
All samples database
Reference decision boundaries
Refined decision boundaries
SML
SSML
(B) SSML + cluster-then-label
Training data + assayed samples
All samples database
Finding 'natural' boundaries
Cluster-then-label

Trends in Microbiology

Figure 3. Proposed Machine Learning Approaches for Next-Generation, More Data-Driven, Routine Biomonitoring. (A) Within an ongoing supervised machine learning (SML) biomonitoring program, the routinely assayed samples are included as unlabeled samples with the labeled dataset into a semisupervised machine learning (SSML) approach. This strategy may improve the accuracy of ecological quality status (EQS) assessment by increasing the sampled space and hence the ability of the algorithm to better capture patterns in the data. Such an approach may refine the decision boundaries of the current classification system. (B) Eventually, when the ecosystem under study has been sufficiently sampled, covering various ecological conditions and associated biological diversity, a pattern-recognition algorithm may help us to identify 'natural' clusters in the data and ultimately refine our management decision-making criteria, by labelling these new classes according to the accumulated ecological knowledge.

In fact, microbial DNA-based biomonitoring shares striking analogies with data-driven medicine, and particularly with the human-associated microbiome research. Indeed, both fields are aiming to unravel ecological communities and biomarkers that could be used as powerful diagnostic tools for the detection of various conditions, namely environmental pollution or human diseases [87–90]. Thus, the approach discussed in our paper is largely inspired from microbiome research and proposes to make conceptual and technical bridges between the two disciplines. By taking full advantage of the latest molecular and computational technologies, routine biomonitoring can enter the big data era and transform our understanding of ecological changes in a long-term perspective.

Outstanding Questions

What are the factors that affect the origin, state, lifetime, and transportation of microbial environmental DNA? How might these factors influence the precision of machine-learning-based predictive models?

'Who' are the microbial environmental bioindicators revealed by eDNA metabarcoding studies? What are their ecological functions?

What are the required steps to implement machine-learning-based environmental genomic tools into policy for routine biomonitoring?

Acknowledgments

Tristan Cordier, Jan Pawlowski, and Laure Apothéloz-Perret-Gentil were supported by the Swiss Network for International Studies, by the Swiss National Science Foundation (grant 31003A_179125), and by the European Cross-Border Cooperation Program (Interreg France-Switzerland 2014-2020, SYNAQUA project). Anders Lanzén was supported by the IKERBASQUE – The Basque Foundation for Science. Thorsten Stoeck was supported by the Deutsche Forschungsgemeinschaft (DFG, grant STO414/15-1). We thank the members of the DNAqua-net COST Action CA15219 'Developing new genetic tools for bioassessment of aquatic ecosystems in Europe' funded by the European Union, for helpful discussions.

Resources

[i] https://eur-lex.europa.eu/resource.html?uri=cellar:5c835afb-2ec6-4577-bdf8-756d3d694eeb.0004.02/DOC_1&format=PDF
[ii] http://data.europa.eu/eli/dir/2008/56/oj
[iii] www.epa.gov/laws-regulations/summary-clean-water-act
[iv] www.un.org/depts/los/convention_agreements/convention_overview_convention.htm

References

1. Borja, A. *et al.* (2016) Overview of integrative assessment of marine systems: the ecosystem approach in practice. *Front. Mar. Sci.* 3, 1–20
2. Grizzetti, B. *et al.* (2016) Assessing water ecosystem services for water resource management. *Environ. Sci. Policy* 61, 194–203
3. Birk, S. *et al.* (2012) Three hundred ways to assess Europe's surface waters: an almost complete overview of biological methods to implement the Water Framework Directive. *Ecol. Indic.* 18, 31–41
4. Borja, A. *et al.* (2013) Good environmental status of marine ecosystems: what is it and how do we know when we have attained it? *Mar. Pollut. Bull.* 76, 16–27
5. Borja, A. *et al.* (2009) Current status of macroinvertebrate methods used for assessing the quality of European marine waters: implementing the Water Framework Directive. *Hydrobiologia* 633, 181–196
6. Hering, D. *et al.* (2018) Implementation options for DNA-based identification into ecological status assessment under the European Water Framework Directive. *Water Res.* 138, 192–205
7. Carroll, M.L. *et al.* (2003) Organic enrichment of sediments from salmon farming in Norway: environmental factors, management practices, and monitoring techniques. *Aquaculture* 226, 165–180
8. Lear, G. *et al.* (2011) A comparison of bacterial, ciliate and macroinvertebrate indicators of stream ecological health. *Aquat. Ecol.* 45, 517–527
9. Borja, A. *et al.* (2000) A marine biotic index to establish the ecological quality of soft-bottom benthos within European estuarine and coastal environments. *Mar. Pollut. Bull.* 40, 1100–1114
10. Rygg, B. and Norling, K. (2013) *Norwegian Sensitivity Index (NSI) for Marine Macroinvertebrates, and an Update of Indicator Species Index (ISI),* Norsk institutt for vannforskning ISBN:978-82-577-6210-0
11. Simboura, N. and Zenetos, A. (2002) Benthic indicators to use in ecological quality classification of Mediterranean soft bottom marine ecosystems, including a new Biotic Index. *Mediterr. Mar. Sci.* 3, 77–111
12. Poikane, S. *et al.* (2016) Benthic macroinvertebrates in lake ecological assessment: a review of methods, intercalibration and practical recommendations. *Sci. Total Environ.* 543, 123–134
13. Pawlowski, J. *et al.* (2018) The future of biotic indices in the ecogenomic era: integrating (e)DNA metabarcoding in biological assessment of aquatic ecosystems. *Sci. Total Environ.* 637/638, 1295–1310
14. Danovaro, R. *et al.* (2016) Implementing and innovating marine monitoring approaches for assessing marine environmental status. *Front. Mar. Sci.* 3, 1–25
15. Goodwin, K.D. *et al.* (2017) DNA sequencing as a tool to monitor marine ecological status. *Front. Mar. Sci.* 4, 1–14
16. Pawlowski, J. *et al.* (2016) Protist metabarcoding and environmental biomonitoring: time for change. *Eur. J. Protistol.* 55, 12–25
17. Maurer, D. (2000) The dark side of taxonomic sufficiency (TS). *Mar. Pollut. Bull.* 40, 98–101
18. Bourlat, S.J. *et al.* (2013) Genomics in marine monitoring: new opportunities for assessing marine health status. *Mar. Pollut. Bull.* 74, 19–31
19. Cowart, D.A. *et al.* (2015) Metabarcoding is powerful yet still blind: a comparative analysis of morphological and molecular surveys of seagrass communities. *PLoS One* 10, e0117562
20. Heiskanen, A.-S. *et al.* (2016) Biodiversity in marine ecosystems – European developments toward robust assessments. *Front. Mar. Sci.* Published online September 23, 2016. http://dx.doi.org/10.3389/fmars.2016.00184
21. Baird, D.J. *et al.* (2012) Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Mol. Ecol.* 21, 2039–2044

22. Taberlet, P. *et al.* (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. *Mol. Ecol.* 21, 2045–2050

23. Chariton, A.A. *et al.* (2014) A molecular-based approach for examining responses of eukaryotes in microcosms to contaminant-spiked estuarine sediments. *Environ. Toxicol. Chem.* 33, 359–369

24. Pawlowski, J. *et al.* (2014) Environmental monitoring through protist next-generation sequencing metabarcoding: assessing the impact of fish farming on benthic foraminifera communities. *Mol. Ecol. Resour.* 14, 1129–1140

25. Lanzén, A. *et al.* (2016) High-throughput metabarcoding of eukaryotic diversity for environmental monitoring of offshore oil-drilling activities. *Mol. Ecol.* 25, 4392–4406

26. Andújar, C. *et al.* (2017) Metabarcoding of freshwater invertebrates to detect the effects of a pesticide spill. *Mol. Ecol.* 27, 146–166

27. Laroche, O. *et al.* (2018) A cross-taxa study using environmental DNA/RNA metabarcoding to measure biological impacts of offshore oil and gas drilling and production operations. *Mar. Pollut. Bull.* 127, 97–107

28. Lejzerowicz, F. *et al.* (2015) High-throughput sequencing and morphology perform equally well for benthic monitoring of marine ecosystems. *Sci. Rep.* 5, 13932

29. Aylagas, E. *et al.* (2016) Benchmarking DNA metabarcoding for biodiversity-based monitoring and assessment. *Front. Mar. Sci.* 3, 96

30. Lobo, J. *et al.* (2017) DNA metabarcoding for high-throughput monitoring of estuarine macrobenthic communities. *Sci. Rep.* Published online November 15, 2017. http://dx.doi.org/10.1038/s41598-017-15823-6

31. Visco, J.A. *et al.* (2015) Environmental monitoring: inferring the diatom index from next-generation sequencing data. *Environ. Sci. Technol.* 49, 7597–7605

32. Vasselon, V. *et al.* (2017) Application of high-throughput sequencing (HTS) metabarcoding to diatom biomonitoring: do DNA extraction methods matter? *Freshw. Sci.* 36, 162–177

33. Elbrecht, V. *et al.* (2017) Assessing strengths and weaknesses of DNA metabarcoding-based macroinvertebrate identification for routine stream monitoring. *Methods Ecol. Evol.* 8, 1265–1275

34. Brown, E.A. *et al.* (2015) Divergence thresholds and divergent biodiversity estimates: can metabarcoding reliably describe zooplankton communities? *Ecol. Evol.* 5, 2234–2251

35. Weber, A.A.T. and Pawlowski, J. (2013) Can abundance of protists be inferred from sequence data: a case study of Foraminifera. *PLoS One* Published online February 19, 2013. http://dx.doi.org/10.1371/journal.pone.0056739

36. Bik, H.M. *et al.* (2013) Intra-genomic variation in the ribosomal repeats of nematodes. *PLoS One* Published online October 11, 2013. http://dx.doi.org/10.1371/journal.pone.0078230

37. Klappenbach, J.A. *et al.* (2001) rrndb: the ribosomal RNA operon copy number database. *Nucleic Acids Res.* 29, 181–184

38. Větrovský, T. and Baldrian, P. (2013) The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PLoS One* Published online February 27, 2013. http://dx.doi.org/10.1371/journal.pone.0057923

39. Lanzén, A. *et al.* (2017) DNA extraction replicates improve diversity and compositional dissimilarity in metabarcoding of eukaryotes in marine sediments. *PLoS One* 12, 1–18

40. Ratnasingham, S. and Hebert, P.D.N. (2007) bold: The Barcode of Life Data System (http://www.barcodinglife.org). *Mol. Ecol. Notes* 7, 355–364

41. Quast, C. *et al.* (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41, D590–D596

42. Guillou, L. *et al.* (2013) The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Res.* 41, D597–D604

43. Keck, F. *et al.* (2016) Phylosignal: an R package to measure, test, and explore the phylogenetic signal. *Ecol. Evol.* 6, 2774–2780

44. Kvist, S. (2013) Barcoding in the dark?: A critical view of the sufficiency of zoological DNA barcoding databases and a plea for broader integration of taxonomic knowledge. *Mol. Phylogenet. Evol.* 69, 39–45

45. Aylagas, E. *et al.* (2014) Environmental status assessment using DNA metabarcoding: towards a genetics based marine biotic index (gAMBI). *PLoS One* 9, e90529

46. Rimet, F. *et al.* (2016) R-Syst::diatom: an open-access and curated barcode database for diatoms and freshwater monitoring. *Database* Published online March 17, 2016. http://dx.doi.org/10.1093/database/baw016

47. Chariton, A.A. *et al.* (2015) Metabarcoding of benthic eukaryote communities predicts the ecological condition of estuaries. *Environ. Pollut.* 203, 165–174

48. Stoeck, T. *et al.* (2018) Metabarcoding of benthic ciliate communities shows high potential for environmental monitoring in salmon aquaculture. *Ecol. Indic.* 85, 153–164

49. Payne, R.J. (2013) Seven reasons why protists make useful bioindicators. *Acta Protozool.* 52, 105–113

50. Rimet, F. and Bouchez, A. (2012) Biomonitoring river diatoms: implications of taxonomic resolution. *Ecol. Indic.* 15, 92–99

51. Almeida, S.F.P. *et al.* (2014) Water quality assessment of rivers using diatom metrics across Mediterranean Europe: a methods intercalibration exercise. *Sci. Total Environ.* 476/477, 768–776

52. Schönfeld, J. *et al.* (2012) The FOBIMO (FOraminiferal BIo-MOnitoring) initiative—towards a standardised protocol for soft-bottom benthic foraminiferal monitoring studies. *Mar. Micropaleontol.* 94–95, 1–13

53. Alve, E. *et al.* (2016) Foram-AMBI: a sensitivity index based on benthic foraminiferal faunas from North-East Atlantic and Arctic fjords, continental shelves and slopes. *Mar. Micropaleontol.* 122, 1–12

54. Foissner, W. and Berger, H. (1996) A user-friendly guide to the ciliates (Protozoa, Ciliophora) commonly used by hydrobiologists as bioindicators in rivers, lakes, and waste waters, with notes on their ecology. *Freshw. Biol.* 35, 375–482

55. Zhang, W. *et al.* (2014) Insights into assessing water quality using taxonomic distinctness based on a small species pool of biofilm-dwelling ciliate fauna in coastal waters of the Yellow Sea, northern China. *Mar. Pollut. Bull.* 89, 121–127

56. McIlroy, S.J. *et al.* (2015) MiDAS: the field guide to the microbes of activated sludge. *Database* 2015, bav062

57. Mateo, P. *et al.* (2015) Cyanobacteria as bioindicators and bioreporters of environmental analysis in aquatic ecosystems. *Biodivers. Conserv.* 24, 909–948

58. Kermarrec, L. *et al.* (2014) A next-generation sequencing approach to river biomonitoring using benthic diatoms. *Freshw. Sci.* 33, 349–363

59. Zimmermann, J. *et al.* (2015) Metabarcoding vs. morphological identification to assess diatom diversity in environmental studies. *Mol. Ecol. Resour.* 15, 526–542

60. Apothéloz-Perret-Gentil, L. *et al.* (2017) Taxonomy-free molecular diatom index for high-throughput eDNA biomonitoring. *Mol. Ecol. Resour.* 17, 1231–1242

61. Pawlowski, J. *et al.* (2016) Benthic monitoring of salmon farms in Norway using foraminiferal metabarcoding. *Aquacult. Environ. Interact.* 8, 371–386

62. Pochon, X. *et al.* (2015) Accurate assessment of the impact of salmon farming on benthic sediment enrichment using foraminiferal metabarcoding. *Mar. Pollut. Bull.* 100, 370–382

63. Laroche, O. *et al.* (2016) First evaluation of foraminiferal metabarcoding for monitoring environmental impact from an offshore oil drilling site. *Mar. Environ. Res.* 120, 225–235

64. Bik, H.M. *et al.* (2012) Dramatic shifts in benthic microbial eukaryote communities following the deepwater horizon oil spill. *PLoS One* 7, e38550

65. Lanzén, A. *et al.* (2016) High-throughput metabarcoding of eukaryotic diversity for environmental monitoring of offshore oil-drilling activities. *Mol. Ecol.* 25, 4392–4406

66. Dowle, E. *et al.* (2015) Assessing the effects of salmon farming seabed enrichment using bacterial community diversity and high-throughput sequencing. *FEMS Microbiol. Ecol.* 91, fiv089

67. Smith, M.B. *et al.* (2015) Natural bacterial communities serve as quantitative geochemical biosensors. *mBio* 6, 1–13

68. Keeley, N. *et al.* (2018) Development and validation of a multi-trophic metabarcoding biotic index for benthic organic enrichment biomonitoring. *Ecol. Indic.* 85, 1044–1057

69. Stoeck, T. *et al.* (2018) Environmental DNA metabarcoding of benthic bacterial communities indicates the benthic footprint of salmon aquaculture. *Mar. Pollut. Bull.* 127, 139–149

70. Aylagas, E. *et al.* (2017) A bacterial community-based index to assess the ecological status of estuarine and coastal environments. *Mar. Pollut. Bull.* 114, 679–688

71. Borja, A. (2018) Testing the efficiency of a bacterial community-based index (microgAMBI) to assess distinct impact sources in six locations around the world. *Ecol. Indic.* 85, 594–602

72. Eren, A.M. *et al.* (2013) Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol. Evol.* 4, 1111–1119

73. Keck, F. *et al.* (2018) Boosting DNA metabarcoding for biomonitoring with phylogenetic estimation of OTUs' ecological profiles. *Mol. Ecol. Resour.* 18, 1299–1309

74. Cordier, T. *et al.* (2017) Predicting the ecological quality status of marine environments from eDNA metabarcoding data using supervised machine learning. *Environ. Sci. Technol.* 51, 9118–9126

75. Cordier, T. *et al.* (2018) Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring. *Mol. Ecol. Resour.* 18, 1381–1391

76. Domingos, P. (2012) A few useful things to know about machine learning. *Commun. ACM* 55, 78

77. Louca, S. *et al.* (2018) Function and functional redundancy in microbial systems. *Nat. Ecol. Evol.* 2, 936–943

78. Mittelstadt, B.D. *et al.* (2016) The ethics of algorithms: mapping the debate. *Big Data Soc.* 3, 1–21

79. Burrell, J. (2016) How the machine 'thinks': understanding opacity in machine learning algorithms. *Big Data Soc.* 3, 1–12

80. Miotto, R. *et al.* (2017) Deep learning for healthcare: review, opportunities and challenges. *Brief. Bioinform.* Published online May 6, 2017. http://dx.doi.org/10.1093/bib/bbx044

81. Ribeiro, M.T. *et al.* (2016) 'Why Should I Trust You?': explaining the predictions of any classifier. *arXiv* February 16, 2016. https://arxiv.org/abs/1602.04938

82. Breiman, L. (2001) Random forests. *Mach. Learn.* 45, 5–32

83. Shi, M. and Zhang, B. (2011) Semi-supervised learning improves gene expression-based prediction of cancer recurrence. *Bioinformatics* 27, 3017–3023

84. Liang, Y. *et al.* (2016) Cancer survival analysis using semi-supervised learning method based on Cox and AFT models with L1/2 regularization. *BMC Med. Genomics* 9, 11

85. Peikari, M. *et al.* (2018) A cluster-then-label semi-supervised learning approach for pathology image classification. *Sci. Rep.* 8, 1–13

86. Gonze, D. *et al.* (2017) Multi-stability and the origin of microbial community types. *ISME J.* 11, 2159–2166

87. Knights, D. *et al.* (2011) Supervised classification of human microbiota. *FEMS Microbiol. Rev.* 35, 343–359

88. Clemente, J.C. *et al.* (2012) The impact of the gut microbiota on human health: an integrative view. *Cell* 148, 1258–1270

89. Beck, D. and Foster, J.A. (2014) Machine learning techniques accurately classify microbial communities by bacterial vaginosis characteristics. *PLoS One* 9, e87830

90. Gilbert, J.A. *et al.* (2018) Current understanding of the human microbiome. *Nat. Med.* 24, 392–400

91. Caroni, R. *et al.* (2013) Combination of multiple biological quality elements into waterbody assessment of surface waters. *Hydrobiologia* 704, 437–451

92. Hering, D. *et al.* (2010) The European Water Framework Directive at the age of 10: a critical review of the achievements with recommendations for the future. *Sci. Total Environ.* 408, 4007–4019

93. Dufrêne, M. and Legendre, P. (1997) Species assemblages and indicator species: the need for a flexible asymmetrical approach. *Ecol. Monogr.* 67, 345–366

94. Young, T. *et al.* (2018) Recent trends in deep learning based natural language processing. *arXiv* 1708.02709 cs.CL, Published online August 9, 2017

95. Chen, T. and He, T. (2015) Higgs Boson discovery with boosted trees. *JMLR Workshop Conf. Proc.* 42, 69–80

96. Silver, D. *et al.* (2017) Mastering the game of Go without human knowledge. *Nature* 550, 354–359

97. Libbrecht, M.W. and Noble, W.S. (2015) Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* 16, 321–332

98. Angermueller, C. *et al.* (2016) Deep learning for computational biology. *Mol. Syst. Biol.* 12, 878

99. Esteva, A. *et al.* (2017) Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 115–118

100. Elith, J. and Leathwick, J.R. (2009) Species distribution models: ecological explanation and prediction across space and time. *Annu. Rev. Ecol. Evol. Syst.* 40, 677–697

101. Crisci, C. *et al.* (2012) A review of supervised machine learning algorithms and their applications to ecological data. *Ecol. Model.* 240, 113–122