CellPress

Trends in Microbiology

Opinion

The forecasting power of the microbiome

Sara Correa-Garcia [1], Philippe Constant, [1] and Etienne Yergeau [1,*]

Microorganisms are informative biological integrators of past and present environmental abiotic and biotic conditions. At the same time, they are directly involved in ecosystem processes. Unfortunately, the complexity of microbial communities has so far resulted in most studies being descriptive. Here, we suggest that signals in the microbiome data can be used to forecast future ecosystem processes. The combination of omics with various statistical learning approaches, selected based on accuracy–interpretability and bias–variance trade-offs, will be key to attain this goal, as exemplified by recent studies. The time is ripe for microbial ecologists to fully exploit the forecasting power of microbiomes.

Highlights

Microbial communities are powerful integrators of past and present ecosystem characteristics.

Several recent studies have used microbial communities as indicators of future ecosystem processes, resulting in high-accuracy models to forecast crop quality, soil health, and susceptibility to infection, among others.

The accuracy versus interpretability and bias versus variance trade-offs, along with many methodological considerations specific to microbiome data, need to be considered when building forecasting models.

There is an untapped forecasting potential in microbiome data that can be harnessed with the application of statistical/machine learning tools.

The untapped forecasting potential of the microbiome

Microbiomes (see Glossary) contain a multitude of information, as they are indicators of past and current ecosystem states and are key players in ecosystem processes. This information can be comprehensively accessed using modern **omics** approaches. Despite that, many microbiome studies remain descriptive, focusing on compositional changes associated with specific treatments or across ecosystems. Some pioneer studies have used **statistical learning** to predict, in the sense of classifying, diagnosing, or discriminating, ecosystem processes/states from microbiome data. Here, we use the definition of statistical learning from James *et al.* [1] which '[…] refers to a vast set of tools for understanding data'. For instance, the origin of various water samples could be predicted by the relative abundance of 30 bacterial operational taxonomic units (OTUs) (mostly Proteobacteria and Bacteroidetes) identified through random forest models [2]. Bacterial richness together with the relative abundance of 39 bacterial genera predicted soil **multifunctionality** in degraded alpine meadows [3]. In boreal forest, fungal richness, community composition, and the relative abundance of 15 fungal genera were used with linear regression to predict soil multifunctionality [4]. In another example, a species balance index based on the log-ratio between the relative abundance of the 140 taxa accurately (77%) predicted potato yields [5]. Litter decomposition in soils has also been accurately predicted (72–80%) as 'high' or 'low' by community descriptors, such as fungal and bacterial richness, using logistic regression models [6]. Classifying healthy versus diseased individuals based on their microbiome has been done successfully using many different datasets and machine learning tools [7]. However, we think that we can take this one step further and use microbiomes to forecast the future ecosystem processes/states. Here, we argue that the application of supervised statistical learning on microbiome-derived omics datasets (Box 1) could result in forecasting models for ecosystem processes. We focus most of our discussion on two cases that we see as ripe for such an approach: forecasting disease based on the human gut microbiome, and forecasting crop yields and quality based on the soil microbiome.

In our opinion, microbiomes are ideally suited to forecast ecosystem processes because (i) they are integrators of the past and current environmental conditions of their habitats, and (ii) they are directly involved in the processes to be modelled (Figure 1). Indeed, microbiomes are composed of thousands of species that are simultaneously affected by pH [8], oxygen [9], nutrients [10], host genotype [11], host diet [12], or contamination [13,14]. The abundance and activity levels of each

[1]Institut national de la recherche scientifique, Centre Armand-Frappier Santé Biotechnologie, 531 boulevard des Prairies, Laval, Québec H7V 1B7, Canada

*Correspondence: Etienne.Yergeau@inrs.ca (E. Yergeau).

 https://doi.org/10.1016/j.tim.2022.11.013

Box 1. Supervised and unsupervised learning in a microbiome context

Supervised learning (SL)

SL is a family of methods that model the relationship between inputs (microbiome features, such as taxa, diversity indices, etc.) and an output (health, yield, degradation, nutrient availability, etc.). Then, the model can be used to predict the output value of new observations based on the inputs. The SL techniques applied in microbiome modelling are grouped in two categories: classification (for discrete variables, e.g., disease/healthy) and regression (for continuous variables, e.g., yield). Some SL algorithms previously used with microbiome data are:

Linear regression

This is the simplest method. It fits a linear equation to explain the relationship between a continuous output variable and one or more input variables. When applicable, it provides highly interpretable models. It was often used to model processes based on community descriptors, such as diversity measurements.

Support vector machines (SVMs)

These are a set of classification algorithms that draw a decision boundary line between points (samples) in an n-dimensional space to create groups. The decision boundary maximises the distance between the closest data points to the line (support vectors) and the line itself. In microbiome studies, it has been used to identify input variables that have a high discriminant power for output variables.

k-nearest neighbours (kNNs)

This is an algorithm for performing classification and regression. The main idea is that the value of a point is calculated based on a k number of nearest neighbour points. In a classification task, the label is assigned democratically according to neighbouring k values. In a regression task, the mean value of the k closest points is calculated for the predicted point.

Random forest (RF)

This is also a common classification and regression algorithm. It combines decision trees with bootstrap sampling (samples selected randomly with replacement). Here, many decision trees are estimated in parallel (bagging), which increases accuracy and reduces overfitting. In microbiome studies, this tool was used to classify phenotypic groups and to identify informative microbial features.

Gradient boosted decision trees (GB)

This combines decision trees with boosting. GB fits a series of decision trees, where each tree is an improved version of the previous one. It is a computationally demanding technique, but it performs particularly well in ecosystem state predicting tasks (i.e., effective at the classification of sex, or country of origin, in a human microbiome context [30]).

Unsupervised learning (UL)

UL methods are used to find relationships and structure in the input data without having output data. These are the most used methods in microbiome studies, where output variables are often not available or are not used. Some common UL algorithms include:

k-means clustering

This finds a structure to group the data into k number of clusters specified by the researcher. The algorithm maximises the distance between clusters while minimising the distance within a cluster. It is relatively fast and easy to interpret.

Hierarchical clustering

This groups similar samples into groups called clusters following an agglomerative or divisive approach and yields a classification tree. The number of clusters does not need to be specified. It is slower than k-means, but it always yields the same clustering result.

Principal component analysis (PCA) (and other ordination methods)

These methods use eigenvalue decomposition to create a new set of (n – 1) variables (eigenvectors) that are orthogonal (i.e., not correlated), represent all the variation in the original dataset, and are ordered by the amount of variation explained. The first few components usually represent most of the variation in the original dataset and are plotted to visually identify patterns among samples. PCA can also be used as a dimension reduction tool for SL algorithms.

For further details about supervised and unsupervised learning algorithms applicable to microbiome data, the readers are referred to [7].

Glossary

Explanatory variable, independent variable, predictor, input: these terms are used interchangeably to refer to parameters, often referred to as features of a model, that influence the variation in a response variable. For example, abundance of a taxa or gene, microbial diversity, etc.

Legacy effect: the impact of previous biotic and abiotic conditions on current processes.

Microbiome: refers to the combination of all the microorganisms (bacteria, fungi, protozoa, archaea, and algae) in an environment.

Multifunctionality: the combination of multiple functions or properties provided by a system. For example, forests providing food, wood and fibre, livelihoods and incomes, carbon sink, environmental and landscape protection, recreation, and habitats for biodiversity.

Multivariate niche: multiple environmental biotic and abiotic characteristics acting together to create a niche.

Omics: methods used to characterize the entire set of biomolecules in a sample, such as genes (metagenomics), RNA transcripts (metatranscriptomics), or proteins (metaproteomics).

Response variable, dependent variable, predicted variable, output: these terms are used interchangeably for a set of qualitative or quantitative variables that depend on the values of the explanatory variables. The aim is generally to model the response variable using the explanatory variables. For example, yields, health status, etc.

Statistical learning: supervised statistical learning is used to predict or estimate, through modelling, an output (ecosystem processes) based on inputs (microbiome omics data). Unsupervised statistical learning is quite common in microbiome studies, it does not involve an output, and is limited to understanding the inherent relationships and structure in the inputs (i.e., the algorithm is asked to identify trends in the input dataset without supervision from a response variable).

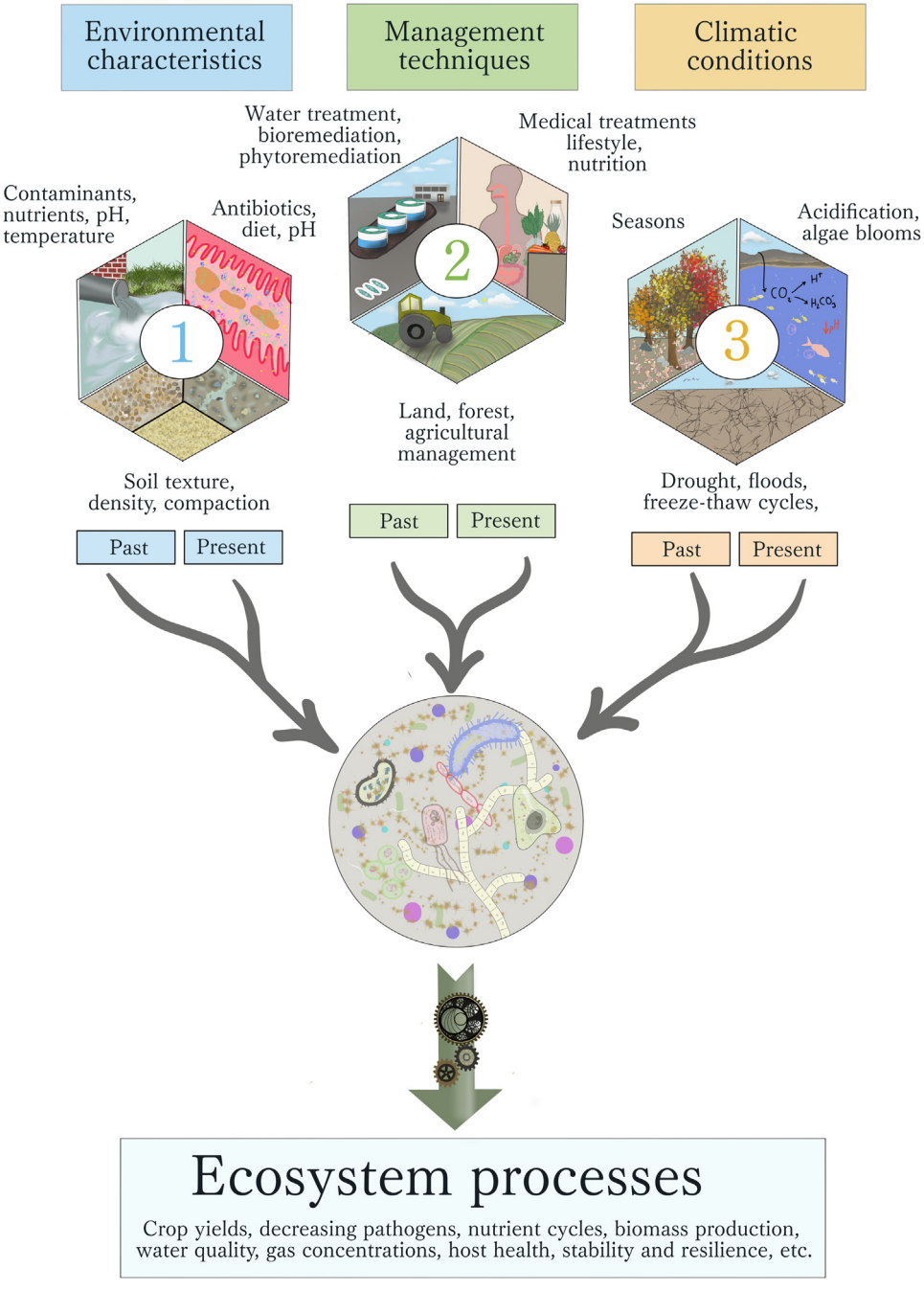

Figure 1. The forecasting power of the microbiome. Traditionally, ecosystem processes have been modelled using various environmental physicochemical predictors. This has often resulted in models with low resolution and accuracy because the parameters mostly influence ecosystem processes indirectly through their influence on the microbiome. Here, we suggest focusing the modelling efforts directly on the microbiome. Microbiomes are integrators of past and present environmental characteristics (1), management techniques (2), and climatic conditions (3), and at the same time they are involved directly in the process to be modelled, giving them an unparalleled forecasting power.

microorganism results from the combined pressures of these abiotic and biotic characteristics, creating a **multivariate niche**. The combined presence of thousands of microorganisms, each with their own multivariate niche, results in a highly integrated description of the biotic and abiotic characteristics of an environment.

On top of integrating many biotic and abiotic environmental characteristics, microbiomes also show a clear signature of past events, called **legacy effect**. Past events of drought [15,16], freeze-thaw and drying-rewetting cycles [17], antibiotic usage [18], and host plant presence [19,20] were shown to influence the composition, diversity, and functions of microbiomes and their response to contemporary events. For instance, the soil microbiome of two fields that had a different history of water stress responded differently to contemporary water stress, which in turn affected wheat growth [15,16]. Plant composition also has strong legacy effects on microbial communities in soil, especially for fungi [20], where past plant community composition outweighed the effect of the current plant community for up to 5 months [19]. In other ecosystems, such as the human gut, there is a strong legacy effect of the use of antibiotics [18] or diet [12] on the microbiome. Vancomycin caused long-term shifts in human gut microbial communities, with only 39% of the previously abundant taxa still present 22 weeks after the vancomycin treatment had ceased, as compared to 90% for control subjects [18]. Taken together, these examples indicate that the legacy effect of past abiotic and biotic environmental conditions can be captured using modern omics tools targeting the microbiome. When predicting actual disease status of individuals from their gut microbiome data, legacy effects were seen as unwanted confounding factors reducing the accuracy of the models [21]. However, since legacy effects have important consequences on future ecosystem processes that we wish to forecast, it is important to be able to take them into account in the models.

Not only are microbiomes integrators of the present and past environmental conditions where they occur, they are also responsible for many ecosystem processes such as nutrient cycling, organic matter decomposition, fermentation of nondigestible plant residues, and pollutant degradation (Figure 1). In that sense, there is a possibility that some microbial **predictors** selected by the models will have a causal relationship with the processes modelled. In view of their capacity to integrate past and present conditions of their habitats and their direct implication in processes, we think that microbiomes also have some predictive power for future ecosystem processes or states (Figure 1).

Microbiomes and the environment where they occur are complex. Microbiome complexity arises from a combination of factors, such as spatial heterogeneity, microscale biotic and abiotic interactions, rapid turnover, high dispersal capacity, and horizontal gene transfer, among others. To further compound this complexity, ecosystem processes often involve multiple species in syntrophic or symbiotic relationships, occupying diverse ecological niches, and possessing various isozymes of different biochemical properties. The nature and the mechanisms behind microbial interactions are poorly understood and difficult to study. One might wonder how it will be possible to forecast future ecosystem processes as this should require an exhaustive knowledge of microbiome diversity, interactions, metabolism, genomic make-up, and physiology, all this varying spatially and temporally at the micrometer scale [22]. In fact, we argue that omics datasets recapitulate, often cryptically, a large part of this information, but it is difficult, if not impossible, to derive it without the help of advanced statistical/modelling tools. In the next section, we discuss the few studies that have used the microbiome as a predictor for future ecosystem processes/states and discuss the rationale behind choosing various statistical learning tools for the purpose of forecasting. For detailed reviews of statistical learning approaches in general and as it applies to the (human) microbiome, the interested reader is referred to James *et al.* [1] and Marcos-Zambrano *et al*. [7], respectively.

Microbial-based forecasting

In subsequent text, we discuss some key criteria that differentiate the various statistical learning approaches, namely, the relative importance of accuracy versus interpretability for the **output** and the goal of the modelling (hypothesis generation, microbiome engineering, diagnosis, monitoring), the bias/variance trade-off, and specific considerations for modelling microbiome data.

Interpretability versus accuracy

An important aspect of forecasting using microbiomes is interpretability, which often comes at the cost of accuracy. Nonparametric statistical learning tools, such as neural networks, are very accurate but difficult to interpret by humans as the inner layers are hidden from the user. At the other end of the spectrum, linear regression, and more so when combined with a dimensionality reduction approach (more information on that in subsequent text), is easier to interpret, but often at the cost of reduced accuracy. In linear regression, the coefficients allow the user to determine the relative effect of the predictor (positive or negative; scale) on the **dependent variable**. For example, both random forest and support vector machine (SVM) models identified the microbiome composition based on 16S rRNA gene sequencing as being an accurate predictor of agricultural soil health in a continental-scale study [23]. However, it was not straightforward to identify the taxa that contributed the most to the predictive power, and a leave-one-out analysis of thousands of taxa had to be performed *a posteriori*. Another study forecasted wheat yield and grain quality in two fields using multiple linear regression based on forward-selected microbial indicators, which allowed direct identification of the most important predictors [24]. Surprisingly, both studies identified a similar subset of important predictors of agricultural soil health/productivity (*Gaiella*, Candidatus *Udeaobacter*, Blastocatellales). With only two studies, it is not possible to draw a strong conclusion, but it may suggest the existence of robust common microbial predictors for important ecosystem services that were captured using models with different interpretability.

Interpretability also enables the elaboration of confirmatory experiments for model validation. These experiments are crucial because when a microbial feature is selected in a microbiome-based forecasting model, it can (i) be unrelated to the process of interest but have an environmental optimum that overlaps with the optimum of the process, (ii) be related causally to the process of interest, such as the abundance of a certain functional guild responsible for a process, or (iii) be indicative of legacy effects that affected the process. An interpretable model could also orient microbiome engineering efforts [25] by identifying parameters to be manipulated. For example, the abundance of ammonia oxidizers at seeding was a key feature selected in stepwise multiple linear regression and random forest models for wheat yields and grain quality at the end of the season [24,26]. Since the coefficients were negative, an independent confirmatory field experiment was set up targeting the ammonia oxidizer using a nitrification inhibitor to increase grain quality [27]. In another example, using unsupervised learning methods, early rhizosphere microbial taxa were identified as key features for the future susceptibility of tomato plants to *Ralstonia solanacearum* wilt [28]. Five bacteria corresponding to these taxa were isolated and, when inoculated on healthy plants, they reduced disease by 30–100%. These studies show the value of model interpretability for some application of forecasting using the microbiome.

The choice between accuracy or interpretability is context dependent, and the researcher may consider accuracy of utmost importance for the question at hand. When forecasting susceptibility to infection or disease, it might be more important to have a highly accurate model, even though the model might not be easily interpretable. For example, using the relative abundance of approximately 100 taxa and SVM, it was possible to forecast the susceptibility of the human gut to invasion by *Vibrio cholerae* [29]. 16S rRNA gene sequencing is becoming a routine tool, which means that highly accurate models containing hundreds to thousands of taxa or general community descriptors (alpha-diversity indices or ordination axes) could be used for high-accuracy forecasting.

However, in the meantime, tools that can select a few variables, such as least absolute shrinkage and selection operator (LASSO), ridge, elastic net, or forward-selected regressions can be useful, when forecasting accuracy is not paramount. Indeed, to forecast the state of a new sample, the few selected taxa could be simply measured using taxa-specific tools, such as qPCR, and the results could then be used in the regression equations. This might not be the best way to proceed in complex ecosystems (e.g., soils) where many processes can be forecasted from the same omic dataset, which might warrant a full sequencing.

Accuracy can increase with model complexity, but at the cost of training time. For instance, when predicting soil health based on 16S rRNA gene amplicon data at the amplicon sequence variant (ASV) level, SVM could be trained in 20 min, whereas random forest models took almost 19 days to compute [23]. In that case, the accuracy was not necessarily improved, and in fact SVM regression often outperformed random forest regression accuracy [23]. Some recent approaches [e.g., extreme gradient boosting (XGBoost)] allow updating of trained datasets with new observation, which can dramatically improve training time of many methods. Code parallelization and the efficient use of graphics processing units (GPUs) could also reduce drastically training time. With the view of 'acting' on predictors identified in forecasting models, training times in the range of weeks is problematic as the opportunity window to modify the ecosystem or make interventions on patients might not be that large.

Bias versus variance

Another issue to keep in mind arises from the limited capacity of models to capture the true relationships of the predicted processes with the measured **explanatory variables**. There are two components to that: bias and variance. Bias is the error in the prediction due to wrong assumptions in the model. A highly biased model will miss the true relationship between the predictors and the **response variables**, failing to capture important patterns of the data, also known as model underfitting. In contrast, the variance component will increase due to an excessive modelling of the noise in the training dataset. This results in a reduced performance of the model with new data. Applying methods with high variance that very accurately represent the training data sets will not be useful to consistently predict new data, a problem also known as model overfitting.

Ideally, the chosen model will accurately capture the regularities of the data used for training (low bias) while consistently offering relatively good fits for new unseen data (low variance). Unfortunately, having the best of both worlds is impossible in practice: decreasing one of these terms will increase the other, a phenomenon called the bias–variance trade-off. Some feature selection methods are prone to overfitting (high variance). For instance, when selecting the most discriminant features using correlation analyses, and then using these features to train a gut microbiome model to classify individuals by disease status, the accuracy of the model to predict disease state decreased when used on a different cohort (high variance) [21]. Similarly, in agricultural soils, linear regression was able to forecast grain baking quality and yield with an accuracy of up to 90% (low bias), using only ten microbial predictors that were preselected by correlation analysis [26]. This model had, however, high variance, as further efforts to forecast wheat grain quality in different fields did not select the same variables [30]. Correlation analysis might not be the optimal method to reduce the dimensionality of the data (more on that in the next section). Regularization, boosting, and bagging are some common methods that can be used to find a balance between bias and variance (the readers are referred to [1] for more details). Reducing the variance of forecasting models will be crucial if these models are to be used routinely to, for instance, forecast agricultural yields or disease susceptibility of patients.

In view of the cost associated with environmental genomics, most studies so far have datasets of less than 1000 samples. This leads to the problem of high dimensionality discussed in the

following section and to high variance in the resulting models, but also to the impossibility to have proper training/test datasets. For instance, a study applying linear models with ridge regularization and gradient boosted decision trees on more than 34 000 gut microbiome samples from US and Israeli individuals accurately predicted various human phenotypes (age, gender, etc.) [30]. However, modelling with smaller subsamples led to highly variable results [30]. This suggests that, at least for highly variable environments, very large datasets might be necessary to generate forecasting models with low variance. The ever-decreasing cost of sequencing might help to solve this issue, though costs associated with sampling and computing might become limiting.

Environmental genomics and forecasting: methodological considerations

An important methodological aspect to consider when dealing with microbiome indicators as **inputs** in forecasting models is the type of genetic material. For instance, the DNA pools in soils can change throughout seasons or years, but they are not especially sensitive to changes over days or weeks, so DNA-based approaches could be good indicators of legacy effects. This can be specifically useful to forecast soil processes that vary through seasons, such as N_2O or other gas emissions [31]. However, the presence of extracellular DNA that persists depending on the environmental conditions and various levels of active versus inactive cells might also blur the picture. In contrast, RNA-based approaches give a snapshot of the current expression profile of the microbiome, but in view of the short half-life of mRNA, they are unlikely to inform on the legacy effect or be useful for long-term forecasting. However, RNA-based approaches such as metatranscriptomics were proven to be useful for monitoring processes, such as microbial activities during phytoremediation of polluted soils [32–35]. In contrast to amplicon sequencing, shotgun omics approaches can inform us, in a single sequencing, about multiple taxa and processes, such as the capacity to fix nitrogen, to degrade contaminants, and to degrade complex organic matter. However, it dramatically increases the dimensionality of the data.

Whether DNA- or RNA-based, amplicon-based, or shotgun, environmental genomic data are generally high dimensional [the number of features (p) far exceeds the number of samples (*n*), p>>*n*]. Using all the features can result in models with a perfect fit, even though the features might be completely unrelated to the response variable. Adding more features does not necessarily result in a better model, and in fact, it often leads to a less optimal model with more noise (high variance), a problem dubbed the ‘curse of dimensionality’. Also, it often results in sparsity in the descriptor space, which is particularly problematic for many nonparametric approaches, such as k-nearest neighbour, that use proximity to other descriptors to predict the output from a new descriptor set. Some approaches were devised to go around this problem, allowing for an improved prediction accuracy and/or model interpretability. The first type of approach consists of selecting the variables to be included in the model. One well-known example is stepwise multiple regression modelling, where variables are entered or removed one by one from the regression equation, until only significant variables are left. Using this approach, it was possible to accurately forecast wheat grain quality and yields using a few selected microbiome predictors measured at the seedling stage [26]. Similarly, shrinkage or regularisation methods, such as ridge, LASSO, or elastic net, include all predictors in the equation but reduce the size of the coefficient estimates towards zero. It was recently used to compare the accuracy of forecasting wheat yield and quality from microbiome datasets collected at different dates [36]. However, for many of the sampling dates close to harvest, the LASSO procedure resulted in a null model, where all coefficients were shrunk to zero, suggesting that the forecasting potential of the microbiome at these dates was mediocre, or that LASSO was not the most appropriate tool for this task [36]. Similarly, LASSO and elastic net resulted in less accurate disease prediction models than when using random forest selection, even though the accuracy of the latter approach was optimal when more

than 60 species were selected [21], which is still a lot of descriptors. Another type of approach is to project the predictors in a lower dimension using eigenvalues decomposition algorithms, such as principal component analysis (PCA). This approach was used in the context of disease prediction from 16S rRNA amplicon data and was shown to improve accuracy, but only for certain diseases [37]. Similarly, PCA was shown to bring no improvement to accuracy in the context of disease prediction [38]. The best method to tackle the high dimensionality of microbiome data will depend on many factors, and feature selection is an active area of research, with many recent approaches developed specifically for microbial omic datasets [39].

Finally, the microbial features selected in a model could consist of genetic material with no known representative in databases. This is not necessarily a problem as it is expected that a high proportion of the genetic material retrieved from environmental samples will have no known attributed function or taxonomy but could still have a predictive power [40]. Using the CoMeta algorithm [41] to calculate similarity between metagenomic reads without prior annotation, it was possible to predict the geographical origin of samples with an accuracy of 87.5%, similar to methods relying on taxonomical/functional information (71–91.2%) [42]. Although using unannotated sequences can be faster, it reduces the interpretability of the results and may be especially sensitive to different genome sequencing methods [43].

Concluding remarks

Microbiomes integrate past and current ecosystem biotic and abiotic characteristics and are major players in ecosystem processes. As such, we think that they contain a signal that can be used to forecast ecosystem processes and states. This signal can be harnessed through the combination of microbial omics and statistical learning, two fields that have tremendously developed in recent years. We discussed the challenges associated with both the data and the methods to fully exploit the combination of these powerful tools for forecasting (see Outstanding questions). We envisage that microbial-based forecasting models will be used in all kinds of environments to (i) help site managers, farmers, foresters, physicians, and veterinarians to decide on the best management approach to optimize ecosystem services, and (ii) guide microbiome manipulation efforts that could help solve humanity's most pressing problems such as environmental degradation, antibiotic resistance, or food shortages. We believe that it is now time to harness the forecasting power of the microbiome.

Outstanding questions

Can we tailor models from microbiome data that suitably predict a range of ecosystem processes across multiple geographical scales?

How do simple models with low dimensionality descriptors compare to complex machine-learning models in terms of interpretability, accuracy, bias, and variance?

How many microbiome features are enough to provide highly accurate models without incurring high complexity? Is there a sweet spot? Does it widely vary depending on the environment/process?

Can we use forecasting models to engineer microbiomes, through agricultural management practices, use of specific inhibitors, prebiotics, or probiotics?

To what extent can we infer causality relationships from forecasting models? What type of confirmatory experiments are needed?

Acknowledgments

This work was supported by Genome Canada and Genome Quebec (2020 Large-Scale Applied Research Project Competition grant 18207).

Declaration of interests

No interests are declared.

References

1. James, G. *et al.* (2013) *An Introduction to Statistical Learning* (2nd edn), Springer, p. 103
2. Wang, C. *et al.* (2021) Machine learning approach identifies water sample source based on microbial abundance. *Water Res.* 199, 117185
3. Wang, J. *et al.* (2021) Bacterial richness is negatively related to potential soil multifunctionality in a degraded alpine meadow. *Ecol. Indic.* 121, 106996
4. Li, J. *et al.* (2019) Fungal richness contributes to multifunctionality in boreal forest soil. *Soil Biol. Biochem.* 136, 107526
5. Jeanne, T. *et al.* (2019) Using a soil bacterial species balance index to estimate potato crop productivity. *PLoS One* 14, e0214089
6. Albright, M.B.N. *et al.* (2020) Soil bacterial and fungal richness forecast patterns of early pine litter decomposition. *Front. Microbiol.* 11, 542220
7. Marcos-Zambrano, L.J. *et al.* (2021) Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment. *Front. Microbiol.* 12, 313
8. Tripathi, B.M. *et al.* (2018) Soil pH mediates the balance between stochastic and deterministic assembly of bacteria. *ISME J.* 12, 1072–1083
9. Singer, J.R. *et al.* (2019) Preventing dysbiosis of the neonatal mouse intestinal microbiome protects against late-onset sepsis. *Nat. Med.* 25, 1772–1782

10. Zwetsloot, M.J. *et al.* (2020) Prevalent root-derived phenolics drive shifts in microbial community composition and prime decomposition in forest soil. *Soil Biol. Biochem.* 145, 107797
11. Morales Moreira, Z.P. *et al.* (2021) Crop, genotype, and field environmental conditions shape bacterial and fungal seed epiphytic microbiomes. *Can. J. Microbiol.* 67, 161–173
12. Jain, A. *et al.* (2018) Similarities and differences in gut microbiome composition correlate with dietary patterns of Indian and Chinese adults. *AMB Express* 8, 1–12
13. Correa-García, S. *et al.* (2021) Soil characteristics constrain the response of microbial communities and associated hydrocarbon degradation genes during phytoremediation. *Appl. Environ. Microbiol.* 87, e02170-20
14. Cavé-Radet, A. *et al.* (2020) Phenanthrene contamination and ploidy level affect the rhizosphere bacterial communities of *Spartina* spp. *FEMS Microbiol. Ecol.* 96, fiaa156
15. Azarbad, H. *et al.* (2018) Water stress history and wheat genotype modulate rhizosphere microbial response to drought. *Soil Biol. Biochem.* 126, 228–236
16. Azarbad, H. *et al.* (2020) Four decades of soil water stress history together with host genotype constrain the response of the wheat microbiome to soil moisture. *FEMS Microbiol. Ecol.* 96, fiaa098
17. Meisner, A. *et al.* (2021) Soil microbial legacies differ following drying-rewetting and freezing-thawing cycles. *ISME J.* 15, 1207–1221
18. Isaac, S. *et al.* (2017) Short- and long-term effects of oral vancomycin on the human intestinal microbiota. *J. Antimicrob. Chemother.* 72, 128–136
19. Hannula, S.E. *et al.* (2021) Persistence of plant-mediated microbial soil legacy effects in soil and inside roots. *Nat. Commun.* 12, 1–13
20. Hellequin, E. *et al.* (2021) Shaping of soil microbial communities by plants does not translate into specific legacy effects on organic carbon mineralization. *Soil Biol. Biochem.* 163, 108449
21. Pasolli, E. *et al.* (2016) Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. *PLoS Comput. Biol.* 12, e1004977
22. Hermans, S.M. *et al.* (2020) Using soil bacterial communities to predict physico-chemical variables and soil quality. *Microbiome* 8, 1–13
23. Wilhelm, R.C. *et al.* (2022) Predicting measures of soil health using the microbiome and supervised machine learning. *Soil Biol. Biochem.* 164, 108472
24. Yergeau, É. *et al.* (2020) Microbial indicators are better predictors of wheat yield and quality than N fertilization. *FEMS Microbiol. Ecol.* 96, fiz205
25. Agoussar, A. and Yergeau, E. (2021) Engineering the plant microbiota in the context of the theory of ecological communities. *Curr. Opin. Biotechnol.* 70, 220–225
26. Asad, N.I. *et al.* (2021) Predictive microbial-based modelling of wheat yields and grain baking quality across a 500 km transect in Québec. *FEMS Microbiol. Ecol.* 97, fiab160
27. Schmidt, R. *et al.* (2022) The nitrification inhibitor nitrapyrin has non-target effects on the soil microbial community structure, composition, and functions. *Appl. Soil Ecol.* 171, 104350
28. Gu, Y. *et al.* (2022) Small changes in rhizosphere microbiome composition predict disease outcomes earlier than pathogen density variations. *ISME J.* 16, 2448–2456
29. Midani, F.S. *et al.* (2018) Human gut microbiota predicts susceptibility to *Vibrio cholerae* infection. *J. Infect. Dis.* 218, 645–653
30. Rothschild, D. *et al.* (2022) An atlas of robust microbiome associations with phenotypic traits based on large-scale cohorts from two continents. *PLoS One* 17, e0265756
31. Graham, E.B. *et al.* (2014) Do we need to understand microbial communities to predict ecosystem function? A comparison of statistical models of nitrogen cycling processes. *Soil Biol. Biochem.* 68, 279–282
32. Gonzalez, E. *et al.* (2018) Trees, fungi and bacteria: tripartite metatranscriptomics of a root microbiome responding to soil contamination. *Microbiome* 6, 53
33. Pagé, A.P. *et al.* (2015) *Salix purpurea* stimulates the expression of specific bacterial xenobiotic degradation genes in a soil contaminated with hydrocarbons. *PLoS One* 10, 1–16
34. Yergeau, E. *et al.* (2014) Microbial expression profiles in the rhizosphere of willows depend on soil contamination. *ISME J.* 8, 344–358
35. Yergeau, E. *et al.* (2018) Soil contamination alters the willow root and rhizosphere metatranscriptome and the root–rhizosphere interactome. *ISME J.* 12, 869–884
36. Asad, N. *et al.* (2023) Early season soil microbiome best predicts wheat grain quality. *FEMS Microbiol. Ecol.* 99, fiac144
37. Oh, M. and Zhang, L. (2020) DeepMicro: deep representation learning for disease prediction based on microbiome data. *Sci. Rep.* 10, 1–9
38. Jasner, Y. *et al.* (2021) Microbiome preprocessing machine learning pipeline. *Front. Immunol.* 12, 1954
39. Ditzler, G. *et al.* (2015) Fizzy: feature subset selection for metagenomics. *BMC Bioinformatics* 16, 1–8
40. Bzdok, D. *et al.* (2018) Statistics versus machine learning. *Nat. Methods* 15, 233–234
41. Kawulok, J. and Deorowicz, S. (2015) CoMeta: classification of metagenomes using k-mers. *PLoS One* 10, e0121453
42. Kawulok, J. *et al.* (2019) Environmental metagenome classification for constructing a microbiome fingerprint. *Biol. Direct* 14, 1–23
43. Anyaso-Samuel, S. *et al.* (2021) Metagenomic geolocation prediction using an adaptive ensemble classifier. *Front. Genet.* 12, 521