**Benchmarking machine learning approaches for microbiome studies**

**Idea**

Compare the performance of different machine learning methods in a regression and classification setting using the Estonian Biobank microbiome cohort (EstMB) data (N = 2509). In addition, consider different data transformations for the microbiome data and evaluate their impact on the performance. Continuous variables such as BMI and age can be used for regression and prevalent diseases for classification setting. The choice of the diseases can be arbitrary, but could include diseases with different numbers of cases, diseases with stronger/weaker associations with the microbiome etc.

**Background**

Machine learning in microbiome studies is widely used and the interest is definitely growing. However, there is no universal understanding of the algorithmic approaches that can best utilize the information present in the microbiome data. Thus, this is an interesting and widely discussed topic that can have a great impact on the potential applications leveraging microbiome data. There exist several consortia/initiations which aim to answer this question. For example, we are part of the COST (European Cooperation in Science and Technology) action “Statistical and machine learning techniques in human microbiome studies”, where one working group focuses on optimizing and standardizing the use of state-of-the-art ML techniques (<https://www.ml4microbiome.eu/working-group-descriptions/>).

A key topic in microbiome research is the sample space of the input data. The sequencing data appears as count data, but in reality, only relative abundance of the microbial features can be observed (<https://www.frontiersin.org/articles/10.3389/fmicb.2017.02224/full>). Thus, transforming the read counts to relative abundances is usually the first step and machine learning methods are usually applied on relative abundances. However, relative abundances raise several limitations, which can have an impact on the performance of the prediction models. For example, the “unit-sum” limitation of the relative abundances can give rise to spurious correlations etc. Relative abundance data and thus also microbiome data are mathematically considered compositional data, with its own mathematical theory which proposes other data transformations with appealing properties (<https://academic.oup.com/bioinformatics/article/34/16/2870/4956011>). Such transformations are called “log-ratio” transformations and several prediction methods based on log-ratio transformations have been proposed for microbiome data showing good predictive performance with high sparsity:

1. Discriminative balance analysis: <https://journals.asm.org/doi/10.1128/msystems.00230-19?permanently=true>
2. Balances (Selbal): <https://journals.asm.org/doi/10.1128/msystems.00053-18?permanently=true>
3. CodaCore: <https://academic.oup.com/bioinformatics/article/38/1/157/6366546>
4. Amalgams: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7671324/>
5. <https://arxiv.org/abs/2111.08953>
6. <https://www.sciencedirect.com/science/article/pii/S0098300420305999?via%3Dihub>
7. <https://onlinelibrary.wiley.com/doi/10.1111/biom.12995>
8. …

One of the key benefits of these methods is the small number of microbial features that they utilize to achieve a competitive predictive power.

Another way to take advantage of the log-ratio transformations is to apply them before using traditional machine learning methods. This has been shown to result in better predictive performance compared to relative abundance data (sometimes referred to as TSS - total sum scaling transformation):

1. <https://www.researchgate.net/publication/334225471_On_machine_learning_algorithms_and_compositional_data>
2. <https://hess.copernicus.org/preprints/hess-2018-584/hess-2018-584.pdf>

Traditional machine learning methods can provide added predictive power with the price of limited explainability. Thus, balancing the predictive power with explainability becomes important for the conclusions.   
  
Considering the special-purpose algorithms and the variety of machine learning approaches in combination with different data transformations, a large-scale benchmark of the performances hasn’t been done. Thus, I propose to apply the most relevant microbiome-specific algorithms and several machine learning methods in combination with different log-ratio transformations on the EstMB data and compare their predictive performance.

**Problems**

1. How to correctly compare the performance of the machine learning models. Which testing scheme to consider? Traditional train/test split + cross validation does not allow correct comparison and is overoptimistic: <https://www.worldscientific.com/doi/abs/10.1142/S0218213015400230>
2. How to deal with class imbalance? Simpler methods can be used and sensitivity analysis can be considered as an extension.
3. How to deal with zero-imputation for the log-ratio transformation? Simpler methods can be used and sensitivity analysis can be considered as an extension.
4. …

**Extensions**

1. Study the effect of various zero-imputation approaches. As log-ratio transformations require zero-imputation before applying the transformation, the approach taken can affect the results. Simplest and most frequently used approach is to use a pseudocount - small non-zero value, but more sophisticated approaches can be used: (<https://www.sciencedirect.com/science/article/abs/pii/S0169743915000490?via%3Dihub>)
2. Comparing different methods for dealing with class imbalance.
3. Integrate log-ratio based dimensionality reduction methods into the machine learning workflow and evaluate their effect on the performance. Commonly, machine learning workflows/packages have integrated most common dimensionality reduction methods such as PCA/ICA/autoencoders, but there are special ones developed for microbiome data such as distal balances or amalgams (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7671324/>). As the methods need to be integrated into cross-validation schemes, this can be an important and practical extension.
4. Log-ratio transformations such as ALR (additive log-ratio transformation) require the selection of the denominator. However, the selection can have an effect on the results. Several methods and proposals have been done for making the selection: (<https://www.biorxiv.org/content/10.1101/2021.05.15.444300v1.abstract>). Sensitivity analysis on the effect of the selection can be done.
5. A bit out of the context, but an impulsive idea: data augmentation using Dirichlet-Multinomial sampling for achieving more robust performance estimates.
6. …