Jennifer Neumaier

Institute of Genomics, University of Tartu

Master Thesis

*Benchmarking machine learning performances with compositional* *data*

Table of Contents

[Introduction 2](#_Toc104917496)

[Characteristics of Compositional Data 3](#_Toc104917497)

[The Simplex Space 5](#_Toc104917498)

[Mapping the Simplex Space into Euclidean Space 7](#_Toc104917499)

[Log-Ratio Transformations 8](#_Toc104917500)

[The Problem 10](#_Toc104917501)

[Machine Learning 10](#_Toc104917502)

[CODACORE 11](#_Toc104917503)

[Implementation/Assessment 12](#_Toc104917504)

[Methodology 15](#_Toc104917505)

[Data set 15](#_Toc104917506)

[Pre-Processing 15](#_Toc104917507)

[Imputation 16](#_Toc104917508)

[Transformations 16](#_Toc104917509)

[Machine LEarning models 17](#_Toc104917510)

[Results 18](#_Toc104917511)

[INFLUENCE OF TRANSFORMATION ON PERFORMANCES 18](#_Toc104917512)

[Model performances in a Binary Classification Setting 18](#_Toc104917513)

[Regression Setting 21](#_Toc104917514)

[INFLUENCE OF DATA LEAKAGE ON PERFORMANCES 23](#_Toc104917515)

[INFLUENCE OF TEST SET ON PERFORMANCES 27](#_Toc104917516)

[Data set size 29](#_Toc104917517)

[Discussion/Conclusion 30](#_Toc104917518)

[Conclusion 32](#_Toc104917519)

[Publication bibliography 35](#_Toc104917520)

[Supplementary 37](#_Toc104917521)

[Correlation Coefficients 37](#_Toc104917522)

[Presence-Absence-Pattern 38](#_Toc104917523)

# Abstract/Introduction

Machine learning in microbiome studies is widely used and the interest is 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. A key topic in microbiome research is the sample space of the input data. The sequencing data appears as count data, but, only relative abundance of the microbial features can be observed, commonly called “compositional data”. 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. Therefore, log-ratio transformations are a proposition made by several studies now, however their impact on machine learning performances has never been tested in large-scale studies. The goal of this benchmarking project is to rectify that and conduct several machine learning models under several log-ratio transformations in comparison to *CoDaCoRe*, an algorithm specifically made with microbiome analysis in mind. This way it will become clearer how log-ratio transformations impact machine learning performances with microbiome data.

# Introduction

Working with mathematical concepts is always a bit out of the comfort zone for most biologists. The field of statistics looms like a big black cave which is only lightened in a few places, but we have to traverse it anyways. Unfortunately, with Machine Learning encroaching also into biological research questions, this black cave just became a bit bigger, without bringing any additional light.

Machine Learning provides a useful tool to exploit information from biological data sets. It brings the possibility of predicting host-phenotypes, finding associations between features and using microbial communities to characterize patients. Several papers (ML pre-processing, 2022 paper, etc.) already gave practical advice on how to use Machine Learning with microbiome data, however a vital concept concerning the mathematical characteristics of microbiome data has been almost completely ignored in the last few years.

~~Ignoring mathematics is just not an option. One concept that became increasingly more important in sequencing data analysis is the concept of “compositional data”.~~ Several papers (Greenacre et al. 2021b; Gloor et al. 2017; Quinn et al. 2018) made it abundantly clear that sequencing data is of compositional nature, which means it has distinct mathematical characteristics than other data types. While some papers already confirmed that log-ratio transformations do affect Machine Learning performances (ML preprocessing, log contrast paper, IBD), it has never been fully analyzed how log-ratio transformations impact Machine Learning performances over several different data sets.

The goal of this master thesis is an attempt in analyzing the properties of compositional data and comparing how they apply in practical Machine Learning pipelines and if they result in improved and beneficial performances. ~~the information around compositional data more approachable, summarizing the achieved solutions, and in a practical part, trying to assess if these achievements are also applicable when combining compositional data and machine learning concepts.~~

As this master thesis uses microbiome sequencing data and was created in a microbiome research group, this text will mostly focus on this field and its papers. All results should be applicable to other high-throughput sequencing data, as well as any data that is in some way confined by an arbitrary sum. Such data is found for example in geochemistry, ecology, sociology, political sciences, etc., and therefore ultimately spans the problematic into various different fields (Greenacre et al. 2021a). However, sequencing data is quite unique as it usually contains way more features than samples, which comes with its own problems in Machine Learning.

# Characteristics of Compositional Data

In order to define and illustrate the concept of compositional data, let’s assume a classical biological example. The following Figure (1A) shows two different ecological fields: A and B. In field A, four rabbits, seven birds, eight bees and one wolf have been counted, whereas field B contains two rabbits, four birds, four bees and one wolf. It becomes clear that, as similar as the diversity may be, the fact that field B seems to have only half of the population of field A. This in itself is already valuable information. The total counts per field can be preserved in our data collection and therefore, the absolute count of each organism in this field matters.



Figure 1: Information Loss of Normalized Data

(A) Illustration of the number of animals found in two different samples. Field A contains four rabbits, eight bees, seven birds and one wolf, whereas field B contains two rabbits, four bees, four birds and one wolf. In (B) the absolute counts have been plotted as a stacked bar plot, with each animal in a different color. (C) shows the stacked bar plot as normalized counts, e.g., percentages.

When using absolute counts, the difference between both fields is easily visible (Fig. 1B). However, when we really want to compare both fields, we need to transform the samples to a common scale. This is called normalization and we can see the effect in (Fig. 1C). As soon as the data is normalized, the particular information of absolute counts gets lost. When collecting ecological data ourselves, we can preserve the fact that field B only contained 11 individuals and field A contained 20, by saving that number somewhere on our Excel sheet. However, the problem with sequencing data is: we get the data in the form of Fig. 1C.

To demonstrate how the method of using a sequencing machine cannot preserve absolute counts, imagine the following situation: We want to sample field A multiple times a day, but in order to be more efficient, we buy a machine to do the counting for us. Three times a day this machine transmits the number of all the different animals coming to this field. However, this machine has one flaw: it can only count to 20. As soon as the 21st animal on this day comes to the field, it is just simply not counted.

This ultimately means, that the overall number of 20 carries no meaning. Every sample has this exact total number, so it carries no valuable information. Of course, a limit of 20 is weird for us to understand, but sequencing machines do the exact same thing. They are limited in their capacity on e.g., the flow cells and even the most effective sequencing machines could never fully sequence the entirety of the organism’s DNA contents. And not only the sequencing machines, but the whole RNA-Seq procedure limits the total number of sequences measured. The total number of sequences measured by sequencing machines ultimately depends on the *chemistry of the assay*, not the input material (Quinn et al. 2018).

The consequence of this sampling problem is, that we have to accept the fact that the sum of counts in sequencing data are irrelevant. This leads us to the problem that we cannot compare groups of data as we are used to do it. If we only have the information of Fig. 1C, then we would assume that field A and B share the same ecology. Or if this were microbiome data, we would maybe assume that these samples came from patients of the same group and calculating parameters like richness are not possible as they require a number of absolute counts of individuals.

~~This in itself is not really a problem, as the gut microbiome has always been analyzed in terms of diversity and how the community changes during stress (resistance) and how quick it turns back into a state of equilibrium after stress (resilience).~~ [~~https://www.sciencedirect.com/science/article/pii/S1931312812003587~~](https://www.sciencedirect.com/science/article/pii/S1931312812003587)

Data that is constraint in that way is called “Compositional Data”, first introduced by John Aitchison (1982), roughly 40 years ago. Thankfully, we can still use sequencing data. We just have to adjust for the fact that the absolute counts are non-informative (Quinn et al. 2018; Greenacre et al. 2021b). We can instead use relative abundances, or the proportions between features in a sample and several statistical analysis tools like ALDEX and ANCOM already accommodate for that. But how about Machine Learning?

In order to understand the potential problems in compositional data and Machine Learning, we first have to take a look at the mathematical concepts behind compositional data and its properties.

## The Simplex Space

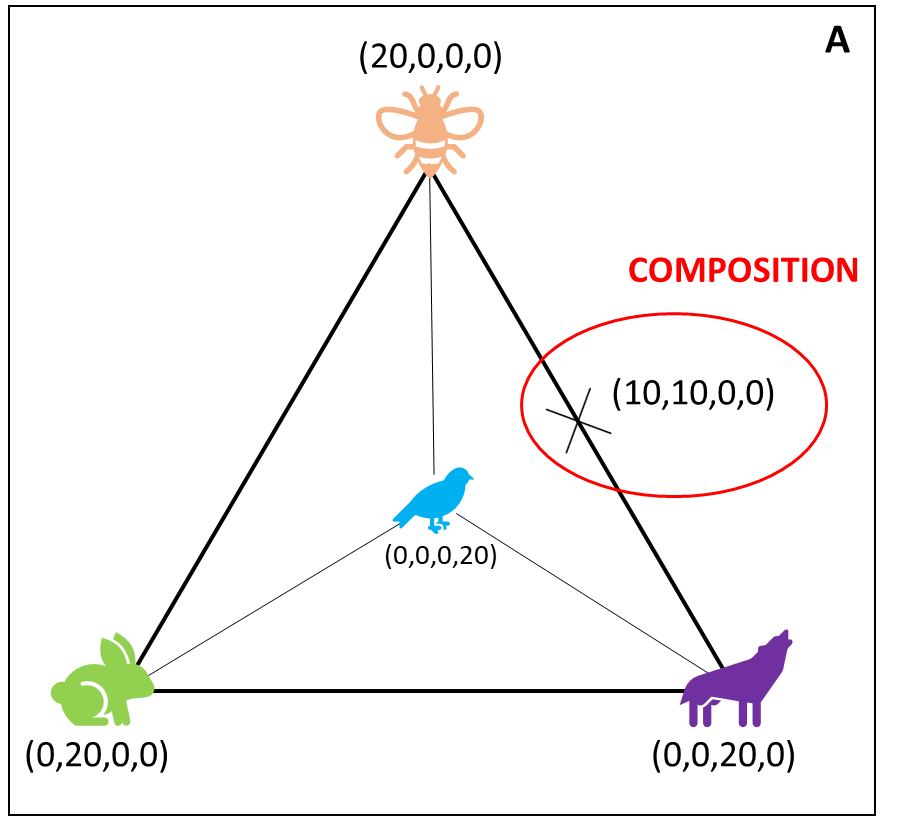
As mentioned, absolute counts in compositional data are unavailable and only relative abundances are of interest. This puts the data in the so-called “simplex space”, instead of the for us more common Euclidean space. The following Figure 2 shows how the data from field A would look like in the simplex space:

Figure 2: Biological Example in the Simplex Space

Assuming the collected ecological data from Figure 1 is compositional, it can be visualized in a S3-Simplex space. Geometrically, a tetrahedron is created with all different components (here animals) placed on the four corners of the polytope. A composition is one possible combination of components confined in the simplex space.

We stick with the ecological example and place all animals as one corner in a geometrical space. With four features, we are able to create a geometric figure called tetrahedron that represents the simplex space visually.

We use the flawed machine, and one day, we sample 20 rabbits in field A. This would lead to a point in the simplex space that sits directly in the left corner, with the coordinates (0,20,0,0), because we only have rabbits, no other animal. Another day, we sample only 20 bees, and no other animals, then we would find our data point where the bee is, at the very top. Marked in red is a sample where the machine counted 10 bees, 10 wolves and no swallows and no rabbits. Every sample round produces one “composition” and the examples show, that the distance between any two variables is sensitive to the presence or absence of other components (Quinn et al. 2018). If a composition is moved from one corner of the animal-simplex, it means that the other values in the composition are directly changed. Consequently, that makes all variables *mutually dependent* on one another and leads to problems in our assumptions about statistical testing. In literature this data is also called “spurious” because it appears as if the data points have a causal relationship when the perceived correlation is only due to data properties. When a composition is moved from bees in the direction of wolves it seems like there is a causal relationship because the increase in the number of wolves, directly decreases the number of bees.

To describe this a bit more mathematically, the problem described above is formally known as “the negative bias problem” (other names are also the constant-sum problem, the closure problem, or the null correlation difficulty) (Aitchison 2003). When the sum of a component is constant, then it can be mathematically proven that the covariance between any two compositions equals 0 (paper Aitchison, geology paper). This has the consequence that some variances would be negative, which is problematic, as variances are by definition always positive. Therefore, negative covariances are presupposed by the limitation of the sum, instead of produced by stochastic factors (Pawlowsky-Glahn and Egozcue 2016; Aitchison 2003). Covariances are similar to correlation coefficients, only that the latter is constraint between -1 and 1, but otherwise they can be interpreted similarly. Translated, a negative covariance means that two factors are negatively correlated, i.e. if one factor goes up, another goes down. Exactly as has been demonstrated by the ecological example.

Thus, using any form of statistical test or machine learning tool seems redundant, as errors in correlations, univariate and multivariate tests are almost preconditioned, and we easily would make false assumptions about the correlation of the data. Thus, a correct handling of compositional data and the simplex space is not optional (Gloor et al. 2017).

## Mapping the Simplex Space into Euclidean Space

In the ecology sampling, the simplex could have been easily overcome by e.g., normalizing to a field size. This indirectly preserves information about the total number of animals. Similarly, it has been tried for sequencing data to calculate an “effective library size” and therefore recover this way the original scale of data. For that, normalization methods like trimmed mean of M-values (TMM) have been introduced, as well as Reads per Kilobase Million (RPKM) and Transcripts per Kilobase Million (TPM) (Quinn et al. 2018). However, all of those methods involve rescaling counts by the library size and these normalizations come with the drawback that some of these methods are sensitive to the removal of low abundant counts, as well as to data symmetry (Quinn et al. 2018).

Furthermore, Aitchison already criticized very early that there is no “magic to open up closed data” (Aitchison 2003), which is what normalization tries to do. Moreover, since information provided in compositional data is essentially about ratios of the components, it seems logical to also think in terms of ratios. Thus, the only way forward is to transform the data in a way that allows us to use it with Euclidean space rules. Aitchison (1982) first proposed several logistic transformations to produce “transformed-normal” models, and later other scientists defined the Aitchison geometry (Pawlowsky-Glahn and Egozcue 2001). The general idea in both cases is, that the simplex space is endowed with a Euclidean space structure. This has several mathematical advantages: if one can map the simplex space into Euclidean space, then all advantages of the Euclidean space can be accessed, i.e., orthogonal projections are possible, the concepts of linear combination, linear dependence, Euclidean distances, as well as all the typical geometrical elements are available (Pawlowsky-Glahn and Egozcue 2016).

Building on top of the Aitchison geometry, methods of analyzing compositional data were proposed by Mateu-Figueras et al., (2011) with the “staying-in-the-simplex” approach or Greenacres (2017) “pragmatic approach”. These methods require a technical understanding of the algebraic-geometric structure of the simplex, therefore they will not be analyzed further in this project. Here, the focus will be more on log-ratio transformations, as they have been more heavily favored in the last decades due to their practicability (Greenacre et al. 2022). They also come with the advantage that they can be easily added to pre-existing Machine Learning pipelines that have been proposed and created over the last years.

## Log-Ratio Transformations

When Aitchison (1982) first tried to overcome the “bounded sum problem” he defined some principals to lay down as fundamentals to the good practice of compositional data. These were important as they paved the way that data could be mapped into Euclidean space correctly.

The most agreed upon and least discussed properties are scale invariance (compositions do not change with e.g., sequencing depth), perturbation invariance (i.e., converting a composition between equivalent units will not change the results), and permutation invariance (i.e., changing the order of the components within a composition will not change the results) (2018 understanding compositional data paper).

It’s clear why these concepts are commonly accepted. Imagine a composition of the ecology example with (8,8,2,2) has been introduced with eight bees, eight wolves, two swallows and two rabbits. Scale invariance means that a transformation should ensure that we get the same ratios even if we had a machine that was able to count to 40 instead of 20 and we would get the composition (16,16,4,4). Similarly, if we would use another unit (instead of counting from one to ten we would count from ten to hundred) should not change the ratios, or if we sampled data in another order (two rabbits, eight wolves, two swallows, eight bees). Aitchison proposed log-normal transformations, because all log-transformations achieve these three properties. Unfortunately, these log-transformations alone do not map the compositional data into Euclidean space. For that a log-ratio transformation is needed.

Three more properties exist that are more heavily discussed in their importance and if they are strictly necessary: isometry (i.e., the distances between ratios should be exactly the same after mapping into Euclidean space), sub-compositional coherence (i.e., overlapping components in two compositions of the same measurement should have the same ratios), and sub-compositional dominance (i.e., using a complete composition carries less information than using the whole (Quinn et al. 2018; Greenacre et al. 2021a; Greenacre et al. 2022, Rivera-Pinto et al. 2018).

Imagine two different scientists observed the same field but scientist A finds eight bees, two swallows, two rabbits and eight wolves, whereas scientist B finds two swallows, eight wolves and two rabbits. Both were not be able to find one animal from the original composition and if one would normalize the results to 1, one would get completely different result (Scientist A: (0.4,0.1,0.1,0,4) and Scientist B: (0.16,0.6,0.16)). However, if ratios are studied, the overlapping components give the same ratios (e.g., the ratio swallow/rabbits in both cases is 1). Given the same example the idea of isometry can also be explained. The distance between eight wolves and two rabbits in the raw compositional data of scientist A is 8-2 = 6, whereas in a log-transformed data set the distance is ln(8)-ln(2) = 1.3 and some advocate that these distances should be preserved in log-ratio transformations. It should be noted that the choice of logarithm is free to choose, as long as the interpretation is adapted respectively. For example, a  
unit increase in the logarithm to base 2 corresponds to a two-fold increase in the original  
magnitude.

One could argue that all mentioned properties are important, unfortunately adhering to all of them has practical implications. The log-ratio transformation that imparts all those properties is called isometric log-ratio (ILR). ILRs are considered the gold standard of log-ratio transformations, as they engender exactly the same multivariate geometric structure of the sample points as that of the formerly mentioned Aitchison geometry (Greenacre et al. 2021b). Ergo, ILR maps a composition in the D-part Aitchison-simplex isometrically to a D-1 dimensional Euclidian vector, which is not just hard to understand but makes it also difficult to interpret (Greenacre et al. 2021a; Greenacre et al. 2022). Additionally, ILRs are particularly problematic when the numbers of components are high as the computational power to calculate ILRs increases linearly (Greenacre et al. 2021b). Thus, considering that microbiome data is usually very high-dimensional, and we strive for interpretable results, ILR is not a good solution for microbiome data.

Text, whiteboard

Description automatically generatedThankfully, there are more types of log-ratio transformation, that are easier to understand and therefore interpret. They do not fully impart sub-compositional coherence and isometry, but it has been mentioned recently that quasi-coherence is sufficient in practice, as well as quasi-isometry (Greenacre et al. 2022), especially in high-dimensional data sets. As a result of this, it was decided to use two types of log-ratio transformation here: ALR (additive log-ratio) and CLR (centered log-ratio).

Figure 3: Equation for CLR

The equation describes the calculation of CLR, with xj as vector of sample features, Dj the total number of features, and g(x) the geometric mean of sample vector x. Log-ratio transformations are applied within a sample (i.e., row-wise).

In CLR, log-ratios are computed between each component and the geometric mean of all components (Gloor et al. 2017). It has the advantage that it is computationally easy to do, which becomes more important with high-dimensional data sets. Furthermore, it reproduces the log-ratio geometry perfectly, but is not sub-compositionally coherent, because the whole composition (i.e., sample) is used to calculate the geometric mean and every sample will therefore use a different geometric mean, as well as every sub-composition. Unfortunately, it is not very easy to interpret and it is not very useful in sparse data containing a lot of zeroes (Gloor et al. 2017).

A picture containing text, watch

Description automatically generatedThe second log-ratio transformation is ALR. Here, the log-ratio is taken of each feature within a composition and divided by a chosen reference feature.

Figure 4: Equation ALR

The equation describes the calculation of ALR, with xj as vector of sample features, D the total number of features, and xref the reference feature. Log-ratio transformations are applied within a feature (i.e., column-wise).

Thus, the interpretation of ALR log-ratios is very straight-forward, as one can interpret a change in all features with respect to the reference. ALR transformed values are also sub-compositional coherent, which is traded for a small loss in isometry. The biggest problem with ALR is the choice of reference. A reference can simply be chosen by the scientist, for example when they want to compare one bacterium to all others. However, if one is unsure about the choice of reference, Greenacre et al. (2021) proposed to use three criteria to find a good reference: (i) the reference component should maximize the Procrustes correlation between the additive log-ratio geometry and the exact log-ratio geometry, (ii) the reference should minimize the variance of log-transformed components, and (iii) it should be a well-populated component.

This means the optimal reference can be determined computationally. Using these guidelines produces additive log-ratios close to being near-isometric, which would make them a favorable log-transformation. The obvious drawback is the computational complexity when a Procrustes analysis is used, which increases especially in higher-dimensional data.

~~In general, log-ratio transformation do not normalize the data (does not “open it”), but makes the interpretation of the transformed data dependent on the reference used and aim for a straight-forward interpretation of the data (Quinn et al. 2018).~~

# The Problems

## Machine Learning

Machine Learning models are of great interest for microbiome analysis, as they allow detection of biomarkers, phenotypes or microbial taxa, as well as other interesting tasks, that are not possible with the standard microbiome tool kit (Marcos-Zambrano et al. 2021). In supervised Machine Learning a model is learned that makes predictions on new data and can identify genomic features that underlie a predictive model (<https://doi.org/10.1038/s41576-021-00434-9>). The hopes are that with the help of AI-based techniques we can transform clinical decision-making and extract more knowledge from the amount of data available to us (<https://doi.org/10.1016/j.artmed.2021.102158>).

However, while modern machine learning algorithms are very powerful, at the moment they come with a great lack of transparency. It is hard to understand why and how the model makes predictions. Especially deep neural networks are considered “black boxes” as the inner workings remain incomprehensible to the outsider and the lack of quality assurances goes against good scientific practice. Thus, more focus is currently put on supervised Machine Learning models like generalized linear models (GLM), support vector machines (SVM) or random forests algorithms (RF), as they are easier to understand and interpret. Still, a correct application and assessment of machine learning models is key to reproducible and interpretable research results.

Now, with compositional data and log-ratio transformations, a new step stone should be introduced to Machine Learning. Some problems become apparent when combining microbiome data, Machine Learning and log-ratio transformations. It has been foreshadowed in the previous section that microbiome data itself is a problem for both log-ratio transformation and machine learning. Microbiome sequencing data is usually very sparse, which means it contains a lot of zeroes and therefore it is left-skewed in its distribution (Gloor et al. 2017).

The zero values are especially problematic because of the log-ratio transformation. The problem is two-fold. First, the logarithm of zero for any type of log is undefined. Secondly, as also ratios are used, placing a zero in the denominator also leads to errors. However, it is quite possible to have features with zero values that have biological relevance, for example when comparing gene expressions between samples. On the other hand, zeros can occur because of technical limitations of the procedure.

For Machine Learning, zeros itself are not that problematic, still the sparse nature of microbiome data affects machine learning models on several technical levels: the most problematic is the increasing complexity of models, as it needs to fit more coefficients or have greater depth to account for all features. It has also been shown that too many features produce noise in the training data which can easily lead to overfitting and negatively impacts the predictive power. Similarly, a machine learning model can also not differentiate between biological or technical zeros and therefore may underestimate the importance of denser features.

Studies in the last years contributed to streamlining Machine Learning pipelines and establishing a good-scientific-practice in order to ensure reproducibility and usability, also specifically for microbiome data (preprocessing papers). Additionally, studies showed how to reduce bias and error by recommending specific cross-validation procedures (CV paper, etc.) and the influence of sample size has been tried to estimate for various Machine Learning models (data size papers). Concerning feature size, some rules of thumb have been proposed by Hua et al. (2005). For uncorrelated features, the optimal feature size is N−1 (with N = sample size) and for highly correlated features . However, these recommendations come from the computer science community and may not be applicable to biological models.

To this day, it is still unclear if and how log-ratio transformation impact the performance in a prediction task. In general, log-ratio transformations seem to outperform raw proportions for classification tasks, but it is not clear how log-ratio transformations relate to the changes in predictive performance and if other factors in a Machine Learning pipeline are affected by log-ratio transformations.

## CODACORE

Therefore, an alternative solution to standard machine learning models could be *CoDaCoRe*. Published by Gordon-Rodriguez et al in 2021 in the paper “Learning Sparse Log-Ratios for High-Throughput Sequencing Data”, *CoDaCoRe* is a novel learning algorithm for finding balances (Compositional Data via Continuous Relaxations). Balances are defined as the log-ratios between geometric means of two features of the input variables. Translated, *CoDaCoRe* finds ratios between groups of features that are explanatory for the given classification task. Such ratios are commonly used as biomarkers of gut health e.g., the Firmicutes-to-Bacteroidetes ratio (Crovesy et al., 2020; Magne et al.; 2020).

As *CoDaCoRe* is able to combine more than two features in its log-ratios, it leads to a richer set of features and therefore more flexible models, compared to e.g. ALR, which only uses one reference. Finding balances is usually very computationally taxing, however, in *CoDaCoRe* Gordon-Rodrigues et al. use a deep learning technology called “continuous relaxation” and only approximate the optimization problem, which has the advantage of greatly reducing the runtime.

Thus, *CoDaCoRe* identifies a sequence of balances, in decreasing order of importance, each of which is sparse and interpretable. This makes it a promising algorithm that is created to also work efficiently on big data sets with a lot of features and could be a good alternative to counter the problems that microbiome data brings into Machine Learning and log-ratio transformation. In their paper, the authors compare *CoDaCoRe* against several machine learning models (Lasso, RF and XGBoost) and show that their algorithm does not sacrifice interpretability nor predictive accuracy.

# Implementation/Assessment

Recent years made it clear that machine learning is a tool that should be available to all biologists, but comes with high complexity and its own pitfalls, even without the addition of mathematical characteristics of compositional data and log-ratio transformations. The above-described examples already mention that machine learning models in combination with log-ratio transformations do not show clear-cut results. Several solutions like *CoDaCoRe* and *selbal* have been introduced by the scientific community, where they showed that they work more efficient than machine learning models. The goal of this project focuses on collecting insights on the performances of machine learning models in direct comparison with *CoDaCoRe*, but also comparing the influence of transformations on performance. Additionally, the goal should be also practicality. At the end it would be good to give some practical guidelines on how to employ machine learning models with compositional data.

As mentioned in the introduction, two log-ratio transformations seem to be the most promising, in terms of applicability and interpretability: ALR and CLR. Additionally, it has been decided to compare ALR transformation in three ways: (i) a random reference will be picked as denominator, to assess the average performance of machine learning models for ALR (random ALR), (ii) find the most optimal denominator (optimal ALR) and (iii) worst ALR denominator (worst ALR).

A standard normalization for raw relative abundances is total sum scaling transformation (TSS). In theory, it removes technical bias related to different sequencing depth by dividing each feature count with the total library size. As mentioned in the introduction, Aitchison criticized these normalization techniques very early on and therefore its impact on Machine Learning performances will be directly compared to chosen log-ratio transformations. Finally, all Machine Learning performances are compared to *CoDaCoRe* in the following conceptual Diagram

Description automatically generatedframework:

Figure 5: Pipeline to compare model performances

The figure shows the proposed pipeline for effective comparison of machine learning performances. Data sets used are chosen by their characteristics large/small, high/low correlations and continuous/discrete variables. Afterwards, data sets will be filtered by abundances and imputed. Microbiome-native methods will be employed and compared to the data being log-transformed and used in machine learning models.

The general pipeline will be constructed of the following building blocks: Pre-processing, Imputation, Transformation, and Machine Learning Models/Microbiome Approaches. Usually, model performances are averaged after several repeats with random training and test splits. However, it was decided here to use the same training/test split for all transformations to directly compare the difference.

As representation for Machine Learning models, Generalized Linear Model (*glmnet*) and Support Vector Machine (*svmRadial*) are used. *Glmnet* represents a common linear machine learning model and *svmRadial* a non-linear one, as several papers already showed different performances for microbiome data depending on the choice of model (papers).

For *CoDaCoRe*, it was decided to train two models per run, one with lambda = 0 and the other with lambda = 1. In *CoDaCoRe* the parameter lambda controls the regularization strength of the model. Lambda = 1 applies the 1-standard-error rule in the discretization step of the log-ratio. This is typically a good choice, leading to models that are both sparse and predictive. Lambda = 0 corresponds to a 0-standard-error rule, in other words choosing the log-ratio that minimizes cross-validation score (Gordon-Rodriguez et al. 2021). Using both models allows for a direct comparison in how *CoDaCoRe* handles different types of microbiome data sets, that are typically very sparse. Additionally, *CoDaCoRe* allows in prediction tasks to include all found log-ratios or just the most descriptive one. It was decided to use both to see how influential one found log-ratio can be for the chosen data sets or if it is better to use several log-ratios.

Performance assessment scores are necessary to compare all models directly. For discrete parameters (e.g., two binary classes like “healthy” and “non-healthy”) AUROC will be used to assess model performances. AUROC tells about the model’s ability to discriminate between cases (positive examples) and non-cases (negative examples). An AUROC above 0.8 generally means excellent discriminatory ability for the model, whereas an AUROC of 0.5 corresponds to a coin flip (i.e., a useless model).

For continuous parameters (e.g., BMI, which can have any real number) a root mean squared error (RMSE) is used for performance evaluation. It is the [standard deviation](https://www.statisticshowto.com/probability-and-statistics/standard-deviation/) of the [residuals](https://www.statisticshowto.com/residual/). Residuals are a measure of how far from the regression line data points are and a RMSE consequently measures how spread out these residuals are. A value of 0 would indicate a perfect fit to the data. Contrary to the AUROC, the lower the RMSE, the better the model.

As several authors pointed out (Quinn and Erb 2020; Gloor et al. 2017), machine learning performance is influenced by data size. Therefore, three data sets were chosen accordingly to include direct comparison of performances of small and large data sets, as well as high and low known correlations between microbiome and host. There is no explicit definition for small datasets in the literature. However, several studies classified data sets ranging from 18 to 1030 samples as small data sets (Althnian et al. 2021) and this guideline is used here, too.

This leads to two small (CRC and PCOS) and one large data set (EstMB) for this project. The CRC data set was first used and described by (Wirbel et al. 2019) in their meta-analysis for colorectal cancer. This data set is well known and contains 7727 features with 695 samples. It shows clear correlations between gut microbiota and colorectal cancer and is therefore helpful to show the behavior of transformations and machine learning algorithms on small but highly specific data sets. The second data set is the Polycystic Ovary Syndrome (PCOS) data set described by Kreete et al. (2020). It observed 312 individuals, with two-thirds of them being healthy, and 72738 features. It is a valuable addition as it is a small data set that shows no correlation between the disease and microbiome structure (paper Kreete). Lastly, is the Estonian Biobank microbiome cohort (EstMB). This data set includes 2509 individuals with several phenotypical markers collected over time and 17180 features overall, which classifies it as a big data set and it contains two phenotypes that are of value for this project: Hypertensive heart disease (HHD) which has a lower to moderate correlation with gut microbiota and Diabetes Type 2 (DT2) which is known to show a higher correlation to gut microbiome compositions (<https://doi.org/10.1038/s41586-022-04567-7>, other papers?).

It has been decided to remove the majority of zeros in all data sets via abundance filtering. This reduced the runtime of computational procedures and also by sub-compositional coherence, this should have no effect on the performance of machine learning models.

* Repeat PAP
* All features Genus in all data sets?

# Methodology

For the data analysis and model pipelines, the script language R (v4.1.3) in combination with RStudio (v2022.02.1+461) has been used. For data cleaning and filtering the main library is “tidyverse” (1.3.1). Imputation was conducted with “zCompositions” (1.4.0.1) (cmultRepl), and transformations were mostly done with “easyCODA” (0.34.3). Models were constructed with “mikropml” (1.2.2), “tidymodels” (0.2.0) and “codacore” (0.0.3).

Additionally, scripts were created for convenience purposes. All scripts can be found on Github JenniferNeumaier/ml\_coda.

## Pre-Processing

First, all data sets were cleaned in order to remove NAs in predictor columns or patients that have no sequencing data. In EstMB data set, 21 rows removed in metadata due to NA and 21 patients respectively cut out of abundance table. This leads to 2485 final sample-size. In CRC, 128 rows were removed due to NA in feature “BMI”, leading to 567 samples overall. In PCOS, 6 rows were removed in the abundance table because no matching patient has been found in metadata, reducing the number of samples to 304.

As discussed, the sparse nature of microbiome data leads to several problems. Therefore, filters were applied to all three data sets. In this benchmarking project, taxa with ≤10% abundance in samples will be discarded. Additionally, a filter of ≤50% abundance in samples will be applied, as well as a mean relative abundance filter for 0.001. For 10% abundance filters CRC keeps 650 features and PCOS 1154 features. Respectively, for 50% abundance filters CRC keeps 189 features and PCOS 120 features. For EstMB data, 90% abundance was used, as the data was otherwise not practically usable without heavy computational power. Even with 90% filtering, EstMB keeps 3062 features.

## Imputation

One of the first steps after filtering and before log-ratio transformation is zero-imputation. Introduced by (Palarea-Albaladejo and Martín-Fernández 2015) is the library “zCompositions”. All three data sets were imputed with Geometric Bayesian Multiplicative (GBM) and output form “p-counts”.

## Transformations

Greenacre et al. (2021) proposed a computational solution of finding an ALR reference. Included in the package “easyCODA” is the function ALR() that assesses the abundances and variances of features in a data matrix, followed by a Procrustes analysis to assess their geometry. This leads to a list of possible good denominators for the respective data set if the top results are chosen or worst denominators, if the bottom results are selected. Similarly, “easyCODA” also contains the function CLR() to compute the centered log-ratio. For TSS() a custom function has been created.

## Machine LEarning models

The focus will be on two standard machine learning models, that are already incorporated in easy-to-use packages in R: generalized linear models (GLMs) and support vector machines (SVM) as non-linear approach. Both are available in the function run\_ml() from the package “mikropml” by Schlosslab (paper). This package nicely compacts the use of standard machine learning models to a few lines of code and supports the use of GLMs (*glmnet*), as well as SVMs (*svmRadial*).

“Tidymodels” was used split the data into 80% training set and 20% test set. A seed was used to ensure that the data splits were the same for every repeat and following test. The split data has then been fed into run\_ml() and codacore(), respectively. Both functions then split the training set into 5 folds and assessed via 10-fold repeated cross-validation the best model. Here, run\_ml() also tunes hyperparameters for glmnet and svmRadial, respectively. In *CoDaCoRe* lambda is not tuned, but instead two models are trained, one with lambda = 0 and the other with lambda = 1.

The best model then is used for predictions on the test set and its AUROC/RMSE scores saved. The functions run\_ml() and codacore() were repeated 10x, which allows to assess statistical fluctuations in a specific data split.

For svmRadial the number of repeats had to be lowered, otherwise the training of the model would have taken too long. It has been decided to use a 2-fold repeated cross-validation and only 5 repeats overall.

# Results

## Influence of Transformations on ML Performances

The most important research question of this thesis was if and how log-ratio transformations influence machine learning performances. The in section () introduced pipeline has been applied to all three data sets and the results separated into binary classification and regression setting results.

### Model performances in a Binary Classification Setting

Figure 6 shows the culminated results of the presented pipeline and combined in one figure to show the differences of machine learning and transformations on several data sets directly.

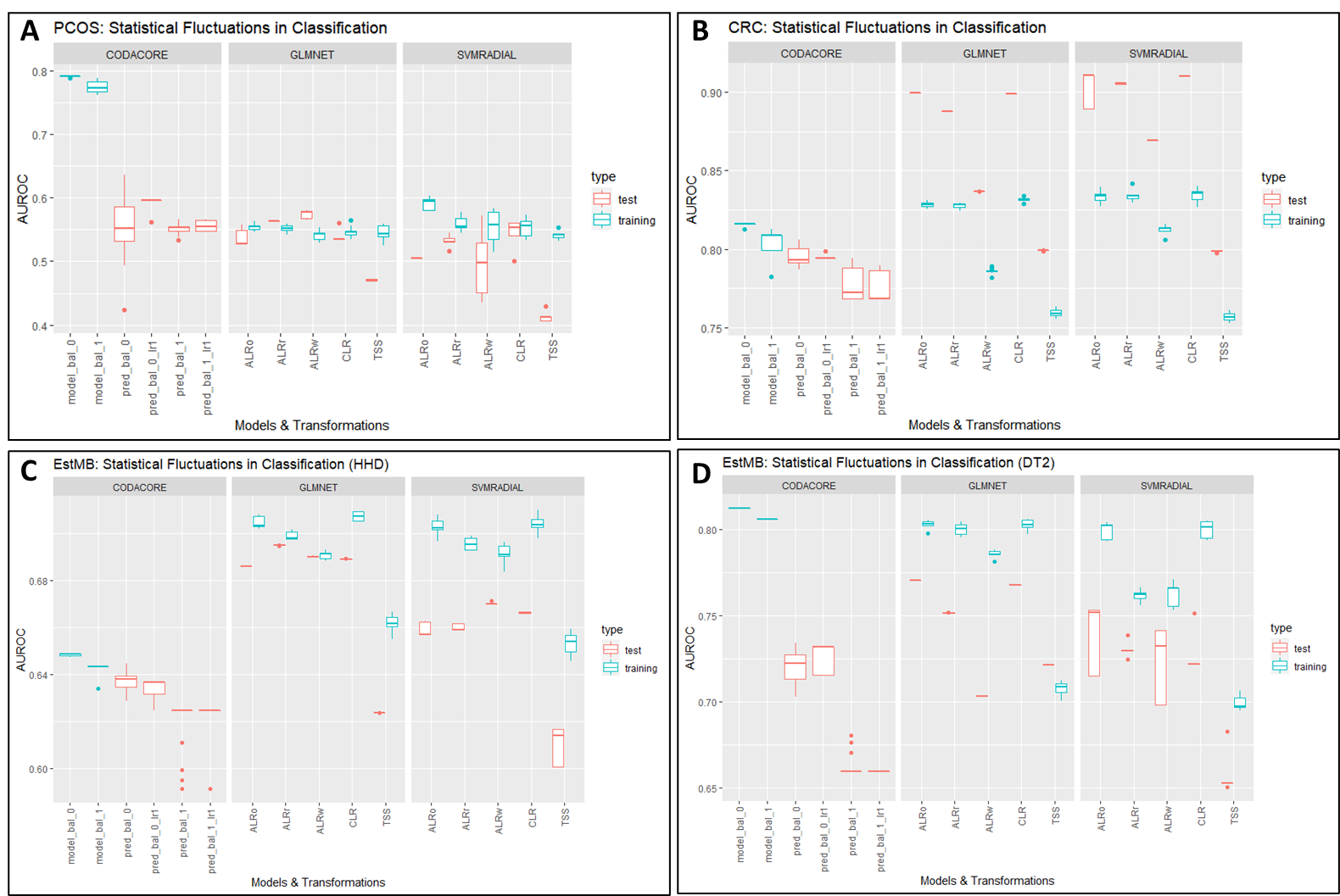


Figure 6: Statistical Fluctuations in Classification

CRC and PCOS use 10% abundance filtered sets, EstMB uses a 90% abundance filtered set. The y-axis contains AUROC scores. The x-axis contains models and transformations, with the most left always being *CoDaCoRe*, the middle portion displaying performances of glmnet and the right svmRadial. In red test performances are plotted and in blue training performances. model\_bal\_0 and model\_bal\_1 correspond to *CoDaCoRe* models with lambda 0 and lambda 1, respectively. Consequently, pred\_bal\_0 and pred\_bal\_0\_lr1 correspond to predictions with all found log-ratios and only the most descriptive log-ratio for model\_bal\_0. The same nomenclature applies to pre\_bal\_1 and pred\_bal\_1\_lr1. ALRo, ALRw and ALRr correspond to the types of ALR transformation applied in this pipeline (optimal, worst and random). A contains data from PCOS data set predicting PCOS phenotype. B contains data from CRC data set predicting CRC phenotype. C and D both contain data from EstMB data set, with C predicting for hypertensive heart disease (HHD) and D for diabetes type 2 (DT2).

For the glmnet models, training data was split into 5 folds, a 10-fold repeated cross-validation has been applied and the models repeated 10 times. For the svmRadial models, training data was split into 5 folds, a 2-fold repeated cross-validation has been applied and the models repeated 5 times. For *CoDaCoRe*, training data was split into 5 folds, a 2-fold repeated cross-validation has been applied and the models repeated 5 times.

As expected for PCOS (Fig.6A), the AUROC scores for test and training performances vary from 0.5 to 0.6 and correspond to the low correlation between microbiome and PCOS. *CoDaCoRe* severely overfits the data set with the test performances corresponding to the performances of both *glmnet* and *svmRadial*.

Contrary to that, Fig. 6B shows the performance of models in a data set with a high correlation between the microbiome and colorectal cancer. All models show good to excellent performances, with ALRo and CLR showing the best performances. ALRr shows surprisingly good results, which is repeated in Fig. 6C and D for the EstMB data set. Although log-ratio transformation not necessarily improve the predictive performance in low correlation data set (Fig. 6A), TSS still shows the lowest performances in all three data sets, independent of correlation or sample size. This trend is also visible in regression settings for all three data sets (see Supplementary).

It is surprising, that *CoDaCoRe* shows lower test and training performances, comparable to ALRw glmnet (Fig. 6B) or TSS (Fig. 6C). In general, its results are very inconclusive throughout different types of data. In the PCOS data set (Fig. 6A) it visibly overfits the data, although its test performances are comparable to glmnet and svmRadial. This trend is also visible in regression settings (see Supplementary). In high correlation data sets, independent of sample size, it does not manage to perform better than Machine Learning models, performances are still considered good to excellent with an AUROC of 0.8 (Fig. 6B and 6D).

In the CRC data set, all machine learning models severely underfit the data, which usually indicates a bias in the data. A possible reason could be the “leakiness” in the pipeline. It is considered good practice to not conducted pre-processing on the whole data set (Whalen et al. 2022). Although imputation and transformation are handled independently from pre-processing in this pipeline, both can be considered part of pre-processing. Indeed, section () shows that splitting the data set before imputation and transformation solves the underfitting in CRC.

Still, the trend depicted in CRC with ALR and CLR performing best, is also found in EstMB data sets, for both moderate correlation (HHD) and high correlation (DT2) between microbiome and disease, and to a small degree in PCOS. However, linear models seem to perform better than non-linear, although that difference is smaller in high-correlation data sets.

~~For HHD~~ *~~CoDaCoRe~~* ~~shows lower performances, corresponding to TSS performances in glmnet and svmRadial. In DT2~~ *~~CoDaCoRe~~* ~~visibly overfits and also shows a big discrepancy between lambda 0 and lambda 1 predictions, with lambda 0 models showing better performances.~~

~~Interestingly, in all data sets and all models that use different transformations, TSS seems to be clearly underperforming, closely followed by ALRw. This is especially visible in high correlation data sets like CRC and EstMB DT2. However, also in low (PCOS) and moderate correlation (EstMB HHD) data sets, models using TSS normalized data significantly underperforms~~. Considering the effect of transformations even in PCOS, it seems to be a clear indication, that log-ratio transformations can indeed affect the performance in machine learning models. Additionally, *CoDaCoRe* only performs as good as log-ratio transformed data, but never better.

## INFLUENCE OF Transformations on Data Leakage

As mentioned in the section above, data leakage is an easy to overlook problem and already lead to performance biases in the section before, seemingly more in smaller data sets. Information leakage occurs when information from the test set is used to pre-process the training set. This is for example happening when the imputation algorithm uses the whole data set, which includes the test set, to replace zeroes (Whalen et al. 2022). To assess the impact of data leakage when adding log-ratio transformations, this section directly compares the model performances of *glmnet* and *CoDaCoRe* of a leaky and non-leaky pipeline.

In theory, the chosen log-ratio transformations ALR and CLR should not lead to any kind of data leakage. CLR transformations are conducted row-wise and therefore do use information between samples. For ALR, if the analysis for the best reference is conducted on the training set and applied on the test set, then data leakage should also not occur. The same applies if a random reference is chosen.

In the non-leaky procedure, processed data is split into train and test set at the start and imputation and transformations are applied separately on the split data. In the case of ALR, the reference was found on the training set and directly applied to the test set. Finally, train and test set are fed into *glmnet* model. The non-leaky procedure for *CoDaCoRe* works similarly but the transformation step is skipped.

In the leaky procedure, imputation and transformations are conducted on the whole data set and afterwards the data is split into train and test set and fed into the *glmnet* model. Again, for *CoDaCoRe* the transformation step is skipped.

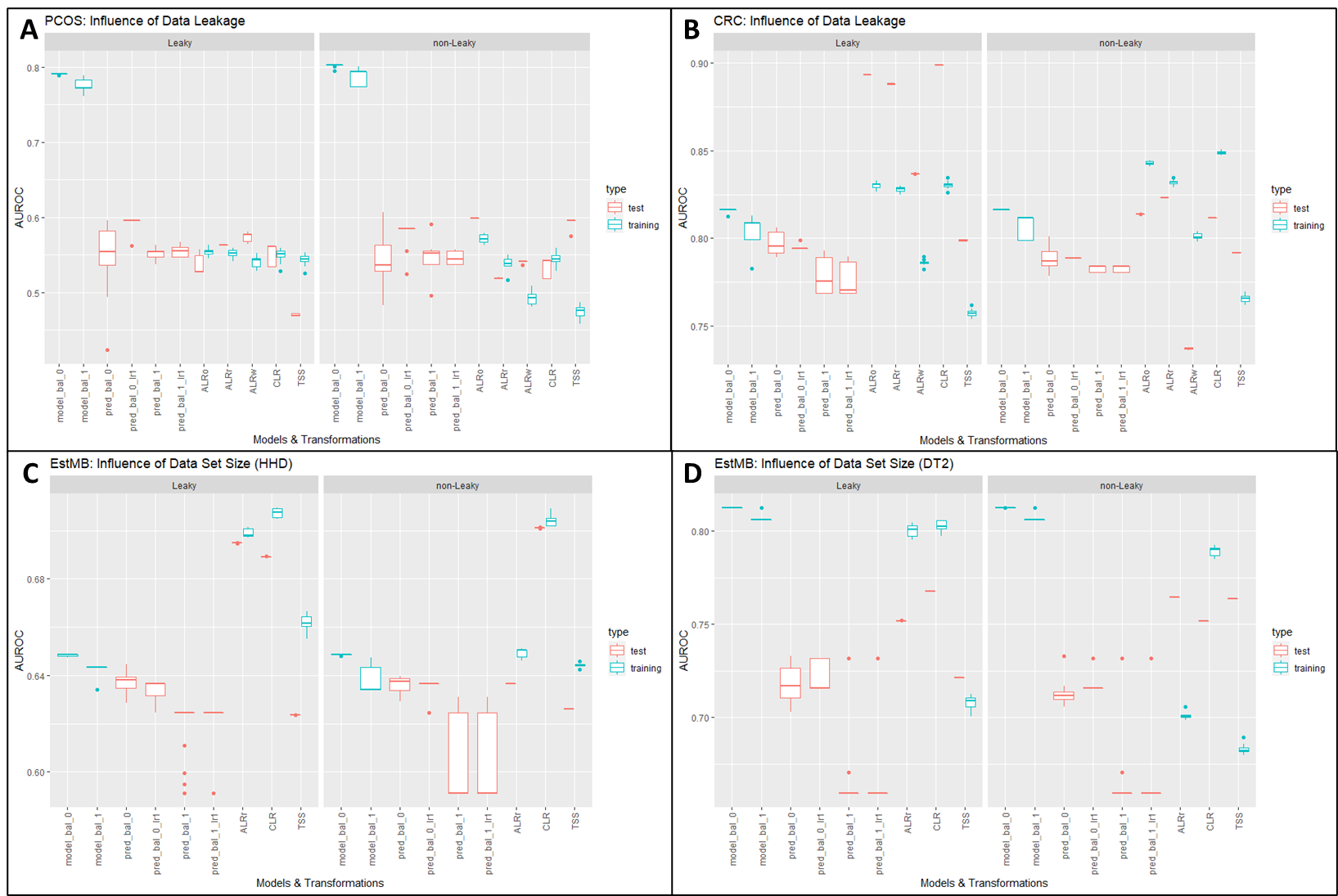


Figure 7: Influence of leaky and non-leaky pipelines on log-ratio transformation

CRC and PCOS use 10% abundance filtered sets, EstMB uses a 90% abundance filtered set. The y-axis contains AUROC scores. The x-axis contains models and transformations, with the most left always being *CoDaCoRe,* followed by *glmnet* results. For EstMB ALRo and ALRw have been skipped because it was too computationally taxing to repeat a Procrustes Analysis. In red test performances are plotted and in blue training performances. model\_bal\_0 and model\_bal\_1 correspond to *CoDaCoRe* models with lambda 0 and lambda 1, respectively. Consequently, pred\_bal\_0 and pred\_bal\_0\_lr1 correspond to predictions with all found log-ratios and only the most descriptive log-ratio for model\_bal\_0. The same nomenclature applies to pre\_bal\_1 and pred\_bal\_1\_lr1. ALRo, ALRw and ALRr correspond to the types of ALR transformation applied in this pipeline (optimal, worst and random). A contains data from PCOS data set predicting PCOS phenotype. B contains data from CRC data set predicting CRC phenotype. C and D both contain data from EstMB data set, with C predicting for hypertensive heart disease (HHD) and D for diabetes type 2 (DT2).

For the glmnet models, training data was split into 5 folds, a 5-fold cross-validation has been applied and the models repeated 5 times.

The results for the leaky pipelines were not repeated and the description for the results can be read in section ().

It becomes apparent that the results for *CoDaCoRe* do not change significantly if imputation is conducted on the whole data set or not. For CRC (Fig. 7B), imputation on the whole data set potentially leads to higher variances in test performances compared to splitting the data before imputation. However, a non-leaky pipeline does not increase performances for *CoDaCoRe* and does not solve the overfitting.

The results for Machine Learning models are a bit more mixed. The results in Fig. 7B show that the underfitting in the CRC data set was most likely due to biases in the pipeline. As soon as a non-leaky was applied, the Machine Learning models are not underfitting the data anymore. The training performances for CRC are very similar to the leaky pipeline, but the test performances have been adjusted. A contrary picture is shown for TSS in PCOS and EstMB DT2 (Fig. 7A and 7D). There, the model is now heavily underfitting the data, whereas no significant changes are seen in Fig. 7B (CRC) and 7C (EstMB HHD) for TSS.

Some trends can be observed for all three data sets: CLR shows very stable performances (sans the underfitting in CRC), which is especially visible in EstMB performances. ALR and TSS are the most impacted by the changed pipeline. In all data sets TSS increases the difference between its test and training performance, independent if it has been under- or overfit the data. In Fig. 7A and B ALRo marginally underfits the data but otherwise shows the highest performances out of all transformations.

Even though a leaky pipeline has been applied in the section before, the trends applied by log-ratio transformations still holds. In both pipelines, log-ratio transformations outperform standard normalization techniques and *CoDaCoRe* is not outperforming standard Machine Learning models.

Nevertheless, in all further tests, the non-leaky pipeline will be applied.

## Do log-ratio transformations influence test set methodology?

The pipeline used to compare all data sets and their performances includes a standard data set split of 80/20, meaning 80% of data goes into the training set, 20% of data into the test set. However, the CRC data set has been designed with holdout sets in mind and the original paper uses this technique (Wirbel et al. 2019).

In Machine Learning theory, a holdout differs from a test set as it provides a final estimate of the machine learning model’s performance after it has been trained and validated by a test set. In the CRC paper they called it leave-one-study-out (LOSO) validation, where data from one study was set aside as an external validation set, while the data from the remaining 4 studies was pooled as a training set (Wirbel et al. 2019).

The study design gives a good opportunity to see if the influence of log-ratio transformations is impacted by the choice of test set methodology. The following figure compares *glmnet* and *CoDaCoRe* performances with a standard 80/20 or Holdout set.

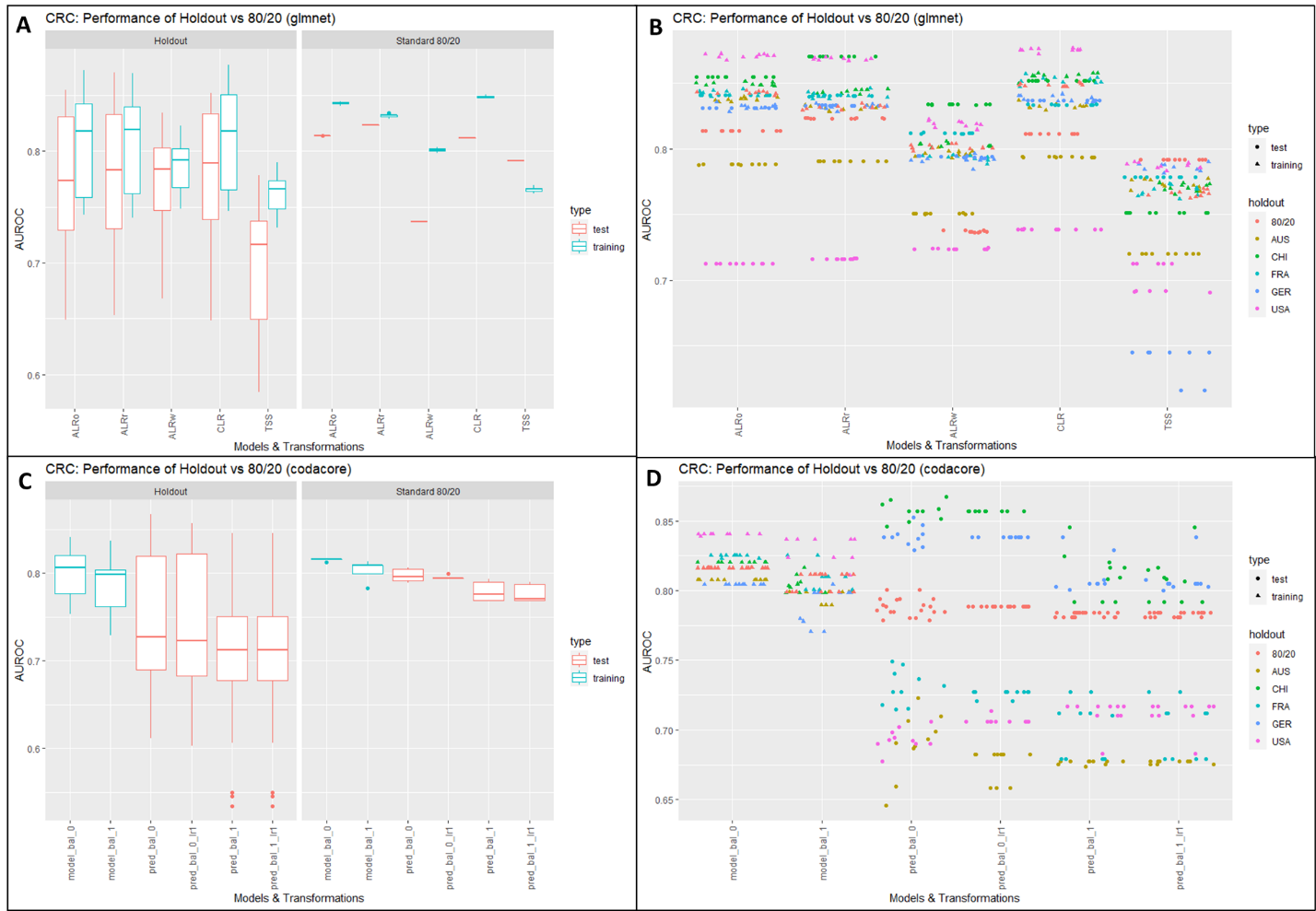
****

Figure : Behavior of log-ratio transformations with test set methodology

CRC uses 10% abundance filtered set. The y-axis contains AUROC scores. The x-axis contains models and transformations, with A and B belonging to glmnet models and C and D belonging to *CoDaCoRe*. In red test performances are plotted and in blue training performances. model\_bal\_0 and model\_bal\_1 correspond to *CoDaCoRe* models with lambda 0 and lambda 1, respectively. Consequently, pred\_bal\_0 and pred\_bal\_0\_lr1 correspond to predictions with all found log-ratios and only the most descriptive log-ratio for model\_bal\_0. The same nomenclature applies to pre\_bal\_1 and pred\_bal\_1\_lr1. ALRo, ALRw and ALRr correspond to the types of ALR transformation applied in this pipeline (optimal, worst and random). A contains data from CRC data set predicting CRC phenotype and B plotting all test and training performances in detail. The same applies in C and D for CRC *CoDaCoRe* performances.

For the glmnet models, training data was split into 5 folds, a 10-fold cross-validation has been applied and the models repeated 10 times.

Fig. 8A and 8B show the difference of test performances as boxplots for different transformations of the same data set and for *CoDaCoRe*, respectively. All five holdout sets have been used separately as test set once and the results pooled. It can be seen clearly that for all transformations, the holdout sets show a lot more variance in the test set performances, compared to the standard 80/20 split, which shows almost no variance at all and arguably higher test performances. This is also visible for *CoDaCoRe* performances.

In Fig. 8B and 8D it becomes clear why the test performances of the holdout sets have such a high variance. For all transformations and holdout sets, the test performance is quite stable. Marked in pink dots is the test set performance when USA has been chosen as holdout set. For all ten repeats of the model, the test performance is stable around 0.72. However, choosing CHI as the holdout set produces test set performances of 0.85. This shows that the high variances in Fig. 8A are not due to statistical fluctuations, but the choice of holdout set. Splitting the data set in an 80/20 manner leads to a performance somewhere in the middle. This is also true when using *CoDaCoRe,* although the test set performances fluctuate more strongly.

Additionally, log-ratio transformations show a similar picture as in the sections above. The performances of ALRo, ALRr and CLR are higher than TSS and ALRw, even when using holdout sets. *CoDaCoRe* doesn’t change its performance.

~~The idea to verify a model by gathering extra test data is sensible, however as seen by the results, it easily introduces a bias and for beginners in machine learning it would be more beneficial to stick to 80/20 data splits.~~

## Data set size

As mentioned in the introduction, the dimensionality and sparse nature of microbiome data is problematic for both Machine Learning models and log-ratio transformations. A common technique to reduce the runtime for pre-processing and zeroes contained in the microbiome data, an abundance filter is applied, or the features are selected via ordination techniques (sources). For all tests above, a 10% abundance filtering has been used, as it is very common in microbiome and ecology (source) and the performances especially for highly correlated data seems promising.

However, it would be of great advantage if the number of features could be reduced even further, with Machine Learning performances staying stable. Therefore, the following test will assess how machine learning models perform if a 50% abundance filter is applied. Additionally, *CoDaCoRe* performances will be again compared to *glmnet*.

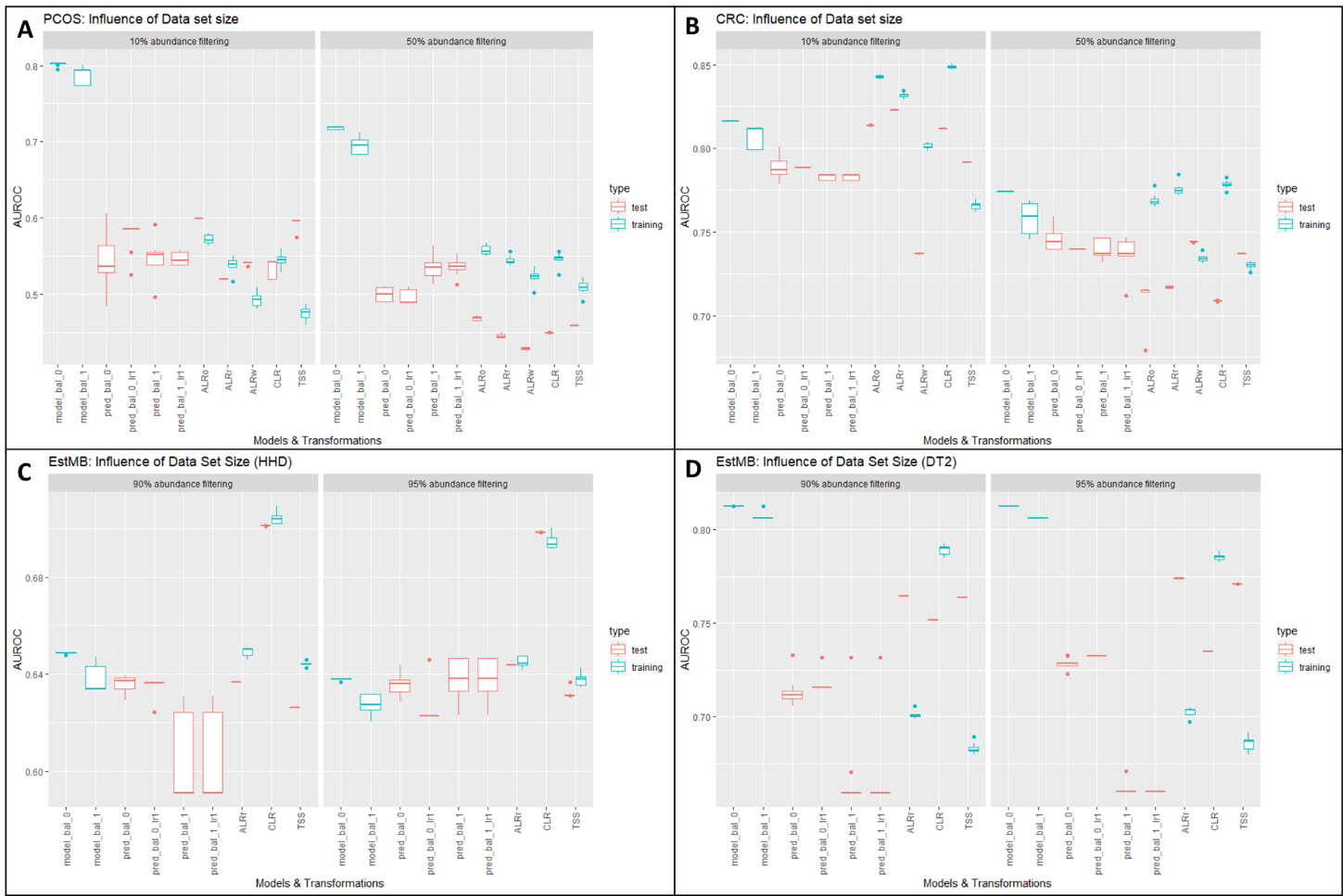


Figure 9: Influence of Data set size

CRC and PCOS use 10% and 50% abundance filtered sets, EstMB uses a 90%/95% abundance filtered sets. respectively. The y-axis contains AUROC scores. The x-axis contains models and transformations, with the most left always being *CoDaCoRe*, the middle portion displaying performances of glmnet and the right svmRadial. In red test performances are plotted and in blue training performances. model\_bal\_0 and model\_bal\_1 correspond to *CoDaCoRe* models with lambda 0 and lambda 1, respectively. Consequently, pred\_bal\_0 and pred\_bal\_0\_lr1 correspond to predictions with all found log-ratios and only the most descriptive log-ratio for model\_bal\_0. The same nomenclature applies to pre\_bal\_1 and pred\_bal\_1\_lr1. ALRo, ALRw and ALRr correspond to the types of ALR transformation applied in this pipeline (optimal, worst and random). A contains data from PCOS data set predicting PCOS phenotype. B contains data from CRC data set predicting CRC phenotype. C and D both contain data from EstMB data set, with C predicting for hypertensive heart disease (HHD) and D for diabetes type 2 (DT2).

For the glmnet models, training data was split into 5 folds, a 10-fold repeated cross-validation has been applied and the models repeated 10 times. For *CoDaCoRe*, training data was split into 5 folds and the models repeated 5 times.

The results for 10% abundance filters were not repeated and the description for the results can be read in section ().

Reducing the number of features has visible impacts on smaller data sets (Fig. 9A and B). For PCOS, *glmnet* is overfitting the data more pronounced compared to 10% filtered data sets and in CRC, the decrease in performance is clearly visible (Fig. 9B), accompanied by heavier overfitting.

The impact on bigger data sets is smaller, however. In Fig. 9C a 95% abundance filter leads to less distance between test and training performances compared to 90% filtering, whereas these distances increase slightly in Fig. 9D. In general. though the performances are very similar.

Again, although the performances in CRC (Fig. 9B) are decreasing, the impact of log-ratio transformations are still visible. ALRw and TSS still perform worse compared to ALRo, ALRr and CLR. Especially the latter is also very stable in EstMB data sets.

The performances of *CoDaCoRe* are not heavily affected, only in CRC do they decrease with a decrease in features.

# Discussion/Conclusion

Several tests have been conducted throughout this whole project to determine machine learning performances with different transformation. Overall, it can be stated that log-ratio transformation indeed influence machine learning performances. (Quinn and Erb 2020) stated that the “choice of log ratio transformation does not impact performance”. where they compared CLR and several balance procedures in LASSO. The results in this project support this statement: In low and moderate correlations, the choice of transformation does not improve performances significantly. Furthermore, in high-correlation data the choice of transformation is almost detrimental as CLR and optimal ALR show similar good performances. Surprisingly, choosing a random feature of ALR shows good performances, which would support recent claims that strict-isometry may not be necessary for machine learning purposes (Greenacre et al. 2022). Therefore, choosing a random ALR feature could be beneficial as it is not as computationally taxing to conduct a full Procrustes analysis to calculate to optimal ALR, especially in bigger data sets. It should be noted however, that ALR random could lead to performances as indicated by ALR worst, with a chance of 1/number of features. This probability becomes more redundant the bigger the data set. Additionally, the impact of a bad reference feature on the performance seems to be bigger in smaller data sets. In CRC, ALR worst shows performances close to TSS, whereas in EstMB data sets the performance is closer to the other ALR performances. This would support the hypothesis by Greenacre et al. (2022) that deviation from isometry (and sub-compositional coherence) might become more diluted in higher-dimensional data.

https://www.frontiersin.org/articles/10.3389/fgene.2022.784397/full :  
- supports results that transformations improve ML performances vs no normalization  
- variance/distribution modifiers and scaling method were significantly better than no normalization.  
- supports claim that choice of transformation is irrelevant (ILR gives same results as CLR)  
- added tests for batch effects (here, no data sets were combined, but their results suggest that using zero-centering can easily be used with transformations)  
- claim that non-linear models show higher performances -> cannot say that with my data however non-linear models could just need more time and data to be more conclusive -> difficult to compare them directly to glmnet

* Also imputed data with zCompositions  
  - interestingly: they also applied 10% abundance filter  
  - also recommend to use genus, as microbiome pipeline  
  - Feature selection was performed with the training set, followed by normalization and batch effect correction with the respective methods to both the training and test sets

Data leakage:

Two factors probably influence the data leakage in the pipeline: imputation and transformations. In the most ideal setting, the data is split into training and test set before imputation and transformation procedures are conducted. The limiting factor on why this is not always possible, is the data size. Splitting data into train and test set is conducted row-wise, which can make especially the test set so small, that imputations are not possible. In the case of transformation, this problem is not as critical, especially when using ALR and CLR.

Interestingly, although EstMB data sets were conducted with the same pipeline as CRC, the models to not overfit the data (comparison Fig. 6B and D). Preliminary results suggested that although variance decreases as training set size increases, this effect does not transfer to the bias. It seems that in bigger data sets the bias introduced by conducting imputation on the whole data set is significantly reduced, which is logical as the underlying distributions that are assumed by the imputation algorithms are probably more detailed.

Therefore, imputation also seems to be the most impacting factor for data leakage when using log-ratio transformations. The scientific community is consistently discussing over the impact of imputation on data analysis (sources), therefore it has been decided to also test how the model performances are influenced if the imputation is conducted on the whole data set.

* Supplementary: only split-imputation shows same results as split-imputation/transformation

The effect of transformations on machine learning performances persists throughout additional tests in the pipeline and its influence is impervious to changes in data leakage and test methodology, at least for ALR and CLR. For CLR, this is expected as CLR is it is by nature conducted sample-wise. In the case of ALR, if the reference feature is calculated by Procrustes analysis on the training set, the transformation is also not affected by data leakage. If this is not possible because of the sample size of the data set, actively choosing a reference feature (or choosing it randomly) does not need any data and can therefore be a valid option.

<https://www.idescat.cat/sort/sort441/44.1.8.Coenders-Pawlowsky-Glahn.pdf>  
For how to interpret log-ratio models

When used as explanatory variables, additive log-ratios are not interpreted as in-

creasing a component at the expense of reducing the last component, as their for-

mulation suggests, but as increasing a component at the expense of reducing all

other components.

• When used as explanatory variables, centred log-ratios are not interpreted as in-

creasing a component at the expense of reducing all other components, as their

formulation suggests, but as increasing a component at the expense of reducing

the component whose log-ratio is omitted.

Classification in Machine Leaning is a challenging task by itself, and it becomes more challenging when dealing with small datasets, as the limited size of training data can lead to unreliable and biased classification models (Althnian et al. 2021). This is important in the biological context, as sample sizes are always limited, and it should be accounted for when using machine learning algorithms. It has already been shown that a high sample-size is not necessary in high-correlation data sets (CRC), which is supported by several other studies in recent years (Althnian et al. 2021; van der Ploeg et al. 2014; Bailly et al. 2022). CRC also supports findings that smaller data set sizes are sufficient for good logistic regression models (Bailly et al. 2022). Furthermore, the non-linear models in some cases still show very high variance in test set performances. (van der Ploeg et al. 2014) found that SVM and NN show instabilities in smaller data sets and mentions that SVM may need over 10 times as many events per variable to achieve a stable AUC compared to linear models. This could also explain the higher variances in SVM performances throughout all data sets.

Literature says sample size around 100 for test set is sufficient for stable test set performances (https://www.sciencedirect.com/science/article/pii/S0003267012016479?via%3Dihub)

Although *CoDaCoRe* shows good results in high-correlation data like CRC and EstMB (DT2), it does not outperform machine learning models. In data sets with low to moderate correlation it severely overfits the data. These results set a contrast to the results in the paper that published *CoDaCoRe* (Gordon-Rodriguez et al. 2021)*.* Of course, in this project the performances in the same data split are directly compared, in contrary to the original paper, that evaluated the performances over 20x random train/test splits. The performances are also not corrected by imputing the data sets on split data as seen in section() and no bias has been introduced with 80/20 split (see section()).

A possible explanation for the different performances could be explained by another important variable for good machine learning performances: the number of features. As can be seen in section (), there seems to be a clash between compositional data theory and machine learning practice, as the performances show a decrease in performance after 50% abundance filtering, compared to 10% abundance filtering.

As mentioned before, sub-compositional is – from a scientific point of view – a very important factor in compositional data. For the sake of reproducibility, it should be possible to get comparable results, even if only a sub-composition is chosen. Therefore, when data is filtered in pre-processing sub-compositional coherence should make it possible that statistical analysis results are similar. Greenacre et al. (2022) confirmed as much when they used several 50% sub-compositions from the same data set and showed that they had very low variability. In the case of abundance filtering, both data sets are sub-compositions of the original, with only keeping features that show up more abundant and should therefore inherit similar characteristics.

More concretely, for 10% abundance filters CRC keeps 650 features and PCOS 1154 features. Respectively, for 50% abundance filters CRC keeps 189 features and PCOS 120 features. With 90% filtering, EstMB keeps 3062 features. Calculating correlation coefficients (see supplementary) shows that only a small number of correlation coefficients show significantly different values between 10% and 50% filtering, which suggests a similar composition for both data sets.

The drop in model performance is also true for *CoDaCoRe*, which is especially interesting as a qualitative analysis (see supplementary) shows that in both 10% and 50% filtering *CoDaCoRe* finds very similar and overlapping log-ratios, but still the model performance, specifically in CRC, is decreasing. This again suggests that finding balances for *CoDaCoRe* is quite stable, independent of feature size and would support the sub-compositional coherence in the data sets.

Filtering also has a practical approach, as it reduces computational time drastically as 10% abundance filtering removes over 90% of data in both PCOS and CRC data sets. Without filtering, using imputation techniques is not possible as some columns contain only zeros and finding balances with *CoDaCoRe* is not possible when the data set contains zeros as well.

Number of features in ML:

<https://academic.oup.com/bioinformatics/article/21/8/1509/249540>  
- for uncorrelated features optimal feature size is N-1  
- for correlated features sqrt(N)  
- if number of features = number of samples model will overfit

<https://www.biorxiv.org/content/10.1101/2020.12.11.422279v1.full.pdf>  
- advocates to stop removing rare ASVs  
- shows that it can have effects on downstream analyses  
-

Suggestions for feature selection before machine learning have been proposed, also in compositional data. Most of them target the search for main factors that maximally explain log-ratio variances (Greenacre et al. 2022). However, such techniques have been introduced in this project with *CoDaCoRe* and also ALR reference selection. As shown, *CoDaCoRe* does not improve performances although it supposedly found predictive balances and the top 20 ALR reference choice differs greatly between 10% and 50% filter of the data set (see supplementary).

These findings suggest that the drop in performance may be of quantitative nature, instead of qualitative. Machine Learning needs a lot of data, and this seems to hold true even if highly predictive. But it also suggests that using compositional data correctly does not solve machine learning problems. This is also supported by the test results in section () and ().

# Conclusion

The problems and solutions discussed in the section before lead to some key notes on transformations in the context of machine learning: (1) Transformations are not able to conjure correlations when there are none, (2) any log-ratio transformation is better than none.

This comes with several advantages when combining compositional data with machine learning, as transformations can be easily integrated into existing machine learning guidelines. The previous section showed that the majority of problems found in this project was due to the character of machine learning algorithms and not compositional data itself.

When using Machine Learning with sequencing data several things should be taken into consideration:

Data set size: for both high and low correlation data, a feature number around 1000 or higher seems to be efficient enough to produce meaningful machine learning performances and it can be lower for high-correlation data as seen in the CRC data set.

Pipeline: Proposed Pipelines for correct Machine Learning application hold true also for compositional data and log-ratio transformations can slot themselves easily into the assortment of pre-processing procedures that should be conducted before data splits.

Log normalization is essential: https://www.frontiersin.org/articles/10.3389/fimmu.2021.677870/full#f4

Choice of Transformation: ALR is the easiest to interpret and should therefore preferred over CLR, although it is computationally more taxing. Using a random feature seems to also be a viable choice. In recent years, a lot more studies suggested to use Spike-Ins as reference feature (sources). Especially in 16S, it is common to add to samples to control the density on flow cells and could therefore work as a standard reference feature in the future and would negate the problem with Procrustes analysis.

Further problems:

As Imputation and Filtering techniques seem to be the most prevalent problems in Machine Learning, further tests would be necessary, if those steps change the geometry of compositional data. Calculating correlation coefficients before and after imputation (see supplementary) does not suggest that imputation changes sub-compositions drastically. Furthermore, it has already been shown to improve machine learning performances when genus level representation was used and all features belonging to the same genus merged through PCA before machine learning (Jasner et al. 2021). Additional tests on different filtering techniques could improve machine learning performances even more.

* Learning curves

Additionally, recent papers showed that imputation may not be necessary in compositional data. Greenacre et al. (2022) claimed that no “ratio” is necessary for isometric behavior and that chi-square normalizations or Box-Cox transformations can also be used as transformation, which would lead to more interpretable results and doesn’t need imputation as no ratio is used.

Furthermore, especially from the geology corner, data-driven alpha-transformations are currently investigated that enables one to deal with the presence of 0s in the compositions, unlike the log-ratio approach which is only suitable for strictly positive compositions.

<https://hal.archives-ouvertes.fr/hal-03379935v2/document>

* More different train/test splits to compare performances
* Including other models like xgboost
* Correct pipeline from the beginning
* Additional feature selection tools
* Additional imputing methods/no imputing methods -> catalogue of zCompositions is big, deciding which one is the right one is difficult -> my choice does not show big changes in data set (correlation coefficients)

Publication bibliography

Aitchison, J. (1982): The Statistical Analysis of Compositional Data. In *J. Royal Statistical Society* 44, pp. 139–177.

Aitchison, J. (2003): A Concise Guide to Compositional Data Analysis. Girona, Italy.

Althnian, Alhanoof; AlSaeed, Duaa; Al-Baity, Heyam; Samha, Amani; Dris, Alanoud Bin; Alzakari, Najla et al. (2021): Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain. In *Applied Sciences* 11 (2), p. 796. DOI: 10.3390/app11020796.

Bailly, Alexandre; Blanc, Corentin; Francis, Élie; Guillotin, Thierry; Jamal, Fadi; Wakim, Béchara; Roy, Pascal (2022): Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models. In *Computer methods and programs in biomedicine* 213, p. 106504. DOI: 10.1016/j.cmpb.2021.106504.

Gloor, Gregory B.; Macklaim, Jean M.; Pawlowsky-Glahn, Vera; Egozcue, Juan J. (2017): Microbiome Datasets Are Compositional: And This Is Not Optional. In *Frontiers in microbiology* 8, p. 2224. DOI: 10.3389/fmicb.2017.02224.

Gordon-Rodriguez, Elliott; Quinn, Thomas P.; Cunningham, John P. (2021): Learning Sparse Log-Ratios for High-Throughput Sequencing Data.

Greenacre, Michael; Grunsky, Eric; Bacon-Shone, John (2021a): A comparison of isometric and amalgamation logratio balances in compositional data analysis. In *Computers & Geosciences* 148, p. 104621. DOI: 10.1016/j.cageo.2020.104621.

Greenacre, Michael; Grunsky, Eric; Bacon-Shone, John; Erb, Ionas; Quinn, Thomas (2022): Aitchison's Compositional Data Analysis 40 Years On: A Reappraisal. Available online at http://arxiv.org/pdf/2201.05197v1.

Greenacre, Michael; Martínez-Álvaro, Marina; Blasco, Agustín (2021b): Compositional data analysis of microbiome and any-omics datasets: a revalidation of the additive logratio transformation.

Hua, Jianping; Xiong, Zixiang; Lowey, James; Suh, Edward; Dougherty, Edward R. (2005): Optimal number of features as a function of sample size for various classification rules. In *Bioinformatics (Oxford, England)* 21 (8), pp. 1509–1515. DOI: 10.1093/bioinformatics/bti171.

Jasner, Yoel; Belogolovski, Anna; Ben-Itzhak, Meirav; Koren, Omry; Louzoun, Yoram (2021): Microbiome Preprocessing Machine Learning Pipeline. In *Frontiers in immunology* 12, p. 677870. DOI: 10.3389/fimmu.2021.677870.

Marcos-Zambrano, Laura Judith; Karaduzovic-Hadziabdic, Kanita; Loncar Turukalo, Tatjana; Przymus, Piotr; Trajkovik, Vladimir; Aasmets, Oliver et al. (2021): Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment. In *Frontiers in microbiology* 12, p. 634511. DOI: 10.3389/fmicb.2021.634511.

Palarea-Albaladejo, Javier; Martín-Fernández, Josep Antoni (2015): zCompositions — R package for multivariate imputation of left-censored data under a compositional approach. In *Chemometrics and Intelligent Laboratory Systems* 143, pp. 85–96. DOI: 10.1016/j.chemolab.2015.02.019.

Pawlowsky-Glahn, V.; Egozcue, J. J. (2001): Geometric approach to statistical analysis on the simplex. In *Stochastic Environmental Research and Risk Assessment* 15, pp. 384–398.

Pawlowsky-Glahn, Vera; Egozcue, Juan José (2016): Spatial analysis of compositional data: A historical review. In *Journal of Geochemical Exploration* 164, pp. 28–32. DOI: 10.1016/j.gexplo.2015.12.010.

Quinn, Thomas P.; Erb, Ionas (2020): Interpretable Log Contrasts for the Classification of Health Biomarkers: a New Approach to Balance Selection. In *mSystems* 5 (2). DOI: 10.1128/mSystems.00230-19.

Quinn, Thomas P.; Erb, Ionas; Richardson, Mark F.; Crowley, Tamsyn M. (2018): Understanding sequencing data as compositions: an outlook and review. In *Bioinformatics (Oxford, England)* 34 (16), pp. 2870–2878. DOI: 10.1093/bioinformatics/bty175.

van der Ploeg, Tjeerd; Austin Peter C; Steyerberg Ewout W (2014): Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. In *BMC Medical Research Methodology* 14, pp. 137–150.

Whalen, Sean; Schreiber, Jacob; Noble, William S.; Pollard, Katherine S. (2022): Navigating the pitfalls of applying machine learning in genomics. In *Nature reviews. Genetics* 23 (3), pp. 169–181. DOI: 10.1038/s41576-021-00434-9.

Wirbel, Jakob; Pyl, Paul Theodor; Kartal, Ece; Zych, Konrad; Kashani, Alireza; Milanese, Alessio et al. (2019): Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. In *Nature medicine* 25 (4), pp. 679–689. DOI: 10.1038/s41591-019-0406-6.

# Supplementary

### Regression Setting

A similar conclusion can be drawn for a regression setting. Figure 8 shows the culminated results for all three data sets in a regression setting. Contrary to the classification setting, all three data sets contained BMI as the continuous variable.

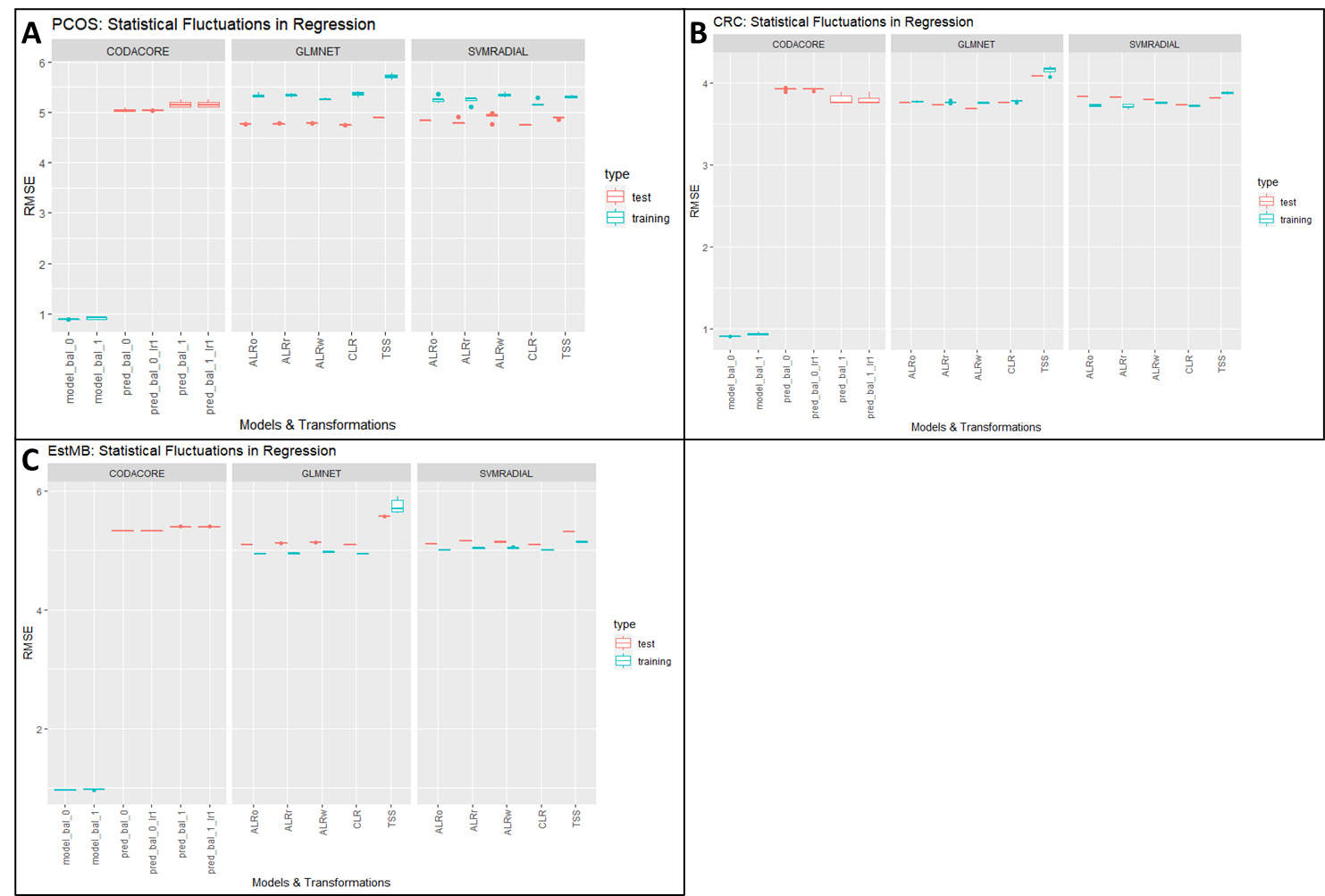
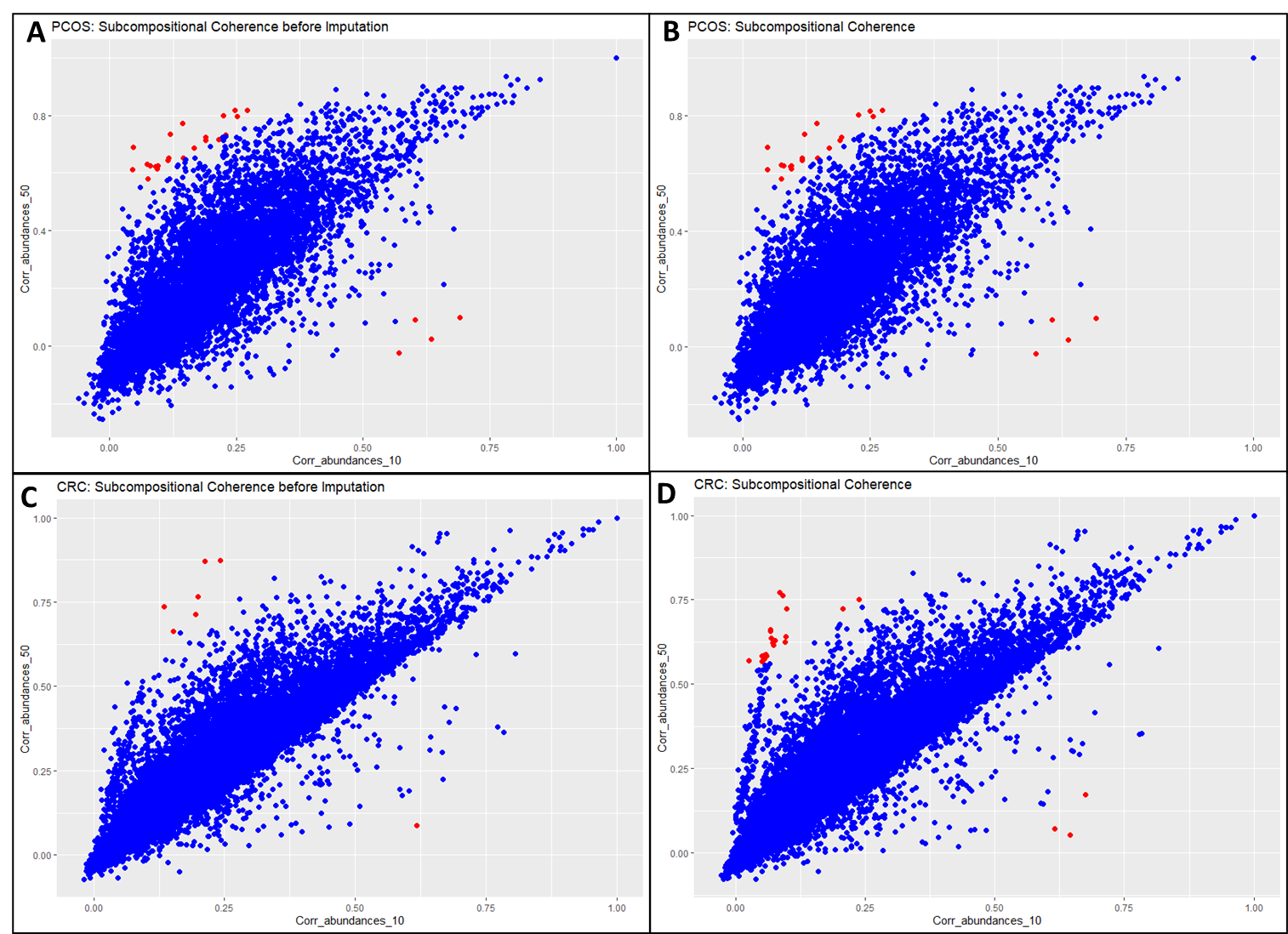


Figure : Statistical Fluctuations in Regression

CRC and PCOS use 10% abundance filtered sets, EstMB uses a 90% filtered set. The y-axis contains RMSE scores. The x-axis contains models and transformations, with the most left always being *CoDaCoRe*, the middle portion displaying performances of glmnet and the right svmRadial. In red test performances are plotted and in blue training performances. model\_bal\_0 and model\_bal\_1 correspond to *CoDaCoRe* models with lambda 0 and lambda 1, respectively. Consequently, pred\_bal\_0 and pred\_bal\_0\_lr1 correspond to predictions with all found log-ratios and only the most descriptive log-ratio. ALRo, ALRw and ALRr correspond to the types of ALR transformation applied in this pipeline (optimal, worst and random). C and D both contain data from EstMB data set, with C predicting for hypertensive heart disease (HHD) and D for diabetes type 2 (DT2).

It is very interesting that all three data sets, having a completely different bioinformatics and sample pipeline in the background and containing very different abundances for species in their data set, show a very coherent picture for BMI prediction. All data sets, PCOS (8A), CRC (8B), and EstMB (8C) show the same behavior over all three models, with RMSE scores between 4.0 and 5.0. PCOS shows the biggest difference between test and training performances which can probably be contributed to the data set size, as CRC and EstMB are significantly bigger data sets, and both show closer test and training performances. In all three data sets, TSS again shows the worst performance, however, more pronounced in glmnet models compared to svmRadial. *CoDaCoRe* is again visibly overfitting.

# Correlation Coefficients



# Presence-Absence-Pattern

