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)

# 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 if a scientist should make the effort in learning about machine learning methods, when automated algorithms perform well enough, and no heavy prior machine learning knowledge is necessary.

Working with mathematical concepts is always a bit out of the comfort zone for most biologists. Unfortunately, with technical improvements and big data encroaching in our field, and statistical methodology being an essential part in data analysis, 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 different mathematical characteristics than other data types. Furthermore, as machine learning concepts become more widespread and useful, their performance in combination with compositional data and its necessary transformations have not been fully analyzed.

The goal of this master thesis is an attempt in making 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. However, 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).

# Characteristics of Compositional Data

In order to define and illustrate the concept (and problems) 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, which is already valuable information in itself. 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 (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 (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 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 the flow cells and even the most effective sequencing machines could never fully sequence the entirety of the organisms RNA 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 concept of “Compositional Data”, first introduced by John Aitchison (1982), roughly 40 years ago. Because 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.

## The Simplex Space

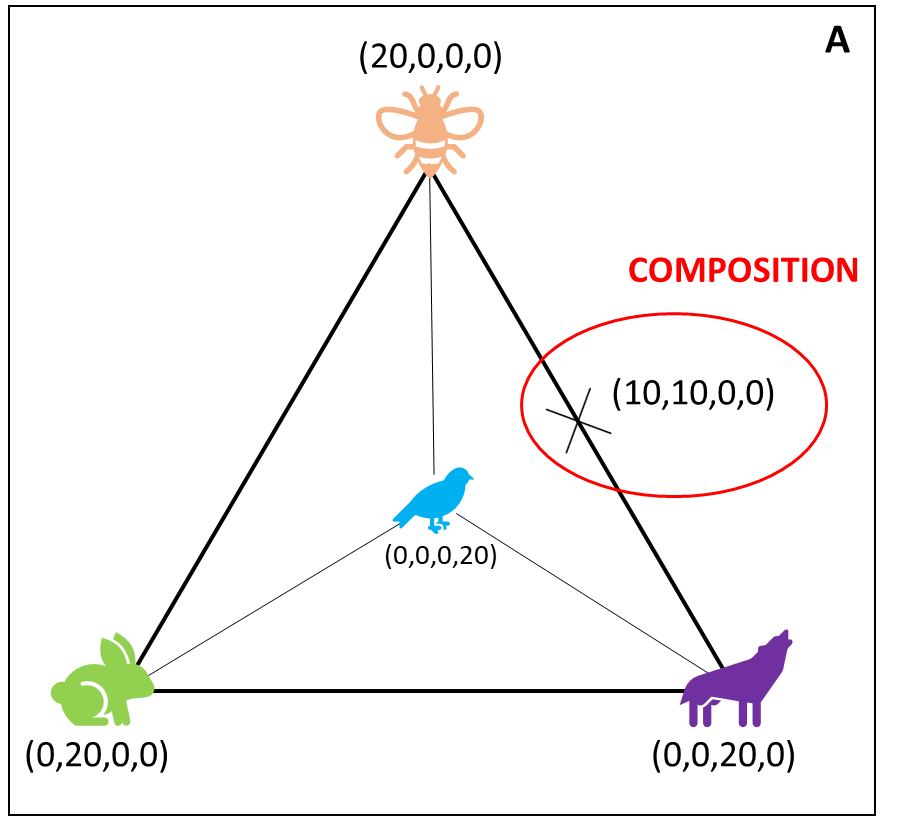
As mentioned, absolute counts in compositional data are irrelevant and only relative abundances are of interest. This puts the data in the so-called “simplex space”, instead of the for us more common Euclidean space. The following Figure 2 shows how the data from field A would look like in the simplex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

Figure : Biological Example in the Simplex Space

Assuming the collected ecological data from Figure 1 is compositional, puts it 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 our ecological example and place all our animals as one corner in a geometrical space. With four features, we are able to create a 3-Simplex and a geometric figure called tetrahedron (otherwise called a pyramid).

We use our 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 bees 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 directly influences the other values in the composition. Consequently, that makes all variables *mutually dependent* on one another and leads to problems in our assumptions about statistical testing. In statistical literature this data is also called “spurious” because it appears as if the data points have a causal relationship. 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), which is the main reason why we have a dependency problem. When the sum of a component is constant, then it can be mathematically proven that the covariance between any two compositions equals 0. This has the consequence that some variances would be negative, which is problematic, as variances are 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).

Additionally, it is commonly assumed - and all experiments are created to accommodate these assumptions – that data is collected IID: independent and identically distributed. The IID assumption is important for e.g., the central limit theorem, Markov sequence, hypothesis testing in general and of course machine learning. Having such an obvious violation in compositional data can have serious consequences on the reproducibility of results. In life sciences, count data are usually modelled using the Poisson distribution or negative binomial, because using anything else would imply that negative and non-integer counts would exist, which is biologically not feasible (Quinn et al. 2018).

Using any form of statistical test or machine learning tool seems redundant, as a type two error is 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

The difficulty of confined data points has already been commented on by Pearson (1897) in the context of spurious correlations and has been taken up by Aitchison 1982 in an attempt to overcome the “bounded sum problem”.

In our ecology sampling, we could easily overcome the simplex by e.g., normalizing to a field size from the start, to indirectly preserve informations 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 RPKM and 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, first by Aitchison (1982) with several logistic transformations proposed to produce “transformed-normal” models, and later with the definition of the Aitchison geometry (Pawlowsky-Glahn and Egozcue 2001). The general idea is, that the simplex space is endowed with a Euclidean space structure, which 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”. In this master thesis, mentioning these methods is as far as I will go here, because they require a technical understanding of the algebraic-geometric structure of the simplex. Here, I will focus more on log-ratio transformations, as they have been more heavily favored in the last decades due to their practicability (Greenacre et al. 2022).

## Log-Ratio Transformations

There are several types of log-ratios, which were proposed over the last 40 years, and I want to take the time and introduce them. Some more in detail than others, as not every log-ratio transforms the data perfectly and it is important to point out here, that there are still ongoing discussions about which log-ratio transformation is preferrable over the other in terms of accuracy, complexity, and interpretability (Greenacre et al. 2022; Quinn and Erb 2020; Rivera-Pinto et al. 2018).

In general, all log-ratio transformations capture the relationship between the features in the data set and taking the logarithm of these ratios makes the data symmetric and linearly related. It moves the simplex into real space and imparts key properties on the data set: 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).

Two more important properties exist that are transformation-specific: sub-compositional coherence (i.e., scientists A and B get identical results for components when these components are included in compositions) and sub-compositional dominance (i.e., using a subset of a complete composition carries less information than using the whole) (Quinn et al. 2018; Greenacre et al. 2021a; Greenacre et al. 2022). From a scientific standpoint, it seems to be a no-brainer to try to adhere to both of these properties, as they are the gold standard of reproducibility. Not following sub-compositional coherence would mean that two sequencing runs from the same patient (and the same bioinformatics pipeline) couldn’t be compared and ignoring sub-compositional dominance would mean we couldn’t filter data before using.

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). The 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, they are also particularly problematic when the numbers of components are high (Greenacre et al. 2021b), which is a quality worth considering as sequencing data is usually very high-dimensional.

Text, whiteboard

Description automatically generatedThankfully, we have more types of log-ratio transformation, that are easier to use and interpret. They do not fully impart sub-compositional coherence, but interestingly, it has been shown 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 in this master thesis: ALR (additive log-ratio) and CLR (centered log-ratio).

Figure : 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).

The CLR uses the geometric mean of the whole composition as the reference feature (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. Unfortunately, it is not very easy to interpret and it is not very useful in sparse data containing a lot of 0s (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 measurement within a composition and divided by a chosen reference feature.

Figure : 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 and it is also sub-compositional coherent, which is traded for a small loss of isometry. The biggest problem with ALR has always been the choice of reference. When choosing a 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. Using these guidelines produces additive log-ratios close to being isometric, which would make them a favorable log-transformation. The obvious drawback is the computational complexity (if a Procrustes analysis is used beforehand), 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 Problem

## Machine Learning

Predictive methods such as random forests (RF), artificial neural networks (ANN), deep learning (DL) or support-vector machines (SVM) and other methods have become in the last years increasingly popular (Tolosana-Delgado et al. 2019). Vital for good machine learning conclusions is the balancing of predictive power with the explainability, similarly to log-ratio transformations.

In terms of statistical analysis, machine learning models are of great interest for microbiome analysis, as they allow predictions 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). Therefore, a correct application of machine learning models is key to reproducible and interpretable research results. However, it is still unclear if any log-ratio transformation improves the performance in a prediction task.

Several studies showed log-ratio transformations in machine learning models with mixed performances (Zhang and Shi 2019; Coenders and Greenacre 2021). In 2019, Zhang and Shi compared several machine learning algorithms on geological compositional data and showed that overall, RF was the best performing model and that ILR and CLR were superior to ALR (Zhang and Shi 2019). Tolosana-Delgado et al. (2019) concluded that ridge regression and SVM both need ILR. More observations were also made by Quinn et al. 2020. They performed linear discriminant analysis (LDR) on ILR-transformed data and partial least squares (PLS) to CLR-transformed data and showed good predictive results (Quinn and Erb 2020). Neural Networks require further research, but does not seem to be equivariant (Tolosana-Delgado et al. 2019), i.e. not any log-ratio transformation works similarly well.

These observations demonstrate the current predicament between compositional data and machine learning. Log-ratio transformation in linear and generalized linear models are not easily chosen and depend heavily on the observations at hand. 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. Furthermore, employing log-ratio transformations leads to an increase in complexity in the correct application of machine learning models. Thus, it is of increasing importance to create a practical guide for all scientists who want to employ such analysis.

~~The question arises if machine learning models are “worth the hassle” considering microbiome-specific algorithms like~~ *~~CoDaCoRe~~* ~~exist, that are faster and do not require a lot of background knowledge to use. The next section describes this algorithm more in detail and their potential effectiveness.~~

## CODACORE

An alternative solution to standard machine learning models as described in the section before, 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 two 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).

Balances are essentially pairwise log-ratios; however, they allow the aggregation of more than one variable in the numerator and denominator of the log-ratio. This leads to a richer set of features and therefore more flexible models. Usually, pairwise log-ratios are computationally very taxing, which is why they are not separately included in the master thesis. 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. 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.

Thus, four log-ratio transformations will be compared: CLR (centered log-ratio), and three ALR methods (worst, random and optimal). As mentioned in the introduction, those two seem to be the most promising, in terms of applicability and interpretability. Including worst and random ALRs should give insights of potential repercussions of using ALR wrong, or at least not in the most ideal way.

The performances will be directly compared to TSS-transformed data (total sum scaling transformation) and *CoDaCoRe* in the following conceptual Diagram

Description automatically generatedframework:

Figure : Used Pipeline

The graph shows the proposed pipeline for the benchmarking project. Data sets will be collected by their characteristics large/small, high/low correlations and continuous/discrete variables. Afterwards, data sets will be pre-processed by zero-imputation methods and filtering. 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. The core idea is to observe statistical fluctuations in two chosen machine learning models, given the same training data set. It has been decided to use a Generalized Linear Model (glmnet) and a Support Vector Machine (svmRadial) as machine learning models representation. Glmnet represents a common linear machine learning model and svmRadial a non-linear one.

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.

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.

# 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 libraries is “tidyverse” (1.3.1). Imputation was conducted with “zCompositions” (1.4.0.1), 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.

### Data set

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, as well as continuous and discrete predictive variables. 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.

First is a Colorectal Cancer (CRC) set. 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. All data sets contain at least one discrete and one continuous response variable.

### 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. ~~Additionally, the column “X.1” has been removed as it is only a sum of all abundances per row.~~ 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 microbiome data usually has a lot of features, the computational work can be taxing. 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 main problems of microbiome data is its sparse nature. When working with relative abundances this is annoying but doesn’t have any mathematical consequences. When using log-ratio transformation however, zeros lead to problems as one: the log(0) is undefined and secondly: when working with ratios, zeros cannot be in the denominator. Therefore, one of the first steps after filtering and before log-ratio transformation is zero-imputation. Introduced by (Palarea-Albaladejo and Martín-Fernández 2015) is pseudocount. It has been frequently used for statistical analysis of microbiome data. It adds a pseudo-count of 1E-05 to avoid non-finite values resulting from log(0). All three data sets were imputed with Geometric Bayesian Multiplicative (GBM) and output form “p-counts”.

### Transformations

As mentioned in the introduction, choosing a log-ratio is not an easy decision. In order to stay with the goal of improving quality management and reducing human error, ILR will left out, as it is the most difficult one to work with and interpret. Similarly, pair-wise log-ratio transformations will also not be tested, as it is included in *CoDaCoRe*. It has been decided to use TSS (total sum scaling transformation), which is a standard approach to normalize relative abundance data and compare it to CLR and ALR transformed data.

As ALR would be the most promising log-ratio transformation in terms of interpretability and its closeness to ILR, we will 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) via Greenacre et al. (2021) proposed way of finding a 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.

### Machine LEarning models

Therefore, the focus will be on 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.

Before modeling, the data set is split 80/20 with stratification with the package “tidymodels”. The train set given to the function run\_ml() from “mikropml” by source schlosslab. 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). As shown in the pipeline, it was of interest to control the initial split into test and train data, which is also allowed by mikropml. The training set is then split into 5 folds and the best model assessed via 10-fold repeated cross validation and the final test and training scores of the best model are saved for plotting. This procedure is repeated 10x for each model and each data set to assess statistical fluctuations of model performances and accuracies.

As it was of interest to compare the machine learning model performances to *CoDaCoRe*, the pipeline includes *CoDaCoRe* directly. For this algorithm, the filtered and imputed data set is used, without any transformation. The training set that is fed to the function codacore() contains the same samples as the other machine learning models. Two *CoDaCoRe* models are trained, one with lambda = 0 and the other with lambda = 1. This creates four performance scores per repetition and saved for further plotting. The codacore() function is also repeated 10x to catch statistical fluctuations under the same data split. For discrete response variables, AUC is chosen as performance score, and for continuous response variables RMSE.

# Results

## INFLUENCE OF TRANSFORMATION ON PERFORMANCES

### Model performances in a Binary Classification Setting

Figure 7 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.

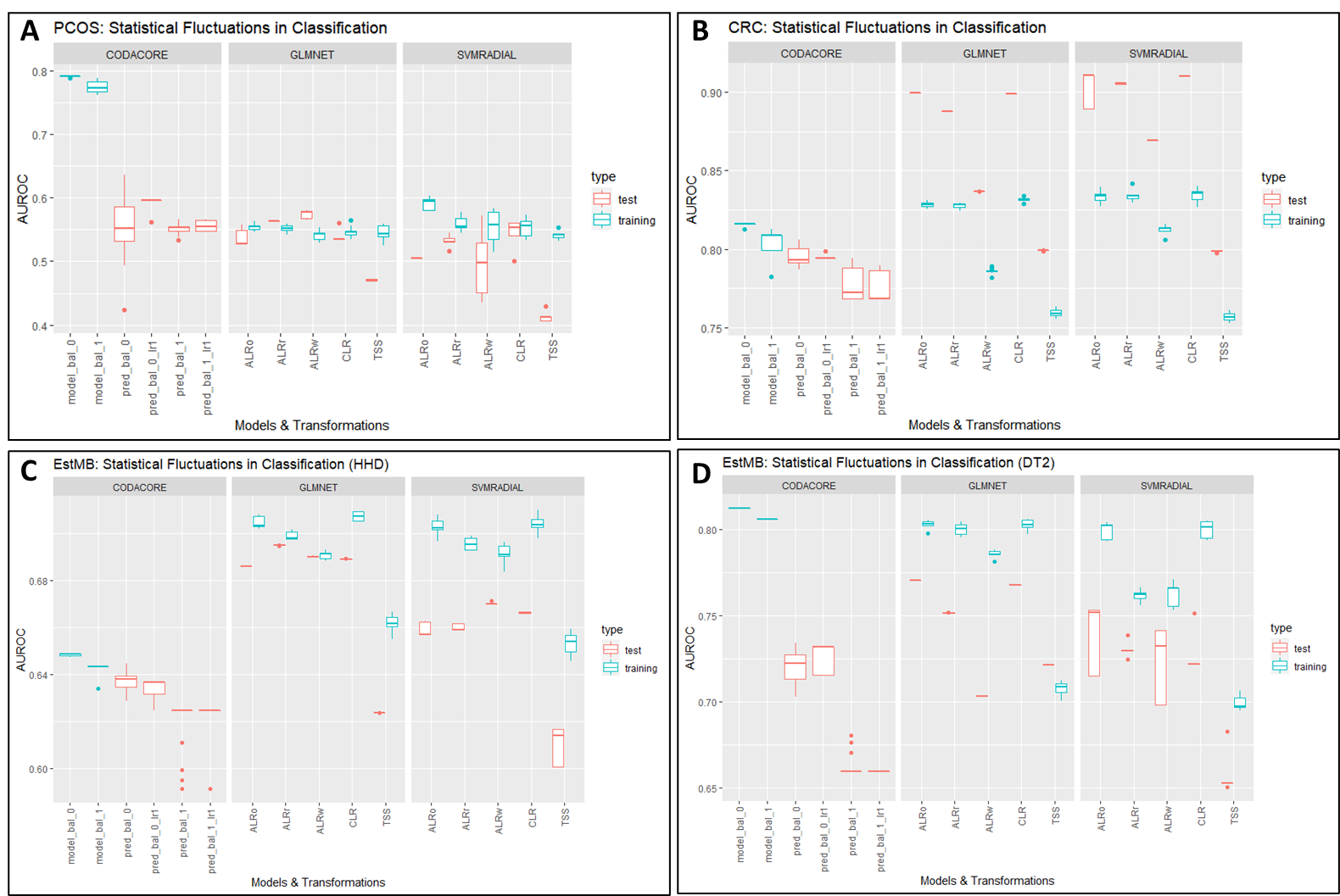


Figure : Statistical Fluctuations in Classification

CRC and PCOS use 10% abundance filtered sets, EstMB uses a 90% 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. 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).

As expected for PCOS (Fig.7A), 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. However, transformations do not improve the predictive performance in low correlation data sets.

Contrary to that, Fig. 7B 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 optimal ALR and CLR showing the best performances. ALR random shows surprisingly good results, which is repeated in Fig. 7C and D for the EstMB data set, but also in PCOS (Fig. 7A) as even in low correlation data, TSS shows the lowest performances. Furthermore, *CoDaCoRe* shows lower test and training performances, comparable to ALR worst in both glmnet. Of course, with an AUROC of 0.8, *CoDaCoRe* performances is still considered excellent. All machine learning models underfit the data, which indicates a bias in the data. This is probably due to the leakiness of the pipeline, considering that imputations and transformations are conducted on the whole data set (Whalen et al. 2022). Indeed, section () shows that splitting the data set before imputation and transformation solves the underfitting.

Interestingly, although EstMB data sets were conducted with the same leaky pipeline as CRC, the models continuously overfit the data, instead of underfitting. 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. However, linear models seem to perform better than non-linear in both 7C and 7D, considering that test and training performances are closer in performances compared to non-linear performances. 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.

### 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.

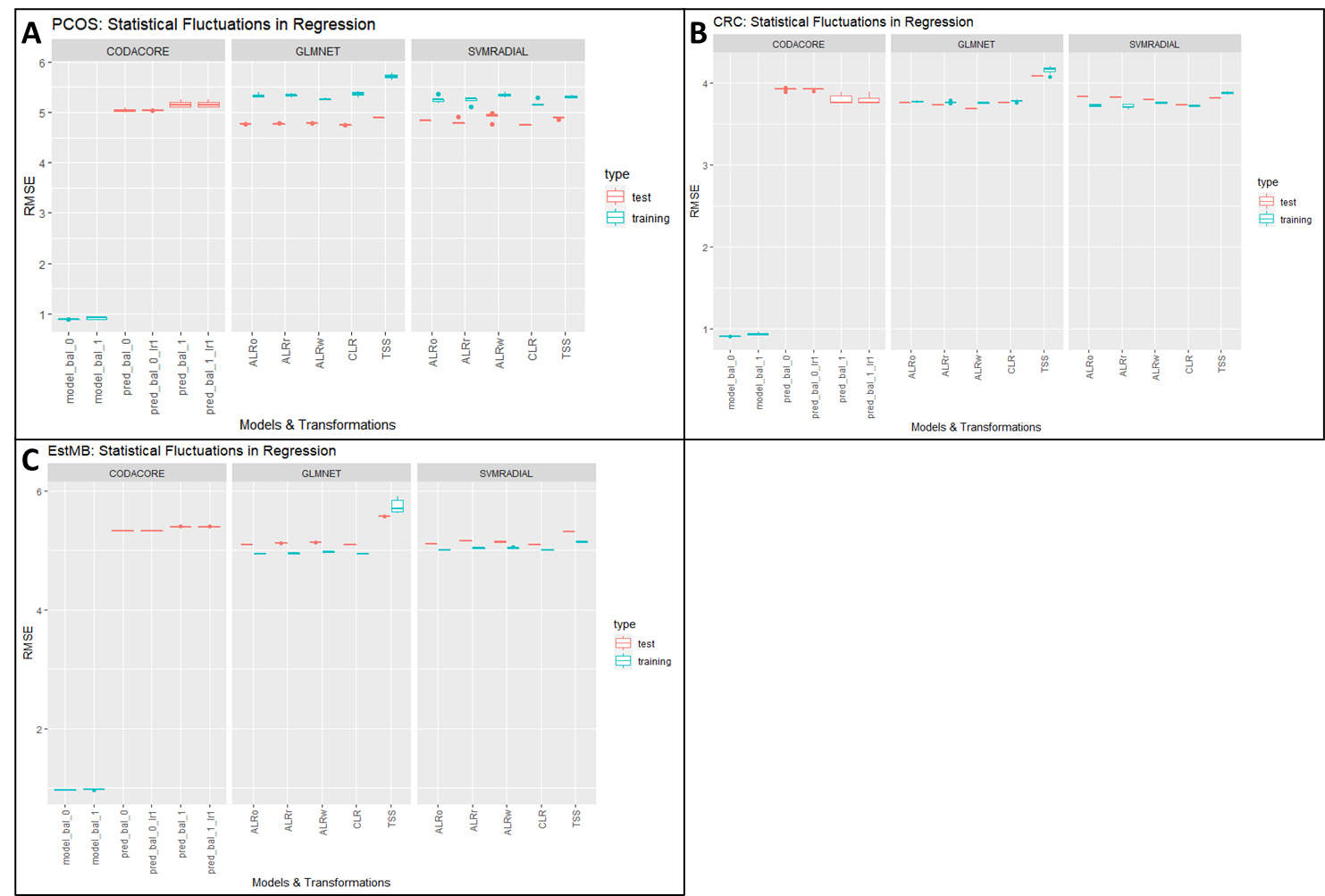


Figure : Statistical Fluctuations in Regression

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.

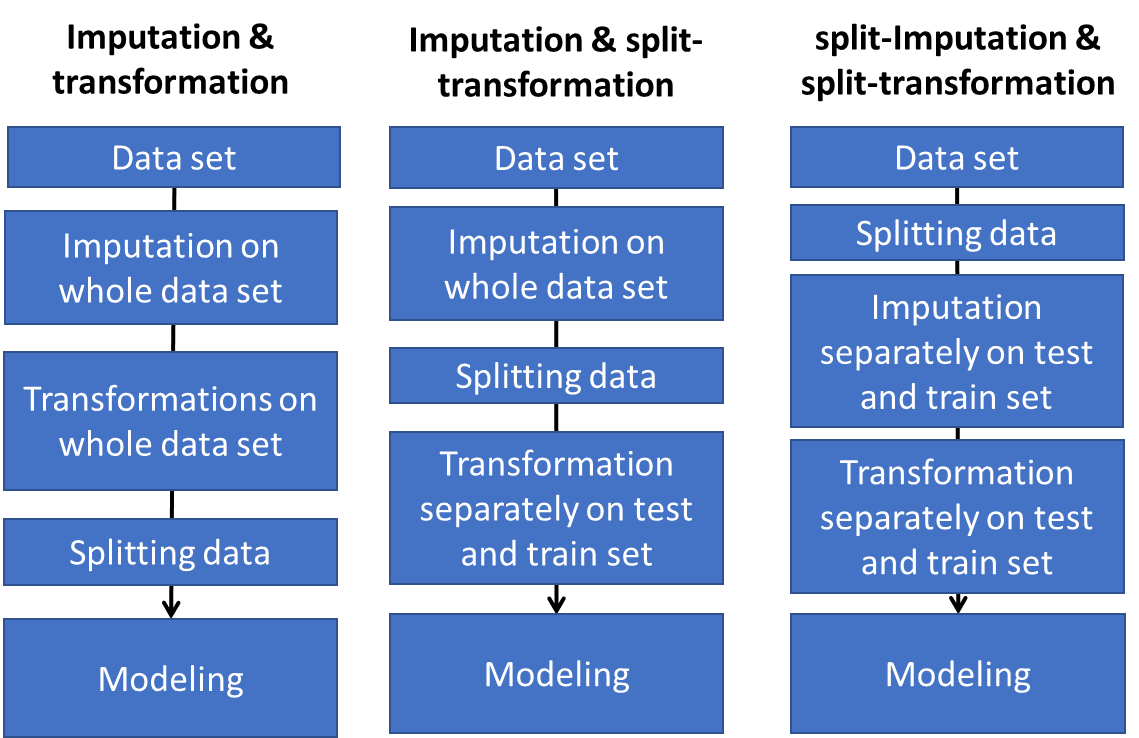
## INFLUENCE OF DATA LEAKAGE ON PERFORMANCES

As mentioned in the section above, data leakage is an easy to overlook problem and already lead to performance biases in the section before. To assess the impact of data leakage when using log-ratio transformations, this section directly compares the model performances of glmnet (and later codacore) under different “leakiness levels”.

Two factors 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. CLR transformations are conducted row-wise and therefore do not impact data leakage and for ALR, the analysis to find the best reference should be conducted on the training set and adapted to the test set.

Therefore, imputation also seems to be the most impacting factor for data leakage when using log-ratio transformations (in theory). 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.

Therefore, intermediate level of data leakage is to conduct imputation on the whole data set and split the data afterwards before transformation, next to the ideal pipeline and the worst pipeline. The model pipeline is visualized in Figure 9:



In the ideal procedure (“split-Imputation & split-transformation”), the processed data is split into train and test set and imputation, as well as all transformations are performed separately on both. 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 a glmnnet model.

In the worst procedure, imputation and transformations are conducted on the whole data set and afterwards the data is merged with the metadata and split into train and test set and fed into the glmnet model. To assess the influence of imputation on data leakiness, the data set has been split after imputation and before transformation.

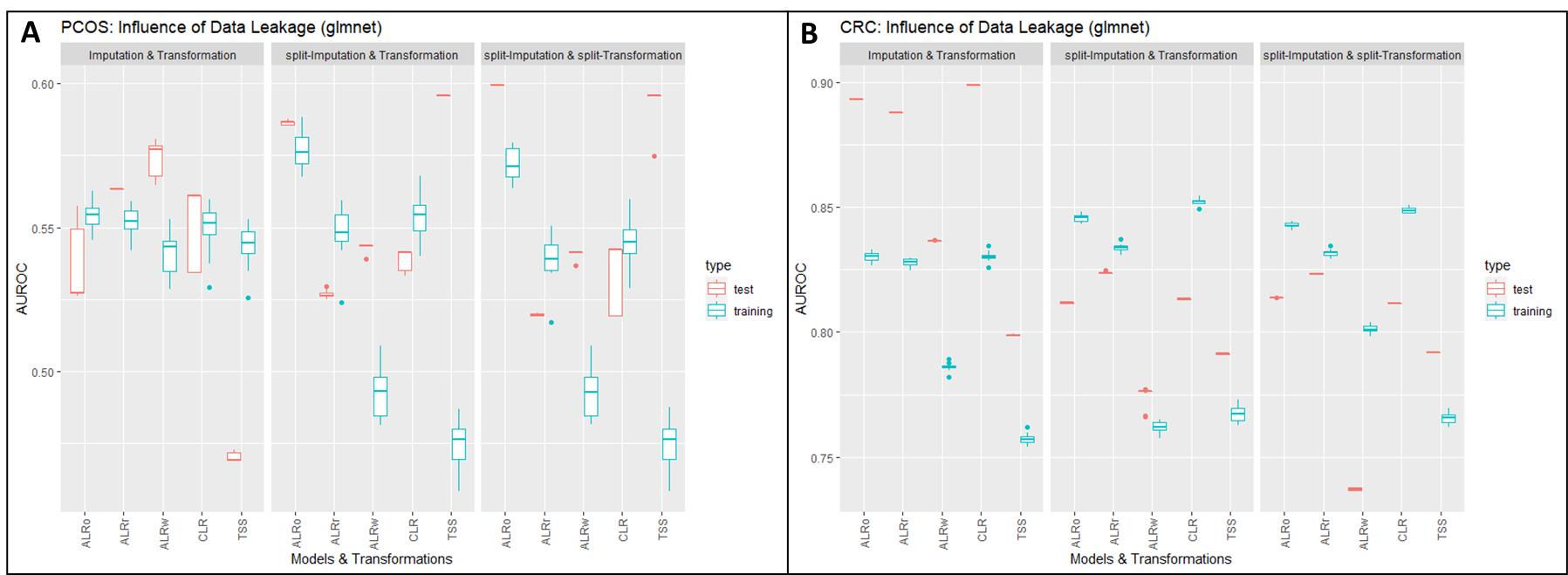


Figure : Influence of Imputation and Transformation on machine learning models

The machine learning model used was glmnet. On the left side, the results after imputation and transformation on the whole data set is shown, the middle part shows the effect of split-imputation on the whole data set and the right side shows the results if the data has been split before imputation and transformation. It should be noted that the split-imputation for PCOS produced an error, as the test set of this data set was too small to use GBM for imputation. Therefore, all zeros have been replaced by 0.5 in this particular data set.

Figure 9A shows the effect of data leakage on small data sets with low correlations. The results for “Imputation & Transformation” are the same as in section (), but stretched out a bit more due to a shorter y-axis. In general, all transformations produce similar results stable around 0.55, with TSS showing the lowest performance around 0.45. Interestingly, introducing split-Imputation and also split-Transformations still mirrors the results from the section above: TSS and ALRw seem to perform worse compared to ALRo, ALRr and CLR. However, the discrepancy in training and test performances between transformations vary greatly. CLR shows similar performances in all three test runs, which is expected as it is a row-wise transformation and therefore is not impacted by the procedures. However, the test performance of CLR shows higher variances when the data is split also before transformation. This could potentially be resolved by a bigger test set, as the PCOS test set is only 60 samples big. ALR and TSS are the most impacted by the changed pipeline. In both split procedures TSS shows a very high discrepancy between its test and training performance, heavily underfitting its data. Compared to that, before it was heavily overfitting its data. ALRw is continuously underfitting the data, which can be explained that the reference chosen explains the least variance. When splitting the data before imputation and/or transformation its performance drops visibly. ALRo marginally underfits the data but otherwise shows the highest performances out of all transformations. As hypothesized before, the biggest difference in machine learning performances seems to indeed stem from splitting the data before imputation. The results can differ when other transformations are chosen, but in the particular case of CLR and ALR, both are not affected by conducting transformations on the whole data set. This can be advantageous especially in smaller data sets, as potentially indicated by the high variance in CLR test performance. For bigger data sets, where the test set can be sufficiently large enough, data can be split without a problem, as is verified by Fig.9B.

The results in 9B show that the underfitting in the CRC data set was due to biases in the pipeline. As soon as split-Imputation was applied, the models are now overfitting the data. As before, splitting before Transformation does not seem to impact the performances heavily. In both pipelines, ALRo, ALRr and CLR show the highest performances, with ALRw and TSS significantly lower.

In summary, data leakage does not impact the general conclusion, that transformations lead to better performances in machine learning models. However, splitting the data before imputation should be included in the future, if the data set permits it. As a last test, it should be verified if imputation is also impacting the performance of *CoDaCoRe* and therefore indirectly pair-wise log-ratio transformation.

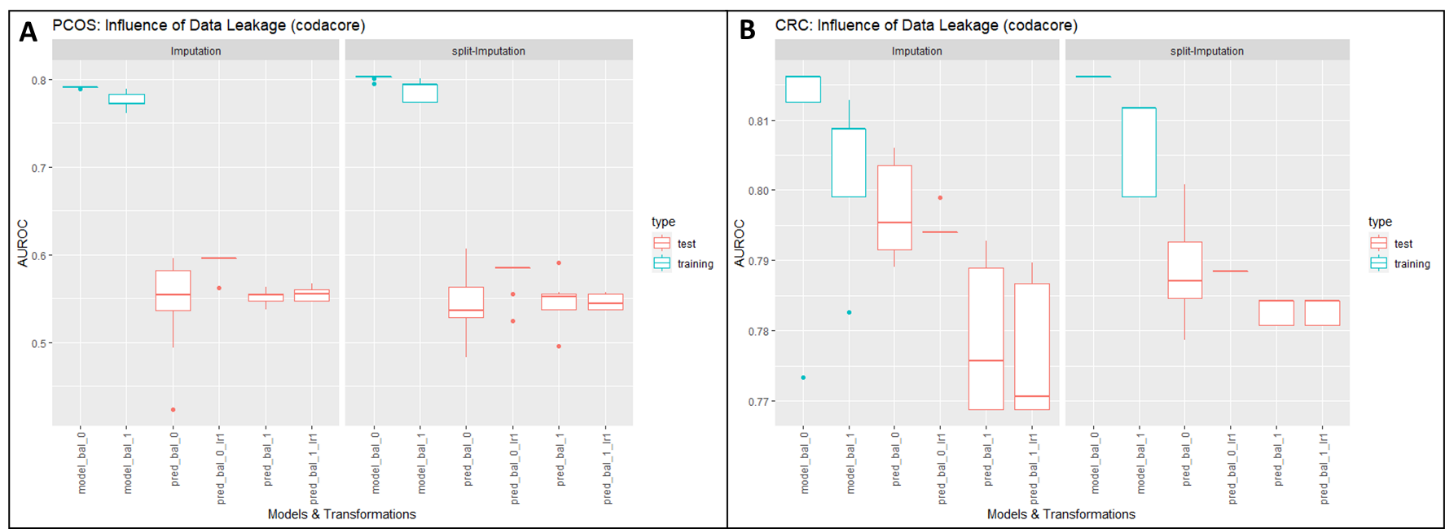


Figure : Influence of Imputation on CoDaCoRe

On the left side, the results after imputation on the whole data set is shown and the right side shows the results if the data has been split before imputation. It should be noted that the split-imputation for PCOS produced an error, as the test set of this data set was too small to use GBM for imputation. Therefore, all zeros have been replaced by 0.5 in this particular data set.

Figure 9 shows the same leakiness concept for *CoDaCoRe*. Here, the data was split before imputation (right) and imputation conducted on the whole data set (left), again for both CRC and PCOS. Similar to Figure 8, having an ideal non-leaky pipeline does not increase performances, as the low correlation in the data set is the most prevalent factor. For CRC, imputation on the whole data set seems to lead to higher variances in test performances compared to split imputation, but the test and training performances are very similar. In all cases, *CoDaCoRe* is overfitting the data.

## INFLUENCE OF TEST SET ON PERFORMANCES

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. Applying Holdout to the other data sets would be very difficult, a standard 80/20 split would be more advantageous, as it makes the comparison between all data sets easier. Therefore, one of the first tests was to assess if using Holdout or an 80/20 split makes a big difference for the CRC data set. Additionally, results from the section before were used. Imputation was conducted on the split data set, the transformation on the whole data set.

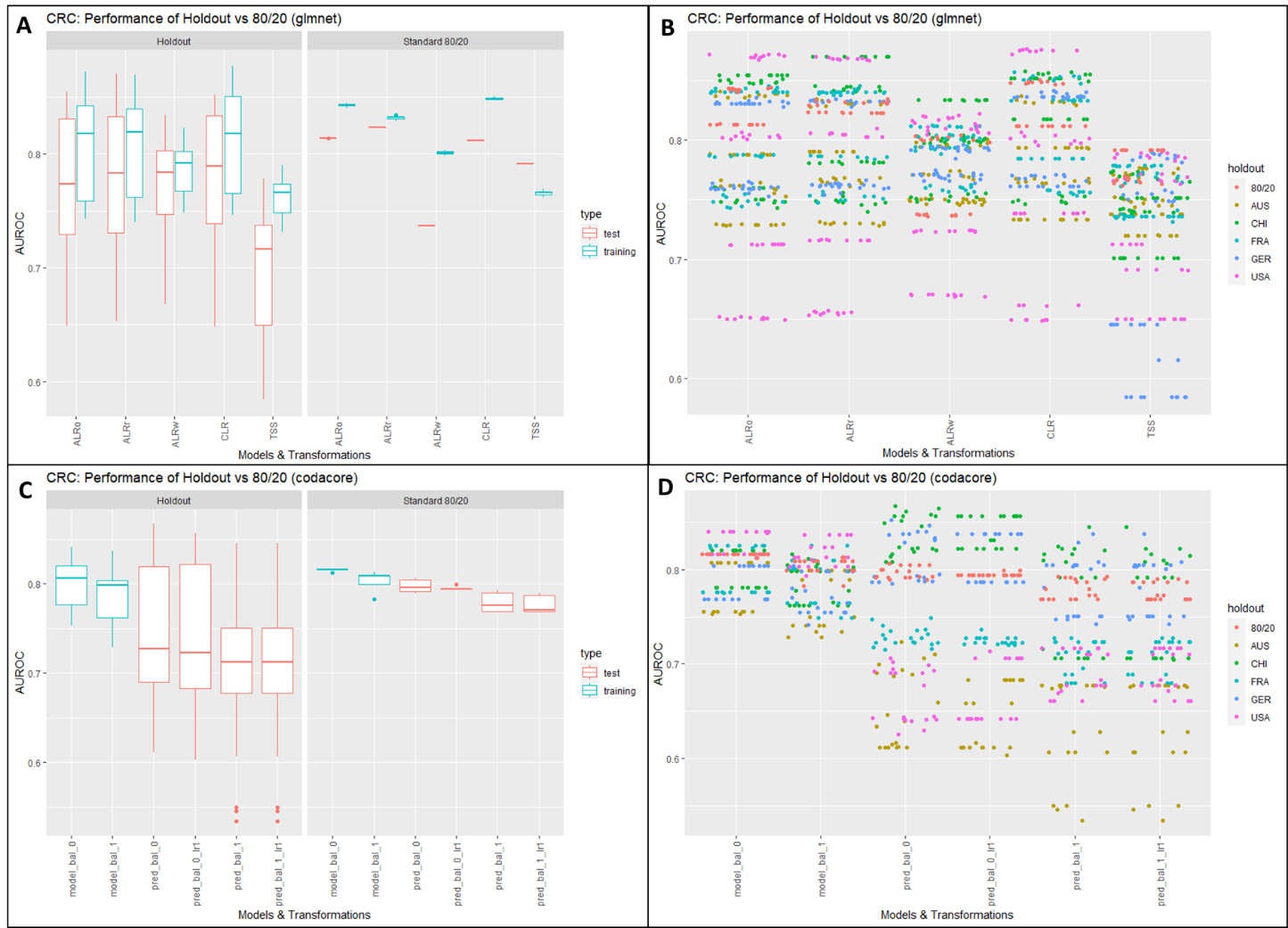
****

Figure 10: Comparison of Holdout vs. 80/20

The graph shows the performances of different test sets using the same glmnet model. A shows the performance of Holdout vs 80/20 with all test performances combined. In B all test performances are separated to show the differences in choosing a test set. C and D follow the same pattern, just for CoDaCoRe performances.

Figure 10A shows the difference of test performances as boxplots for different transformations of the same data set. All five holdout sets have been used separately as test set once. 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 also higher test performances.

In Figure 10B it becomes clear why the test performances of the Holdout set have such a high variance. For all transformations and holdout sets, the test performance is quite stable. For example, 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 big boxplots in 10A 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.*

Again, the holdout test performance has a high variance also in *CoDaCoRe* compared to 80/20. 7B plots again the details and the same results as before are visible. The high variance results in the choice of holdout set, with USA and AUS producing very low test performances and GER and CHI showing very high test performances. Standard 80/20 sits more in the middle.

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

For all tests above, a 10% abundance filtering has been used, as it is very common in microbiome and ecology (source). However, machine learning models work best if they have a big amount of data available (source). Therefore, the following test will assess how machine learning models perform if only 50% of data is available. Additionally, *CoDaCoRe* performances will be assessed.

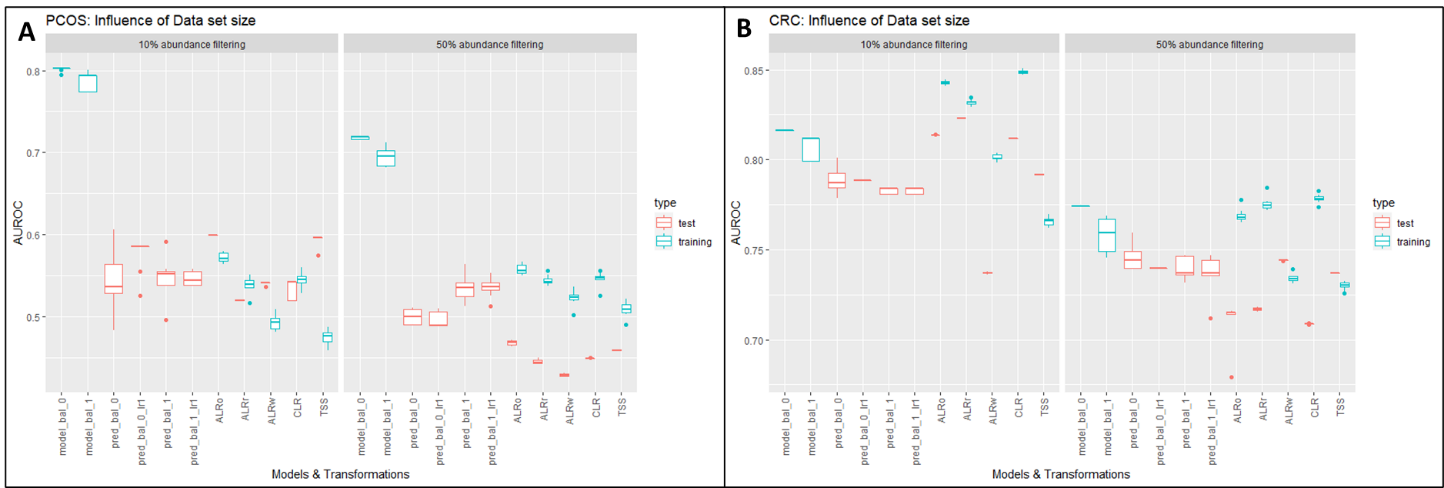


Figure : Influence of Data set size

For A and B x-axis contains *CoDaCoRe* models and several transformations, separated into 10% and 50% abundance filtering. A contains performances for PCOS data set and B for CRC data set. GLMNET model has been used.

Filtering the data sets has visible impacts in machine model performances. 10% abundance filter performances are the same as described in section() for both PCOS and CRC. In PCOS, using 50% abundance filter leads to a severe overfit of the data for both CoDaCoRe and glmnet, with TSS and ALRw again underperforming slightly compared to ALRo, ALRr and CLR.

CRC shows a similar picture, with 50% abundance filtering clearly performing worse than 10% abundance filter and an even bigger overfit.

# 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

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.

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.

Coenders, Germa; Greenacre, Michael (2021): Three approaches to supervised learning for compositional data with pairwise logratios. Available online at http://arxiv.org/pdf/2111.08953v1.

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.

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.

Rivera-Pinto, J.; Egozcue, J. J.; Pawlowsky-Glahn, V.; Paredes, R.; Noguera-Julian, M.; Calle, M. L. (2018): Balances: a New Perspective for Microbiome Analysis. In *mSystems* 3 (4). DOI: 10.1128/mSystems.00053-18.

Tolosana-Delgado, Raimon; Khodadadzadeh, Mahdi; Talebi, Hassan (Eds.) (2019): On machine learning algorithms and compositional data.

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.

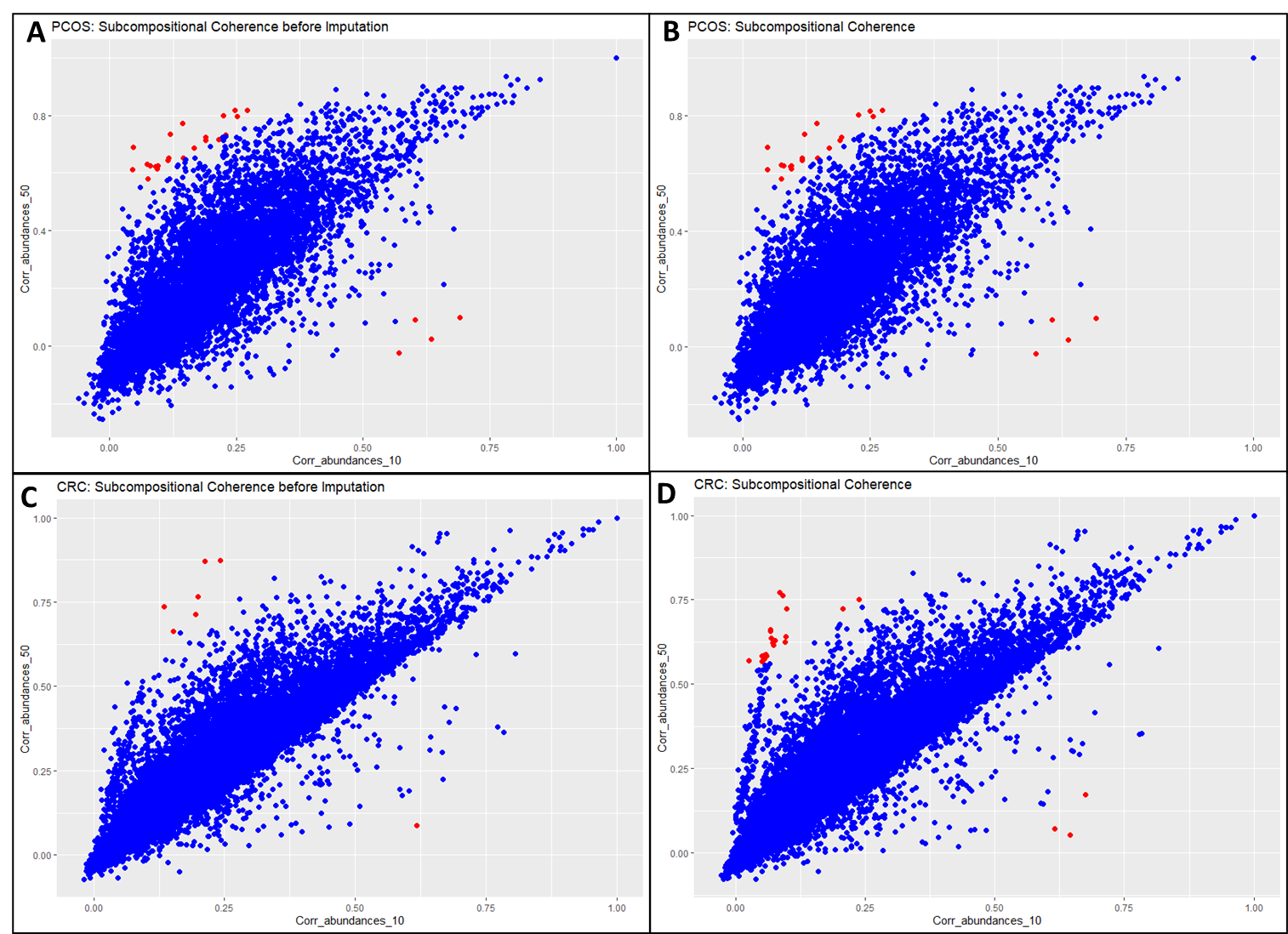
Whalen, Sean; Schreiber, Jacob; Noble, William S.; Pollard, Katherine S. (2022): Navigating the pitfalls of applying machine learning in genomics. In *Nature reviews. Genetics* 23 (3), pp. 169–181. DOI: 10.1038/s41576-021-00434-9.

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.

Zhang, Mo; Shi, Wenjiao (2019): Systematic comparison of five machine-learning methods in classification and interpolation of soil particle size fractions using different transformed data.

# Supplementary

# Correlation Coefficients



# Presence-Absence-Pattern

