University of Tübingen, Chair of Integrative Transcriptomics

Project Report for ML-4998

Exploring Linear and Nonlinear Embeddings for Microbiome Data

Patrick Köhler

February 18, 2022

Advisors: Prof. Kay Nieselt, Susanne Zabel

Contents

Abstract

High dimensional microbiome data (number of variables $\sim 10^4$) requires a low dimensional representation to improve statistical power in clinical studies and leverage Machine Learning methods for data analysis Bukin et al., 2019. Tataru and David (2020) demonstrate the superiority of GloVe (Pennington et al., 2014) an embedding method developed for Natural Language Processing over PCA for this task.

To this end they measure the quality of an embedding algorithm by predicting the Inflammatory Bowel Disease (IBD) from the embedded data.

We replicate their experiments to account for uncertainty in their evaluation. Subsequently, we investigate why GloVe retains more predictive information than PCA and compare the latter to UMAP McInnes et al., 2018, a non-linear manifold based approach.

We find that PCA is not suitable for the data at hand because variables with extremely large variances mask the predictive information in low-variance variables. GloVe's higher predictive capacity arises from its ability to focus on certain regions in input space. The coordinate axes cluster into four major directions. Each axis within one cluster distinguishes nuances in the cluster-specific region of the input space.

Finally we show that distances along the manifold carry even more information than GloVe's projection onto a linear subspace.

1 Introduction

Microbiomes are often represented in terms of occurrence counts (abundance) of all observed microbes. Depending on the resolution of phylogenetic relatedness among the microbes the size of the data matrix may explode rapidly. The quantification of one host organism (e.g. one human) then requires a vast amount of variables (Bukin et al., 2019). The number of variables in a typical data matrix may surpass the magnitude of 10^4 easily.

Many studies target the V4 region of ribosomal RNA, identifying one microbe with approximately 150 nucleotides. Defining the abundance of each unique nucleotide sequence as one variable gives a high dimensional representation of each microbiome.

In such high dimensional spaces, statistical analyses of any kind suffer from the Curse of Dimensionality (Hastie et al., 2009). This leads to a lack of power to detect small effects (Tataru and David, 2020) and any kind of general structures. Thus, in order to explore microbiome data statistically efficient it is necessary to reduce the number of variables in a meaningful way. To this end recent approaches have employed Machine Learning algorithms for dimensionality reduction.

Most methods, including Principal Component Analysis (PCA), Principal Correspondance Analysis and Global Vectors for Word Representation (GloVe) (Pennington et al., 2014) focus on projections onto linear subspaces. In contrast, non-linear manifold methods, e.g. UMAP (McInnes et al., 2018), have not been considered as an alternative.

In the present work we compare the qualities of three different embedding methods. Additionally we provide empirical evidence as well as theoretical arguments supporting the manifold hypothesis: Reducing the dimensionality by retaining the distances along the data manifold is possibly superior to projections on a linear subspace.

Tataru and David (2020) introduce GloVe, an embedding method from Natural Language Processing, for microbiome data. They design a disease prediction experiment to measure the embedding quality and compare it to PCA and non-embedded (raw) data. The results show that PCA performs worse than GloVe in this setup. Moreover they claim that GloVe retains as much information about the Inflammatory Bowel Disease (IBD) as the raw data.

We replicate their experiments to account for variance in the train/ test procedure. Our findings suggest that their conclusion is an artefact of randomness: GloVe does not retain as much

information about IBD as the raw data. Nevertheless, we confirm that PCA performs worse and investigate the reasons thereof. Extremely large variances in the abundance of some microbes biases the Principal Components (PCs) towards those. Discriminative information in important low variance microbes is lost after projecting onto the PCs.

An inspection of the embedding coordinate system shape suggests that GloVe focuses on a handful of non-orthogonal regions in input space to represent data points. Given the fact that GloVe outperforms the orthogonal Principal Coordinate System, we hypothesize that the relatedness of two microbiomes can be measured more accurately in terms of their distance along the data manifold. Our preliminary experiment with UMAP embeddings provide first evidence that this is indeed the case.

The following section provides thorough motivations and explanations for the GloVe and UMAP embedding methods, emphasizing the inherent difference between linear and non-linear embeddings. Section 3 introduces the data set and outlines the design of the prediction experiments. In Section 4 we compare the predictive performances. Afterwards we investigate the properties of those embeddings and finally present an independent experiment to evaluate the potential of the UMAP embedding. The Appendix includes a brief overview over the Code repository and mentions major implementation challenges.

2 Embedding Theory

This section outlines the dimensionality reduction algorithms Global Vector Word embeddings and Uniform Manifold Approximation and Projection. The underlying concept of the latter algorithm is contrasted to a set of dimensionality reduction methods known as linear embeddings, to which Principal Component Analysis (PCA) and GloVe belong.

2.1 Linear vs. Non-linear Embeddings

A class of dimensionality reduction methods, which will be referred to as linear embeddings, searches for the linear subspace of the input vector space that yields the best representation of the data in a fixed number of dimensions.

While different notions of *best representation* give rise to different algorithms and consequently different embedding structures, they follow the same mathematical paradigm. Let D_{original} be the data matrix and D_{embedded} its lower dimensional representation.

1. Define a subspace of input space that will be the new coordinate system in which the embedded points will be represented.
   Denote each axis by a vector in input space and write it as a column of the embedding matrix $E \in \mathbb{R}^{d \times p}$, where d is the number of original features and p the reduced number of dimensions.

2. Project the data points onto the new coordinate system and obtain the matrix of embedded points:

   $$D_{\text{original}} E =: D_{\text{embedded}} \in \mathbb{R}^{n \times p}.$$

   The matrix product notation denotes the projection of each point onto each dimension. To illustrate this, consider one data point $x \in \mathbb{R}^n$ in input space , i.e. a row of D_{original}. Projecting x (orthogonally) onto an axis $a \in \mathbb{R}^n$ mathematically corresponds to computing the dot product between the two vectors. The resulting scalar $s = x^\top a$ indicates how much x points towards the given axis.
   The matrix notation computes s for each new axis, for each data point and denotes them as new coordinates. A row in D_{embedded} contains the s for all new axes for one data point.

3. Define the notion of best representation in terms of a loss function comparing the original with the embedded data sets:

   $$J: \quad D_{\text{original}} \times D_{\text{embedded}} \to \mathbb{R}.$$

   J either encodes the belief that distances between columns of D_{embedded} should be similar to distances in D_{original} or expresses a global notion of information retention. The latter is the case for PCA, which builds on the premise that explaining variance in the data corresponds to retaining information.

4. In order to find the best embedding E^*, minimize J via E:

   $$E^* = \text{argmin}_E \quad J(D_{\text{original}}, D_{\text{original}} E), \quad E \in \mathbb{R}^{d \times p}.$$

Note that the axes of the new coordinate system, i.e. the columns in E^*, are linear combinations of the axes in input space. For all columns, the i-th entry corresponds to the weight the i-th variable in input space contributes to the direction of this new axis.
After projecting onto E^* non-linear structure in the data might be lost.
Furthermore the distribution of meta variables, e.g. disease status (for microbiome data), may change in an uncontrollable manner. Assume there exists a feature in input space , e.g. the abundance of a particular microbe, which allows for clear discrimination between sick and healthy

hosts. If this variable is outweighed in J by the others, because it does not contribute much to the heuristic definition of a good embedding its discriminative value is lost. Note that the specific effects strongly depend on the choice of the inductive bias induced by the choice of J as well as the data set itself.

Some non-linear methods circumvent this problem by embedding points implicitly. Since distances between points are the ultimate aspect of interest [1] there is no need to find an explicit coordinate system if distances between points in a lower dimensional representation can be characterized otherwise.
UMAP does so by measuring the distances *along* the cloud of points (more formally: the underlying manifold around which the data set concentrates). Points are considered similar if there exists a short path along the manifold connecting them. In contrast to that, linear embeddings usually aim to assign high similarity to points which are close in input space as measured by some metric distance function.

Figure 1 illustrates the manifold concept in $\mathbb{R}^3$. The right side visualizes a synthetic data set where each point corresponds to one observation, while the left side displays the manifold around which the data set concentrates.

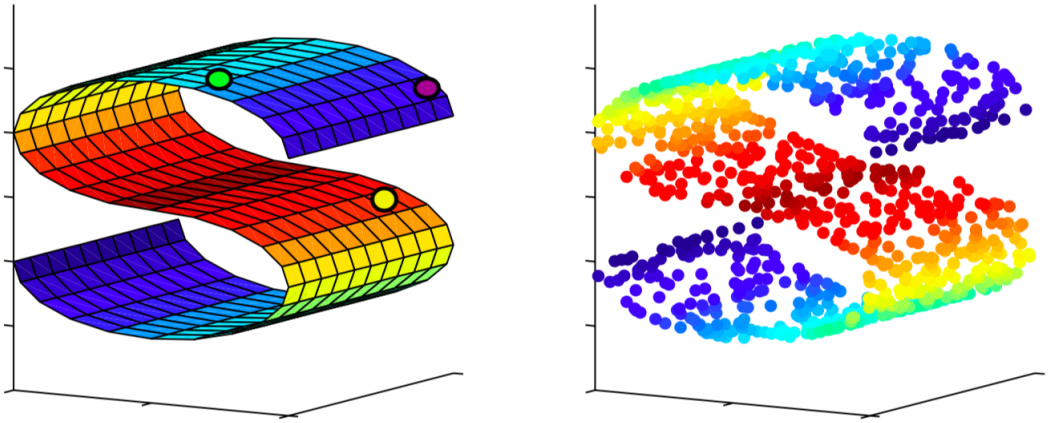

Figure 1: A data set in $\mathbb{R}^3$ (right) and its underlying 2D Manifold (left) (Luxburg, 2020). Colours only emphasize the 3D visualization. In the left image the green data point is closer to the purple one than to the yellow one, if distance is measured along the manifold. The converse relation holds if distances are measured with the euclidean metric.

Observe on the left that the purple point might be closer to the yellow one below than to the green one in a euclidean sense but further away from it measured by geodesic distance, i.e. the shortest path along the manifold connecting the two points. While non-euclidean distance functions can reflect the same property as the geodesic distance in this example, metric distance measures are invariant to shifts in general. If all three points were to be shifted in the same direction (add the same vector to all points), the relation *purple closer to yellow than to green* would not change for any fixed metric distance measure. However, if the points are shifted from a flat region on the manifold into a region where the manifold exhibits more spikes and valleys, the relationships may change. It might happen that two points which had a short path connecting them on the flat region are far apart when shifted into the spiky region because connecting them along the manifold requires to move through local minima and maxima.

This observation is the key difference between linear and non-linear embeddings. Whereas linear

[1] Many powerful ML analysis techniques (clustering, classification/ regression) only require distance between the data points.

embeddings measure similarities independent of other data points, manifold approaches take into account the density of data in the neighbourhood around the points of interest.

In the realm of microbiome data the geodesic notion of distance reflects the belief that ecological systems change smoothly in input space. For an existing community of microbes, the abundance counts of microbes may not be changed arbitrarily but only in such a way that the resulting community is stable enough to appear in nature. Increasing the count of one particular microbe (e.g. a pathogen) by an unreasonable amount may result in a data point which is impossible to observe in reality due to biological constraints. Assume that the set of points on the right side of Figure 1 represents the population of all realistic gut microbiomes exhaustively. Shifting the purple point on the left side downwards such that it lies in the void between the blue and red areas of the manifold results in a *theoretically possible* but *realistically not observable* data point.

2.2 Uniform Manifold Approximation and Projection (UMAP)

UMAP (McInnes et al., 2018) is one instance of a manifold learning algorithm which, in contrast to others (e.g. Tenenbaum et al., 2000), builds the specific parameter choices onto a solid mathematical framework rooted in Riemannian geometry and algebraic topology.
The high level paradigm of manifold learning can be described as follows: Choose a distance measure d in input space. Represent the data set as an undirected graph where each observation is a node, which is connected to its k nearest neighbours. The weights of the edges w_{ij} are proportional to dissimilarities between the observations i and j. Given this neighbourhood graph, compute pairwise shortest distances between all nodes and define these as the geodesic distances, i.e. the proximity along the manifold.

Manifold Approximation

Observe that the output is strongly dependent of the choice of d, k and the relation between d_{ij} and w_{ij}. k is not modelled by any of the manifold algorithms and remains to be optimized by the user.

For each observation x_i define

$$\rho_i = \min\{d(x_i, x_j) \quad | \quad 1 \leq j \leq k\}$$

and a normalization factor σ_i such that

$$\sum_{j=1}^{k} \exp\left(\frac{-\max(0, d(x_i, x_j)) - \rho_i}{\sigma_i}\right) = \log_2(k).$$

In particular the definition of ρ_i ensures that each observation is connected to at least one other point with weight 1. Both choices were derived from a category theory point of view capturing local structures (ρ_i) in terms of fuzzy simplical sets and individual smoothing according to the Riemannian metric local to each point.

With these parameters a directed graph $\bar{G}_i$ is constructed per observation, reflecting the local geometry of the manifold. Define x_{i_j} as the j-th nearest neighbour of x_i. The weights of $\bar{G}_i$ are:

$$w(x_i, x_{i_j}) = \exp\left(\frac{-\max(0, d(x_i, x_{i_j})) - \rho_i}{\sigma_i}\right).$$

To develop an intuition for this intermediate step one can think of these weights as the probability that the given edge exists. It resembles the confidence by which a close by point was observed because of structure and not due to random variation.

Combining these directed subgraphs into a single undirected one which faithfully captures the geometry in the data is not straight forward. However, the authors' abstraction of these graphs as fuzzy simplical sets provides a natural way to do so. Translated into graph operations this is

achieved by forming one large graph

$$\bar{G} = \bigcup_{i=1}^{n} \bar{G}_i.$$

The corresponding adjacency matrix A contains all locally computed weights. Finally the symmetric adjacency matrix B of the neighbourhood graph as used in the algorithm is obtained by:

$$B = A + A^\top - A^\top \circ A^\top,$$

with $\circ$ being the pointwise (Hadamard) product. B_{ij} can be conceived as the probability that there exists at least one edge between x_i and x_j in $\bar{G}_i$ or $\bar{G}_j$. As an output of UMAP geodesic distances are computed as shortest paths on G.

Projection onto Lower Dimensional Space

Beyond the geodesic distances, UMAP provides an explicit representation of points by embedding into any m-dimensional space. The embedding is performed by an iterative algorithm minimizing the cross entropy between the weighted graph G and its equivalent H in the lower dimensional space. Cross entropy is defined implicitly for these graphs, since the entries of their adjacency matrices can be conceived as probabilities. The non-convex problem is guaranteed to arrive at local minimum by a repulsive force algorithm.

2.3 Word Embeddings with GloVe

Manifolds might capture the topology of microbiome data most accurately but provide no direct insight into its biological meaning. The Global Vectors for Word Representation (GloVe) algorithm (Pennington et al., 2014) does so by searching for properties of microbes perceived as co-occurrence patterns. GloVe was designed to embed texts in natural language processing by co-occurrence patterns of words. Texts are often represented by word occurrences which matches the data type of Amplicon Sequence Variant (ASV) reads, namely discrete and finite count values. Moreover microbiome and language data share high level structure. In the same way as a text corpus can be used to model a document as a set of topics which itself is modelled as a set of words, microbiomes may be modelled as sets of microbial neighbourhoods consisting of sets of taxa (Sankaran and Holmes, 2019).
For illustrative reasons the GloVe algorithm is described in the context of language. Applying GloVe to microbiomes only requires to replace words by ASVs and documents by samples.

The following explanation and illustration is in accordance with Pennington et al. (2014). Let X_{ij} denote the number of times word i appears together with word j in a document, and let X be the corresponding matrix. $X_i = \sum_k X_{ik}$ be the number of times any word appears together with i and $P_{ij} = P(j|i) = X_{ij}/X_i$ the probability that j appears in the context of i.
The core of GloVe lies in the observation that the importance of a word in the context of a particular aspect of interest may be obvious from ratios of co-occurrences but not from co-occurrences themselves.

To illustrate this notion, consider the words $i =$ ice and $j =$ steam in the context of thermodynamics. Table 1 shows the co-occurrence probabilities of several words k with i and j obtained from a 6 billion token corpus.

Probability and Ratio	k = solid	k = gas	k = water	k = fashion
$P(k\|\text{ice})$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|\text{steam})$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|\text{ice})/P(k\|\text{steam})$	8.9	8.5×10^{-2}	1.36	0.96

Table 1: Co-occurrence probabilites of words in the context of thermodynamics from a six billion token corpus. Excerpt from Pennington et al., 2014.

Both words have the highest co-occurrence with k = water, even though k = gas is evidently closer to j = steam in the context of thermodynamics. The same holds for i = ice and k = water, suggesting that both i and j are closest associated with k = water.
For k that are associated with i but not so much with j the ratio P_{ik}/P_{jk} is large. The table reflects this notion by a high value for k = solid and a low value for k = gas. Values close to 1 indicate similar associations with the probe word.

Let $w_i \in \mathbb{R}^d$ be an embedding of word i. In order to preserve the context information require that some function of the embedding reproduces the co-occurrence ratios:

$$F(w_i, w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}, \tag{1}$$

where $\tilde{w}_k$ is a separate embedding of the context word, which facilitates the optimization procedure of the algorithm. The inherently linear structure of vector spaces suggests F to depend on differences of the embeddings and to be linear in all arguments. Context ratios being real numbers narrow the reasonable choices of F down to a form as:

$$F((w_i - w_j)^\top \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}, \tag{2}$$

i.e. the difference of embeddings affects the context ratio linearly in the context embedding. In terms of co-occurrences there is no distinction between w and $\tilde{w}$, as they reflect the same words only in different embedding spaces; thus requiring them to be interchangeable in (2). Since this is not the case, some algebraic manipulations are necessary to enforce this invariance under relabeling. Introducing bias terms b_j as offset parameters for a linear function the resulting condition amounts to:

$$w_i^\top \tilde{w}_k + b_i + \tilde{b}_k = \log(X_{ik}),$$

which is a strongly simplified, but explicitly parameterized version of (1). With this parameterization of the co-occurrence ratio preservation the loss function of a weighted log bilinear least squares regression is defined:

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^\top \tilde{w}_j + b_i + \tilde{b}_j - \log(X_{ik}))^2,$$

where V is the number of distinct words (vocabulary) and

$$f(x) = \begin{cases} (x/100)^{0.75} & \text{if} \quad x < 100 \\ 1 & \text{else,} \end{cases}$$

a weighting function that satisfies $f(0) = 0$, non-decreasing and $f(x)$ relatively small for large x so that frequent co-occurrences are not overweighted. Note that the product inside the sum of J is defined to be 0 for $X_{ij} = 0$, despite the log being not defined.
The problem is convex thus a global optimum can be found through gradient descent.

Denote the embeddings $w_i \in \mathbb{R}^d$ which minimize J in rows of a matrix W. The columns of W

(dimensions of embedding space) can be interpreted as properties of microbiomes. To calculate the score of sample s on property p, compute the dot product between the s-th row of the sample by ASV matrix, D, and the p-th column of W. Doing so for all samples and all properties amounts to the matrix product DW and results in a matrix that represents the samples in terms of its property scores in rows.

2.4 Statistical Concepts

In the following some statistical tools are discussed which are used to evaluate the conducted experiments. The explanations serve as a reminder of the underlying concepts, not as a detailed review.

Precision Recall vs. Receiver Operator Characteristic

The commonly used Receiver Operator Characteristic (ROC) is a loss criterion for binary classification that captures the tradeoff between specificity and sensitivity. It is a curve that plots *fp/N* against *tp/P*, where *fp, tp* denote the number of false and true predictions of the positive class and the capital letters denote the number of observations in the class. Each point corresponds to a different threshold for predicting the positive class and the area under this curve is computed to assess the quality of the entire model.
For increasing imbalance of data (increasing N) the false positive rate decreases, increasing the AUC. In this case it might be informative to consult the precision recall curve as well: Instead of *fp/N* it considers the precision *tp/(tp+fp)*, which is independent of N (Hastie et al., 2009).

F_β Statistic

Another performance measure which trades off precision and recall (*tp/P*) is called F_β. It's a weighted harmonic mean between the both where β denotes the factor by which recall is considered more important as precision.

$$F_\beta = \frac{(1+\beta^2)tp}{(1+\beta^2)tp + \beta^2 fn + tp} = (1+\beta^2)\frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}.$$

F_β is commonly used to evaluate fixed threshold classifiers for different weightings of misclassification errors (Pope and Webster, 1972).

Importance Measures for Variables in Input Space

Please refer to this subsection when interpreting Figures 3 and 4. For the raw data, we consider an ASV important, if its variable importance measure (cf. (Breiman, 2001))ranks among the top 100 in more than 60% of the 27 replications.

Identifying important ASVs for PCA embedded data is more intricate. Now, the Random Forest's inputs are not ASVs themselves but scores on 100 PCs. We first identify the 10 most important PCs per replication in terms of variable importance measure. Given one of the 27 replications, we select the 10 ASVs per PC which contribute most to the axis direction. The magnitude of contribution for any ASV to a PC is given by the absolute value this ASVs coefficient on the PC-axis. Again, we filter for the 60% most frequent ASVs.

Such a selection is not possible for GloVe, as the columns of the embedding matrix have no deterministic order. Two fits for different subsets of the data may find the same axes but denote them in a different order.

3 Related Work

Since out conducted experiments build heavily upon Tataru and David (2020), we give a summary of their work, in this Section. We used the same data set for all our experiments.

3.1 Data Set

The publicly accessible American Gut Project (AGP) (McDonald et al., 2018) data set was used. It is gathered in a crowd driven manner, where study-participants are recruited on the internet. They pay $100 to receive the tools to extract a microbial probe from a self-chosen body-site (oral, fecal, skin). In return, they get access to an interface on which they can compare their own microbiome to those of the other study participants.

Sequencing ribosomal RNA (rRNA) is used to measure the abundance of the different microbes. This is not as accurate as DNA-DNA hybridization but provides a reasonable trade-off between taxonomic resolution and data acquisition cost. Since sequencing all nine regions is too costly for microbiome analyses the semi conserved V4 region is chosen to be sequenced.

The DADA2 method was applied to correct for sequencing errors and obtain reads with single nucleotide resolution referred to as Amplicon Sequence Variants (ASVs) (Callahan et al., 2017). The resulting ASVs are strings of an approximate length of 150 nucleotides. After preprocessing, the data matrix contains one variable per unique ASV, which counts its occurrence frequency.
Additionally vast amounts of life-style meta variables are available that study participants fill out in the form of an extensive survey. Aside from standard demographic information (age, gender, residency) these contain information on dietary habits, allergies, various diseases and physical exercise.

In total, the data set contains around 19,000 samples and 335,000 different ASVs. Samples with less than 5,000 reads as well as ASVs with relative frequencies lower than 0.7% were discarded, resulting in an 18,480 by 26,726 sample by ASV matrix.
Classifiers were only trained on self reported healthy (5018) or professionally with IBD diagnosed (856) subjects. Unlabelled data points were only used to compute the embeddings, not the classifiers.

3.2 Experimental Setup

In order to compare the quality of different embeddings, the authors propose to predict whether a patient has IBD or not only from embedded data. Predicting IBD status from the raw data serves as a "no loss of information" baseline.

Interpreting predictive performance as the quality of an embedding seems fairly arbitrary at first glance. However, it reflects an important aspect of what one would expect from a good embedding: It allows for the reconstruction of small effects with strong semantic importance.

In this context *small effect* refers to the fact that IBD is a rare disease.

Hence, retaining the information whether a microbial community is associated with IBD renders as a useful notion of embedding quality. Further details on IBD are provided in the Appendix.

Note that the experiments were not carried out to create an optimal diagnostic tool. Predictive power is rather to be viewed as a proxy of how much structure is captured by the embedding. In other words: Low performance is interpreted as loss of relevant information caused by the dimensionality reduction.

Three Random Forest models were trained with different inputs: Projected onto GloVe embedding space, projected onto the first 100 principal components and normalized by the inverse hyperbolic sin. Hyperparameters were optimized for all models individually via cross validation.

3.3 Results

Even though the GloVe Embedding outperforms the normalized raw data by 2% in Area under the ROC (AUROC) the inverse is true for Area under Precision Recall Curve (AUPRC). The model

with principal component data performs worst with regard to both metrics: The difference of Area to the raw-data model amounts to 2% for ROC and for PR. Table 2 shows exact numbers.

Thus GloVe embeddings are neither clearly superior nor inferior to the normalized raw data in terms of predictive performance. However, yielding comparable performances to the raw data suggests that there was no substantial loss of information, with regard to IBD status. Even though the number of features only amounts to 1/200 of the number of raw data features. GloVe outperforming PCA in both metrics (+4% in AUROC, +2% in AUPRC) shows that embeddings preserving co-occurence structures convey more information, with regard to IBD status, than linear combinations of ASV counts with maximal variance.

4 Experiments and Analysis

The prediction experiments by Tataru and David (2020) were solely based on one embedding that was optimized for a randomly drawn subset of the original data. Since it is not clear how strongly the embeddings change for different training data, their findings might exhibit artefacts of randomness. Thus we evaluate their claim that GloVe is the superior embedding on a statistically by accounting for the variability.

4.1 Uncertainty Assessment through Replication

In order to assess the reproducibility and robustness of the results, we replicated the experiment 27 times and evaluated the statistical uncertainty of the performances. In contrast to the original experiment, we did not provide any meta variables to the Random Forest (RF). Thus we make sure that our results can be traced back to the microbial community and not to confounding meta variables such as dietary habits.

Table 2 compares the median AUC to the point estimates in Tataru and David (2020). The median AUCs, with dietary variables included, are generally lower than the reported estimates. Without any meta variables, the performance decreases 5-7%.

	Dietary variables included		Dietary variables excluded
	Median over 27 reps	Estimate from Tataru and David (2020)	Median over 27 reps
Raw	0.74	0.79	0.74
GloVe	0.74	0.81	0.72
PCA	0.71	0.77	0.69

Table 2: AUCs from all prediction experiments compared: Single replication from Tataru and David (2020) compared with median from our 27 replications. Very right column only uses microbiome data whereas the others include dietary information.

Figure 2 shows the distribution of various performance measures for different embeddings. Each column corresponds to a different input to the Random Forest: GloVe embedded, PCA embedded and normalized raw data. In addition to the AUC, which is the only performance metric considered by Tataru and David (2020), we compute the precision and the F_β statistic, as described in Section 2 to account for the class imbalance (roughly five times more IBD negative than positive).
The raw data outperforms GloVe in terms of AUC and not vice versa as in the original experiment. The median is higher by 2% and the Inter Quartile Range is substantially smaller. Raw data Random Forests also achieve higher median precision by 2% with comparable box-size, i.e. variability across the replications.

The results agree with the original experiments in so far that PCA embedded data generally performs worst. The AUCs and precisions from PCA based predictions exhibit similar variability as for GloVe, although the F-measures vary much stronger.

In summary, the replications support the original hypothesis that GloVe embeddings can retain most of the information on the IBD status, while PCA loses much more. However, the larger variability in GloVe AUC clearly reflects a decrease of robustness if compared to the raw data.

4.1.1 Bias in Original Experiment

Figure 9 in the Appendix shows the equivalent of Figure 2 including the 13 dietary variables used in the original paper. Comparing both, we observe that the rank order of performance does not change by dropping the meta variables. However, we observe lower performances in Figure 2 without the dietary information. Furthermore, GloVe performs better in comparison to the raw data baseline with the meta variables included.

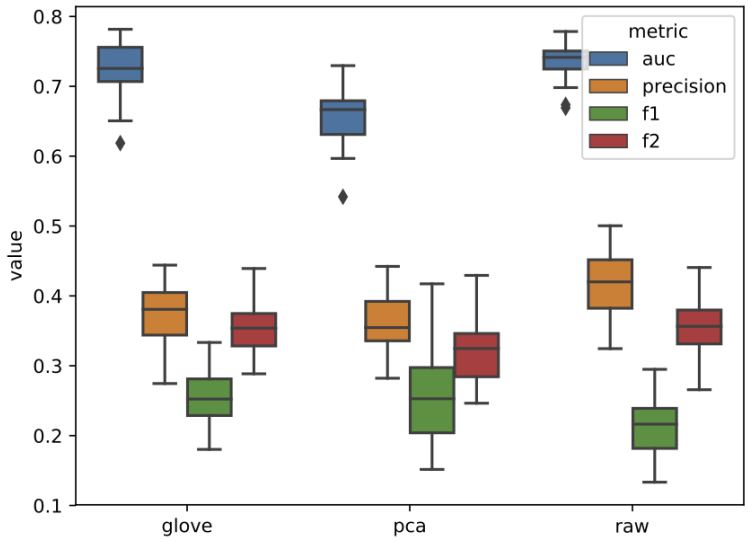

Figure 2: Performance Measures for 27 replications of the IBD-Prediction Experiment. Horizontal axis indicates the embedding algorithm which transformed the raw data to the input of the Random Forest. $\beta \in \{1, 2\}$ are commonly used choices for F_β.

From Figure 9, we infer that the results from (Tataru and David, 2020) are artefacts of randomness. The median AUC of GloVe is actually slightly lower than the one for the raw data. With substantially larger variability in GloVe AUC, we conclude that they happened to sample training data which is beneficial for GloVe.

From those results we can not conclude that GloVe contains a similar amount of information as the raw data, because the performance estimates exhibit biases from the dietary variables.

4.2 Embedding Properties

In order to reason about why PCA embedded data loses more information than GloVe, we analyze the predictive properties of the embedding spaces.

We show that the Principal Components are strongly biased towards a small set of microbes with extremely large variances. These do not carry as much predictive power as the non-orthogonal linear combinations derived from co-occurrence patterns in GloVe.

4.2.1 Principal Component and Raw Data Predictors Leverage Different Microbes

First, we inspect why PCA embedded data loses predictive information when compared to the raw data. To this end, we identify the most predictive ASVs for both data types separately and compare whether the predictors leverage the same microbes for classification.

Important ASVs are selected as defined in Section 2.4.

Given those most important ASVs for raw and PCA embedded data, we plot the distribution of their abundances Figures 3 and 4. Each pair of boxplots represents one microbe and shows the distribution per class.

Both Figures show that some ASVs carry discriminative information even when considered independent of the others. The first three microbes in Figure 3 tend to appear more frequently in sick hosts. We also observe the other direction of influence: The first microbe which is marked with an asterisk is found less frequently in sick hosts.

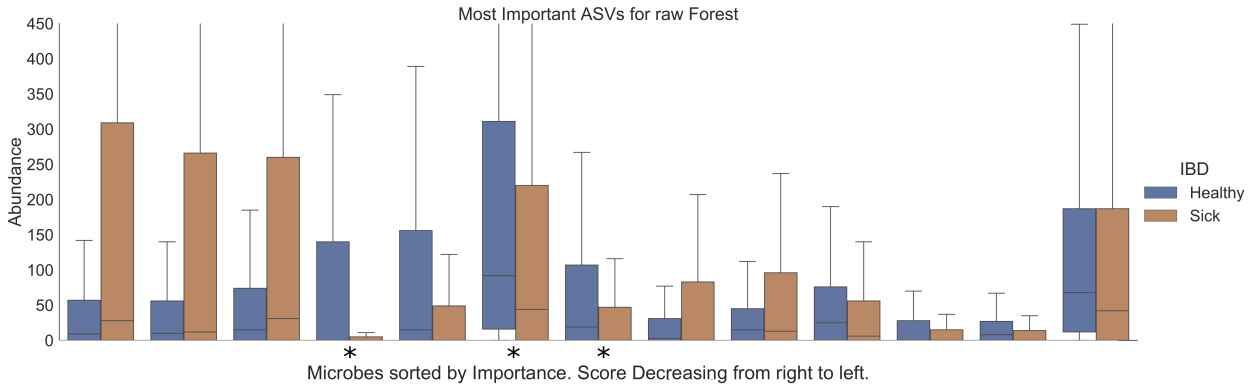

Figure 3: Per class distribution of abundances for most important ASVs in raw data. Asterisk indicates that ASV has also been considered important for PCA.

While we observe similar patterns in Figure 4, the relative differences in abundance are generally not as large as for the first three ASVs in Figure 3. These three in particular have not been found to be important, i.e. contribute much to the classification, when PCA-embedded data was used.

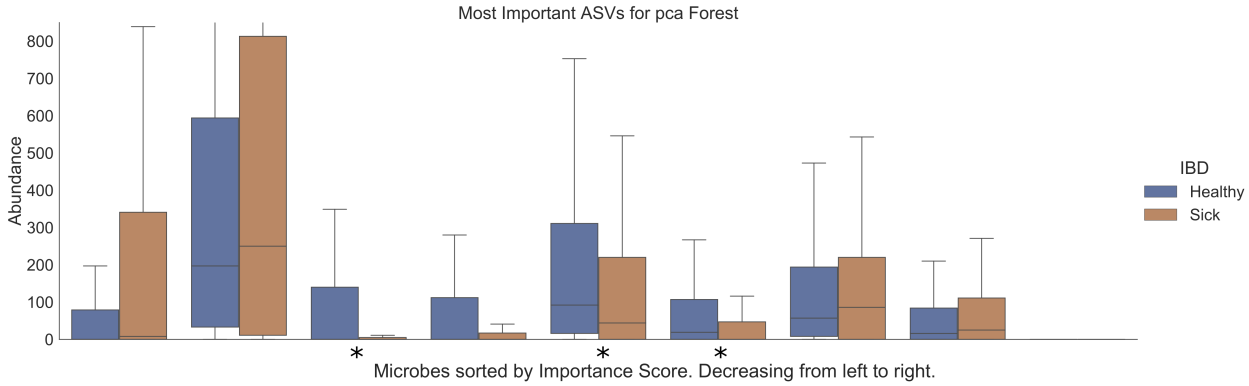

Figure 4: Per class distribution of abundances for most important ASVs in PCA embedded data. Asterisk indicates that ASV has also been considered important for the raw data.

PCA-embedded data can not recover the information whether there is a high or low abundance of a certain microbe, if the respective loading on the PCs is close to 0. We suspect that some ASVs simply can not be taken into account properly because their variances are substantially smaller than some of the others which results in small loadings on the principal axes. This claim is investigated in the following subsection.

4.2.2 Relationship Between Variance and Predictive Information

We compare variances of the most important ASVs between PCA and raw data. Figure 5 visualizes the distribution of their variances. The variances for PCA-important ASVs are substantially higher than those for the raw data. Their median value of approximately 200×10^3 is twenty times larger. Additionally the ASVs which are leveraged for PCA are distributed with a much wider range, while most of the distribution mass concentrates around the median of 10×10^3 for the raw ASVs. Notice, that the minimum variance for PCA important microbes is larger than 100×10^3.

However, Figure 6 shows that around 99% of *all* microbes exhibit variances smaller than 100×10^3. Most ASVs have much smaller variances as the aforementioned orders of magnitude. Regardless of their importance for classification, around 90% of all ASVs have variances even smaller than 100. Thus, data in 100 dimensional PC-space only considers the microbes with the highest variances for classification.

We conclude that most ASVs can not be used for the classification with PCA embedded data,

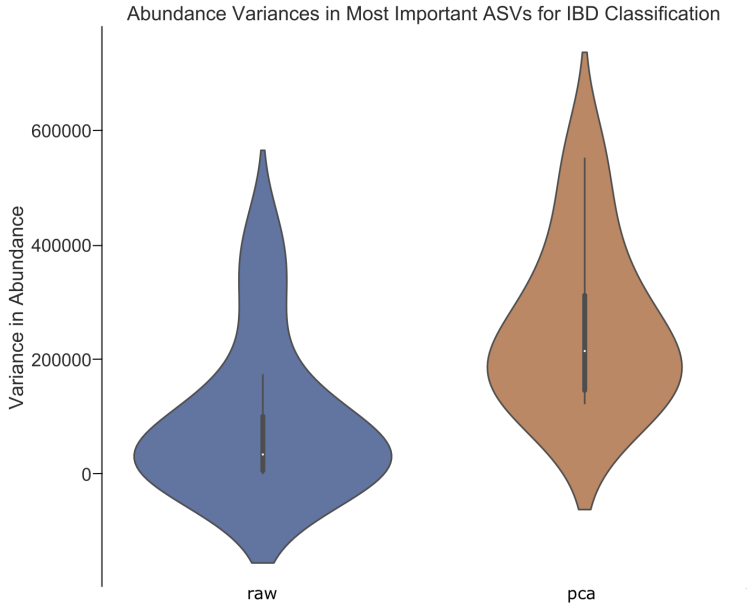

Figure 5: Distribution of variances for most important ASVs per embedding type.

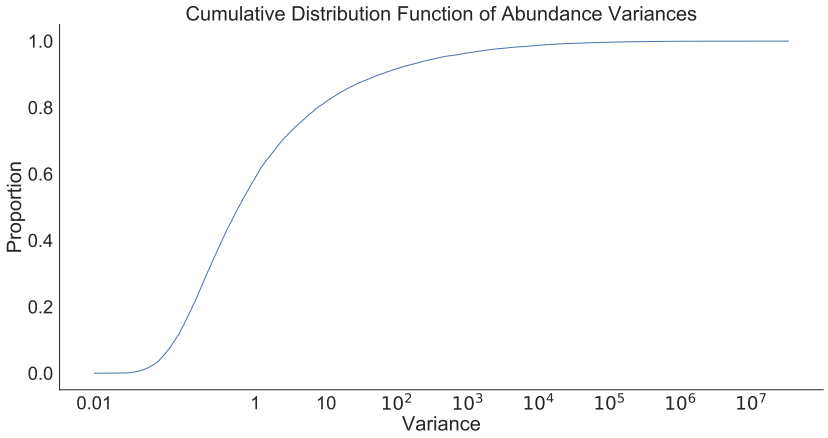

Figure 6: Cumulative Distribution Function of Abundance Variances with logarithmic scale.

since their information does not contribute much to the overall variance.

4.2.3 Data Concentrates in Non-Orthogonal Regions

Now, we investigate why GloVe embedded data retains more predictive power than PCA. To this end, we compare the coordinate systems onto which the raw data is projected. We visualize those with heatmaps that color code the weight an ASV has on each embedding dimension in. Each column represents one of the approximately 27×10^3 microbes from the raw data. The 100 rows correspond to the axes in the coordinate system that results after applying the embedding algorithm. Thus, the pixel-color in row i and column j encodes the coefficient of microbe i in embedding dimension j. In other words: If this pixel is dark purple, a subject's i-th value in the embedded representation is strongly influenced by the abundance count of the j-th ASV in the raw data. Rows were clustered with hierarchical clustering to improve visual interpretation.
In Figure 7 we observe clusters of dimensions which point towards similar directions in input space. All rows above the green segment show more or less homogeneous colors, indicating that the corresponding axes have similar weights for all ASVs. Dimensions inside the green rectangle attach great weight to the first few ASVs but low weights to most others. The third group of

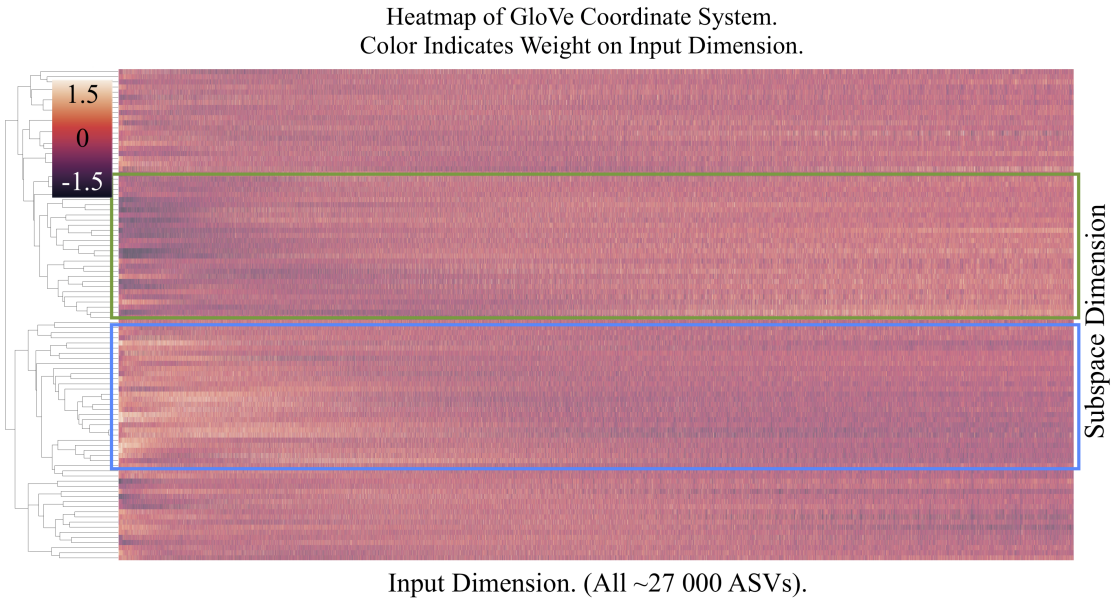

Figure 7: Heatmap of the GloVe Embedding Coordinate System in Input Space.

embedding dimensions (which point to similar directions in input space) lies inside the blue box. They implement a pattern inverse to the one in the second group: High weights on the microbes which receive low weights from the second group and vice versa.

In Figure 8 we mostly observe white, black or pink values. Therefore, Principal Components attach either high (in an absolute sense) or no weight to every ASV. Most of the high values appear on the very left, i.e. for a small fraction of all microbes. As all other columns to the right exhibit very similar color patterns, the few ASVs on the left determine the Principal Component Coordinate System. Those are the 10-15% ASVs that exhibit variances in much larger orders of magnitude than the vast majority (cf. Figure 6).

Notice that the rows, i.e. the Principal Components, can not be clustered in a meaningful way, since they are orthogonal to each other.

Recall that the dimensionality reduction happens by projecting onto the vectors in the high dimensional input space which are illustrated here as rows. Identifying those clusters in Figure 7 tells us that the embedding focuses on a few distinct regions in input space to generate the low dimensional representation. Given that the embedding retains a substantial amount of information, it seems that raw data concentrates around those few regions. Nuances within those regions are further distinguished by the small differences between the axes within a cluster.

PCA fails to identify those regions by design. Maximizing the variability of the PCs does not guarantee that we focus on regions in the input space where the majority of the data concentrates. In the case of this data set, where a few variables carry most of the variance, projecting onto Principal Components discards the structural information that comes with the shape and density of most of the data point cloud.

4.3 Distances Along the Manifold are Meaningful

The previous findings indicate that the shape and topology of the point cloud play a crucial role for lower-dimensional representations. Hence we suspect that an embedding which compares subjects by measuring their distance along the data manifold in the high dimensional space leads to a more holistic dimensionality reduction.

We compare GloVe to UMAP embeddings in an independent experiment. The prediction task

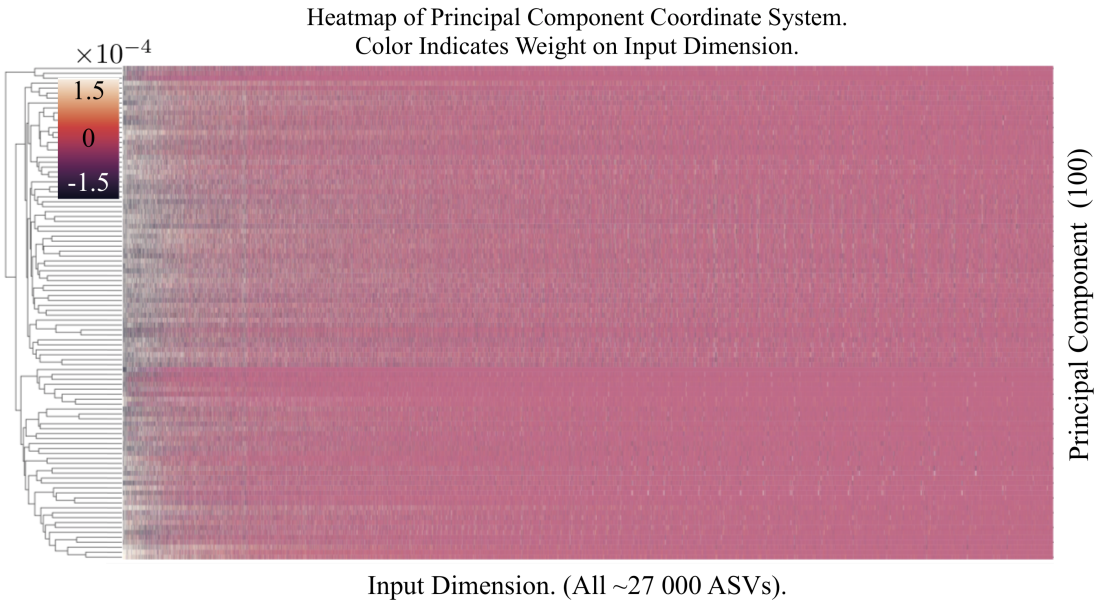

Figure 8: Heatmap of the Principal Component Coordinate System in Input Space.

remains the same as before, however we only use a subset of the previous data. After discovering that the initial experiments have trained joint embeddings for all body sites, we only consider fecal probes now. Although we can not compare the performances to earlier runs, we choose this setup to simulate a more application oriented scenario. The data set comprises of approximately 9000 observations with 500 (5.5%) positive labels.

After splitting it into five folds, we optimize hyper parameters for Random Forests and k-NN for GloVe embedded data. For UMAP embedded data we only train k-NN to assure that the predictive performance arises only from the distances along the manifold and not from a function of those. Since Random Forests achieve the best results for GloVe, we compare them to UMAP k-NN in Table 3. We evaluate the performance with F_1 and precision along with their standard deviations to account for the – now even stronger – class imbalance.

	GloVe RF	UMAP k-NN
F_1	0.2 ± 0.11	0.34 ± 0.14
Precision	0.26 ± 0.12	0.28 ± 0.1

Table 3: Comparison between UMAP and k-NN of Prediction Performances for Fecal Data Only. Average and Standard Deviation from 5 fold Cross Validation. For GloVe we select Random Forests as they perform better than with k-NN. UMAP is only trained on k-NN.

Compared to earlier experiments, all performance scores are substantially lower indicating that classifying from this subset of data is harder. k-NN with UMAP data outperforms GloVe in both metrics. Note that F_1 (the harmonic mean of precision and recall) does not favour a more sensitive or a more specific predictor. The substantially higher F_1 for UMAP data indicates that more patients *with* IBD were classified as such.

UMAP k-NN performing better than GloVe RF despite the fact that Random Forests generally allow for more sophisticated classifiers emphasizes the usefulness of a manifold representation. Note that GloVe k-NN performs even worse. This indicates that distances in GloVe space carry less semantic meaning than those along the manifold.

5 Conclusion and Discussion

In this project we analyzed the very high dimensional 16S-V4 rRNA microbiome data. It requires a lower dimensional representation to improve statistical power for clinical studies and overcome the Curse of Dimensionality for advanced data analysis methods. Finding an appropriate embedding method is non trivial, since it is not entirely clear which type of statistical information shall be retained and which can be discarded.

5.1 Conclusion

Tataru and David (2020) measure embedding quality by the performance a disease-classifier can achieve with embedded data. Even though predicting whether a person was diagnosed with Inflammatory Bowel Disease (IBD) is far from an exhaustive characterization of embedding quality, it serves as a criterion to comapre different methods.

Their comparison suggests that GloVe, an algorithm developed in Natural Language Processing, outperforms classical approaches. However, they include dietary information as variables which may have introduced severe bias, and their evaluation was not conducted statistically rigorous.

We start out by replicating their experiment 27 times to account for uncertainty of predictive performance estimates. In general, we can confirm their finding that the GloVe embedding retains more information than PCA. In contrast to the original experiment we find that the raw data allows for a more accurate prediction with lower variability in the performance. This indicates that GloVe is a suboptimal embedding already in this very synthetic scenario.

We gain further insight into why PCA loses so much information about the IBD status of a subject. To this end we compared which microbes were considered most important for the raw data Random Forest (RF) and the PCA data RF. We identify three Amplicon Sequence Variants (ASVs) which are highly discriminative for IBD but are not leveraged by the PCA RF. PCA can not reconstruct the abundance of these ASVs since they do not contribute much to the overall variance in the data set.

We visualize the low dimensional coordinate systems and find that the non-orthogonal axes of the GloVe system cluster around four regions of the high dimensional space. The Principal Component Coordinate System covers a wider volume in the input space due to its orthogonality constraint. However, the better performance of GloVe RFs indicates that data set concentrates around the axes-clusters rather than being scattered all over input space.

Finally, we measure distances between subjects along the data manifold to embed the data with UMAP. In a similar prediction experiment as before, we find that k-NN on UMAP data outperforms GloVe RFs. We interpret this as first evidence to support the theoretical arguments for manifold methods.

In summary, PCA is not a suitable embedding for this data set, since variances range over different orders of magnitude. GloVe's retention of co-occurrence ratios retains on average most of the information about the IBD status but is not robust to distribution shifts in the training data. First experiments indicate that a manifold embedding retains even more structure of the original data with less inductive bias and a straight forward interpretation of distances.

5.2 Discussion

Our work focused around mathematical properties of the data and its embedded versions. It remains unclear how the interpretation of the results change, when the data is preprocessed with in depth microbiological knowledge. The findings that the most highly discriminative microbes in Figure 3 were not leveraged extensively by PCA suggests that declaring a lot of variance in the raw data is not necessarily a useful criterion to embed microbiome data. However, we did not transform

the original data before applying PCA. A logarithmic transformation could improve the quality of the PCA embedding.

As there is no explicit notion of what type of statistical information is relevant, it also remains unclear how representative the prediction of the Inflammatory Bowel Disease is for other relevant properties of a microbiome.

Our comparison of important microbes for IBD prediction was only based on a heuristic selection. For example, the threshold of when a microbe was considered important was not altered to check the robustness of the results.

The experiment in Section 4.3 only yields preliminary results. Similar experiments should be conducted for other body sites, different diseases on various data sets, which are beyond the scope of this work. However,the results support theoretical arguments for the necessity of a non-linear embedding.

While GloVe builds a powerful representation, it arises from the belief that co-occurrence ratios correlate with properties of microbiomes. Additionally, GloVe is restricted by the linearity assumption: Each subject must be represented as a linear combination of prototypic subjects (axes of the embedding space). For manifold methods, such as UMAP, the inductive bias (prior assumption) is much weaker. The only assumption is that similarity between two microbiomes can be measured by the distance along the data manifold. Less inductive bias may prove helpful if the embedding shall be used for different downstream tasks. Since almost all Machine Learning methods leverage distances between data points, manifold embeddings could serve as a powerful representation with the least amount of inductive bias.

Ultimately, the evaluation of other data sets and different prediction tasks is necessary to support or falsify our conclusion that manifold distances may be the most informative embedding. In general such kind of prediction experiments shall focus on ruling out confounding effects and biases inherent to the task at hand. In particular, we are ought to consider other magnitudes of class imbalance, e.g. through up/downsampling approaches as well as predicting labels which are understood in more depth by microbiome experts.

References

Breiman, L. (2001). "Random Forests". In: *Machine Learning* 45.1, pages 5–32.

Bukin, Yu S, Yu P Galachyants, IV Morozov, SV Bukin, AS Zakharenko, and TI Zemskaya (2019). "The effect of 16S rRNA region choice on bacterial community metabarcoding results". In: *Scientific Data* 6.1, pages 1–14.

Callahan, Benjamin J, Paul J McMurdie, and Susan P Holmes (2017). "Exact sequence variants should replace operational taxonomic units in marker-gene data analysis". In: *The ISME Journal* 11.12, pages 2639–2643.

Hastie, T., R. Tibishirani, and J. H. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Second. Springer.

Huttenhower, Curtis, Aleksandar D Kostic, and Ramnik J Xavier (2014). "Inflammatory bowel disease as a model for translating the microbiome". In: *Immunity* 40.6, pages 843–854.

Kemp, Karen, Jane Griffiths, and Karina Lovell (2012). "Understanding the health and social care needs of people living with IBD: a meta-synthesis of the evidence". In: *World journal of gastroenterology: WJG* 18.43, page 6240.

Luxburg, U. (2020). *Statistical Machine Learning Leture Script*. Unpublished at University Tuebingen.

McDonald, Daniel, Embriette Hyde, Justine W Debelius, James T Morton, Antonio Gonzalez, Gail Ackermann, Alexander A Aksenov, Bahar Behsaz, Caitriona Brennan, Yingfeng Chen, et al. (2018). "American Gut: an open platform for citizen science microbiome research". In: *Msystems* 3.3, e00031–18.

McInnes, Leland, John Healy, and James Melville (2018). "Umap: Uniform manifold approximation and projection for dimension reduction". In: *arXiv preprint arXiv:1802.03426*.

Molodecky, Natalie A, Shian Soon, Doreen M Rabi, William A Ghali, Mollie Ferris, Greg Chernoff, Eric I Benchimol, Remo Panaccione, Subrata Ghosh, Herman W Barkema, et al. (2012). "Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review". In: *Gastroenterology* 142.1, pages 46–54.

Pennington, Jeffrey, Richard Socher, and Christopher D Manning (2014). "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Podolsky, Daniel K (1991). "Inflammatory bowel disease". In: *New England Journal of Medicine* 325.13, pages 928–937.

Pope, PT and JT Webster (1972). "The use of an F-statistic in stepwise regression procedures". In: *Technometrics* 14.2, pages 327–340.

Sankaran, Kris and Susan P Holmes (2019). "Latent variable modeling for the microbiome". In: *Biostatistics* 20.4, pages 599–614.

Tataru, Christine A and Maude M David (2020). "Decoding the language of microbiomes using word-embedding techniques, and applications in inflammatory bowel disease". In: *PLoS computational biology* 16.5, e1007859.

Tenenbaum, Joshua B, Vin De Silva, and John C Langford (2000). "A global geometric framework for nonlinear dimensionality reduction". In: *science* 290.5500, pages 2319–2323.

Appendix

5.3 Inflammatory Bowel Disease: Medical Background Knowledge

The inflammations are mediated by immune activity, with underlying causes of various kinds: host genetics, environmental factors and microbial activity, which can not be explained comprehensively so far (Huttenhower et al., 2014). IBD has become a popular condition to study microbiome host interactions (Huttenhower et al., 2014).

Crohn's disease (CD) and ulcerative colitis (UC) are the most common type within the group of IBDs. The majority of studies related to either one of these (75% for CD, 60% for UC) report a general upward trend in incidence (Molodecky et al., 2012).

While both diseases are characterized by widely unpredictable acute inflammations, UC is restricted to the colon's mucosa whereas CD patients may exhibit inflammations in the whole gastrointestinal tract (Podolsky, 1991). Acute phases of diarrhea, gastrointestinal bleeding and cramp like strong muscular contractions (colics) are among the most common symptoms which might go into full remission during non active phases of the disease (Kemp et al., 2012).

5.4 Experimental Results

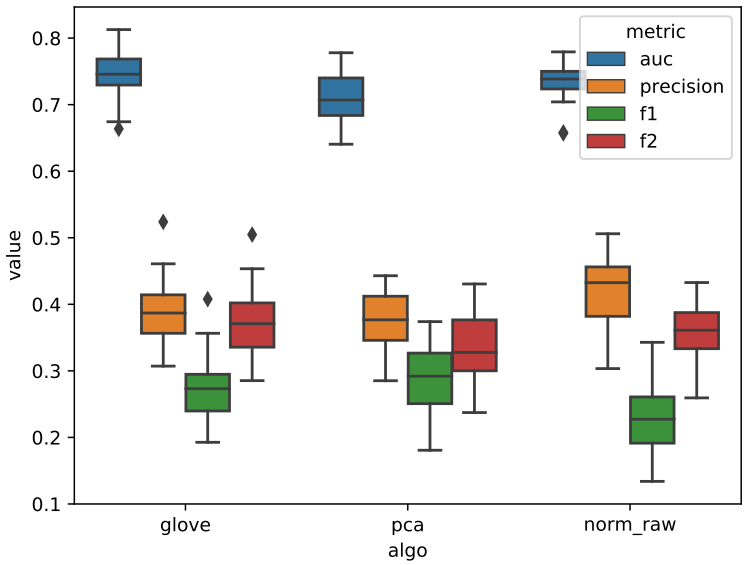

Figure 9: Performance Measures for 27 replications of the IBD-Prediction Experiment. x-axis indicates the embedding type of the input to the Random Forest. Meta Variables included, as in Tataru and David (2020).

5.5 Implementation Details

For ease of reproducibility we give a brief overview over the implementation pipeline. All code can be found at `https://github.com/pat-rig`. All figures and data are provided in respective folders on the toplevel directory. The other three directories `make_embeddings`, `post_process_embeddings`, `prediction_experiments` divide the project chronologically.

5.5.1 `make_embeddings`

In this directory, we subsample the training data and fit the GloVe embeddings. Starting with a list of nucleotide sequences per subject, we convert the data to a conventional *sample* by *ASV-abundance*

matrix. We create a co-occurrence table in `make_glove_input.py`, which is used as the input to the C implementation of GloVe.

Since the C code is executed from a shell script, the major challenge was to automate the GloVe fitting with sample-specific file names. To this end, `make_shell_script.py` creates one `shell` script for each resampling, where the appropriate file names are retrieved from the data directory. Execution on the TCML computing cluster required to retrieve all paths dynamically from temporary working directories.

The resulting embeddings can be found in `./../data/embeddings`.

5.5.2 `post_process_embeddings`

Unfortunately, the output from `runGloVe.sh` is not in an appropriate format to perform the required matrix multiplication for embedding unseen points. For instance, we need to remove entries for microbes that only existed in the GloVe training data. Thus we perform further cleaning and relabelling in this directory.

Finally we embed all test sets into their respective (individual!) embedding in `predict_AGP_testset.py`.

As we replicate their experiments, we leverage Code from Tataru and David (2020), which required some debugging to begin with. Their repository and a description of one major bug can be found at `https://github.com/MaudeDavidLab/gut_microbiome_embeddings`.

5.5.3 `prediction_experiments`

All experimental results and plots were created with scripts from this directory. `PredictIBD.py` fits all Random Forests, evaluates their performances and creates Figures 2 and 9. We extract and aggregate the feature importances from the `sklearn.ensemble.randomForestClassifier` objects in `feature_importances.py`. Evaluating those and plotting Figures 3 - 8 happens in `backtrace_signal.py`.

The UMAP experiment with all of its independent preprocessing pipeline can be found in `optimize_prediction.py`.