DSE Capstone Report - American Gut Project

Neural Embeddings for 16S Microbiome Classification

Ryan Conrad[1], Ryan Inghilterra[1], Sean Rowan[1], Brandon Westerberg[1]
Advisors: Rob Knight[2], Daniel McDonald[2]

1 University of California San Diego, Master of Advanced Study, Data Science and Engineering
2 University of California San Diego, Departments of Pediatrics and Computer Science & Engineering

***Abstract* — The American Gut Project (AGP) [1] is the largest citizen crowd-sourced collection of gut microbiome samples available today. Knowledge of the microbiome is in its beginning stages and the enormous amount of organism and gene effects that are ill-understood makes accurately interpreting results difficult. Reducing this high dimensional space with fundamentally different embedding techniques can be effective in capturing different aspects of the microbiome data to aide in research. Dimensionality reduction techniques like Word2Vec, Hyperbolic Embeddings, and Principal Coordinates Analysis (PCoA) were used to reduce a single sample's dimensionality and explore their different strengths. Embeddings were validated by using them as features for a supervised machine learning model that classifies microbiome body sites (e.g. sebum, feces, saliva). Competing against the state of the art of PCoA using underlying phylogeny distances, the different embeddings kept the baseline logistic regression model's F1 score within acceptable margins at +/- 0.1. These reduction comparisons included actual dimension sizes, metrics of the model prediction, and a representation of samples' clusters. This paper will discuss the analysis, architecture, and visualization of the project that approached this main technical challenge of gaining a better understanding of microbiota.**

I. INTRODUCTION & QUESTION

The study of the human microbiome is a relatively new focus in health research. Valid findings through research in this realm have great potential to improve quality of life and understanding of human bodies for all. The human body contains thousands of human genes but also millions of human microbiome genes, dwarfing our own gene makeup [9]. As discussed, this large crowd-sourced repository of microbiome data has been collected by the American Gut Project (AGP) [1], and is the main source of data for this project. A major challenge in microbiome data studies is the pure amount of microbial imprints found within the gut, sebum, saliva, and other body sites where these organisms live. These imprints are hereby referred to as operational taxonomic units (OTUs) which for this study are 100 nucleotide truncated V4 regions of the 16S SSU rRNA gene referenced using the Greengenes taxonomy database [15]. From a data science perspective, these large amounts of diverse microorganisms makes it difficult to interpret and model for underlying research. In its post-sequenced OTU form this challenge becomes especially difficult in studies with small sample sizes even for the AGP that has tens of thousands of samples. Providing alternative solutions to this problem can aide future interpretability of the microbiome for biological researchers to unveil more of what comprises our human microbiome makeup. The team's approach throughout this paper will support this main hypothesis, that different embedding techniques can be used in symphony to allow further research to discover new unknowns about our microbiomes than before.

B. Related Work

Although fundamentally different to the approaches described in this study previous research have utilized embeddings such as word2vec as a dimensionality reduction technique for body site classification. In 2018 Woloszynek *et al.* described a method for using skip-gram word2vec to embed k-mer sequences of 16S amplicon data for the purpose of body site classification. [2] Using this method body site classification performance was shown to be comparable to using unprocessed OTU abundances. The implementation of word2vec and hyperbolic embeddings in this study take a

different approach where the embedding techniques are generated using OTU co-occurrence information rather than raw nucleotide sequence data.

Microbiome studies leveraging next generation sequencing have revealed an important relationship between microbes and their environment. This close relationship between environment and microbiome composition means that more variation is seen between environments, e.g. mouth or skin, than is seen from person to person. Utilizing this information classification of body site based on prevalence of OTUs can be accomplished at high accuracy. In the context of our research we use this classification task to validate the quality of the generated embeddings to ensure that the important semantics of a given microbiome are retained after undergoing dimensionality reduction.

II. TEAM ROLES

The Capstone project's team structure was taken from advisor guidance and divided among the four student team members. The four roles were Treasurer, Project Coordinator, Record Keeper, and Rotational Support of Roles. B. Westerberg was the Treasurer who maintained the Amazon Web Service (AWS) credits and in summary was a principal investigator of the hyperbolic embeddings, drug clustering, drug integration, and body site classification lead. R. Inghilterra was the Project Coordinator who kept the team up to date on milestones, meetings, and deliverables. In summary, he was the principal investigator of the QIIME [6] library usage, drug permanova testing, PCoA analysis, and lead on visualizations. S. Rowan was the Record Keeper who submitted reports and deliverables, and in summary also was investigator of word2vec embeddings, pipeline data integration, and model hyperparameter execution/tuning. R. Conrad was the Rotational Support of Roles and aided in each role if a member was unavailable or ill. In summary, he also investigated data imputation, data cleaning, pipeline execution/scalability, and experimental metadata modeling.

III. DATA ACQUISITION

A. Data Sources and Sizes

The data sources for the project and project experiments contained in the source code repository are illustrated in Table 1.

Ref. #	Dataset Name	Data Size
1	American Gut Project Drug Questionnaire	200 MB
2	XML Drugbank Database	1.3 GB
3	American Gut Project Metadata & Vioscreen Diet Data	150MB
4	Raw Microbiome Sample Data	101 GB

Table. 1. Data Sources

The main problem statement and results illustrated in this paper solely use dataset #4, the Raw Microbiome Sample Data ultimately used to make embeddings. This data contains 19,762 samples of raw 16S rRNA genomic data sequences. The rest of the data is pulled into an integrated pipeline for the final product that visualizes patient drug, diet, survey, and microbiome data in one view, as well as appearing in other experiments not reported.

B. Data Collection

The data sources #3 and #4 were made available under study 10317 in the Qiita [5] web portal that allows easy downloads of files and ingestion into a storage bank like AWS S3 buckets. The #1 set of drug questionnaires was supplied privately by advisors and #2 was downloaded from the Drugbank [8] web pages into .csv files. All of this data was uploaded to an S3 AWS bucket as pickle files for accessibility from the luigi[1] python pipeline using boto3[2] libraries.

D. Data Setup and Pipeline

All of the teams data was stored in the cloud on AWS to be either run locally or remotely on more powerful servers using a data integration pipeline powered by luigi. There were no databases used in the processing, modeling, or visualizations of the project. All data was downloaded and streamed into memory for processing which had checkpoints to have results loaded back onto S3. For instance, clean debloomed and rarefied Greengenes OTUs were saved to S3 after their laborious

[1] luigi: https://luigi.readthedocs.io/en/latest/
[2] boto3: https://boto3.amazonaws.com/v1/documentation/api/latest/index.html

conversion from raw sequences. Additionally, the team could save embedding runs of Word2Vec or Hyperbolic into S3 for inclusion in the pipeline at anytime. The data pipeline below in Figure 2. and later sections describe this in more detail.

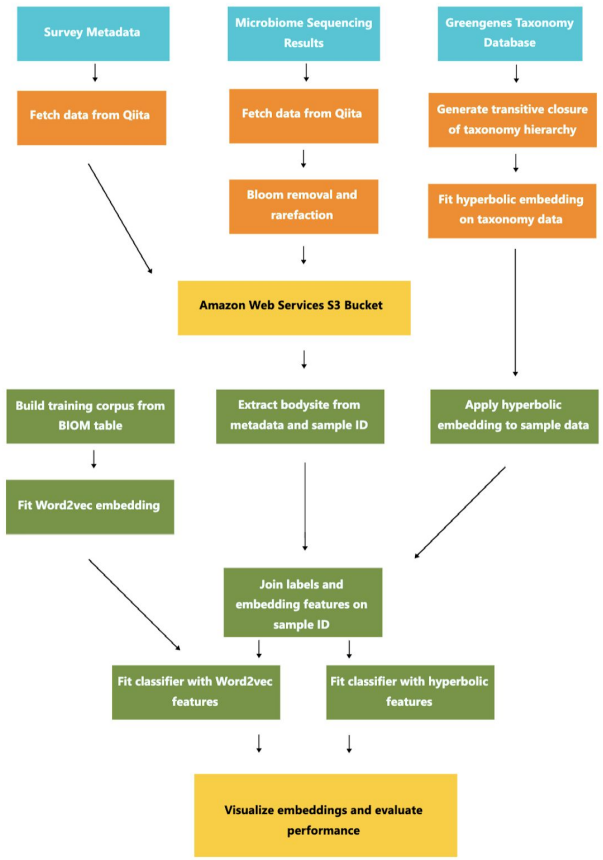

Fig. 1. Abstract Data Integration and Execution Pipeline using Luigi

IV. DATA PREPARATION

A. Microbiome Gene Sequences

The raw 16S data sequences are converted to the OTUs in this study via "97% pairwise identity using as reference the latest release of the GreenGenes (GG) taxonomy" aided by the use of the QIIME libraries. As noted in Section I, these OTUs are 100 nucleotide truncated V4 regions of the 16S SSU rRNA gene using the Greengenes taxonomy database as reference [1]. The preprocessing of these OTUs involves removing blooms and using rarefactions at 1000 sequences/sample. The resulting feature size of OTUs used at 1000 seq/sample was 19,105 per sample.

i. Bloom Removal

Blooms are bacterial growths which can appear on room temperature DNA samples even in sterile environments. Recent work on the American Gut Project has been done to help identify the DNA sequences (OTUs) that correspond to these bloom bacterias. Studies from the American Gut Project show that these blooms can affect the alpha and beta diversity of microbiome samples significantly in some cases. Code was added to our data pipeline to identify any bloom OTUs and remove them from our microbiome dataset. See reference for more details on blooms [11].

ii. Rarefaction

As part of the microbiome data processing, rarefaction is applied to the microbiome OTU data. Rarefaction randomly samples the data we get after sequencing a patient's sample. This is done because the more you sequence the more you will observe. If in one sample you sequence a lot, you are more likely to observe organisms at a 1:100000 relative ratio for instance. This problem is analogous to sampling a forest for example. Counting the number of different types of plants in a square mile of a forest and comparing that to the number you would observe if you were operating at ten square miles. Rarefaction helps to keep the sampled space onto a similar playing field. This rarefaction can be run at different sampling depth levels where different sampling levels will change the number of OTUs produced in the end microbiome dataset [13]. This rarefaction sampling depth level in a sense is a hyperparameter that can be optimized, so our data processing workflow is generic enough to easily pass in different rarefaction levels with the ability to generate new datasets from these different rarefaction scenarios.

iii. Alpha Diversity

Alpha diversity has been shown to be a metric of interest in recent microbiome related research as it helps capture the amount of different observed OTU's that occupy each sample. This can be a basic microbial metric to show differences of 'counts' between samples. We calculated the alpha diversity of each microbiome sample by taking the number of unique operational taxonomic units (OTUs) found in each sample OTU (Observed OTU). This number was integrated into the finalized data

collection and shown for each sample in a visualization among other patient information.

iv. Pythonic Organization

Additional processing included creating 'OTU_ids' for our microbiome dataset and exporting a dictionary of OTU identifiers linked to the raw unique DNA sequence of each OTU. Previously the raw DNA sequences were being used as the ids which was making our dataset larger than necessary and slowing down our analysis work in the python pipeline.

The final format exported after processing the microbiome is a Pandas Sparse Dataframe. This worked well for our large dimensional sparse nature of the OTU data, allowing for storage in a pickle file and easy transportation and ingestion by our other python scripts. Logic was required in our processing steps to convert the microbiome data from qiime2 libraries 'BiomTable' format into a correctly formatted Pandas Sparse Dataframe format.

B. Sample's Survey Metadata

The sample metadata comprises about 350 columns of vast and varying information about a person's life, living, and well-being. The metadata additionally includes 250 columns that comprise Vioscreen diet questionnaires. A subset of this information like age, Body Mass Index (BMI), country of residence, latitude, longitude accompanied patient microbiome and drug data in the final visualization. There are times that survey answers were not filled out and resulted in missing information, showing 'null'. For the most part, the above metadata features have a high percentage of appearance for all samples.

C. Drug Data

In the scope of our research the drug survey data was primarily used to enrich the data visualization aspect of the project. Initial exploratory analysis was able to reproduce some of the widely established correlations between antibiotic use and alpha diversity. In total there were 1,500 logged supplements or drugs used by sample providers. In other experiments, a hierarchical cluster of drugs was produced using the Drugbank database, as well as Omeprazole analysis using permanova tests.

V. ANALYSIS METHODS

The size of all AGP sample submittals at publication of this paper was 19,762, and the feature dimension was 19,105. To make classifiers and principal component interpretation more approachable for this data the techniques to be used as experiments against raw OTU abundances were PCoA, Word2Vec, and Hyperbolic embeddings.

A. Word2Vec Embeddings

The first embedding technique applied to the dataset leverages a technique commonly used in NLP, Word2Vec [14]. This model attempts attempts to learn the contextual meaning of words from a training corpus. Once the model is trained, the model outputs a vector, typically ranging between 50 to 1000 dimensions, corresponding to each word. Similar words should be closer in proximity in vector space. For example, the distance between "dog" and "cat" in this vector space should be smaller than the distance between "dog" and "mountain."

Since the OTU occurrence in each sample is sparse, we hypothesized that a corpus could be created from the OTUs within each sample. Each sample could be seen as a sentence and the value of the OTU as the number of times the word appears in the sentence. In process to create the embedding from the sample is the following: sum the word vector for each word in the sentence and divide by the number of words in the sentence.

Hyperparameter search was performed in order to optimal embedding. The luigi data pipeline made it easy and efficient to parallelize the execution of the hyperparameter search. A few hyperparameters were number of dimensions, epochs, and minimum amount of times a word appears in the corpus. These embeddings also used the continuous bag of words (CBOW) implementation of Word2Vec. Training a logistic regression model to classify body site and comparing the test F1 score was how we determined the effectiveness of the embedding.

PCA was run on the embedded samples in order to visualize the data in both two and three dimensions in order to examine the embeddings. It is clear by the separation of body site clusters in Figure 2 that Word2Vec is able to construct logical embeddings using only OTU data.

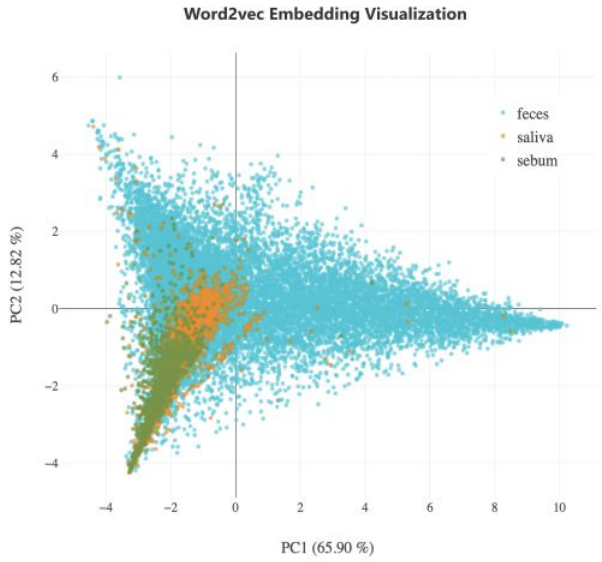

Fig. 2. Word2Vec Embeddings

B. Hyperbolic Embeddings

The goal of embeddings is to describe data in a low dimensional space that maintains the original semantic information contained in the data's uncompressed form. For data that comes in the form of a tree, graph, or any other structure with a latent hierarchy embedding in a low dimensional space, it becomes challenging to do this while faithfully representing the structure of the original data. In 2017 Maximilian Nickel and Douwe Kiela of Facebook AI Research proposed an embedding function to overcome this challenge [12]. Their solution to the problem of representing hierarchical data was to project it into hyperbolic space rather than Euclidean space. The negative curvature of hyperbolic space with its inherent nonlinearity lends itself to effectively representing hierarchical data that contain several layers of edges and vertices. Nickel and Kiela propose several variants of this technique known as Poincare and Lorentz embeddings, which differ in implementation details, both work by projecting data into hyperbolic space.

As originally described by Nickel and Kiela, this type of embedding is highly effective at representing large scale taxonomies in low dimensional space and a prime candidate for describing the taxonomy of each OTU in a way that could be consumed by a machine learning model. Using the Greengenes database that maps OTUs to their estimated taxonomic description we were able to generate a ten dimensional embedding that described the full taxonomy, i.e. kingdom, phylum, class, order, family, genus, species, of each OTU. To create a final embedding for each sample a log weighted average of all OTU embeddings were averaged together to create a final ten dimensional vector describing a sample.

Construction of the embedding was done using the open-source library, poincare-embeddings, published by Facebook Research. To exploit the native GPU support of the library an g3.4xlarge EC2 instance from AWS was used to fit the embedding.

This embedding similarly to the Word2vec one was validated by using the product embedding vectors as input features to supervised body site classification. The embedding was also qualitatively validated by visualizing it in two and three dimensional space using principal component analysis. Despite the different shape of the bodysite clusters it was clear from the visualization exercise that the embedding was highly effective at representing the data using only taxonomic information associated with each OTU.

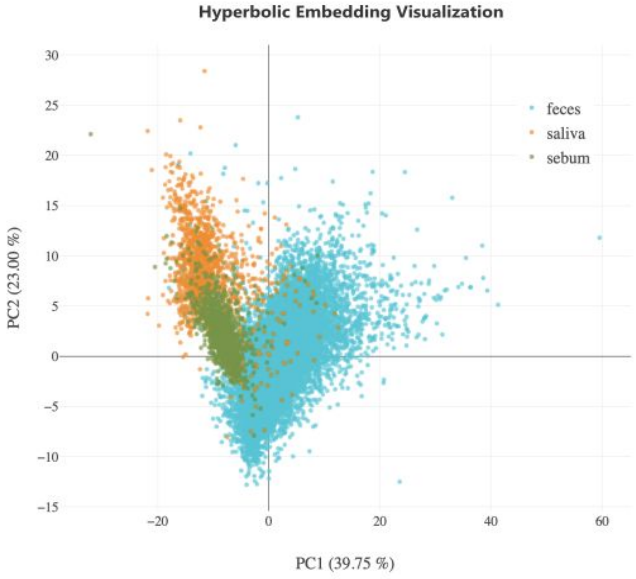

Fig. 3. 2D projection of hyperbolic embeddings obtained from Greengenes taxonomy data

C. PCoA Embeddings

Principal Coordinates Analysis (PCoA) is an extensively researched and powerful dimensionality reduction technique to identify factors explaining differences among microbial communities [7]. PCoA is similar to PCA with the only difference being that the input matrix is a distance matrix instead of a covariance matrix. PCoA has shown to be most effective when used with a UniFrac distance metric [7]. UniFrac is a β-diversity measure that uses phylogenetic information to compare samples. β-diversity was measured between all

of the samples using the Unifrac distance metric with the result being a matrix containing beta diversity similarity measurements between each sample. PCoA was then applied to the matrix, which gave a 3-dimensional embedding of all samples, as well as the percent explained variance of each of these top three dimensions.

The same body site classification and qualitative visual inspection process used on the other embeddings was also used on the PCoA results. From visualizing the top two principal components, it was clear from the distinct separation in body sites why using PCoA with Unifrac is currently the de facto standard for dimensionality reduction in microbial analysis.

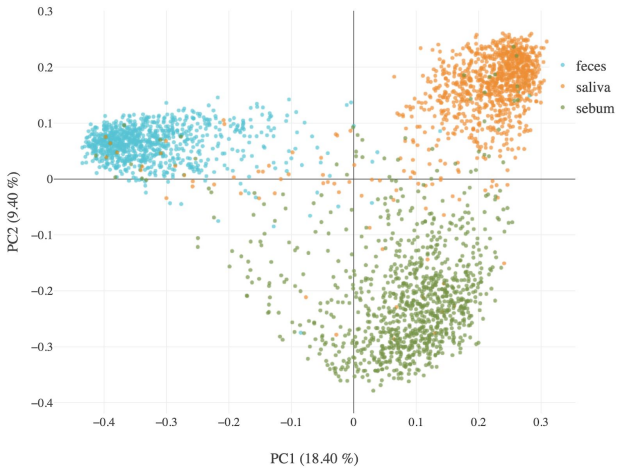

Fig. 4. 2D projection of PCoA embeddings

D. Body Site Classification Model

To provide a baseline of comparison against PCoA embeddings and no embeddings, a linear regression model was implemented at the end of the pipeline to predict the body site locations of where the microbiome OTU's occurred. These sites are sebum (skin), saliva, and feces. The total sample size was 19,762 and was split into train and test sets as 66% and 33% respectively. All of the embeddings were run through the same linear regression model that used L2 penalty, the solver algorithm Limited Memory (LM) Broyden-Fletcher-Goldfarb-Shanno (BFGS) [lbfgs], and C or inverse of regularization strength parameter of $1e^{-3}$. This allowed comparisons of F1 scores and accuracy of prediction in the next section.

VI. FINDINGS

A. Model Metrics

The final results shown in Table 2 are the F1 scores of the body site classification model run with 19,762 row samples. This used a balanced F1 score metric, which can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal. The formula for the F1 score is: F1 = 2 * (precision * recall) / (precision + recall). Since this model predicted multi-class of body sites (i.e. sebum, saliva, feces), a weighted average of the F1 score of each class is the final result.

Embedding	**Dimensions**	**Training Acc. %**	**Test Acc. %**	**Test F1 Score**
No embedding	19105	99.0	98.3	.982
PCoA	3	95.5	95.8	.958
Hyperbolic	10	96.6	96.2	.961
Word2Vec	80	88.9	88.7	.848

Table. 2. Body Site Classification with Embeddings Test Results

These findings show that the neural embeddings perform well compared to the current state of the art which is PCoA. Hyperbolic embeddings for this particular model slightly outperform the PCoA embedding in F1 score, but it also has 7 more dimensions available and could be seen as equal or even slightly less powerful because of this fact. The Word2Vec embeddings, which are most divergent in terms of function from the other embeddings, performs approximately 0.1 less than the other F1 scores. The word2vec model, despite its dimensions size of 80, was not able to capture as much informative data compared to the other embeddings. This is most likely a factor of having no taxonomic information, but makes this approach even more interesting. As seen in the embedding clusters of body site results in Figure 2, the reduction appears to capture different structures of the microbiome data than PCoA and Hyperbolic. The model classification accuracy of word2vec trails behind the other

models with at test accuracy of 88.7%. These are all compared against the pure feature set, which understandably outperforms them all in F1 score and accuracy because of the vast amount of features available for the model to utilize.

A. Embeddings Explained Variance

Embedding	PC1 Explained Variance %	PC2 Explained Variance %	PC3 Explained Variance %
No embedding	40.1	5.0	4.3
PCoA	18.4	19.4	4.4
Hyperbolic	65.8	12.8	4.2
Word2Vec	39.7	23.0	9.3

Table. 3. Embeddings PCA Explained Variance

PCA was performed on each of the embeddings to evaluate the explained variance captured in the top three eigenvectors. The hyperbolic and word2vec embeddings yielded a higher total explained variance in the first three eigenvectors compared PCoA. However this difference in explained variance was uncorrelated to classification accuracy and emphasises the point that capturing high explained variance does not imply high classification accuracy. To see how the embedding results for samples were visualized see Section VII Part C.

VII.TECHNICAL SOLUTION

A. Model Evaluation

The word2vec and hyperbolic embedding models were compared against the control method of PCoA. PCoA represents the gold standard dimensionality reduction technique as it excels at showing distinct separation between body site locations in low dimensions.

B. Performance and Scaling

The UCSD Knight Lab had granted the team usage rights on its datacenter cluster called `barnacle`. These servers have many different options of configuration if running within a Jupyter context or from a job submission on the command line (using PBS [Portable Batch System] qsub). The Jupyterhub instances offer up to 32 cores and 64 GB of RAM for 12 hours, otherwise you can use the `barnacle` datacenter and submit jobs for larger needs.

Due to the modular nature of Luigi it is extensible, flexible, and offers idempotence with checkpointing. This allowed the team to take the Luigi python pipeline and load it onto barnacle as a library. The pipeline was able to run instances of raw data preprocessing and model execution fast and in parallel when spinning up a 32 core and 64 GB RAM instance on barnacle's Jupyterhub. This was much for efficient than using commodity local hardware or computers. The team had used this capability to scan through many hyperparameter variables at once to capture low hanging fruit of optimizing models, even in some experiments we did not report in this paper.

The team also extensively used the checkpointing feature of Luigi and saved post-processed data onto the S3 buckets, like embeddings for example. This allowed the team to focus on iterative improvements of particular parts of the entire execution and therefore resulted in runtime performance not being as challenging compared to the other aspects of the project.

C. Reporting Results Interface

An interactive dashboard for visualizing the different types of embedding results was created using Plotly's Dash framework, with the full implementation done using Python. A user can select the embedding type, a two or three dimensional scatterplot, and the principal components/coordinates of the embeddings are plotted for all samples, with the points colored by body site. Zooming, rotating, and information-on-hover are all supported, providing a powerful exploratory experience to the user. Additionally, a user can select a specific sample from a dropdown menu, and the dashboard will dynamically update the grey section of the user interface with additional sample information such as Alpha Diversity, BMI, medications taken, etc. When a sample is selected from the dropdown, that sample is highlighted with a red marker in the embedding scatter plot, allowing for visual analysis of the sample relative to the samples.

Plotly allows for rendering scatter plots using 'webgl', a highly scalable rendering solution for web browsers. 'Webgl' rendering was used for the dashboard's scatterplot, allowing for quick and scalable rendering of over 17,000 points.

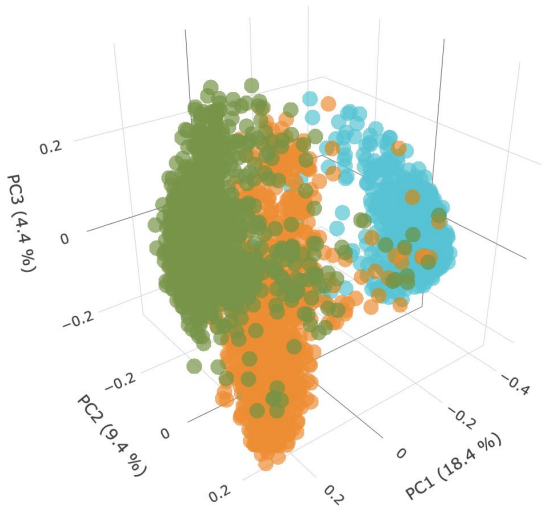

Fig. 5. 3D scatterplot of PCoA embedding

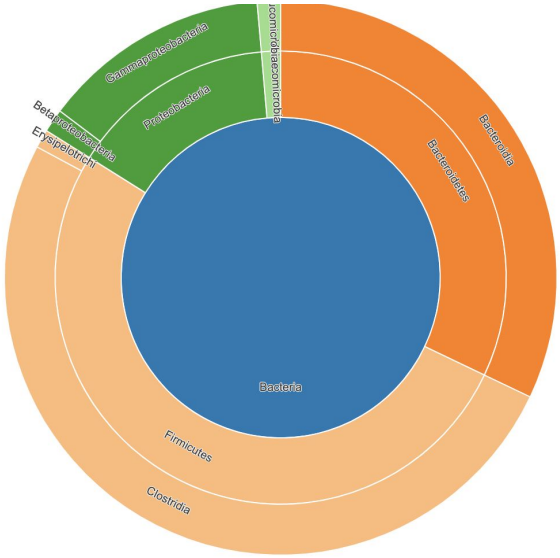

Fig. 6. Sunburst plot showing relative abundances of bacteria found in sample_id: 1042, broken down by phylum and class

In order to visualize the taxonomic breakdown of each sample a sunburst plot was included in the dashboard to complement the 3D embedding scatter plot. The sunburst plot shows the top two levels of the taxonomy breakdown for a selected sample, i.e. phylum and class. This helps recapture the native structure of the data that was lost during the embedding process and helps highlight the power of the hyperbolic embedding when it comes to transforming hierarchical data into a low dimensional continuous space. Like the rest of the visualization dashboard the sunburst plot was created using Plotly. In its current implementation only two taxonomy levels are shown due to the challenges associated with plotting hierarchical data that contains significant depth. The sunburst supports interactivity meaning certain levels of the tree can be collapsed if the user deems there to be too much information being shown. Overall, we found the sunburst plot to be an expressive and effective tool for communicating the taxonomic breakdown of each sample.

VIII.CONCLUSION

The goal set out for the original problem statement was successful. The OTUs from one of the largest available collections of microbiome data was reduced in size by multiple magnitudes by current and experimental embedding techniques that ended up fundamentally capturing different structures of information. These low dimensional vector space features were tested in comparison with the state of the art and comparably performed. The true semantics of these embeddings are most likely not currently appreciated, and future work could shed light on the items they highlight within the microbiome.

An AGP applicable ingestion pipeline that can drastically scale, speed up performance, and allow iterative data analysis of models, features, and data was set up. This pipeline could integrate varying sources of data easily and resulted in a rich dashboard for citizen scientists, researchers, and hobbyists to explore and add to for years to come. Users can view the difference of embeddings at a granular scale, even down to a single sample and its microbiome taxonomy makeup.

IX.REFERENCES

[1] *American Gut: an Open Platform for Citizen Science Microbiome Research* Daniel McDonald, Embriette Hyde, Justine W. Debelius, James T. Morton, Antonio

Gonzalez, Gail Ackermann, Alexander A. Aksenov, Bahar Behsaz, Caitriona Brennan, Yingfeng Chen,Lindsay DeRight Goldasich, Pieter C. Dorrestein, Robert R. Dunn, Ashkaan K. Fahimipour, James Gaffney, Jack A. Gilbert, Grant Gogul, Jessica L. Green, Philip Hugenholtz, Greg Humphrey,Curtis Huttenhower, Matthew A. Jackson, Stefan Janssen, Dilip V. Jeste, Lingjing Jiang, Scott T. Kelley, Dan Knights, Tomasz Kosciolek, Joshua Ladau, Jeff Leach, Clarisse Marotz,Dmitry Meleshko, Alexey V. Melnik, Jessica L. Metcalf, Hosein Mohimani, Emmanuel Montassier, Jose Navas-Molina, Tanya T. Nguyen, Shyamal Peddada, Pavel Pevzner, Katherine S. Pollard,Gholamali Rahnavard, Adam Robbins-Pianka, Naseer Sangwan, Joshua Shorenstein, Larry Smarr, Se Jin Song, Timothy Spector, Austin D. Swafford, Varykina G. Thackray,Luke R. Thompson, Anupriya Tripathi, Yoshiki Vázquez-Baeza, Alison Vrbanac, Paul Wischmeyer, Elaine Wolfe, Qiyun Zhu, The American Gut Consortium, Rob Knight - Casey S. Greene, Editor **DOI:** 10.1128/mSystems.00031-18

[2] *Conducting a Microbiome Study* Julia K. Goodrich, Sara C. Di Rienzi, Angela C. Poole, Omry Koren, William A. Walters, J. Gregory Caporaso, Rob Knight, and Ruth E. Ley Cell. 2014 Jul 17; 158(2): 250–262. **DOI**: 10.1016/j.cell.2014.06.037

[3] *A Guide to Enterotypes across the Human Body: Meta-Analysis of Microbial Community Structures in Human Microbiome Datasets* Omry Koren, Dan Knights, Antonio Gonzalez, Levi Waldron, Nicola Segata, Rob Knight, Curtis Huttenhower, and Ruth E. Ley - Jonathan A. Eisen, Editor **DOI**: 10.1371/journal.pcbi.1002863

[4] *Supervised classification of human microbiota.* Knights D, Costello EK, Knight R. **DOI**: 10.1111/j.1574-6976.2010.00251.x

[5] *Qiita: rapid, web-enabled microbiome meta-analysis,* (Specifically Study ID 10317 Used) Antonio Gonzalez, Jose A. Navas-Molina, Tomasz Kosciolek, Daniel McDonald, Yoshiki Vázquez-Baeza, Gail Ackermann, Jeff DeReus, Stefan Janssen, Austin D. Swafford, Stephanie B. Orchanian, Jon G. Sanders, Joshua Shorenstein, Hannes Holste, Semar Petrus, Adam Robbins-Pianka, Colin J. Brislawn, Mingxun Wang, Jai Ram Rideout, Evan Bolyen, Matthew Dillon, J. Gregory Caporaso, Pieter C. Dorrestein & Rob Knight. Nature Methods, volume 15, pages 796–798 (2018); **DOI**: https://doi.org/10.1038/s41592-018-0141-9

[6] *QIIME 2™ is a next-generation microbiome bioinformatics platform* Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet C, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope E, Da Silva R, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley G, Janssen S, Jarmusch AK, Jiang L, Kaehler B, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MG, Lee J, Ley R, Liu Y, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton J, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson, II MS, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CH, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, Caporaso JG. 2018. QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. PeerJ Preprints6:e27295v2 **DOI:** https://doi.org/10.7287/peerj.preprints.27295v2

[7] *UniFrac: a new phylogenetic method for comparing microbial communities.* Lozupone C, Knight R **DOI**: 10.1128/AEM.71.12.8228-8235.2005

[8] *DrugBank 5.0: a major update to the DrugBank database for 2018.* Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, Wilson M. Nucleic Acids Res. 2017 Nov 8. **DOI**: 10.1093/nar/gkx1037.

[9] *The relationship between the human genome and microbiome comes into view,* Julia K. Goodrich, Emily R. Davenport, Andrew G. Clark, and Ruth E. Ley, **DOI**: 10.1146/annurev-genet-110711-155532

[10] *16S rRNA sequence embeddings: Meaningful numeric feature representations of nucleotide sequences that are convenient for downstream analyses* Stephen Woloszynek, Zhengqiao Zhao, Jian Chen, Gail L. Rosen. bioRxiv 314260; doi: https://doi.org/10.1101/314260. Now published in *PLOS Computational Biology* doi: 10.1371/journal.pcbi.1006721

[11] *Correcting for Microbial Blooms in Fecal Samples during Room-Temperature Shipping,* Amnon Amir, Daniel McDonald, Jose A. Navas-Molina, Justine Debelius, James T. Morton, Embriette Hyde, Adam Robbins-Pianka, Rob Knight, *Mani Arumugam, Editor,* **DOI:** 10.1128/mSystems.00199-16

[12] *Poincare Embeddings for Learning Hierarchical Representations*, Nickel, Maximilian and Kiela, Douwe, Advances in Neural Information Processing Systems 30. I. Guyon and U. V. Luxburg and S. Bengio and H. Wallach and R. Fergus and S. Vishwanathan and R. Garnett. 6341--6350. 2017. Curran Associates, Inc.

[13] *Normalization and microbial differential abundance strategies depend upon data characteristics*, Sophie Weiss, Zhenjiang Zech Xu, Shyamal Peddada, Amnon Amir, Kyle Bittinger, Antonio Gonzalez, Catherine Lozupone, Jesse R. Zaneveld, Yoshiki Vázquez-Baeza, Amanda Birmingham, Embriette R. Hyde and Rob Knight **DOI:** https://doi.org/10.1186/s40168-017-0237-y

[14] *Distributed Representations of Words and Phrases and their Compositionality*. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean DOI: **DOI:**https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

[15] *An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea* Daniel McDonald, Morgan N Price, Julia Goodrich, Eric P Nawrocki, Todd Z DeSantis, Alexander Probst, Gary L Andersen, Rob Knight, and Philip Hugenholtz **DOI:** 10.1038/ismej.2011.139

APPENDIX A

MAS DSE Attributions

Our team would like to thank our advisors Rob Knight and Daniel McDonald for the extensive knowledge shared about the American Gut Project and microbiome domain. We would also like to attribute the final project solution and success to Professor Ilkay Altintas for her support and guidance throughout the breadth of the project, we could not have done it without these key personnel.

The entire team wholly used the teachings from UCSD's Jacobs School of Engineering in the Masters of Advanced Study Degree for Data Science and Engineering. The statistical knowledge from classes aided greatly during effect size, permanova, and Faith PD tests and trials. Teachings on Principal Component Analysis (PCA) and multidimensional scaling (MDS) aided in the overall investigations of this project, as it was part of the main hypothesis' goal. The visualization class was key in teaching the team the best ways to share difficult information to the end user in our visualizations comparing different embeddings and the microbial makeup of samples. The hyperparameter tuning of the models on linear regression, SVM, random forests, and XGBoost would not have been as simple if it weren't for machine learning classes, as well as knowing the extensibility and power of XGBoost itself that was used in experiments contained in the project's repository. To the entire staff of the MAS DSE program, we thank you and we will carry our knowledge steadfast and into the future.

APPENDIX B

UCSD Library Archive

MAS DSE Capstone - American Gut Project Cohort 4 2019, Ryan Conrad, Ryan Inghilterra, Sean Rowan, Brandon Westerberg, Daniel McDonald, Rob Knight, **DOI**: https://doi.org/10.6075/J0HT2MN3