Learning a deep language model for microbiomes: the power of large scale unlabeled microbiome data

Quintin Pope[1◑], Rohan Varma[2◑], Chritine Tataru [3], Maude David[3,4], Xiaoli Fern[1]

1 School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, Oregon, United States

2 Computer Science and Engineering, University of Michigan, Ann Arbor, Michigan, United States

3 Department of Microbiology, Oregon State University, Corvallis, Oregon,United States

4 Department of Pharmaceutical Sciences, Oregon State University, Corvallis, Oregon, United States

◑These authors contributed equally to this work.

*popeq@oregonstate.edu

Abstract

We use open source human gut microbiome data to learn a microbial "language" model by adapting techniques from Natural Language Processing (NLP). Our microbial "language" model is trained in a self-supervised fashion (i.e., without additional external labels) to capture the interactions among different microbial species and the common compositional patterns in microbial communities. The learned model produces contextualized taxa representations that allow a single bacteria species to be represented differently according to the specific microbial environment it appears in. The model further provides a sample representation by collectively interpreting different bacteria species in the sample and their interactions as a whole. We show that, compared to baseline representations, our sample representation consistently leads to improved performance for multiple prediction tasks including predicting Irritable Bowel Disease (IBD) and diet patterns. Coupled with a simple ensemble strategy, it produces

a highly robust IBD prediction model that generalizes well to microbiome data independently collected from different populations with substantial distribution shift.

We visualize the contextualized taxa representations and find that they exhibit meaningful phylum-level structure, despite never exposing the model to such a signal. Finally, we apply an interpretation method to highlight bacterial species that are particularly influential in driving our model's predictions for IBD.

Author summary

Human microbiomes and their interactions with various body systems have been linked to a wide range of diseases and lifestyle variables. To understand these links, citizen science projects such as the American Gut Project (AGP) have provided large open-source datasets for microbiome investigation. In this work we leverage such open-source data and learn a "language" model for human gut microbiomes using techniques derived from natural language processing. We train the "language" model to capture the interactions among different microbial species and the common compositional patterns that shape gut microbiome communities. By considering the entirety of species within a sample and their interactions, our model produces a representation that enables contextualized interpretation of individual bacterial species within their microbial environment. We demonstrate that our sample representation enhances prediction performance compared to baseline methods across multiple microbiome tasks including prediction of Irritable Bowel Disease (IBD) and diet patterns. Furthermore, our learned representation yields a robust IBD prediction model that generalizes well to independent data collected from different populations. To gain insight into our model's workings, we present interpretation results that showcase its ability to learn biologically meaningful representations.

1 Introduction

Identifiable features of the human microbiome and its interactions with various body systems have been associated with a wide range of diseases, including cancer [1], depression [2, 3] and inflammatory bowel disease [4–6]. As our knowledge of such

connections has advanced, research on the human microbiome has undergone a shift in focus, moving from establishing links to unraveling the underlying mechanisms and utilizing them to develop clinical interventions [7]. This transition has sparked interest in applying statistical methods to microbiome data, leading to the launch of open source projects such as the American Gut Project (AGP) and Human Food Project (HFP), which provide open source datasets for microbiome investigation [8]. These repositories offer data in the form of raw genetic reads, which, even after being processed into taxa counts, still present thousands of features per sample. Consequently, researchers often employ dimension reduction techniques to transform this data into a more manageable feature space.

Significantly, the relevance of microbes to any particular analysis is often intertwined with the presence and potential interactions of other microbes in the environment. Unfortunately, most dimension reduction techniques commonly used in microbiome data analyses, such as binning based on phylogenetic relationships [9, 10], clustering based on gene similarity [11], or ordination methods utilizing PCA and other techniques [12], are unable to consider these interactions when producing lower dimensional representations of samples. Consequently, a significant challenge in microbiome data analysis is to produce lower dimension representations (embeddings) of samples that not only take into account the presence of specific taxa but also how their interactions and overall functioning as a whole.

Fortunately, a similar challenge has been investigated in the natural language processing (NLP) domain, which shares many similarities with the microbiome domain. Just as a sample comprises numerous microbes, a sentence consists of multiple words. Similarly, certain microbes hold greater relevance for specific analyses, while certain words are more important for different NLP tasks. Furthermore, just as a microbe can assume different functional roles under varying conditions, a word can possess different meanings in different contexts.

Given the strong similarities between the two domains and the shared goal of producing quality lower-dimensional sample / sentence representations, there is a growing interest in applying NLP techniques to microbiome analysis. Notably, previous work has successfully applied NLP word embedding algorithms to microbiome data, generating taxa embeddings that have shown promising results surpassing the

performance of traditional dimension reduction techniques like PCA for various microbiome prediction tasks [13].

However, this prior work [13] has several limitations. First, the embeddings are learned based on aggregated global microbe-to-microbe co-occurrence statistics — in reality, microbe interactions can be dynamic and context-dependent. Second, given a sample containing many taxa, the embedding for the sample is computed by taking an abundance-weighted-average of the taxa embeddings without considering the context-specific roles of individual microbes in the sample. Consider the word "fly", which has drastically different meanings in the two sentences "I like to fly" and "I caught a fly". Using the weighted average of word embeddings as sentence representation will fail to capture the significant differences between the two sentences as they share half of their words.

Transformers, an exciting recent development from NLP [14], provides a potential solution to above issues. Past work [15–19] has applied transformers to biological data. However, such work has focused on learning a sequence encoder for representing DNA [19] or, more commonly, protein amino acid sequences [15–18] (e.g., each token might represent a k-mer in such a sequence). In contrast, we focus on representing entire microbial communities and their interactions, using each token to represent a single microbe in such a community.

We present the first use of transformers to learn representations of microbiome at the taxa level by adapting "self-supervised" pre-training techniques from NLP, allowing the model to learn from vast amounts of unlabeled 16S microbiome data and mitigating the required amount of expensive labeled data. The pre-trained models can be viewed as a form of "language model" for microbiome data, capturing the inherent composition rules of microbial communities, which we can easily adapt to downstream prediction tasks with a smaller amount of labeled "finetuning" data.

We show that using a transformer model pre-trained on data from the American Gut Project (AGP) as the starting point, we can achieve state of the art performance for multiple downstream host phenotype prediction tasks including IBD disease state prediction. These results showcase the remarkable capability of the pre-trained microbial "language" model in generating enhanced representation of the microbiome.

Focusing on the IBD prediction task, we demonstrate that our IBD prediction model,

trained on the IBD data from the American Gut Project, with a simple ensemble strategy, exhibits robust generalization across several IBD studies with notable distributional shifts. We further visualize the contextualized taxa embeddings produced by our pre-trained language model and show that they capture biologically meaningful information. Finally, we analyze the learned IBD prediction model to identify taxa that strongly influence the model's prediction.

2 Materials and methods

We begin by introducing the general workflow of applying a transformer model for generating a sample embedding (Fig. 1) and explaining each step of the work flow, including a detailed look into the transformer architecture. We then explain how we perform the pretraining, followed by finetuning for specific down stream tasks (Fig. 2). This section will also explain how we identify those taxa that most affect the model's classification decisions (Eq. 1) and conclude with a description of the datasets used in this paper.

2.1 Transformers for microbiome data: workflow overview

Since their introduction in 2017, [14] transformers have emerged as one of the most powerful classes of neural models invented to date, demonstrating state-of-the-art performance in many domains, though different tasks and data types require specific adaptations. Figure 1 summarizes the basic workflow of applying transformer to microbiome data for generating sample representations and context-sensitive taxa embeddings.

Preprocessing steps.

We assume that microbiome samples are represented as vectors of relative taxa abundances (Fig. 1A). To prepare our input for the transformer model, we perform a pre-process step (Fig. 1B) to transform the microbiome sample into 'text-like' inputs (Fig. 1C). Specifically, we rank all the taxa present in the sample in decreasing order of abundance to create an ordered list of taxa (truncated to contain no more than the 512

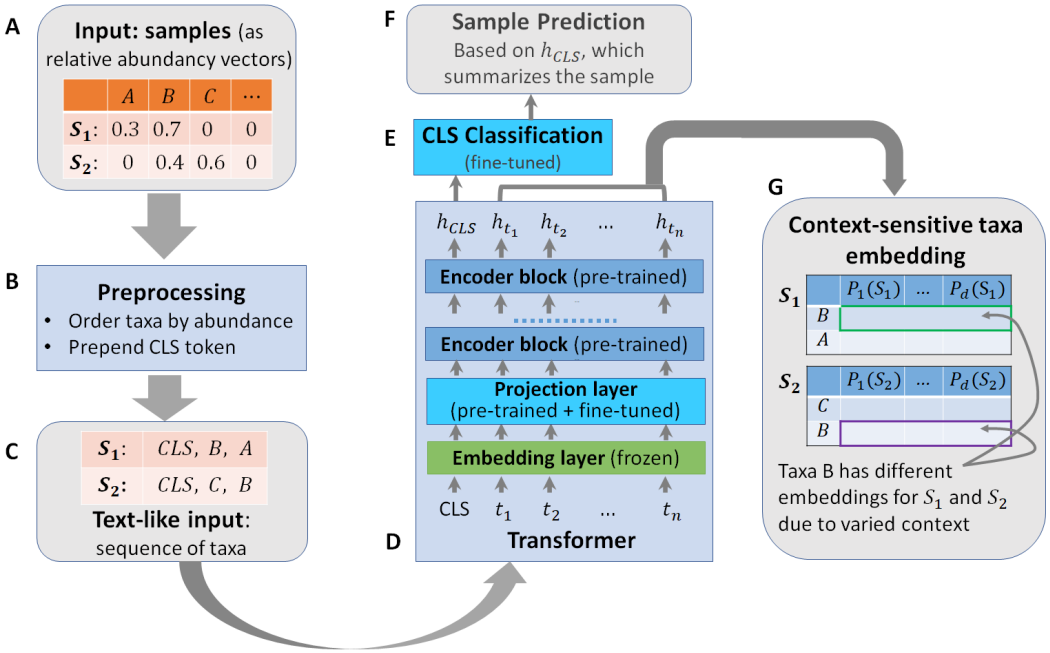


Fig 1. Workflow of using a transformer model for generating sample embedding/classification and context sensitive taxa embeddings. The inputs (A), which are samples represented as relative abundance vectors, first go through the preprocessing step (B) to generate text-like inputs (C) for the transformer model (D). The transformer model generates a sample embedding (h_{cls}) that goes through a sample classification layer (E) to produce task specific sample level predictions (F). The transformer model also generates context sensitive embedding (G) for each taxa in the sample. Same taxa appearing in different samples can have different embedding because of contextual differences.

most abundant taxa). This step creates inputs that are analogous to texts, which are ordered lists of tokens of variable length capped at 512.

Similar to what is done in processing textual inputs, we prepend a special 'classification (CLS)' token to our input list. The 'CLS' token serves the critical role of summarizing the full sample during the later processing steps and provides the final sample representation for classification purposes.

The transformer model

Fig. 1D provides a sketch of our transformer architecture for performing a sample classification task. The input to the transformer model is an ordered list of taxa. The list first goes through an embedding layer and a projection layer. The output of the projection layer then feeds into a sequence of multiple encoder blocks (we use 5 encoder blocks in this work), where each encoder block produces a new representation based on outputs of the previous block. Below we explain the individual components.

Embedding layer. The embedding layer maps from discrete tokens/taxa to their corresponding vector representations. We use absolute positional embeddings [14] to encode the abundance-based taxa order into the taxa embeddings. We preset the embedding layer using the 100-dimensional embedding from [13] and freeze them during training, except for the 'CLS' token embedding, which is initialized randomly and learned during fine-tuning.

Projection layer. The projection layer is a linear transformation from the vocabulary embedding space to the model's hidden representation space. The projection layer allows the model to process inputs of different dimensionality than the model's hidden space. In this work, the projection layer projects from the 100 dimensional vocabulary embedding into a richer 200 dimensional hidden space used by the model.

Encoder blocks. This is where the transformer begins incorporating "context" into the representation of each ASV in the sample. Here we provide an intuitive explanation of the encoder block. Please see [14] for concrete mathematical definitions.

An encoder block consists of a multi-headed self-attention layer [14] and a fully connected layer. The multi-headed attention layer computes a set of self-attention

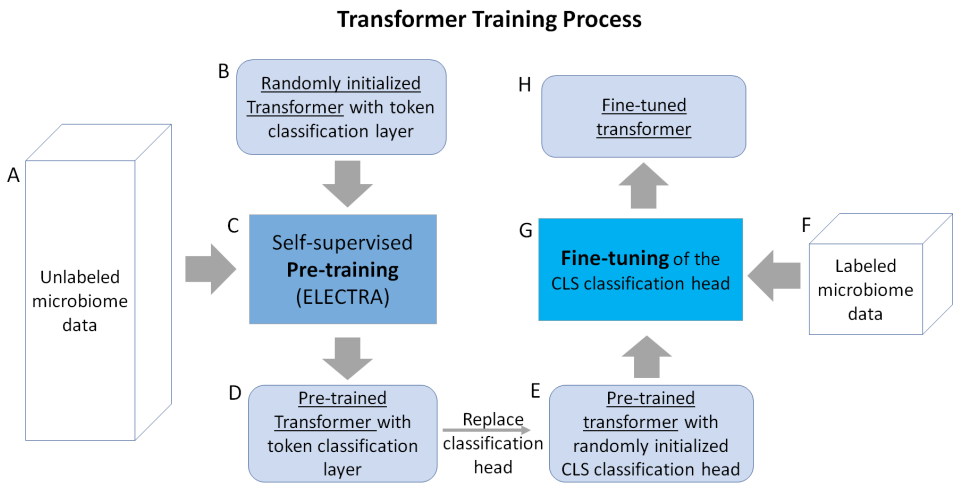


Fig 2. Training of the transformer model. Unlabeled microbiome data (A) is fed into a randomly initialized transformer (B) as inputs to the self-supervised pre-training process (C), which produces a pre-trained transformer that generates token-level classifications (D). We replace the token-level classification head with a randomly initialized CLS classification head (E), and use labeled microbiome data (F) to fine-tune the CLS classification head (G), which produces the fine-tuned transformer (H).

scores (one per head). Each attention head tracks its own set of all-pairs interactions between every taxon in the sample, allowing different heads to track different patterns of interactions between taxa.

The network modulates how much 'attention' is paid to each context taxon when updating the representation for a particular taxon in the sample. For example, in the context of language and given a sentence such as "I waved at the band, but they didn't see me", a properly trained encoder block could update the embedding of word "they" to reflect that it is referencing "the band". Analogously, in microbiome data, if bacteria A performs a functional role conditioned on the presence of bacteria B, a properly trained encoder block could update the embedding for bacteria A to reflect the presence or absence of bacteria B.

Classification head. We rely on a special 'CLS' token to summarize information from all the other taxa / tokens. The CLS token then feeds into a classification head, which is a standard two-layer feed forward neural network with 200 hidden nodes, to produce a prediction for a specific classification task.

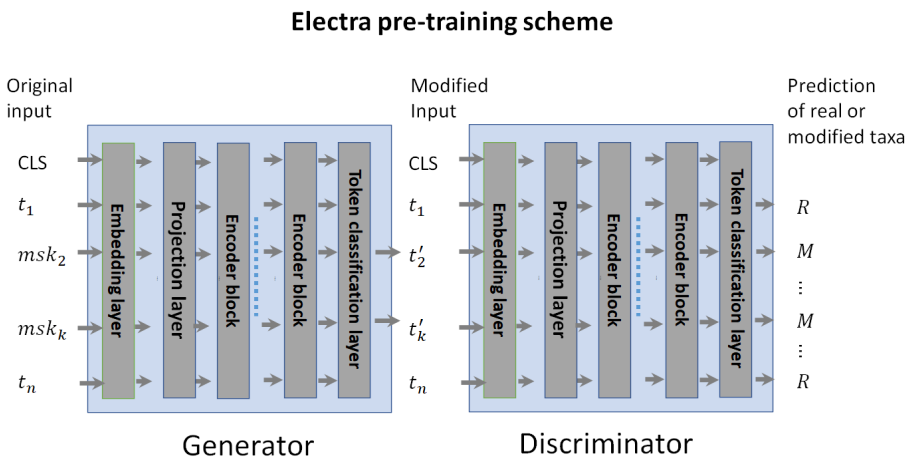


Fig 3. Electra pre-training diagram. A generator is trained to predict the masked taxa from a sample. A discriminator is trained to differentiate taxa filled in by the generator from the original taxa in the sample. Both use the same transformer architecture, and have token level classification heads. The generator token level classification head predicts the taxa ID whereas the discriminator token level classification head predicts the input taxa as "Real" or "Modified"

2.2 Transformer training

A critical challenge in applying complex deep learning models like transformers is the lack of large amounts of labeled training data. This can be addressed, however, using a technique referred to as self-supervised pre-training [20], which leverages readily available unlabeled data. In this work, we follow this approach and our training process is described in Fig. 2.

Pre-training

We begin with a randomly initialized transformer and first train a task-agnostic transformer using unlabeled data via self-supervised pre-training. Specifically, We use ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) [21] to pre-train the encoder layers of the transformer model. We chose ELECTRA because it reaches comparable performance to other popular pre-training approaches (BERT [22] and its various flavors) while being computationally efficient.

The ELECTRA pre-training approach has two steps. The first step trains a generator model by randomly masking out 15% of taxa in microbiome samples and training the generator model to predict the missing taxa based on the remainder of the

sample. For the second step, we use the trained generator to produce perturbed microbiome samples by replacing all the masked taxa with generator predictions and train a discriminator model to differentiate the original taxa of the sample from those replaced by the generator. Essentially, the generator attempts to fill in the masks with taxa predictions and the discriminator takes in the predicted sequence and attempts to identify which taxa are modified by the generator.

Both the generator and discriminator models have the same general architecture as shown in Fig. 3. To train the generator, the inputs are randomly corrupted by replacing 15% of taxa IDs with a special 'mask' ID, and the embedding of each masked taxa after the final encoder layer is fed into a classification head to predict the ID of the masked taxa.

To train the discriminator model, we take the masked sample completed by the generator as input to the discriminator, and feed the embedding of each taxa after the final encoder layer into a classification head that differentiates 'real' (original taxa) from 'modified' (generated taxa).

At the end of pre-training, we have two transformer models, the generator and discriminator. Following the practice of the original work, we use the encoder of the pre-trained discriminator as the initial model to be fine tuned for downstream tasks.

Pre-training details. We perform pre-training on 18480 gut microbiome samples from the American Gut Project Database using the ELECTRA scheme as described above. Specifically, the generator was trained for 240 epochs to predict masked microbe embeddings, and checkpoints of the model were saved every 30 epochs. The discriminator was then trained on the replacement prediction task for 120 epochs with replacements generated by the increasingly trained generator. Specifically, for every 15 epochs of discriminator training, we replace the generator used to produce inputs for the discriminator with a stronger generator using the previously mentioned checkpoints. For example, the generator trained for 30 epochs provided inputs for the first 15 epochs of discriminator training. Then, for epochs 16-30, discriminator inputs were provided by the generator trained for 60 epochs. This was done to gradually ramp up the difficulty of the replacement prediction task.

Task specific fine-tuning

Given a specific prediction task and the pre-trained discriminator, we remove the token classification head and add a new (randomly initialized) sequence classification head to the 'CLS' token. In addition to the embedding layer, we also freeze the parameters of the encoder blocks such that only the classification head and the projection layer was trained during fine-tuning. In other words, the pre-trained discriminator encoders are used as a universal encoder for representing microbiome samples for different prediction tasks. Empirically we have found this practice reduces overfitting and produces more robust generalization performance across different tasks.

Fine-tuning details. We perform fine-tuning using Stochastic Gradient Descent (SGD) optimization with a learning rate of 0.01 and the mean squared error loss, which we found gave better results than the more traditional negative cross-entropy loss, potentially because mean squared error is more robust to noise and outliers. Furthermore, during training, we perform data augmentation by randomly deleting 10% of the input taxa in each training sample to increase the robustness of the trained model and reduce overfitting. The SGD optimization is performed for a total of 50 epochs. Instead of using the final model, we use a separate set of validation data to evaluate the model's performance at the end of each epoch and select the best model to use.

2.3 Feature ablation attribution: finding the important taxa

Given a machine learning model making a prediction for microbial samples, we are interested in finding which microbe species whose presences push the model towards making a positive or negative classification for samples containing those species. To this end, we use feature ablation attribution [23].

Consider a sample X containing n microbial species, which the model predicts as being positive (for some property, e.g, IBD) with probability $M(X)$. Feature ablation individually deletes each microbe species from the original X, then records how much each species' removal reduces the model's predicted probability of being positive. We average these changes across every sample in which a species appears, giving the expected change in classification probability caused by deleting the species in question

from a random sample containing the species.

Given a dataset $\mathbf{D}$, let $\mathbf{D}_m$ denote the set of samples that contains a specific microbe m, we can calculate m's attribution $a(m)$ as:

$$a(m) = \frac{1}{|\mathbf{D}_m|} \sum_{X \in \mathbf{D}_m} \mathbf{M}(X) - \mathbf{M}(X \setminus m) \tag{1}$$

where $\mathbf{M}(\cdot)$ denotes the model's probabilistic output for the given input and $X \setminus m$ denotes sample X with microbe m removed.

2.4 Datasets

We use three different datasets over the course of this study. We now describe them and summarize where they are used.

American Gut Project (AGP). The American Gut Project (AGP) [8] is a crowdsourced microbiome data gathering effort. From it, we used 18480 microbiome samples sequenced from the v4 hypervariable region of the 16S gene that were curated by the authors of [13]. The sample sequences come with metadata information on the subject the sample originates from, providing information about their diet, medical status on inflammatory bowel disease and more. We used all 18480 samples for our pre-training and relevant portions in our evaluation of downstream tasks. We now describe the three downstream tasks we ran experiments on.

- Inflammatory Bowel Disease (IBD). This task aims to predict whether a given microbiome sample belongs to an individual diagnosed with IBD or not. Samples originating from individuals with IBD are the positive class. Label information was drawn from AGP metadata producing 435 samples from IBD positive individuals and 8136 healthy controls.

- Frequency of fruit in diet. This task aims to determine the frequency with which an individual consumes fruits based on their microbiome sample. The label is derived from AGP metadata, which ranks fruit consumption frequency on a one to five scale. For this experiment, samples ranked 3-5 are grouped to form the positive (frequent) class. Samples ranked 0-2 are considered negative (infrequent). Out of 6540 AGP examples with fruit metadata, 4026 were labeled positive.

- Frequency of vegetable in diet. This task aims to determine the frequency with which an individual consumes vegetables based on their microbiome sample. In the same manner as the fruit task, label information was drawn from the AGP metadata and frequency ranks from 0-5 were grouped to form the "frequent" (3-5) and "infrequent" (0-2) classes. Out of 6549 AGP examples containing vegetable frequency metadata, 5654 were labeled positive.

Table 1 provides the summary statistics for the three classification tasks.

Table 1. Three classification tasks derived from the AGP data and meta data.

Datasets	AGP-IBD	AGP-Fruit	AGP-Vegetable
# of samples	8571	6540	6549
# of positive samples	435	4026	5654
# of negative samples	8136	1514	895

Halfvarson (HV). This dataset comes from an IBD study performed in [24]. We used the curated dataset produced in [13], which contains 564 microbiome samples, with 510 of them IBD positive.

HMP2. This dataset comes from an IBD study performed as part of phase 2 of the Human Microbiome Project [25]. Again, we used the curated dataset produced in [13], which contains 197 microbiome samples with 155 IBD positive examples.

3 Results and Discussion

3.1 Transformer representations outperform baseline on multiple microbiome tasks

In this section, we empirically compare transformer-produced sample representations against the baseline method of abundance-weighted-averaging [13]. As mentioned previously, our method applies a standard multi-layered perceptron (MLP) classifier to the transformer-produced sample representations for classification. To remove the impact of using different classifiers and ensure a fair comparison, we replaced the random forest classifier used in prior work with the same MLP classifier. We evaluate both methods using the AGP dataset on three microbiome classification tasks.

For each task, we perform evaluation using 5-fold Cross Validation. For each cross-validation run, 80% of the available labeled data was used for training, and the remaining 20% was split between a validation and test set. We use the validation set to determine the stopping point for training the MLP classifier and use the test set to evaluate the resulting classifiers. We consider two different evaluation criteria: the Area Under the ROC Curve (AUROC) and the Area Under the Precision-Recall curve (AUPR).

Table 2 shows the performance of both methods on three tasks. We see that for all three tasks, the transformer produced representation achieved substantially improved performance for both AUROC and AUPR. This confirms that our approach learns a transformer model that produces robust sample representation that performs well across multiple prediction tasks.

Table 2. Average performance (standard deviation) on Three Tasks

	IBD Task		Fruit Diet Task		Vegetable Diet Task	
	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR
Baseline	0.777(.03)	0.391(.05)	0.581(.02)	0.697(.02)	0.663(.01)	0.927(.004)
Transformer	0.802(.02)	0.529(.07)	0.676(.01)	0.777(.01)	0.735(.03)	0.949(.01)

3.2 Generalization to independent datasets

One of the largest challenges in working with microbiome data is that there is large variance in the distributions and characteristics of data used from study to study. Therefore it is important to test how well our transformer based prediction models generalize on independent datasets that come from different population/sample distributions. To test this, we applied our transformer model trained for the IBD prediction task using the AGP data on the Halfvarson and HMP2 datasets from independent studies.

An issue that arises when performing such cross-study tests is the need to decide a stopping point during finetuning to pick the best model to use on the test data. In the previous single study experiments, using a held-out validation set for this purpose proved to be an effective strategy. However, due to the substantial distributional shift between the AGP data used for training/validation and the independent test set of Halfvarson and HMP2, using a held-out AGP validation set for stopping is observed to

lead to highly unstable performance. We address this problem by introducing a simple ensemble strategy. During fine tuning, we train an ensemble of k classifiers using different random initializations of the classification head. Similar to the standard practice when applying transformer to language [22], we found that each individual classifier only needs to be fine-tuned for a single epoch, i.e., going over all of the training once, and that training more epochs often leads to overfitting. In our experiments, we used ensemble size $k = 10$. Similarly, we compare our ensemble performance with the baseline of [13] using a ensembled MLP classifier. In this comparison we gave the baseline an advantage by reporting the best *testing* performance achieved by the baseline method during training. We report the performance of both methods averaged across five random runs with different initialization in Table 3. The results show that our method achieves consistently better performance on the Halfvarson dataset and comparable performance on the HMP2 dataset compared to the best performing baseline model selected using testing data.

Table 3. Average performance (standard deviation) on independent IBD datastes. Baseline's standard deviation is close to zero, thus omitted.

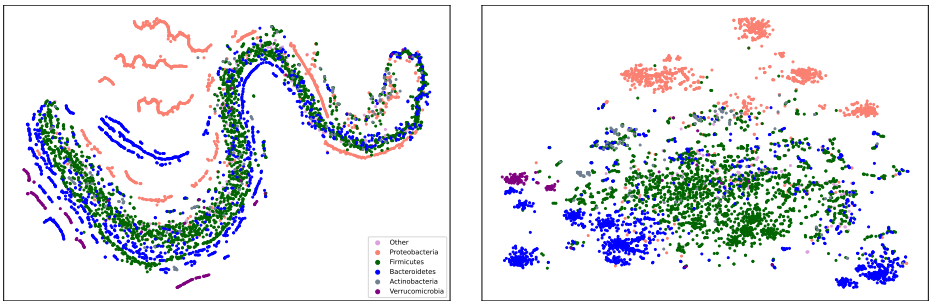
	HMP2		Halfvarson	
	AUC	AUPR	AUC	AUPR
Baseline (ensemble)	0.668	0.863	0.752	0.962
Transformer (ensemble)	0.6822 (.02)	0.855 (.01)	0.805 (.01)	0.973 (.001)

In this section we take a closer look at the pre-trained language model to interpret the learned context-sensitive representations of microbial species.

3.3 Context sensitive taxa embedding captures biologically meaningful information

We hypothesize that the superior predictive performance of our model is because our pre-trained language model transforms the input taxa embedding into a more meaningful latent space capturing context sensitive information (Fig. 1G), making biologically relevant features of the taxa more readily extracted and applied to downstream tasks. Here we compare the geometry of the original taxa embedding with the geometry of the transformed context sensitive representations.

We focus on the top 5000 (out of 26726) most frequent taxa from the IBD dataset and computer their averaged contextualized embeddings across every entry in the IBD



(a) (b)

Fig 4. t-SNE visualization of (a) original taxa vocabulary embeddings and (b) contextualize taxa embeddings. Both are colored by phylum.

dataset. Fig. 4 shows the t-SNE [26] visualization of the taxa using the original vocabulary embedding from [13] (a) and the averaged contextualized embeddings produced by our model (b), colored by the phylum of the taxa assigned by the DADA2 tool [27].

From Fig. 4, we see that the original embedding space in panel (a) displays a degree of clustering by phylum. In particular, Proteobacteria (red) tend to cluster in distinct manifolds from the rest of the taxa. However, most of the taxa lie in a single large but stratified manifold of mixed phyla. In contrast, the contextualized representations in panel (b) appear to have more consistent clustering by phylum in this reduced 2-D space. We further verify that the contextualized embedding does cluster more strongly in the full-dimensional embedding spaces with Supporting Information Fig. 9, which shows that K-means clustering in the contextualized embedding space achieves significantly higher phylum purity than clustering in the original vocabulary embedding space.

To highlight the differences between the two representations, Figure 5 explores the mapping between them by highlighting the same group of taxa in both figures, where the left column shows the t-SNE visualization of the original taxa embeddings, and the right column shows the t-SNE of the contextualized taxa embeddings. From the comparison, we can see that the phyla that are well separated in the original embedding space as distinct manifolds are well preserved and further compacted into tighter clusters (see Fig. 5 a).

The data in the original vocabulary embedding space appear to lie on long “strands”, rather than clump together in clusters. In particular, we see a large strand in the middle that contains most of the data, and seems to be made up of smaller “threads”

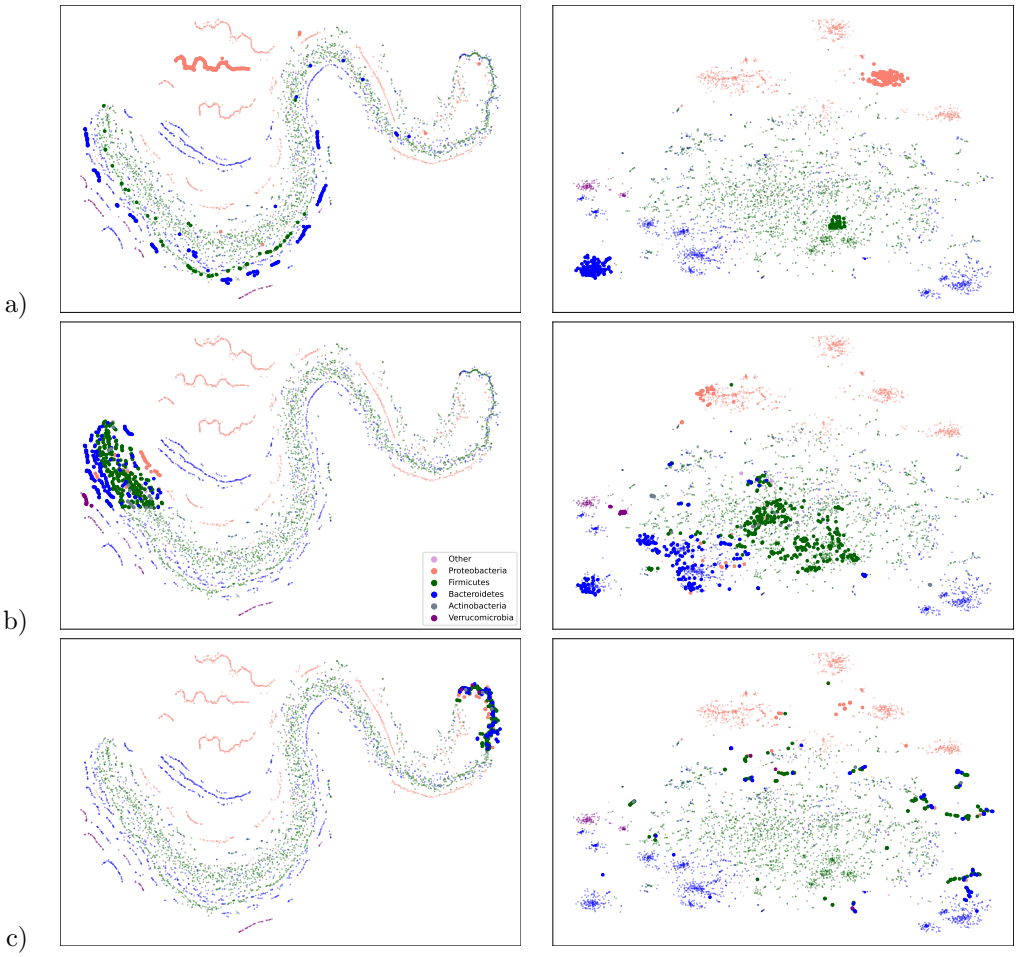


Fig 5. Mapping between the original vocabulary and contextualized embedding spaces. Figure a) shows how the contextualized embeddings can extract "threads" of a single phylum from the vocabulary embedding space, and map those taxa to tight clusters in the contextualized embeddings. Figure b) shows that the mapping to the contextual embedding space is able to more cleanly separate taxa by phylum. Figure c) contrasts Figure b) and shows that taxa which are very tightly clustered in the vocabulary embeddings may not map to meaningful clusters or phylum-level separation in the contextualized embedding space.

very close together. The contextual representations appear to "unwind" the large strand so that the smaller threads can be extracted and grouped together in their own isolated clusters, which more cleanly separate by phylum (see Fig. 5 a). This highlights the capability of self-supervised representation learning to flexibly extract important features from unlabeled data.

Our model's ability to cluster taxa by phylum seems to degrade for taxa whose vocabulary embeddings are too close together. Figure 5 b) highlights taxa in a less compact region of the original embedding space, and highlights the same taxa in the contextualized embedding space, where the taxa show reasonable separation by phylum. In comparison, Figure 5 c) highlights a more compact region of the vocabulary space, as well as the corresponding taxa in the contextualized embedding space, which appear to show worse separation than we see in Figure 5 b).

3.4 Understanding taxa importance for IBD prediction

In this part, we focus on the fine-tuned IBD ensemble prediction model to understand what taxa play critical roles in our model's IBD prediction by studying their attribution. We first consider the 5000 most frequent taxa shown in Figure 4 and compute for each taxa its average attribution toward the model's IBD prediction using the AGP IBD data, as described in Sec. 2.3.

Figure 6 (a) presents the t-SNE visualizations of the contextualized embeddings colored by taxa attribution strength. We can see one cluster of highly attributing taxa appearing close to the center of the figure, another at the bottom left corner. It is important to note that the contextualized embeddings generated by our pre-trained language model have never been trained on any IBD labels, yet these highly attributing taxa exhibit strong clustering behavior, suggesting that our pre-trained language model indeed captures meaningful biological information.

Next, we wish to find the most important taxa for our model's correct IBD classifications across different study populations. We therefore filter the data to focus on samples for which our model makes confident and correct predictions. Specifically, we filter each of the three IBD datasets (American Gut Project (AGP) [8], the curated [13] versions of the Human Microbiome Project phase 2, (HMP2) [25], and

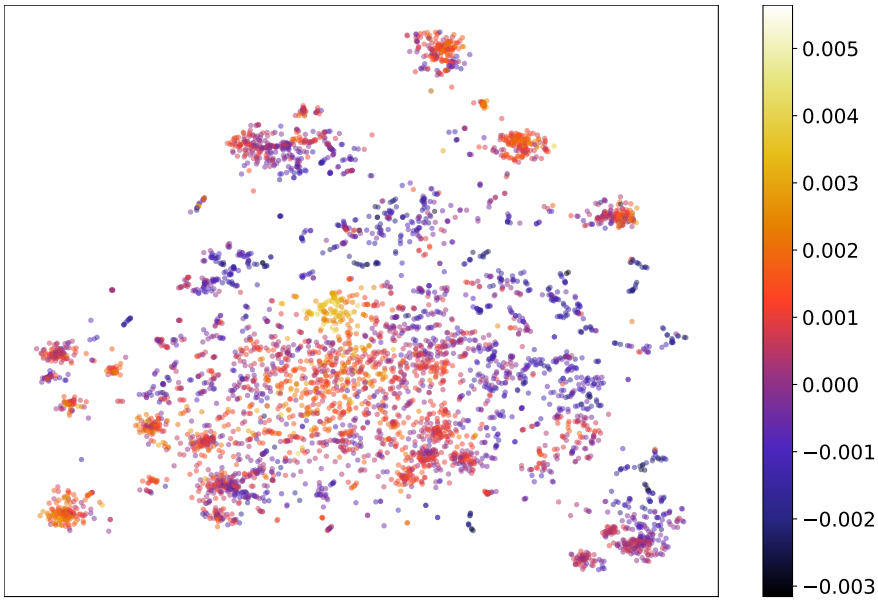


Fig 6. t-SNE visualization of the contextualized embeddings colored by attribution to IBD. The taxa associated with IBD are visualized in lighter color (yellow) and the taxa associated with no-disease state are in dark purple.

Halfvarson (HV) [24]) and include only correctly classified samples with a predicted probability ranking within the top 50%, regardless of being positive or negative. To focus on reasonably common microbial species, we also filter out taxa that appear in less than 5% of all samples across all three IBD datasets (AGP, HMP2 and HV).

To allow for independent validation of our attribution estimation, we combine HV and HMP2, into a single dataset (HV+HMP2), and compute the average attribution on APG and HV+HMP2 separately, and reduce noise by filtering out any taxa that appear in less than five samples in each dataset. The attribution for a taxon is considered validated if it has two estimates from AGP and HV+HMP2 respectively, and they have the same sign. This ensures that we only identify these microbial species that have consistent impact on the model in two different populations. We then compute the average attribution of all validated taxa across the combined (filtered) datasets. We show the 10 taxa most attributed to negative IBD classification (Table 4) and the 10 taxa most attributed to positive IBD classification (Table 5).

General observations

Overall, taxa positively and negatively associated with IBD show phylogenetic consistency. For example, the genera *Prevotella* was identified as a taxa negatively

Table 4. Top 10 Taxa associated with negative (non-disease) IBD classification ordered by attribution strength.

Phylum	Class	Order	Family	Genus
Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella
Bacteroidetes	Bacteroidia	NA	NA	NA
Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus
Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Allisonella
Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Roseburia
Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcus_1
Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Paraprevotella
Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella_9
Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnoclostridium

Table 5. Top ten Taxa associated with positive IBD classification ordered by attribution strength.

Phylum	Class	Order	Family	Genus
Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	NA
Euryarchaeota	Methanobacteria	Methanobacteriales	Methanobacteriaceae	Methanosphaera
Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	NA
Firmicutes	Clostridia	Clostridiales	NA	NA
Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Oscillibacter
Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	NA
Actinobacteria	Coriobacteriia	Coriobacteriales	Atopobiaceae	Atopobium
Bacteroidetes	Bacteroidia	Bacteroidales	Tannerellaceae	Parabacteroides
Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	NA
Firmicutes	Clostridia	Clostridiales	NA	NA

associated with IBD, alongside Paraprevotella, a genera closely phylogenetically related and belonging to the same family. We also observed that five of the top ten ASVs positively associated with IBD were identified as belonging to the family Ruminococcaceae (Table 4). Additionally, we observe no overlap at the genus level between taxa identified as positively or or negatively associated with IBD. At the family level, there is only one overlap: Ruminococcaceae appears in both groups.

Positively associated taxa include classic IBD biomarkers

Our analysis revealed several key players in IBD. ASVs, belonging to distinct bacteria, have already been associated in the literature with bowel inflammation pathways. In the genera *Atopibium*, the species *A. parvulum* has been shown to induce pancolitis in colitis-susceptible interleukin-10-deficient mice. It is hypothesized to play a key role in controlling H2S producers and interact with mitochondrial proteins implicated in H2S detoxification, which are downregulated in Crohn's Disease patients [28].

The Parabacteroides family has been reported to be a non pathogenic commensal

gut microorganism [29], however, evidence indicates that *P. distasonis* could contribute to IBD [30]. The bacteria *Parabacteroides*, more specifically the species *P. distasonis*, has been directly isolated from the gut wall of a patient who underwent surgical removal of the ileum as treatment for CD [31].

Exploration of the role of archaea in IBD has recently intensified, and *Methanosphaera* has been identified early on as a strong inducer of the inflammatory response with a strong immunogenic potential [32].

Taxa with major role in gut homeostasis were negatively associated with IBD

The *Lachnoclostridium* genus contains many butyrate producers, and, as such, plays a major role in the gut homeostasis [33]. But it has also been found as a biomarker of several diseases, such as atherosclerosis [34] or colorectal cancer [35].

The *Roseburia* genus has been shown in general to be beneficial for gut health [36], plays a role in homeostasis [36], and to regulate the immune response (for example *R. intestinalis* ameliorates colitis induced in mice [37]).

Reduced incidence of *Bacteroides* genus has been associated with UC [38]. For example, six *Bacteroides* species exhibited significantly lower relative abundance in patients with UC [39], and the scientific community speculates that the loss of key *Bacteroides* species may exacerbate UC activity, serving as potential biomarkers [38].

The Veillonellaceae family, including the *Allisonella* genus, is known as a group of propionate-producing bacteria that utilize lactate as a substrate [38], and the *Allisonella* genus also an histamine-producing genus [40]. It is worth noting as this compound is pro-inflammatory, but, no study has yet reported this genus associated with IBD. There have been, however, mixed reports in the literature regarding its association with Irritable Bowel Syndrome (IBS) (a set of symptoms similar to IBD such as chronic diarrhea, constipation, or both, without intestinal wall destruction). This genus has been found in a study associated with IBS when mixed symptoms appeared [41], but otherwise decreased in other IBS subtypes [42, 43].

The genus *Lactobacillus* presents mixed evidence as of its contribution to IBD. On one hand, IBD patients seem to show an increase in *Lactobacillus* [44]. On the other hand, *Lactobacillus* has been shown to be effective in reducing inflammation in animal

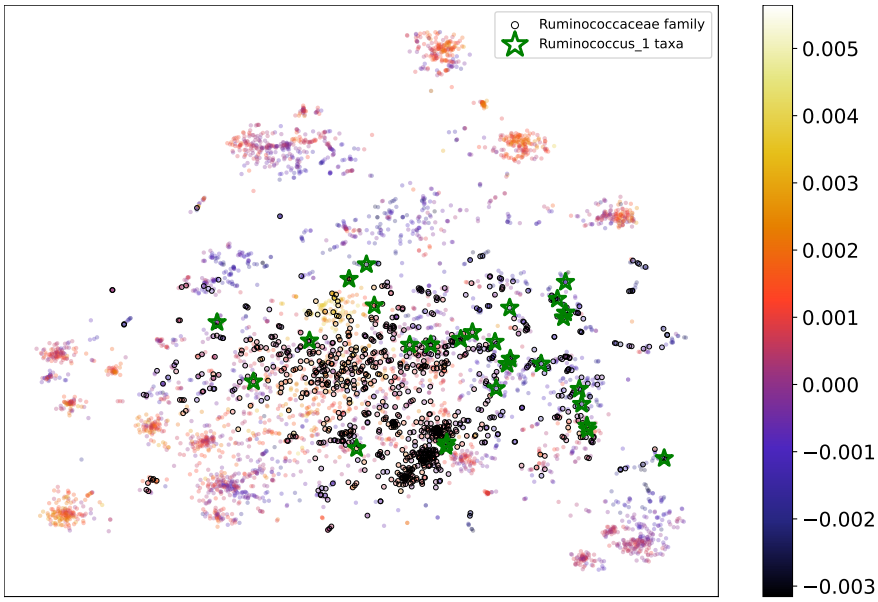


Fig 7. t-SNE visualization of the contextualized embeddings colored by attribution to IBD. Black circles highlight members of the Ruminococcaceae family, and green stars indicate members of the *Ruminococcus_1* genus.

models [45], and is even under consideration as a potential therapeutics for IBD [46]. The mechanisms of action of *Lactobacillus* species on IBD are complex and not well understood, and it is hypothesized that *L. plantarum* modulates the intestinal microbial composition by for example suppressing pathogens [47].

Contextualized embeddings unravel the complex association of Ruminococcaceae, *Bacteroides* and Prevotellaceae families in IBD.

The family Ruminococcaceae is present in both sets positively and negatively associated with IBD. Most of the ASVs associated with this family could not be assigned a taxonomy at the genus level—it is difficult to conduct a detailed bibliographic research on their possible implication with IBD. Exploring their position in the representation of the contextualized embeddings 7 shows that many Ruminococcaceae's ASVs were widespread throughout the plot. However, *Ruminococcus_1*, the only Ruminococcaceae present in the group negatively associated with ASD, was not present in the clusters strongly associated with IBD on the plot (in yellow in Figures 6 and 7).

The same observation can be done with ASVs belonging to the Prevotellaceae family: they seem to group into an island outside the main IBD related cluster (Fig 8), and while historically associated with inflammation susceptibility in animal studies [48],

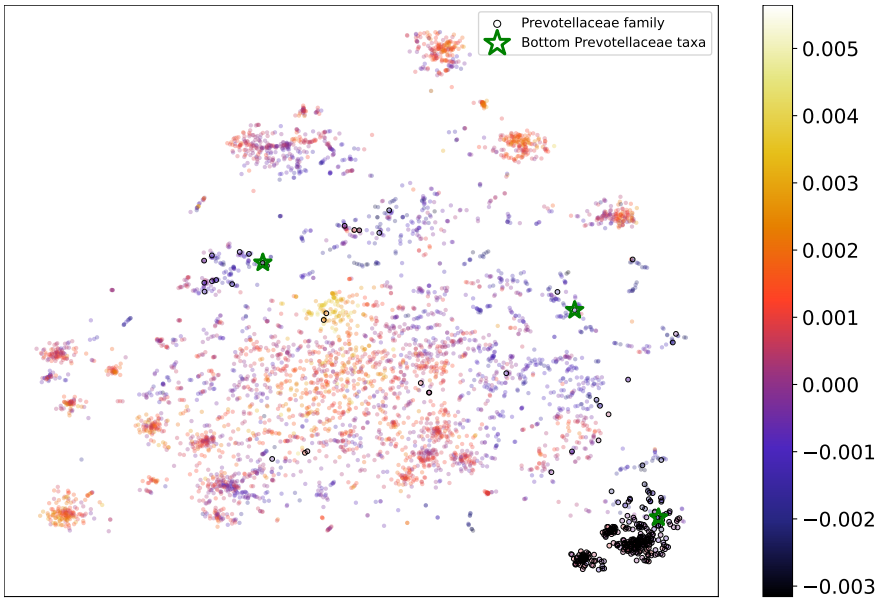


Fig 8. t-SNE visualization of the contextualized embeddings colored by attribution to IBD. Black circles highlight members of the Prevotellaceae family, and green stars indicate the three taxa of the Prevotellaceae family present in Table 4.

our analysis does not show them as the key ASV in our model's prediction of IBD.

A closer inspection of the (center yellow) cluster of ASVs in Fig. 6 with the most positive attribution to our model's IBD prediction revealed that it does not have a strong phylogeny homogeneity (see Figure 4.b). Rather, it contains taxa belonging to a total of 21 families. Our results suggest that phylogenetic assignment of microbial biomarkers of IBD (here our ASVs) may not be sufficient, as ASVs belong to the same genus appears on different part of our t-SNE embedding representation. Considering the context of other ASVs, and potentially their interactions (between them or with the host, indirectly or directly, through metabolic network or ligand-receptor interactions) is important in deciphering the etiology of IBD.

4 Conclusion

We apply recent natural language processing techniques to learn a language model for microbiomes from public domain human gut microbiome data. The pre-trained language model provides powerful contextualized representations of microbial communities and can be broadly applied as a starting point for any downstream prediction tasks involving human gut microbiome. In this work, we show the power of

the pre-trained model by fine-tuning the representations for IBD disease state and diet classification tasks, achieving strong performance in all tasks. For IBD, our ensemble model demonstrates competitive performance that is robust across study populations even with strong distributional shifts.

We visualize the contextualized taxa embedding learned by our pre-trained language model and show that it captures biologically meaningful information including phylogenetic structure and IBD association without any prior training on such signals. We employ an interpretability technique to investigate the basis for our models' IBD classification decisions and identify sets of taxa that negatively and positively attribute to the model's predictions. We find known biomarkers of both IBD and gut homeostasis, as well as evidence that our embeddings learn to separate ASVs by their pathogenicity, even among ASVs sharing the same family and genus level phylogenetic classifications.

Our investigation suggests that NLP techniques like deep language models represent a promising direction to better understand the microbiome. However, our effort is limited in both volume and breath of the data that is used for training the microbial language model. A direct consequence of this limitation is that our model can only be applied to tasks involving human gut microbiomes. Future work could extend this line of research to include data beyond simply taxa assignments computed from the 16S data of the V4 region, e.g., by using direct nucleotide sequences as inputs to the language model. This would vastly expand the scope of data usable for training such a model, including the content of massive genomics data repositories such as the Sequence Read Archive [49].

Supporting information

Quality of the embeddings in high dimensional space

Figure 4 shows 2-D projections of the vocabulary embedding space learned by [13] using GloVe [50] (a) and of the contextualized embedding space generated by our method (b). Visual inspection suggests that our method show stronger clustering of taxa by phylum. We confirm that this holds true in the full-dimensional embedding spaces of the two models by applying K-means clustering in both embedding spaces for K ranging from 1

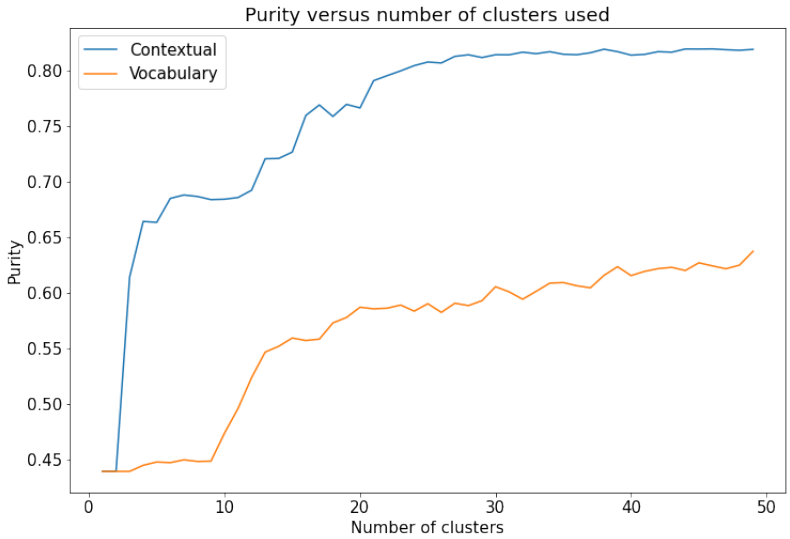


Fig 9. Cluster purity versus K for K-means clustering in the vocabulary and contextualized embedding spaces.

to 50. Figure 9 shows that clusters in the contextualized embedding space consistently have less cross-phylum contamination (higher phylum purity) as compared to clusters in the GloVe embedding space.

Acknowledgments

We thank the National Science Foundation for the funding of this work under grant number URoL:MTM2 2025457.

References

1. Kostic AD, Chun E, Robertson L, Glickman JN, Gallini CA, Michaud M, et al. Fusobacterium nucleatum Potentiates Intestinal Tumorigenesis and Modulates the Tumor-Immune Microenvironment. Cell Host & Microbe. 2013;14(2):207–215. doi:https://doi.org/10.1016/j.chom.2013.07.007.

2. Jiang H, Ling Z, Zhang Y, Mao H, Ma Z, Yin Y, et al. Altered fecal microbiota composition in patients with major depressive disorder. Brain, Behavior, and Immunity. 2015;48:186–194. doi:https://doi.org/10.1016/j.bbi.2015.03.016.

3. Zheng P, Zeng B, Zhou C, Liu M, Fang Z, Xu X, et al. Gut microbiome remodeling induces depressive-like behaviors through a pathway mediated by the host's metabolism. Molecular Psychiatry. 2016;21(6):786–796. doi:10.1038/mp.2016.44.

4. Frank DN, St Amand AL, Feldman RA, Boedeker EC, Harpaz N, Pace NR. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. Proceedings of the National Academy of Sciences. 2007;104(34):13780–13785. doi:10.1073/pnas.0706625104.

5. Gevers D, Kugathasan S, Denson L, Vázquez-Baeza Y, Van Treuren W, Ren B, et al. The Treatment-Naive Microbiome in New-Onset Crohn's Disease. Cell Host & Microbe. 2014;15(3):382–392. doi:10.1016/j.chom.2014.02.005.

6. Ni J, Shen TCD, Chen EZ, Bittinger K, Bailey A, Roggiani M, et al. A role for bacterial urease in gut dysbiosis and Crohn's disease. Science Translational Medicine. 2017;9(416):eaah6888. doi:10.1126/scitranslmed.aah6888.

7. Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. Current understanding of the human microbiome. Nature Medicine. 2018;24(4):392–400. doi:10.1038/nm.4517.

8. McDonald D, Hyde E, Debelius JW, Morton JT, Gonzalez A, Ackermann G, et al. American Gut: an Open Platform for Citizen Science Microbiome Research. mSystems. 2018;3(3). doi:10.1128/mSystems.00031-18.

9. McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PloS one. 2013;8(4):e61217.

10. Brooks AW, Priya S, Blekhman R, Bordenstein SR. Gut microbiota diversity across ethnicities in the United States. PLoS biology. 2018;16(12):e2006842.

11. Washburne AD, Morton JT, Sanders J, McDonald D, Zhu Q, Oliverio AM, et al. Methods for phylogenetic analysis of microbiome data. Nature microbiology. 2018;3(6):652–661.

12. Woloszynek S, Zhao Z, Chen J, Rosen GL. 16S rRNA sequence embeddings: Meaningful numeric feature representations of nucleotide sequences that are

convenient for downstream analyses. PLoS computational biology. 2019;15(2):e1006721.

13. Tataru CA, David MM. Decoding the language of microbiomes using word-embedding techniques, and applications in inflammatory bowel disease. PLOS Computational Biology. 2020;16(5):e1007859. doi:10.1371/journal.pcbi.1007859.

14. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. arXiv:170603762 [cs]. 2017;.

15. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, et al. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2022;44(10):7112–7127. doi:10.1109/tpami.2021.3095381.

16. Heinzinger M, Elnaggar A, Wang Y, Dallago C, Nechaev D, Matthes F, et al. Modeling aspects of the language of life through transfer-learning protein sequences. BMC Bioinformatics. 2019;20(1). doi:10.1186/s12859-019-3220-8.

17. Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M. ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics. 2022;38(8):2102–2110. doi:10.1093/bioinformatics/btac020.

18. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences. 2021;118(15):e2016239118. doi:10.1073/pnas.2016239118.

19. Ji Y, Zhou Z, Liu H, Davuluri RV. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics. 2021;37(15):2112–2120. doi:10.1093/bioinformatics/btab083.

20. Ericsson L, Gouk H, Loy CC, Hospedales TM. Self-Supervised Representation Learning: Introduction, advances, and challenges. IEEE Signal Processing Magazine. 2022;39(3):42–62. doi:10.1109/MSP.2021.3134634.

21. Clark K, Luong MT, Le QV, Manning CD. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. arXiv:200310555 [cs]. 2020;.

22. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:181004805 [cs]. 2019;.

23. Zeiler MD, Fergus R. Visualizing and Understanding Convolutional Networks. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, editors. Computer Vision – ECCV 2014. Cham: Springer International Publishing; 2014. p. 818–833.

24. Halfvarson J, Brislawn CJ, Lamendella R, Vázquez-Baeza Y, Walters WA, Bramer LM, et al. Dynamics of the human gut microbiome in inflammatory bowel disease. Nature Microbiology. 2017;2(5):17004. doi:10.1038/nmicrobiol.2017.4.

25. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature. 2019;569(7758):655–662. doi:10.1038/s41586-019-1237-9.

26. van der Maaten L, Hinton G. Visualizing Data using t-SNE. Journal of Machine Learning Research. 2008;9(86):2579–2605.

27. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nature methods. 2016;13(7):581—583. doi:10.1038/nmeth.3869.

28. Mottawea W, Chiang CK, Mühlbauer M, Starr AE, Butcher J, Abujamel T, et al. Altered intestinal microbiota–host mitochondria crosstalk in new onset Crohn's disease. Nature Communications. 2016;7(1):13419. doi:10.1038/ncomms13419.

29. Dziarski R, Park SY, Kashyap DR, Dowd SE, Gupta D. Pglyrp-Regulated Gut Microflora Prevotella falsenii, Parabacteroides distasonis and Bacteroides eggerthii Enhance and Alistipes finegoldii Attenuates Colitis in Mice. PLOS ONE. 2016;11(1):e0146162. doi:10.1371/journal.pone.0146162.

30. Lopetuso LR, Petito V, Graziani C, Schiavoni E, Sterbini FP, Poscia A, et al. Gut Microbiota in Health, Diverticular Disease, Irritable Bowel Syndrome, and

Inflammatory Bowel Diseases: Time for Microbial Marker of Gastrointestinal Disorders. Digestive Diseases. 2017;36(1):56–65. doi:10.1159/000477205.

31. Yang F, Kumar A, Davenport KW, Kelliher JM, Ezeji JC, Good CE, et al. Complete Genome Sequence of a Parabacteroides distasonis Strain (CavFT hAR46) Isolated from a Gut Wall-Cavitating Microlesion in a Patient with Severe Crohn's Disease. Microbiology Resource Announcements. 2019;8(36). doi:10.1128/mra.00585-19.

32. Lecours PB, Marsolais D, Cormier Y, Berberi M, Haché C, Bourdages R, et al. Increased Prevalence of Methanosphaera stadtmanae in Inflammatory Bowel Diseases. PLoS ONE. 2014;9(2):e87734. doi:10.1371/journal.pone.0087734.

33. Dandachi I, Anani H, Hadjadj L, Brahimi S, Lagier JC, Daoud Z, et al. Genome analysis of Lachnoclostridium phocaeense isolated from a patient after kidney transplantation in Marseille. New Microbes and New Infections. 2021;41:100863. doi:10.1016/j.nmni.2021.100863.

34. Cai YY, Huang FQ, Lao X, Lu Y, Gao X, Alolga RN, et al. Integrated metagenomics identifies a crucial role for trimethylamine-producing Lachnoclostridium in promoting atherosclerosis. npj Biofilms and Microbiomes. 2022;8(1). doi:10.1038/s41522-022-00273-4.

35. Liang JQ, Li T, Nakatsu G, Chen YX, Yau TO, Chu E, et al. A novel faecal iLachnoclostridium/i marker for the non-invasive diagnosis of colorectal adenoma and cancer. Gut. 2019;69(7):1248–1257. doi:10.1136/gutjnl-2019-318532.

36. Nie K, Ma K, Luo W, Shen Z, Yang Z, Xiao M, et al. Roseburia intestinalis: A Beneficial Gut Organism From the Discoveries in Genus and Species. Frontiers in Cellular and Infection Microbiology. 2021;11. doi:10.3389/fcimb.2021.757718.

37. Luo W, Shen Z, Deng M, Li X, Tan B, Xiao M, et al. Roseburia[U+FFFD]intestinalis supernatant ameliorates colitis induced in mice by regulating the immune response. Molecular Medicine Reports. 2019;doi:10.3892/mmr.2019.10327.

38. Nomura K, Ishikawa D, Okahara K, Ito S, Haga K, Takahashi M, et al. Bacteroidetes Species Are Correlated with Disease Activity in Ulcerative Colitis. Journal of Clinical Medicine. 2021;10(8):1749. doi:10.3390/jcm10081749.

39. Zhou Y, Zhi F. Lower Level ofi Bacteroides/i in the Gut Microbiota Is Associated with Inflammatory Bowel Disease: A Meta-Analysis. BioMed Research International. 2016;2016:1–9. doi:10.1155/2016/5828959.

40. Garner MR, Flint JF, Russell JB. Allisonella histaminiformans gen. nov., sp. nov. Systematic and Applied Microbiology. 2002;25(4):498–506. doi:10.1078/07232020260517625.

41. Schmulson M, Bielsa MV, Carmona-Sánchez R, Hernández A, López-Colombo A, Vidal YL, et al. Microbiota, gastrointestinal infections, low-grade inflammation, and antibiotic therapy in irritable bowel syndrome (IBS): an evidence-based review. Revista de Gastroenterología de México (English Edition). 2014;79(2):96–134. doi:10.1016/j.rgmxen.2014.01.001.

42. Kassinen A, Krogius-Kurikka L, Mäkivuokko H, Rinttilä T, Paulin L, Corander J, et al. The Fecal Microbiota of Irritable Bowel Syndrome Patients Differs Significantly From That of Healthy Subjects. Gastroenterology. 2007;133(1):24–33. doi:10.1053/j.gastro.2007.04.005.

43. Hills R, Pontefract B, Mishcon H, Black C, Sutton S, Theberge C. Gut Microbiome: Profound Implications for Diet and Disease. Nutrients. 2019;11(7):1613. doi:10.3390/nu11071613.

44. Ghosh TS, Arnoux J, O'Toole PW. Metagenomic analysis reveals distinct patterns of gut lactobacillus prevalence, abundance, and geographical variation in health and disease. Gut Microbes. 2020;12(1):1822729. doi:10.1080/19490976.2020.1822729.

45. Madsen KL, Doyle JS, Jewell LD, Tavernini MM, Fedorak RN. Lactobacillus species prevents colitis in interleukin 10 gene–deficient mice. Gastroenterology. 1999;116(5):1107–1114. doi:10.1016/s0016-5085(99)70013-2.

46. Wang H, Zhou C, Huang J, Kuai X, Shao X. The potential therapeutic role of lt;igt;Lactobacillus reuterilt;/igt; for treatment of inflammatory bowel disease. American Journal of Translational Research. 2020;12(5):1569—1583.

47. Arena MP, Silvain A, Normanno G, Grieco F, Drider D, Spano G, et al. Use of Lactobacillus plantarum Strains as a Bio-Control Strategy against Food-Borne Pathogenic Microorganisms. Frontiers in Microbiology. 2016;7. doi:10.3389/fmicb.2016.00464.

48. Iljazovic A, Roy U, Gálvez EJC, Lesker TR, Zhao B, Gronow A, et al. Perturbation of the gut microbiome by Prevotella spp. enhances host susceptibility to mucosal inflammation. Mucosal Immunology. 2021;14(1):113–124. doi:10.1038/s41385-020-0296-4.

49. Leinonen R, Sugawara H, Shumway M. The Sequence Read Archive. Nucleic Acids Research. 2010;39:D19 – D21.

50. Pennington J, Socher R, Manning C. Glove: Global Vectors for Word Representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics; 2014. p. 1532–1543. Available from: `http://aclweb.org/anthology/D14-1162`.