ML4FG Report: Using Transfer Learning to Enhance Microbiome Analysis

George Austin, gia2105@columbia.edu

April 2021

1 Abstract

The human gut microbiome is an important factor contributing to many diseases' developments. However, it is generally challenging to apply deep-learning structures to microbial datasets, as these datasets are typically high-dimensional and sparse, with small amounts of samples. We alleviate this issue by creating a flexible framework to transfer information across different studies and prediction tasks, thus allowing neural networks to capture more signals for a given task. By applying this framework across a collection of six datasets, we outperform standard modelling approaches, with larger improvements when the target task's dataset is small.

2 Introduction

Analyses of human microbiomes have improved our understanding of many diseases, including Inflammatory Bowel Disease, Type 2 Diabetes, and colorectal cancer (1). However, there remains a challenge in applying deep neural networks to this kind of data, which is typically high dimensional with a small number of datapoints. This means that any network needs lots of parameters, but can rarely have enough data to extract signals without overfitting. While it has been suggested that as more datasets become publicly available, there will be an opportunity to alleviate this issue by applying transfer learning methods (2), there are very few examples to date.

In this work, we explore the impact of transfer learning techniques such as hard parameter sharing (3) and Model Agnostic Meta-Learning (MAML) (4) on publicly available microbial datasets. This is done by extending the results of DeepMicro(5), a study applying various deep learning methods to 6 different collections of fecal gut samples, demontrating that signals can be improved by using autoencoders to reduce the datasets' dimension. We apply transfer learning on feedforward neural networks, in addition to their autoencoder structures, to determine if transfer learning techniques can lead to measurable improvements in predictive performance.

3 Methods

3.1 Data Summary

In total, we are working with six datasets, spanning five different prediction tasks: inflammatory bowel disease (6), type 2 diabetes(7)(8), obesity(9), liver cirrhosis(10), and colorectal cancer(11). The datasets are from six distinct studies, all of which provide publicly available data. All datasets were processed during the DeepMicro (5) study using MetaPhlAn2 (12) and MetAML (13) to create species abundance representations of the data, which have relative abundance values for 771 bacterial species. All six datasets are summarized in Figure 1.

Figure 1: Dataset Summaries [1]

	Nationality	Year	Positive Samples	Negative Controls
IBD	European	2010	25	85
EW-T2D	European Women	2013	53	43
C-T2D	Chinese	2012	170	174
Obesity	Danish	2013	164	89
Cirrhosis	Chinese	2014	118	114
Colorectal	French	2014	74	47

[1] Our Colorectal positive/negative numbers are different from what DeepMicro uses. This is because we classify our adenomas as positive, as these detections might warrant a colonoscopy, even if they are not cancerous.

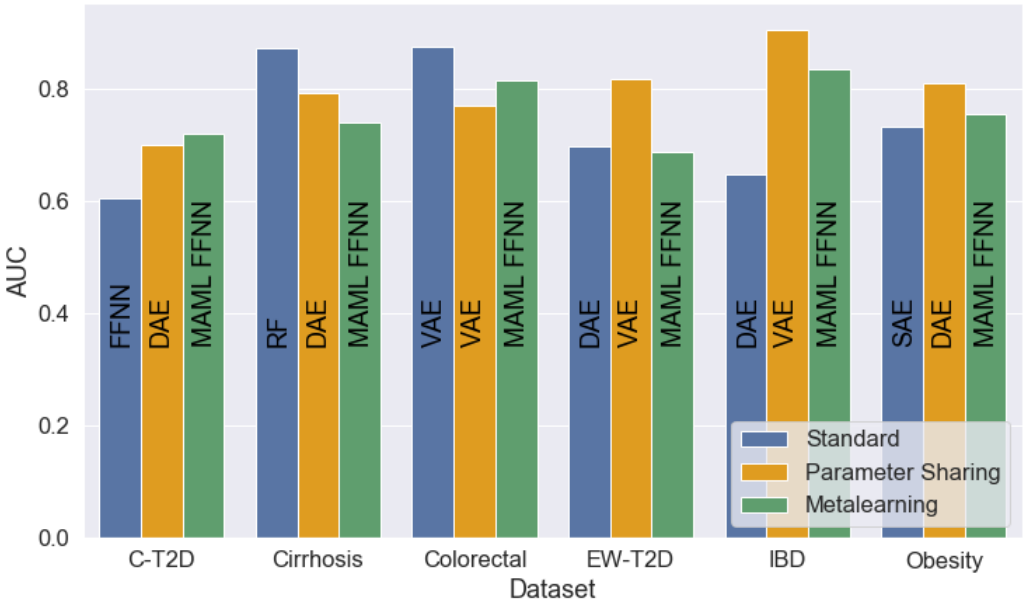
3.2 Hyperparameter Tuning

We trained each modelling approach for each prediction task. The baseline models for this analysis are random forest (RF), support vector machines (SVM), feedforward neural network (FFNN), short autoencoder (SAE), deep autoencoder (DAE), and variational autoencoder (VAE). For each model-dataset pair, we ran 15 iterations of Ax tuning (14)) across a predefined hyperparameter seach space (Supplementary Figure S1), which mimicked those used in DeepMicro, when applicable. For the autoencoders, either random forest or support vector machines were tuned using 5-fold cross validation on the encoded train and validation sets (Supplementary Figure S2). Once each model's 15 iterations were finished, their best observed AUC and hyperparameter space were recorded. The best-performing baseline models and corresponding AUC for each structure are illustrated in Supplementary Figure S4.

We trained parameter sharing autoencoders using a scheme similar to the baseline autoencoders' approach (using SAEs, DAEs, and VAEs). The only difference is that the autoencoders were trained using a train – validation split encompassing multiple datasets, so our encodings could capture patterns spanning multiple training tasks. In addition, we wanted to consider different possible groups of datasets that might be useful for transfer learning. To explore this, we ran tuning for each encoder using the different groups of datasets outlined in Supplementary Figure S5. Some groups, such as EU and CH, were used as a way to mitigate the impact that batch effects (illustrated in Supplementary Figure S3) might have on our performance. Other groups, such as T2D and Colon, were attempts to consider only tasks that were similar, to potentially facilitate the transferal of relevant information. We also considered grouping all datasets together, to determine if we could potentially get further improvements with a brute force approach of using all available data.

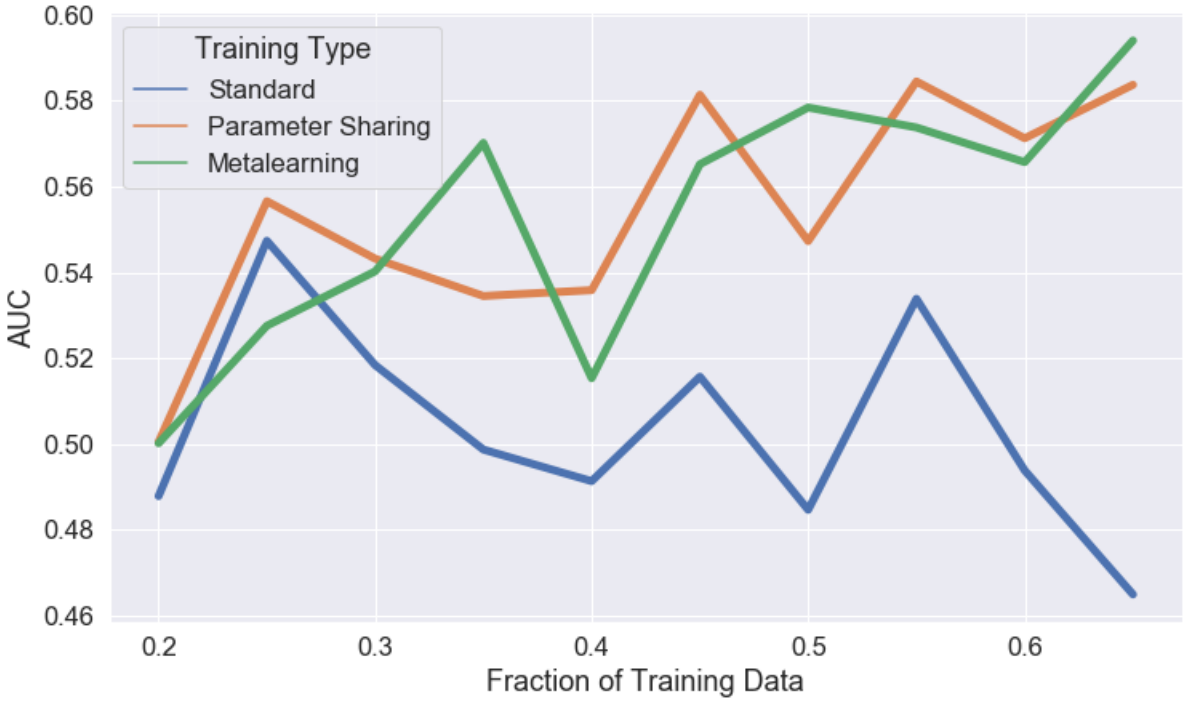
Using the different groups of datasets, we also meta-trained FFNNs using MAML (4)(17), followed by standard SGD. During the metalearning stage, we trained to improve our k-shot learning performance by distinguishing between two randomly selected groups of positive or negative samples, from randomly selected datasets. The comparison of MAML FFNN to standard FFNN's results is shown in supplementary figure S6. The overall best observed model and AUC for each training type (Standard, Parameter Sharing, and Metalearning) are displayed in Figure 2.

Figure 2: Best Observed Model and AUC for Each Prediction Task



Displaying the best model and AUC observed during our tuning process. In 3 out of 6 datasets, the parameter sharing approach lead to the best result, while the metalearning approach outperforms the standard results on half of the datasets.

Figure 3: Prediction Performance on Samples of the Obesity Dataset



Each datapoint illustrates average AUC across five tests for the correspondng fraction of training data used. Both the Parameter Sharing and the Metalearning approaches outperforming the standard learning in this low-data regime ($p < 10^{-5}$ for both comparisons, by wilcox test)

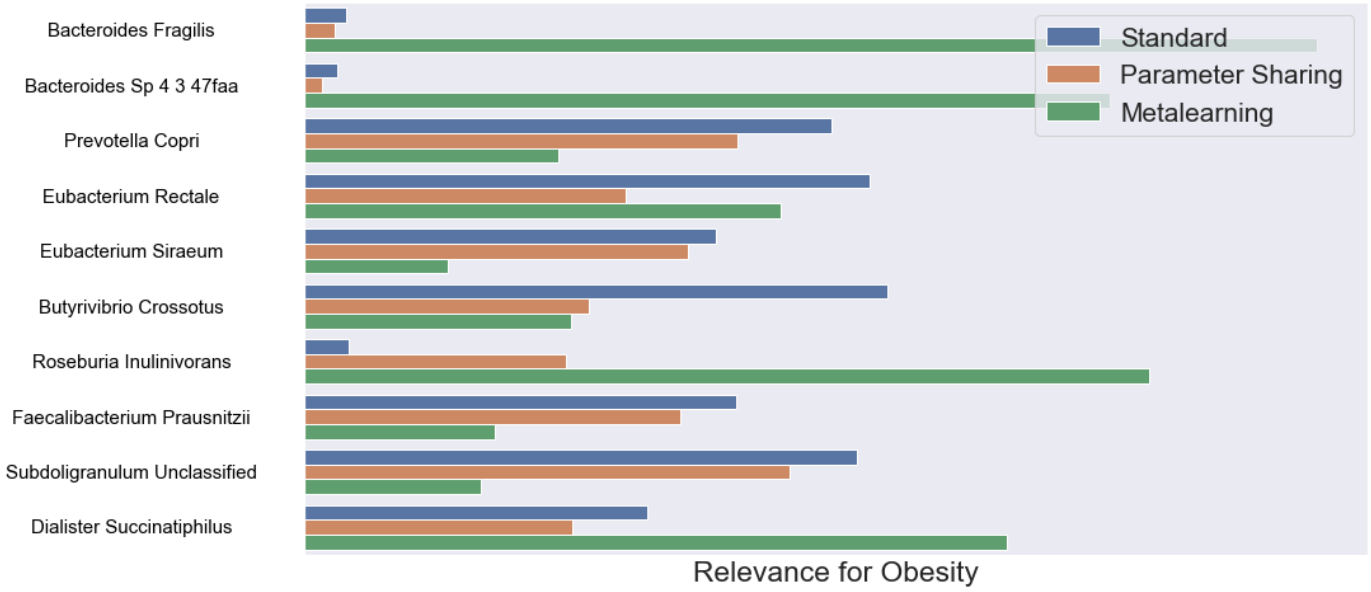
3.3 Testing Reduced Data

To understand how effectively the transfer learning techniques performed for different amounts of samples, we trained using different fractions of data from the Obesity dataset. The reduced data was trained using an 80-20 train-validation split, incorporating full datasets from other tasks when using transfer learning. We did not downsample our test dataset (always a 20% split). Using the best hyperparameters (and transfer learning groups) for each of the three training types, we trained models on five separate train-test splits for each level of downsampling. Figure 3 shows the mean AUCs on the test set for each training type, across different fractions of training data.

3.4 Inference

To understand why we might see improvements from transfer learning, we analyzed different models' most important bacterial species using saliency maps (15)(16). For a given prediction task, we trained the best model from each training type with its best observed hyperparameter space. For the encoder based models, we considered the gradient of the inputs with respect to their encodings. We then took the absolute value of the attribution to each encoded element, and averaged the attributions across elements. The same scheme was utilized for the MAML FFNN's outputs. This was repeated for each point in the test set, averaging the final attributions for each model across each input. We normalized the feature importance estimates so their relative values could be visually compared across model types. The four features with the largest overall influence for each Obesity model are visualized in Figure 4 (an IBD version of this plot is in Supplementary Figure S7).

Figure 4: Visualizing the Most Important Features for Models' Predictions



Showing our Obesity models' most import predictors, via saliency maps. For all cases, we averaged the absolute value of attirbutions across all elements of the models' encoding (or the MAML FFNN's output). These final importance values are scaled for each model, so the relative importances can be visualized on the same plot.

4 Results

Our results suggest that there are potential improvements from applying transfer learning methods. We see the biggest improvement in AUC on the IBD dataset, which is also the dataset with the fewest number of positive samples. In general, the groups of datasets that performed best for the transfer learning were the EU and CH groups. This implies that limiting batch effects are important to successfully perform transfer learning on microbial datasets. Furthermore, limiting to just similar tasks (such as T2D or colon-related diseases) usually did not offer as much predictive power as a more general group.

The down-sampling analysis shows strong improvements when we apply transfer learning to smalller samples of the Obesity dataset ($p < 10^{-5}$ between Parameter Sharing/Standard AUCs, $p < 10^{-5}$ between Metalearning/Standard AUCs, via wilcox test). Parameter sharing and Metalearing models reach AUCs close to 0.6 when the standard model is generally performing as well as a random guess (AUC=0.5). Therefore, our transfer learning allowed models to pick up a predictive signal when a standard encoder and ML model could not.

We see some variation in the most important features for the different models, further suggesting that there are signals being transferred across training tasks when we apply our transfer learning techniques. For example, in Figure 4, neither encoder attributes much importance to Baceteroides Fragilis, but it is one of the metalearning model's top features. This taxon, which has been associated with our other tasks, such as Colorectal Cancer (18), and appears in large amounts within most of our datasets, is a good example of a species whose signal we would hope to see transferred across tasks.

We also see some differences between the two encoders (whose saliency maps are easier to compare than to a FFNN), which further illustrates that the transfer learning is leading to better detection of certain signals. For example, both the parameter sharing encoder and metalearning FFNN assign a greater importance to Roseburia Inulinivorans (known to be decreased in T2D patients (19)) than the standard encoder. This species does appear in varying relative abundance across all our datasets' positive/negative samples (including Obesity), which means

this is a signal that we made easier to capture through our parameter sharing framework. We see similar differences across models in the IBD importance plot, with our transfer learning approaches assigning more importance to Bactereoides Fragilis and Coprococcu, the latter of which has been associated with both colorectal cancer (20) and type 2 diabetes (21). Fragilis has also been linked to both obesity (22) and IBD (23), which indicates that it is a signal we would expect a well-performing model to identify for these training tasks. The fact that we are getting better performance from picking up this known signal through our transfer learning methods is encouraging.
For all code, results and figures, see our github repository: `https://github.com/gaustin15/Microb_Transer_Learning`

5 Discussion

We built a deep learning framework to apply transfer learning methods across different microbial datasets and prediction tasks, both through parameter sharing and model-agnostic meta-learning. Furthermore, we show that our framework outperforms deep learning approaches that don't use transfer learning, particularly in the low-data regime. This low-data regime is highly relevant to the field, as most microbiome datasets fall into this category. While we apply our methods to encoders and feedforward neural networks, it is flexible enough to easily be applied to other deep learning structures, such and CNNs, RNNs, or transformers. There are also many other possible transfer learning methods which can be explored in this field, including soft parameter sharing, ANIL (24), TreeMAML (25), and MetaPred(26), to name a few.

Of course, applying transfer learning methods does come with an additional computational cost, since researchers will need to obtain, process and store more data. It is also important to carefully consider which datasets to use when performing transfer learning, as we show that using more datasetes does not necessarily improve performance. In our case, we observed severe batch effects which, when included in our transfer learning, negatively impacted our results. Some batch effects, which result from contamination, have been shown to be alleviated with methods like decontam (27). We could not apply this method in this study, as contamination samples were not available, but using such a decontamination method could lead to better results in future applications.

While we focus on transferring information across different predictions tasks, many diseases, such as type 2 diabetes and colorectal cancer, have lots of publicly available datasets, suggesting there is less need to transfer information across different tasks if those diseases are the focus (our methods could still be used across multiple datasets spanning the same task). On the other hand, there is a greater value for metalearning and parameter sharing across different tasks when the goal is to analyze a rarer condition, for which there are fewer available datasets. While there are many possible applications for this work within the microbiome field, similar transfer learning frameworks could also be applied to genome and proteome data.

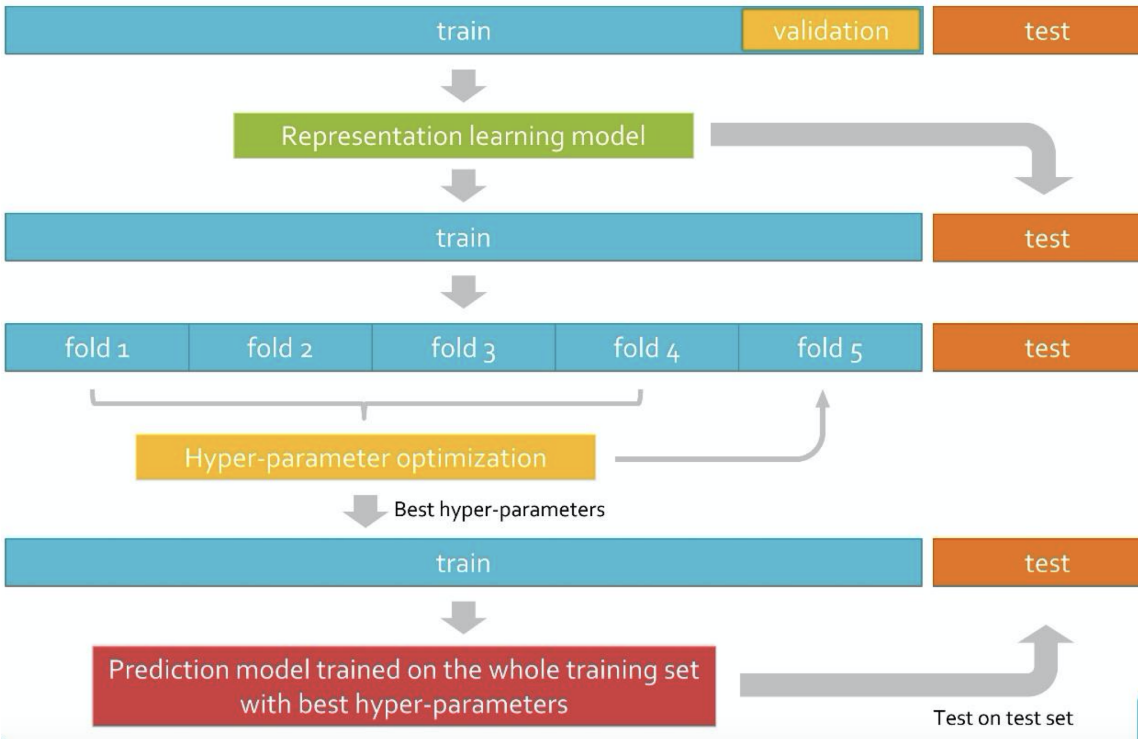
6 Appendix

6.1 Supplementary Figures

Figure S1: Each Model's Hyperparameter Search Space

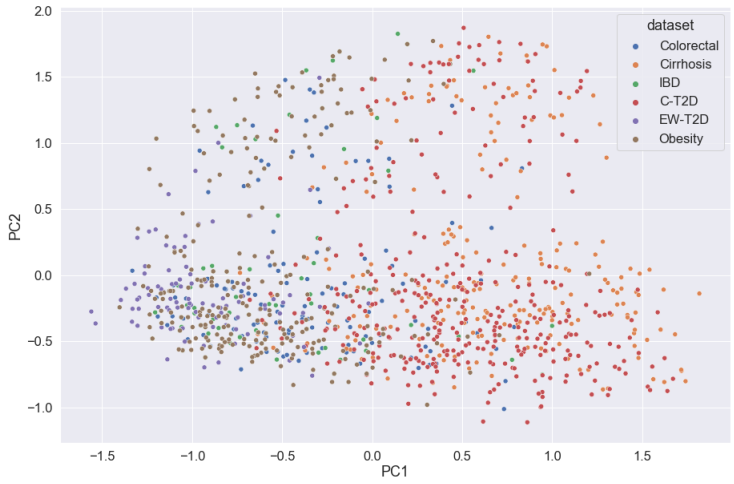
Model Structure	Hyperparameters Tuned	Values Considered
Random Forest	n_estimators	[100, 300, 500, 700, 900]
Random Forest	min_samples_leaf	[1, 2, 3, 4, 5]
Random Forest	criterion	[gini, entropy]
SVM	C	[-5, -3, -1, 1, 3, 5]
SVM	gamma	[-15, -13, -11, -9, -7, -5, -3, -1, 2, 23]
FFNN	layer_1_size	[128, 256, 512, 1024]
FFNN	layer_2_size	[256, 128, 64]
FFNN	layer_3_size	[128, 64, 32]
FFNN	learning_rate	[0.001, 0.01, 0.1]
FFNN	dropout	[0.1, 0.3, 0.5]
DAE	layer_1_size	[64, 128, 256, 512, 1024]
DAE	layer_2_size	[32, 64, 128, 256, 512]
DAE	classifier_model	[svm, rf]
SAE	layer_size	[32, 64, 128, 256, 512]
SAE	classifier_model	[svm, rf]
VAE	layer_1_size	[32, 64, 128, 256, 512]
VAE	layer_2_size	[4, 8, 16]
VAE	classifier_model	[svm, rf]
MAML_FFNN	layer_1_size	[128, 256, 512, 1024]
MAML_FFNN	layer_2_size	[256, 128, 64]
MAML_FFNN	layer_3_size	[128, 64, 32]
MAML_FFNN	learning_rate	[0.0001, 0.1, (Considered the range using baye...
MAML_FFNN	dropout	[0.1, 0.3, 0.5]
MAML_FFNN	adaptation_lr	[0.0001, 0.1, (Considered the range using baye...
MAML_FFNN	k	[1, 2, 4, 8, 16]

Figure S2: The Encoder Evalutaion Scheme



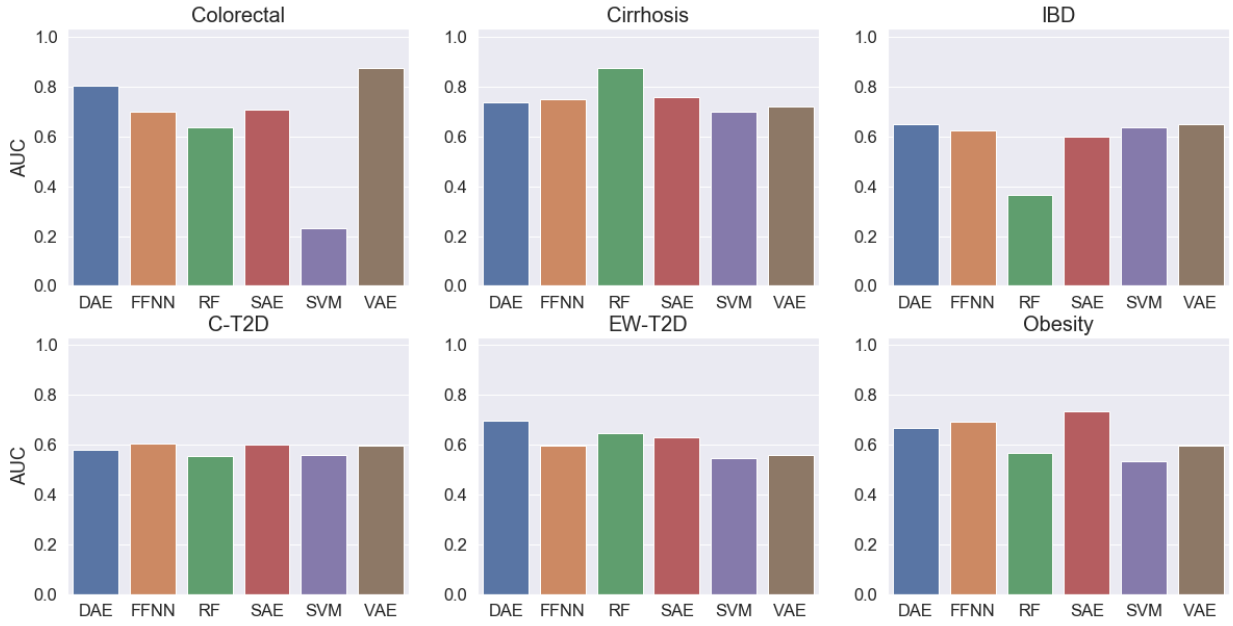
This figure was taken from DeepMicro's Suplementary Materials (5)

Figure S3: PCA of Datasets



We see that the Cirrhosis and the C-T2D datasets are separate from the other four groups along the first principal component. PCOA with Bray-Curtis similarity gives similar results.

Figure S4: Baseline Results



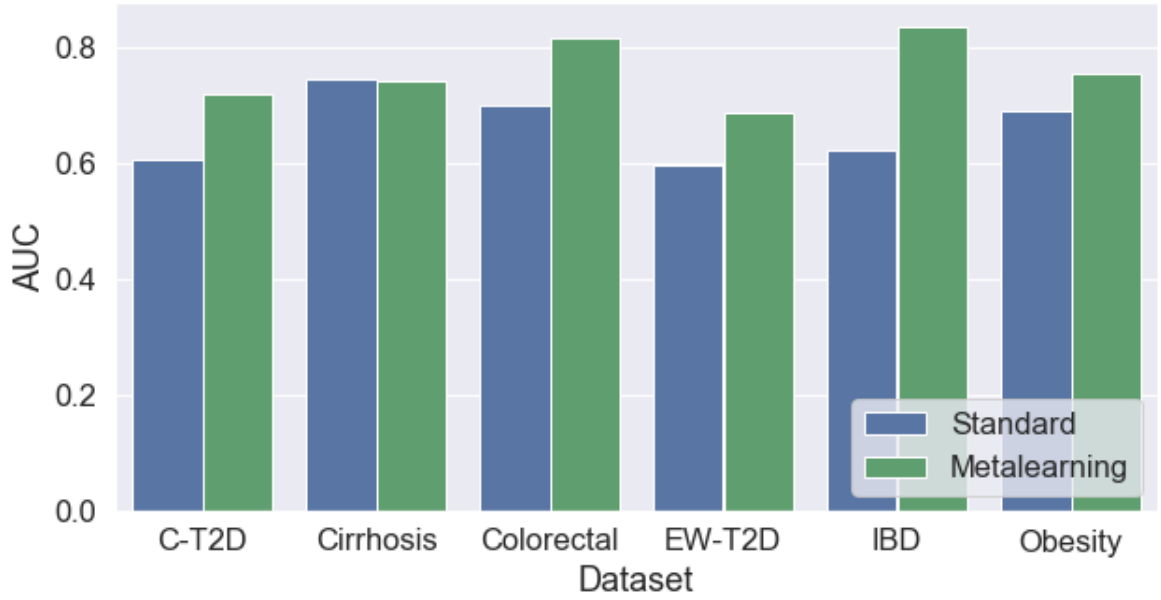
Ths figure shows the results for all baseline models. These metrics are generally within the uncertainty ranges reported by DeepMicro (on their 'Abundance' datasets. In some cases, we adjust our definitions of positive/negative samples (such as defining adenomas as positive), which accounts for some deviations from DeepMicro's results.

Figure S5: Transfer Learning Groups

	Datasets in Group
ALL	['Obesity', 'IBD', 'Colorectal', 'EW-T2D', 'C-T2D', 'Cirrhosis']
EU	['Obesity', 'IBD', 'Colorectal', 'EW-T2D']
CH	['C-T2D', 'Cirrhosis']
T2D	['EW-T2D', 'C-T2D']
Colon	['IBD', 'Colorectal']

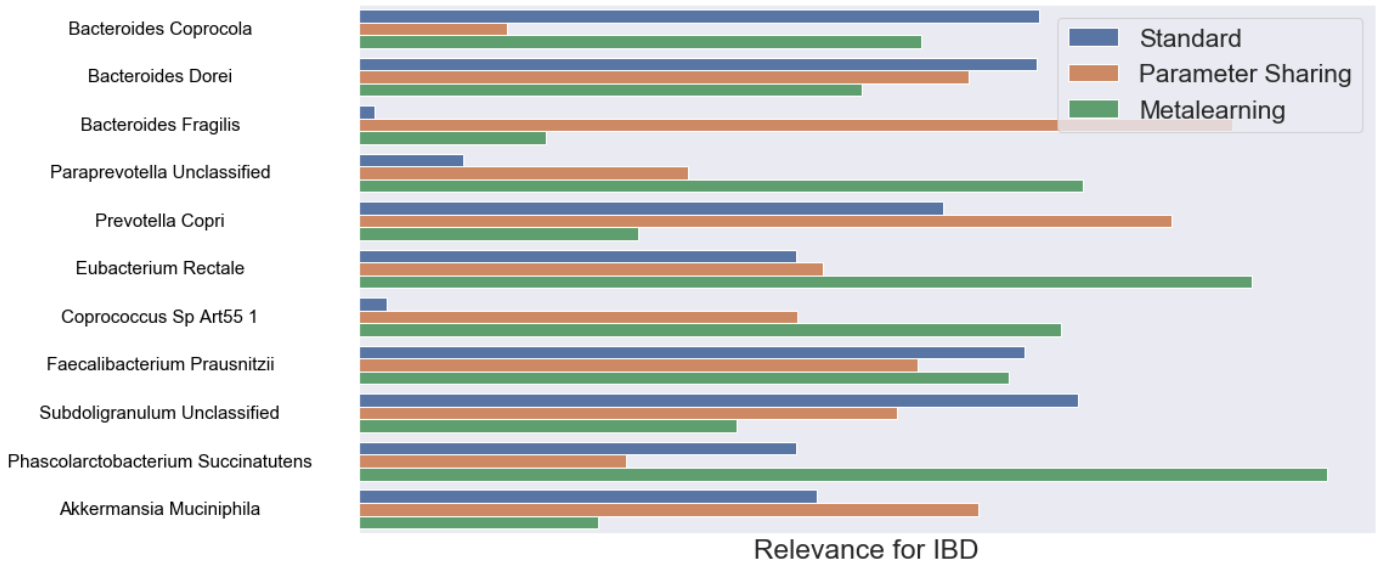
The different groups of datasets we tested transfer learning techniques on.

Figure S6: Comparing Metalearning's FFNN performance to standard FFNN Models



We are consistently getting better results on the same structure when usng the MAML scheme for parameter initialization.

Figure S7: Visualizing the Most Important Features for IBD Predictions



Showing models' most import predictions, via saliency maps. This figure was constructed with the same approach described in Figure 4.

References

[1] Marcos-Zambrano, Laura Judith, Kanita Karaduzovic-Hadziabdic, Tatjana Loncar Turukalo, Piotr Przymus, Vladimir Trajkovik, Oliver Aasmets, Magali Berland, et al. 2021. "Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment." Frontiers in Microbiology 12 (February). https://doi.org/10.3389/fmicb.2021.634511.

[2] Metwally, Ahmed A., Philip S. Yu, Derek Reiman, Yang Dai, Patricia W. Finn, and David L. Perkins. 2019. "Utilizing Longitudinal Microbiome Taxonomic Profiles to Predict Food Allergy via Long Short-Term Memory Networks." Edited by Yana Bromberg. PLOS Computational Biology 15 (2): e1006693. https://doi.org/10.1371/journal.pcbi.1006693.

[3] Sebastian Ruder. An Overview of Multi-Task Learning in Deep Neural Networks, 2017; arXiv:1706.05098.

[4] Chelsea Finn, Pieter Abbeel and Sergey Levine. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks, 2017; arXiv:1703.03400.

[5] Oh, Min, and Liqing Zhang. 2020. "DeepMicro: Deep Representation Learning for Disease Prediction Based on Microbiome Data." Scientific Reports 10 (1). https://doi.org/10.1038/s41598-020-63159-5.

[6] Qin, Junjie, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, et al. 2010. "A Human Gut Microbial Gene Catalogue Established by Metagenomic Sequencing." Nature 464 (7285): 59–65. https://doi.org/10.1038/nature08821.

[7] Karlsson, Fredrik H., Valentina Tremaroli, Intawat Nookaew, Göran Bergström, Carl Johan Behre, Björn Fagerberg, Jens Nielsen, and Fredrik Bäckhed. 2013. "Gut Metagenome in European Women with Normal, Impaired and Diabetic Glucose Control." Nature 498 (7452): 99–103. https://doi.org/10.1038/nature12198.

[8] Qin, Junjie, Yingrui Li, Zhiming Cai, Shenghui Li, Jianfeng Zhu, Fan Zhang, Suisha Liang, et al. 2012. "A Metagenome-Wide Association Study of Gut Microbiota in Type 2 Diabetes." Nature 490 (7418): 55–60. https://doi.org/10.1038/nature11450.

[9] Le Chatelier, Emmanuelle, Trine Nielsen, Junjie Qin, Edi Prifti, Falk Hildebrand, Gwen Falony, et al. 2013. "Richness of Human Gut Microbiome Correlates with Metabolic Markers." Nature 500 (7464): 541–46. https://doi.org/10.1038/nature12506.

[10] Qin, Nan, Fengling Yang, Ang Li, Edi Prifti, Yanfei Chen, Li Shao, Jing Guo, et al. 2014. "Alterations of the Human Gut Microbiome in Liver Cirrhosis." Nature 513 (7516): 59–64. https://doi.org/10.1038/nature13568.

[11] Zeller, Georg, Julien Tap, Anita Y Voigt, Shinichi Sunagawa, Jens Roat Kultima, Paul I Costea, Aurélien Amiot, et al. 2014. "Potential of Fecal Microbiota for Early-stage Detection of Colorectal Cancer." Molecular Systems Biology 10 (11): 766. https://doi.org/10.15252/msb.20145645.

[12] Truong, Duy Tin, Eric A Franzosa, Timothy L Tickle, Matthias Scholz, George Weingart, Edoardo Pasolli, Adrian Tett, Curtis Huttenhower, and Nicola Segata. 2015. "MetaPhlAn2 for Enhanced Metagenomic Taxonomic Profiling." Nature Methods 12 (10): 902–3. https://doi.org/10.1038/nmeth.3589.

[13] Pasolli, Edoardo, Duy Tin Truong, Faizan Malik, Levi Waldron, and Nicola Segata. 2016. "Machine Learning Meta-Analysis of Large Metagenomic Datasets: Tools and Biological Insights." Edited by Jonathan A. Eisen. PLOS Computational Biology 12 (7): e1004977. https://doi.org/10.1371/journal.pcbi.1004977.

[14] Eytan Bakshy, Lili Dworkin, Brian Karrer et at, AE: A domain-agnostic platform for adaptive experimentation; https://konstantinkashin.com/assets/papers/AE-NeurIPS2018.pdf

[15] Karen Simonyan, Andrea Vedaldi and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, 2013;

[16] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for PyTorch, 2020; arXiv:2009.07896.

[17] Arnold, Sebastien M. R., Praateek Mahajan, Debajyoti Datta, Ian Bunner, and Konstantinos Saitas Zarkias. 2020. "learn2learn: A Library for Meta-Learning Research." arXiv [cs.LG]. http://arxiv.org/abs/2008.12284.

[18] Ulger Toprak, N., A. Yagci, B.M. Gulluoglu, M.L. Akin, P. Demirkalem, T. Celenk, and G. Soyletir. 2006. "A Possible Role of Bacteroides Fragilis Enterotoxin in the Aetiology of Colorectal Cancer." Clinical Microbiology and Infection 12 (8): 782–86. https://doi.org/10.1111/j.1469-0691.2006.01494.x.

[19] Delzenne, N.M., Cani, P.D., Everard, A. et al. Gut microorganisms as promising targets for the management of type 2 diabetes. Diabetologia 58, 2206–2217 (2015). https://doi.org/10.1007/s00125-015-3712-7

[20] Ai, Dongmei, Hongfei Pan, Xiaoxin Li, Yingxin Gao, Gang Liu, and Li C. Xia. 2019. "Identifying Gut Microbiota Associated With Colorectal Cancer Using a Zero-Inflated Lognormal Model." Frontiers in Microbiology 10 (April). https://doi.org/10.3389/fmicb.2019.00826.

[21] Upadhyaya, Smitha, and Gautam Banerjee. 2015. "Type 2 Diabetes and Gut Microbiome: At the Intersection of Known and Unknown." Gut Microbes 6 (2): 85–92. https://doi.org/10.1080/19490976.2015.1024918.

[22] Castaner, Olga, Albert Goday, Yong-Moon Park, Seung-Hwan Lee, Faidon Magkos, Sue-Anne Toh Ee Shiow, and Helmut Schröder. 2018. "The Gut Microbiome Profile in Obesity: A Systematic Review." International Journal of Endocrinology 2018: 1–9. https://doi.org/10.1155/2018/4095789.

[23] Rabizadeh, Shervin, Ki-Jong Rhee, Shaoguang Wu, David Huso, Christine M. Gan, Jonathan E. Golub, XinQun Wu, Ming Zhang, and Cynthia L. Sears. 2007. "Enterotoxigenic Bacteroides Fragilis: A Potential Instigator of Colitis." Inflammatory Bowel Diseases 13 (12): 1475–83. https://doi.org/10.1002/ibd.20265.

[24] Aniruddh Raghu, Maithra Raghu, Samy Bengio and Oriol Vinyals. Rapid Learning or Feature Reuse? Towards Understanding the Effectiveness of MAML, 2019; arXiv:1909.09157.

[25] Jezabel R. Garcia, Federica Freddi, Feng-Ting Liao, Jamie McGowan, Tim Nieradzik, Da-shan Shiu, Ye Tian and Alberto Bernacchia. Meta-Learning with MAML on Trees, 2021; arXiv:2103.04691.

[26] Xi Sheryl Zhang, Fengyi Tang, Hiroko Dodge, Jiayu Zhou and Fei Wang. MetaPred: Meta-Learning for Clinical Risk Prediction with Limited Patient Electronic Health Records, 2019;

[27] Davis, N.M., Proctor, D.M., Holmes, S.P. et al. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome 6, 226 (2018). https://doi.org/10.1186/s40168-018-0605-2