A Additional experimental results

Tables 5 and 6 show the performance of Compositional Feature Dropout and Compositional CutMix on our microbiome benchmarks and are discussed in the main text. Tables 7 and 8 show the performance of Compositional Feature Dropout and Compositional CutMix on our non-microbiome benchmarks.

Tables 9, 10, and 11 show the effect on expected calibration error (ECE) of adding data augmentation to our various models. Notice that overall calibration error stays the same or improves slightly after incorporating data augmentation. Tables 14, 15, and 16, show error bars for Tables 2, 5, and 6, respectively.

In Tables 12 and 13, we show the performance of our contrastive models defined with Aitchison Mixup and Compositional CutMix, respectively. The latter performs equally strongly as Compositional Feature Dropout, as shown in Section 4.2, the former performs noticeably worse, however still no worse than DeepMicro, thus demonstrating the robustness of our contrastive framework to multiple augmentation strategies. Further gains may be obtained by combining augmentations in a contrastive loss, a direction which is left to future work. Table 3 shows error bars for Table 17, respectively.

Table 5: Data augmentation performance for Compositional Feature Dropout, similar to Table 2. Training sets augmented with Compositional Feature Dropout consistently performed better than those without.

Task	RF	Aug	XGB	Aug	mAML	Aug	DeepCoDa	Aug	MetaNN	Aug
1	0.72	**0.78**	0.76	**0.77**	**0.72**	**0.72**	0.73	**0.78**	0.74	**0.76**
2	0.78	**0.81**	**0.81**	**0.81**	**0.80**	**0.80**	0.78	**0.85**	0.74	**0.76**
3	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
4	0.60	**0.62**	**0.57**	**0.57**	**0.56**	**0.56**	**0.58**	0.52	**0.50**	**0.50**
5	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
6	0.81	**0.82**	0.82	**0.83**	**0.84**	0.83	0.78	**0.82**	0.75	**0.78**
7	**0.68**	**0.68**	**0.67**	**0.67**	0.73	**0.74**	0.63	**0.74**	**0.59**	0.53
8	0.62	**0.64**	0.66	**0.69**	**0.64**	0.63	0.45	**0.61**	0.64	**0.65**
9	**0.93**	0.92	**0.94**	**0.94**	**0.92**	**0.92**	0.84	**0.90**	0.76	**0.82**
10	0.53	**0.58**	0.57	**0.61**	**0.61**	**0.61**	**0.62**	0.55	0.62	**0.63**
11	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	0.98	**0.99**	**0.96**	**0.96**
12	0.55	**0.60**	0.58	**0.63**	**0.61**	**0.61**	**0.66**	0.65	0.58	**0.59**
Mean	0.77	**0.79**	0.78	**0.79**	**0.78**	**0.78**	0.75	**0.78**	0.74	**0.75**

B Contrastive learning implementation details

DeepMicro [41] is composed of an encoder and a decoder; the encoder has 2 hidden layers with 256 and 128 units respectively, and ReLU nonlinearities, and a 64-dimensional output. The decoder architecture is a mirror image of the encoder. The decoder output is of the same dimension as the encoder input, and the model is trained to minimize the mean squared error between the two, using the Adam optimizer with default parameters for 2,000 epochs.

To ensure a fair comparison, our contrastive model uses the same encoder architecture from DeepMicro. The decoder layers are discarded; instead, the 64-dimensional latent representation is passed to a multilayer projection head that is trained jointly with the encoder using a contrastive loss, as is done in SimCLR [9]. Our projection head contains one 32-dimensional hidden layer with ReLU activations, and a 16-dimensional projection output, which is normalized to 1 unit in L2 norm and passed to a temperature-scaled cross-entropy loss. Intuitively, this loss function is designed to draw the representations of positive pairs of examples close together, and push those of negative pairs far apart. Positive pairs refer to two synthetic samples generated by random data augmentations of the same training example; negative pairs refer to synthetic samples generated from different training examples. In our implementation, at each training step we sample two Compositional Feature Dropout from each training example, and compute the contrastive loss over all such pairs. In particular, our batch size corresponds to the entire training set, which is reasonable given the small sample sizes in our data. We again used the Adam optimizer with default parameters for 2,000 epochs.

Table 6: Data augmentation performance for Compositional CutMix, similar to Table 2. Training sets augmented with Compositional CutMix consistently enjoyed better test performance than those without, for all of our predictive models. Note that models trained with this data augmentation set a new state-of-the-art on 8 out of 12 tasks, including disease prediction for colorectal cancer, type 2 diabetes, and Crohn's disease.

Task	RF	Aug	XGB	Aug	mAML	Aug	DeepCoDa	Aug	MetaNN	Aug
1	0.72	**0.78**	0.76	**0.77**	0.72	**0.74**	0.73	**0.79**	**0.74**	**0.74**
2	0.78	**0.81**	0.81	**0.82**	0.80	**0.81**	0.78	**0.83**	0.74	**0.77**
3	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
4	0.60	**0.65**	0.57	**0.59**	0.56	**0.59**	**0.58**	0.57	**0.50**	**0.50**
5	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
6	0.81	**0.82**	0.82	**0.83**	**0.84**	**0.84**	0.78	**0.82**	0.75	**0.78**
7	0.68	**0.69**	0.67	**0.68**	**0.73**	0.72	0.63	**0.76**	**0.59**	0.55
8	0.62	**0.66**	0.66	**0.72**	**0.64**	**0.64**	0.45	**0.69**	0.64	**0.65**
9	**0.93**	**0.93**	0.94	**0.95**	0.92	**0.93**	0.84	**0.91**	0.76	**0.81**
10	0.53	**0.57**	0.57	**0.59**	**0.61**	**0.61**	**0.62**	0.61	0.62	**0.65**
11	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	0.98	**0.99**	**0.96**	**0.96**
12	0.55	**0.57**	0.58	**0.63**	**0.61**	**0.61**	**0.66**	0.65	0.58	**0.61**
Mean	0.77	**0.79**	0.78	**0.80**	0.78	**0.79**	0.75	**0.80**	0.74	**0.75**

Table 7: Data augmentation performance for Compositional Feature Dropout on non-microbiome CoDa benchmarks. We show the test AUC, averaged over 20 train/test bootstraps, for each non-microbiome dataset and predictive model, trained with and without data augmentation. Bold numbers indicate whether the version with or without augmentation performed best.

Dataset	n/p	RF	Aug	XGB	Aug	DeepCoDa	Aug	NN	Aug
Glass	213/8	0.84	**0.85**	0.84	**0.87**	0.83	**0.87**	**0.84**	**0.84**
Bayesite	20/5	0.86	**0.89**	**0.84**	0.81	**0.69**	0.52	0.53	**0.56**
Serum	30/4	0.81	**0.82**	0.80	**0.82**	0.62	**0.69**	**0.87**	0.80
Hydrochem	246/14	**1.00**	**1.00**	**1.00**	**1.00**	0.96	**0.99**	**0.95**	**0.95**
Jura	147/7	0.93	**0.94**	0.92	**0.93**	**0.83**	0.71	**0.75**	0.68
Metabolites	220/885	**0.95**	0.94	**0.94**	0.93	0.84	**0.94**	**0.92**	**0.92**
MicroRNA	717/188	**0.90**	**0.90**	**0.91**	0.90	**0.87**	**0.87**	**0.90**	**0.90**
Coffee	30/15	0.92	**0.93**	0.80	**0.91**	0.71	**0.79**	0.65	**0.75**
Mean	-	0.90	**0.91**	0.88	**0.90**	0.79	**0.80**	**0.80**	**0.80**

C Other augmentation strategies

We propose an additional augmentation strategy, *Multinomial Resampling*, which can be thought of as a CoDa analogue of image blurring. The goal of this augmentation is to inject noise across the coordinates of $\mathbf{x}$ in a principled manner. Additive Gaussian noise is clearly ill-suited for CoDa, as it is not constrained to the simplex. Instead, inspired by the recording mechanism for microbiome CoDa, we propose generating new datapoints from a multinomial distribution. Note that high-throughput sequencing technologies record the microbial composition present in a specimen by subsampling the larger population. The total number of reads in this subsample, known as the sequencing depth, is an artifact of the measurement process. Assuming the subsample is small relative to the population, the multinomial distribution provides a crude approximation for this generative process, with each trial being drawn according to the true proportions in the underlying population. In this way, we can sample new datapoints from a multinomial distribution where the number of trials corresponds to the sequencing depth. Namely, each new datapoint is generated as follows:

1. Draw a training point i uniformly at random.
2. Draw $\tilde{\mathbf{x}} \sim \text{Multinomial}(L_i, \mathbf{x}_i)$, where L_i is the sequencing depth (a.k.a. library size).
3. Set $\mathbf{x}^{\text{aug}} = \tilde{\mathbf{x}}/(\sum_{j=1}^{p} \tilde{x}_j)$ and $y^{\text{aug}} = y_i$.

Table 8: Data augmentation performance for Compositional CutMix on non-microbiome CoDa benchmarks. We show the test AUC, averaged over 20 train/test bootstraps, for each non-microbiome dataset and predictive model, trained with and without data augmentation. Bold numbers indicate whether the version with or without augmentation performed best.

Dataset	n/p	RF	Aug	XGB	Aug	DeepCoDa	Aug	NN	Aug
Glass	213/8	0.84	**0.86**	0.84	**0.88**	0.83	**0.85**	0.84	**0.85**
Bayesite	20/5	**0.86**	0.84	**0.84**	**0.84**	**0.69**	0.48	**0.53**	**0.53**
Serum	30/4	**0.81**	0.68	**0.80**	0.71	0.62	**0.75**	**0.87**	0.81
Hydrochem	246/14	**1.00**	**1.00**	**1.00**	**1.00**	0.96	**0.99**	**0.95**	0.94
Jura	147/7	**0.93**	**0.93**	**0.92**	**0.92**	0.83	**0.92**	**0.75**	0.73
Metabolites	220/885	0.95	**0.96**	0.94	**0.95**	0.84	**0.94**	0.92	**0.93**
MicroRNA	717/188	0.90	**0.91**	0.91	**0.92**	0.87	**0.92**	0.90	**0.91**
Coffee	30/15	0.92	**0.97**	0.80	**0.88**	0.71	**0.78**	0.65	**0.66**
Mean	-	**0.90**	**0.90**	0.88	**0.89**	0.79	**0.83**	**0.80**	**0.80**

Table 9: Augmentation performance for Aitchison Mixup. We show the test ECE, averaged over 20 train/test bootstraps, for each learning task and predictive model, trained with and without data augmentation.

Task	RF	Aug	XGB	Aug	mAML	Aug	DeepCoDa	Aug	MetaNN	Aug
1	0.13	0.13	0.20	0.22	0.21	0.18	0.15	0.26	0.30	0.31
2	0.14	0.14	0.19	0.19	0.15	0.18	0.18	0.21	0.31	0.31
3	0.03	0.02	0.00	0.00	0.02	0.02	0.01	0.00	0.00	0.00
4	0.10	0.08	0.27	0.29	0.23	0.19	0.15	0.27	0.46	0.47
5	0.02	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	0.11	0.11	0.13	0.15	0.10	0.11	0.14	0.12	0.26	0.26
7	0.13	0.13	0.23	0.24	0.16	0.18	0.20	0.22	0.40	0.45
8	0.08	0.12	0.24	0.25	0.20	0.19	0.43	0.23	0.38	0.38
9	0.15	0.13	0.11	0.11	0.13	0.13	0.20	0.16	0.28	0.23
10	0.16	0.13	0.31	0.32	0.22	0.22	0.13	0.12	0.25	0.28
11	0.07	0.08	0.04	0.04	0.07	0.08	0.07	0.06	0.05	0.06
12	0.18	0.14	0.31	0.28	0.15	0.22	0.09	0.13	0.24	0.27
Mean	0.11	0.10	0.17	0.17	0.14	0.14	0.15	0.15	0.24	0.24

Notice that the number of trials L_i controls the noise level; as $L_i \to \infty$, $\mathbf{x}^{\text{aug}} \to \mathbf{x}_i$. When applied to CoDa that does not arise from high-throughput sequencing, L_i can be specified arbitrarily, or treated as a hyperparameter. In our microbiome datasets, the sequencing depth is typically on the order of $L \sim 10\,000$.

D Limitations of our work

In addition to those mentioned in Section 3, we note the following limitations of our work:

- Although rare, there were some datasets where data augmentation hurt the classification performance for certain models.
- We did not study the effect of combining multiple augmentations in a single pipeline, which is left to future study.

Table 10: Augmentation performance for Compositional Feature Dropout. We show the test ECE, averaged over 20 train/test bootstraps, for each learning task and predictive model, trained with and without data augmentation.

Task	RF	Aug	XGB	Aug	mAML	Aug	DeepCoDa	Aug	MetaNN	Aug
1	0.13	0.13	0.20	0.22	0.21	0.21	0.15	0.24	0.30	0.26
2	0.14	0.13	0.19	0.16	0.15	0.15	0.18	0.13	0.31	0.27
3	0.03	0.01	0.00	0.00	0.02	0.02	0.01	0.00	0.00	0.00
4	0.10	0.10	0.27	0.27	0.23	0.21	0.15	0.12	0.46	0.46
5	0.02	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00
6	0.11	0.10	0.13	0.13	0.10	0.11	0.14	0.13	0.26	0.24
7	0.13	0.11	0.23	0.24	0.16	0.17	0.20	0.12	0.40	0.44
8	0.08	0.09	0.24	0.26	0.20	0.19	0.43	0.23	0.38	0.37
9	0.15	0.11	0.11	0.12	0.13	0.13	0.20	0.18	0.28	0.24
10	0.16	0.16	0.31	0.29	0.22	0.22	0.13	0.06	0.25	0.24
11	0.07	0.08	0.04	0.04	0.07	0.10	0.07	0.05	0.05	0.05
12	0.18	0.17	0.31	0.27	0.15	0.17	0.09	0.10	0.24	0.25
Mean	0.11	0.10	0.17	0.17	0.14	0.14	0.15	0.11	0.24	0.24

Table 11: Augmentation performance for Compositional CutMix. We show the test ECE, averaged over 20 train/test bootstraps, for each learning task and predictive model, trained with and without data augmentation.

Task	RF	Aug	XGB	Aug	mAML	Aug	DeepCoDa	Aug	MetaNN	Aug
1	0.13	0.13	0.20	0.22	0.21	0.19	0.15	0.24	0.30	0.29
2	0.14	0.13	0.19	0.19	0.15	0.17	0.18	0.17	0.31	0.29
3	0.03	0.01	0.00	0.00	0.02	0.02	0.01	0.00	0.00	0.00
4	0.10	0.10	0.27	0.27	0.23	0.15	0.15	0.25	0.46	0.48
5	0.02	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	0.11	0.09	0.13	0.16	0.10	0.12	0.14	0.12	0.26	0.24
7	0.13	0.14	0.23	0.24	0.16	0.16	0.20	0.18	0.40	0.44
8	0.08	0.11	0.24	0.23	0.20	0.22	0.43	0.27	0.38	0.38
9	0.15	0.11	0.11	0.10	0.13	0.12	0.20	0.15	0.28	0.25
10	0.16	0.16	0.31	0.31	0.22	0.19	0.13	0.13	0.25	0.26
11	0.07	0.06	0.04	0.04	0.07	0.10	0.07	0.05	0.05	0.06
12	0.18	0.18	0.31	0.29	0.15	0.19	0.09	0.12	0.24	0.25
Mean	0.11	0.10	0.17	0.17	0.14	0.14	0.15	0.14	0.24	0.24

Table 12: Similar to Table 3, but the contrastive model is now defined using Aitchison Mixup.

	Linear Evaluation			Finetuning		
Task	No pretrain	DeepMicro	Contrastive	No pretrain	DeepMicro	Contrastive
1	0.59	0.68	**0.72**	0.72	0.75	**0.77**
2	0.67	0.76	**0.77**	0.76	0.76	**0.77**
3	**1.00**	**1.00**	0.96	**1.00**	**1.00**	**1.00**
4	0.47	**0.53**	**0.53**	0.50	**0.53**	**0.53**
5	**1.00**	**1.00**	0.99	**1.00**	**1.00**	**1.00**
6	0.75	**0.76**	**0.76**	0.77	0.76	**0.78**
7	0.59	0.68	**0.75**	0.59	0.59	**0.65**
8	0.65	**0.66**	0.60	0.66	**0.68**	0.66
9	0.72	0.71	**0.77**	0.76	0.77	**0.79**
10	0.53	**0.58**	0.54	0.60	**0.61**	**0.61**
11	0.96	**0.98**	0.97	**0.98**	**0.98**	0.97
12	0.66	**0.68**	0.63	0.62	**0.64**	0.62
Mean	0.72	0.75	0.75	0.74	0.76	0.76

Table 13: Similar to Table 3, but the contrastive model is now defined using Compositional CutMix.

Task	Linear Evaluation			Finetuning		
	No pretrain	DeepMicro	Contrastive	No pretrain	DeepMicro	Contrastive
1	0.59	0.68	**0.76**	0.72	0.75	**0.76**
2	0.67	0.76	**0.80**	0.76	0.76	**0.77**
3	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
4	0.47	0.53	**0.57**	0.50	0.53	**0.54**
5	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
6	0.75	0.76	**0.80**	0.77	0.76	**0.78**
7	0.59	0.68	**0.72**	0.59	0.59	**0.61**
8	0.65	**0.66**	**0.66**	0.66	**0.68**	**0.68**
9	0.72	0.71	**0.85**	0.76	0.77	**0.81**
10	0.53	0.58	**0.63**	0.60	0.61	**0.63**
11	0.96	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**
12	0.66	**0.68**	0.64	0.62	**0.64**	0.63
Mean	0.72	0.75	**0.78**	0.74	0.76	**0.77**

Table 14: Error bars for Table 2.

Task	RF	Aug	XGB	Aug	mAML	Aug	DeepCoDa	Aug	MetaNN	Aug
1	0.02	0.01	0.01	0.01	0.02	0.02	0.01	0.01	0.02	0.02
2	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.02
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
7	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
8	0.02	0.03	0.02	0.02	0.03	0.02	0.02	0.02	0.03	0.03
9	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02
10	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01
12	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.02	0.02
Mean	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02

Table 15: Error bars for Table 5.

Task	RF	Aug	XGB	Aug	mAML	Aug	DeepCoDa	Aug	MetaNN	Aug
1	0.02	0.01	0.01	0.01	0.02	0.02	0.01	0.01	0.02	0.02
2	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.02
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
7	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
8	0.02	0.02	0.02	0.03	0.03	0.03	0.02	0.03	0.03	0.03
9	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02
10	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01
12	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
Mean	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02

Table 16: Error bars for Table 6.

Task	RF	Aug	XGB	Aug	mAML	Aug	DeepCoDa	Aug	MetaNN	Aug
1	0.02	0.01	0.01	0.01	0.02	0.02	0.01	0.01	0.02	0.01
2	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.01
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
7	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
8	0.02	0.02	0.02	0.03	0.03	0.03	0.02	0.02	0.03	0.02
9	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.02	0.02
10	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01
12	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.02	0.02
Mean	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01

Table 17: Error bars for Table 3.

	Linear Evaluation			Finetuning		
Task	No pretrain	DeepMicro	Contrastive	No pretrain	DeepMicro	Contrastive
1	0.02	0.02	0.01	0.02	0.02	0.01
2	0.03	0.02	0.02	0.02	0.02	0.02
3	0.00	0.00	0.00	0.00	0.00	0.00
4	0.02	0.02	0.02	0.01	0.01	0.02
5	0.00	0.00	0.00	0.00	0.00	0.00
6	0.01	0.01	0.01	0.01	0.01	0.01
7	0.02	0.02	0.02	0.02	0.02	0.02
8	0.02	0.02	0.02	0.02	0.02	0.02
9	0.02	0.02	0.02	0.02	0.02	0.02
10	0.01	0.02	0.02	0.02	0.02	0.02
11	0.00	0.00	0.00	0.00	0.00	0.00
12	0.02	0.01	0.01	0.02	0.02	0.02
Mean	0.01	0.01	0.01	0.01	0.01	0.01