Supplement for:
DeepCoDA: personalized interpretability for compositional health data

0.1. Study 4: 2-levels of interpretability (expanded)

The **DeepCoDA** framework offers 2-levels of interpretability: (1) the "weights" of the self-explanation module (layer w in Figure 1 of the main paper) tell us how the classifier predicts a class label; (2) the weights of the log-bottleneck module (θ_z in Figure 1 of the main paper) tell us which features contribute to each log-contrast.

At the **first level**, the product scores ($w_i * z_i$) can be interpreted directly by a clinical laboratory or researcher to identify which features drive the final prediction. When a classifier makes a decision, the patient-specific weights (w_i) are multiplied with the log-contrast values (z_i), then added together. In data set “3”, a patient is predicted to have inflammatory bowel disease if this sum exceeds zero; otherwise, the patient is healthy. The largest product scores contribute most to the decision.

Consider two patients, chosen randomly:

- Patient 26 is healthy: their product scores are [2.0, -15.6, -3.6, -1.1, 1.8].
- Patient 13 is unhealthy: their product scores are [0.5, -7.3, 8.8, 7.4, 0.1].

For patient 26, the second term is highly negative, suggesting that the patient is healthy. Since the second term is derived from the second log-contrast, we can infer that log-contrast 2 is most important for this prediction.

For patient 13, the third and fourth terms are highly positive, suggesting that the patient is unhealthy. Interestingly, the second is highly negative (like patient 26), suggesting that the second log-contrast “looks” healthy. However, this negative score is not enough to sway the final decision.

At the **second level**, the log-bottleneck weights define how each bacteria contribute to the ratios. For log-contrast 3, the bacteria *Gordonibacter pamelaeae* makes the largest contribution to the numerator, while *Bacteroides cellulosilyticus* makes the largest contribution to the denominator. For patient 13, the log-contrast values are [0.61, -2.98, -4.64, 5.79, -1.76]. The signs of the log-contrasts reveal which bacteria dominate. Negative values mean that the denominator bacteria outweigh those in the numerator; positives mean that the numerator outweighs the denominator.

Meanwhile, our canonical correlation analysis reminds us that the importance of log-contrast 3 (w_3) depends on the value of log-contrast 2 (z_2), via an interaction learned automatically from the data. The top panels in Figure 5 summarize the distribution of these patient-specific weights and log-contrast values across the entire patient cohort.

Data set ID	# Samples	# Features	# Classes	Class 1	Class 2
1	975	48	2	Crohn’s disease	Without
2	128	60	2	Men who have sex with men	Without
3	220	153	2	Control	IBD
4	164	158	2	Crohn’s disease	Ulcerative colitis
5	220	885	2	Control	IBD
6	164	885	2	Crohn’s disease	Ulcerative colitis
7	182	278	2	Case	Diarrheal control
8	247	610	2	Case	Non-Diarrheal control
9	292	1133	2	Colorectal cancer (CRC)	Without
10	318	1302	2	Colorectal cancer (CRC)	Non-CRC control
11	1182	188	2	Primary solid tumor	Solid tissue normal
12	1004	188	2	Her2 Cancer	Not Her2 Cancer
13	718	188	2	LumA Cancer	LumB Cancer
14	140	992	2	Crohn’s disease (ileum)	Without (ileum)
15	160	992	2	Crohn’s disease (rectum)	Without (rectum)
16	2070	3090	2	GI tract	Oral
17	180	3090	2	Female	Male
18	404	3090	2	Stool	Tongue (dorsum)
19	408	3090	2	Subgingival plaque	Supragingival plaque
20	172	980	2	Healthy	Colorectal cancer
21	124	2526	2	Without	Diabetes
22	130	2579	2	Cirrhosis	Without
23	199	660	2	Black	Hispanic
24	342	660	2	Nugent score high	Nugent score low
25	200	660	2	Black	White

Table 1. Characteristics of the 25 data sets used in Section 4.3.

Data set ID	Original source	Retrieved via
1	doi: 10.1016/j.chom.2014.02.005	doi: 10.1128/mSystems.00053-18
2	doi: 10.1016/j.ebiom.2016.01.032	doi: 10.1128/mSystems.00053-18
3	doi: 10.1038/s41564-018-0306-4	supplemental materials
4	doi: 10.1038/s41564-018-0306-4	supplemental materials
5	doi: 10.1038/s41564-018-0306-4	supplemental materials
6	doi: 10.1038/s41564-018-0306-4	supplemental materials
7	doi: 10.1128/mBio.01021-14	doi: doi.org/10.1038/s41467-017-01973-8
8	doi: 10.1128/mBio.01021-15	doi: doi.org/10.1038/s41467-017-01973-9
9	doi: 10.1186/s13073-016-0290-3	doi: doi.org/10.1038/s41467-017-01973-10
10	doi: 10.1186/s13073-016-0290-3	doi: doi.org/10.1038/s41467-017-01973-11
11	doi: 10.1038/ng.2764	labels from doi: 10.1186/s13058-016-0724-2
12	doi: 10.1038/ng.2764	labels from doi: 10.1186/s13058-016-0724-2
13	doi: 10.1038/ng.2764	labels from doi: 10.1186/s13058-016-0724-2
14	doi: 10.1016/j.chom.2014.02.005	doi: 10.1093/gigascience/giz042
15	doi: 10.1016/j.chom.2014.02.005	doi: 10.1093/gigascience/giz043
16	doi: 10.1038/nature11209	doi: 10.1093/gigascience/giz044
17	doi: 10.1038/nature11209	doi: 10.1093/gigascience/giz045
18	doi: 10.1038/nature11209	doi: 10.1093/gigascience/giz046
19	doi: 10.1038/nature11209	doi: 10.1093/gigascience/giz047
20	doi: 10.1101/gr.126573.111	doi: 10.1093/gigascience/giz048
21	doi: 10.1038/nature11450	doi: 10.1093/gigascience/giz049
22	doi: 10.1038/nature13568	doi: 10.1093/gigascience/giz050
23	doi: 10.1073/pnas.1002611107	doi: 10.1093/gigascience/giz051
24	doi: 10.1073/pnas.1002611107	doi: 10.1093/gigascience/giz052
25	doi: 10.1073/pnas.1002611107	doi: 10.1093/gigascience/giz053

Table 2. Sources for the 25 data sets used in Section 4.3.

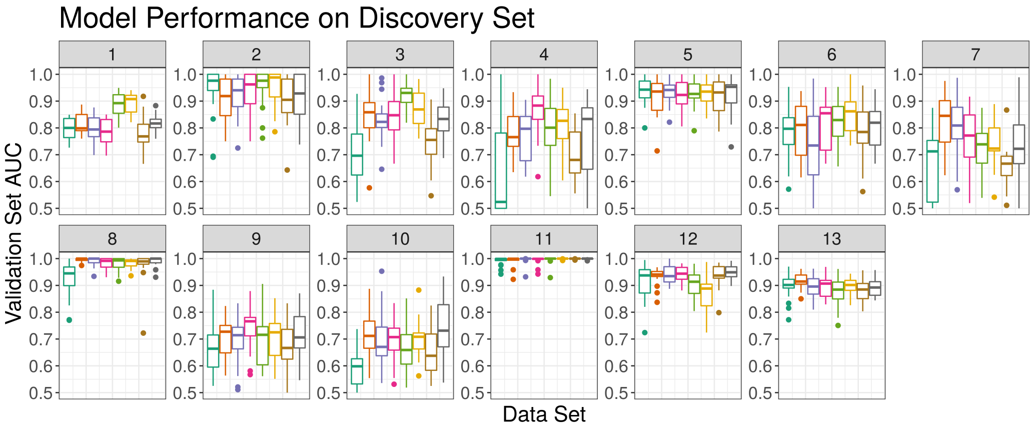

Figure 1. This figure shows the AUC for several models, organized by the method (x-axis) and data set source (facet). For all models, the boxplot shows the AUC distribution across 20 random 90%-10% training-test set splits. All **DeepCoDA** models use 5 log-bottlenecks and an L1 penalty of 0.01, chosen based on the "discovery set". Our model achieves appreciable performance across the 25 data sets. However, our aim is not to improve performance, but to extend personalized interpretability to compositional data.

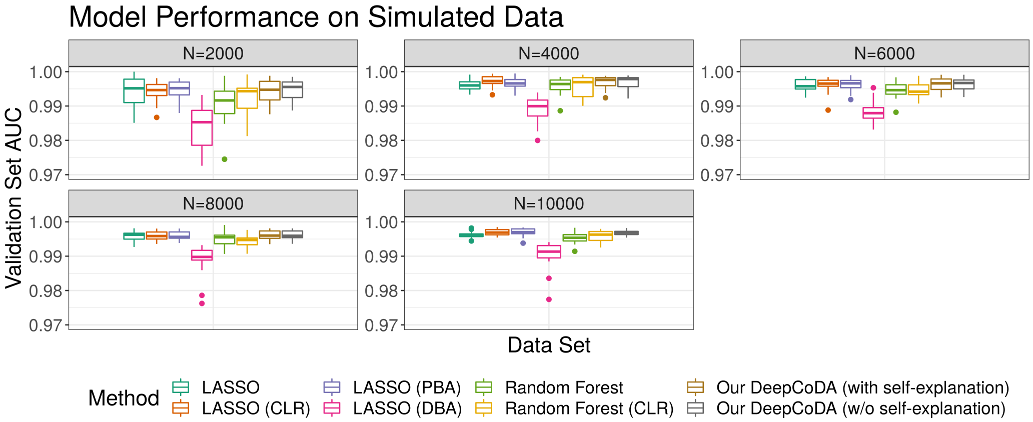

Figure 2. This figure shows the AUC for several models, organized by the method (x-axis) and number of samples in the second synthetic data set (facet). For all models, the boxplot shows the AUC distribution across 20 random 90%-10% training-test set splits. This figure confirms that the **DeepCoDA** model can scale to larger data sets with many samples.