Supplementary Material for 'TaxoNN: Ensemble of Neural Networks on Stratified Microbiome Data for Disease Prediction'

Contents

List of Tables

List of Figures

1 Supplementary Tables

1.0.1 Analysis of changes in CNN parameters

Supplementary Table 1: Results evaluating performance of our model by changing network parameters on the simulated data. The boldfaced attributes represent the parameter values for which the model performs the best.

Performance Analysis		
Parameter	Values	AUC
Stride Size	**1**	**0.887**
	2	0.853
	3	0.822
	4	0.812
	5	0.800
Window Size	3	0.834
	4	0.851
	5	**0.886**
	6	0.857
	7	0.842
No. of filters	16	0.842
	32	**0.877**
	64	0.852
No. of OTUs associated with risk of disease	8	0.792
	16	0.837
	32	**0.872**

We tried to compare the performance of our CNN model by changing the parameters associated with the network such as, stride size, number of causal OTUs, number of filters and window size to see if accuracy improves. Results are shown in the Supplementary Table 1. It was observed that as we increased the stride size (the number by which the window slides) in the CNN network, the model performance reduced, as the correlations between some of the adjacent OTUs were dropped in each slide. We obtained the best performance when stride size was 1 (AUC=0.887). Increasing the window size on the other hand, increased the AUC value as we observed mean AUC value reaching a high of 0.886 on window size 5. But as we went on increasing the window size, we noticed a drop in performance. Similarly, we chose the number of filters in the CNN model in a standard manner as suggested in [4]. As already discussed, filters are equal to the number of features in every layer of the network. We obtained and AUC of 0.877 with 32 filters, however, when we increased the number of filters from 32 to 64 we observed that the performance dropped. Finally by changing the number of OTUs associated with risk of disease in the model, we observed the best AUC with 32 associated OTUs (AUC=0.872) and decreasing the number of associated

OTUs reduced the prediction performance.

Supplementary Table 2: Table detailing the clusters in the T2D study [1] based on the phyla containing maximum number of OTUs. The right handside represents the genera in each cluster. The numbering provided to each genus provides a unique identifier to each OTU which is further used in Heatmaps as labels for the x and y axis, in Supplementary Figure 13, 14 and 15 to illustrate the correlations between the OTUs.

OTUs in T2D study		
Cluster	**Phylum**	**Genus**
Cluster 1	p_Firmicutes	1. g_Abiotrophia 2. g_Acidaminococcaceae_unclassified 3. g_Acidaminococcus 4. g_Alicyclobacillus 5. g_Allobaculum 6. g_Anaerococcus 7. g_Anaerofustis 8. g_Anaeroglobus 9. g_Anaerostipes 10. g_Anaerotruncus 11. g_Bacillus 12. g_Blautia 13. g_Bulleidia 14. g_Butyricicoccus 15. g_Butyrivibrio 16. g_Catenibacterium 17. g_Cellulosilyticum 18. g_Clostridiaceae_noname 19. g_Clostridiales_Family_XIII_Incertae_Sedis_noname 20. g_Clostridiales_Family_XIII_Incertae_Sedis_unclassified 21. g_Clostridiales_noname 22. g_Clostridium 23. g_Coprobacillus 24. g_Coprococcus 25. g_Dialister 26. g_Dorea 27. g_Eggerthia 28. g_Enterococcus 29. g_Erysipelotrichaceae_noname 30. g_Eubacterium 31. g_Faecalibacterium 32. g_Finegoldia 33. g_Flavonifractor 34. g_Gemella 35. g_Granulicatella

		36. g_Holdemania 37. g_Lachnoanaerobaculum 38. g_Lachnospiraceae_noname 39. g_Lactobacillus 40. g_Lactococcus 41. g_Leuconostoc 42. g_Marvinbryantia 43. g_Megamonas 44. g_Megasphaera 45. g_Mitsuokella 46. g_Oribacterium 47. g_Oscillibacter 48. g_Parvimonas 49. g_Pediococcus 50. g_Peptoniphilus 51. g_Peptostreptococcaceae_noname 52. g_Peptostreptococcus 53. g_Phascolarctobacterium 54. g_Pseudoflavonifractor 55. g_Pseudoramibacter 56. g_Roseburia 57. g_Ruminococcaceae_noname 58. g_Ruminococcus 59. g_Selenomonas 60. g_Shuttleworthia 61. g_Solobacterium 62. g_Staphylococcus 63. g_Stomatobaculum 64. g_Streptococcus 65. g_Subdoligranulum 66. g_Turicibacter 67. g_Veillonella 68. g_Weissella
Cluster 2	p_Proteobacteria	69. g_Acinetobacter 70. g_Actinobacillus 71. g_Aeromonas 72. g_Aggregatibacter 73. g_Bartonella 74. g_Bilophila 75. g_Brevundimonas 76. g_Buchnera 77. g_Burkholderia

78. g_Burkholderiales_noname
79. g_Campylobacter
80. g_Candidatus_Zinderia
81. g_Cardiobacteriaceae_unclassified
82. g_Caulobacter
83. g_Chromobacterium
84. g_Citrobacter
85. g_Citromicrobium
86. g_Comamonas
87. g_Cronobacter
88. g_Cupriavidus
89. g_Desulfovibrio
90. g_Enhydrobacter
91. g_Enterobacter
92. g_Enterobacteriaceae_noname
93. g_Erythrobacteraceae_unclassified
94. g_Escherichia
95. g_Gallionellaceae_unclassified
96. g_Haemophilus
97. g_Halomonas
98. g_Helicobacter
99. g_Kingella
100. g_Klebsiella
101. g_Lautropia
102. g_Limnohabitans
103. g_Mesorhizobium
104. g_Morganella
105. g_Neisseria
106. g_Oxalobacter
107. g_Pantoea
108. g_Paracoccus
109. g_Parasutterella
110. g_Plesiomonas
111. g_Polaromonas
112. g_Proteus
113. g_Providencia
114. g_Pseudoalteromonadaceae_unclassified
115. g_Pseudoalteromonas
116. g_Pseudomonas
117. g_Pseudoxanthomonas
118. g_Raoultella
119. g_Rheinheimera

		120. g_Rhodanobacter 121. g_Rhodobiaceae_unclassified 122. g_Serratia 123. g_Shewanella 124. g_Shigella 125. g_Shinella 126. g_Sinobacteraceae_unclassified 127. g_Sphingobium 128. g_Sphingopyxis 129. g_Spiribacter 130. g_Succinatimonas 131. g_Sutterella 132. g_Sutterellaceae_unclassified 133. g_Variovorax 134. g_Vibrio 135. g_Xanthomonas 136. g_Yersinia
Cluster 3	p_Actinobacteria	137. g_Actinomyces 138. g_Adlercreutzia 139. g_Agromyces 140. g_Alloscardovia 141. g_Atopobium 142. g_Bifidobacterium 143. g_Brachybacterium 144. g_Brevibacterium 145. g_Collinsella 146. g_Coriobacteriaceae_noname 147. g_Corynebacterium 148. g_Cryptobacterium 149. g_Dermatophilaceae_unclassified 150. g_Eggerthella 151. g_Gardnerella 152. g_Gordonibacter 153. g_Kocuria 154. g_Leifsonia 155. g_Leucobacter 156. g_Microlunatus 157. g_Mobiluncus 158. g_Nocardioides 159. g_Olsenella 160. g_Parascardovia 161. g_Propionibacteriaceae_unclassified

		162. g_Propionibacterium 163. g_Rothia 164. g_Scardovia 165. g_Slackia 166. g_Tropheryma 167. g_Varibaculum
Cluster 4	1. p_Spirochaetes 2. p_Synergistetes 3. p_Tenericutes 4. p_Verrucomicrobia 5. p_Bacteroidetes 6. p_Candidatus_Saccharibacteria 7. p_Chlorobi 8. p_Deinococcus_Thermus 9. p_Acidobacteria 10. p_Fusobacteria	168. g_Brachyspira 169. g_Fretibacterium 170. g_Pyramidobacter 171. g_Synergistes 172. g_Mycoplasma 173. g_Akkermansia 174. g_Naumovozyma 175. g_Saccharomyces 176. g_Saccharomycetaceae_unclassified 177. g_Alistipes 178. g_Alloprevotella 179. g_Bacteroidales_noname 180. g_Bacteroides 181. g_Bacteroidetes_noname 182. g_Barnesiella 183. g_Butyricimonas 184. g_Cellulophaga 185. g_Coprobacter 186. g_Dysgonomonas 187. g_Odoribacter 188. g_Parabacteroides 189. g_Paraprevotella 190. g_Pedobacter 191. g_Porphyromonas 192. g_Prevotella 193. g_Riemerella 194. g_Sphingobacterium 195. g_Zunongwangia 196. g_Candidatus_Saccharibacteria_noname 197. g_Candidatus_Saccharibacteria_noname_unclassified 198. g_Chlorobium 199. g_Deinococcus 200. g_Meiothermus

		201. g_Methanocaldococcaceae_unclassified 202. g_Acidobacteriaceae_unclassified 203. g_Granulicella 204. g_Cetobacterium 205. g_Fusobacterium 206. g_Leptotrichia 207. g_Leptotrichiaceae_unclassified 208. g_Rhodopirellula

Supplementary Table 3: Table detailing the clusters in the Cirrhosis study [2] based on the phyla containing maximum number of OTUs. The right handside represents the genera in each cluster. The numbering provided to each genus provides a unique identifier to each OTU which is further used in Heatmaps as labels for the x and y axis, in Supplementary Figure 16, 17 and 18 to illustrate the correlations between the OTUs.

OTUs in Cirrhosis study		
Cluster	**Phylum**	**Genus**
Cluster 1	p_Firmicutes	1. g_Abiotrophia 2. g_Acidaminococcaceae_unclassified 3. g_Acidaminococcus 4. g_Aerococcus 5. g_Anaerococcus 6. g_Anaerofustis 7. g_Anaeroglobus 8. g_Anaerostipes 9. g_Anaerotruncus 10. g_Anoxybacillus 11. g_Bacillus 12. g_Blautia 13. g_Bulleidia 14. g_Butyricicoccus 15. g_Butyrivibrio 16. g_Catenibacterium 17. g_Catonella 18. g_Centipeda 19. g_Clostridiaceae_noname 20. g_Clostridiales_Family_XIII_Incertae_Sedis_noname 21. g_Clostridiales_Family_XIII_Incertae_Sedis_unclassified 22. g_Clostridiales_noname 23. g_Clostridium 24. g_Coprobacillus 25. g_Coprococcus 26. g_Dialister 27. g_Dorea 28. g_Eggerthia 29. g_Enterococcus 30. g_Erysipelotrichaceae_noname 31. g_Eubacterium 32. g_Faecalibacterium 33. g_Filifactor 34. g_Finegoldia 35. g_Flavonifractor

		36. g_Gemella 37. g_Granulicatella 38. g_Holdemania 39. g_Lachnoanaerobaculum 40. g_Lachnospiraceae_noname 41. g_Lactobacillus 42. g_Lactococcus 43. g_Leuconostoc 44. g_Megamonas 45. g_Megasphaera 46. g_Mitsuokella 47. g_Oribacterium 48. g_Oscillibacter 49. g_Parvimonas 50. g_Pediococcus 51. g_Peptoniphilus 52. g_Peptostreptococcaceae_noname 53. g_Peptostreptococcus 54. g_Phascolarctobacterium 55. g_Pseudoflavonifractor 56. g_Roseburia 57. g_Ruminococcaceae_noname 58. g_Ruminococcus 59. g_Selenomonas 60. g_Shuttleworthia 61. g_Solobacterium 62. g_Staphylococcus 63. g_Stomatobaculum 64. g_Streptococcus 65. g_Subdoligranulum 66. g_Turicibacter 67. g_Veillonella 68. g_Weissella
Cluster 2	p_Proteobacteria	69. g_Acinetobacter 70. g_Actinobacillus 71. g_Aeromonas 72. g_Aggregatibacter 73. g_Bartonella 74. g_Bilophila 75. g_Bordetella 76. g_Burkholderia 77. g_Burkholderiales_noname

78. g_Campylobacter
79. g_Cardiobacteriaceae_unclassified
80. g_Cardiobacterium
81. g_Chromobacterium
82. g_Citrobacter
83. g_Comamonas
84. g_Cronobacter
85. g_Desulfovibrio
86. g_Eikenella
87. g_Enterobacter
88. g_Enterobacteriaceae_noname
89. g_Escherichia
90. g_Gallionellaceae_unclassified
91. g_Haemophilus
92. g_Halomonas
93. g_Helicobacter
94. g_Kingella
95. g_Klebsiella
96. g_Kosakonia
97. g_Lautropia
98. g_Morganella
99. g_Neisseria
100. g_Oxalobacter
101. g_Pantoea
102. g_Parasutterella
103. g_Pectobacterium
104. g_Plesiomonas
105. g_Proteus
106. g_Providencia
107. g_Pseudomonas
108. g_Pusillimonas
109. g_Ralstonia
110. g_Raoultella
111. g_Rhodopseudomonas
112. g_Rhodospirillum
113. g_Serratia
114. g_Shewanella
115. g_Shigella
116. g_Sinobacteraceae_unclassified
117. g_Succinatimonas
118. g_Sutterella
119. g_Sutterellaceae_unclassified

		120. g_Yersinia
Cluster 3	p_Actinobacteria	121. g_Actinomyces 122. g_Actinopolyspora 123. g_Adlercreutzia 124. g_Alloscardovia 125. g_Atopobium 126. g_Bifidobacterium 127. g_Brevibacterium 128. g_Collinsella 129. g_Coriobacteriaceae_noname 130. g_Corynebacterium 131. g_Cryptobacterium 132. g_Eggerthella 133. g_Gardnerella 134. g_Gordonibacter 135. g_Kocuria 136. g_Olsenella 137. g_Parascardovia 138. g_Propionibacteriaceae_unclassified 139. g_Propionibacterium 140. g_Pseudonocardia 141. g_Rothia 142. g_Saccharomonospora 143. g_Saccharopolyspora 144. g_Scardovia 145. g_Slackia
Cluster 4	1. p_Spirochaetes 2. p_Synergistetes 3. p_Tenericutes 4. p_Verrucomicrobia 5. p_Bacteroidetes 6. p_Candidatus_Saccharibacteria 7. p_Chlorobi 8. p_Deinococcus_Thermus 9. p_Acidobacteria 10. p_Fusobacteria	146. g_Brachyspira 147. g_Fretibacterium 148. g_Pyramidobacter 149. g_Synergistes 150. g_Akkermansia 151. g_Naumovozyma 152. g_Saccharomyces 153. g_Saccharomycetaceae_unclassified 154. g_Alistipes 155. g_Alloprevotella 156. g_Bacteroidales_noname 157. g_Bacteroides 158. g_Bacteroidetes_noname

		159. g_Barnesiella 160. g_Butyricimonas 161. g_Cellulophaga 162. g_Coprobacter 163. g_Dysgonomonas 164. g_Odoribacter 165. g_Parabacteroides 166. g_Paraprevotella 167. g_Pedobacter 168. g_Porphyromonas 169. g_Prevotella 170. g_Riemerella 171. g_Sphingobacterium 172. g_Zunongwangia 173. g_Candidatus_Saccharibacteria_noname 174. g_Candidatus_Saccharibacteria_noname_unclassified 175. g_Chlorobium 176. g_Deinococcus 177. g_Meiothermus 178. g_Methanocaldococcaceae_unclassified 179. g_Acidobacteriaceae_unclassified 180. g_Granulicella 181. g_Cetobacterium 182. g_Fusobacterium 183. g_Leptotrichia 184. g_Leptotrichiaceae_unclassified

Supplementary Table 4: Association of age and sex to outcome of disease status in the T2D and Cirrhosis studies.

	T2D			Cirrhosis		
Variables	**Cases**	**Controls**	**p-value**	**Cases**	**Controls**	**p-value**
Age Mean (standard deviation)	54.5 (13.7)	41.6 (12.7)	<0.001	49.9 (11.3)	42.5 (9.3)	<0.001
Male (Frequency (%))	106 (62.4%)	84 (48.3%)	0.009	80 (67.8%)	72 (63.2%)	0.457
Female (Frequency (%))	64 (37.6%)	90 (51.7%)		38 (32.2%)	42 (36.8%)	

1.0.2 Analysing effect of interaction terms

We have considered 3 interaction terms while simulating our OTU data to approximate the possible OTU interactions that may be present in the the real studies. These interaction terms introduce non-linearity in the OTU data and disease outcome. To analyse whether, $taxoNN$ is efficiently capturing this non-linearity, we compared the performance of $taxoNN$ with other machine learning methods with and without the 3 interaction terms during the simulations (Supplementary Table 5). We observed that if we removed the interaction terms, the AUC obtained through $taxoNN_{corr}$ on the test set was observed to be 0.891 whereas, $taxoNN_{dis}$ gave an AUC value of 0.884. However, it was interesting to note that, eliminating the non-linearity in the data improved the performance of other methods as well. RF gave an AUC value of 0.865, SVM's AUC was 0.844, Ridge regression's AUC was 0.841, Lasso regression gave an AUC of 0.838, GBC gave an AUC value of 0.827, NB's AUC value improved to 0.815 and CNN_shuffle and CNN_basic gave AUC values of 0.844 and 0.812 respectively. On the other hand, the results of the performance of each method with interaction terms is shown in Figure 4. We observed that there was a significant improvement in AUC values of $taxoNN_{corr}$ and other machine learning methods, ranging from difference in AUC from 0.037 to 0.13 when we introduced non-linearity in the simulation study.

Supplementary Table 5: AUC values tabulated for various machine learning methods on test set of simulation studies. The results are reported on considering model performance without (w/o) interactions and with interactions. Note that the last row shows the consistent improvement in the performance of the proposed model $taxoNN_{corr}$ for both scenarios.

Method	**AUC w/o interaction**	**AUC with interaction**
Random Forest	0.865	0.846
Gaussian Bayes Classifier	0.827	0.792
Support Vector Machines	0.844	0.825
Lasso Regression	0.838	0.799
Ridge Regression	0.841	0.823
Naive Bayes	0.815	0.790
CNN_basic	0.844	0.753
CNN_shuffle	0.812	0.822
$taxoNN_{dis}$	0.884	0.874
$taxoNN_{corr}$	**0.891**	**0.883**

Supplementary Table 6: Mean AUC values tabulated for various machine learning methods on training set of T2D and Cirrhosis studies. The results are reported on considering 10 times 10-fold cross-validation on both studies. Note that the last row shows the consistent improvement in the performance of the proposed model $taxoNN_{corr}$ for both studies.

Method	**AUC for T2D**	**AUC for Cirrhosis**
Random Forest	0.740	0.892
Gaussian Bayes Classifier	0.684	0.874
Support Vector Machines	0.721	0.881
Lasso Regression	0.687	0.862
Ridge Regression	0.699	0.877
Naive Bayes	0.682	0.870
CNN_basic	0.667	0.832
CNN_shuffle	0.736	0.895
$taxoNN_{dis}$	0.741	0.919
$taxoNN_{corr}$	**0.753**	**0.921**

1.0.3 Validation on external cohort

We used the Type 2 Diabetes study evaluated by Karlsson et al. in their 2013 Nature Paper [5], which comprises of metformin confounding information along with OTU data (We call it T2D II).

Supplementary Table 7: Association of age and metformin to outcome of disease status in the T2D II study

	T2D II study (53 cases and 43 controls)		
Variables	**Cases**	**Controls**	**p-value**
Age Mean (standard deviation)	70.4 (0.78)	70.3 (0.69)	0.286
Number of individuals with metformin intake (In %))	20 (37.7%)	0 (0%)	<0.001
Male (Frequency (%))	0 (0%)	0 (0%)	
Female (Frequency (%))	53 (100%)	43 (100%)	

This study had 53 cases and 43 controls, all of which were females. In the cases, 20 individuals had taken metformin medication, while none of the controls had taken metformin. A table describing the T2D II cohort in terms of age, metformin medication intake and number of samples is shown in the presented in the supplementary file, Supplementary Table 7.

We carried additional experiments on this dataset:

- 1^{st} Experiment: To externally validate our results as an independent cohort, we divided the new T2D study (T2D II) into 4 major clusters based on the phyla containing majority OTUs, in a similar fashion, as we had done for T2D study in our manuscript. We applied $taxoNN$ trained on T2D dataset [1] mentioned originally in our manuscript, to T2D II. We obtained robust results on comparing $taxoNN$ to other methods on this new validation set, shown in the first column of Supplementary Table 8.
- 2^{nd} Experiment: To understand the effect of metformin in the T2D II study we stratified the OTUs based on the phylum level into 4 clusters. After putting the OTU data into 4 clusters for all subjects, we provided metformin information in a column along with the relative abundance of OTUs in each of the clusters and trained the model. This was done in a similar manner as we had included age and sex as covariates along with OTU data in the original manuscript for T2D dataset. The AUC value obtained using metformin as covariate in the $taxoNN_{corr}$ model provided consistently better performance, in comparison to other

conventional machine learning models as shown in the second column of Supplementary Table 8.

Supplementary Table 8: AUC values tabulated for various machine learning methods on T2D II study. The results are reported considering model performance using only OTU data and with metformin information as a covariate alongwith OTU data. Note that the last row (values in bold) shows the consistent improvement in the performance of the proposed model $taxoNN_{corr}$ for both cases.

Method	**AUC on T2D II study**	
	OTU data	**OTU data with metformin covariate**
Random Forest	0.702	0.696
Gaussian Bayes Classifier	0.611	0.602
Support Vector Machines	0.641	0.655
Lasso Regression	0.667	0.658
Ridge Regression	0.689	0.688
Naive Bayes	0.662	0.667
CNN_basic	0.602	0.598
CNN_shuffle	0.658	0.649
$taxoNN_{dis}$	0.693	0.701
$taxoNN_{corr}$	**0.709**	**0.711**

These two experiments show that our method is stable and gives good performance on an external validation set, as well as, is robust when metformin is chosen as a covariate.

1.0.4 Stratification based on class level in the taxonomy tree

Choosing phylum level for our clustering was a strategic choice because we wanted to have adequate number of OTUs per cluster. This was done to ensure proper training of our model after stratification, and at the same time be able to find an association between the OTUs to arrange them for giving them as an input to the CNN.

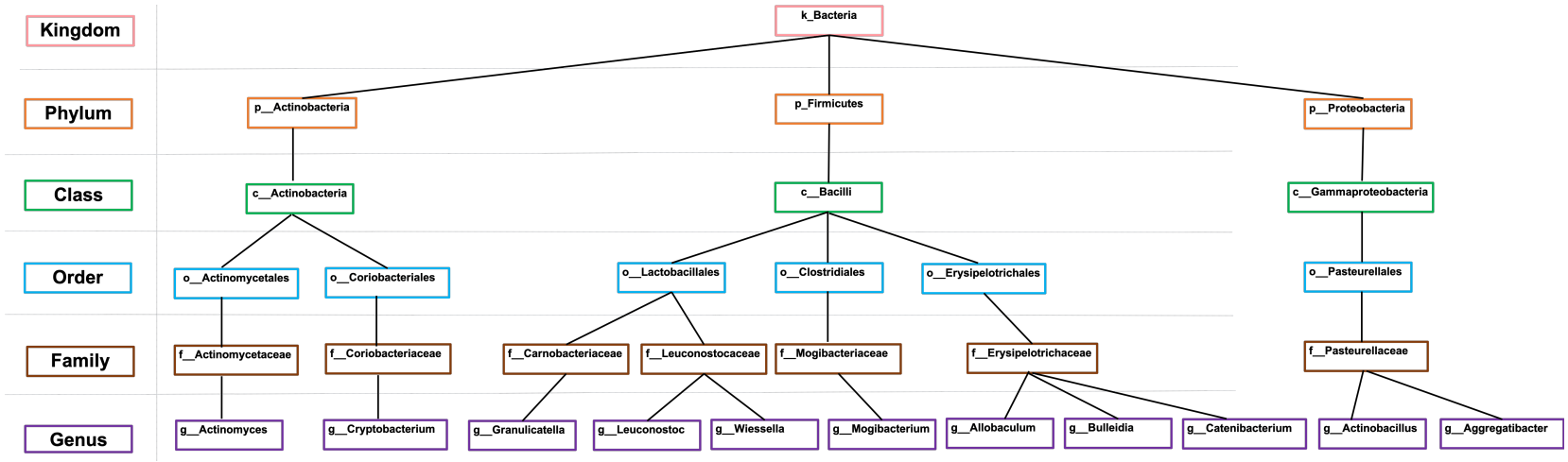
As we divided the clusters based on phyla with majority OTUs, we were able to determine 4 main clusters, each containing adequate number of OTUs for training our network. But when we went a level down in the taxonomy tree, to class level we noticed that there were fewer OTUs in each class.

Supplementary Table 9: AUC values tabulated for various machine learning methods upon class based stratification for T2D and Cirrhosis studies.

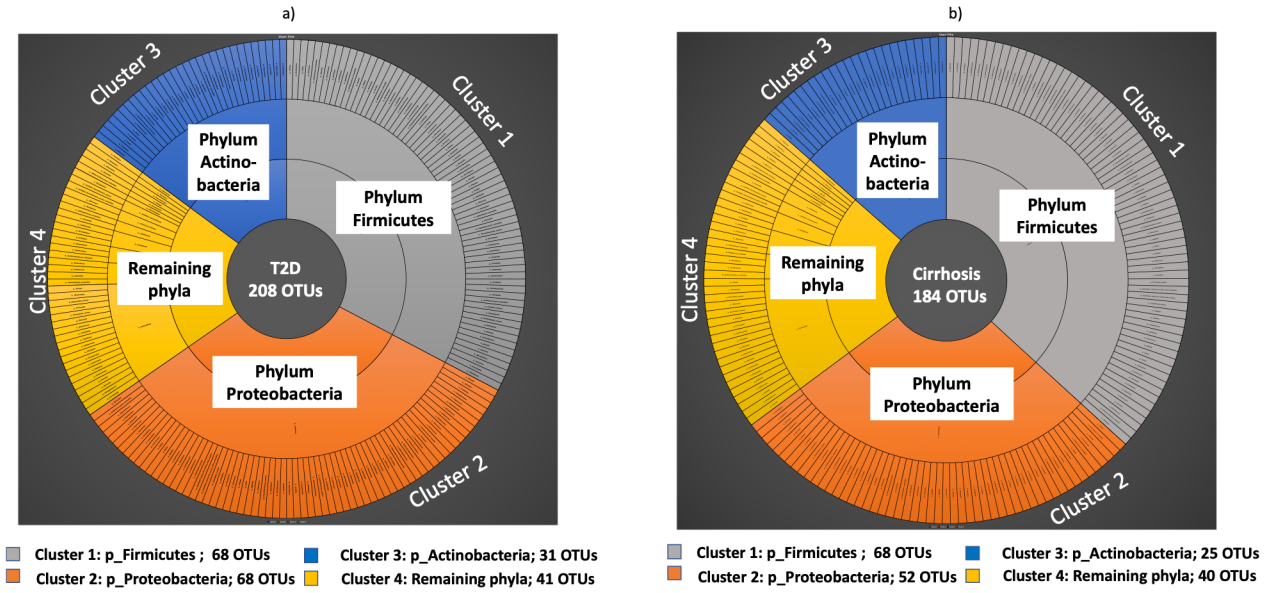
Method	**AUC T2D**	**AUC Cirrhosis**
Random Forest	0.703	0.893
Gaussian Bayes Classifier	0.642	0.816
Support Vector Machines	0.701	0.877
Lasso Regression	0.665	0.823
Ridge Regression	0.700	0.842
Naive Bayes	0.682	0.802
$taxoNN_{dis}$	0.700	0.887
$taxoNN_{corr}$	0.706	0.892

For example, in Cirrhosis dataset, we had 3 major phyla, namely, p_Actinobacteria with 38 OTUs, p_Firmicutes with 91 OTUs and p_Proteobacteria with 91 OTUs. Going down the taxonomy tree we had 40 different classes in class level and 60 different orders in order level. In such a case, in stratification based on classes, we could identify 5 major classes which had number of OTUs that were more than 20. Class c_Actinobacteria contained 38 OTUs, c_Bacilli contained 27 OTUs, c_Betaproteobacteria contained 25 OTUs, c_Clostridia contained 44 OTUs, c_Gammaproteobacteria contained 44 OTUs and the rest of the classes were clubbed in another cluster. The performance of each method on this approach for both studies is given in Table 9. We observed a drop in our model performance by stratifying in terms of classes, which we attribute to the fact that there were not enough OTUs in each cluster for the algorithm to learn well.

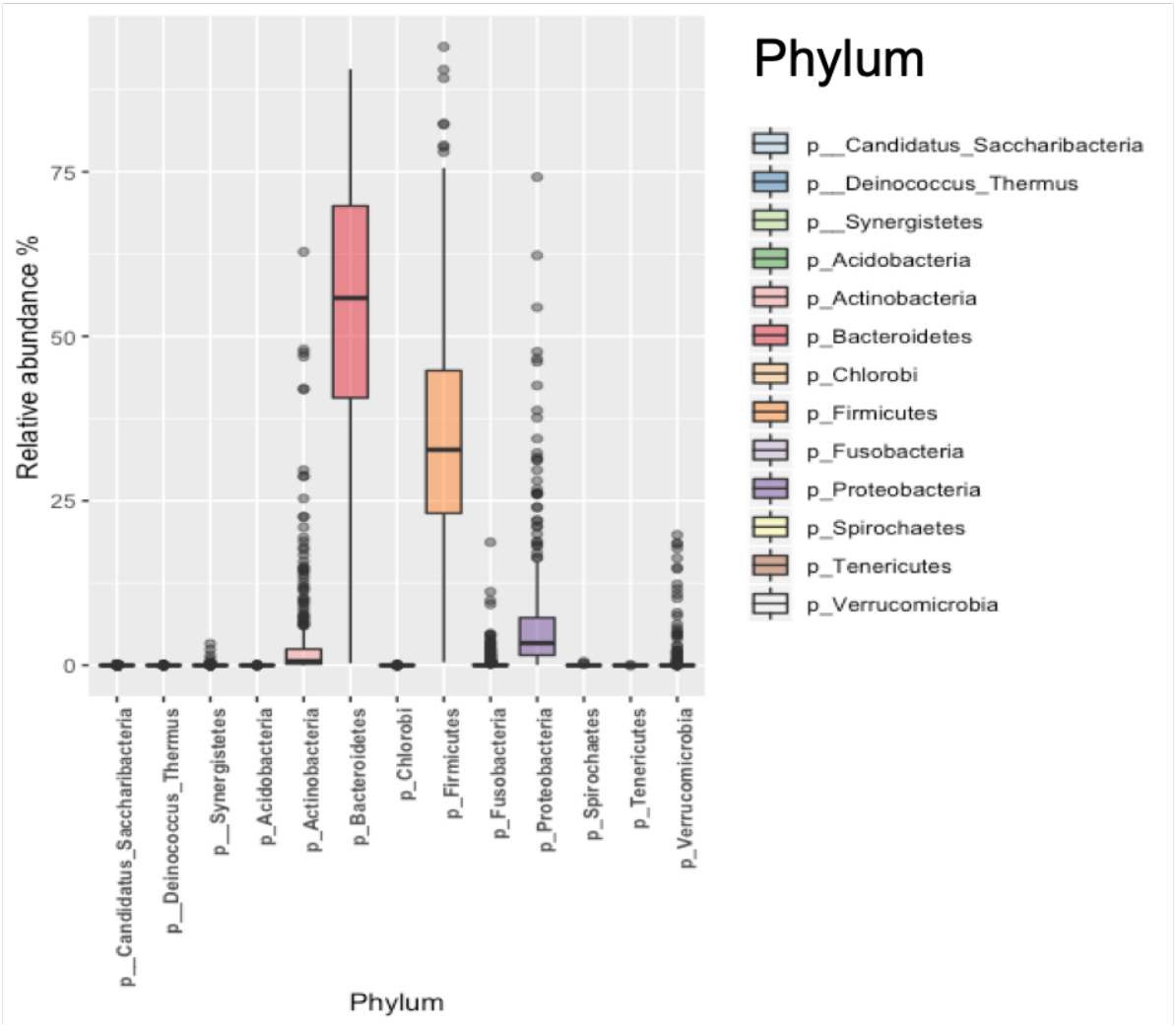
2 Supplementary Figures



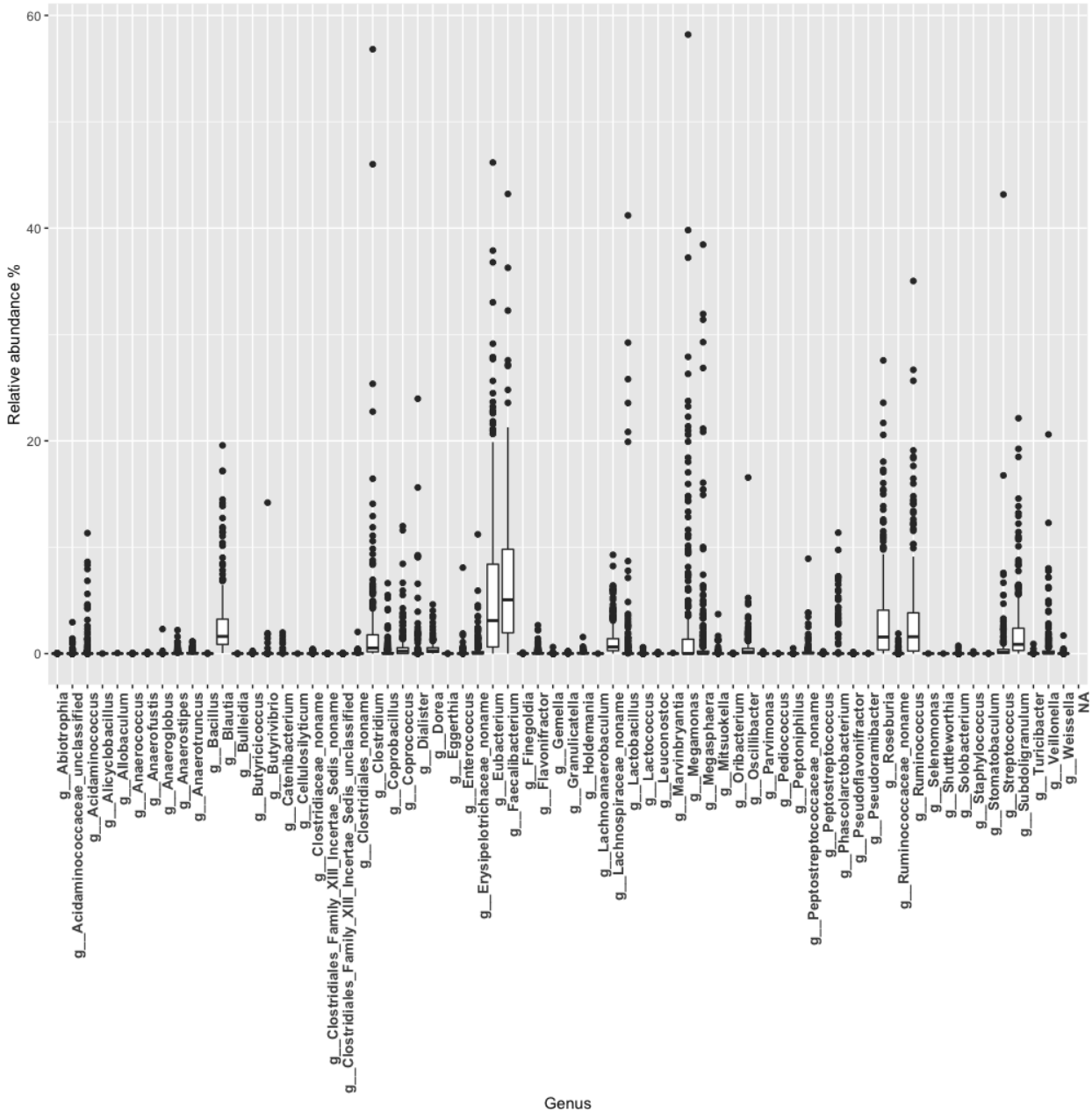
Supplementary Figure 1: Example of taxonomy levels in the OTU data illustrated using 29 OTUs taken from the OTU data publicly available in [3]



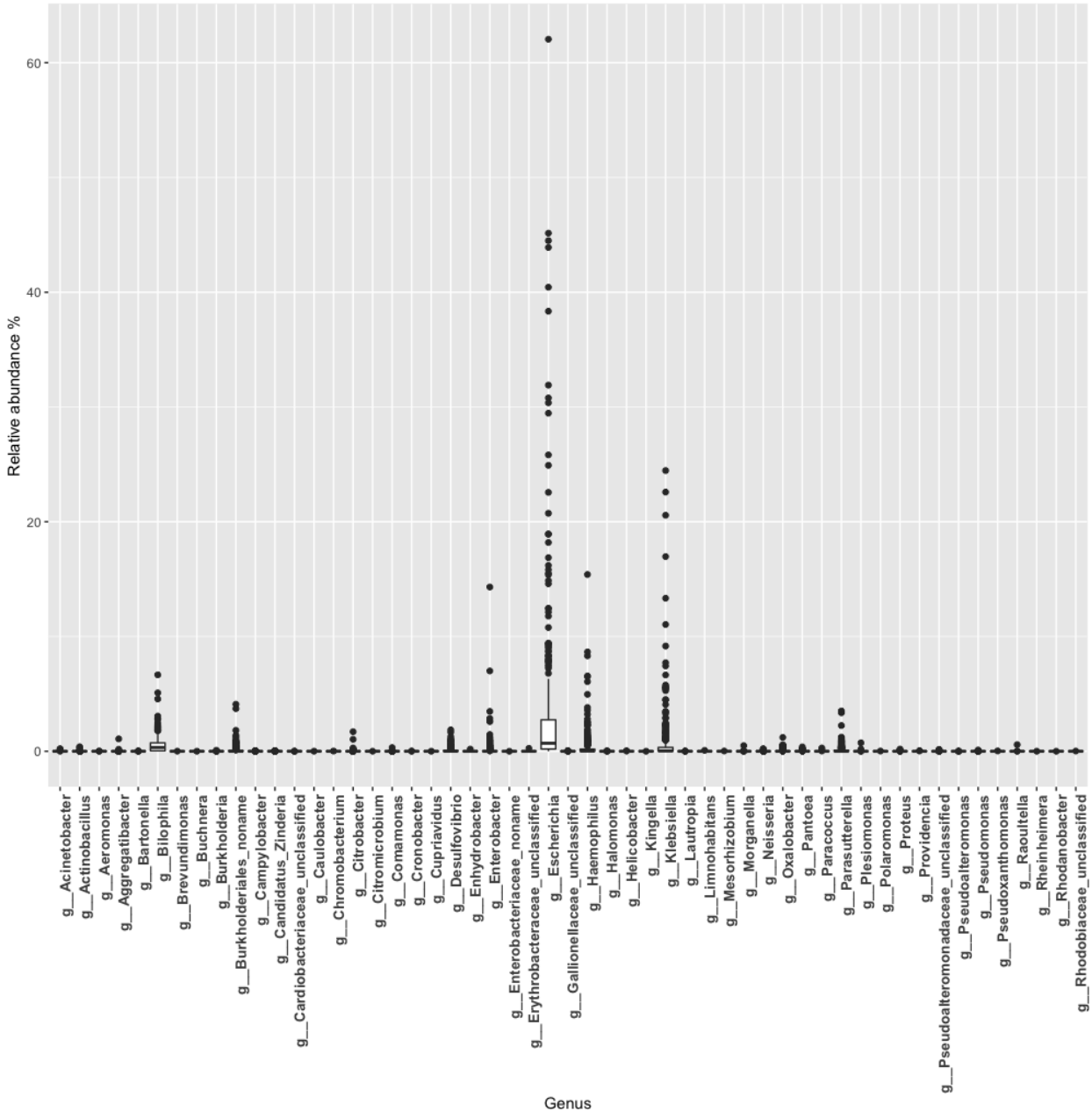
Supplementary Figure 2: OTU clustering in a) T2D study (208 OTUs) b) Cirrhosis study (184 OTUs). Outer circle represents the OTUs at the genus level for each cluster. Note that in both studies Proteobacteria, Actinobacteria and Firmicutes played as the phyla with highest number of OTUs in a single phylum, leading to forming the three major clusters for $taxoNN$.



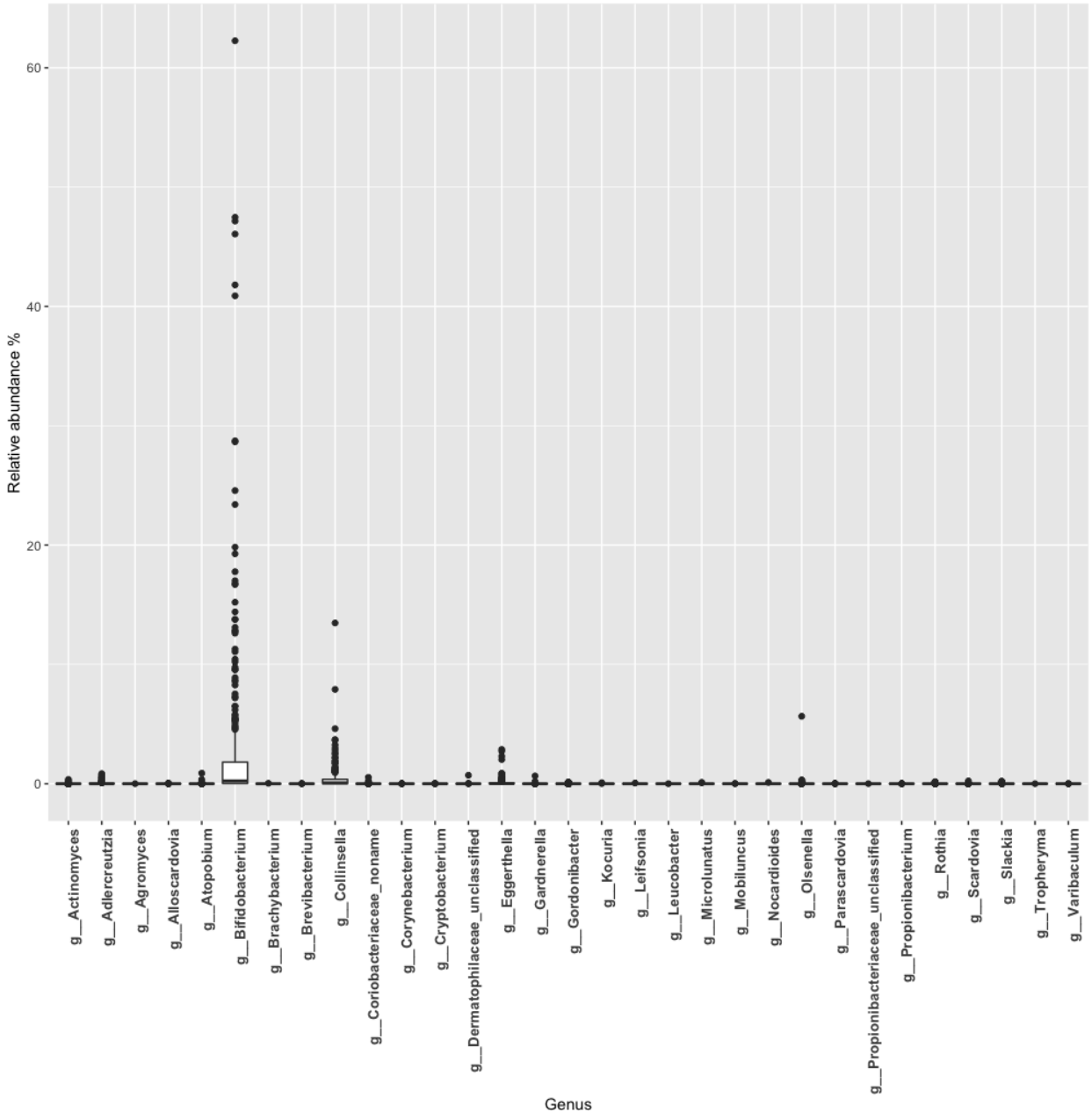
Supplementary Figure 3: Boxplot illustrating relative abundance percentage of OTUs in each phylum of the T2D study. The upper whisker extends from the hinge to the largest value no further than 1.5 * IQR from the hinge (where IQR is the inter-quartile range, or distance between the first and third quartiles). The lower whisker extends from the hinge to the smallest value at most 1.5 * IQR of the hinge. Data beyond the end of the whiskers are called ”outlying” points and are plotted individually.



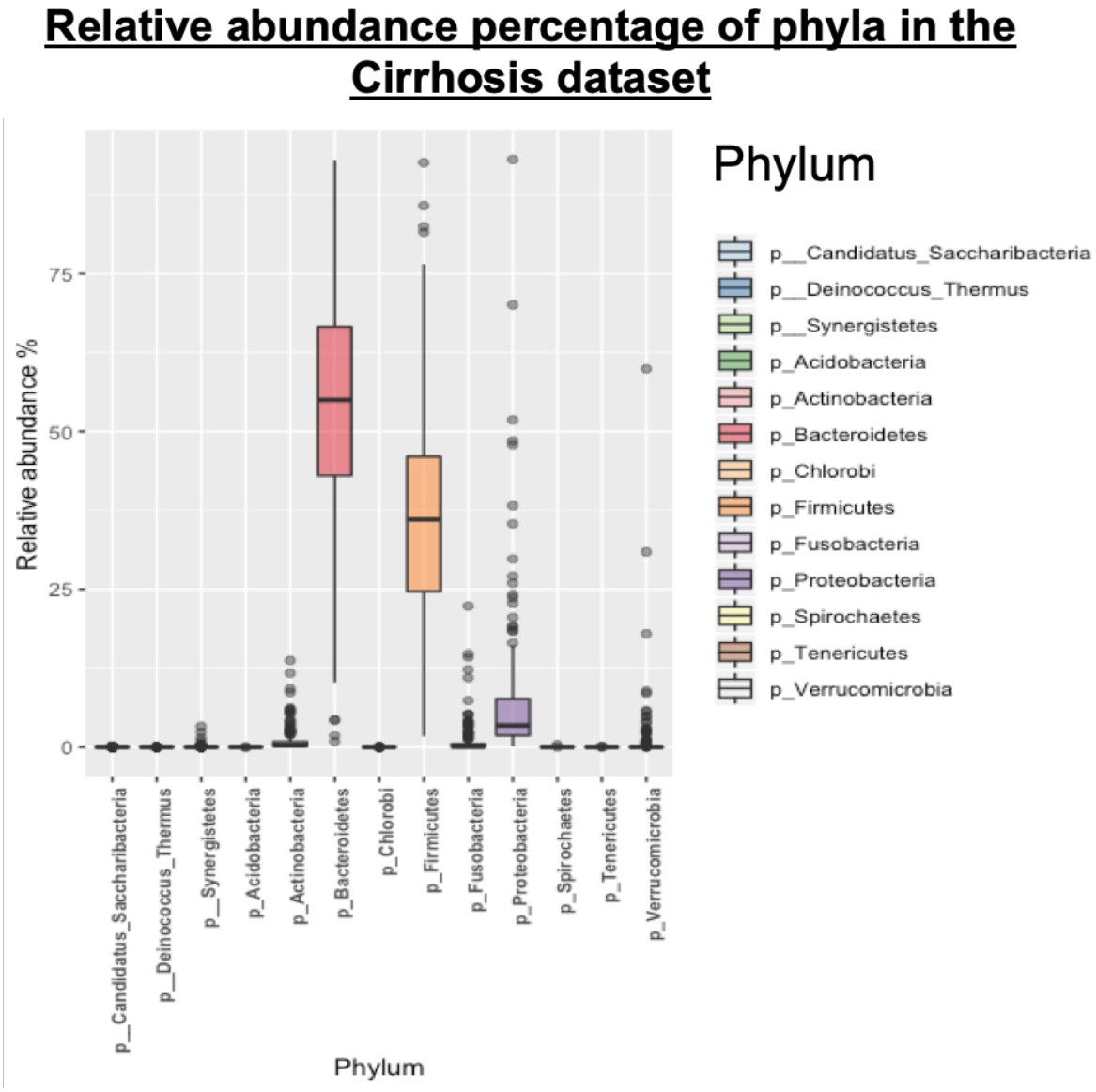
Supplementary Figure 4: Relative abundance percentage of OTUs at genus level in the Firmicutes phylum of the T2D study



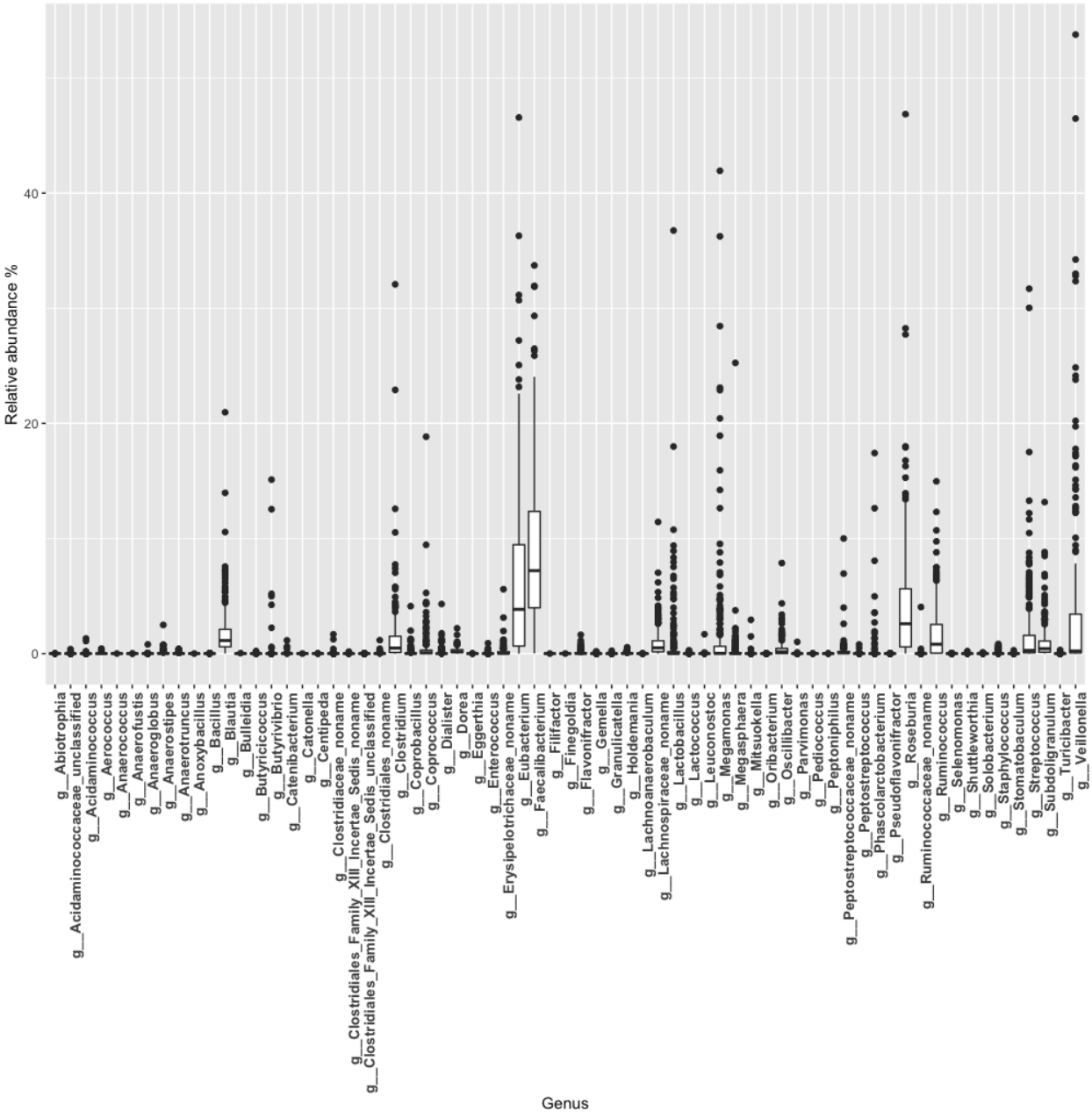
Supplementary Figure 5: Relative abundance percentage of OTUs at genus level in the Proteobacteria phylum of the T2D study



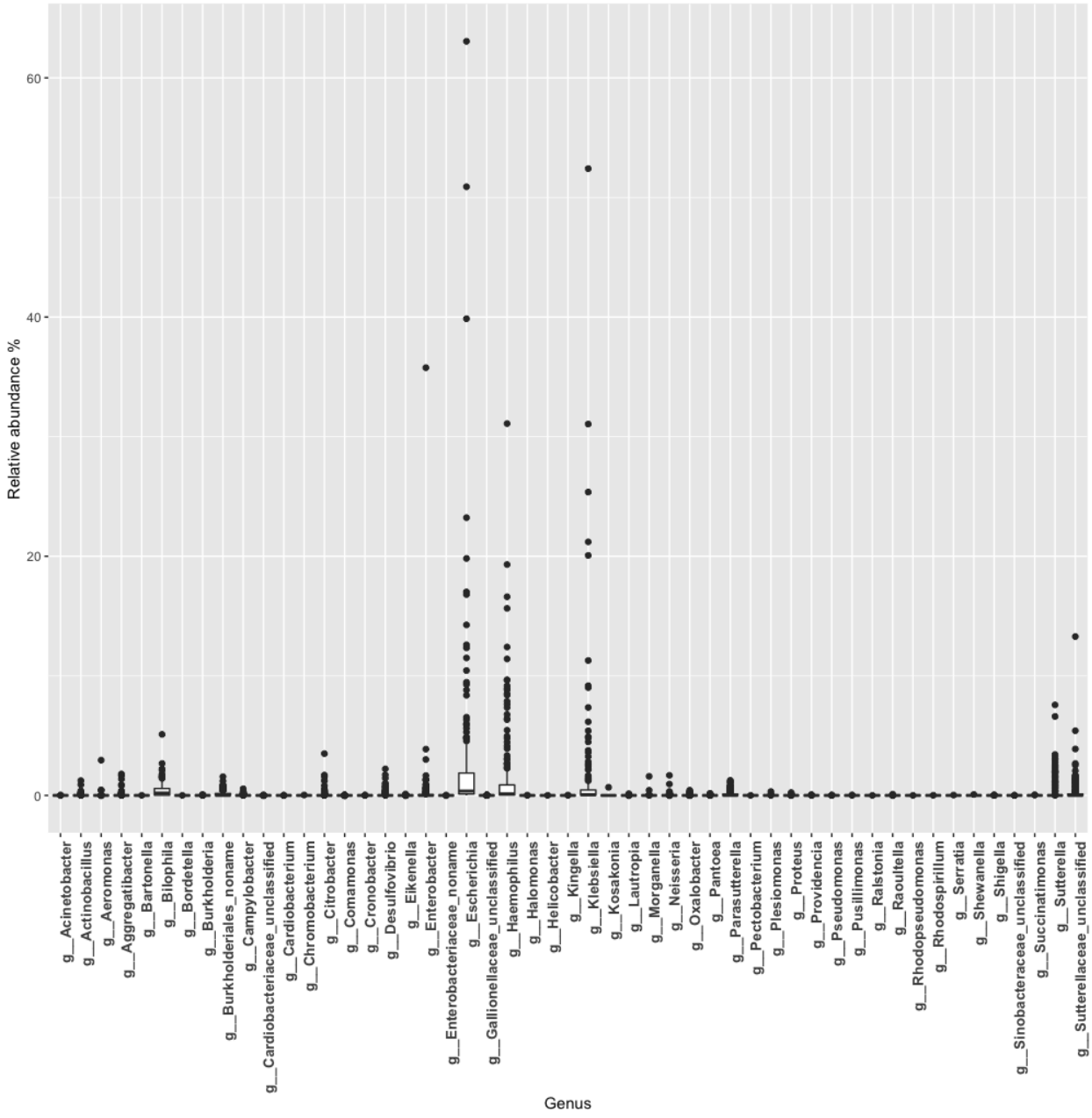
Supplementary Figure 6: Relative abundance percentage of OTUs at genus level in the Actinobacteria phylum of the T2D study



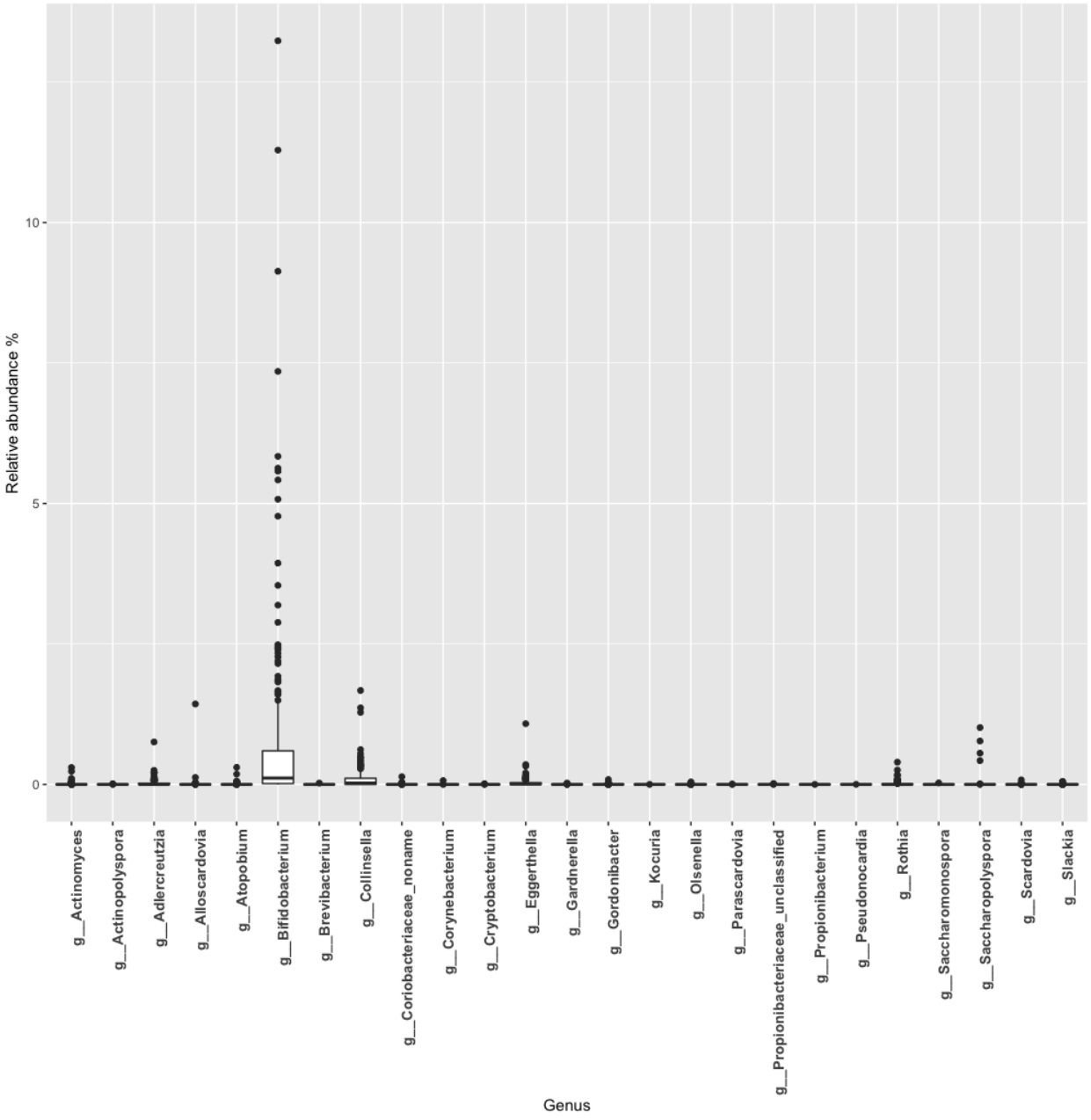
Supplementary Figure 7: Boxplot illustrating relative abundance percentage of OTUs in each phylum of the Cirrhosis study. The upper whisker extends from the hinge to the largest value no further than 1.5 * IQR from the hinge (where IQR is the inter-quartile range, or distance between the first and third quartiles). The lower whisker extends from the hinge to the smallest value at most 1.5 * IQR of the hinge. Data beyond the end of the whiskers are called ”outlying” points and are plotted individually.



Supplementary Figure 8: Relative abundance percentage of OTUs at genus level in the Firmicutes phylum of the Cirrhosis study



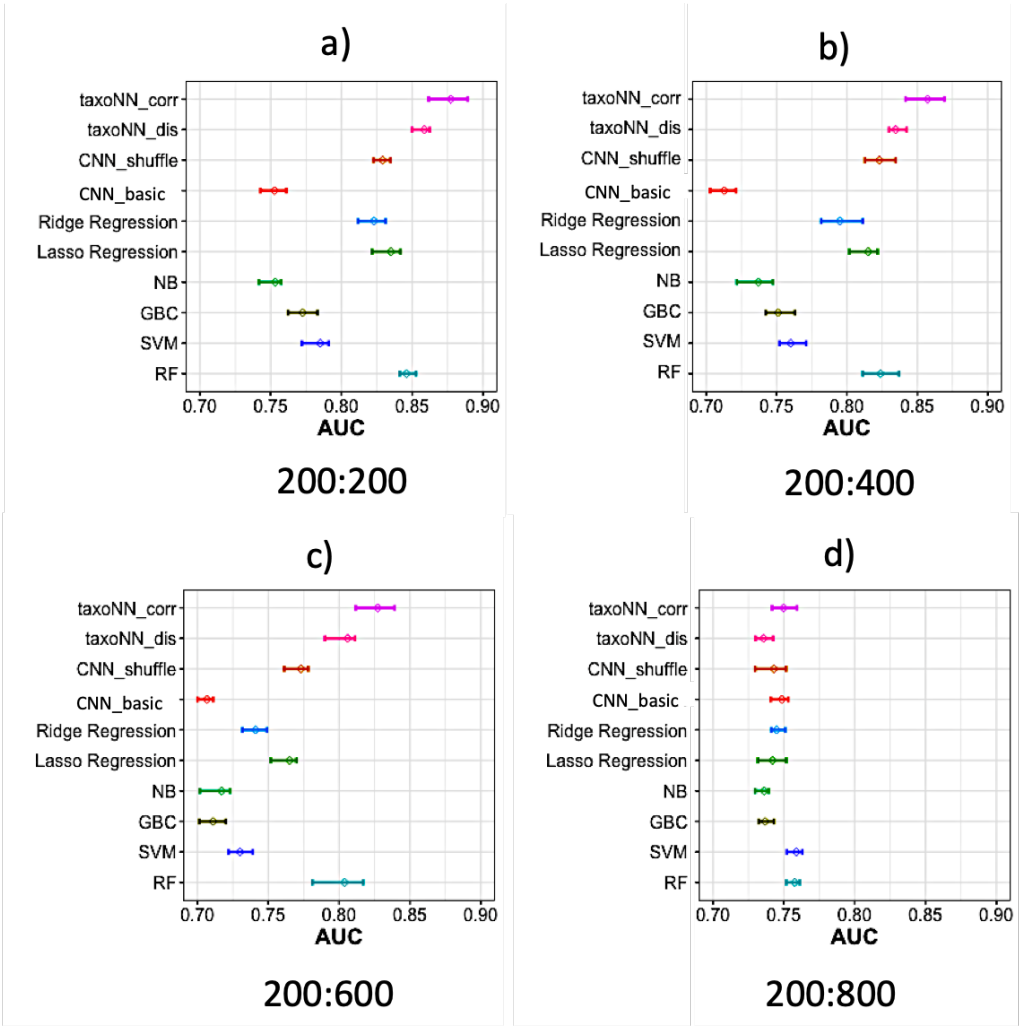
Supplementary Figure 9: Relative abundance percentage of OTUs at genus level in the Proteobacteria phylum of the Cirrhosis study



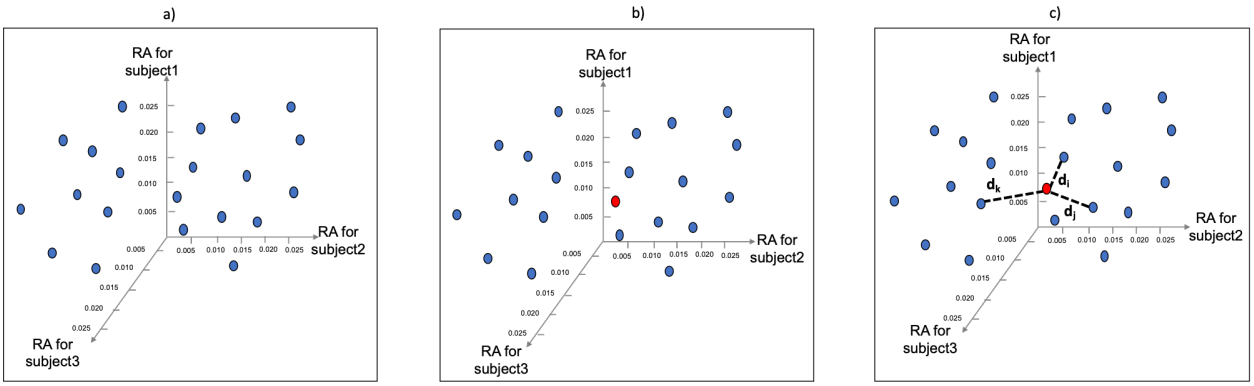
Supplementary Figure 10: Relative abundance percentage of OTUs at genus level in the Actinobacteria phylum of the Cirrhosis study

2.0.1 Robustness in imbalance of case and controls

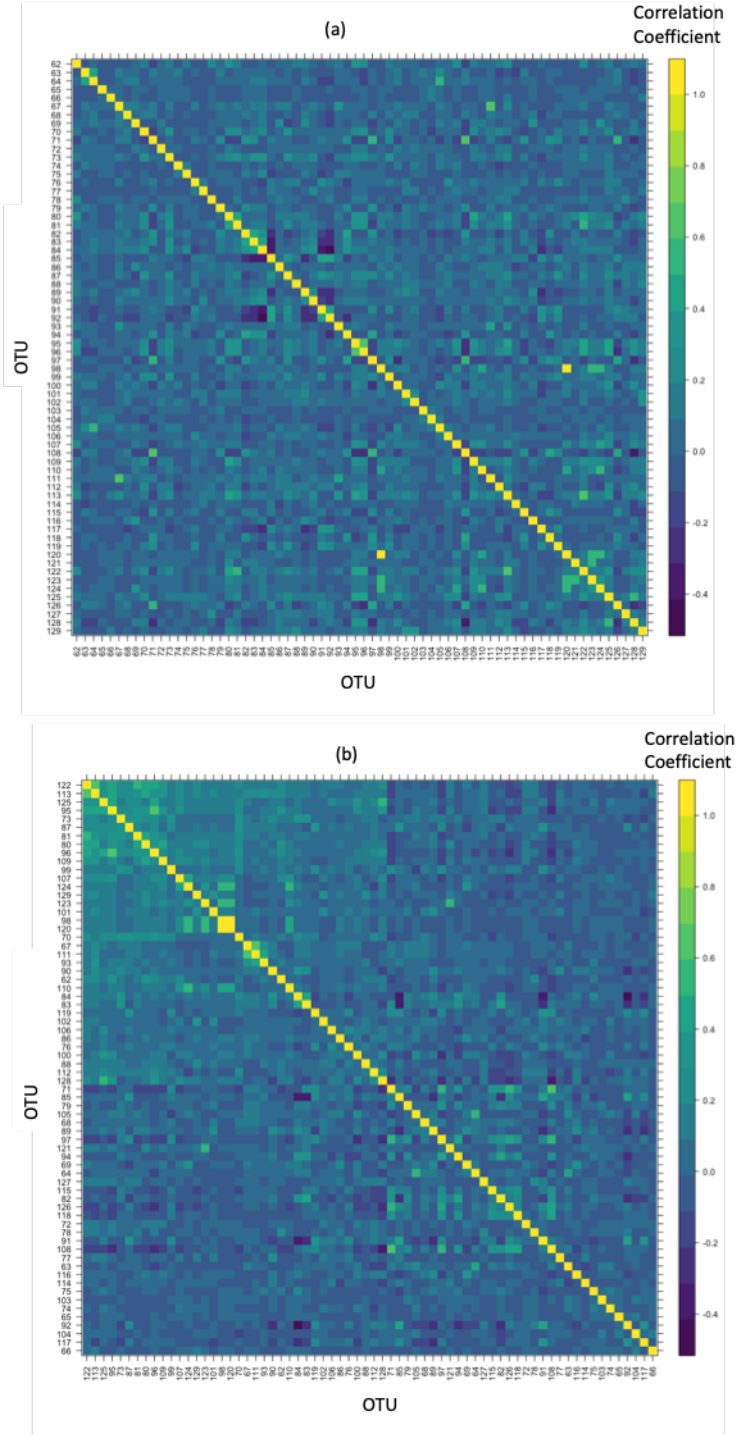
To provide a comprehensive analysis, we also examined how robust $taxoNN$ was in the scenario of imbalance of controls and cases in the input data. As can be seen in Supplementary Figure 11(a) with 200 cases and 200 controls, both the variations of the proposed model $taxoNN_{corr}$ and $taxoNN_{dis}$ perform well with a mean AUC of 0.877 and 0.858 respectively. In the case of 1:2 ratio (Supplementary Figure 11(b)) and 1:3 ratio (Supplementary Figure 11(c)) of cases and controls, $taxoNN_{corr}$, seemed to perform better than other machine learning models with AUC equal to 0.857 and 0.827 respectively. However, as we increased the number of controls to 800 (Supplementary Figure 11(d)), we saw that the performance of other methods became comparable to our technique with the difference in AUC values between $taxoNN_{corr}$ and RF method reducing to 0.007.



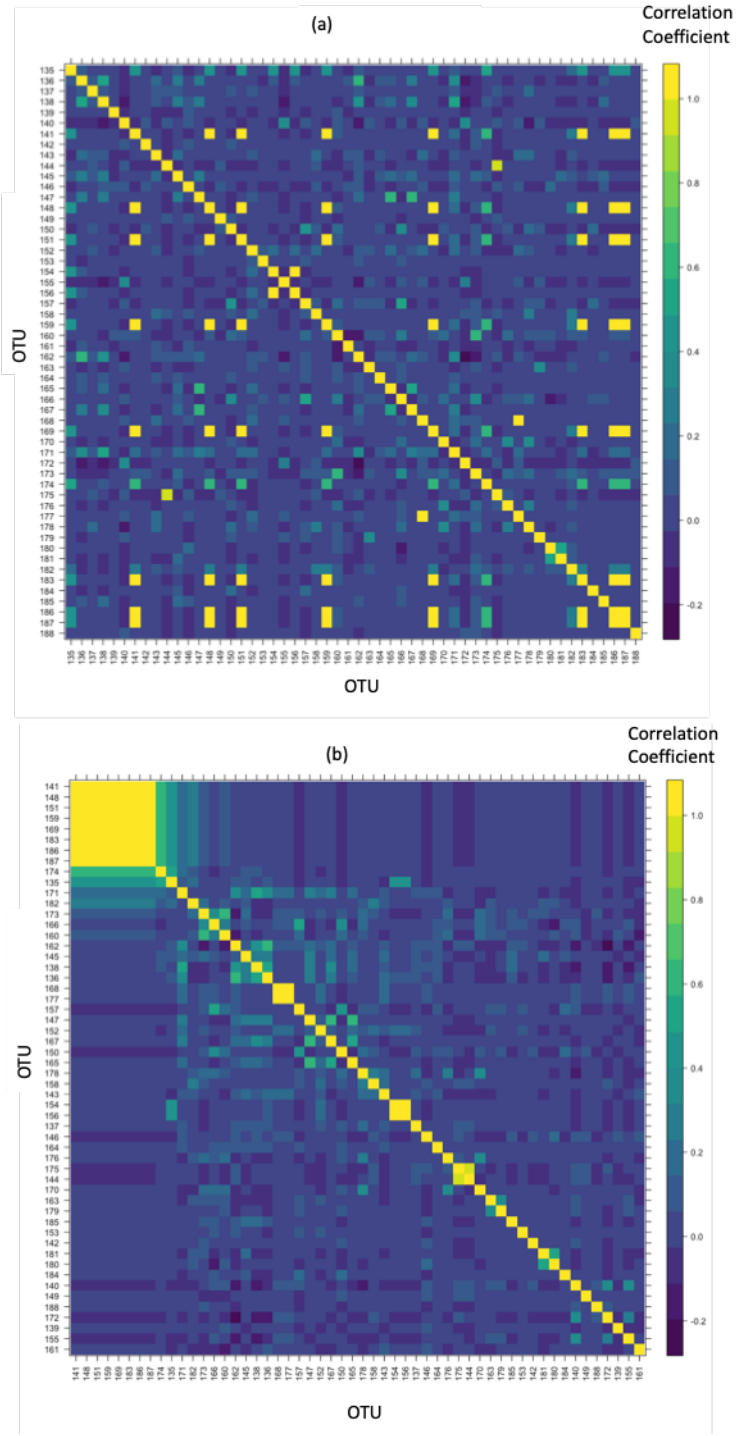
Supplementary Figure 11: Analysing performance of model in the scenario of case and control imbalance in the simulated data. a) Case and control data is properly balanced with 200 cases and controls each. b) Case and control ratio increasing to 1:2 c) 200 cases to 600 controls d) 1:4 ratio between case and control samples



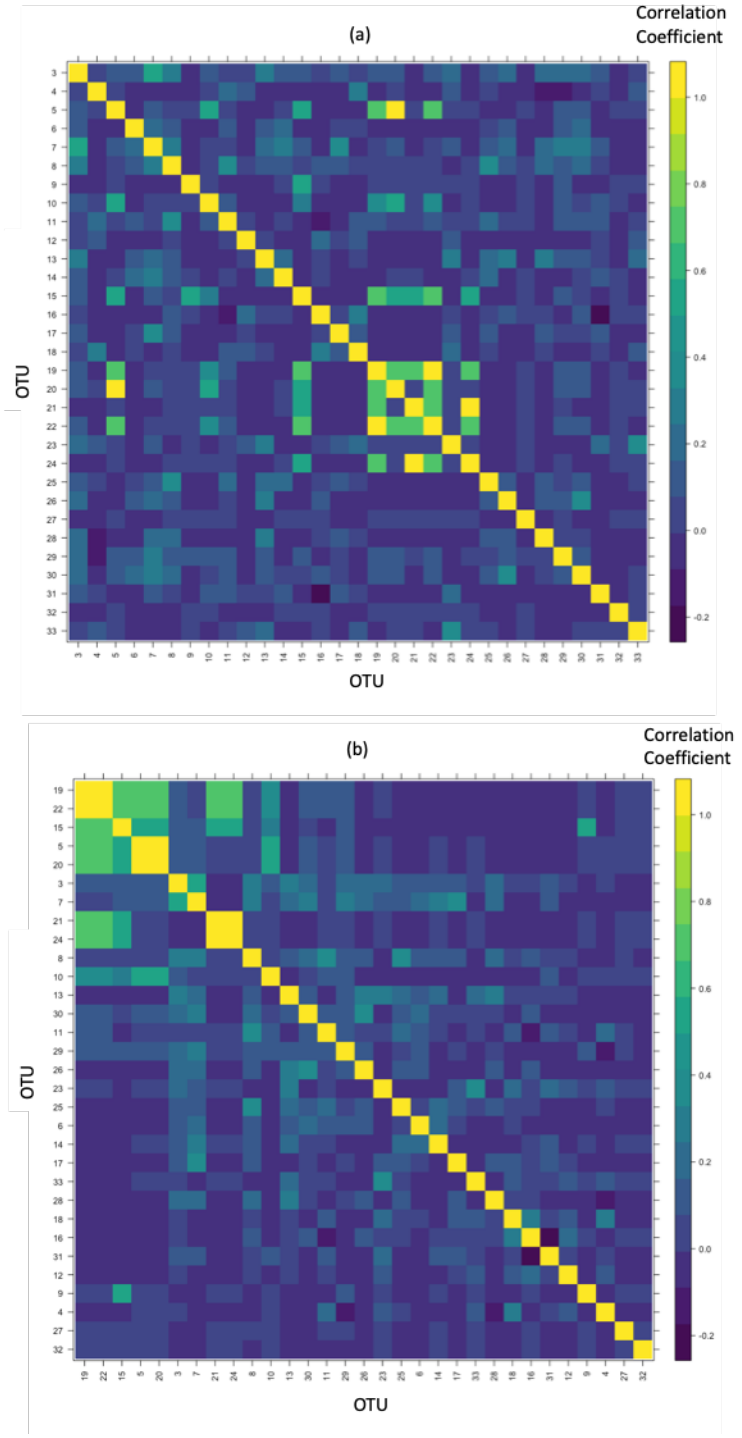
Supplementary Figure 12: An example plot to illustrate Euclidean distance based ordering in the OTUs in a cluster. (a) relative abundance of 21 OTUs for 3 subjects represented as blue dots. (b) red dot represents the medoid of the cluster. (c) black dashed lines represent the Euclidean distance of three OTUs from the medoid. As d_i is the smallest followed by d_j and d_k, therefore, OTU with distance d_i will be ordered first in the cluster as compared to OTU with distance d_j, followed by OTU with distance d_k. For the ease of understanding, this illustration is an example for only 3 subjects. However, in reality, there are multiple individuals (sample size = I) in a study leading to this 3-D plot being extended into an I-Dimensional space.



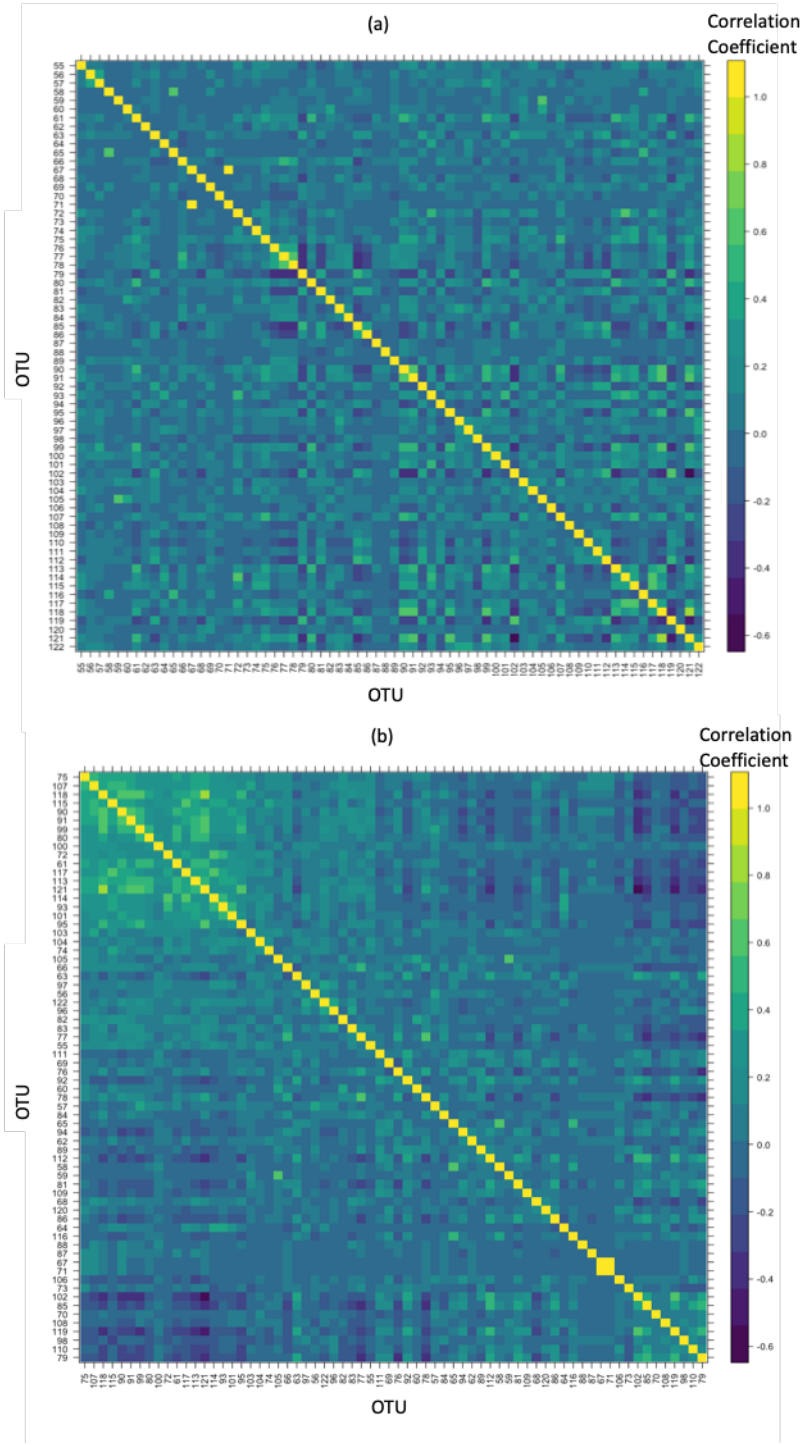
Supplementary Figure 13: Heatmaps for the Spearman rank of the OTUs in the cluster, Phylum Firmicutes, (a) before ordering and (b) after the ordering based on correlation of the OTUs in the T2D study



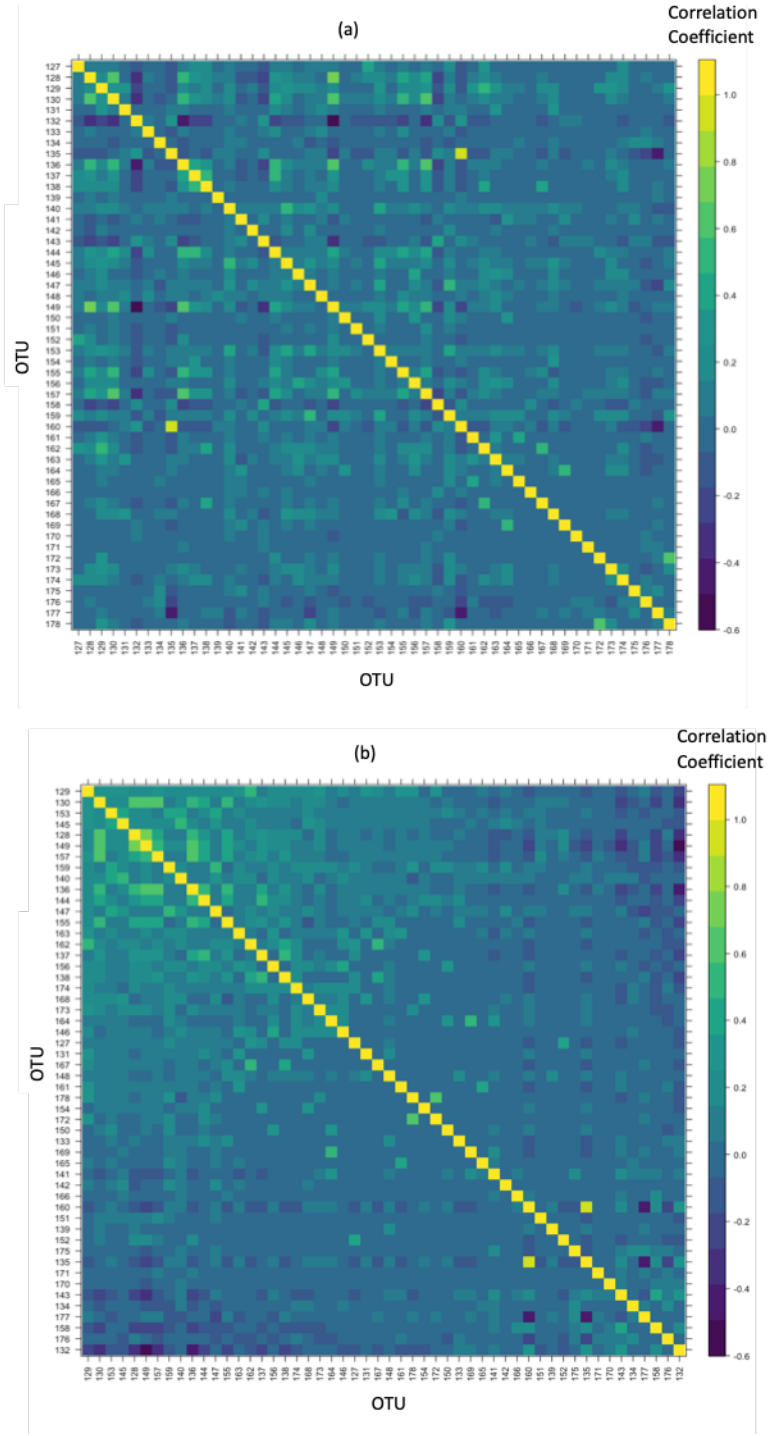
Supplementary Figure 14: Heatmaps for the Spearman rank of the OTUs in the cluster, Phylum Proteobacteria, (a) before ordering and (b) after the ordering based on correlation of the OTUs in the T2D study



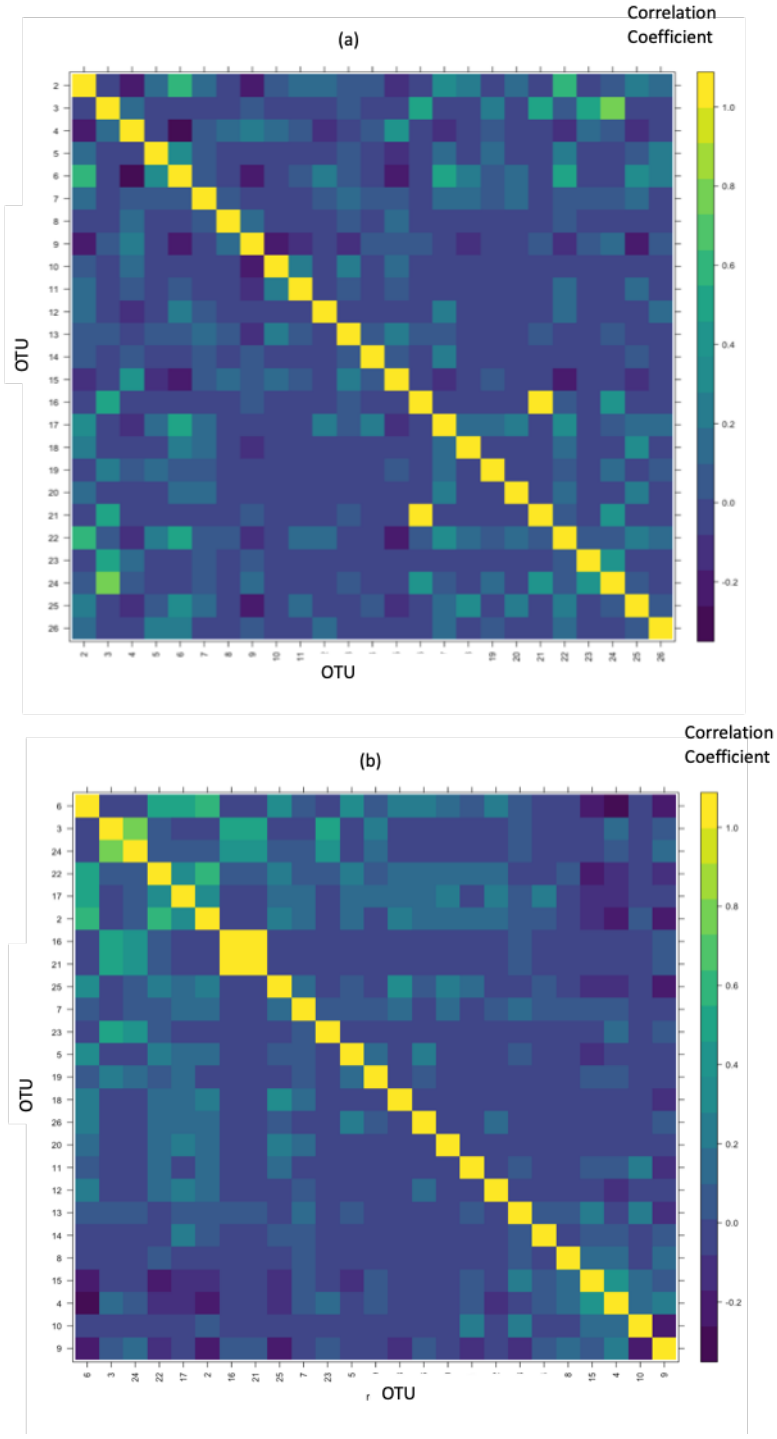
Supplementary Figure 15: Heatmaps for the Spearman rank of the OTUs in the cluster, Phylum Actinobacteria, (a) before ordering and (b) after the ordering based on correlation of the OTUs in the T2D study



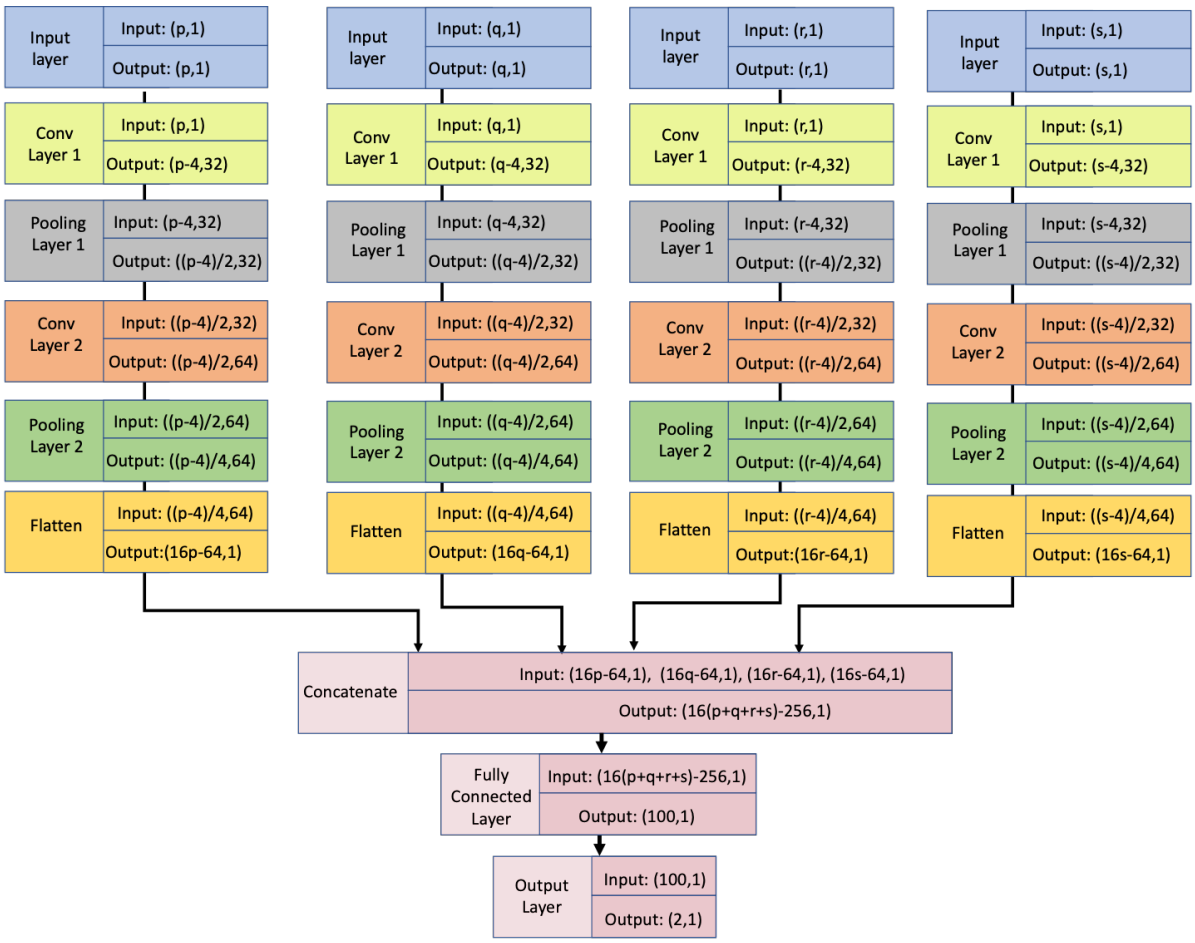
Supplementary Figure 16: Heatmaps for the Spearman rank of the OTUs in the cluster, Phylum Firmicutes, (a) before ordering and (b) after the ordering based on correlation correlation of the OTUs in the Cirrhosis study



Supplementary Figure 17: Heatmaps for the Spearman rank of the OTUs in the cluster, Phylum Proteobacteria, (a) before ordering and (b) after the ordering based on correlation correlation of the OTUs in the Cirrhosis study



Supplementary Figure 18: Heatmaps for the Spearman rank of the OTUs in the cluster, Phylum Actinobacteria, (a) before ordering and (b) after the ordering based on correlation correlation of the OTUs in the Cirrhosis study



Supplementary Figure 19: Functional working of the layers of $taxoNN$ on 4 clusters of an example dataset containing 'p', 'q', 'r' and 's' OTUs in the respective clusters (where p+q+r+s = N). Each block corresponds to a layer acting on the cluster. Input signifies the dimension of the input to the layer. The input at each step is represented as (k,l) where, 'k' is the number of rows in the input and 'l' represents the number of columns. As the initial input was a vector therefore, l in this case was '1'. Output signifies the dimension of the result after certain operations in that particular layer. Further, as the number of filters increases from 32 in the first Conv layer to 64 in the second Conv layer, the number of columns in the nodes vary from 32 to 64. Finally, in the concatenation step we obtain a single column concatenation vector by stacking flattened vectors from all clusters together.

3 References

[1] Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55–60 (2012).

[2] Qin, N. *et al.* Alterations of the human gut microbiome in liver cirrhosis. *Nature* **513**, 59–64 (2014).

[3] Turpin, W. *et al.* Association of host genome with intestinal microbial composition in a large healthy cohort. *Nature Genetics* **48**, 1413–1417 (2016).

[4] Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 1097–1105 (2012).

[5] Karlsson, F. H. *et al.* Gut metagenome in european women with normal, impaired and diabetic glucose control. *Nature* **498**, 99–103 (2013).