Integrative Data Analysis for the Prediction of Metagenomic Functions

Jyotsna Talreja Wassan

Research conducted within the:
School of Computing,
Faculty of Computing, Engineering and the Built Environment

This dissertation is submitted for the degree of
Doctor of Philosophy

March 2020

I confirm that the word count of this thesis is less than 100,000 words excluding the title page, contents acknowledgements, summary or abstract, abbreviations, footnotes, diagrams, maps, illustrations, tables, appendices, and references or bibliography.
Declaration

I hereby declare that with effect from the date on which the thesis deposited in the Library of the University of Ulster, I permit the Librarian of the University to allow the thesis to be copied in whole or in part without reference to me on the understanding that such authority applies to the provision of single copies made for study purposes or inclusion within the stock of another library. This restriction does not apply to the British Library Thesis service (which is permitted to copy the thesis on demand for loan or sale under the terms of a separate agreement) nor to the copyright or publication of the title and abstract of the thesis. IT IS A CONDITION OF USE OF THIS THESIS THAT ANYONE WHO CONSULTS IT MUST RECOGNISE THAT THE COPYRIGHT RESTS WITH THE AUTHOR AND THAT NO QUOTATION FROM THE THESIS AND NO INFORMATION DERIVED FROM IT MAY BE PUBLISHED UNLESS THE SOURCE IS PROPERLY ACKNOWLEDGED”
This thesis is dedicated to

almighty God who blessed my life spiritually, emotionally, and relationally during the journey of PhD;

beloved Parents;

&

caffeine (my Husband) and sugar (my Son), the best companions very much needed during this journey. My Husband and little Son have shown that real love is based on respect, compromise, care and trust.
Microbes Matter and so do our lives

Microbial genes that we carry are the main reason on which our life drives

These tiny creatures make us who we are

The mystery behind the Microbial realm is a bizarre

Let’s explore it…

Jyotsna Talreja Wassan
Table of Contents

CHAPTER 1 ........................................................................................................... 1
1.1 Introduction ................................................................................................. 1
1.2 Metagenomics: Basic concepts ...................................................................... 4
    1.2.1 Metagenomics pipeline ......................................................................... 5
1.3 Machine learning for downstream analysis of metagenomes ................9
1.4 Microbial phylogeny ..................................................................................... 12
1.5 Metagenomic functions ................................................................................ 18
    1.5.1 Human microbiome studies .................................................................. 19
    1.5.2 Cattle rumen microbiome studies ...................................................... 21
1.6 Problem domain under investigation ......................................................... 21
    1.6.1 Key characteristics of microbiome data .............................................. 22
1.7 Research focus ............................................................................................. 23
    1.7.1 Rationale and Objectives ...................................................................... 24
1.8 Key contributions of current research ......................................................... 26
1.9 Overview of the thesis .................................................................................. 30
1.10 Publications ................................................................................................. 33
1.11 Summary ..................................................................................................... 36

CHAPTER 2 ........................................................................................................... 37
2.1 Introduction .................................................................................................. 37
2.2 Approaches to infer metagenomic functions using supervised learning ........................................................................... 40
    2.2.1 Abundance-aware methods for downstream analysis of metagenomes ................................................................................ 41
    2.2.1.1 Random Forest (RF) ........................................................................ 42
    2.2.1.2 Support Vector Machines (SVM) .................................................... 44
2.2.1.3 Elastic Net (ENet) ......................................................... 47
2.2.1.4 Multilayer Perceptron (MLP) ...................................... 49
2.2.2 Phylogeny-aware methods for downstream analysis of metagenomes ................................................................. 55
  2.2.2.1 Homology-based approach ........................................ 56
  2.2.2.2 Supervised model-based approaches ......................... 58
  2.2.2.3 Rank-based approach .............................................. 63
  2.2.2.4 Feature space transformation-based approach ............. 64
  2.2.2.5 Biostatistics-based approach .................................. 66
  2.2.2.6 Subcategorization of different phylogenetic approaches .. 68
2.3 Feature selection in metagenomic studies ......................... 72
2.4 Current limitations of predictive approaches and future scope .......................................................................................... 73
2.5 Summary ............................................................................. 79
CHAPTER 3 .................................................................................. 83
3.1 Conceptual framework .......................................................... 83
  3.1.1 Description and presentation of data inputs ..................... 85
    3.1.1.1 Abundance-based profiling with OTU counts ............ 87
    3.1.1.2 Knowledge-based profiling ..................................... 88
  3.1.2 Data pre-processing ......................................................... 90
  3.1.3 Data analysis .................................................................. 98
  3.1.4 Computational strategies to assess supervised ML models 100
3.2 Possible research approaches formalised based on the conceptualised framework ......................................................... 106
3.3 Data sources ........................................................................ 110
3.4 Summary ............................................................................. 117
CHAPTER 4 .................................................................................. 119
4.1 Rationale ............................................................................. 120
4.2 Methodology ......................................................................... 121
4.3 Implementation details ................................................................. 124
4.4 Experiments and results ................................................................. 125
  4.4.1 Abundance-driven analysis of DS5 and DS6 metagenomic data sources................................................................. 126
    4.4.1.1 Discussion of results obtained over the abundance-driven analysis of DS5 ................................................................. 129
    4.4.1.2 Discussion of results obtained over the abundance-driven analysis of DS6 ................................................................. 136
4.4.2 A comparison with an extant study ........................................... 141
4.4.3 A study of dynamics of the embedded methods and their impact in current research ................................................................. 143
4.5 Key Findings ................................................................................. 146
4.6 Limitations .................................................................................... 150
4.7 Summary ....................................................................................... 151

CHAPTER 5 ........................................................................................ 154
5.1 Rationale ....................................................................................... 155
5.2 Methodology ................................................................................ 156
  5.2.1 Construction of PAAM feature space incorporating biological domain knowledge of phylogeny and abundance count profiles 159
  5.2.2 Feature selection over the hierarchical feature space of PAAM .................................................................................. 163
  5.2.3 Supervised ML over selected features from PAAM feature space .................................................................................. 164
  5.2.4 Assessment of supervised ML algorithms ................................ 165
5.3 Implementation of PAAM-ML ....................................................... 166
5.4 Predictivity of microbial features with PAAM construction 168
5.5 Experiments and results ................................................................. 170
  5.5.1 Distinguishing microbial profiles using abundance and phylogeny-driven Measures ................................................................. 171
5.5.2 Classification of the microbiome using feature selection strategy of RFI ................................................................. 172
5.5.3 Classification of the microbiome using feature Selection strategy of XGB-pen-LR-Classification .............................. 177
5.5.4 Classification of the microbiome using feature selection strategy of correlation-based filtering (CFS) ...................... 180
5.5.5 A comparison with other phylogeny-aware supervised ML Methods ........................................................................... 182
5.6 Discussions ............................................................................................................................................................. 186
5.7 Summary ................................................................................................................................................................. 189
CHAPTER 6 ................................................................................................................................................................. 193
6.1 Introduction .............................................................................................................................................................. 194
6.2 Methodology and implementation ............................................................................................................................ 196
6.3 Experimental results and discussions ........................................................................................................................ 205
6.3.1 Experimental results .............................................................................................................................................. 205
6.3.2 Discussions ............................................................................................................................................................. 218
6.4 Summary ................................................................................................................................................................. 221
CHAPTER 7 ................................................................................................................................................................. 224
7.1 Introduction .............................................................................................................................................................. 225
7.2 A use case: Determine functional phenotype of the cattle rumen microbiome ............................................................ 226
7.2.1 MetaPlat project .................................................................................................................................................... 227
7.2.2 MetaPlat data sources .......................................................................................................................................... 228
7.2.3 Integrative analysis of MetaPlat data using proposed methods of PAAM-ML and Phylogeny-RF ......................... 230
7.2.3.1 Analysis based on the application of PAAM-ML to differentiate supplemented diet phenotype ......................... 230
7.2.3.2 Analysis based on the application of Phylogeny-RF to differentiate supplemented diet phenotype ......................... 234
7.2.3.3 Phylogeny-aware analysis of circulating glucocorticoid on the cattle rumen microbiome with forage or concentrated diet ..235

7.2.4 Additional contributions towards a comprehensive analysis of MetaPlat generated cattle rumen data..........................240

7.2.4.1 A study conducted using Approach 1 (abundance-driven) over 40 samples of cattle rumen ................................................240

7.2.4.2 A study conducted using Approach 2 over 80 samples of cattle rumen .................................................................245

7.2.5 Contributory scripts ..................................................................251

7.3 Key findings ............................................................................252

7.4 Limitations ............................................................................254

7.5 Summary .................................................................................254

CHAPTER 8 ......................................................................................257

8.1 Introduction ............................................................................257

8.2 Contribution to the knowledge ................................................258

8.2.1 A comprehensive assessment of predictive models for inferring phenotypes related to two use cases of the human microbiome (DS5,6) [Chapter 4].............................................262

8.2.2 Proposal and application of a novel PAAM-ML framework for the prediction of metagenomic functions [Chapter 5] ..........264

8.2.3 Phylogeny-aware RF model-based metagenomic predictions [Chapter 6] ..................................................................267

8.2.4 Application of data analysis models for the prediction of cattle microbiome functions in MetaPlat project [Chapter 7] ........268

8.2.5 Development of contributory tools .........................................270

8.3 Limitations and future work .....................................................270

8.3.1 Limitations ..........................................................................271

8.3.1.1 Compositional nature of metagenomic data that was not addressed by this research ................................................272

8.3.1.2 Constructing signatures at multiple levels in an integrative way in an ML model .......................................................273
8.3.1.3 Some other limitations embraced by the methods developed in this thesis. ................................................................. 273

8.3.2 Future work .............................................................................................................................................................. 275

8.3.2.1 Extension of study based on deep ML models over phylogeny and abundance-aware analysis ........................................... 278

8.3.2.2 An application of dimensionality reduction and visualisation techniques to an enhanced study of the microbiome ............... 278

8.3.2.3 Incorporation of node-by-node information in a tree-based ML modelling ........................................................................ 279

8.3.2.4 Compositional effect of metagenomes .................................................. 279

8.4 Conclusions ................................................................................................................................................................. 280

References ........................................................................................................................................................................ 282
List of Figures

Figure 1.1 The basic principle involved in the emerging field of metagenomics. .............................................................3
Figure 1.2 An overview of the typical metagenomics pipeline related to the 16SrRNA gene sequencing. ..............................6
Figure 1.3 A universal phylogeny (tree of life) of the living world obtainable by comparative sequencing of DNA. The evolutionary measure between two micro-organisms is proportional to the cumulative distance(length) of the branches that join two groups of organisms. .............................................................................13
Figure 1.4 An illustrative representation of a detailed sample phylogenetic tree. ..............................................................15
Figure 1.5 An illustration of UniFrac Distance measure between two samples (shown as pink and blue microbial samples). ..........17
Figure 1.6 An illustration of possible ways to integrate phylogeny in predictive ML modelling. .............................................18
Figure 1.7 4D’s highlighting major factors playing an important role in key contributions made in this thesis. ..........................27
Figure 1.8 Mapping of key contributions to chapter-wise distribution in the scope of the current thesis. .................................30
Figure 2.1 An illustrative scheme of functional inference from metagenomic environments. ..................................................38
Figure 2.2 A diagrammatic illustration of an RF Classifier classifying input samples into two Classes X and Y with the help of ‘t’ decision trees. .....................................................................................................44
Figure 2.3 A diagrammatic representation of state-of-the-art SVM classifier differentiating two classes. .................................46
Figure 2.4 A diagrammatic illustration of MLP Classifier with two hidden layers. .................................................................50
Figure 2.5 Categorical differentiation of phylogenetic tools for metagenomics downstream analysis. .................................56
Figure 2.6 Phylogeny-aware approaches characterised into three different ways to integrate phylogeny in predictive ML modelling.69
Figure 2.7 Timeline of Popular phylogeny-aware approaches; PICRUSt [134], MetaPhyl [87], A study by Ning & Beiko [98], PhyloRelief [92], aMiSPU [149], PhILR [133], DL methods such as PopPhy-CNN [142], A study by Wang et al. [85]; used in predictive ML modelling of functional metagenomes with useful insights and limitations. ...........................................................................71

Figure 2.8 Feature selection methods dealing with filter, wrapper and embedded approaches.................................................................73

Figure 2.9 A Comparative Analysis with Kappa Statistic (κ) [177] between i) RF applied over raw abundances and ii) RF applied over PICRUSt scaled data; in the Study by Xu et al., 2014 [131]; over public datasets CBH, CS, CSS, FSH and FS used in [69]. .................78

Figure 2.10 Areas addressed in the current dissertation based on the observations made from the literature..............................................81

Figure 3.1 Graphical representation of the methodological framework for the prediction of metagenomic functions. (a) Interpretation of heterogeneous data of abundance count table and the phylogenetic tree, being sourced from metagenomic pipelines (b) Feature engineering is performed to identify important taxa. (c) Selected features are input to classification algorithms for predicting metagenomic functions (d) The predictive performance is evaluated using different assessment measures..............................................84

Figure 3.2 An illustration of a binary rooted phylogenetic tree.......89

Figure 3.3 Feature selection methods falling within each category utilised in this thesis across various studies.................................92

Figure 3.4 An example to discuss Kappa and Accuracy over class imbalances in classifying metagenomic input dataset. .................103

Figure 3.5a A graphical representation of Approach 1 indicating possible workflow for abundance-driven analysis. ......................107

Figure 3.5b A graphical representation of Approach 2 indicating possible workflow for phylogeny-driven analysis. ......................108

Figure 3.5c A graphical representation of Approach 3 indicating possible workflow for phylogeny-driven ML modelling. ..............109

Figure 4.1 A generalised abundance-driven methodological workflow for functional metagenomics..............................................123
Figure 4.2 Selecting top features from glmnet Model (α = 0.3) over DS5 by measuring AUC-ROC. .......................................................... 132

Figure 4.3 Statistical analysis conducted for measuring the difference between Accuracies obtained over 10-fold of validation using SVM, SVMXGPLR (SVM applied over feature-set obtained from XGBoost and Pen-LR using glmnet), and SVMPLR (SVM applied over Pen-LR using glmnet). .......................................................... 134

Figure 4.4 Some important taxa identified with EM-based feature selection over DS5 playing an important role in associating the human diet with the microbiome. .......................................................... 136

Figure 4.5 Analysis conducted for measuring the difference of significance between AUC-ROC obtained over 10-fold of cross-validation using SVM (employed directly over the abundance profiles, SVMXG (SVM employed over feature-set obtained from XGBoost), RF (employed directly over the abundance profiles, and RFXG (RF employed over feature-set obtained from XGBoost). 140

Figure 4.6 Important taxa identified with EM-based Feature selection over DS6 playing an important role in classifying IBD diseased state of the human microbiome. 141

Figure 4.7 Comparing RF_RF and RF_XGBoost in terms of time and predictive performance over DS5 and DS6. 142

Figure 4.8a Comparative results of predictive performance obtained before and after application of Feature Selections with RF and SVM over DS5 and DS6. .......................................................... 144

Figure 4.8b Comparative Results of execution time (in secs) over DS5 and DS6. .......................................................... 145

Figure 4.9 Comparative results of Predictive performance obtained by EM methods as classifiers over the abundance profiles, and; their application as feature selectors precursor to models of RF and SVM over DS5 and DS6. .......................................................... 146

Figure 4.10 A summary of the proposed ensembled approach XGB-pen-LR-Classification as a contribution made from this study. .......................................................... 149

Figure 4.11 A summary of key observations, strengths and limitations of this study.......................................................... 153
Figure 5.1 Graphical representation of the PAAM-ML prediction framework. (A) heterogenous inputs of abundance count and prior knowledge of phylogenetic relationships are combined for the construction and engineering of the phylogeny and abundance-aware Matrix (PAAM). (B) Feature ranking or selection is performed over the PAAM structure. (C) supervised ML is applied over selected features. (D) assessment of the predictive performance.

Figure 5.2 An illustration of the construction of the PAAM feature space involving internal (ancestral) nodes.

Figure 5.3 An Illustration of the GUI interface of the PAAM generation. This screenshot demonstrates the selection of abundance count table and phylogenetic tree to attain PAAM for the downstream analysis.

Figure 5.4 An Illustration of a toy example of how PAAM feature space could result in better classification of functional metagenomes.

Figure 5.5 An illustration of the comparison of variation in performance of Pen-LR applied over OTU count table and PAAM (DS1).

Figure 5.6 An illustration of the comparison of variation in performance of three phylogeny-aware models.

Figure 5.7 Graphical representation of the generative PAAM-ML prediction framework proposed as a useful tool in functional metagenomics.

Figure 5.8 A summary of key observations, strengths and limitations of this study.

Figure 6.1 A graphical representation of the stepwise procedure followed in the construction of the proposed new Phylogeny-RF classifier for classifying microbiome into functional phenotype.

Figure 6.2a A graphical representation with Box Plots indicating variation in AUC-ROC obtained over 3 trial runs of 5-fold cross-validation with (i) Phylogeny-RF over DS4 (ii) traditional RF over DS4.
Figure 6.2b A graphical representation with Box Plots indicating variation in AUC-ROC obtained over 3 trial runs of 5-fold cross-validation with (i) Phylogeny-RF over DS1 and (ii) traditional RF applied over DS1. .......................................................... 210

Figure 6.3 A comparison of Phylogeny-RF, RF, and RF applied over PhILR transformed data over different numbers of trees (1-500). ........................................................................................................ 215

Figure 6.4 A summary of key observations, strengths and limitations of this study.......................................................................................................................... 223

Figure 7.1 Results obtained from PAAM-ML, PhILR and MetaPhyl over 80 cattle rumen samples relating to diet phenotype. ......................................................... 234

Figure 7.2 Cattle rumen samples differentiation as coloured by the (a) DEX treatment and (b) Diet (PCoA plot, as shown in this Figure). .................................................................................................. 239

Figure 7.3 A Computational approach followed to conduct a case study for functional analysis of cattle rumen microbiome associated with 40 samples and diet phenotype. ...................................................... 241

Figure 7.4 Results obtained from the preliminary case study over 40 cattle rumen samples. ................................................................................................. 243

Figure 7.5 A summary of the stepwise procedure of Phylogeny-PINA Relief approach. ........................................................................................................ 249

Figure 7.6 A summary of key observations, strengths and limitations of this study................................................................................................................ 256

Figure 8.1 High-level pipeline design to classify microbiome (genome structure) into functional phenotypes (genome function). .......................... 260

Figure 8.2 4D’s highlighting major contributory factors in the context of the current thesis. ............................................................................................... 262

Figure 8.3 A proposed path for combining Approaches 2 and 3. .. 277
List of Tables

Table 1.1 An overview of methods for constructing phylogeny. The first column states the name of the method followed by a description [42], [43]. ................................................................. 13

Table 1.2 An example of phylogenetic similarity distance matrix corresponding to a sample tree shown in Figure 1.4. ....................... 16

Table 2.1 A review of studies involving predictive modelling to infer metagenomic functions using abundances of OTUs and reference to each study is listed; data source, related phenotypes, number of samples, taxonomic level, and size of feature space is mentioned; different classifiers used in respective study are listed, and best predictive model is reported over the benchmarks....................... 52

Table 2.2 The widely used phylogeny-aware methods in each category (Figure 2.5) to infer metagenomic functions are listed. Reference to each study is listed with their implementation source; datasets used are summarised, performance and various insights are reported over the data in use........................................................................ 74

Table 3.1 A list of popular tools from the literature that is useful for data encoding of 16S rRNA data............................................. 86

Table 3.2 An illustration of an OTU count table with phenotype. ... 87

Table 3.3 An illustration of a confusion matrix depicting two classes positive and negative............................................................ 100

Table 3.4 An illustration of an ANOVA table where SS represents the sum of squares; df is the degree of freedom and MS is mean squares, and F is the f-statistics............................................................ 105

Table 3.5 A summary of metagenomic datasets used in this thesis for functional analyses.............................................................. 116

Table 4.1 A summary of data sources (with predictors >> the number of samples) used in the current study............................................ 126

Table 4.2 Performance of various ML models (with/without feature selection) employed over DS5 (associated with human diet phenotype) with 10-fold cross-validation. NFS represents the number of features. Best performance attained is bold-faced......................... 130


Table 4.3 Performance of various ML models employed over DS6 (phenotype of IBD states associated with human microbiome) with 10-fold cross-validation. NFS represents the number of features. Best performance attained is bold-faced. .......................................................... 137

Table 4.4 Performance of RF over RF (i.e., RF_RF) method with 10-fold cross-validation over DS5 and DS6.......................................................... 142

Table 5.1 Summary of performance results obtained by the application of RF and SVM classifiers over PAAM and OTU abundance matrix obtained from different metagenomic data sources with cross-validation. Results are obtained by choosing the best value after resampling results across different tuning parameters over different folds (in caret package). .......................................................... 172

Table 5.2 Summary of performance results obtained by application of SVM over the feature space obtained after RFI ranking of features in PAAM and OTU abundance count of DS1, dealing with human throat microbiome. Results are averaged over all LOOCV folds by choosing the best value after resampling results across different tuning parameters. .......................................................... 173

Table 5.3 Summary of performance results obtained by the application of SVM over the feature space obtained after RFI ranking of features in PAAM and OTU abundance count of DS2, dealing with human microbiome (body sites). Results are averaged over all 10-fold by choosing the best value after resampling results across different tuning parameters.................................................. 174

Table 5.4 Summary of performance results obtained by application of SVM over the feature space obtained after RFI ranking of features in PAAM and OTU abundance count of DS4, dealing with soil microbiome (substrate treatment). results are averaged over all LOOCV folds by choosing the best value after resampling results across different tuning parameters.................................................. 175

Table 5.5 Summary of performance results obtained by application of SVM over the number of features (NFS) obtained after XGBoost and/or Pen-LR ranking of features in PAAM and OTU abundance count of DS1, dealing with human throat microbiome. Results are averaged over all folds and choosing the best value after resampling results across tuning parameters. .................................................. 178
Table 5.6 Summary of performance results obtained by application of SVM over the number of features (NFS) obtained after XGBoost and/or Pen-LR ranking of features in PAAM and OTU abundance count of DS2, dealing with human body sites. Results are averaged over all folds and choosing the best value after resampling results across tuning parameters.

Table 5.7 Summary of performance results obtained by application of SVM over the number of features (NFS) obtained after XGBoost and/or Pen-LR ranking of features in PAAM and OTU abundance count of DS4, dealing with soil microbiome. Results are averaged over all folds and choosing the best value after resampling results across tuning parameters.

Table 5.8 Summary of performance results obtained by application of RF, SVM over the number of features (NFS) obtained after CFS employed over PAAM and OTU abundance count of DS1, DS2 and DS4. Results are averaged over all folds and choosing the best value after resampling results across different tuning parameters.

Table 5.9 Summary of performance results obtained by application of PAAM-ML (represented as PAAM_featureselection_classifier) and MetaPhyl employed over DS1, DS2 and DS4.

Table 5.10 Summary of performance results obtained by application of PAAM-ML (represented as PAAM_featureselection_classifier) and ML (RF/SVM) employed over PhILR transformed data obtained from DS1, DS2 and DS4.

Table 6.1 Summary of datasets used in the current chapter for functional analyses.

Table 6.2 Summary of average performances obtained by application of Phylogeny-RF and RF with 5-fold (3 trial runs) cross-validation dealing with the different number of trees in RF over DS4 relating to soil microbiome (best values attained in each column are highlighted in bold in each category).

Table 6.3 Summary of average performances obtained by application of Phylogeny-RF and RF with 5-fold (3 trial runs) cross-validation dealing with a different number of trees in RF over DS1 relating to human throat microbiome (best values attained in each column are highlighted in bold in each category).
Table 6.4 Summary of average performances obtained by application of Phylogeny-RF and RF with 5-fold cross-validation dealing with a different number of trees in RF over DS3 relating to human body sites (Costello body habitats) microbiome (best values attained in each column are highlighted in bold in each category).

Table 6.5 Summary of performance results obtained by application of Phylogeny-RF with 5-fold cross-validation and MetaPhyl with default settings over DS1, DS3, DS4.

Table 6.6 A comparison of average performance results obtained by application of Phylogeny-RF and classical RF applied over PhILR transformed data with 5-fold cross-validation (3 trials) on DS4 (soil microbiome) (best value in each column is bold-faced).

Table 6.7 A comparison of average performance results obtained by application of Phylogeny-RF and classical RF applied over PhILR transformed data with 5-fold cross-validation (3 trials) on DS1 (Throat microbiome) (best value in each column is bold-faced).

Table 6.8 A comparison of average performance results obtained by application of Phylogeny-RF and classical RF applied over PhILR transformed data with 5-fold cross-validation on DS3 (Costello Body Habitats) (best value in each column is bold-faced).

Table 6.9 Results of ANOVA and post hoc analysis employed over AUC-ROC values obtained over a different number of trees in RF (applied over the abundance counts), Phylogeny-RF (PhyRF) and RF employed over PhILR (PHILRF) over DS1.

Table 6.10 Results of ANOVA and post hoc analysis employed over AUC-ROC values obtained over the different number of trees in RF (applied over the abundance counts), Phylogeny-RF (PhyRF) and RF employed over PhILR (PHILRF) over DS3.

Table 6.11 Results of ANOVA and post hoc analysis employed over AUC-ROC values obtained over the different number of trees in RF (applied over the abundance counts), Phylogeny-RF (PhyRF) and RF employed over PhILR (PHILRF) over DS4.

Table 7.1 Summary of datasets sampled as part of MetaPlat project and used in current analyses.
Table 7.2 Predictive performance of ML models employed over PAAM and ASV abundance profiles of cattle rumen dataset using LOOCV settings. The dataset was characterised into 4 functional categories of diet (oil, nitrate, oil-nitrate combined, and controls). The best performance in each column is bold-faced.

Table 7.3 Summary of predominant features reported by RFI method in current analyses.

Table 7.4 Predictive performance of Phylogeny-RF (with 5-fold cross-validation) over cattle microbiome rumen to differentiate samples (best performance is bold-faced in each column).

Table 7.5 Predictive performance of PAAM-ML (with LOOCV settings) over cattle microbiome rumen to differentiate 118 cattle rumen samples (best performance is bold-faced).

Table 7.6 A summary of important features selected by PAAM-ML method for classifying diet phenotype associated with cattle microbiome.

Table 7.7 Predictive performance of Phylogeny-RF with 5-fold cross-validation over cattle microbiome rumen to differentiate samples (best performance in each column is bold-faced).

Table 7.8 A summary of predominant features selected by WFS method for classifying diet phenotype associated with cattle microbiome.

Table 7.9 Predictive performance of state-of-the-art RF model applied over (a) relief-selected features from PTAM, (b) PINA-driven relief selected features from PTAM, (c) UniFrac-driven relief selected features from PTAM and (d) relief-selected features from PAAM; with LOOCV settings. the dataset was characterised into 4 functional categories of diet (oil, nitrate, oil-nitrate combined, and controls). The best performance in each column is bold-faced.

Table 7.10 Summary of integrative methods employed over the MetaPlat data.
Acknowledgements

This thesis is the result of three years of research work conducted at the School of Computing, Ulster University, Jordanstown, Northern Ireland. During this time, I have had the privilege of working with many domain experts and very supportive people. First and foremost, I would like to thank my wonderful supervisors Professor Huiru Zheng, Dr. Haiying Wang and Dr. Fiona Browne for their invaluable supervision throughout the PhD process. They guided me with pragmatic approaches to research and gave me a wonderful opportunity to work on the MetaPlat Project related to an emerging and exciting field of Metagenomics. Their valuable discussions on creative ideas, conversations regarding algorithmic techniques, and suggestions have made vital contributions to this research. Without their support, at both the academic and personal levels, I would not have completed my research studies. I would like to thank my progress review panel at stages of confirmation and 100-day viva to help me in building the knowledge base of my research context. I acknowledge Professor Christopher Nugent, Head of School of Computing, for his valuable support throughout and providing useful research training such as on thesis writing. I also thank Professor Hui Wang, Research Director, for his support in approving for various dissemination opportunities. I also thank Ulster University School of Computing staff and Doctoral College for their support during my PhD studies. I thank researchers of MetaPlat project for their help, technical discussions, and friendship during my secondment periods whilst attaining goal of the project. I gratefully acknowledge Research Strategy Fund of Ulster University and the MetaPlat project, (www.metaplat.eu), under H2020-MSCA-RISE scheme for providing generous financial support of this PhD.

I would also like to thank my colleagues and friends at Ulster University for their warm friendship, support, humour and social arrangements.

My deepest gratitude goes to my family and friends. To my husband especially for his patience, encouragement and most of all his understanding during this PhD journey by living in a different continent alone; and my son who travelled with me to Northern Ireland at the age of 2 years but provided me with constant love, motivation and unforgettable moments in this whole journey. Last but not the least I pay all my respects to my beloved parents for their blessings and shower unconditional love. I owe a lot to my husband, son and parents.
Abstract

The emergence of High-throughput sequencing (HTS) techniques has revolutionised the field of “Metagenomics” which deals with studying the genomic structure and function of uncultured microbial communities in an ecosystem. The field helps in understanding the composition, diversity and functioning of complex microbial communities. The outcome of sequencing is large, complex, heterogeneous, sparse and biologically rich metagenomic datasets. The unprecedented advances in sequencing have necessitated the development of computational methods for analysing such data, thereby reducing the computational costs and increasing the predictive performance of methods. This thesis has applied Machine Learning (ML) techniques to address the task of computationally inferring functions associated with the genes present in microbial communities (in humans, cattle and soil). The aim of this research is twofold, dealing with investigating, developing, and evaluating ML classification approaches for: (i) abundance-driven analyses, and; (ii) phylogeny-driven analyses of microbial genomes in an integrative way. The current thesis has utilized embedded ML techniques to detect and classify microbiome into functions dealing with its high-dimensional and sparse nature and informing the development of a new abundance-driven framework (Chapter 4). The integrative approaches take advantage of the biological evolutionary characteristics (i.e. phylogeny). Phylogenetically similar microbial species could share similar characteristics and henceforth similar functional traits. The novel integrative approaches involving modelling over phylogeny and abundance profiles are proposed to predict metagenomic functions effectively with a key idea of integration of phylogeny at either at the data pre-processing level as a precursor to ML model (Chapter 5) or in an ML model itself (Chapter 6). An additional case study involving the prediction of functions in cattle microbial genes have been presented in this thesis (linked to MetaPlat¹, European Commission Project) (Chapter 7). The thesis includes key contributions, novel findings, limitations in the current context and future work with a summary.

¹ http://www.metaplat.eu/
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>16S rRNA</td>
<td>16S ribosomal Ribonucleic Acid</td>
</tr>
<tr>
<td>aMISPU</td>
<td>adaptive Microbiome-based Sum of Powered Score Statistical test</td>
</tr>
<tr>
<td>AN</td>
<td>Actual Negative</td>
</tr>
<tr>
<td>ANOVA</td>
<td>Analysis of Variance</td>
</tr>
<tr>
<td>AP</td>
<td>Actual Positive</td>
</tr>
<tr>
<td>ASV</td>
<td>Amplicon Sequence Variant</td>
</tr>
<tr>
<td>AUC-ROC</td>
<td>Area Under Receiver Operator Characteristics Curve</td>
</tr>
<tr>
<td>BIOM</td>
<td>Biological Observation Matrix</td>
</tr>
<tr>
<td>BLAST</td>
<td>Basic Local Alignment Search Tool</td>
</tr>
<tr>
<td>CFS</td>
<td>Correlation-based Feature Selection</td>
</tr>
<tr>
<td>CD</td>
<td>Crohn's Disease State</td>
</tr>
<tr>
<td>CDF</td>
<td>Cascade Deep Forest</td>
</tr>
<tr>
<td>CNN</td>
<td>Convolution Neural Networks</td>
</tr>
<tr>
<td>DL</td>
<td>Deep Learning</td>
</tr>
<tr>
<td>DMs</td>
<td>Dirichlet Multinomial Regression Models</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic Acid</td>
</tr>
<tr>
<td>DS1</td>
<td>Data Source 1</td>
</tr>
<tr>
<td>DS2</td>
<td>Data Source 2</td>
</tr>
<tr>
<td>DS3</td>
<td>Data Source 3</td>
</tr>
<tr>
<td>DS4</td>
<td>Data Source 4</td>
</tr>
<tr>
<td>DS5</td>
<td>Data Source 5</td>
</tr>
<tr>
<td>DS6</td>
<td>Data Source 6</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>DS7</td>
<td>Data Source 7</td>
</tr>
<tr>
<td>EM</td>
<td>Embedded Methods</td>
</tr>
<tr>
<td>ENet</td>
<td>Elastic Net</td>
</tr>
<tr>
<td>FFS</td>
<td>Filter-based Feature Selection Strategies</td>
</tr>
<tr>
<td>FN</td>
<td>False Negative</td>
</tr>
<tr>
<td>FP</td>
<td>False Positive</td>
</tr>
<tr>
<td>HSD</td>
<td>Tukey-test-Honest-Significant-Difference</td>
</tr>
<tr>
<td>HTS</td>
<td>High-throughput Sequencing</td>
</tr>
<tr>
<td>IBD</td>
<td>Inflammatory Bowel Disease</td>
</tr>
<tr>
<td>IC</td>
<td>Indeterminate Colitis</td>
</tr>
<tr>
<td>IG</td>
<td>Information Gain</td>
</tr>
<tr>
<td>LASSO</td>
<td>Least Absolute Shrinkage and Selection Operator</td>
</tr>
<tr>
<td>LOOCV</td>
<td>Leave-One-Out-Cross-Validation</td>
</tr>
<tr>
<td>LSD</td>
<td>Least Significant Difference</td>
</tr>
<tr>
<td>LR</td>
<td>Logistic Regression</td>
</tr>
<tr>
<td>ML</td>
<td>Machine Learning</td>
</tr>
<tr>
<td>NCBI</td>
<td>National Centre for Biotechnology Information</td>
</tr>
<tr>
<td>OTU</td>
<td>Operational Taxonomic Unit</td>
</tr>
<tr>
<td>PAAM</td>
<td>Phylogeny and Abundance-aware Matrix</td>
</tr>
<tr>
<td>PCoA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>PDM</td>
<td>Phylogenetic Distance Matrix</td>
</tr>
<tr>
<td>Pen-LR</td>
<td>Penalized Logistic Regression</td>
</tr>
<tr>
<td>Phylogeny-RF</td>
<td>Phylogeny driven RF</td>
</tr>
<tr>
<td>PINA</td>
<td>Phylogenetic INteraction-Adjusted indices</td>
</tr>
<tr>
<td>PN</td>
<td>Prediction Negative</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>PP</td>
<td>Prediction Positive</td>
</tr>
<tr>
<td>QIIME</td>
<td>Quantitative Insights into Microbial Ecology</td>
</tr>
<tr>
<td>RBS</td>
<td>Relief-based Measure</td>
</tr>
<tr>
<td>RF</td>
<td>Random Forest</td>
</tr>
<tr>
<td>RFE</td>
<td>Recursive Feature Elimination</td>
</tr>
<tr>
<td>RFI</td>
<td>Random Forest Importance</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machines</td>
</tr>
<tr>
<td>TN</td>
<td>True Negative</td>
</tr>
<tr>
<td>TP</td>
<td>True Positive</td>
</tr>
<tr>
<td>UC</td>
<td>Ulcerative Colitis</td>
</tr>
<tr>
<td>WFS</td>
<td>Wrapper-based Feature Selection Strategies</td>
</tr>
<tr>
<td>XGBoost</td>
<td>eXtreme Gradient Boosting</td>
</tr>
</tbody>
</table>
The pioneer Antonie van Leeuwenhoek designed a microscope for the first time introducing us to the whole world of unseen microorganisms. This revolutionized molecular investigation underpinning the role of microbes in environmental ecology and biological systems. The current thesis is built upon studying the structure and functions of microorganisms intriguing to the biological world. This chapter presents the background information relevant to understanding the contributions of current research. The chapter further provides the main research questions that are addressed in this thesis. It ends with a brief overview of this thesis.

1.1 Introduction

Microorganisms are ubiquitous. Microbial communities are found almost everywhere, such as in living organisms, soil, water, etc. Microbes play a pivotal role in the biological cycles. To leverage the relationships between microbes and their host environment, the development of novel therapeutic approaches targeting microbial genetics has surged in recent years. The combined genetic material of the microorganisms in an ecosystem is known as the microbiome.
[1]. Genomic sequencing is a crucial technology for generating microbiome and studying its impact. The unprecedented development of Next Generation Sequencing (NGS) techniques [2] has accelerated the analysis of genomic sequences of the microbiome to determine their function repertoire. NGS is a more efficient way to sample genomic sequences of uncultivated microorganisms. It supports parallel sequencing and is unlike traditional methods such as the Sanger approach [3], which deals with the cloning of a single Deoxyribonucleic acid (DNA) fragment in laboratory cultivation. NGS has increased the speed and reduced the cost of deep DNA sequencing [2]. Due to this revolutionary breakthrough, a popular “omics” field of Metagenomics [4]-[5] has emerged, which deals with the characterization of abundant gene sequences obtained directly from the whole microbial communities present in an environmental niche (Figure 1.1). Handelsman et al. [6] coined the science of metagenomics. Chen & Pachter [7] defined metagenomics as "the application of modern genomics technique without the need for isolation and lab cultivation of individual species." Metagenomics helps in understanding the composition, diversity, and functions of non-cultured microbial communities to determine their structural and functional capability. The science of metagenomics is revolutionizing the understanding of biological systems through the lens of microbial analysis [8].
Learning over the structure of genes present in microbial communities is critical for understanding their biological functions. As an example, the inhabitant microbial genes in the human body and their patterns of occurrence have indicated connections to diseases such as diabetes, obesity, Inflammatory Bowel Disease (IBD), and vaginosis [9]. The study of microbial genes obtained from a variety of sampling environments has other potential implications in biological domains such as food and nutrition, agriculture, soil nutrition, environmental monitoring, microbial forensics, and production of bio-products [10]. However, metagenomic datasets as sequenced by NGS are large, complex, and are biologically rich. With this plethora of data, the development of therapeutic approaches for analysing metagenomic sequences holds the key to an in-depth and better understanding of the functions of the microbiome. This chapter further provides the background information and some essential biological terminology, involved in the analysis of metagenomic sequences. It
also discusses the ongoing challenges about the inference of metagenomics data, the rationale, objectives, and key contributions of this research.

1.2 Metagenomics: Basic concepts

In this section, a brief overview of the biological concepts involved in microbiome analysis is provided as background information. Microbiome sampled from an environmental site is a collection of millions of biological units (cells) present in unicellular organisms, including bacteria, archaea, eukarya[11]. The shared vocabulary of microbiome research used in this thesis is provided below[11].

- **Microbiota**: The assemblage of microorganisms present in a defined environment, is referred to as the microbiota.

- **Metagenome**: The collection of genomes and genes obtained from constituting organisms present in the microbiota.

- **Microbiome** Metagenome, together with its environmental functions, is known as the microbiome.

- **Functional metagenomics**: A study of host-to-metagenome relationships to infer the association between the microbial community and its functional roles.

- **Descriptive metagenomics**: A study of community structure and variation present in the microbiome.

A series of steps are involved in a typical metagenomics pipeline and are discussed in the next section.
1.2.1 Metagenomics pipeline

A typical metagenomics pipeline can be described by following the workflow of processes [12], as shown in Figure 1.2. The stages are discussed below.

(a) Pre-processing

DNA is extracted from a metagenomic environment, fragmented and amplified to create a DNA library that is fed into the sequencing machine. The massive amount of short sequencing reads generated by sequencing technology is further pre-processed to filter gene sequences by quality and length by identifying and removing any chimeric sequences (or contaminants) using tools and techniques listed in [12]. There are two conventional metagenomic approaches to sequence the microbiome. The first approach is amplicon sequencing of a universal marker gene 16S ribosomal RNA (16S rRNA) for characterizing the microbial composition and its diversity [13]. 16S rRNA is present in all microorganisms, and it is a stable genetic marker with highly conserved regions. This versatile target gene for metagenomics typically has nine variable regions V1-V9 for attaining a fine level of taxonomic classification [14]. These regions establish considerable sequence diversity to classify microorganisms and facilitate sequencing with universal primers. There exist databases of 16S rRNA gene reference sequences and taxonomies, such as Ribosomal Database Project (RDP), Greengenes, National Centre for Biotechnology Information (NCBI), and SILVA [15]. The second approach is to sequence DNA with random primers in the metagenomic samples (rather than the universal markers) and is known as shotgun metagenomic sequencing [13]. Whole genomes are produced based on the assembly of overlapping regions of microbial gene sequences in
such an approach. 16S rRNA is utilised in current research because it is a universally present gene and is a stable genetic marker serving as a reliable molecular clock.

Figure 1.2 An overview of the typical metagenomics pipeline related to the 16SrRNA gene sequencing.

(b) Compositional profiling

This step often creates a microbial profile by assembling the pre-processed sequence reads into 16S genome profiles. The profiling typically is driven majorly by two categories: (a) Taxonomic Profiling and (b) Functional Profiling. The taxonomic profiling maps the sequencing reads to reference 16S
genomes collected from already known gene databases, using standard tools such as Basic Local Alignment Search Tool (BLAST) [16]. The taxonomic composition of the microbiome is obtained by profiling the sequenced genomic data against different levels of the taxonomy (*kingdom, phylum, class, order, family, genus, species*). This taxonomic ancestral relates to the microbial phylogeny [17]. On the other hand, functional profiling includes grouping highly similar microbial sequences (For example, 97% similar sequences) into a functional taxonomic unit known as Operational Taxonomic Unit (OTU) [18]. The term OTU can be used interchangeably with species or microbial genes or taxon. The information obtained from microbiome profiling is then converted to quantitative profiles such as abundance count vectors of mapped gene reads (i.e., how many OTUs are present in a sample) and a tree of microbial phylogeny (i.e., how the OTUs are biologically related to each other) [12]. This derived knowledge is useful in the downstream analysis of the microbiome. A variety of computational tools are available for analysing metagenomic sequences to extract meaningful these compositional profiles from the microbial community. An example of such computation tools includes Mothur [19] and QIMME [20] considering marker 16S rRNA genes and; other tools such as MEGAN [21] and MG-RAST [22] finding whole-genome sequences.

Very recently, a newer alternative method [23],[24] has been proposed for the construction of microbial units that resolve individual sequence differences without grouping of sequences at a similarity threshold. Instead, these intend to determine sequence differences at a single nucleotide level. The pipelines such as QIIME2 employing dada2 [25],[26] supports this kind of analysis to derive functional Units. These units are termed as Amplicon
Sequence Variants (ASVs) [23],[24]. ASVs can be used for identifying microbial composition independently of a 16S genome reference databases. ASVs provide high phylogenetic resolution, supporting deeper levels of phylogeny related to a 16S rRNA gene. However, ASVs may not be as ideal as OTU references for linking microbial composition with its functions at deeper levels of phylogeny by facing difficulty in identifying specific subsets of taxa, that could change in their abundances in the due course of time across the microbial samples. An increased resolution could make the downstream analysis more difficult than the OTU referencing.

The microbial profiles in current research are assumed to be known a priori.

(c) **Downstream analysis**

The downstream analysis forms a promising direction of metagenomic research by uncovering knowledge from the information of compositional profiles obtained in step (b) [12]. Typical questions addressed in downstream are: - "How diverse is the microbiome?"; "What microbes are present?"; "How differentially abundant the microbes are in microbial samples?"; "How microorganisms are related in terms of phylogenetic measures?"; "What metagenomes co-occur in the samples of interest?"; or "How to infer association of microbial genomic composition with its functions?". Microbial abundance count, relationships between microbes by their evolutionary taxonomy (phylogeny), and sample microbe-microbe interactions play an important role in analysing metagenomic functional roles [27]. Hence, downstream analysis is useful in both descriptive and functional metagenomics.
Figure 1.2 provides a graphical overview of the steps involved in a 16SrRNA metagenomics pipeline in a summarised way, and it is included as background information only. The current thesis is primarily focused on developing computational methods to address some of the tasks in the downstream analysis, aiding in inferring metagenomic functions from the microbial profiles. Research in [28] suggested that the use of computation techniques using Machine Learning (ML) [29], could improve our understanding of microbial profiles and their functions. The next section introduces potential ML techniques that are useful for downstream analysis.

1.3 Machine learning for downstream analysis of metagenomes

An increasing pace of development of NGS technologies generates enormous metagenomic data producing challenges in terms of data processing, analysis, and interpretation as well as sequencing quality control [30],[31]. Due to this data-intensive nature and inherent complexity, studying the association of sequenced microbial genes with their functional environment is challenging. However, on the other hand, this provides a significant opportunity for the development of novel computational models using ML [29] techniques for downstream analysis of the microbiome. ML deals with the construction and evaluation of computational algorithms to identify, classify, or predict patterns from the metagenomic data. Hence these techniques aim to make a large amount of metagenomic data comprehensible and usable. There are a variety of ML methods available for the downstream analysis, which could roughly be divided into the following main categories.
• **Predictive Modelling.** It is useful for distinguishing metagenomic samples based on known environmental functions (also known as the phenotype). In predictive modelling, samples of microbial communities with a known phenotype are used to train an ML algorithm; the algorithm identifies discriminative microbial signatures and generates a predictive model that can subsequently be used to predict the phenotypic class of other microbial samples [32]. This kind of ML modelling typically is referred to as “supervised learning” or classification [33].

• **Dimensionality Reduction.** The technique reduces many microbiome attributes (features) into a handful of valuable attributes, with minimal loss of information. This plays a key role in identifying genes or compositions that are most relevant for characterizing metagenomic data and its response to the environmental functions [34].

• **Clustering.** It is useful for grouping microbiome profiles while performing exploratory data analysis to test whether the groups or clusters created, signal the similarities between microbial communities. The approach progresses by calculating a distance matrix obtained by calculating the pairwise similarity between metagenomic samples, and ML algorithms learn a phenotypic label from the statistical regularities or patterns existing in the distance matrix [33]. Unlike the supervised way of learning, there exists no predefined phenotypic label in a clustering-based approach.

• **Visualisation.** It is useful in analysing the diversity of microbiome by projecting a diversity measure into a lower-dimensional representative plane [35]. The diversity measures are typically driven by either (i) alpha diversity that summarised different microorganisms present in a sample or ii) beta
diversity that summarise the differences in microorganisms present across different samples [36]. Conventional approaches to visualise microbiome is the use of one of several ordination methods, such as non-metric multidimensional scaling (NMDS), principal component analysis (PCA), and principal coordinates analysis (PCoA) [37]. This facilitates the top-level view of metagenomic data across different environments.

- **Correlation Analysis.** It is useful for statistical analysis of an overall association between different constituting microbial genes/biological variables present in the composition of a microbial community or their correlation with an outcome of interest (functional phenotype) [38]. This analysis employs univariate or multivariate statistical tests to identify relationships between metagenome and its environment [38],[39] and is practised as one of the ways of predictive modelling.

The research in this thesis provides advanced methods to support the tasks of predictive modelling with the dimensionality reduction phenomena. Detailed literature on such relevant ML tools and techniques useful in downstream analysis of metagenomics is provided in Chapter 2.

It has also been suggested that domain knowledge of microbial phylogeny could be useful in discovering new microbial functions and could support downstream analysis in a biologically meaningful way [40],[41]. OTUs present in microbial samples are not independent units. They are related by the common ancestral of phylogeny [17]. This thesis targets to develop phylogeny informed strategies for predictive modelling of metagenomes. The next subsection describes microbial phylogeny and its related concepts as background information for such integrative microbial analysis.
1.4 Microbial phylogeny

The way of constructing OTUs, by the binning of 16S sequences in metagenomic pipelines, tends to ignore the natural structure present in phylogeny as it considers only the similarity between gene sequences (at a threshold) but not the ancestor-descendant microbial relationships [12]. Hence, downstream analysis based on only raw abundances of OTUs would neglect the biological relationships between them. However, phylogenies serve as an essential tool to structure and understand the microbiome data by uncovering evolutionary ties between microorganisms [17]. The closely related microorganisms tend to have similar genomic sequences, whereas distantly related microorganisms have more dissimilar gene sequences [41],[42]. The possible ways to construct phylogeny from gene sequences from literature are summarised in Table 1.1 [43].

The literature in [40] defines phylogeny as “A phenomenon which indicates how various groups of microorganisms are genetically related and aids in tracing their evolution.” The science that relates to the study of phylogeny is known as 'Phylogenetics,' and the outcome of the process is a 'Phylogenetic Tree,' which serves as an established form of a data structure for sequence-based genetic similarity in microbial species with three possible domains of bacteria, archaea, and eukarya (Figure 1.3).
Table 1.1 An overview of methods for constructing phylogeny. The first column states the name of the method followed by a description [42], [43].

<table>
<thead>
<tr>
<th>Type of Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance-based Methods</td>
<td>Phylogeny is constructed by similarities of genomic sequences based on pairwise distances that are calculated by summing of all base pair differences between two neighbouring gene sequences.</td>
</tr>
<tr>
<td>Maximum Parsimony</td>
<td>The method aims to find the best tree(s) of evolution that minimise/s the number of changes between ancestors and their children.</td>
</tr>
<tr>
<td>Maximum Likelihood</td>
<td>This method uses a general statistical technique for estimating probabilities of the sequence observations given a model of their evolution over a tree of evolutionary history. It aims to select the evolutionary tree that maximises the likelihood of the sequence occurrence.</td>
</tr>
<tr>
<td>Bayesian Approach</td>
<td>The approach works closely to Maximum Likelihood but differs in assuming a prior probability distribution of the possible evolutionary trees in phylogenetic construction.</td>
</tr>
</tbody>
</table>

Figure 1.3 A universal phylogeny (tree of life) of the living world obtainable by comparative sequencing of DNA. The evolutionary measure between two micro-organisms is proportional to the cumulative distance(length) of the branches that join two groups of organisms.
The structure of a phylogenetic tree (also known as the tree of life) [17] encodes information on the common ancestry from \textit{phylum} to \textit{species} taxonomic levels of the microbiome. A phylogenetic tree is an acyclic line graph representing a set of nodes and branches $T\{\text{nodes, branches}\}$. The data structure is comprised of $n$ leaf nodes representing contemporary taxa or OTUs; $n - 1$ internal nodes representing ancestral nodes placing themselves at branching points; 1 root node (last common ancestor of all taxa); where each node is connected to other nodes by a set of branches (Figure 1.4) [44],[45]. The two descendants of a common ancestor are sister or brother nodes, indicating how close they are to each other than other taxa [44],[45]. A group containing a common ancestor and all its descendants based on grouping of phylogenetic lineages is known as a clade. The term ‘clade’ was coined by Huxley in 1957 [46]. The clades are nested within one another. The pattern of branching of clades in a phylogenetic tree is indicative of how OTUs have evolved from their common ancestors (Figure 1.3,1.4). Commonly the branches of phylogenetic trees are labelled with evolutionary biological distances, also known as phylogenetic distances [47]. An example of a sample phylogenetic tree is illustrated in Figure 1.4. The bar at the bottom of the figure depicts a scale that represents an amount of evolutionary or genetic change (0.08 in the example illustrated in Figure 1.4). It is calculated by the number of variations in genomic sequences or 'substitutions' divided by the length of the nucleotide sequence.
Microbial phylogeny

Figure 1.4 An illustrative representation of a detailed sample phylogenetic tree.

The Newick format noticed in [48], is widely used for representing phylogenetic trees in a computer-readable form. It represents the correspondence between trees and the nested parentheses. The sample tree shown in Figure 1.4 could be represented by the following pattern of parenthesis in the Newick tree format.

(((A: 0.95, B: 0.64): 0.88, ((C: 0.85, D: 0.28): 0.4, E: 1.0): 0.51): 0.0)

An important piece of information that can be extracted from phylogeny is a matrix depicting phylogenetic distances between various OTUs or a matrix that records the amount of shared branch length between each pair of OTUs [49]. Such phylogenetic matrices capture information on evolutionary distances from tree topology and serve as a basis to evaluate richness, evenness, and diversity within the microbial communities and their relation to the biological functions.
An example of a phylogenetic similarity distance matrix corresponding to the sample tree shown in Figure 1.4 is shown in Table 1.2.

Table 1.2 An example of phylogenetic similarity distance matrix corresponding to a sample tree shown in Figure 1.4.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>1.59</td>
<td>3.59</td>
<td>3.02</td>
<td>3.34</td>
</tr>
<tr>
<td>B</td>
<td>1.59</td>
<td>0</td>
<td>3.28</td>
<td>2.71</td>
<td>3.03</td>
</tr>
<tr>
<td>C</td>
<td>3.59</td>
<td>3.28</td>
<td>0</td>
<td>1.13</td>
<td>2.25</td>
</tr>
<tr>
<td>D</td>
<td>3.02</td>
<td>2.71</td>
<td>1.13</td>
<td>0</td>
<td>1.68</td>
</tr>
<tr>
<td>E</td>
<td>3.34</td>
<td>3.03</td>
<td>2.25</td>
<td>1.68</td>
<td>0</td>
</tr>
</tbody>
</table>

One of the popular phylogeny-aware distance matrices is UniFrac (Unique Fractions of a phylogeny tree unique to a sample or the other sample) [50]. It measures the biologically meaningful distance between microbial communities in comparison to the other standards in literature, such as Euclidean and the Bray–Curtis distances [51]. To calculate UniFrac, microbial taxa found in one or both samples are located on a phylogenetic tree to find out tree branches leading to that taxa from both samples (i.e., shared branches) and; branches leading to taxa present in only one of the samples (i.e., unshared branches). The UniFrac distance between the two samples shows a fraction of total unshared branch lengths and calculated as shown in Eq.1.1.

\[
UniFrac = \frac{\text{Sum of unshared branch lengths}}{\text{Sum of all tree branches}}
\]  

An illustration of UniFrac distance between samples is shown in Figure 1.5. Incorporating phylogenetic distance measure in downstream analysis has the potential to quantify for detecting functional differences between metagenomes. Overall, phylogeny could guide microbiome research to (i) categorise
Unclassified microorganisms; (ii) to test trait or habitat associations of the microorganisms and their biology; (iii) to study how microbial communities evolve in time and space; and (iv) to better understand the diversity in microbial communities [50],[52].

Figure 1.5 An illustration of UniFrac Distance measure between two samples (shown as pink and blue microbial samples).

The three possible ways to integrate phylogeny into ML predictive models are characterised in Figure 1.6.
This sub-section discussed the basic concepts involved in the phylogenetic profiling of microbiome, opening the avenues to integrate phylogeny in downstream analysis of the quantitative microbiome.

The next section is dedicated to how downstream analysis of microbiome profiles and linking them to functional roles is a captivating research community in general. The key milestones achieved in this are listed, suggesting a plethora of opportunities to model over the microbiome.

### 1.5 Metagenomic functions

Metagenomics has gained prominence due to the emergence of important projects such as the Human Microbiome Project (HMP) (http://hmpdacc.org/) [53], Earth Microbiome Project (http://www.earthmicrobiome.org/) [54], and American Gut Project (http://americangut.org/) [55]. Efforts to map the structure of microbiome (genotype) to its functions (phenotype) have been made across different environments such as the human body[53], marine life [56], soil [57].
Several studies have indicated relations between microbial composition with its functions, leading to novel insights into the world of the microbiome. Hadrich et al. [59] recently summarised a large number of key projects across the world that are useful for studying the complexity, variability, and effects of interdisciplinary microbial communities, holding a lot of potential and possibilities. The few other key milestones achieved in this direction are discussed below. Giovannoni et al. [60] analysed 16S rRNA microbial genes obtained from the natural populations of Sargasso Sea picoplankton. The analysis identified a novel microbial group, the *SAR11 cluster*, and a group of cyanobacteria, prochlorophytes, and chloroplasts, playing essential roles in the biogeochemical change in marine ecosystems. Sunagawa et al. [61], studied the structure and function of the microbiome in TARA oceans samples from 68 locations in epipelagic and mesopelagic waters, highlighting the fact that the microbial diversity is responsible for various biogeochemical processes in oceans. Aislabie et al. [62] examined the functional, metabolic, and phylogenetic diversity of soil microbiome to study their effect on nutrient cycling and organic matter production in soil production. In addition to these key studies, the other two key potential areas of metagenomics research have embraced unprecedented growth and utilised in this thesis are -i) Human Microbiome and ii) Cattle Microbiome.

**1.5.1 Human microbiome studies**

HMP, an initiate of the National Institute of Health, USA [53], uncovered that the microorganisms residing in humans even outnumber the genes present in the human body; and hence are responsible for controlling various metabolic
functions, their health, and predisposition to various diseases [9]. HMP has accelerated the study and computational analysis of metagenomic studies. As part of this initiative, more than 5,000 microbiome profiles were sampled from 15 (men) to 18 (women) body sites of 242 healthy US volunteers [53]. Dynamics of human microbiome could also be defined by another key milestone of the American Gut Project [55]; which primarily studied the cohorts of microbiome at different body sites, effects of diet on microbial composition, effects of antibiotic treatments and impact of s lifestyle patterns such as sleep, stress and exercise levels of human beings. The Integrative Human Microbiome Project (iHMP) [63], is a dedicated second phase of HMP based on studies of relating the microbiome with: i) pregnancy and preterm birth; ii) IBD, and iii) stressors that affect diabetes patients. The compositional distribution of human microbiome varies across different body habitats. Human microbiome studies have revealed that the composition of microbial communities is responsible for physiological or patho-physiological states [9]. In a series of studies coming out, important breakthroughs indicated the linkage of the human microbiome with obesity, the effect of diet intake [64],[65], IBD [66], cancer [67], and antibiotic treatment [68] for medical applications. In addition to the above studies, microbial analysis has been used for the classification task of human body sites [69]. Such classification aims to study the role of microbial communities associated with different human body sites (phenotypes), such as skin, oral cavities, lungs, gut, and vagina sampled from the human subjects.
1.5.2 Cattle rumen microbiome studies

Metagenomics serves as a powerful method for studying cattle (Bos taurus) rumen microbiome [58]. Cattle rumen contains microorganisms that support anaerobic and methanogenic fermentation activities over cattle feeds relating to cattle productivity. The studies in [70]–[74] suggested that diverse cattle rumen microbiome works symbiotically to break down the diet consumed by cattle controlling the production, feed efficiency, and methane production. It is well known that diet has an impact on microbial composition in cattle microbiome [75]. The authors in [76] recently reviewed the application of metagenomics to study the impact of the microbiome in cattle functions such as lower methane emissions, higher feed efficiencies, and responses to different feeding regimes. The research in [58], [70]–[72], [77], [78] indicates the relation of cattle microbiome with its different phenotypic functions such as an effect of diet, stress hormones, and feed intake or residual efficiency.

The current thesis primarily focusses on human and cattle rumen microbiome profiles for their functional analysis. Additionally, one of the use cases in the current research utilised the soil microbiome.

1.6 Problem domain under investigation

The steep decrease in the cost of generating metagenomic sequences with unprecedented advances in NGS techniques has created a gap between the pace of their generation and computational analysis. Metagenomic sequences are usually massive, high-dimensional, sparse, heterogeneous, over-dispersed, and highly varied [79]. This poses computational challenges [79]. The main aim for
most microbiome studies is to discover new knowledge about the characteristics and functions of the microbial community and their relationships or interactions with the host environment, considering the challenges of metagenomic data.

The vital biological questions of **Who?** **How?** and **What?** in microbiome research are majorly related to - i) who are present in a metagenomic sample; ii) how they are related, and iii) what they are doing in an ecosystem. The methods developed in this dissertation are motivated by computational analysis of the structure of microbiome (i.e., **who** are present and **how** are they related) and further linking it to environmental phenotype (i.e., **what** are they doing). Such computational analysis is typically referred to as "**functional metagenomics**" in this thesis. In the current research, novel computational methods are developed to address the above biological questions, considering the key characteristics of the microbiome data (as detailed below).

**1.6.1 Key characteristics of microbiome data**

The key characteristics of the microbiome data are listed below [31], [80]–[82].

- The abundance count data of microbial features (OTUs or ASVs) in metagenomics are high-dimensional. Mostly, the number of features outnumber the number of metagenomic samples. Feature selection becomes vital for the analysis of such high-dimensional data.

- Abundance count profiles generated by a metagenomic pipeline are often incomplete, over-dispersed, noisy or erroneous, due to the probable existence of high variability in species of a microbial community. The metagenomics pipelines provide such microbial count data, albeit with a lot of zero values due to limited sequencing depth.
• Construction and synthesis process of microbial features in standard pipelines, such as in [19],[20], ignore the natural structure present in phylogeny. Prior knowledge of the evolutionary relationships among OTUs/ASVs is essential. It is assumed that closely related OTUs/ASVs share their evolutionary history and tend to show similar responses to biological functions [42],[83],[84],[85]. Therefore, integrating biological domain knowledge of phylogeny is important in microbiome analysis.

• Heterogenous data input formats are inherent in metagenomics with the generation of abundance count matrix and phylogenetic tree of OTUs/ASVs.

• The distribution of the abundance of microbial features, as well as their occurrence across metagenomic samples, is skewed.

• Compositional nature of microbiome indicates the relative weight or importance of each feature in a metagenomic sample.

The current research targets to employ significant analysis in the field of functional metagenomics by considering the key characteristics of microbiome data being heterogenous, high-dimensional and sparse.

1.7 Research focus

The focus of this thesis is to investigate, develop, and evaluate the performance of computational methods that can be employed for determining functions of microbial compositions with a two-folded aim:

• To study the overall properties of the microbiome profiles for differentiating microbiome samples into the functional phenotypes by finding ways to
determine relevant microbial features that are associated with such differentiation.

- To study the biological domain knowledge of phylogenetic relationships between microorganisms at different levels of taxonomy and integrating them in deciding functions, under the assumption that the natural hierarchical grouping of OTUs (from phylogeny) could help the functional classification of metagenomes better [42],[83],[84],[85].

The remainder of this section will discuss the rationale and objectives of current research seeking answers to key research questions.

1.7.1 Rationale and Objectives

The steep decrease in the cost of generating metagenomic sequences with unprecedented advances in NGS technology created a gap between the pace of generation and its analysis. Due to the key characteristics of metagenomic data, and the necessity of identifying the related microbiome (from phylogeny) and their functions, the employment of computational approaches could advance our understanding of how the microorganisms are linked to their environment [86]. Predictive modelling of high-dimensional, complex, heterogeneous, and biological intensive microbiome data to determine its functions is a challenging task in terms of attaining high predictive performance and lowering the computational cost. Additionally, jointly modelling the phylogenetic profiles with high-dimensional abundance profiles of OTUs in ML frameworks pose a further computational challenge. The design of new methods is still emerging to analyse metagenomic data indicating their critical applications and implications in biological sciences. Gaining insights from this rationale, the
main objective of this thesis is to address the following questions in this research.

1. How to design an efficient computational approach to extend beyond the conventional methods of downstream functional analysis in metagenomics, for attaining better outcomes while dealing with high-dimensional metagenomic data?

2. How to include biological domain knowledge (i.e., taxonomical structure or phylogeny or depth covering lineages from phylum to species levels or the natural characteristics of the microbiome) into the perspective of computational metagenomics analysis for predicting functions?

3. How to integrate diverse and heterogenous metagenomic data sources in a structured way for the downstream analysis of metagenomes?

4. How to develop a novel knowledge-driven ML Method, which could govern the functional derivations from metagenomics profiling data?

5. To perform a case-study evaluation to investigate essential questions (Who? How? and What?) of the microbiome research.

The thesis has targeted to implement solutions to the above research questions in the form of novel ML-based methods. This dissertation presents new computational frameworks developed for microbiome data analysis centering around the two themes of (i) abundance-only analysis and (ii) phylogeny-aware analysis to cater to the limitations in the literature (discussed in Chapter 2). These are developed for 16S rRNA sequence data, although they can also be adopted for shotgun metagenomic data in the future.
1.8 Key contributions of current research

Microbiome research is on the unobtrusive path to relate the structure and functions of non-cultured microbial communities. The current thesis has contributed to the development of new predictive models for such analysis. The thesis progresses by the construction of novel ML-based frameworks having an emphasis on integrative structural profiles for enhancing the quality of functional predictions over the varied metagenomic data, finding solutions to the key research questions in hand (Section 1.7.1). The contributions made in current research revolve around the proposed 4D’s as major influencing factors (illustrated in Figure 1.7). The proposed 4D structure here summarises the objectives of current research.

The key contributions of this research are outlined below.

1. **Key Contribution 1.** The thesis proposed a new framework using an ensemble of embedded ML models for classifying microbiome use case sample datasets constituting the high-dimensional OTU abundances (quantitative profiles), into their respective phenotypes. The focus of this framework is to investigate, develop, and evaluate the performance of embedded ML methods that can be used for functional classification of the microbiome with p>>n, where p denotes the number of microbial features and n denotes the number of microbial samples. Conventional models lack in using embedded methods (such as the gradient boosting methods) to analyse abundance profiles, and; could be improved with the use of embedded methods to attain high predictive power by reducing the computational cost involved.
Figure 1.7 4D’s highlighting major factors playing an important role in key contributions made in this thesis.

2. **Key Contribution 2.** The second and significant contribution of this thesis is the development of a novel ML-based integrative framework that captures the evolutionary relationships between the OTUs rather than considering them as independent features in microbiome data formulation. It builds on the assumption that groups of closely related OTUs share key functional similarities [42],[83],[84],[85]. A new data structure involving the creation of microbial feature space at multiple levels of phylogenetic granularity (as obtained from the phylogenetic tree), is proposed. It is termed as PAAM
(i.e., Phylogeny and Abundance-aware Matrix). The data structure maps to node-by-node information obtained from a phylogenetic tree. This data structure is further used as an input into the predictive models. The corresponding study revealed the possibility to distinguish microbiome functions better by incorporating biological phylogenetic lineages in an integrative way with the abundance count profiles.

3. **Key Contribution 3.** Additionally, modelling of phylogeny within ML models has not been explored in the literature sufficiently. To the best of knowledge gained from literature, currently, phylogenetic distance has been utilised to regularize the logistic regression model in [85],[87]. However, efforts of integrating biological knowledge in tree-based models are lacking, which could be the potential direction of work. In this thesis, for the first time, the phylogenetic information is used to guide the popular tree-based predictive model of Random Forest [88],[89], accounting for useful functional classification of the microbiome. Closely related microbes by phylogeny are highly correlated and tend to have similar genetic and phenotypic traits [83]–[85]. Such microbes behave similarly and hence tend to be selected together, or one of these could be dropped from the analysis, to make ML with a popular state-of-the-art RF better. Following this principle, a new classifier of Phylogeny-RF is developed for functional metagenomics.

4. **Key Contribution 4.** Furthermore, a case study is conducted to analyse the community composition of *Bos taurus* (cattle) rumen microbiome to link it to the phenotype of supplemented diet or synthetic cortisol indicating its essential role in cattle productivity, health, and immunity. The study
reported interesting findings using different phylogeny and abundance-driven ML methods and characterised OTUs or ASVs, that are useful in identifying important cattle functional roles. The case study supports that the presence of microbiome in cattle ruminants is linked to their dietary supplements. Along with this case study, a novel method of ‘Phylogeny-PINA Relief’ is proposed which explored the integration of a phylogenetic measure at the feature selection level. In this method, the parental nodes are weighted and ranked based on Relief measure [90] using the phylogenetic interactions of a parent with the child node features (obtained from a phylogenetic tree) [91], deciding their contribution in the differentiation of the phenotypic classes associated with the microbiome. Comparison with state-of-the-art methods [90],[92],[93] highlights the advantages of using such phylogenetic information in the ranking of features in microbial space.

Each of the above research contributions (1-4) correspondingly maps to each of the contributory Chapters in this thesis (Figure 1.8).

Additionally, computational tools (scripts) developed as part of this thesis are contributed to a common portal of GitHub to facilitate the research community.
1.9 Overview of the thesis

The thesis is organized into 8 Chapters. The background information is provided at the beginning of the thesis. An overview of the basic metagenomics concepts and the problem under consideration is discussed in Chapter 1. The key characteristics of metagenomic data are also listed. Furthermore, the objectives and key contributions of this research are provided in Chapter 1. A summary of the rest of the chapters in this thesis is outlined below.

Chapter 2. A Literature Review on Predictive Modelling of the Microbiome using ML techniques. An introduction to classifying the microbiome into environmental functions is provided, highlighting the computational principle of supervised ML involved. A literature review is conducted that focuses on a
variety of ML techniques for such classification. This chapter reviews novel approaches which integrated phylogenetic knowledge in the classification of microbial samples into environmental functions. The methods are examined, and their limitations and advantages are highlighted to obtain more in-depth insights into the research questions investigated in this thesis.

Chapter 3. Research Methodology and Materials. The chapter illustrates a generalised conceptual framework proposed in this thesis to infer metagenomic functions (functional metagenomics). It is useful in problem formulation across this thesis to ensure the development and application of computational algorithms in the context of functional metagenomics across all the contributory chapters (Chapter 4-7). A description of a variety of data sources utilised throughout the thesis for inferring metagenomic functions is also provided.

Chapter 4. An Abundance-driven Framework for the Prediction of Metagenomic Functions. The chapter provides a novel abundance-driven framework for functional metagenomics, following the framework conceptualised in Chapter 3. The chapter presents a comprehensive evaluation of different ensembled ML models employed for the predictive task of determining metagenomic functions over the high-dimensional abundance counts. Embedded methods of Gradient Boosting and Penalized Logistic Regression are utilised as potential methods in the functional classification of metagenomes. Results on the performance of the different methods are presented along with the relevant conclusions.
Chapter 5. A Novel Phylogeny-driven Framework for the Prediction of Metagenomic Functions. A novel phylogeny-driven framework for the prediction of metagenomic functions is proposed. An overview of the methodology is presented. A description of how prior knowledge of phylogeny is incorporated into microbial abundances and insight on how heterogenous metagenomic data is integrated is provided. An overview of the implementation of the framework over the metagenomic data from different environments is presented. The framework primarily deals with integrating biological knowledge of phylogeny at the data pre-processing level, deriving a new data structure to be further used as input to the predictive models. The next steps in the subsequent chapter would explore integrating phylogeny in the predictive ML model itself.

Chapter 6. A New Phylogeny-driven Random Forest Classifier for Functional Metagenomics. This chapter aims to integrate prior knowledge of phylogeny in a popular predictive ML model of Random Forest (RF) to construct a novel Phylogeny-driven RF framework for functional metagenomics [88]. The RF model is discussed in Chapter 2 as one of most of the popular methods in functional metagenomics. Preliminary results obtained when applying the Phylogeny-driven RF prediction paradigm are presented. The novel feature encoding scheme of RF is proposed, and the rationale is summarised. Experimental settings, including the RF, datasets, and feature-encoding technique, are presented. An overview of the classification performance is provided, along with insightful conclusions.
Chapter 7. Prediction of Metagenomic Functions in Cattle Rumen using the Abundance-driven and Phylogeny-Driven Analysis. In this chapter, the application of the abundance and phylogeny-driven ML models to infer metagenomic functions in cattle microbiome is presented. The results are presented, together with a comprehensive evaluation of phylogeny-driven and abundance-driven analysis. The data is sourced from the MetaPlat project (http://www.metaplat.eu/) under the Horizon 2020 EU project. The biologically relevant cattle rumen microbiome is identified, providing various novel insights.

Chapter 8. Summary and Future Work. In this chapter, a summary of the main findings and research contributions from each chapter is presented, including the integration of prior knowledge of phylogeny and the development of new proposed frameworks. A discussion on related potential improvements and possible future work to extend the research in this thesis is presented.

1.10 Publications

The following is a list of publications that were disseminated in the course of the three years of the research presented in this thesis.


1.11 Summary

Metagenomics dealing with the study of microbial communities has become an unobtrusive science. Microbial communities are responsible for various biological functions within the environment. Analysing the profiles of microbial communities is an integral part of metagenomics and could assist in the understanding of biological applications better. The speed of such a metagenomics analysis has been accelerated by the development of large-scale NGS projects. This has resulted in the enormous production of biologically rich and varied data. Data analysis and modelling techniques are required to analyse such heterogeneous data, inferring metagenomic functions, and evaluating their predictive performance. This chapter has provided an overview of a generalised metagenomics pipeline for 16S rRNA genes for determining their functional roles. The chapter provided background information on the basic concepts of metagenomics, highlighting microbial phylogeny, and emphasizing the importance of functional metagenomics. It listed the number of characteristics of metagenomic data that require addressing while modelling. This chapter further discussed the motivation for the research presented in this thesis. This chapter highlighted the research focus and main research questions. The key contributions achieved (in response to the key research questions) are stated along with a chapter-wise distributed outline of the thesis providing a high-level overview.
This chapter provides an overview of relevant literature on current ML-based solutions to predict metagenomic functions in downstream analysis of microbiome profiles. The goal of this overview is to introduce various current ML techniques that could be applied for the prediction task, highlighting ongoing challenges, their advantages, and limitations. The chapter firstly introduces a general learning methodology for the downstream analysis involved in understanding and linking the role of the microbiome to its functions. The following sections describe general exemplary literature categorised into - i) abundance-aware methods and ii) phylogeny-aware methods, making it a self-contained discussion.

2.1 Introduction

The rapid generation of metagenomic sequences has led to intense research in developing ML-based computational workflows for interpreting biological roles of metagenomes. As discussed in Chapter 1, prediction of microbiome functions (i.e., phenotype) and linking them to the microbiome structure is
essential in downstream analysis of metagenomic sequences. Such a genotype-phenotype association is involved in the development of therapeutic approaches relating to functional metagenomics. For example, metagenomic studies of human gut bacteria have shown signs of connections to diseases and pathogenies and hence is useful in developing corresponding medical treatment and therapeutics [94].

Figure 2.1 represents a general methodology applied in metagenomic use cases for deriving functional inferences. This methodological flow offers a path to interpret microbiome structure, variation, and its connection to the biological systems.

**Figure 2.1 An illustrative scheme of functional inference from metagenomic environments.**

The input data comprises of microbiome samples containing information on 16S gene sequences of the microbial species. A “tree of life” that describes the phylogenetic relationships among the species is used as the prior domain knowledge in this thesis [17]. A learning objective provides a functional mapping from each metagenomic sample comprising of microbial features, to the set of functional predictions obtainable from metadata (phenotype) information. Metadata provides researchers in the metagenomics field with the
ability to study the association of natural environment upon the microbial community structure and its related functions. The human microbiome has been implicated as a gold standard studying the relations between the microbial structure and phenotypic functions[53]. Investigation of microbial communities obtained from water bodies, soil, cattle, etc. has also been accelerated [60],[70],[95]. The outcome provided by the learning algorithm needs to be evaluated to be confirmed. Robust methods of assessing Accuracy using bootstrapping or cross-validation [96] are useful to validate the functional predictions. Bootstrapping is achievable by estimating the performance of a learner over the smaller data samples (with replacement) whereas cross-validation is done by splitting the input data iteratively into k different parts (say k = 10 parts) to train the learning algorithm on k-1 parts, and predicting the result for the k\text{th} part. [96]. The functional predictions achieving high performance and providing relevant insights on biological processes lead to the development of an appropriate computational model for functional inference. A predictive scheme summarised in Figure 2.1, serves as a useful tool for understanding vital microorganisms, distinguishing microbial host sites or relating microbiome to its functions in biological systems.

This predictive scheme maps to the standard framework of “supervised learning or classification” [97], well as the functional mapping f learned from the input data (structure of microbial genome) to the set of predefined output class labels (Eq.2.1).
where $X$ represents samples containing the quantitative profile of OTUs or ASVs or taxa and $Y$ represents the set of predefined functional labels (also known as metadata or phenotype). The training set $(X_1, Y_1), \ldots, (X_t, Y_t)$ of $n$ independent samples associated with $t$ labels, is distributed as $(X, Y)$. A high-quality solution attains good predictive values of actual output $y \in Y$ on a given input $x \in X$. In this sense, predictive learning intuitively infers metagenomic functions as part of the problem under investigation. The goal of supervised learning is to derive functions of training sample data that can further be used to assign correct functional labels on unseen sample inputs. The following section provides a review of relevant ML-based supervised modelling techniques that are useful for determining the biological roles of the microbiome.

2.2 Approaches to infer metagenomic functions using supervised learning

To identify the most suitable model for predicting functional metagenomes, various supervised ML classification algorithms are broadly separated into two groups – a) those using only abundance counts of OTUs or ASVs in analysis (i.e. abundance-aware methods) and b) those combining biological or evolutionary relationships between various OTUs/ASVs in the analysis (i.e. phylogenetic-aware methods). The range of models in each category is discussed below with insights on their advantages and limitations.
2.2.1 Abundance-aware methods for downstream analysis of metagenomes

Related prior work on comparison of classifiers performing predictive modelling with the environment or host phenotype using 16S gene sequencing data from human body niches [69], [98]; diseased state/s in humans [99], [100]; prediction of subjects in forensic studies [101] and other environments such as in [102], has been observed. The abundance count of OTUs served as the primary input feature vector for downstream analysis using ML algorithms in these studies. The prior work on the human body-site classification used the following datasets: Costello Body Habitat (CBH) with six body sites, Costello Skin Sites (CSS) with 12 class labels, and Pei Body Site (PBS) 4 sites as the functional classes [69],[93]. Furthermore, the datasets of Fierer et al. Subject (FS) dealing with keyboard data and Fierer et al. Subject with Hand (FSH) dealing with the diversity of microorganisms in human hands were also analysed in the study by [69],[93]. A comprehensive comparison of classifiers for body-site classification over CBH, CSS, and PBS over the OTU counts, has been performed by Statnikov et al.[98], using 18 classification methods with various parameter settings. The research in [98], [103], conducted experiments for the prediction of functional states of Psoriasis and IBD disease. The studies in [69],[93], [98], [103],[104] indicated Random Forest (RF), Support Vector Machines (SVM), Elastic Net (ENet) and Multilayer Perceptron (MLP) as widely explored state-of-the-art supervised learners, for classifying 16S rRNA data into metagenomic functions. A key finding from these studies reported RF as consistently outperforming the other algorithms probably due to, modelling over the nonlinear effects and diversified features (For example, RF attained
93.8% and 99% Accuracy over CBH and FS datasets respectively). Also, the high predictive Accuracy of 92% and 98% was achieved by SVM when predicting functions of CBH and FS human microbiome datasets [93],[98].

A discussion on these popular classification techniques in microbiome studies is presented below. Insights on their pros and cons are also provided.

**2.2.1.1 Random Forest (RF)**

RF has been a popular ML algorithm applied in the analysis of biological systems [89]. In functional metagenomics, RF has become the most popular method [69], [98], [99]. RF is an ensemble learning method, which progresses by constructing a set of $t$ decision trees $\{DT(x, f)\}_t$. Each $DT(x, f)$ represents a decision tree which is grown on $f$ features with bagging, or bootstrap aggregating [88], [105] to decide on sample $x$. The method outputs the functional class $i$ for a sample $x$ based on majority voting decision by all constituting $t$ decision trees $DT(x, f)$ grown in an RF (Figure 2.2) [88],[105]. RF tends to attain a robust combination of different decision trees, minimizing the risk of the possible weak decision of a single tree. Henceforth, it tends to minimise the generalization error of classification on the test feature set. Scientific interpretability in the structure of the $DT(x, f)$ of RF is primarily driven by randomly picking set of features and ranking them according to the class distribution in input data using Gini Impurity (GI) [88].

GI of a node in the decision tree of RF tree is formulated as shown in Eq.2.2 [88],[89],[105].

$$GI = \sum_{i=1}^{C} \text{prob}_i (1 - \text{prob}_i)$$  \hspace{1cm} (2.2)
where \( prob_i \) denotes the probability of choosing an element with class \( i \), and; \( C \) denotes the total number of class labels. The weighted average of the total decrease in the GI weighted by the proportion of samples reaching a given node in a decision tree is termed as the mean decrease in GI. It is useful in deciding the best split in the constituting decision tree as best split is the one that maximises the impurity reduction. The step-wise functionality of RF is summarised in Algorithm 2.1 [88].

<table>
<thead>
<tr>
<th>Algorithm 2.1 Random Forest</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input.</strong> The input set consists of ‘x’ samples, with feature space of ‘f’ dimensions and ‘p’ phenotypic class feature with C distinct classes.</td>
</tr>
<tr>
<td><strong>Method.</strong></td>
</tr>
<tr>
<td>1. Generates a random sub-sample space using bagging or bootstrapping method to generate ( t ) subsets ( \mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_t ) from the input data.</td>
</tr>
<tr>
<td>2. For each training sub-sample set ( \mathcal{D}_i ) (( 1 \leq i \leq t )), a decision tree is grown. For each decision tree, a subset of features ( m ) is generated (( m &lt;&lt; f ) and usually ( m = \sqrt{f} ) for classification problems), splits using GI are computed, and; the node with the best split is chosen as a splitting feature to generate a child node. This is until the stopping criteria are met or tree is built fully[88].</td>
</tr>
<tr>
<td>3. Builds a forest by repeated steps in 1-2.</td>
</tr>
<tr>
<td>4. Uses the forest built-in step 3 to predict output class for the samples in the test set. Calculate the votes for each predicted class by each tree. Consider the majority class target as the final prediction from the RF algorithm.</td>
</tr>
<tr>
<td><strong>Output.</strong> Predicted Class of Test Samples</td>
</tr>
</tbody>
</table>

The strength of RF lies in that it combines several classification trees with random splitting method whilst learning from the input data and increasing the interpretability of the model [88]. It maintains good classification Accuracy when a large proportion of the datasets have missing values or when datasets are high-dimensional[106]. RFs are not dependent on any prior data distribution assumption and could be used on any types of data. Hence, RF seems to be more suitable for metagenomic datasets about its fundamental characteristics [106].
Also, there is not much need for feature normalization before applying RF. RF can estimate the importance of variables. RFs are usually robust methods avoiding overfitting, depending on several trees in the forest or the maximum depth of the constituent trees. On the contrary, a single decision tree would tend to overfit the data. The main disadvantage of RFs is their complexity while building many trees. RFs may prove more time-consuming over large datasets with many trees.

![Random Forest Diagram](image)

**Figure 2.2** A diagrammatic illustration of an RF Classifier classifying input samples into two Classes X and Y with the help of ‘t’ decision trees.

### 2.2.1.2 Support Vector Machines (SVM)

SVM is a well-known classification technique that maps training set in high-dimensional feature space to a separating hyper-plane [107]. It primarily behaves as a non-probabilistic binary linear classifier establishing the maximization of the margin between two functional class categories in
Approaches to infer metagenomic functions using supervised learning

consideration (Figure 2.3) [108]. SVM searches for the points in each class that are the closest to the margin. Those points are called “support vectors”. SVM tends to calculate the optimal hyper-plane by using kernel functions [107] over these support vectors. Solving SVM is a quadratic programming problem [107], [108]. The training dataset can be divided into two different sets S1 and S2 representing the class labels +1 and -1 respectively. The separating plane between the classes +1 and -1 can be constructed as shown in Eq.2.3 a-c (Figure 2.3).

\[
\begin{align*}
\mathbf{w}^T \mathbf{d}_i + b &\geq 1 \quad \text{if} \ c_i = 1 \tag{2.3a} \\
\mathbf{w}^T \mathbf{d}_i + b &\leq -1 \quad \text{if} \ c_i = -1 \tag{2.3b} \\
\text{Margin between two hyperplanes} &= \frac{2}{||\mathbf{w}||} \tag{2.3c}
\end{align*}
\]

where \( \mathbf{w} \) represents a parameter of the decision hyperplane, \( b \) is an offset in the equation of the plane, \( \mathbf{d}_i \) is a data point, and \( c_i \) is the class of data point \( i \) (+1 or -1), \( ||\mathbf{w}|| \) is the norm of vector \( \mathbf{w} \). The problem formulation targets to maximise the margin of \( \frac{2}{||\mathbf{w}||} \). SVM support a variety of other kernel functions [108], which can transform the original feature space into kernel space, constructing a decision boundary to separate samples into different class categories. As an example, a polynomial kernel is used for degree-n polynomials and is defined in Eq.2.3d.

\[
\text{PolyKernel} \ (x, y) = (xy + b)^n \tag{2.3d}
\]

where \( x \) and \( y \) are vectors in the input space, and \( b \geq 0 \) is an offset parameter.

A sample radial kernel is defined in Eq.2.3e.
RadialKernel \( (x, y) = \exp\left(-\frac{||x - y||^2}{2\sigma^2}\right) \)  \hspace{1cm} (2.3e)

where \( ||x - y||^2 \) is a squared Euclidean distance between the two vectors and \( \sigma \) is a free parameter.

Since SVM is originally designed to discriminate two different categories; hence for multiclass classification, one-against-one or one-against-all policy is used. However, if the interest is inclined more towards relative importance of the features in specifying this decision boundary instead of formulation of decision boundary itself, SVM with Recursive Feature Elimination (i.e. SVM-RFE) serves as an alternative method [109].

In the case of high-dimensional metagenomic data, SVM-RFE could act as a specialised backward weighing based strategy to rank the OTUs, and it deletes the OTU feature with the lowest weight before the next SVM training [109]. It has been recommended as one of the effective methods in [110] for metagenomic functional classification.

\[\text{Figure 2.3 A diagrammatic representation of state-of-the-art SVM classifier differentiating two classes.}\]
SVM does not rely on prior distribution assumption as well and focusses on only support vectors (i.e., samples close to margin) while ignoring the rest of the samples that are far away from the decision margin. This serves as optimization and saves the computational cost. However, SVM predominantly depends on the choice of kernel functions [108]. SVM was initially designed for two-class boundary problems [107]. They utilise one-against-one or one-against-all strategy to determine labels in multiclass problems. SVM works better for linearly separable classes. SVM may perform poorly in the case of datasets with large numbers of features. In such cases, wrapper methods of RFE with SVM have been proposed [109] to select high-quality feature subsets for a classifier. RFE with SVM, however, is computationally extensive since it scans through all features one by one recursively and it does not consider any correlation the features. The filter methods such as the ratio of the between-class sum-of-squares to the within-class sum-of-squares (BSS/WSS) have also been used before the application of SVM in a comprehensive study by Knights et al. [69],[93]; for classifying the metagenomic data. This indicates a combination of a feature selection technique and SVM could prove more useful in metagenomic studies.

### 2.2.1.3 Elastic Net (ENet)

ENet is one of the representative methods for solving regularization problems with an optimization strategy [111]. The method tends to fit a generalised logistic regression (LR) model for classification by regularizing the penalty associated with maximum likelihood in LR [112], [113]. A general framework
for solving regularization problems with optimization strategy is shown in Eq.2.4 [111], [113].

\[
P\left(Y = \frac{1}{x}\right) = \frac{1}{1 + e^{-h(x)}}
\]

(2.4)

where \( h(x) = w_0 + \sum_{j=1}^{d} w_j x_j \), and \( d \) is the number of dimensions; \( w_j \) are regression coefficients and \( x_j \) are input features. ENet, constraint the regression coefficients (w’s) by imposing a penalty on their size (Eq.2.5) [113].

\[
\text{ENet} = \arg \max_w \sum_{k=1}^{n} (y_k(h(x)) - \log(1 + h(x))) - \lambda \sum_{j=1}^{d} (\alpha w_j + (1 - \alpha) (w_j^2))
\]

(2.5)

where \( \alpha \in [0,1] \), \( j \) denotes the number of features (dimensions); \( \lambda \) is the regularization parameter that controls the overall strength of the penalty; \( n \) is the number of samples; \( y \) is the response at sample \( k \) (i.e. a vector of \( d \) values at observation \( k \)).

ENet has embedded feature selection support driven by penalizing the regression coefficients. It provides a powerful penalty function to efficiently select important features on high-dimensional sparse data combining the effects of other regularization strategies of Least Absolute Shrinkage and Selection Operator i.e., LASSO [114] which uses l1-norm penalty (\( \alpha = 1 \ in \ Eq.2.5 \)) on coefficients and the ridge which considers the grouping effect of correlated features (l2-norm penalty with \( \alpha = 0 \ in \ Eq.2.5 \) ) [115]. This allows the ENet to leverage the tendency of LASSO to set many coefficients to zero and the capability of the ridge penalty to consider the correlated variables [113]. Due to this capability, the applicability of ENet suits in classifying microbiome samples having OTUs that have correlated patterns of abundance profiles. By
taking the advantages of coordinate descent optimization algorithm in the process [116], ENet could be used for classifying the data-intensive genomic domains. The study in [117] has successfully implemented ENet, LASSO, and ridge for a benchmark study. Nonetheless, the performance of ENet regularization, as assessed in the study by Knight et al. [69] is not significantly better than the other two methods of RF and SVM.

ENet can automatically select features while the model is constructed and hence saves the cost of explicit feature selection, unlike SVM in high-dimensional use cases. The advantage of the approach is that there are no limitations on the number of selected variables as part of its optimization strategy. Also, ENet provides an efficient solution to the regularization problem when a sparse solution is desired. It encourages grouping effect of variables in the case of highly correlated variables. However, in practice, it is advisable to try out several values of $\alpha$ and $\lambda$ (Eq. 2.5), regularization methods, while using cross-validation to guide the feature selection strategies [69], [117]. Regularization-based methods appear to be useful and efficient for high-dimensional metagenomic datasets, saving computational cost.

### 2.2.1.4 Multilayer Perceptron (MLP)

MLP is built upon an interconnected feed-forward network of nodes known as "neurons" and each edge in the network is annotated with weights (Figure 2.4) [118]. The architecture of MLP consists of: i) a set of inputs, ii) set of outputs, iii) several hidden layers and number of neurons in each hidden layer [118]. This structure allows the network to form a mapping between the input features
and the outcome of interest. Each neuron is associated with an activation function and a threshold value, computing output for other neurons at the next layer [118]. Various mappings in this MLP result in different network behaviours to generate an output value. Various mappings in this MLP result in different network behaviours to generate an output value. A weighted sum of \(n\) inputs \((x_1, x_2, x_3, \ldots x_n)\) is computed by the neuron by considering their corresponding activation weights \((w_1, w_2, w_3, \ldots w_n)\). If the calculated sum is above a given threshold, the output of +1 is generated otherwise an output of -1 is generated. This can be summarised in the following Eq.2.6.

\[
    f(x) = \begin{cases} 
    +1 & \text{if } w_1x_1 + w_2x_2 + w_3x_3 \ldots w_nx_n > \text{threshold} \\
    -1 & \text{otherwise} 
    \end{cases} \tag{2.6}
\]

MLP structure can be illustrated as a weighted directed graph, as shown in Figure 2.4. Particularly in metagenomic datasets, a set of OTUs are used as input layer variables; and the hidden layers optimise the weights of the input variables to predict different metagenomic traits [104].

![Diagram](image)

**Figure 2.4** A diagrammatic illustration of MLP Classifier with two hidden layers.
MLP could learn and model complex data relationships. MLP learn events and make decisions by considering similar events. Nevertheless, the parameters, such as hidden layers, threshold, and activation values in the network are usually set by a user. There are no specific rules for determining the structure of MLP. The solution obtained may depend upon unexplainable behaviour of the MLP. MLP proved to be computationally expensive and depends highly upon hardware configurations as well. Studies in [119],[120],[121] have also explored the use of other predictive models of Naïve Bayes (NB) [122]; Adaptive Boosting (AdaBoost) [123]; k-nearest neighbours (k-NN) [124] and Nearest Shrunken Centroids (NSCs) [125] for the benchmark analysis. Liu et al.[126], proposed a novel sparse distance learning method named MetaDistance for unbalanced multiclass metagenomic classification. This method combined k-NN and SVM learning methods to simultaneously maximise the interclass distance, feature selection, and to leverage the classification, especially in the case of multiclass environments with possible unbalanced classes. A summary of key publications encompassing supervised learning methods for functional metagenomic analysis is provided in Table 2.1.
Table 2.1 A review of studies involving predictive modelling to infer metagenomic functions using abundances of OTUs and reference to each study is listed; data source, related phenotypes, number of samples, taxonomic level, and size of feature space is mentioned; different classifiers used in respective study are listed, and best predictive model is reported over the benchmarks.

<table>
<thead>
<tr>
<th>Publication</th>
<th>Microbiome Source</th>
<th>Phenotypic Traits</th>
<th>Sample Size</th>
<th>Level of Study</th>
<th>No: of Features</th>
<th>Supervised Classifiers Used for Benchmark Studies</th>
<th>Best Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knights et al., 2011 [69],[93]</td>
<td>Human</td>
<td>Body Habitats of External Auditory Canal (EAC), Gut, Hair, Nostril, Oral cavity, and Skin (CBH)</td>
<td>612</td>
<td>Species</td>
<td>2741</td>
<td>RF, SVM, ENet, NB, NSC</td>
<td>NB</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Skin Sites such as the volar forearm, forehead, and palm (CSS)</td>
<td>401</td>
<td></td>
<td>2227</td>
<td>RF, SVM, ENet, NB, NSC</td>
<td>RF</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The niche of arms, hands, and Fingers (CS)</td>
<td>144</td>
<td></td>
<td>1592</td>
<td>RF, SVM, ENet, NB, NSC</td>
<td>RF</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fingerprints of three experimental subjects obtained over the keyboard (FS)</td>
<td>101</td>
<td></td>
<td>565</td>
<td>RF, SVM, ENet, NB, NSC</td>
<td>RF</td>
</tr>
<tr>
<td>Study Ref</td>
<td>Organism</td>
<td>Sample Type</td>
<td>Sample Size</td>
<td>OTU Count</td>
<td>Feature Type</td>
<td>Feature Count</td>
<td>Model Types</td>
</tr>
<tr>
<td>-----------</td>
<td>----------</td>
<td>-------------</td>
<td>-------------</td>
<td>-----------</td>
<td>--------------</td>
<td>---------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Liu et al., 2011 [126]</td>
<td>Human</td>
<td>CBH</td>
<td>815</td>
<td>101</td>
<td>Family, Genus</td>
<td>Identified 11 important OTU features</td>
<td>RF, SVM-RFE, ENet, NB, NSC</td>
</tr>
<tr>
<td>Statnikov et al., 2013 [98]</td>
<td>Human</td>
<td>CBH</td>
<td>552</td>
<td>565</td>
<td>Species</td>
<td>6979</td>
<td>SVM, SVM-RFE, k-NN, LR, RF</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CSS</td>
<td>140</td>
<td>565</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CS</td>
<td>357</td>
<td>565</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>FS</td>
<td>104</td>
<td>565</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>FSH</td>
<td>98</td>
<td>565</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>PBS</td>
<td>200</td>
<td>565</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Blaser Psoriasis (BP)</td>
<td>151</td>
<td>565</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gevers et al., 2014 [127]</td>
<td>Human</td>
<td>Crohn’s disease</td>
<td>1,359</td>
<td>104</td>
<td>Genus</td>
<td>9,511</td>
<td>RF, SVM</td>
</tr>
<tr>
<td>Ditzler et al., 2015 [104]</td>
<td>Soil</td>
<td>pH</td>
<td>22</td>
<td>104</td>
<td>Species</td>
<td>500</td>
<td>RF, MLP</td>
</tr>
<tr>
<td>Human</td>
<td>Body Site Niche</td>
<td>1,967</td>
<td>104</td>
<td></td>
<td></td>
<td></td>
<td>RF, MLP</td>
</tr>
<tr>
<td>Pasolli et al., 2016 [128]</td>
<td>Human</td>
<td>Body Site Niche</td>
<td>1,967</td>
<td>104</td>
<td>Species</td>
<td>500</td>
<td>RF, MLP</td>
</tr>
<tr>
<td></td>
<td>IBD</td>
<td>110</td>
<td>104</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Liver cirrhosis</td>
<td>232</td>
<td>104</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cancer</td>
<td>121</td>
<td>104</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Type II diabetes</td>
<td>344</td>
<td>104</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wassan et al., 2016</td>
<td>Cattle Rumen</td>
<td>Various</td>
<td>40</td>
<td>27</td>
<td>Phylum</td>
<td></td>
<td>RF, SVM</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Class</td>
<td>52</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Approaches to infer metagenomic functions using supervised learning

<table>
<thead>
<tr>
<th>Year</th>
<th>Study</th>
<th>Disease</th>
<th>OTU/ASV</th>
<th>Methods at Family Level</th>
<th>Methods at Genus Level</th>
<th>Methods at Phylum Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>2017</td>
<td>Diet with oil &amp; nitrate supplements</td>
<td>Human Obesity</td>
<td>18</td>
<td>LR, NB, k-NN, NN, MLP, AdaBoost</td>
<td>386</td>
<td>SVM, NN, AdaBoost</td>
</tr>
<tr>
<td>2017</td>
<td>Human Cancer</td>
<td>141</td>
<td>Species</td>
<td>NB, RF, LR</td>
<td>SVM, LR</td>
<td></td>
</tr>
<tr>
<td>2018</td>
<td>Human Diabetes, rheumatoid arthritis, and liver cirrhosis</td>
<td>806</td>
<td>Genus</td>
<td>RF, SVM, k-NN, LR, AdaBoost</td>
<td>300</td>
<td>SVM, LR</td>
</tr>
</tbody>
</table>

In a very recent study, Bahadorinejad et al. [129], proposed a Bayesian paradigm making use posterior probability distribution calculation [130] for characterization of 16S rRNA metagenomic sequences. The proposed approach is in a very nascent stage and could be explored in the future as a potential benchmark. One of the bottlenecks of all the above-discussed methods is that the application of none of these methods over OTU/ASV abundance count profiles can explain the hidden biological relationships between OTUs/ASVs. By using only, the abundance profiles of OTUs/ASVs to calculate inter-community similarities/dissimilarities/response to environmental functions; it assumes that all OTUs/ASVs are equally related to one another or are independent. However, phylogeny allows measuring the degree of similarity of two microbial species by their biological relatedness of evolution. In the light of studying relationships between various microbial features derived from 16S communities, researchers [69],[129],[131],[138] have further highlighted the
need of development of phylogeny-aware methods for functional analysis of metagenomic data. The methods developed in this context are listed below.

2.2.2 Phylogeny-aware methods for downstream analysis of metagenomes

Integration of microbial phylogeny in the field of functional metagenomics is an emerging domain, taking advantage from the assumption that phylogenetically similar microbial species tend to behave similarly [42], [83], [84], [132], [133]. To review current tools driven by phylogeny for functional metagenomic classification, they are categorised into homology-based, predictive model tuning-based, ranking-based, feature space transformation-based, and biostatistics-based approaches; in the scope of the current thesis. As shown in Figure 2.5, each of these category levels can be considered as a phylogeny-aware modelling approach for determining the metagenomic response to a functional phenotype. The popular approaches from literature, under each category, are discussed in this section highlighting their advantages and limitations.
Approaches to infer metagenomic functions using supervised learning

Figure 2.5 Categorical differentiation of phylogenetic tools for metagenomics downstream analysis.

2.2.2.1 Homology-based approach

The approach involves the use of an existing database of microbial reference genomes [15] for predicting the functional repertoire of 16S genes based on a phylogenetic tree and; use it for normalizing the OTU table. This creates a metagenome profile by using the phylogenetic tree of reference genomes.

i) PICRUSt

Langille et al. [134], proposed a method of Phylogenetic Investigation of Communities by Reconstruction of Unobserved States (i.e., PICRUSt), a predictive approach for determining the functional repertoire of 16S genes based on their taxonomic composition and the phylogenetic diversity of reference genomes derived from a phylogenetic tree. PICRUSt [134] supports a process for determining profile of a microbial community with the aid of
Approaches to infer metagenomic functions using supervised learning

following stages: - i) an initial gene-content inference step computing gene content for 16 S species from evolutionary ancestry of OTUs present in a reference phylogenetic tree (input) by ancestral state reconstruction [134]; ii) normalizing and linking OTU table with gene contents of obtained 16 S genes by dividing each OTU by its predicted 16S quantitative content; iii) metagenome inference step by linearly combining the microbial gene profiles of organisms from the reference database with the normalised OTU tables obtained in step ii). The microbial profiles created are subsequently mapped to the functional phenotype for functional classification. PICRUSt software package and related data are available at http://picrust.github.com/. PICRUSt was applied to a variety of datasets e.g. human microbiome sites, soils microbiome, other guts; to show the correlation between reference genomes and the Accuracy of PICRUSt [134]. The method provided useful biological insights. Nonetheless, PICRUSt is based on reference genomes only in Greengenes [15]. Hence, its limitation lies in the analysis of microbial communities that are under presented in such genome databases. Xu et al. [131], further studied whether analysis based on PICRUSt provided better or worse ability to classify samples with the application of RF [89] classifier for determining the functional roles of microbiome communities. Intriguing results were produced in their study, indicating that PICRUSt-predicted microbiome profile does not seem to improve the classification Accuracy over the benchmark datasets of CBH, CSS, FSH, FS, derived from the study by Knights et al. [69]. A significant limitation of this approach is its dependency on reference genomes. There exists an ongoing need for constructing extended
reference genome databases with a rate of evolution of microbiome. Another limitation is that the approach is computationally expensive.

2.2.2.2 Supervised model-based approaches

The approach involves the tuning of supervised ML predictive models of classification [135] with the use of phylogeny-aware parameters, for determining the microbiome functions.

i) Supervised ML Modelling Based on Phylogeny-driven kernel of SVM

Ning & Beiko [103], performed ML-based experiments for classifying nine subsites of the oral human microbiome [53] dealing with dental plaque using SVM, tuned with a phylogeny-based kernel. The authors employed weighted and unweighted UniFrac distances [50] derived from a phylogenetic tree to custom the kernel for SVM. A comparative analysis was performed with SVM kernels tuned with different non-phylogenetic kernels such as Euclidean and Canberra distances [136]; SVM with radial basis function kernel [108] over raw the OTU abundances; and PICRUSt [134]. The highest performance was achieved by ML framework of SVM with radial basis function kernel [103]. The authors [103] reported customizing kernels based on the UniFrac measure of phylogenetic similarity did not improve the overall performance of functional metagenomic classification over HMP oral data. It was also observed by Ning & Beiko [103], that despite the capability for PICRUSt to identify phylogenetic connections between taxa and clades, it did not improve the Accuracy of classification.
The study in [103] also reported an experiment using phylogeny in creating a feature space by combining an abundance of children OTUs to create ancestral profiles in clade lineages improved the classification performance with feature selection in this study [103]. SVM allowed authors in this study [103], to consider biologically informed encodings of the input data. The study suggests that clade-based feature encoding provided the most substantial increase in performance Accuracy. Instead of ML methods over input OTU features which are considered independent, it is beneficial to combine these features along the taxonomical lineage. The authors indicated that further work could explore a multitude of different other ML approaches that can be applied to such formulation of microbiome feature space.

ii) MetaPhyl

The approach in [87], proposed a multi-class metagenomic classifier “MetaPhyl”, regularizing ML model of multinomial LR [112] with a tree-based penalty function [137]. The penalty was derived from prior knowledge on biological relatedness of microbial species on a phylogenetic tree [45]. A tree shows the hierarchical relationship among the microbial features. The hierarchical structure of the phylogenetic tree captures the similarities and differences among the microbial features by representing the groups of similar sequences at the internal nodes. Thereby, Tanaseichuk et al. [87] incorporated this hierarchy in optimizing the coefficients of the LR model based on the maximum log-likelihood with the tuning of the phylogeny-driven tree-guided penalty function [137]. The penalty function was formulated with the weight associated with the node of a tree and reflects a correlation within the group of features descending from that node. Coefficients in the tree penalty
corresponded to multiple hierarchically overlapping groups. The authors in [87], indicated that using such hierarchal phylogenetic information in LR improves the classification Accuracy in comparison to state-of-the-art classification algorithms in literature such as RF [88], SVM [107], LR [112] and MetaDistance [126].

MetaPhyl considered input feature space of abundance profiles of OTUs defined at the leaves of a phylogenetic tree. However, microbial community abundance may also be characterised by lineages at varying phylogenetic depth from *phylum* to *genus*; and ancestral nodes may lead to better classification models for microbiome analysis as indicated in a study by Knights *et al.* [69],[138]. Hence, it is worthwhile to model features at multiple granularity levels of taxonomy rather than just at the leaf level of phylogenetic trees, in future work.

iii) **Regression Modelling based on Phylogenetic Levels**

A study in [85] suggested developing predictive models using the phylogenetically close OTUs along with the different levels of the tree by considering them to have similar effects on the phenotypic outcome of interest. The predictive model in this study comprised of microbial taxa at different levels of the taxonomy. The authors in [85] proposed tree-guided penalties in the regression model using the tree topology as well as branch lengths. The authors indicated internal nodes obtained from a phylogenetic tree as potentially important features.
iv) **Deep Learning ML Models based on Phylogenetic Measures**

Deep learning (DL) techniques have facilitated significant advances in analysing images, signals, and texts, often using Convolution Neural Networks (CNN) [139]. CNN works by learning filters that detect the local patterns. The application of DL techniques is needed to map input data in a meaningful way to kind of 1D or 2D images, to predict various functions. CNN tends to generate convolution layers with multiple feature maps dealing with the spatial information in training metagenomic data. Ditzler et al. [104], proposed that DL may not be suitable for metagenomics applications; however, recent attempts [140]–[143] have counter-arguments to the research. These attempts indicate that layers of phylogeny could be used to map them to CNN modelling. For example, T. Nguyen, 2018 [141], proposed an approach named Met2Img, involving colour coding scheme for ranking OTUs in taxonomic orders from phylum to species. The presence and relationships of OTUs were marked by the coloured pixels, and the absence was marked by white pixel. This coding scheme was used to generate feature space for metagenomic classification. Reiman et al. [142], on the other hand, proposed a framework PopPhy-CNN based on the architecture of CNN to predict functions from microbial abundance profiles [142]. The method used phylogenetic trees to indicate the spatial relationship of the microbes by embedding microbial abundance counts at each taxonomical level. A study in [143] evaluated the use of phylogenetic trees along with ASV abundance profiles in samples for the classification of a processed cattle metagenomics data as part of MetaPlat project [144] using CNN [145] for the functional classification of metagenomic profiles. CNN learned small filters that detect “local patterns” in the data. The idea behind the
approach was to determine a biologically pertinent sentence describing the path passing by through the phylogenetic tree from the root to that element. The abundance of each microbial feature was also added to the end of each sentence in terms of high, low, dominant. This way the feature set was generated to be passed to CNN filter. However, the cattle dataset was too small and acknowledged an issue of overfitting given a large number of trainable parameters and the small amount of training data available.

The development of more effective approaches using CNN is likely necessary as the current approaches' obscure the integration of phylogenetic relatedness fully. Also, the concern arises in the application of CNN, when there are more features than samples, which is often the case in predictive modelling of metagenomes.

In a recent study [146], Cascade Deep Forest (CDF) was proposed as an alternative to include spatial structure between nodes through the embedding of phylogenetic tree information. CDF proved better in cases of small-scale training data unlike CNN which require structured and large-scale data. LaPierre et al. [147], recently reviewed a critical evaluation of DL in predicting diseased metagenomes from HMP and concluded disease prediction as a challenging task, and; further observed that only minor or no performance improvements were noticed with the DL methods over the traditional methods. The authors also suggested that the scope exists for developing new design principles, improving the performance results, and providing valuable insights into the future.
2.2.2.3 Rank-based approach

The approach ranks and identifies the top-rated features from the microbial compositions, based on their phylogeny.

i) PhyloRelief

PhyloRelief proposed in [92], is driven by the Relief strategy [90] of selecting OTU features based on the phylogenetic weights annotated on tree branches. The approach progresses by associating phylogenetic weights with the clades (i.e., the branches connection two OTUs) and these weighted clades are ranked according to their contribution to the differentiation of the metagenomic samples. The workflow in PhyloRelief strategy comprises of two main conceptual steps: i) a scoring function based on Relief measure that ranks the clades of the subtrees in a phylogenetic tree, according to their participation in differentiating the microbiome samples, and ii) a merging function that merges nested subtrees and coalesces the corresponding clades sharing the similar tree-nodes [92]. Relief measure (Eq.2.7) identifies two nearest neighbour instances of the target microbial sample (S); one with the same class, termed as the nearest hit (H) and the other with the opposite class, called the nearest miss (M).

For each subtree t in phylogenetic tree T:

\[
W_t = W_t - \frac{\text{diff}(t,S,H)}{m} + \frac{\text{diff}(t,S,M)}{m}
\]  

(2.7)

where \(W_t\) is, the weight associated with subtree \(t\). In general \(\text{diff}(t,X,Y)/m\) is computed by calculating the UniFrac [50] distance between the sample \(X\) and \(Y\) in a subtree \(t\), where \(m\) denotes the total number of samples. In the case of unweighted UniFrac distance, the unweighted update function is utilised in the
exploratory analysis of the metagenomic samples, while the weighted update function is utilised in case of weighted UniFrac. With the development of the information contained in the weight $W_t$, independent clades (internal nodes) were ranked by identifying the sub-tree $t$ with the highest phylogenetic weight. This way, PhyloRelief provides an iterative procedure leading to the potentially better classification of the set of uncorrelated clades. The study [92] reported that application of PhyloRelief improved ability to identify microbial features in the analysis that are biologically important but are lost if phylogenetic relationships are not considered. The authors in [92], compared PhyloRelief coupled with RF to MetaPhyl [87] and RF without feature selection [89]. PhyloRelief showcased a high predictive value, which is competitive or better, in comparison to these state-of-the-art methods [87],[89]. The main improvement of this analytical method is, it defines the ranking of the microbial clades to determine significant OTUs based on their distribution amongst the samples; weighted by their annotated phylogenetic distance. Since the approach calculates sample similarity/dissimilarity based on the phylogenetic distances, the choice of a phylogenetic distance metric is critical. In the current scope, UniFrac [50] has been used as an exemplifier. These results encouraged more in-depth investigations on determining biologically relevant OTUs to predict functional phenotype from microbial compositions.

2.2.2.4 Feature space transformation-based approach

The approach enables the metagenomic analysis by structuring, normalizing, or transforming microbial feature space via the use of a phylogenetic measure to define a new set of features or dimensions for the downstream analysis.
i) **Phylogenetic Isometric Log-Ratio Transform**

Advancing on the methods to identify essential clades driven by phylogeny [92] that characterise similar patterns in the OTUs as per the metadata associated with samples. Silverman *et al.* [133] proposed Phylogenetic Isometric Log-Ratio (PhILR). The method explained the existence of inherently compositional nature of metagenomes yielding Isometric Log-Ratio (ILR) [148] system to transform the compositional space of metagenomic profiles into a new coordinate system with an orthonormal basis. This captured the evolutionary relationships between metagenomes well. The resulting coordinates in the new feature space were termed as “balances”. The balances obtained from an ILR transform typically represent the log-ratio of the geometric means of the corresponding data (induced by the edges of the tree) in their two subtrees under a given node while disregarding abundances in the respective remainder of the tree. The branch length information was used to linearly scale ILR balances by the phylogenetic distance between neighbouring clades. Following this, PhILR method identifies those clades in the phylogeny that characterise separating patterns in the sample meta-data, considering the compositional artefacts. Hence it serves as an alternative to PhyloRelief using the UniFrac distance metric [50]. Silverman *et al.* [133], focused on the various environmental and human microbiome, 16S rRNA datasets and revealed that PhILR transformed data overcomes the challenges associated with the compositional nature of OTU data, by including the branch lengths of the underlying phylogeny. PhILR transformed data significantly improved the Accuracy of supervised classification with LR in 7 of the 12 benchmark tests while compared to raw OTU abundance counts. It was observed that balances covary between more
phylogenetically similar clades and hence indicated the importance of evolutionary history in structuring the microbiome profiles. Overall, the approach served as efficient modelling to handle statistical artefacts over the compositional metagenomic data [133], [148]. The authors reported that a prior count normalization is required to deal with removing rare species, choice of weights, and the handling of zero values; before the applicability PhILR. However, a challenge does exist in the occurrence of the information loss during such kind of count normalization.

2.2.2.5 Biostatistics-based approach

The approach gains insights from statistical methods for discriminating microbial samples into functional phenotypes.

i) the adaptive microbiome-based sum of the powered score (aMiSPU)

Wu et al. [149] devised a new multivariate regression test-driven statistical test method called the adaptive microbiome-based sum of the powered score (aMiSPU), for studying an overall association between the composition of a microbial community and a functional phenotype. aMiSPU test is based on regularizing OTU abundance counts with the weights associated with microbial phylogeny, in a generalised regression framework. It involves the application of an adaptive test called aSPU test [150] driven by a family of a sum of powered score (SPU) tests, and subsequently incorporating the variable weighting obtained from UniFrac to regularize abundance values in the test [50]. Each SPU test controls the extent of weighting on the variables. The permutation scheme was followed to calculate the statistical \( p \)-value for the performance analysis. In comparison with other multivariate joint tests [150], aMiSPU was
proved to be convincing. aMiSPU test also supported the ranking of importance of OTUs and did perform much better at detecting true positives and reducing false positives unlike another state-of-the-art test [151]. The weighting scheme of aMiSPU is driven by UniFrac [50]. It's worthy of exploration of this intriguing approach further with different similarity measures based on phylogeny.

ii) Dirichlet-Tree Multinomial Regression Model

The work in [130], introduced the use of Dirichlet multinomial regression models (DMs) [152] into the microbiome analysis. The embedding of tree information into such regression models employed an alternative to imposing the tree-guided penalties on the parameters for LR based ML modelling over the structured sparsity in metagenomic data [87], [153]. A Dirichlet tree multinomial model (DTM) was introduced to aggregate the OTU abundance counts along the phylogenetic tree and to emphasize on each how an ancestral or internal node of the phylogenetic tree are statistically related to its child nodes.

The authors in [153] proposed a phylogenetic scan test (i.e. PhyloScan) for investigating microbiome compositions using the DTM model, cascading the local statistical models at each internal node of the phylogenetic tree and calculating statistical approximation on the scan statistic for each DTM. The statistics were defined over the triplets, formulated by a particular internal node, its parent, and one of its children [153]. This aided in investigating the compositional differences between cross-microbial communities well. The study in [154] emphasized that the descendants of each internal node on the phylogenetic tree share evolutionary characteristics, and it is meaningful to
apply local DTM at an internal node for associating covarying functions with sparse, over-dispersed, multivariate and tree-structured count data. This sets the stage for ML modelling over internal nodes covering information on shared taxonomy of OTUs. Nevertheless, the approach primarily is focussed on the taxonomy of internal nodes and tends to select the internal node with the statistics upon which all its descendant OTUs agree. In future, incorporating phylogenetic distances in addition to the hierarchal taxonomy may prove beneficial to estimate biological relationships.

2.2.2.6 Subcategorization of different phylogenetic approaches

Overall, by the actual state of integrating phylogeny, the above-discussed approaches could be further categorised into three ways of phylogenetic integration extending on Figure 1.6. to construct Figure 2.6.
Approaches to infer metagenomic functions using supervised learning

**Figure 2.6** Phylogeny-aware approaches characterised into three different ways to integrate phylogeny in predictive ML modelling.

The approaches depending on only taxonomy would ignore the evolutionary distances ((A) in Figure 2.6), which may otherwise prove useful in predictive modelling. The approaches calculating sample to sample similarity/dissimilarity based on the phylogenetic distances would require a critical choice of a similarity/dissimilarity measure for calculating distances ((B) in Figure 2.6). On the other hand, approaches using direct evolutionary distances annotated on branches of the phylogenetic tree are useful in creating a feature space of environmental microbial communities using their biological assemblages ((C) in Figure 2.6). The current thesis gains insights from Figure 2.6, to model microbiome features and predictive learning using both the taxonomy and evolutionary distances as driven by phylogeny, intending to gain insights from (A), (B) and (C).
Figure 2.7 gives a summary of the timeline of the emergence of popular phylogeny-aware ML methods for functional metagenomics with valuable insights and their possible limitations. The studies (discussed in section 2.2.2) have indicated that including phylogeny may or may not improve significantly in the field of functional metagenomics. It depends on an intriguing way of incorporating phylogeny in functional metagenomics (Figure 2.6). The scope of the current thesis lies in further expanding the use of the taxonomic and phylogenetic resolution of 16S rRNA sequences in functional metagenomics. The current research extends upon the insights that constructing microbial features at different levels of taxonomy may prove useful rather than performing analysis at leaf level OTUs.
Approaches to infer metagenomic functions using supervised learning

Figure 2.7 Timeline of Popular phylogeny-aware approaches; PICRUSt [134], MetaPhyl [87], A study by Ning & Beiko [98], PhyloRelief [92], aMiSPU [149], PhILR [133], DL methods such as PopPhy-CNN [142], A study by Wang et al. [85]; used in predictive ML modelling of functional metagenomes with useful insights and limitations.
2.3 **Feature selection in metagenomic studies**

Metagenomic studies are influenced by the choice of microbial features and their functional capabilities across different samples. A natural question to ask is “which microbial features are significant for differentiating the functional roles in metagenomic analysis in high-dimensional metagenomic data?” Feature selection is usually a precursor step to the predictive analysis of the microbiome for improving the quality of predictions [128]. Three standard techniques for feature selection are: i) filter-based feature selection (FFS) [155] that evaluate the worth of features by heuristic functions over general statistical characteristics of data; ii) wrapper-based feature selection (WFS) techniques [34],[156] evaluating the quality of features by iterating an ML algorithm over the input features and; iii) embedded (EM) methods that learn the features contributing best towards attaining higher performance while the model is being constructed [157]. The functionality of these three applied techniques is differentiated in Figure 2.8 for any future reference in this thesis.
Based on the methods shown in Figure 2.8, feature selection has been carried out in metagenomic analysis by studies such as in [120], [128], [158]–[163].

2.4 Current limitations of predictive approaches and future scope

The predictive approaches discussed in section 2.2., differ in terms of data sources employed, feature selection strategies, supervised learning techniques, predictive tasks, and evaluation methods. Such an example of phylogeny-aware approaches is indicated in Table 2.2 with varied data sources and ML methods with their assessments. Due to these differences, it is difficult to compare the approaches directly.
Table 2.2 The widely used phylogeny-aware methods in each category (Figure 2.5) to infer metagenomic functions are listed. Reference to each study is listed with their implementation source; datasets used are summarised, performance and various insights are reported over the data in use.

<table>
<thead>
<tr>
<th>Phylogenetic Tool and Source</th>
<th>Datasets Under Use</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PICRUSt [134]</strong></td>
<td>Diverse Metagenomic Use Cases of mammalian guts, soils from different geographic locations and a hypersaline mat community. The study used PICRUSt to generate biological insights from: ✓ i) 6,431 samples from the HMP to differentiate supragingival and subgingival plaque [164] ✓ ii) 335 coral mucus samples to study algae growth and its role in eutrophication [165] ✓ iii) 993 samples from time courses covering ~16 weeks each from the vaginal microbiomes of 34 individual subjects [166]</td>
<td>Performance Metric: Used classification Accuracy with Nearest Sequenced Taxon Index (NSTI) [134] and Spearman Correlation coefficient [167] to predict the agreement of abundances of metagenomes with different environments. Conclusive Results: The application of PICRUSt indicated that phylogenetic information derived from 16S marker gene sequences correlates the genomic content to yield accurate functional predictions better in the case when related reference genomes are available in databases. PICRUSt neither precludes nor outperforms the classification of samples from a wide range of habitats.</td>
</tr>
<tr>
<td><strong>MetaPhyl [87]</strong></td>
<td>✓ Simulated datasets of 2, 3, 5 and 10 classes with 20 samples for each class ✓ Three real-world datasets were used in the study. ✓ CBH; Turnbaugh et al. [64] dataset contains human gut samples having lean, obese, and overweight physiological states and; Human Microbiome Project Consortium, 2012, containing samples from five body habitats: oral, nasal, skin,</td>
<td>Performance Metric: Classification error rates (%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Performance Benchmarks: Compared the error rates (%) of proposed method benchmarked with LR, SVM, RF and MetaDistance classifiers on datasets generating features at different OTU similarity cut-offs and resolution levels in QIIME [20].</td>
</tr>
<tr>
<td><strong>PhyloRelief</strong> [92]</td>
<td>Implemented in Python (source available at <a href="http://compmetagen.github.io/phylorelief">http://compmetagen.github.io/phylorelief</a>)</td>
<td>✓ CBH dataset differentiating forehead (FH) vs. external nose (EN); volar forearm (VF) vs. popliteal fossa (PF) [69]; ✓ FS[69] ✓ IBD (cases and controls) [168]</td>
</tr>
<tr>
<td><strong>Phylogenetic approaches to microbial community classification</strong> [98]</td>
<td>✓ In this study, 2706 human oral cavity samples from nine oral sites were collected from the HMP database [53].</td>
<td>Performance Metric: Classification Accuracy.</td>
</tr>
</tbody>
</table>
### Current limitations of predictive approaches and future scope

| aMiSPU [149] R package (https://github.com/ChongWu-Biostat/MiSPU and CRAN) | √ Human throat microbiome dataset for smoking effects with 60 samples and 856 OTUs [172]. √ Human gut microbiome for IBDs such as Crohn’s disease (CD) and ulcerative colitis (UC) in 40 twin pairs [173] | Improved the predictive accuracy. Feature selection played an important role in such functional classification. |
| PhILR Transform [133] R package (https://bioconductor.org/packages/devel/bioc/vignettes/PhILR/inst/doc/PhILR-intro.html) | ✓ The study focused on 3 primary datasets: i) Costello Skin Sites (CSS), a dataset of 357 samples from 12 human skin sites [69]; ii) Human Microbiome Project (HMP), a dataset of 4743 samples from 18 human body sites (e.g., skin, vaginal, oral, and stool) and [53], iii) Global Patterns (GP), a dataset of 26 samples from nine human or environmental sites [174] | Performance Metric: The tests significance (quoting p-values) in accepting or rejecting the null hypothesis (i.e., no association between taxon and environmental outcome) | Performance Benchmarks: MiRKAT yielded p values from weighted UniFrac, unweighted UniFrac, and Bray–Curtis has driven kernels to show the test significance whilst benchmarking with aMiSPU. |
| | | Conclusive Results: Combining taxon abundance information with UniFrac [50] provides an efficient computational approach to perform covariate analysis with the outcome of interest (phenotype) in a wide range of environmental applications. | |
| | | Performance Metric: PERMANOVA R² values [175], for measuring sample pairwise distances between metagenomic samples. Accuracy of supervised classification methods was tested on benchmark datasets with raw-relative and log-relative OTU abundances. | Performance Benchmarks: Distances were calculated by using weighted and unweighted UniFrac, Bray-Curtis, Binary Jaccard, and Euclidean distance [51] Models used for benchmark classification were: - LR, SVM, k-NN, RF |
The choice of a single "best tool" is not straightforward in analysing functional metagenomes. Depending on the structural and functional compositions, varying levels of predictive performance have been reported by these abundance and phylogeny-aware methods of metagenomic analysis in their publications [69], [98], [103], [133], [142]. No accepted gold standard method exists within this field of functional metagenomics except RF, which has been reported as one of the favourable methods for functional metagenomics.

| Deep Learning with PopPhy-CNN [142] | ✓ The study used HMP microbiome data [53] sets related to human diseases: cirrhosis (114 cases vs. 118 controls); (2) type 2 diabetes (223 cases vs. 217 controls), and (3) obesity (164 cases vs. 89 controls) | Performance Metric: Area Under Receiver operating characteristic curve (ROC-AUC) | Performance Benchmarks: Compared to classical algorithms of SVM, RF, and LASSO | Conclusive Results: PopPhy-CNN (with 2D-CNN), models outperformed the other methods. The 1D-CNN model was as competitive as RF and performed better than SVM and LASSO. However, there is an urgent need to extend it to large-scale metagenomes.

Implemented in Python using Theano library with source available at [https://github.com/derekreiman/PopPhy-CNN](https://github.com/derekreiman/PopPhy-CNN) |
over raw OTU abundances (Table 2.1). The high dimensionality of the metagenomic abundance data, along with the sparse nature of such data and; the probable existence of high variability in species of a microbial community makes the use of classical models of ML more challenging.

Recently developed methods for high-dimensional datasets such as XGBoost [176] have not been explored much in the field of metagenomics which deals with high-dimensional microbial features in functional analysis. Integrating domain knowledge of phylogeny into the microbial analysis is also facing computational and performance challenges, and the field is until emerging. Ning & Beiko [103] and Xu et al. [131], presented intriguing results indicating that adding phylogenetic information did not improve classification Accuracy while deriving functional phenotype from the microbiome profiles. An example of comparative analysis by Xu et al. is indicated in Figure 2.9.

Figure 2.9 A Comparative Analysis with Kappa Statistic (κ) [177] between i) RF applied over raw abundances and ii) RF applied over PICRUST scaled data; in the Study by Xu et al. , 2014 [131]; over public datasets CBH, CS, CSS, FSH and FS used in [69].
A key finding of the study by Wassan et al. [163], also indicated that ML models of MetaPhyl [87] and PhILR [133] did not necessarily produce the best results when compared to non-phylogenetic predictive models, that are purely based on the abundance of microbial species. Due to inadequacies exhibited by both the abundance and phylogeny-aware experimental methods, advancements in computational methods are needed in the prediction of metagenomic functions. Hence, this research aims to identify and explore more accurate alternative ML models for classifying microbiomes by further progressing on these two primary directions: i) abundance-aware analysis and ii) phylogeny-aware analysis. Construction of novel methods and their use in predictive modelling should provide the performance gains over the existing methods or a new interpretation of modelling biological features for functional predictions.

Phylogenetically similar species tend to function similarly in the microbial ecosystem. However, how should the integration of phylogenetic measure be designed in predictive modelling, is still an active area of research, with a plethora of opportunities. Based on the combination of strengths of (A), (B) or (C) ways in Figure 2.6, construction of novel approaches may play an essential role in functional metagenomics. The incorporation of the biological and structural context has the potential to provide more realistic and significant modelling for predicting functions in the field of metagenomics.

2.5 Summary

In this chapter, a subset of relevant and accessible approaches to the predictive modelling of metagenomes in literature are discussed. The review in this chapter
highlights the essential development of computational methods for the predictive analysis of vast quantities of metagenomic data being generated by NGS pipelines. Several supervised learning techniques have been reviewed for predicting metagenomic functions from the microbial compositional profiles. However, they face limitations that require addressing. Overall, the RF method over the abundance count profiles of the microbiome is proposed as one of the favourable methods for inferring metagenomic predictions (Table 2.1). The ability to predict metagenomic functions with prior biological domain knowledge of phylogeny makes predictive modelling more attractive. Evolutionarily related microbes by phylogeny, tend to have similar lifestyle patterns, reflecting their ability to share the environmental traits as supported in a study by [41]. By partitioning the literature into distinct classes of methods, i.e., abundance-driven and phylogeny-driven methods emphasize their application in functional metagenomics. The main findings from the literature that forms the motivation of studies in this thesis are:

- OTUs/ASVs are related to one another by a hierarchical structure as derived from a phylogenetic tree [40].

- Distance measures that use tree information, such as UniFrac [50] are effective at figuring out ecological patterns better.

- The phenotypic trait may be characterised better by considering several taxa (internal nodes and leaf nodes both) of varying phylogenetic depth [85],[138].
• The studies in this[103], [131] indicated computational performance does not always improve with designed phylogenetic models which otherwise are expected to produce more significant biological results.

• More efficient modeling would be preferred over large-scale high-dimensional abundance counts of metagenomes.

**Figure 2.10 Areas addressed in the current dissertation based on the observations made from the literature.**

This dissertation is focused on improving the current-state-of-the-art in predictive modelling of high-dimensional metagenomes. The chapter highlights the areas in which more research could be applied and how these are addressed by contributions highlighted in this thesis are summarised in Figure 2.10.
The next chapter provides an overview of the conceptual framework for the interpretation of essential components involved in predictive modelling of metagenomes. The general demonstration of the standard framework would pave a path to building novel computational algorithms for a detailed understanding of the linkage of microbiome profiles with their functions. This may provide potential contributions to the field. In this thesis, several diverse metagenomic data sources are selected, pre-processed, and integrated using ML to infer metagenomic functions. The next chapter would also list these data sources used for data modelling and analysis throughout this thesis.
CHAPTER 3

RESEARCH METHODOLOGY AND MATERIALS

This chapter aims to formalise a conceptual methodological framework for inferring the functional potential of the microbiome. It outlines the fundamental stages involved in the conceptualized framework. Three main research approaches are suggested along with the conceptual framework. The subsequent chapters in the thesis will expand on this conceptual framework to detail its application and propose novel methods in the field of functional metagenomics. The current chapter also describes the diverse data sources utilised throughout the thesis for the inference of metagenomic functions.

3.1 Conceptual framework

The description of the conceptual framework for functional predictions in metagenomic studies is divided into four main steps, as illustrated in Figure 3.1. These stages are (a) representation of the metagenomic input data (features); (b) pre-processing of data using techniques such as feature engineering or transformation or feature selection; (c) inference of metagenomic functions using ML classification techniques and (d) assessment of the predictive performance of ML models employed. The purpose of this section is to describe
notations and present the whole framework further in detail for determining the functional repertoire of microbial compositions throughout this thesis.

Figure 3.1 Graphical representation of the methodological framework for the prediction of metagenomic functions. (a) Interpretation of heterogeneous data of abundance count table and the phylogenetic tree, being sourced from metagenomic pipelines (b) Feature engineering is performed to identify important taxa. (c) Selected features are input to classification algorithms for predicting metagenomic functions (d) The predictive performance is evaluated using different assessment measures.
3.1.1 Description and presentation of data inputs

The source biological 16Sr RNA sequence data is transformed into a heterogeneous feature space providing useful information for linking genotype (metagenomes) to the phenotype (environmental functions). The microbial sequences are laid into specialized data structures of (a) abundance count matrix, and (b) a phylogenetic tree for their downstream analysis. The characterization of microbial communities by 16S rRNA gene sequencing is possible by using accessible pipelines such as QIIME [20], dada2 [26],[178], or mothur [19] (Table 3.1). These pipelines could provide information on the number of metagenomic features and the topological structure connecting various features biologically. The interpretation of obtained microbial feature space with tools such as phyloseq[179], vegan [180], and ape R package [181] (Table 3.1) supports a further comprehensive analysis.

Although microbiome profiles can be learned from 16S data using a variety of methods (some listed in Table 3.1), it is assumed that this information is known a priori in the contextual analysis presented in this thesis. The topological structure of phylogeny was read and interpreted with the help of ape package[181], and; phyloseq allowed us to perform operations on abundance counts, phylogenetic tree, taxonomy, and the metadata; as a comprehensive integrative object [179]. Profiling metagenomic feature attributes is an essential factor to take into consideration as it may affect the quality of functional predictions. The two primary profiling measures serving as inputs to the downstream analysis in the studies conducted through this thesis are - i) using
abundance profiles of OTUs (quantitative method) and; ii) integrating biological domain knowledge of the phylogeny of OTUs (qualitative method).

Table 3.1 A list of popular tools from the literature that is useful for data encoding of 16S rRNA data.

<table>
<thead>
<tr>
<th>TOOL</th>
<th>Description</th>
<th>Source URL</th>
<th>Type</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>QIIME</td>
<td>16S sequences are clustered into OTUs using open reference OTU picking method in QIIME at certain similarity threshold (usually 97 %). The RDP [182] Classifier and UCLUST [183] implemented in QIIME are used for defining the taxonomic classification of 16S rRNA genes using the Greengenes database [15].</td>
<td><a href="http://qiime.org/">http://qiime.org/</a></td>
<td>Comprehensive Tool Suite Pipeline</td>
<td>[20]</td>
</tr>
<tr>
<td>dada2</td>
<td>Sequence variances are detected. The end-product is Amplicon Sequence Variants (analogous to OTUs).</td>
<td><a href="https://bioconductor.org/packages/release/bioc/html/dada2.html">https://bioconductor.org/packages/release/bioc/html/dada2.html</a></td>
<td>R package</td>
<td>[178]</td>
</tr>
<tr>
<td>phyloseq</td>
<td>OTU tables, phylogenetic tree, taxonomy, and sample data are combined into a comprehensive phyloseq objects for downstream analysis.</td>
<td><a href="http://www.bioconductor.org/packages/release/bioc/html/phyloseq.html">http://www.bioconductor.org/packages/release/bioc/html/phyloseq.html</a></td>
<td>R package</td>
<td>[179]</td>
</tr>
<tr>
<td>vegan</td>
<td>Consists of useful tools for detecting differences in microbiome community structure amongst the microbial samples</td>
<td><a href="http://cran.r-project.org/web/packages/vegan/index.html">http://cran.r-project.org/web/packages/vegan/index.html</a></td>
<td>R package</td>
<td>[185]</td>
</tr>
<tr>
<td>ape</td>
<td>Useful for reading and interpreting phylogeny.</td>
<td><a href="https://cran.r-project.org/web/packages/ape/index.html">https://cran.r-project.org/web/packages/ape/index.html</a></td>
<td>R package</td>
<td>[181]</td>
</tr>
</tbody>
</table>
The following sub-sections set up some notations and provide useful insights on both abundance-driven and phylogeny-driven profiling.

### 3.1.1.1 Abundance-based profiling with OTU counts

Considering a microbiome dataset with \( n \) OTUs denoted by a set \( \Omega = \{OTU_1, OTU_2, \ldots, OTU_n\} \) in \( m \) number of samples obtained from a sample set \( S = \{S_1, S_2, \ldots, S_m\} \); an abundance count matrix \( X \) is usually annotated, as shown in Eq.3.1.

\[
X(m,n) = \begin{bmatrix}
x_{1,1} & \cdots & x_{1,n} \\
\vdots & \ddots & \vdots \\
x_{m,1} & \cdots & x_{m,n}
\end{bmatrix}
\]  

(3.1)

where ‘\( x_{i,j} \)’ represents the abundance count of \( j^{th} \) OTU in the \( i^{th} \) sample; \( 1 \leq i \leq m \); \( 1 \leq j \leq n \). A set of phenotypic labels \( L = \{L_1, L_2, \ldots, L_t\} \) is associated with the abundance count matrix to create a full microbiome predictive profile for the downstream modelling. Table 3.2 illustrates the quantitative microbial profile comprising of OTU counts for each group of \( m \) samples with phenotypic metadata (where \( t \ll m \)).

<table>
<thead>
<tr>
<th></th>
<th>( OTU_1 )</th>
<th>( OTU_2 )</th>
<th>( \ldots )</th>
<th>( OTU_n )</th>
<th>Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>( S_1 )</td>
<td>( x_{1,1} )</td>
<td>( x_{1,2} )</td>
<td>( \ldots )</td>
<td>( x_{1,n} )</td>
<td>( L_1 )</td>
</tr>
<tr>
<td>( S_2 )</td>
<td>( x_{2,1} )</td>
<td>( x_{2,2} )</td>
<td>( \ldots )</td>
<td>( x_{2,n} )</td>
<td>( L_2 )</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( S_m )</td>
<td>( x_{m,1} )</td>
<td>( x_{m,2} )</td>
<td>( \ldots )</td>
<td>( x_{m,n} )</td>
<td>( L_t )</td>
</tr>
</tbody>
</table>

More recently, efforts have also been made to construct Amplicon Sequence Variants (i.e., ASVs) [24], as another standard unit of marker-gene analysis. ASVs reportedly resolve the differences in biological genomic sequences at a
low level (such as single nucleotide), unlike imposing the arbitrary dissimilarity thresholds that define molecular OTUs. The authors in [24] proposed that ASVs have a similar intrinsic biological meaning as an RNA genomic sequence and are identified independently from a reference database, unlike OTUs.

The features (i.e., OTUs or ASVs) usually exceed the number of samples in abundance profiles (n >> m). Hence, feature selection or transformation becomes essential as the next step in their downstream analysis. The count data contains various low-abundant taxa. However, it is inappropriate to infer that the low-abundant taxa are not necessary or important in microbiome classification. The computational methods relying on low abundant taxa may not understand functions correctly if a biologically relevant change occurs during evolution in less abundant lineages. Henceforth, including a phylogenetic measure for the estimation of biological relationships between different microbial features is significant.

3.1.1.2 Knowledge-based profiling

The information on evolutionary distances obtained from tree topology is useful in recording the amount of shared branch length between each pair of OTUs and in the formulation of their importance in downstream analysis. Considering, T as a rooted full binary phylogenetic tree describing the evolutionary relations of the n OTUs; it consists of \{LN_i\}, 1 ≤ i ≤ n representing a set of leaf nodes of T with each leaf corresponding to an OTU; \{IN_j\}, 1 ≤ j ≤ n-1 denoting the collection of internal nodes (i.e., non-leaf nodes), each having two children, and the root contains lineage information from all the OTUs (shown in Figure 3.2).
Phylogeny enriches understanding of metagenomes, adding biological meaning to data, and allowing a better understanding of biological functions.

Figure 3.2 An illustration of a binary rooted phylogenetic tree.

Phylogenetic profiling is exploited in the below-mentioned contexts in the scope of the current thesis.

i) A phylogenetic tree is an informative tool to define a biologically meaningful distance measure. If the two sets of OTUs are closely related and share much of their evolutionary traits, the distance between them would be small whereas OTUs located far apart on the phylogenetic tree, would have a larger distance between them highlighting their unique evolutionary history. This could prove useful in designing a phylogeny-informed feature space.

ii) The tree provides a hierarchical ancestry of the OTUs and hence allows to make use of several OTU lineages at different levels of the taxonomy (*species*, *genus*, *family*, *order*, *class*, or *phylum*) in modelling the downstream analysis. Several microbial lineages sharing common ancestors at different depths could affect phenotype.
The tree may prove useful for phylogenetically informed feature selection. Phylogenetically close OTUs are highly correlated and tend to have similar genetic traits. They behave and function in the same manner and share biological mechanisms [42],[83]–[85]. Hence, they tend to be selected together. Also, when two OTU features have a high correlation and similarity based on phylogeny, one of the two features could be dropped to make the learning algorithm faster.

These three contributing factors have prompted the need for the development of integrative computational methods for the detection of metagenomic functions in the context of the current thesis.

3.1.2 Data pre-processing

The thesis applies pre-processing techniques to the data inputs for attaining an appropriate final subset of microbial features from the metagenomic datasets. This aims to maximise predictive performance while training a computational ML model in the downstream analysis. Pre-processing techniques primarily include: (i) data formatting and trimming to determine the intersecting indices such that both data structures of abundance count table and phylogenetic tree describe precisely the same taxa; (ii) data transformation method to reduce the between-sample variabilities in input data [133], [186], [187] or; (iii) data filtering with feature selection strategies to extract useful microbial features from high-dimensional metagenomes [188]. Different possible ways employed in this thesis for processing the data before the application of supervised ML are listed below.
a) The taxa labels are required to be matched between phylogenetic (i.e., the tips/leaves of the tree) and the OTU table (i.e., columns of OTU tables containing abundances). Hence, the trimming function of the `drop.tip()` available in R package `ape` [181] is useful in determining common taxa and pruning a phylogeny tree to match the corresponding OTUs in abundance count data.

b) Different data inputs may cover a diverse range of biological coverage and subtle correlations may exist between features. Henceforth, the integration of diverse heterogeneous data types is preferred to transform the abundance feature space for improving the inference of metagenomic functions.

c) Data filtering is applied in this thesis for removing irrelevant and redundant features from the high-dimensional metagenomic datasets [189]. Three standard techniques are utilised through the thesis: i) filter-based techniques (FFS) [155] evaluated the worth of features by heuristic functions over general statistical characteristics of data; ii) wrapper-based (WFS) [34],[156] techniques evaluated the quality of features by iterating an ML algorithm over the input features and; iii) embedded (EM) methods learned the features contributing best towards attaining higher performance while the model is being constructed [157].

This section further gives an overview of what different feature selection strategies in each category has been used for extracting meaningful features from microbiome datasets throughout this thesis (as shown in Figure 3.3).
Figure 3.3 Feature selection methods falling within each category utilised in this thesis across various studies.

(i) FFS Methods

- Information Gain (IG)

IG method depicts how much “information,” a feature in the input data, tells about the outcome of interest. Features that perfectly partition the phenotypic class labels provided maximal information, whereas features unrelated with class gives no information. This information is derived by calculating the entropy (Eq.3.2) over the set of data samples $S$ with $n$ distinct classes $[190],[191]$.

$$H(S) = - \sum_{c=1}^{n} p_c \log_2 p_c$$

(3.2)

where $p_c$ denotes the proportion of the input samples belonging to class $c$ (ranging from 1, 2, ..., $n$ classes). Assuming a feature, $F$ having $v$ possible values and $S$ be the set of samples $\{X\}$. $S_v$ denotes the subset where $X_F = v$. The entropy-based on the partitioning into subsets by $F$ is given by (Eq.3.3).

$$E(F) = \sum_{v \text{ belongs to values of feature } F} \frac{S_v}{S} \ast H(S_v)$$

(3.3)

IG is calculated as a change in entropy or mutual information between the feature $F$ and class labels of $S$, as shown in Eq.3.4 $[190]$. 

(i) FFS
- Correlation-based Filtering
- Information Gain
- Relief-based Measure

(ii) WFS
- Using RF
- Using LR

(iii) EM
- Penalized Logistic Regression
- Gradient Boosting
- Random Forest Importance
\[ IG(S, F) = H(S) - E(F) \] (3.4)

A feature with high IG value is considered more important for classifying the data.

- **Correlation-based Filtering (CFS)**
  
The rationale behind CFS is: “a good feature subset contains features that are highly correlated with the class (i.e., the outcome of interest), yet uncorrelated with each other” [192]. CFS evaluates a heuristic score of a subset of input features by considering the degree of redundancy between them (as indicated in Eq.3.5) [191], [192].

\[
CFS_{\text{Subset}} = \frac{n \cdot r_{cf}}{\sqrt{n + n \cdot (n-1) \cdot r_{ff}}} \] (3.5)

where \( CFS_{\text{Subset}} \) is the score of a feature subset \( S \) containing \( n \) features; \( r_{cf} \) is the mean feature to class correlation (\( f \) belongs to \( S \)), and \( r_{ff} \) is the average feature to feature inter-correlation. In the current thesis context, feature subset search is executed by the best first search strategy [193] heuristics employing back-tracking the feature search path for finding feature subsets till noticeable results are attained [186],[192].

- **Relief-based Measure (RBS)**
  
Kira and Rendell [194] suggested the use of the RBS feature weighting algorithm to rank features according to how their quantitative values contribute towards differentiating between instances of different classes. The feature filter in RBS is particularly an instance-based measure.
It repeatedly chooses a random sample instance from the input data and then tends to locate its nearest neighbour of the same class, and that of different classes to calculate the weighted scores or a proxy statistic for each feature [90]. Feature scores estimate the quality of features. The summary of the score calculation is shown in Eq.3.6. The features containing similar scoring information are treated as redundant features and needs to be filtered [90].

\[
W_f = W_f - \frac{\text{diff}(f, S, H)}{m} + \frac{\text{diff}(f, S, M)}{m}
\]  

(3.6)

where \(S\) is the selected instance; the other two nearest neighbour instances of the target; the instance the same class known as nearest hit \((H)\), and the instance with the opposite class the nearest miss \((M)\); \(\text{diff}\) function calculates the distance between instances when finding nearest neighbours; \(m\) represents the number of random training instances out of total samples used to update \(W_f\), the feature score value.

(ii) **WFS Methods**

WFS methods search for the best feature with regards to an ML model performance and successively add them to create a feature subset. WFS methods utilised popular state-of-the-art RF and LR in the context of the current thesis. Additionally, Recursive Feature Elimination (RFE) with RF served as a greedy optimization with the WFS method, which constructs successive model iterations with the best features from the previous model iteration. This way, WFS methods aided in ranking features in order of their elimination.
(iii) **EM Methods**

- **Penalized LR Based Feature Selection**

Regularization methods introduce additional constraints known as penalties into the predictive model of generalised LR by choosing the parameters that maximise the log-likelihood of observed sample values (Eq.3.7) to lower the complexity of the model [112], [116].

\[
P \left( Y = \frac{1}{x} \right) = \frac{1}{1 + e^{-h(x)}}
\]  

(Eq. 3.7)

where \( h(x) = w_0 + \sum_{j=1}^{d} w_j x_j \), and \( d \) represent the number of dimensions/features; \( w_j \)'s are the regression coefficients and \( x_j \)'s are input features. The three penalties of Least Absolute Shrinkage and Selection Operator (LASSO), Elastic Net (ENet), and Ridge are utilised in this thesis. LASSO and ENet penalties perform inherent feature selection by setting the coefficients of weak taxonomical features to zero. This also aids in attaining a solution to the sparse metagenomic data by increasingly setting coefficients to zero in case of many features than samples. The Ridge penalty set up similar coefficients for abundance-wise correlated taxa.

The three penalties are shown in Eq.3.8a-c.

\[
\text{LASSO} = \arg\max_w \sum_{k=1}^{n} (y_k(h(x)) - \log(1 + h(x))) - \lambda \sum_{j=1}^{d} w_j
\]  

(Eq. 3.8a)

\[
\text{Ridge} = \arg\max_w \sum_{k=1}^{n} (y_k(h(x)) - \log(1 + h(x))) - \lambda \sum_{j=1}^{d} w_j^2
\]  

(Eq. 3.8b)

\[
\text{ENet} = \arg\max_w \sum_{k=1}^{n} (y_k(h(x)) - \log(1 + h(x))) - \lambda \sum_{j=1}^{d} (\alpha w_j + (1 - \alpha) (w_j^2))
\]  

(Eq. 3.8c)
where \( j \) is the number of features; \( \lambda \) controls the overall strength of the penalty to avoid overfitting and \( \alpha \in [0,1] \); \( \alpha = 0 \) implies Ridge and \( \alpha = 1 \) implies LASSO.

- **eXtreme Gradient Boosting (XGBoost)**

The boosting method of XGBoost [195], [196] serve as a sparsity-aware algorithm that provided scalable tree boosting for the prediction of metagenomes under the Gradient Boosting framework, as proposed by Friedman [197]. XGBoost comprises of a linear boosting model solver and a tree ensemble learning algorithm which used \( n \) additive functions over the training input vectors \( x_i \) (with multiple features) to predict the output phenotype target \( Y_i \). The XGBoost model is represented in (Eq.3.9).

\[
Y_{in} = \sum_{n=1}^{N} f_n (x_i), \quad f_n \in F, \tag{3.9}
\]

where \( i \) is the number of samples; \( N \) is the number of trees; \( f \) is the function in \( F \), and \( F \) is a regression tree space in which each regression tree is a function that maps the feature attributes to a score value to judge its importance. The additive training in the XGBoost procedure henceforth involves the addition of new function each time, as shown below (according to Eq. 3.9).

\[
Y_{i0} = 0 \\
Y_{i1} = f_1(x_i) = Y_{i0} + f_1(x_i) \\
Y_{i2} = f_2(x_i) = Y_{i1} + f_2(x_i) \\
\ldots \\
Y_{iT} = \sum_{n=1}^{T} f_n (x_i) = Y_{iT-1} + f_T(x_i)
\]
The objective of XGBoost is to find \( f_T \) to minimise the following (Eq.3.10)

\[
\sum_{n=1}^{N} l(Y_{iT}, Y_{iT-1} + f_T (x_i)) + \Omega (T) + \text{constant}
\]  

(3.10)

where \( l \) represents a training loss function, and \( \Omega \) represents a regularization function [196].

- **Random Forest Importance**

RF is one of the promising approaches for predicting the functional role of high-dimensional metagenomes due to its good Accuracy, robustness, and ease of use (as described in Chapter 2). Additionally, it provides an essential variable importance measure to rank the microbial features, according to their predictive power, as introduced in [88],[105]. Breiman [88] proposed that the importance of each feature could be calculated by Gini Importance (GI) or Mean Decrease in Impurity (MDI), estimating the impurity of a decision node weighted by the probability of samples reaching to a node in a decision tree of RF. The impurity function is primarily considered as a Gini Impurity indicated in Eq.3.11[88].

\[
\text{Gini Impurity (GIP)} = \sum_{i=1}^{L} -f_i(1-f_i)
\]  

(3.11)

where \( f_i \) denotes the frequency of features labelled with the class label \( i \) in an RF decision tree, and \( L \) is the total number of class labels.

The value of weighted impurity (GIP) is averaged over all the constituting decision trees of the RF model to calculate importance (GI) (Eq.3.12) [88].

\[
GI = \frac{1}{N} \sum_{t \in \text{Nodes of a tree in } N} p(t) * \text{GIP}
\]  

(3.12)

where \( N \) is the number of decision trees in the forest and node probability \( p(t) \) is calculated by the number of samples that include the feature, proportionally to the number of samples it splits. The higher the value of GI, the more
influential the feature is considered. The procedure is termed as random forest importance (RFI). It is considered as one of useful feature selection measures in this thesis over a variety of metagenomic environments.

EM methods, however, could be used as both the feature selector or a classifier.

### 3.1.3 Data analysis

Relevant features are selected by any of the strategies discussed in the previous step to best discriminate between the phenotypic classes. These features are used to train a classifier model. The trained classifier is used on the test data to classify microbial samples into phenotype. This kind of classification utilised with supervised learning falls into one of the following primary tasks [198].

- **Binary Classification Task**: The pre-processed data is to be classified into one of the two non-overlapping phenotypic classes (For example, classifying the human microbiome communities into the smoking and non-smoking individuals) [94].

- **Multi-class Classification Task**: The pre-processed microbiome input is to be classified into one of the $n$ non-overlapping phenotypic classes where $n \geq 2$. Multiclass problems include the identification of the microbiome into several categories, such as organizing the human microbiome into several host sites (e.g., skin, gut, oral cavity, vagina, etc.).

The construction of predictive models which detail binary or multi-class functions associated with microbial communities may improve our understanding of biological processes. Supervised learners are now essential for
the task of predicting high-dimensional metagenomic data. These methods are trained to classify metagenomic functions as a “phenotypic class” while microbial taxa are encoded as feature vectors. Previous research includes the employment of supervised learners such as SVM, RF, LR and MLP to infer metagenomic functions (discussed in Chapter 2).

In this PhD research project, a range of computational methods is developed for classifying or associating microbial features (either abundance-driven or phylogeny-driven) into their respective functional phenotypes. The proposed studies throughout this thesis primarily focus on upon recent development of embedded methods such as XGBoost and penalized LR (pen-LR) (discussed in section 3.1.2 as feature selectors) to handle high-dimensional and sparse metagenomic data. Also, the most popular RF and SVM models (a review on these classifiers is provided in Chapter 2), for classification tasks are mediated through the integration of hierarchical ancestral derived from phylogeny into abundance profiles for the functional microbiome analysis in the context of current thesis. An effort also has been made in this thesis to regularize the classical method of RF with a phylogenetic measure to develop a new phylogeny-aware RF classifier.

The analysis demonstrated several applications of these approaches to determine biological phenotypes of the human microbiome in smoker and non-smoker individuals; human microbiome across different body sites; soil microbiome treated with sugar substrates; and cattle microbiome associated with supplemented diet in this thesis. This increases the ability of the research community to understand the association of structural and functional profiles of microbiome communities.
3.1.4 Computational strategies to assess supervised ML models

The strategies used in this thesis to assess the employed predictive models inferring metagenomic functions from microbiome are summarised in this section. The performance of the functional predictions in the current thesis is primarily driven by statistical assessment strategies, based on a confusion matrix of size $c \times c$ with regards to the “$c$” functional classes [169]. A sample confusion matrix in the case of binomial classes is shown in Table 3.3.

**Table 3.3 An illustration of a confusion matrix depicting two classes positive and negative.**

<table>
<thead>
<tr>
<th>Total Population</th>
<th>Actual Positive Sample (AP)</th>
<th>Actual Negative Sample (AN)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prediction Positive (PP)</strong></td>
<td>True Positive (TP) Correct Classification of Positive examples into the positive class</td>
<td>False Positive (FP) (Type I Error) Incorrect classification of negative examples into the positive class</td>
</tr>
<tr>
<td><strong>Prediction Negative (PN)</strong></td>
<td>False Negative (FN) (Type II Error) Incorrect classification of positive examples into the negative class</td>
<td>True Negative (TN) Correct Classification of negative examples into the negative class</td>
</tr>
</tbody>
</table>

For a binary classification (where $c = 2$), assessment strategies of Accuracy, and Area Under Receiver Operator Characteristics Curve (AUC-ROC) are constructed using the TP and TN for predictions (Table 3.3) [199], [200]. Accuracy is calculated as shown in Eq.3.13a.

$$\text{Accuracy} = \frac{TP + TN}{P + N} \quad (3.13a)$$

where, $P = TP + FN$, and $N = TN + FP$.

AUC-ROC analysis is primarily based on the Accuracy by which an ML
model could distinguish positive and negative samples. It is a measurement of the area under the ROC plot in a single line graph that measures Sensitivity (Eq.3.13b) against 1-Specificity (Eq.3.13c) over a range of different thresholds.

\[
\text{Sensitivity} = \frac{TP}{P} \quad (3.13b)
\]

\[
\text{Specificity} = \frac{TN}{N} \quad (3.13c)
\]

where, \( P = TP + FN \), and \( N = TN + FP \)

A perfect classifier will have an AUC and Accuracy value of 1.0.

The multi-class classification is handled with a one-vs-all loss in which for each class, a binary classifier is trained such that the class under consideration is taken as a positive class and rest of the classes as a negative class [201].

With \( c \) classes, \( c \times c \) confusion matrix \( C = m_{i,j} \) matrix was constructed, where \( m_{i,j} \) represents several samples predicted as class \( i \) but belonging to class \( j \). Considering, \( F_i = \sum_{1 \leq j \leq c} m_{i,j} \) be the number of input features predicted to be in functional class \( i \); the assessment metric of Accuracy is generalised in Eq.3.14 [200].

\[
\text{Accuracy} = \frac{\sum_{1 \leq i \leq c} m_{i,i}}{\sum_{1 \leq i \leq c} F_i} \quad (3.14)
\]

To extend the AUC-ROC to multi-class classification, a one-versus-all strategy calculated the operating points for individual class and then averaged it out for the entire classifier. The studies such as in [202], [203] introduced the calculation of AUC-ROC for multi-class problems by calculating an AUC-ROC for every pair of classes in the problem domain under consideration, and a similar idea is being applied in this thesis.

Another widely used measure of Kappa Coefficient (Eq. 3.15) [177]
serves as a logical extension to Accuracy supporting the interpretation of classification performance in a better way over the class imbalances in metagenomics datasets. The Kappa Coefficient is a statistical test to evaluate the Accuracy of a classification process by measuring the reliability between two different raters of the class who each classify input samples into mutually exclusive class categories. The value of Kappa ranges from -1 to 1.

For imbalanced datasets, positive and negative cases (e.g., TP and TN in Table 3.3) are unequally distributed, and one of them is present heavily in the dataset. In such cases, Accuracy may not be the right assessment as it could predict the training model to be good, which may not be a case in imbalanced datasets. For such cases, the use of Kappa is preferred as it adds some probability parameters to the Accuracy shown in Eq.3.15.

\[
\text{Kappa} = \frac{\text{Accuracy} - P_e}{1 - P_e} \tag{3.15}
\]

where \(P_e\) is the hypothetical probability of a chance agreement with \(P_e = P_{\text{positive}} + P_{\text{negative}}\) and; \(P_{\text{positive}} = \frac{(TP+FP)(TP+FN)}{(P+N)(P+N)}\) and \(P_{\text{negative}} = \frac{(FN+TN)(FP+TN)}{(P+N)(P+N)}\), \(P = TP + FN\), and \(N = TN + FP\).

In the case of imbalanced datasets, \(P_e\) converges to higher values. This brings down the overall Kappa value. A degenerative case of such an example based on Table 3.3, is shown in Figure 3.4, indicating the value of Kappa in the range of -1 to 1. It indicates Kappa’s ability to differentiate between imbalanced classes better. In this thesis, Kappa is utilised in addition to Accuracy to calculate predictive performance over varied metagenomic datasets in terms of the imbalanced number of classes. Furthermore, Kappa is also extensible to be used for multiple-class classification use cases, as indicated in [204].
The use of a validation scheme to evaluate the prediction performance forms an essential component of ML. The counts of TP, TN, FP and FN (Table 3.3) are calculated using cross-validation methods. The conventional validation methods to govern the performance of ML models used in this thesis are listed below [205].

- **k-fold Cross-Validation** 
  - The original dataset is randomly partitioned into $k$ sets where $k-1$ sets are used for training the classifier, and one set is used as a test set, and the process is repeated $k$ times to get overall results [205],[206].

- **Leave-One-Out-Cross-Validation (LOOCV)**: This serves as a degeneration case of $k$-fold cross-validation, where $k$ is equal to the number of samples

---

**Figure 3.4 An example to discuss Kappa and Accuracy over class imbalances in classifying metagenomic input dataset.**

<table>
<thead>
<tr>
<th></th>
<th>AP</th>
<th>AN</th>
</tr>
</thead>
<tbody>
<tr>
<td>PP</td>
<td>90</td>
<td>0</td>
</tr>
<tr>
<td>PN</td>
<td>0</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Acuracy</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pe</td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td>Kappa</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>AP</th>
<th>AN</th>
</tr>
</thead>
<tbody>
<tr>
<td>PP</td>
<td>90</td>
<td>10</td>
</tr>
<tr>
<td>PN</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Acuracy</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pe</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>Kappa</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>AP</th>
<th>AN</th>
</tr>
</thead>
<tbody>
<tr>
<td>PP</td>
<td>0</td>
<td>50</td>
</tr>
<tr>
<td>PN</td>
<td>50</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Acuracy</th>
<th>0.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pe</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>Kappa</td>
<td>-1</td>
<td></td>
</tr>
</tbody>
</table>
LOOCV is potentially useful in the case of a limited number of samples for training.

Based on the performance Accuracy obtained over different folds or different datasets or different values of parameters of ML classifier, the significance of the differences between the performance of different groups has been established using statistical t-test [208]. It tests the hypothesis that the two groups have equal means. t-test statistically compares the means from two or more independent groups assuming a normal distribution. The calculation of the t-test over two groups is as shown in Eq.3.16.

\[
t = \frac{M_1 - M_2}{\sqrt{\frac{Sd_1^2}{N_1} + \frac{Sd_2^2}{N_2}}}
\]

where \(M_1\) and \(M_2\) represent the means of two groups 1 and 2; \(Sd_1^2\) and \(Sd_2^2\) represent the standard deviation of two groups, and; \(N_1\) and \(N_2\) represent the numbers of observations in each group. This test commonly reports a tail probability of t distribution (according to the degree of freedom) known as p-value. The p-value less than or equal to 0.05 rejects the null hypothesis, which indicates that a true difference in means between two groups is equal to 0 at a 95 % confidence interval [209].

Based on the predictive values, the significance of the differences between the multiple ML methods (in a benchmark analysis) has been performed using Analysis of Variance (ANOVA) [210]. ANOVA is commonly applied for comparing the difference between means from two or more independent groups [210]. ANOVA depicts whether different groups (e.g. classification techniques) significantly contribute to the variations in prediction
outcomes (i.e., means are significantly different with $p$-value to be less than or equal to 0.05). A sample ANOVA table is shown in Table 3.4. However, it does not indicate amongst which methods significant differences lie. For this purpose, post-hocs tests are conducted to determine specific information on which means are significantly different [211].

Table 3.4 An illustration of an ANOVA table where SS represents the sum of squares; df is the degree of freedom and MS is mean squares, and $F$ is the $f$-statistics.

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>$p$-value</th>
<th>$F$ critical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>$SS_b$</td>
<td>$df_b$</td>
<td>$MS_b = SS_b /df_b$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within Groups</td>
<td>$SS_w$</td>
<td>$df_w$</td>
<td>$MS_w = SS_w /df_w$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The three post-hocs tests utilised in current research are listed in Eq.3.17a-c.

i. Least-significant-difference (LSD) = $t \sqrt{MS_w \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}$ (3.17a)

where $t$ is the critical, tabled value of the t-distribution over the $df_w$ associated with $MS_w$; $MS_w$ is the Mean Square within and $n_i$ is the number in the $i^{th}$ group of interest.

ii. Tukey-test-honest-significant-difference (HSD) = $q \sqrt{\frac{MS_w}{n_k}}$ (3.17b)

where $q$ is the value in student range table corresponding to the number of groups of observations and $df_w$ : $MS_w$ is the Mean Square within groups and $n_k$ is the number in each group of interest.

iii. Scheffe’s Test = $\sqrt{df_w \cdot F \text{ critical} \cdot MS_w \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}$ (3.17c)
where $df_b$ is between samples degrees of freedom, $F$ critical value (from ANOVA as illustrated in Table 3.4), and $MS_w$ is mean square error within groups and $n_i$ is the number in the $i^{th}$ group of interest.

Both ANOVA and the post-hocs test make assumptions about data distribution and equality of variance. A description is available at this link\(^2\) to follow. Interested readers are referred to this source.

The above-discussed stages distribute and maximise the applicability of computational models and, therefore, are responsible for a walk through a generalised workflow to determine metagenomic functions (Figure 3.1). To conclude, this section particularly has suggested a general experimental framework to integrate diverse metagenomic data and to obtain metagenomic predictions, setting up the stage for research approaches explored throughout the current thesis.

### 3.2 Possible research approaches formalised based on the conceptualised framework

This PhD research project follows the conceptualised framework introduced in section 3.1 and proposes three main research approaches along with this

Possible research approaches formalised based on the conceptualised framework, leading to three different key studies in the current thesis. These three approaches are proposed and listed below.

- **Approach 1. Abundance-Aware Analysis.** It deals with the use of ML analysis techniques over the microbial abundance profiles to ascertain better metagenomic predictions (as shown in Figure 3.5a). This approach led exploratory direction to the study conducted in Chapter 4 within this thesis in determining phenotypes related to the high-dimensional and sparse human microbiome.

![TAXA Table (with their Abundance Count)](image)

**Figure 3.5a** A graphical representation of Approach 1 indicating possible workflow for abundance-driven analysis.

- **Approach 2. Phylogeny-Informed Data Encoding and Processing.** This approach is driven by combining heterogeneous data of the OTU abundance count table and the phylogenetic tree into a new data structure. ML-based methods are applied over the newly constructed data structure, and their
assessment is performed with an insightful benchmark analysis (as shown in Figure 3.5b). The related study is proposed and explained in Chapter 5 of the current thesis.

Figure 3.5b A graphical representation of Approach 2 indicating possible workflow for phylogeny-driven analysis.

- **Approach 3. Integrating Phylogeny in ML Supervised Learner itself.**
  
  Along with this approach, an idea of hierarchical grouping obtained from phylogeny is proposed to be further used by a supervised learner for encoding relevant phylogenetic relationships between dependent (or phylogenetically close) OTU variables. The supervised learner tuned with phylogenetic information is further used for classifying metagenomic samples (as shown in Figure 3.5c). The study related to this is presented in Chapter 6 of this thesis.
Figure 3.5c A graphical representation of Approach 3 indicating possible workflow for phylogeny-driven ML modelling.

These proposed approaches (Approach 1-3) provide an opportunity for the organization and the prediction of environmental functions in metagenomic studies. The next section describes the diverse metagenomic data sources utilised by these research approaches in the scope of this thesis for the inference of microbiome functions. The data have majorly been extracted from the real-world datasets curated with QIIME [20] versions in their source publications. One of the data sources is sampled as part of MetaPlat project[144].
3.3 Data sources

The advances in experimental high-throughput technologies and the rapid development of metagenomics tools have increased the generation and study of thousands of microbial taxa across hundreds of microbial samples. The microbiome compositions could help better in studying their functional properties exhibiting great potential toward revolutionizing biological studies. In the current thesis, we target to employ the ML-based framework/s for detecting the functions of microbiome differences by analysis of diverse datasets related to linking microbiome composition to functional phenotype.

The data sources utilised for inferencing the role of metagenomes from different environments in the scope of the current thesis are listed below.

• Data Source 1 (DS1)

A real-world human microbiome dataset consisting of 60 microbiome samples, and 856 OTUs obtained from a study by Charlson et al. [172] investigated the effect of cigarette smoking on the oropharyngeal and nasopharyngeal microbial communities present in the human body. The samples in the study were collected from the right and left nasopharynx and oropharynx of smoking and non-smoking individuals. The multiplexed pyrosequencing amplified the V1-V2 region of the microbial 16S rRNA genes sequences, which were subsequently processed using the QIIME pipeline with default parameter settings. An OTU count table of 60 samples (consisting of 28 smokers vs. 32 non-smokers) and 856 OTUs were generated. The phylogenetic tree was produced by QIIME computing phylogenetic distances between different
OTUs. The associated dataset is available as part of the R package of MiSPU (https://cran.r-project.org/web/packages/MiSPU/) and described in the study by [149]. This has been used in the current thesis to study the microbiome differences between smokers and non-smokers individuals by using both the abundance profiles and the phylogenetic information of OTUs.

- **Data Source 2 (DS2)**

HMP provided a reference collection of 16S rRNA gene sequences collected from various sites (habitats) across the human body to better study the changes in the microbiome composition and associating them with human health. HMP Consortium has reported the human microbiome in 300 healthy adults at 18 body sites in total. The two-time point 16S window spanning over V3-V5 regions was conducted for the human microbiome samples. The dataset consisting of an OTU count table with 3285 samples and 5830 OTUs, and the related phylogenetic tree was obtained by pyrosequencing of these 16S rRNA sequences obtained from multiple human body sites and using QIIME pipeline. The dataset derived from HMP has been used to study dynamics and associations of the microbial communities across four major human body habitat groups of Oral (1818), Skin (966), Vaginal (291), and Stool (210) samples in current research. The source is available from a shared GitHub repository linked to https://github.com/twbattaglia/MicrobeDS/blob/master/data/HMPv13.rda.
• **Data Source 3 (DS3)**

The work described in [93] ties together high levels of variability among microbial individuals sampled from six body habitats of External Auditory Canal (EAC) (44 samples), Gut (45 samples), Hair (14 samples), Nostril (46 samples), Oral cavity (54 samples), and Skin (419 samples), generating *Costello Body Habitats* (CBH) datasets. It is used as one of the example datasets for microbial analysis in the scope of the current research. Microbial communities are highly differentiable between these different body habitats. This example dataset obtained from [93], is pre-processed using *ape* [181] and *phyloseq* [179] packages in R to attain an OTU table with 622 samples and 2683 OTUs and the corresponding phylogenetic tree with intersecting indices.

• **Data Source 4 (DS4)**

Combining the multiplexed High Throughput DNA sequencing (HTS) with Stable Isotope Probing (SIP) of DNA or RNA nucleic acids (i.e., HTSSIP), is a powerful culture-independent method for linking microbial functioning to their taxonomic profiles. SIP experiments in [212], employed incubating aliquots of soil with substrates of either 13C-glucose or 13C-cellulose (treatments) or the 12C-glucose (the corresponding control), to reveal patterns of the inhabitant microbes and their role in converting labelled gradients into biomass. The main goal is to compare labelled-treatment gradients to their corresponding unlabelled (controls) gradients for determining which 16 S rRNA genes contribute to the biomass breakdown in the soil and serving the microbial food-web better. The dataset is available as part of the *HTSSIP package* in R (https://cran.rstudio.com/web/packages/HTSSIP/vignettes/HTSSIPintro.html),
developed for analysing high throughput sequence (HTS) data obtained from DNA- or RNA-based SIP experiments. The dataset used in this thesis has been extracted from this study [213] with the help of the phyloseq package in R. The dataset consists of 139 samples (almost equally distributed in 3 class categories) and 1072 OTUs with their corresponding phylogenetic profile.

- **Data Source 5 (DS5)**

  An interesting study in [64] focused on the interrelationships between human diet and the gut microbiome. The curated version of this human gut microbiome dataset used in the current research was sourced from [http://www.exploredata.net/Downloads/Microbiome-Dataset](http://www.exploredata.net/Downloads/Microbiome-Dataset) in July 2017. The dataset curated to 658 human microbial samples with 6696 OTUs and the diet functions of - i) LF/PP (standard low-fat, plant polysaccharide-rich diet) and ii) Western (high fat, high sugar diet). The study created a well-defined, representative prototype of the human gut microbial ecology by transplanting human fecal microbial communities into germ-free C57BL/6J mice. The study indicated that switching from a low-fat, plant polysaccharide-rich diet to a high-fat, high-sugar “Western” diet affects the microbial composition and subsequently affects human health.

- **Data Source 6 (DS6)**

  The study in [173] has shown the linkage of IBD with the human intestinal microbiome. The research in [173], focused on long-term dynamics of the human gut microbiome from an IBD cohort of 128 individuals sampled at intervals of 3 months, generating a total of 673 samples. For the classification
of IBD disease into ulcerative colitis (UC), healthy controls (HC), Crohn's disease (CD), collagenous colitis (CC) states, the corresponding 16S rRNA abundance counts are availed from the source QIITA study of https://qiita.ucsd.edu/study/description/1629. The data extracted from 16S rRNA gene sequences annotated from the V4 hypervariable region, containing 673 microbial samples and 10996 OTUs, is employed in the construction of ML models for the predictive task of inferring the metagenomic function of relating microbiome abundance profiles to IBD diseased states in the current research.

- **Data Source 7 (DS7)**

This thesis also involves a case study over the real-world cattle rumen microbiome datasets sampled as part of the MetaPlat³ Project under European Horizon H2020-MSCA-RISE-2015 for the first time [144]. The project aims to study microbial ruminal environment in cattle (Bos taurus) and assess the effect of the dominance of feeding behaviours on the dietary patterns, feed efficiency, or methane emissions. The Bos taurus microbiota plays a vital role in cattle productivity, health, and immunity. 40 cattle rumen samples were collected at the beginning of the project to investigate the effects of (i) addition of the nitrate and (ii) the dietary oil in beef cattle receiving these two different diets. The corresponding dataset consisted of 20 samples from an oil-based treatment and 20 samples from a nitrate-based treatment. 5 OTU tables, with different taxonomic levels (phylum to genus) of classification, were generated with the help of QIIME. The tables consist of 27, 52, 101, 194, and 386 OTU feature vectors for phylum, class, order, family, and genus levels, respectively.

³ https://cordis.europa.eu/project/id/690998
To extend upon the sampling of data, rumen samples from 80 steers at the end of a feed efficiency measurement period with four diet treatments (control, nitrate, a combination of nitrate-oil and oil) and; two different cattle breeds; were sampled as part of MetaPlat project. This dataset consisted of two breeds of *Aberdeen Angus*, or *Limousine sired* steers; and four dietary treatments of control (443 g concentrate and 25 g lipid/kg diet DM); nitrate (18 g nitrate/kg DM); oil (maize distiller’s dark grains, 37 g lipid/kg diet DM) and combined (18 g nitrate and 37 g lipid/kg dietary DM). The dataset consisted of 20 samples from each of the dietary treatments and the controls. ASV table (analogues to an OTU table), containing an abundance count of 727 microbial taxa, was obtained using the dada2 workflow along the QIIME2 [25] pipeline under the MetaPlat project [144].

Furthermore, in the third phase of sampling of data, glucocorticoid was circulated on the ruminal environment to study the effect of stress provoked microbiota that might change fermentation patterns in cattle affecting various digestive functions. 679 ASVs were obtained at species level using the dada2 along the QIIME2 pipeline with 118 samples. 70 samples were treated with glucocorticoid amongst the 118 examples. However, the phenotypic effects of the dominance of feeding behaviour were applied as concentrated (in 60 samples) and forage diet (in 58 samples). The different datasets discussed above are summarised in Table 3.5.
### Table 3.5 A summary of metagenomic datasets used in this thesis for functional analyses.

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Data Structures</th>
<th>Host Source</th>
<th>OTUs or ASVs Count</th>
<th>Functional Phenotype (Sample Size in each phenotypic Category)</th>
<th>Source URL</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Source 1 (DS1)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OTU table, Phylogenetic Tree</td>
<td>Human Microbiome</td>
<td>856</td>
<td>Smoking (28) vs. Non-smoking individuals (32)</td>
<td><a href="https://cran.r-project.org/web/packages/MiSPU/">...</a>[172]</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Data Source 2 (DS2)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OTU table, Phylogenetic Tree</td>
<td>Human Microbiome</td>
<td>5830</td>
<td>Body habitats of Oral (1818), Skin (966), Vaginal (291) and Stool (210)</td>
<td><a href="https://github.com/twbattaglia/MicrobeDS">...</a>[53]</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Data Source 3 (DS3)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OTU table, Phylogenetic Tree</td>
<td>Human Microbiome</td>
<td>2683</td>
<td>Body habitats of External Auditory Canal (EAC) (44), Gut (45), Hair (14), Nostril (46), Oral cavity (54), and Skin (419)</td>
<td><a href="https://www.knightslab.org/data">...</a>[69, [93]</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Data Source 4 (DS4)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OTU table, Phylogenetic Tree</td>
<td>Aliquots of soil</td>
<td>1072</td>
<td>Substrates treatment of glucose (47) or cellulose (46) or controls (46)</td>
<td><a href="https://cran.rstudio.com/web/packages/HTSSIP/vignettes/HTSSIP_intro.html">...</a>[212]</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Data Source 5 (DS5)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OTU table</td>
<td>Human Microbiome</td>
<td>6696</td>
<td>Functions of Low-fat plant diet (389) and High-fat Western diet (269)</td>
<td><a href="http://www.exploredata.net/Downloads/Microbiome-Dataset">...</a>[64]</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Data Source 6 (DS6)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OTU table</td>
<td>Human Microbiome</td>
<td>1099</td>
<td>IBD diseased states of ulcerative colitis (UC) (324), Healthy controls (HC) (62), Crohn’s disease (CD) (251), collagenous colitis (CC) (36)</td>
<td><a href="https://qiita.ucsd.edu/study/description/1629">...</a>[173]</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Data Source 7 (DS7)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OTU table o ASV Table, Phylogenetic Tree</td>
<td>Cattle Gut Rumen Microbiome</td>
<td>386; 727; 679</td>
<td>The supplemented diet of Cattle with 40 samples (20 from each phenotype of oil-based diet or nitrate -based diet); 80 samples with 20 each in oil, nitrate</td>
<td><a href="https://cordis.europa.eu/project/rcn/199933/factsheet/en">...</a>[144]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3.4 Summary

The unprecedented advancement in functional metagenomics has accelerated the development of ML techniques that can be applied in the computational prediction of microbial functions. This kind of analysis is crucial for understanding useful biological functions in microbial ecology by identifying useful microbial features and defining the relationships between microbial features and biological functions (outcome). This chapter contributes towards providing an overview of a conceptualised workflow employed to detect metagenomic functions and linking phenotype to metagenomes. The typical workflow for such classification of metagenomic functions involves data pre-processing and feature selection before the learning phase of modelling, as shown in Figure 3.1. The thesis aims to extend upon this workflow for modelling metagenomic functions using three different approaches.

Approach 1 (Figure 3.5a) explores the recently developed EM methods over the abundance count of human microbiome data. Additional knowledge of phylogeny is integrable and important in predictive modelling. Henceforth, the
thesis further aims to find the best possible way to incorporate microbial phylogeny as next steps to aid in functional predictions, following Approach 2 (Figure 3.5b) in an experimental framework. In the current context, a new idea of the fusion of OTUs along the tree topology is implemented, integrating their evolutionary distances with their abundance counts. The idea has roots in the biological assumption that closely related species in a tree may behave similarly towards their biological function.

In Chapter 5, a novel integrative framework based on phylogeny and abundance of OTUs is proposed and assessed using techniques from the field of ML. The chapter highlights how the combination of phylogeny and quantitative profiles of abundances effectively account for determining metagenomic functions. An attempt to integrate biological knowledge at ML classifier (following Approach 3 in Figure 3.5c), is made in Chapter 6. It is driven from a natural assumption that closely related OTUs by phylogeny is highly correlated and tend to have similar genetic traits and hence, to be selected together.

To conclude, the chapter lists potential methods and materials utilised throughout this thesis for analysing the microbiome sourced from different environments.
CHAPTER 4

AN ABUNDANCE-DRIVEN FRAMEWORK FOR THE PREDICTION OF METAGENOMIC FUNCTIONS

This chapter comprehensively assesses various ML approaches to characterise high-dimensional microbiome datasets adequately into their functional phenotypes. An important research question getting attention from the research community in microbiome studies is: “Which ML modelling suits best for dealing with sparse and high-dimensional metagenomes? and “Which OTUs or taxa play an important role in characterizing the environmental phenotype?”.

Not much work has been done applying EM methods, especially gradient boosting in the field of functional metagenomics. However, there seems a great potential in using these techniques to derive accurate metagenomic functions from high-dimensional metagenomes. The chapter introduces that more research could be applied to the area of functional metagenomics using the recently developed EM methods of gradient boosting and the penalized regression. The study conducted in this chapter highlights that the microbial features obtained from the EM models contributed well to the functional predictions, enhancing the interpretability of the popular state-of-the-art methods of RF and SVM over the functional metagenomic studies. This has led the path to the development of a new abundance-driven framework.
4.1 Rationale

Studies in [93], [98],[120], [127]–[129] have suggested that supervised ML techniques of RF and SVM could be applied successfully to abundance count data of microbial taxa for the predictive task of functional inference in metagenomic studies. These techniques have been discussed in Chapter 2 on the Literature Review. However, the approach presented in this chapter aims to extend upon these traditional supervised ML approaches to effectively classify the sparse and high-dimensional metagenomes with faster and scalable methods. This could provide an improvement in terms of the predictive performance of functional interference. This chapter employs different computational methods to predict phenotype in two of the abundance-driven high-dimensional, and large-scale data sources (DS5 and DS6) related to the human microbiome and assess their performance. These data sources have much more microbial feature predictors (p) in comparison to the number of microbial samples (n) with $p >> n$. DS5 associates human microbiome with diet (high-fat or low-fat) having 6696 features $>>$ 658 samples, and; DS6 associates human microbiome with IBD states (having 10996 features $>>$ 673 samples). EM methods have been developed for improving the predictability of large datasets with the “large p small n” problem cases [196],[214]. This research has introduced the use of EM methods of eXtreme Gradient Boosting (XGBoost) [196] and Penalized Logistic Regression (Pen-LR using glmnet) [112], [214], [215] as these embedded methods can determine important features as well as to classify high-dimensional metagenomes well. The use of EM methods has been proposed to improve upon the classical methods [93], [98], [128]. The
efforts in this chapter have explored utilizing EM methods in the area of functional metagenomics, which have been applied successfully in several other non-metagenomic areas, such as in analysing medical records [216]; protein prediction [217],[218]; intrusion detection[219],[220]; and image classification [221],[222]. EM methods are designed for speed and performance [196],[215]. A few attempts have been made to apply LASSO and ENet (the variants of Pen-LR) in functional metagenomics [93],[128], but XGBoost has not been explored much yet in functional metagenomics. A rationale behind this study is to identify a scalable, time-efficient ML-based framework, which provides good predictive performance and is fast at classification over the high-dimensional metagenomic abundance profiles. The framework follows Approach 1, as suggested in Chapter 3 (Figure 3.5a).

This work presents experimentation for functional metagenomics with features derived from the EM methods. Specifically, the work is targeted towards finding which OTU features are relevant in such functional classification and investigate which classifier provides more accurate results over the microbial (OTU) features. The work presented in this chapter is based on the published work in [223].

4.2 Methodology

This section presents an experimental framework that employs various ML models for the functional classification of microbial samples comprising of OTU abundances. The stepwise methodology along the framework is described as follows (illustrated in Figure 4.1).
(i) **Input:** An OTU table containing samples associating them with metadata of environmental traits, was used as the primary input in this study. The experiments performed were exploratory in investigating high-dimensional and large-scale data sources of DS5 and DS6 on: a) how diet associated with human microbiome samples can be used to provide dietary management support in terms of analysing the impact of high-fat western or low-fat plant diet; and b) the relation of the human microbiome samples with different states of IBD.

(ii) **Pre-processing:** The OTU tables were pre-processed to fit the ML procedures. It is preferable to select relevant features from a variety of feature attributes in an OTU table to maximise the performance of the experimental design. Feature selection was aimed at removing irrelevant and redundant features. For this purpose, EM techniques were applied with the regularization methods and a boosting method. EM techniques are known to embed feature selection in the ML classifier itself (as discussed in Chapter 3). Additionally, in the scope of the current chapter, filter selection (FFS) with IG and WFS with recursive feature elimination over RF and SVM was employed over the quantitative profiles of the human microbiome use cases for a benchmark analysis.

(iii) **Cross-Validation:** The approach partitioned the input dataset to train and test sets for *k*-fold cross-validation where *k* = 10.
(iv) **Classification**: To identify the most suitable model for predicting the functional phenotype of metagenomes, popular supervised ML classification algorithms were evaluated for their fitness in the prediction task against the microbial samples consisting of filtered OTU abundances. These include the popular state-of-the-art RF and SVM. Additionally, EM methods served as both a feature selector and the classifier in this study.

(v) **Performance Evaluation**: The performance of the classifier was assessed using the widely used assessment measures of Accuracy, Kappa and AUC-ROC [199], as listed in Chapter 3.

![A generalised abundance-driven methodological workflow for functional metagenomics.](Image)

**Figure 4.1** A generalised abundance-driven methodological workflow for functional metagenomics.
4.3 Implementation details

The methods proposed in this study [223] were implemented over the R platform (http://www.R-project.org). The various packages related to ML models used and the related set of different configurational hyper-parameters utilised in this study are listed below.

a) **XGBoost.** Implementation: package *XGBoost*. Configurational parameters: objective" = "binary: logistic" for binary classes and objective" = "multi: SoftMax" for multinomial classes, "nthread” = 8, “max_depth” = 3, "gamma loss reduction " = 0

b) **Pen-LR with LASSO, Ridge, ENet logistic classifier.** Implementation: package *glmnet*. Configurational parameters: family = “binomial”, “multinomial”, alpha regularization penalty (ɑ)=0,0.3,0.5,1

c) **RF.** Implementation: package *randomForest*. Configurational parameters: ntree=100

d) **svm.** An SVM with the configurational kernel. package e1071. Configurational parameters: kernel="radial"/"poly"/"linear", cost=1, gamma=0.5, scale=TRUE

e) **Recursive Feature Elimination.** Implementation: package *caret*. Configurational parameters: rfeControl = rfFuncs for RF classifier and rfeControl = caretFuncs, with number = 10 and method = "svmRadial"

f) **Information Gain Filter.** Implementation: package *FSelector*. Configurational parameters: Top 20 Features

g) **Confusion Matrix.** Implementation: package *caret*. Evaluation metrics over $overall and $byClass associated parameters
h) **AUC-ROC.** Implementation: package pROC. Evaluation over binomial and multinomial classes by the area under curve (AUC-ROC) values.

i) **glmnetRank.** Implementation: package SurvRank, Rank order of Coefficients in glmnet

j) **Random Forest Importance.** Implementation: package FSelector. Configurational parameters: Top 10 Features

10-fold cross-validation was performed for all experiments. Each of the datasets was divided into 10 partitions known as folds. One-fold was used for testing, and the remaining 9 were used for training the data. The process was repeated for every fold. The time recorded for ML models is the User (CPU) time charged for the execution of user instructions of the calling process (in seconds). The running environment consisted of a system configured with AMD processor A8-7410 @2.20 GHz, Quad-Core, 8 GB RAM.

### 4.4 Experiments and results

The proposed abundance-driven framework was employed over DS5 and DS6 circumventing an objective of determining “Which ML models attain better prediction performance of classification in metagenomic studies?” The study applied different classifiers and evaluated their performance using the methods described in section 4.2.
4.4.1 Abundance-driven analysis of DS5 and DS6 metagenomic data sources

RF and SVM classification methods were reported as most effective approaches by the classical studies in the field of metagenomics [93],[98],[99],[103] for the prediction of metagenomic functions over human microbiome. It was recommended that the RF classifier performed robustly in inferring metagenomic functions. These two classical ML methods of RF and SVM were applied over DS5 and DS6 for a baseline comparison in the current research. The datasets employed under current study are summarised in Table 4.1.

**Table 4.1 A summary of data sources (with predictors >> the number of samples) used in the current study.**

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Samples</th>
<th>OTUs</th>
<th>Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS5</td>
<td>658</td>
<td>6696 (relative abundances)</td>
<td>Functions of Low-fat plant diet (389) and High-fat Western diet (269)</td>
</tr>
<tr>
<td>DS6</td>
<td>673</td>
<td>10996 (absolute abundances)</td>
<td>IBD diseased states of ulcerative colitis (UC) (324), Healthy controls (HC) (62), Crohn’s disease (CD) (251), collagenous colitis (CC) (36)</td>
</tr>
</tbody>
</table>

DS5 is a use case with a binary classification requirement with balanced classes, whereas DS6 deals with the multiclass classification (imbalanced classes).

The methods of RF and SVM would allow us to predict the related phenotypes and to observe their performance values as a benchmark for future analysis of results. The results were obtained from these classifiers with per-class metrics averaged over all the classes and overall the 10-fold under implementation.
RF progresses with a bagging strategy that averages the results over several independent decision trees modelled over the metagenomic sub-samples. On the contrary, a boosting strategy involves leveraging patterns in predictors sequentially instead of independently [197]. XGBoost [196] combines advantages from both strategies of boosting and bagging, attaining competitive Accuracy, time-efficiency, and feasibility to be applied over metagenomic data in use. Using the Pen-LR with glmnet [112],[214],[215] aided in reducing information redundancy by dropping out some unimportant features. This, in turn, is useful in avoiding the overfitting of the predictive model and reducing the complexity of training. Merited by the performance superiority and affordable time of recently developed EM methods of XGBoost [196] and Pen-LR (with glmnet) [112], these were explored and employed to the abundance profiles of the human microbiome (related to DS5, DS6) with parametric settings described in section 4.3. In the process of predicting metagenomic functions, the processing of high-dimensional data by a traditional classifier (such as RF and SVM) could pose model-complexity issues. Therefore, dimensionality reduction is an indispensable part of the data pre-processing process applied before the supervised classifier over DS5, DS6. Further research into approaches that leverage feature selections were conducted to determine an efficient abundance-driven framework for the classifying human microbiome use cases in this study. EM methods of XGBoost and Pen-LR were emphasized for performing feature selections due to their time-efficiency and scalability. The study investigated how selecting a reduced subset of relevant discriminative OTU features impacts the predictive performance in functional metagenomics. The detection of OTUs with high
predictive power with the help of feature selection is potentially an interesting area in the field of functional metagenomics. Based on the 6697, 10997-dimensional feature vectors obtained from DS5 and DS6 respectively, selecting an appropriate method to improve the performance rate is of great significance to the establishment of the functional classification of the metagenomes in these data sources. This study compares the impact of feature selection strategies of XGBoost, Pen-LR, IG, RF, and RFE on the prediction results obtained over DS5 and DS6, and targets to choose an optimal dimensionality reduction strategy. The dimensionality reduction was performed using the EM method of XGBoost (over DS5, DS6) with the optimal set of parameters, as listed in section 4.3 (over both DS5, DS6). Furthermore, using Pen-LR (\textit{glmnet} fit with $\alpha = 0.3$ as it attained the best performance) over DS5, the feature set giving the best performance (Accuracy) amongst Top-5,100,120,150,200,250,300,350, and 400 features, was selected.

The multiclass use case of DS6 involved explicitly extracting Top 100 features obtained from the \textit{glmnet} fit for each class (with glmnet $\alpha = 0.5$ as it attained the highest performance) and designing feature set with features unique to every class. The process over DS6 was designed as one of the heuristics in current research due to the limited capability of \textit{glmnetRank()} to rank-order coefficients of \textit{glmnet} fit for only binary class use cases in the R package [224]. When using Pen-LR for dimensionality reduction, features corresponding to non-zero feature coefficients have been selected. As a further heuristic, the features from glmnet were uniquely combined with XGBoost to gain the best of the features obtained from both EM models, in current research. IG selected top 20 features (as a default setting), and; the RF selected top 10 features in the
scope of the current study. The subsequent classification, in this study with RF and SVM models, was carried out with 10-fold cross-validation to predict and evaluate model performance avoiding overfitting. The results obtained from the above-stated various modelling combinations of feature selector and classifier, in current research are discussed below.

4.4.1.1 Discussion of results obtained over the abundance-driven analysis of DS5

This sub-section discusses the results obtained by an application of abundance-driven ML models over the DS5 (as reported in Table 4.2). From Table 4.2, it can be inferred that the prediction performance using the RF model is better than the SVM, however, RF seems more time-intensive.

Supervised modelling over DS5 (Table 4.2) containing relative abundance counts of OTUs (with the small number of samples and many OTU features), posed likelihood of an SVM classifier’s overfitting in classification leading to devious results. Even after tuning cost, gamma and SVM kernels, no major change was observed in the predictive performance of SVM over DS5.

The analysis in Table 4.2, showed that the EM method of Pen-LR using glmnet (α = 0.3) provided relatively better classification performance than RF and much better performance than state-of-the-art of SVM and, in comparatively very less execution time. RF produced relatively better classification performance than XGBoost over DS5. Nonetheless, it proved to be time-consuming in comparison to XGBoost. Pen-LR and XGBoost proved to be faster than RF and SVM over high-dimensional datasets.
Table 4.2 Performance of various ML models (with/without feature selection) employed over DS5 (associated with human diet phenotype) with 10-fold cross-validation. NFS represents the number of features. Best performance attained is bold-faced.

<table>
<thead>
<tr>
<th>Model</th>
<th>Feature Selection</th>
<th>NFS</th>
<th>Time (secs)</th>
<th>Accuracy</th>
<th>Kappa</th>
<th>AUC-ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>No</td>
<td>6696</td>
<td>2085</td>
<td>0.940</td>
<td>0.876</td>
<td>0.947</td>
</tr>
<tr>
<td>SVM (radial kernel)</td>
<td>No</td>
<td>6696</td>
<td>793</td>
<td>0.551</td>
<td>0.001</td>
<td>0.500</td>
</tr>
<tr>
<td>SVM (poly kernel)</td>
<td>No</td>
<td>6696</td>
<td>640</td>
<td>0.539</td>
<td>0.049</td>
<td>0.544</td>
</tr>
<tr>
<td>SVM (linear kernel)</td>
<td>No</td>
<td>6696</td>
<td>580</td>
<td>0.552</td>
<td>0.091</td>
<td>0.569</td>
</tr>
<tr>
<td>XGBoost</td>
<td>No</td>
<td>6696</td>
<td>62</td>
<td>0.931</td>
<td>0.852</td>
<td>0.926</td>
</tr>
<tr>
<td>Pen-LR using glmnet-(\alpha = 1) (LASSO)</td>
<td>No</td>
<td>6696</td>
<td>29</td>
<td>0.924</td>
<td>0.836</td>
<td>0.908</td>
</tr>
<tr>
<td>Pen-LR using glmnet-(\alpha = 0.5) (ENet)</td>
<td>No</td>
<td>6696</td>
<td>28</td>
<td>0.950</td>
<td>0.893</td>
<td>0.942</td>
</tr>
<tr>
<td>Pen-LR using glmnet-(\alpha = 0) (Ridge)</td>
<td>No</td>
<td>6696</td>
<td>244</td>
<td>0.567</td>
<td>0.111</td>
<td>0.571</td>
</tr>
<tr>
<td>Pen-LR using glmnet-(\alpha = 0.3)</td>
<td>No</td>
<td>6696</td>
<td>32</td>
<td>0.951</td>
<td>0.897</td>
<td>0.944</td>
</tr>
<tr>
<td>RF</td>
<td>XGBoost</td>
<td>83</td>
<td>81</td>
<td>0.955</td>
<td>0.907</td>
<td>0.950</td>
</tr>
<tr>
<td>RF</td>
<td>XGBoost and glmnet</td>
<td>270</td>
<td>160</td>
<td>0.960</td>
<td>0.916</td>
<td>0.952</td>
</tr>
<tr>
<td>RF</td>
<td>glmnet</td>
<td>120</td>
<td>78</td>
<td>0.962</td>
<td>0.916</td>
<td>0.957</td>
</tr>
<tr>
<td>RF</td>
<td>IG</td>
<td>20</td>
<td>290</td>
<td>0.940</td>
<td>0.878</td>
<td>0.939</td>
</tr>
<tr>
<td>SVM (linear kernel)</td>
<td>XGBoost</td>
<td>83</td>
<td>32</td>
<td>0.923</td>
<td>0.840</td>
<td>0.920</td>
</tr>
<tr>
<td>SVM (linear kernel)</td>
<td>XGBoost and glmnet</td>
<td>270</td>
<td>44</td>
<td>0.963</td>
<td>0.923</td>
<td>0.961</td>
</tr>
<tr>
<td>SVM (linear kernel)</td>
<td>glmnet</td>
<td>48</td>
<td>120</td>
<td>0.928</td>
<td>0.849</td>
<td>0.928</td>
</tr>
<tr>
<td>SVM (linear kernel)</td>
<td>IG</td>
<td>20</td>
<td>288</td>
<td>0.911</td>
<td>0.800</td>
<td>0.907</td>
</tr>
<tr>
<td>RF</td>
<td>RFE</td>
<td>----</td>
<td>~ 40K</td>
<td>0.940</td>
<td>0.862</td>
<td>0.934</td>
</tr>
<tr>
<td>SVM (radial)</td>
<td>RFE</td>
<td>----</td>
<td>~20K</td>
<td>0.939</td>
<td>0.858</td>
<td>0.932</td>
</tr>
</tbody>
</table>
However, Pen-LR using glmnet $\alpha = 0$ (Ridge) attained a low classification performance representing a probable case of overfitting which was observed to attain some good results on some training data but did not generalise to the good predictive ability on test data.

Table 4.2 suggests that the application of RF and SVM over XGBoost and Pen-LR (using glmnet) selected features further enhanced the performance of supervised classification with RF and SVM respectively over DS5. With this ensemble setting (RF or SVM as a classifier, and; XGBoost combined with Pen-LR as feature selectors), RF’s performance (Time: 2085 secs, Accuracy: 0.937, AUC-ROC:0.947, Number of Features (NFS): 6697), improved to (Time: 160 secs, Accuracy: 0.960, AUC-ROC:0.952, No of Features (NFS): 270), and; SVM’s (with linear kernel) performance (Time: 580 secs, Accuracy: 0.552, AUC-ROC:0.569, Number of Features (NFS): 6697), improved greatly to (Time: 44 secs, Accuracy: 0.963, AUC-ROC:0.961, No of Features (NFS): 270), over DS5. This embraced the remarkable performance improvement in the case of classification with SVM. glmnet was tuned with $\alpha = 0.3$ in this ensemble setting, as this alpha value provided the highest Accuracy than glmnet regularized with other $\alpha$ values of 0, 0.5, and 1 (Table 4.2); and top 120 features were selected by glmnet as they verified high AUC-ROC on this value (Figure 4.2). In the XGBoost model, importance was calculated by how many times an OTU feature is selected as a splitting point across all trees in the XGBoost ensemble.
It seems that SVM encountered overfitting in determining functional phenotypes of high-dimensional relative abundance omics data of DS5 (Table 4.2). It is essential to avoid deceptive functional results and enhance phenotype decision making in such a case. In this work, the problem has comprehensively investigated that functional prediction using SVM classifier inevitably embraced overfitting under polynomial, linear and radial kernels and various tuning of cost and gamma parameters; possibly because of DS5’s special characteristic i.e., the small number of samples and many microbial features with their relative abundances. SVM achieved good performance results on training data but was not able to generalize the good predictive ability to test data. It could have been trapped in one or a few patterns in high-dimensional input data and produced irrelevant functional predictions. The application of EM feature selection precursor to SVM provided a small subset of features curtailing the problem of the small number of samples and many microbial features and avoided overfitting in metagenomic functional predictions over DS5 (Table 4.2).
The process of feature selection before an SVM classifier, not only robustly conquered the overfitting problem but also achieved good classification performance (bold-faced in Table 4.2). Furthermore, the application of RF or SVM over the feature set created by EM methods, provided significantly better results than the application of RF or SVM directly over the abundance count profiles of DS5. The results of t-tests over test accuracies (in each of the 10-fold) obtained by classical models with or without the use of embedded methods of feature selection reported a p-value < 0.01 signifying that EM methods have the potential to significantly improve the performance of state-of-the-art classifiers of RF and SVM. Further to this, an ANOVA analysis with post-hocs procedures were conducted over three groups: RF (employed directly over the abundance profiles), RFXGPLR (RF applied over feature-set obtained from XGBoost and Pen-LR using glmnet from abundance profiles), and RFPLR (RF applied over features obtained by Pen-LR using glmnet); for measuring the difference of significance between test Accuracies obtained by 10-fold of validation. ANOVA provided p-value of 0.089 > 0.05; showing not much significant difference between the three groups. However, the pair RF and RFPLR reported Least Significant difference (LSD) value of 0.022 marginally greater than the threshold of 0.021. By performing ANOVA and post hoc analysis, highly significant differences were observed between the classification models of SVM and SVM applied over feature set obtained by EM methods of Pen-LR and/or XGBoost. Figure 4.3 also shows a great variation in performance (Accuracy across different folds) of different models.
Experiments and results

Figure 4.3 Statistical analysis conducted for measuring the difference between Accuracies obtained over 10-fold of validation using SVM, SVMXGPLR (SVM applied over feature-set obtained from XGBoost and Pen-LR using glmnet), and SVMPLR (SVM applied over Pen-LR using glmnet).
From Table 4.2, it can also be seen that EM methods as feature selectors provided relatively better performance in comparison to FFS (with IG) and WFS (with RFE) over DS5. RFE as a wrapper proved to be time-intensive (Table 4.2) for high-dimensional metagenomes. The features selected by EM methods fit the RF and SVM model better and has more obvious advantages in comparison to other methods. (Table 4.2).

Based on EM-based feature selection, it is observed the EM models selected the following taxa as indicative of playing an important role associating microbial samples into diet phenotype (Figure 4.4).

*phylum: class: Bacteroidetes: Bacteroidetes; Firmicutes: Bacilli; Firmicutes: Clostridia; Firmicutes: Erysipelotrichi;*

These identified microbial taxa coincide with previous research [64] in analysing the effect of diet on mice injected with human fecal microbial communities to support dietary management. Overall, the results attaining high predictive performance (Table 4.2) indicate that the nutritional value of food has the potential to influence the human gut microbiome.
Experiments and results

Figure 4.4 Some important taxa identified with EM-based feature selection over DS5 playing an important role in associating the human diet with the microbiome.

4.4.1.2 Discussion of results obtained over the abundance-driven analysis of DS6

The use case of DS6 relates to the study of the dynamics of IBD disease in the context of the human microbiome with multinomial imbalanced classes (related to different IBD states). The results of ML models applied over DS6 are reported in Table 4.3. Since it is a use case related to imbalanced classes, Kappa or AUC-ROC would serve as a more effective performance measure in comparison to Accuracy [225] [226].
Table 4.3 Performance of various ML models employed over DS6 (phenotype of IBD states associated with human microbiome) with 10-fold cross-validation. NFS represents the number of features. Best performance attained is bold-faced.

<table>
<thead>
<tr>
<th>Model</th>
<th>Feature Selection</th>
<th>NFS</th>
<th>Time (secs)</th>
<th>Accuracy</th>
<th>Kappa</th>
<th>AUC-ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>No</td>
<td>10996</td>
<td>3229</td>
<td>0.746</td>
<td>0.552</td>
<td>0.703</td>
</tr>
<tr>
<td>SVM (radial kernel)</td>
<td>No</td>
<td>10996</td>
<td>946</td>
<td>0.481</td>
<td>0.250</td>
<td>0.500</td>
</tr>
<tr>
<td>SVM (poly kernel)</td>
<td>No</td>
<td>10996</td>
<td>920</td>
<td>0.511</td>
<td>0.271</td>
<td>0.653</td>
</tr>
<tr>
<td>SVM (linear kernel)</td>
<td>No</td>
<td>10996</td>
<td>846</td>
<td>0.611</td>
<td>0.379</td>
<td>0.672</td>
</tr>
<tr>
<td>XGBoost</td>
<td>No</td>
<td>10996</td>
<td>198</td>
<td>0.770</td>
<td>0.613</td>
<td>0.730</td>
</tr>
<tr>
<td>Pen-LR using glmnet $\alpha = 1$(LASSO)</td>
<td>No</td>
<td>10996</td>
<td>254</td>
<td>0.695</td>
<td>0.468</td>
<td>0.657</td>
</tr>
<tr>
<td>Pen-LR using glmnet $\alpha = 0.5$ (ENet)</td>
<td>No</td>
<td>10996</td>
<td>299</td>
<td>0.728</td>
<td>0.537</td>
<td>0.728</td>
</tr>
<tr>
<td>Pen-LR using glmnet $\alpha = 0$ (Ridge)</td>
<td>No</td>
<td>10996</td>
<td>5109</td>
<td>0.770</td>
<td>0.610</td>
<td>0.709</td>
</tr>
<tr>
<td>Pen-LR using glmnet $\alpha = 0.3$</td>
<td>No</td>
<td>10996</td>
<td>406</td>
<td>0.747</td>
<td>0.570</td>
<td>0.712</td>
</tr>
<tr>
<td>RF XGBoost</td>
<td></td>
<td>449</td>
<td>46</td>
<td>0.795</td>
<td>0.654</td>
<td>0.746</td>
</tr>
<tr>
<td>RF XGBoost and glmnet</td>
<td></td>
<td>775</td>
<td>35</td>
<td>0.775</td>
<td>0.644</td>
<td>0.718</td>
</tr>
<tr>
<td>RF glmnet</td>
<td></td>
<td>384</td>
<td>61</td>
<td>0.787</td>
<td>0.602</td>
<td>0.745</td>
</tr>
<tr>
<td>RF IG</td>
<td></td>
<td>20</td>
<td>151</td>
<td>0.726</td>
<td>0.542</td>
<td>0.706</td>
</tr>
<tr>
<td>SVM (linear kernel) XGBoost</td>
<td></td>
<td>449</td>
<td>30</td>
<td>0.742</td>
<td>0.592</td>
<td>0.743</td>
</tr>
<tr>
<td>SVM (linear kernel) XGBoost and glmnet</td>
<td></td>
<td>775</td>
<td>104</td>
<td>0.652</td>
<td>0.436</td>
<td>0.710</td>
</tr>
<tr>
<td>SVM (linear kernel) glmnet</td>
<td></td>
<td>384</td>
<td>40</td>
<td>0.719</td>
<td>0.539</td>
<td>0.735</td>
</tr>
<tr>
<td>SVM (linear kernel) IG</td>
<td></td>
<td>20</td>
<td>152</td>
<td>0.685</td>
<td>0.480</td>
<td>0.670</td>
</tr>
<tr>
<td>RF RFE</td>
<td>---</td>
<td>~25K</td>
<td>0.710</td>
<td>0.510</td>
<td>0.695</td>
<td></td>
</tr>
<tr>
<td>SVM (radial) RFE</td>
<td>---</td>
<td>~70K</td>
<td>0.600</td>
<td>0.420</td>
<td>0.662</td>
<td></td>
</tr>
</tbody>
</table>
In this use case, XGBoost achieved better performance (in terms of predictive performance metric of Kappa and AUC-ROC) for classifying the human microbiome into IBD diseased states (i.e., data source DS6 with imbalanced classes) in comparison to RF and SVM models (Table 4.3).

\[ \text{glmnet with } \alpha = 0 \text{ provided better predictive performance than classical models (Table 4.3), but at the cost of running time when compared with RF (with 5109 secs > 3229 secs). The results indicate that over DS6, glmnet with } \alpha = 0.5 \text{ (ENet) provided better AUC-ROC than the } \alpha = 0.3 \text{ parametric settings.} \]

Overall, the advantage of XGBoost and Pen-LR using glmnet circumvented around their speed of execution and ability to handle sparse values in high-dimensional metagenomic datasets effectively [176], [196],[214] over RF and SVM. The results highlight that EM methods served as scalable and faster models when benchmarked with state-of-the-art RF and SVM. Investigating the predictive quality performance of different combinations of feature selection and classifiers was also performed over the DS6.

Interestingly, from Table 4.3, it is observed that obtaining features from the XGBoost model and applying classifiers of RF or SVM proved as an effective strategy in comparison to the strategies involving other feature selectors (Table 4.3). The ensemble setting of XGBoost as feature selector and RF for classification produced best performance results in this use case with (Time: 46 secs, Accuracy: 0.795, Kappa: 0.654, AUC-ROC: 0.746, No: of Features (NFS): 449) (bold-faced in Table 4.3), which indicates an improvement over RF applied on source abundance dataset with (Time: 3229 secs, Accuracy: 0.746, Kappa: 0.552, AUC-ROC: 0.703, No: of Features (NFS): 10997) ( Table 4.3).
For this dataset (DS6), the performance of the SVM model applied over XGBoost selected features improved to (Time: 30 secs, Accuracy: 0.742, Kappa: 0.592, AUC-ROC: 0.743, No: of Features (NFS): 449) (Table 4.3) from (Time: 846 secs, Accuracy: 0.611, Kappa: 0.379, AUC-ROC: 0.672, No: of Features (NFS): 10996) (Table 4.3).

RFE over SVM appeared to time-intensive and did not perform better than other feature selection strategies (Table 4.3). The FFS method of IG also does not appear to perform better than EM methods with parametric settings listed in section 4.3 (Table 4.3).

A t-test was carried out to determine if there exists a statistically significant difference between the RF predictions over OTU abundance count data (with no feature selection) and RF predictive estimations over XGBoost selected features for classifying the human microbiome into IBD diseased states (DS6). The results reported $p$-value <0.05 indicate a significant difference in their performances. A similar analysis was applied in the context of the SVM classifier. SVM employed over XGBoost selected features provided significant improvement ($p$-value <0.05). ANOVA with post-hocs methods was conducted to confirm further the statistical significance between the different groups: (i) classification over the abundance profiles and (ii) the classification of EM selected features. With AUC-ROC, ANOVA indicated $p$-value = 0.022, whilst comparing RF, RF over XGBoost selected features (RFXG), SVM and SVM over XGBoost selected features (SVMXG). The marginal significant differences were observed between SVMXG and SVM with Tuckey’s HSD value of 0.073 > calculated threshold of 0.700. The overall differences in the performance were observed as indicated in Figure 4.5.
Figure 4.5 Analysis conducted for measuring the difference of significance between AUC-ROC obtained over 10-fold of cross-validation using SVM (employed directly over the abundance profiles, SVMXG (SVM employed over feature-set obtained from XGBoost), RF (employed directly over the abundance profiles, and RFXG (RF employed over feature-set obtained from XGBoost).

The analysis confirms that the feature-set obtained by the XGBoost model is worthwhile to perform predictive analysis over DS6. SVM and RF applied over XGBoost selected features improved the predictive ability of these models (can be seen in Box Plots in Figure 4.5). The best modelling combination of XGBoost (as feature selector) and RF (as a classifier) showed high predictive quality. The analysis using RF over the XGBoost selected features suggested some taxa playing a predominant role in classifying IBD states of the human microbiome (shown in Figure 4.6).
Figure 4.6 Important taxa identified with EM-based Feature selection over DS6 playing an important role in classifying IBD diseased state of the human microbiome.

4.4.2 A comparison with an extant study

A further analysis was completed to determine a reduced subset of features by the RF algorithm before its application as a classifier (as an Embedded RF approach). This approach was suggested by Pasolli et al. [128] for the meta-analysis of large metagenomic datasets. The results of this ensemble (RF as feature selector and RF as a classifier abbreviated as RF_RF) over DS5 and DS6 has been obtained and are summarised in Table 4.4.
Table 4.4 Performance of RF over RF (i.e., RF_RF) method with 10-fold cross-validation over DS5 and DS6.

<table>
<thead>
<tr>
<th>Data</th>
<th>Time (secs)</th>
<th>Accuracy</th>
<th>Kappa</th>
<th>AUC-ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS5</td>
<td>1680</td>
<td>0.947</td>
<td>0.900</td>
<td>0.941</td>
</tr>
<tr>
<td>DS6</td>
<td>1315</td>
<td>0.762</td>
<td>0.598</td>
<td>0.730</td>
</tr>
</tbody>
</table>

The results in Table 4.4. are further compared with the application of RF over the XGBoost selected feature subset. Comparing the performance of RF over RF (i.e., RF_RF) with RF over XGBoost (i.e., RF_XGBoost) with 10-fold cross-validation, the ensembled method of RF_XGBoost exhibited better results than RF_RF (Figure 4.7), especially in terms of the computational time. Here, the top 10 features were selected as part of RF_RF modelling.

Figure 4.7 Comparing RF_RF and RF_XGBoost in terms of time and predictive performance over DS5 and DS6.
4.4.3 A study of dynamics of the embedded methods and their impact in current research

The chapter further studied the dynamics of empirical results. The impact of tuning the number of trees in the RF model applied over XGBoost selected features from the human microbiome use cases were studied. It was discovered that the RF classifier with a different number of trees performed robustly over the XGBoost-driven features in inferring metagenomic functions. Tuning the number of trees to 100, 200, 300 and 500 in RF over XGBoost-driven features (keeping XGBoost settings intact) from DS5 and DS6, did not bring major variation in the predictive performance. However, changing the parametric configuration of RF and XGBoost may impact the overall performance of predictive modelling. Predictive Accuracy of RF was improved to 0.962 (with ntree in RF set to 500 and XGBoost depth, i.e., max_depth set to 5) when the ensemble of XGBoost and RF is applied over DS5. Also, the ensemble of XGBoost and RF further improved the performance from the Accuracy of 0.746 to 0.817 when it was tuned to 500 as the number of trees (ntree) and depth as 5 in RF over DS6. It is suggested that the EM methods of feature selection with parametric tunings (such as the number of trees: 100/300/500 in RF, and max depth in XGBoost as 3-6), may have potential to further enhance the performance of predictive analysis and worthy of exploring. SVM (with the linear kernel) also proved robust when applied over features selected by XGBoost, as not much variations in predictive performance were observed by tuning the cost and gamma parameters. Overall, the results presented in this study indicate the importance of feature selection in microbiome studies. A comparison summary between the performance of state-of-the-art classifiers
without feature selection and with the feature selection (with EM methods) on DS5 and DS6 is shown in Figure 4.8(a, b). Figure 4.8(a, b) indicates that feature selection plays an important role in classifying the high-dimensional human microbiome into phenotypes.

Figure 4.8a Comparative results of predictive performance obtained before and after application of Feature Selections with RF and SVM over DS5 and DS6.
Experiments and results

Figure 4.8b Comparative Results of execution time (in secs) over DS5 and DS6.

It is interesting to note that EM methods have good potential to serve as feature selectors. To seek best performing EM methods from each Table 4.2.4.3, a comparison is conducted of their direct application as classifier over the abundance profiles, and; their role as a feature selector precursor to
classification by classical methods. Results in Figure 4.9 shows the utility of EM methods as feature selecting strategies.

**Figure 4.9** Comparative results of Predictive performance obtained by EM methods as classifiers over the abundance profiles, and; their application as feature selectors precursor to models of RF and SVM over DS5 and DS6.

### 4.5 Key Findings

As metagenomics data are usually represented with small sample size and many features, with many zeros, EM methods proved beneficial in handling such data due to their inherent properties. The advantage of using EM methods lies in the fact that XGBoost and Pen-LR methods have in-built regularization, which prevents the ML model from overfitting [112], [196]. Also, these methods have shown the good capability to handle sparse input metagenomic data. The work described in this chapter discusses the use of EM methods to select features and to predict the phenotypic function of metagenomes accurately. The work has been encouraged by the need to make the classification of metagenomes more accurate, faster, and scalable. The results (in section 4.4) show that the EM
methods were able to select discriminative features on the problem of functional classification in metagenomics of the human microbiome. The features selected with these methods were noticed to be more suitable to be used as a precursor step to RF and SVM, the classical classifiers. The EM methods of XGBoost and Pen-LR proved to be highly scalable for efficient calculation of phenotypes from the metagenomic data (DS5, DS6). Based on the observed results in section 4.4., a new abundance-driven workflow is proposed in this section. The new ensembled construction combining boosting of trees (XGBoost) and/or Pen-LR as EM methods with RF or SVM classifiers is proposed in this chapter for functional metagenomics classification. This workflow has successfully been applied to the human microbiome (with phenotype: diet and IBD disease states) in this chapter and could inform future microbiome research. The workflow is named as XGB-pen-LR-Classification for future research purposes (Figure 4.10). The specific steps recommended as part of experimentation over XGB-pen-LR-Classification for metagenomic functional prediction are summarised below:

1. An abundance count table of the microbiome with phenotype-target datasets are considered as input.

2. Pen-LR with either LASSO, ENet, or Ridge or other alpha such as $\alpha = 0.3$, and/or XGBoost methods are used to reduce the dimension of the features space by ranking features based on their functionality and selecting the unique combination of features from both.

3. The optimal feature vectors obtained after the dimensionality reduction step are input into either of the two popular state-of-the-art classification algorithms: RF or SVM to analyse the effect of feature reduction.
4. Based on the optimal parameters of the models obtained in (2) or (3), Accuracy, Kappa, and AUC-ROC are calculated, and; the predicted performance of the model is compared with other methods.

It is noted that, while functional prediction has been challenging in human microbiome use cases [93], [98], [128], the ML approaches are still limited in their ability to deal with problems involving a greater number of microbial features. The proposed approach intends to solve these problems. The proposed framework XGB-pen-LR-Classification seems to perform extremely well in comparison to SVM at classifying the microbial samples from the human microbiome into the phenotype of diet and IBD states. The workflow provided competitive Accuracy to RF over DS5 and improved Accuracy over RF as applied on DS6. These methods greatly improved in computational time (in comparison to the RF) over DS5 and DS6. These methods even enhanced the performance of the classical methods of RF and SVM significantly with embedded feature modelling ($p$-value < 0.05). Based on this observation, the proposed abundance-driven workflow for functional metagenomics (in Figure 4.10), is worthwhile to explore in other microbiome use case studies. This could prove as an optimal approach for predicting functions in futuristic metagenomic studies.
The key findings in current research along the proposed workflow are listed as follows.

i. The results presented in this work suggest that EM-based feature selection provided significantly better results over the state-of-the-art of RF and SVM applied over high-dimensional datasets. For example, over DS5 with only 270 features, SVM applied over XGBoost and glmnet selected features attained 96% Accuracy in 44 secs in comparison to SVM applied directly over 6696 features (with no feature selection) attaining only 55% Accuracy.
in 580 secs. This indicates a remarkable improvement in the analysis of high-dimensional metagenomes with the proposed framework.

ii. XGBoost and Pen-LR (using glmnet) could also serve as a fast and alternative method for RF and SVM that has been established as the classical classifier in literature [93],[98].

iii. The ensemble of EM-based feature selection using XGBoost and/or Pen-LR (using glmnet) and RF as the classifier provided better predictive performance in metagenomic case studies in comparison to filter and wrapper methods over RF.

iv. The ensembled method of RF over XGBoost provided a marginal improvement in Accuracy but a major improvement in computational time over RF applied over RF [128].

v. The study reported useful microbial features (taxa) identified with feature selection strategies.

4.6 Limitations

About the limitation of this work, experiments were designed over the abundance profiles of OTUs, assuming them as independent features. However, OTUs are related biologically by their phylogeny. For the DS5 and DS6 data sources, no phylogenetic information was available. The challenges faced in efforts to characterise metagenomic functions are compounded by the fact that the observed abundance profiles of microbial communities are dominated by a small number of important taxa, whereas a recent important biological or phylogenetic change could have occurred in uncommon populations. Henceforth, more research is needed to integrate phylogeny in the functional
microbiome classification by adding evolutionary or taxonomical aspects at either data pre-processing or ML modelling levels in cases where phylogenetic information is available. It is expected that the incorporation of biological domain knowledge of phylogeny could inform metagenomics research with better biological-driven insights.

The predictive modelling in current research covers the overall procedure following Approach 1 (as proposed in Chapter 3), beginning with the essential data pre-processing with feature selections, data splitting, and foundations of extensive model tuning. However, more exhaustive parametric tunings could be explored in the future.

4.7 Summary

The explosive growth of metagenomic data has prompted the development of new computational methods for analysing phenotype targets based on ML. This study proposes a new prediction method of XGB-pen-LR-Classification. The method has been used to extract the features of the human microbiome data, and the redundant information in the data source is processed based on the EM dimensionality reduction methods. The selection of optimal parameters for models in XGB-pen-LR-Classification can significantly improve the prediction Accuracy rate. In this study, metagenomic human microbiome data was uniformly evaluated from 2 studies using 10-fold cross-validation to evaluate and predict the metagenomic functions. The results proposed the use of embedded methods for dealing with high-dimensional metagenomic functional data. For example, the proposed framework using embedded methods
overcomes SVM overfitting in predicting diet association with the human microbiome (DS5). Embedded method of XGBoost has not been explored extensively in metagenomic studies in the literature. To the best of knowledge, the current research provides the first rigorous use of XGBoost in metagenomic studies. It is proposed that in terms of computational cost, embedded methods of XGBoost and Pen-LR outperformed RF and SVM by scaling well to high-dimensional metagenomic data.

Embedded methods were combined with the popular methods of RF and SVM to further embrace performance improvements along XGB-pen-LR-Classification. Overall results indicated significant improvement in the performance of models following the XGB-pen-LR-Classification framework and the classical models of RF and SVM (p-value <0.05). Along the workflow, some relevant microbiome taxa were identified for each use case of DS5 and DS6. *Firmicutes* was emphasized as an important *phylum* in the human microbiome. The results presented in this chapter would inform human microbiome analysis in the context of the human diet, and IBD disease states and creates a baseline knowledge for any future research. Figure 4.11 summarizes the key observations made in the current study.
Summary

Key Scientific Discovery

• An abundance-driven framework is proposed based on the embedded (EM) methods of XGBoost and Pen-LR to improve the classical methods of RF and SVM for classifying high-dimensional metagenomes into their functional phenotypes.

• It was discovered that functional analysis of metagenomes benefitted from including the use of EM methods as feature selectors in handling metagenomic use cases of high dimensionality (p features >> n samples). As an example, with the use of EM methods as feature selectors, the best performance (96 % Accuracy) was attained with SVM applied on only 270 microbial features of DS5 in just 44 secs in comparison to the application of classical method of SVM over 6696 original features in ~500 seconds attaining only 55% Accuracy (Table 4.2). Similarly over DS6, RF applied over 449 features obtained from embedded feature selection attained more accurate performance in just 46 secs; in comparison to the application of RF over the 10,996 microbial features in ~ 3000 seconds (Table 4.3).

• Enterobacteriales were indicated as a likely microbial driver of classifying IBD disease states in humans.

• Erysipelotrichi and Clostridia were identified as important microbial taxa supporting dietary management in humans.

Strengths

• The framework (XGB-pen-LR-Classification) proposed in this chapter is efficient in terms of both the computational cost (time) and predictive performance (Accuracy, Kappa and AUC-ROC) over the high-dimensional metagenomes.

Limitations

• The proposed framework is applied only to abundance profiles assuming microbial features as independent without any integration of phylogenetic relationships.

Figure 4.11 A summary of key observations, strengths and limitations of this study.
CHAPTER 5

A NOVEL PHYLOGENY-DRIVEN FRAMEWORK FOR THE PREDICTION OF METAGENOMIC FUNCTIONS

This chapter provides the rationale, methodology, and construction of a novel phylogeny-driven framework for the prediction of phenotypic functions in different metagenomic environments. An algorithm is proposed to create a novel data structure of Phylogeny and Abundance-aware Matrix (PAAM) exploiting the phylogenetic ancestry and integrating taxonomical evolution of microorganisms with their abundances at the data-pre-processing level for downstream microbiome classification. The algorithm is further embedded in a comprehensive metagenomics classification framework, termed as PAAM-ML (Phylogeny and Abundance-aware Matrix-based Machine Learning). The framework is applied to diverse metagenomic samples distinguishing associated environmental phenotypes. The detailed development of the proposed framework is discussed in this chapter, highlighting its advantages over the conventional approaches employed for phenotypic predictions in metagenomic studies. Furthermore, this chapter describes how: (1) PAAM is constructed via integration of prior knowledge of phylogenetic tree with abundance count of OTUs for constructing quantitative profiles; (2) the feature selection methods are employed to select relevant features; and (3) the supervised classification of
metagenomic samples is performed and assessed. A description of the implementation of the PAAM-ML is presented, outlining the tools and techniques employed to construct the framework. The results obtained from the classification framework are discussed with a relevant summary. The work presented in this chapter is based on published work in [227],[228].

5.1 Rationale

Intense research in developing ML-based computational methods for interpreting the biological phenotypic roles of metagenomes is ongoing [28], [32]. As described in Chapter 2 and 4, the ML and statistical approaches have great potential to be applied to the abundance profiles of microbiome data facilitating the prediction of phenotypic functions. The framework proposed in Chapter 4 applied ML models on the abundance profiles of microbial OTUs in a novel way. Nonetheless, these methods face limitations due to their assumption that the microbial taxa or OTUs/ASVs or microbial features are independent of each other, and do not affect each other. However, subtle biological relationships between microbial features could affect the strength of functional predictions [41], [69], [131],[138]. The proposed method in the current chapter considers these relationships and is based on the biological assumption that phylogenetically close taxa may respond similarly to the host phenotype as supported by studies in [42], [83], [84], [132], [133].

While a substantial amount of current research involves supervised ML shaped by abundance profiles of OTUs, research on creating phylogeny distance-aware features at different levels of taxonomy is still evolving [85], [92], [121], [133]. The popular state-of-the-art phylogeny-driven ML model of
MetaPhyl [87] lacks in analysing the environment-specific microbial features at multiple granularity levels of the taxonomy. Overall, the ML methods have the potential to be improved by including phylogeny-aware search space for finding the best subset of microbiome features at different taxonomic levels in microbiome studies. Constructing feature space at multiple levels by fusing children nodes into a parent node is driven by the assumption that close taxa may behave similarly in response to functional traits as supported in studies by [42], [83]–[85]. Henceforth, an alternative novel approach of the PAAM-ML framework is presented as a significant contribution to this chapter. The framework aims to demonstrate a novel idea of creating the feature space bearing a phylogeny-driven explanatory value over the contemporary microbiome classification tools [28],[32]. The feature selection techniques are used to reduce the high-dimensional phylogeny-driven feature space to contribute towards the more targeted phylogeny-aware therapeutics. The current work is dedicated to formalizing the microbiome analysis by systematically creating and searching a suitable phylogeny-aware feature space for linking microbiome profiles to phenotype.

5.2 Methodology

A description of the proposed framework PAAM-ML for microbiome classification is detailed in Figure 5.1. The methodology follows Approach 2, which was proposed in Chapter 3 (Figure 3.5b). PAAM-ML is composed of following main phases, including (A) prior knowledge of phylogenetic tree is integrated into quantitative abundance profiles of microbial features to create a feature space of the Phylogeny and Abundance-aware Matrix (PAAM); (B)
feature selection is employed over the engineered feature space constructed by integrating the hierarchical relationships between diverse microbial features; (C) supervised ML models; (D) assessing performance of the supervised ML model/s applied; as illustrated in Figure 5.1. Additional details of each phase are presented below.
Figure 5.1 Graphical representation of the PAAM-ML prediction framework. (A) heterogenous inputs of abundance count and prior knowledge of phylogenetic relationships are combined for the construction and engineering of the phylogeny and abundance-aware Matrix (PAAM). (B) Feature ranking or selection is performed over the PAAM structure. (C) supervised ML is applied over selected features. (D) assessment of the predictive performance.
5.2.1 Construction of PAAM feature space incorporating biological domain knowledge of phylogeny and abundance count profiles

A significant contribution in this thesis has been the integration of both abundances and the phylogenetic relationships (from the phylogenetic tree) of all taxa into a set of informative features, including both leaf-level taxa and the ancestral elements present at different levels of the taxonomy. In a phylogenetic tree, leaves correspond to OTUs, and ancestral/internal nodes represent parental taxa at different levels. The quantitative profiles of OTUs are derived from the abundance count table at a phylogenetic level. However, the quantitative profiles of ancestral parent nodes are constructed by combining the abundances of their children OTU nodes and the evolutionary distances embarked on the branches connecting the parent node to its corresponding children. This approach is employed to define a novel data structure of PAAM (abbreviated as $P_{n \times 2m-1}$) containing $n$ samples, and $2m-1$ microbial features with $m-1$ as ancestral nodes and $m$ OTUs. This matrix $P_{n \times 2m-1}$, contains the taxa information considering the biological domain knowledge along a phylogenetic tree. An illustration of the creation of the PAAM feature space is depicted in Figure 5.2. In Figure 5.2, leaves are annotated as a set of OTUs, i.e., $\{\text{OTU}_j\}$ where $1 <= j <= m$; samples are annotated as a set of samples, i.e., $\{\text{Sample}_i\}$ where $1 <= i <= n$; internal nodes connecting $\text{OTU}_j$ and $\text{OTU}_l$ are annotated as a set of $\{\text{IN}_{j,l}\}$. The quantitative profile of OTUs is interpreted as $x_{i,j}$ representing the abundance of $j^{th}$ OTU in the $i^{th}$ sample. The quantitative profiles of internal nodes are annotated as $x_{i,j,l}$, representing the quantitative profile of an internal node $x$ related to a sample $i$ and created by the regularizing
abundances of children \( j \) and \( l \) with their respective phylogenetic distances (as indicated in Figure 5.2). The method for PAAM construction has been designed with a purpose to map the original OTU abundance-based feature space into a novel tree-based feature space, integrating the phylogenetic information.

The phylogenetic relationships could potentially be useful in influencing the weighted profiles of taxa and play an important role in defining the quantitative strength of the ancestral nodes. In a phylogenetic tree, two taxa are more related if evolutionary distance embarked on the branch connecting taxa and their ancestral node is less, and; are less related if the distance embarked on the branch connecting taxa and their ancestral node is more [229],[230]. The branch length (i.e. evolutionary distance annotated on the branch), is introduced hereby to regularize the quantitative profiles of children node to create profiles of parental nodes. This would further aid the computational modelling over the phylogeny and abundance-aware microbiome data, as there exists the possibility that a phylogenetic change could occur in less abundant taxa. Phylogenetic proximity between OTUs provides a structural relationship between the taxa (OTUs). The assumption that closely related taxa are likely to function similarly considers that distance between two close taxa as small [42], [83]–[85], [229]. The lower-level taxa with smaller distance or smaller branch length are more likely to be grouped as a comprehensive ancestral level feature on the phylogenetic tree and function similarly. Owing to these assumptions, the method for the construction of the PAAM feature space is summarised in Algorithm 5.1. In this way (Algorithm 5.1), biological knowledge obtained from phylogeny has been used to model relationships between the OTU features.
Methodology

**Figure 5.2** An illustration of the construction of the PAAM feature space involving internal (ancestral) nodes.

Quantitative Profile of an Internal Node $IN_{ij}$ in $ith$ sample is calculated using abundance count of its children $i$ and $j$ and their respective phylogenetic distances.

For example,

Quantitative profile of new Feature $IN_{12}$ in Sample 1 would be \( x_{1,12} = \frac{x_{1,1}}{pd_{1,12}} + \frac{x_{1,2}}{pd_{2,12}} \)

Quantitative profile of new Feature $IN_{1234}$ in Sample 1 would be \( x_{1,1234} = \frac{x_{1,1}}{pd_{1,12}pd_{12,1234}} + \frac{x_{1,2}}{pd_{2,12}pd_{12,1234}} + \frac{x_{3,1}}{pd_{3,14}pd_{34,1234}} + \frac{x_{3,3}}{pd_{3,34}pd_{34,1234}} + \frac{x_{3,4}}{pd_{4,34}pd_{34,1234}} \)
Algorithm 5.1 Construction of PAAM Feature Space

Input: OTU Abundance Count Matrix $C_{n \times m}$, Phylogeny Tree ‘T’

Output: PAAM denoted as $P_{n \times 2^m - 1}$

procedure PAAMConstruction (OTU Abundance Count Matrix ‘C,’ Phylogeny Tree ‘T’)

# Quantitative profiles of leaf-level features remain intact
for each sample row, ‘i’ in original Abundance Count Matrix $C_{n \times m}$ with $n$ samples and $m$ OTU features do
   for each feature $j$ in $C$ and $T$
      $P_{i,j} \leftarrow C_{i,j}$
   end for
end for

# Quantitative Profile of Ancestral Nodes
for each sample row, ‘i’ in original Abundance Count Matrix $C_{n \times m}$ with $n$ samples and $m$ OTUs do
   j $\leftarrow m + 1$
   for each ancestral node $v$ in a phylogeny tree $T$ do
      $P_{i,j} \leftarrow 0$
      for each OTU node ‘u’ in $C_{n \times m}$ do
         if OTU ‘u’ in the sample ‘i,’ is descendant of node ‘v’ in T then
            $PD_{u,v} \leftarrow$ Phylogeny distance of OTU ‘u’ from ‘v’
            $X_{u,i} \leftarrow$ Abundance count of OTU ‘u’ in the sample ‘i’
            The weighted abundance of ancestral node ‘v,’
            i.e., $WA(v)$ is formulated as $WA(v) \leftarrow X_{u,i} / PD_{u,v}$
            $P_{i,j} \leftarrow P_{i,j} + WA(v)$
         end if
      end for
   end for
   j $\leftarrow j + 1$
end for

return PAAM Feature Space, $P_{n \times 2^m - 1}$

As detailed in Algorithm 5.1, the abundance of each taxon is weighted by the phylogenetic distances annotated on branches of a phylogenetic tree to span ancestral nodes at each level of a tree, forming a hierarchical topology. This topology is useful to expand the research presented in this thesis and the prediction of metagenomic functions when integrating the biological knowledge.
of evolutionary distances. In this work, both quantitative knowledge of OTU abundances and evolutionary distances and; qualitative knowledge of taxonomy is incorporated in the functional analysis to determine phenotype as a significant contribution. Theoretically, with this construction, a successive ML model could be applied to different levels of taxonomy using the hierarchical feature space of PAAM. However, the feature space generated is high-dimensional. Feature selection strategies over such high-dimensional feature space are expected to yield advantage in the classification of metagenomes by reducing the complexity. Hence, it is useful to employ feature selection for the prediction of phenotypes and to study their effect on the predictive performance of supervised machine learning methods.

5.2.2 Feature selection over the hierarchical feature space of PAAM

A minimal set of features is preferred to determine phenotypic functions with excellent predictive performance. A reduction in the number of features has the potential advantage of reducing the computational complexity of the adoption of the supervised ML model. The power of feature selection is twofold, as it aids in knowledge discovery from PAAM feature space and deals with its curse of dimensionality [31]. Heuristics to select features from hierarchical feature space are applied using different feature selection strategies in the current study. These involve embedded feature selection techniques of Random Forest Importance (RFI) [88], [189], [231], XGBoost, and Pen-LR(using glmnet). It is further benchmarked with feature selection technique of feature filters of Correlation-based filtering (CFS) [192]. CFS is chosen as it works by evaluating the merit of a correlation between a feature and the functional class considering
the usefulness of individual features for phenotypic class. The techniques are employed in this study to distil and characterised informative features from the metagenomic data of PAAM. The emphasize is put on using embedded and a filter approach as wrapper strategies are time intensive [189],[232]. This would aid in making the framework easily interpretable for subsequent downstream analysis.

5.2.3 Supervised ML over selected features from PAAM feature space

In the next phase of the proposed framework, PAAM-ML, a classification model is learned and assessed over the selected features with k-fold cross-validation. To identify the most suitable model for predicting functional metagenomes, the most popular supervised ML classification algorithms of RF and SVM in functional metagenomics (as discussed in Chapter 2), are employed to determine their fitness in the prediction task [93],[98], [128]. Additional methods, including the MetaPhyl [87] and PhILR [133], are employed for a comparative assessment and benchmark study of the phylogeny distance-driven methods. It is worthwhile to determine whether PAAM-ML would outperform the phylogeny-driven method of MetaPhyl, which does not consider multiple granularity levels of phylogeny and the other means of PhILR, which models the various levels of the taxonomy. In this way, a rigorous comparison of PAAM-ML with different algorithms is conducted, in terms of obtaining successful classification models for predicting metagenomic functions.
5.2.4 Assessment of supervised ML algorithms

The cross-validation estimates are useful for an overall performance assessment of classification. A supervised learning algorithm is trained and tested k times; by testing on the $k^{th}$ fold whilst using remaining $k-1$ folds for model training. A procedure is carried out in which PAAM data is randomly split into $k$ exclusive subsets where $k = 10$ (k-fold cross-validation), or $k =$ number of sampled instances (i.e., LOOCV). The results primarily are presented in terms of the performance assessment metrics of Accuracy and Kappa for evaluating classification models in this study, because of their simplicity and successful application in predictive tasks using caret package in R [225],[233],[234]. Kappa is more robust than scalar metrics of Accuracy, in the case of imbalanced datasets, and when the small number of samples are associated with different classes [177]. The main difference between Accuracy and Kappa scores is that the Accuracy measures all correct classifications (i.e., successes) over all the phenotypic classes, whereas Kappa calculates successes independently for each functional type and aggregates them later. Kappa measures the agreement between two raters ranging from -1 (an example of total disagreement) through 0 (an example of random classification) to 1 (an example of perfect agreement) [177]. Statistical tests are performed to compare PAAM-ML with other phylogeny-aware methods.

The above four sub-sections (5.2.1- 5.2.4) corresponds to the four steps (A-D) of the proposed methodological framework in the current study (Figure 5.1).
5.3 Implementation of PAAM-ML

The input metagenomic data sources of the OTU abundance count table and phylogenetic tree are stored in a comma-separated file format (.csv) and Newick format [48], respectively. Data sources have been pre-processed using R packages of ape [181] in case the taxa in the OTU table and tips in the phylogenetic tree do not match. Newick tree formats are readable and downloadable in the R platform as a with the help `read.tree()` function in `ape` package [181]. `drop.tip()` function in the ape package helps in selecting taxa from the tree that matches taxa in the OTU table in R [181]. After the OTU table and phylogenetic tree are made consistent by matching the tip labels of tree and taxa in the OTU table, these are integrated for the prediction of phenotypes.

Perl-based pipeline [235], is developed to integrate the two different data types of OTU table and a phylogenetic tree into a single data structure of ‘PAAM.’ The construction of PAAM is implemented in Perl [235], a general-purpose scripting language, which aided in reading, and processing of the Newick tree format [48] and to read the corresponding evolutionary distances of respective OTUs. Using these evolutionary distances to regularize microbial abundances is implemented as part of this script. Figure 5.3 illustrates the interface of the PAAM generation.
Implementation of PAAM-ML

Figure 5.3 An Illustration of the GUI interface of the PAAM generation. This screenshot demonstrates the selection of abundance count table and phylogenetic tree to attain PAAM for the downstream analysis.

After the construction of PAAM, features are selected for prediction of metagenomic phenotypes. The approach progressed by (i) measuring the importance of PAAM features using RFI (*randomForest* package in R) and (ii) application of different ML models, using the *caret* package in R [233],[234],[236] using LOOCV settings over DS1 and DS4; and 10-fold cross-validation over DS2. The package implements models using resampling of results across different tuning parameters: SVM (svmPoly) with *degree*, *scale* and *c*; RF(rf) implemented with tuning *mtry*; XGBoost(xgbTree) with nrounds, max_depth, *eta*, *gamma*, colsample_bytree, min_child_weight and subsample;
and glmnet with alpha and lambda values [237]. It reports an optimal model by estimating Accuracy and Kappa using the largest value attained while iterating through the parametric settings. By default, XGBoost and glmnet reported 20 important features using varImp() function in caret package for an ML-based analysis. PhILR is implemented as an R package sourced from [238]. MetaPhyl [87] is implemented as a C++ executable available at [239].

5.4 Predictivity of microbial features with PAAM construction

Certain analysis has indicated that incorporating information on evolutionary distances between taxa is useful in microbiome analysis [50], [87], [92], [133], [240]. This may be due to the assumption that related microbes may be more likely to share similar phenotypic traits [42],[83],[84],[85]. With PAAM construction, one-to-one correspondence is established between microbial features profiles and the nodes on the phylogenetic tree. The evolutionary information is integrated into the PAAM feature space by scaling abundances of a node by using the phylogenetic distance between its direct descendants. These phylogenetic weights are chosen in the heuristic design to model the integrative feature space. These could also serve as a heuristic as they might influence microbial features with many zero abundance values. Due to introducing scaling by phylogenetic weights, parental nodes could get a non-zero value, even if one of the dependent child nodes is having zero count values. Applying feature selection strategies over the PAAM feature space could identify microbial clades with internal nodes (ancestors) that may help in differentiating functional phenotypes better in comparison to the microbial
features of leaf-level OTUs. It is expected that the Accuracy of supervised classification methods on the sourced data sources is matched or improved with feature engineering over the PAAM feature space data in comparison to applying the same feature selection models on the abundance count profiles (only). A toy example illustrating PAAM feature space (based on Algorithm 5.1) and its utility is shown in Figure 5.4.

---

**Figure 5.4** An Illustration of a toy example of how PAAM feature space could result in better classification of functional metagenomes.

Figure 5.4 demonstrates the utility of PAAM for the downstream analysis. Additional information in discriminating microbial samples into phenotype
could be gained by including the generated profiles of internal (ancestral) nodes. The quantitative profile of $IN_{12}$ better discriminates between classes A and B. For example, samples having the weight profile of $IN_{12} < 300$ could be classified as A, otherwise, as the class B. Similarly, if the weight of $IN_{34} > 350$, samples could be classified as A otherwise class B; and if the $IN_{345} > 90$, then samples are classified as A otherwise as B. A sample of such patterns provides better opportunities aiding in the functional classification. However, no such explicit patterns were observed while classifying samples based on only the original OTUs in this example (Figure 5.4). Henceforth, there is likely potential for improvement in functional metagenomics with the development of a novel feature space that accounts for the phylogeny and evolutionary characteristics of metagenomes including the ancestral nodes. Also, the possibility of sparseness in ancestral node profiles is lesser in comparison to the raw abundance profiles of OTUs. Feature selection over such feature space could help in reducing the dimensionality and sparsity of metagenomes. Various experiments conducted, and the results obtained are discussed in the next section.

5.5 Experiments and results

To show the applicability of PAAM-ML over the diverse and heterogeneous datasets, human and soil microbiome are utilised in this chapter. The chosen environments for this study are listed as DS1, DS2, and DS4 in the context of the current thesis (Table 3.5 in Chapter 3). DS1 deals with throat microbiome with 60 samples associated with smokers and non-smokers human beings, DS2 deals with human microbiome with 3285 samples associated with 4 body sites
and DS3 deals with soil microbiome having 139 samples with 3 glucose treatment as phenotypes. 10-fold cross-validation is applied over DS2. Due to the limited sample size in DS1 and DS4, LOOCV is used over these data sources to be used in ML analysis. The training data have an imbalanced representation of sample classes in DS2. The detailed empirical analysis of these diverse metagenomic environments, along with the proposed PAAM-ML framework, employs a variety of experiments. The conceived experiments are discussed below.

### 5.5.1 Distinguishing microbial profiles using abundance and phylogeny-driven Measures

In the first experiment, state-of-the-art classifiers of RF [88], SVM [108] has been applied over microbial profiles in all three use cases DS1, DS2 and DS4. ML models are benchmarked when built using PAAM (i.e. with a phylogenetic measure), against the results obtained using only the OTU abundance profiles (i.e. non-phylogenetic measure). The results are compared in Table 5.1.

In use cases of DS1 and DS2, RF and SVM applied over PAAM produced relatively better performance results than when applied over the raw OTU abundances only (bold-faced values in Table 5.1). These findings supported that the observed internal nodes may lead to better functional predictions. However, no such improvement was observed over DS4 (Table 5.1).
Table 5.1 Summary of performance results obtained by the application of RF and SVM classifiers over PAAM and OTU abundance matrix obtained from different metagenomic data sources with cross-validation. Results are obtained by choosing the best value after resampling results across different tuning parameters over different folds (in caret package).

<table>
<thead>
<tr>
<th>Input</th>
<th>PAAM</th>
<th>OTU ABUNDANCE MATRIX</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Accuracy</td>
</tr>
<tr>
<td>Data</td>
<td>Classifier</td>
<td>NFS</td>
</tr>
<tr>
<td>DS1</td>
<td>SVM</td>
<td>1711</td>
</tr>
<tr>
<td>DS2</td>
<td>SVM</td>
<td>11660</td>
</tr>
<tr>
<td>DS2</td>
<td>RF</td>
<td>11660</td>
</tr>
<tr>
<td>DS4</td>
<td>SVM</td>
<td>2143</td>
</tr>
<tr>
<td>DS4</td>
<td>RF</td>
<td>2143</td>
</tr>
</tbody>
</table>

Table 5.1 reports the results of classifiers applied over the OTU abundance count table and PAAM for a future benchmark. However, PAAM has almost twice as many features as in the OTU abundance table and hence have higher dimensionality. The increased dimensionality may influence building an accurate model in Table 5.1. Therefore, applying feature selection (as detailed in the next steps of analysis), has the potential to improve ML modelling over PAAM.

5.5.2 Classification of the microbiome using feature selection strategy of RFI

The chapter further investigates the use of embedded feature selection strategies for selecting important microbial taxa from the PAAM. The studies such as in [128], [241] suggested the potential use of the RF methodology to obtain the variable importance. RFI has been employed in the current study to rank the
importance of features in the microbial feature space. The analysis is compared further with other embedded feature selection strategies of XGBoost and/or Pen-LR; and; with the feature filter of CFS. The goal of this task is to analyse whether feature selection considering relationships among microbial features in PAAM might lead to better prediction performance. The classification method of SVM has been employed over the top 5, 10, 20, 40 % of the feature set obtained from embedded RF importance ranking (denoted as RFI_5%, RFI_10%, RFI_20%, RFI_40% respectively), from PAAM in an exploratory study. It is observed that an ensemble of RF as a feature selection strategy and SVM for classification, produced a relatively better predictive performance in comparison to its application on OTU abundance table in all the three use cases (comparing results bold-faced in Table 5.2-5.4).

Table 5.2 Summary of performance results obtained by application of SVM over the feature space obtained after RFI ranking of features in PAAM and OTU abundance count of DS1, dealing with human throat microbiome. Results are averaged over all LOOCV folds by choosing the best value after resampling results across different tuning parameters.

<table>
<thead>
<tr>
<th>PAAM</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature Selection</td>
<td>Accuracy</td>
</tr>
<tr>
<td>RFI_5%</td>
<td>0.867</td>
</tr>
<tr>
<td>RFI_10%</td>
<td>0.900</td>
</tr>
<tr>
<td>RFI_20%</td>
<td>0.783</td>
</tr>
<tr>
<td>RFI_40%</td>
<td>0.783</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OTU Count Matrix</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature Selection</td>
<td>Accuracy</td>
</tr>
<tr>
<td>RFI_5%</td>
<td>0.817</td>
</tr>
<tr>
<td>RFI_10%</td>
<td>0.783</td>
</tr>
<tr>
<td>RFI_20%</td>
<td>0.867</td>
</tr>
<tr>
<td>RFI_40%</td>
<td>0.833</td>
</tr>
</tbody>
</table>
Table 5.3 Summary of performance results obtained by the application of SVM over the feature space obtained after RFI ranking of features in PAAM and OTU abundance count of DS2, dealing with human microbiome (body sites). Results are averaged over all 10-fold by choosing the best value after resampling results across different tuning parameters.

<table>
<thead>
<tr>
<th>PAAM</th>
<th>Feature Selection</th>
<th>Accuracy</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>RFI_5%</td>
<td></td>
<td>0.937</td>
<td>0.895</td>
</tr>
<tr>
<td>RFI_10%</td>
<td></td>
<td>0.941</td>
<td>0.901</td>
</tr>
<tr>
<td>RFI_20%</td>
<td></td>
<td>0.942</td>
<td>0.903</td>
</tr>
<tr>
<td>RFI_40%</td>
<td></td>
<td>0.940</td>
<td>0.899</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OTU Count Matrix</th>
<th>Feature Selection</th>
<th>Accuracy</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>RFI_5%</td>
<td></td>
<td>0.928</td>
<td>0.881</td>
</tr>
<tr>
<td>RFI_10%</td>
<td></td>
<td>0.934</td>
<td>0.890</td>
</tr>
<tr>
<td>RFI_20%</td>
<td></td>
<td>0.938</td>
<td>0.896</td>
</tr>
<tr>
<td>RFI_40%</td>
<td></td>
<td>0.935</td>
<td>0.891</td>
</tr>
</tbody>
</table>
Table 5.4 Summary of performance results obtained by application of SVM over the feature space obtained after RFI ranking of features in PAAM and OTU abundance count of DS4, dealing with soil microbiome (substrate treatment). Results are averaged over all LOOCV folds by choosing the best value after resampling results across different tuning parameters.

<table>
<thead>
<tr>
<th>Feature Selection</th>
<th>PAAM SVM</th>
<th>OTU Count Matrix</th>
<th>SVM SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>RFI_5%</td>
<td>0.986</td>
<td>Accuracy</td>
<td>Kappa</td>
</tr>
<tr>
<td>RFI_10%</td>
<td><strong>0.993</strong></td>
<td>0.988</td>
<td>0.989</td>
</tr>
<tr>
<td>RFI_20%</td>
<td><strong>0.993</strong></td>
<td>0.989</td>
<td>0.989</td>
</tr>
<tr>
<td>RFI_40%</td>
<td>0.986</td>
<td>0.978</td>
<td>0.978</td>
</tr>
</tbody>
</table>

The integration of phylogeny with OTU abundance resulted in an improvement of microbiome sample classification by using a smaller number of features over the three use cases DS1, DS2, and DS4 (Table 5.2-Table 5.4). Overall, the results in Table 5.2-Table 5.4, indicate that the features obtained by the application of RFI over PAAM could play an important role in determining the functional roles in metagenomics. The selected features included both the internal nodes as well as the OTUs. The best performance obtained over DS1 (Table 5.2) is the Accuracy of 0.900 and Kappa of 0.79, with an application of svmPoly with degree =1, scale = 0.1 and C=0.25 (optimal results achieved through caret package) over 10 % features obtained (by RFI) from PAAM (i.e., ~170 microbial features). Over DS4 (Table 5.4), the best modelling with
svmPoly (Accuracy of 0.993 and Kappa 0.989) was achieved with degree =1, scale = 0.1 and C=0.25 over top 10 % features of PAAM (~215 in number). The percentages of internal nodes proved to be dominant in features selected by RFI in these use cases with DS1 having ~88% of internal nodes, and DS4 having ~70 % of internal nodes.

The best performance was achieved by RF (at mtry = 48) applied over RFI_n% (where n =10) (i.e., ~1165 features from PAAM), with an Accuracy of 0.977 and Kappa of 0.961 with most internal nodes (90 %) over source DS2. However, the application of SVM over PAAM selected features also indicated a relative improvement in functional prediction (Table 5.3). Over DS1 and DS4, the highest performance of RF employed over RFI_n% (where n=5/10/20/40) selected features were noticed as Accuracy of 0.767 and Kappa of 0.529 (in use case DS1 with 170 features); and Accuracy of 0.935 and Kappa of 0.903 (in use case DS4 with 109 features). SVM performed relatively better over RFI selected features in the use cases DS1 and DS4 (Table 5.2, Table 5.4) in comparison to RF.

Overall, results in Table 5.2-5.4 indicated that the phylogeny-aware modelling produced relatively better or competent results in comparison to application of SVM or RF models over the OTU abundance count table.

Additionally, the performance improvement was observed by application of Pen-LR over RFI selected features (RFI_10%) from PAAM of DS1, (Figure 5.5).
Figure 5.5 An illustration of the comparison of variation in performance of Pen-LR applied over OTU count table and PAAM (DS1).

5.5.3 Classification of the microbiome using feature Selection strategy of XGB-pen-LR-Classification

In this sub-section, empirical results obtained by assessing the two representative embedded feature strategies of XGBoost [196] and Pen-LR [112] (as suggested in XGB-pen-LR-Classification framework in Chapter 4) are presented. These are further compared with the results obtained in the above set of experiments (section 5.4.2) along with the PAAM-ML framework. The main goal of these methods is to achieve scalability over high-dimensional PAAM. The results in Table 5.5-5.7, have been reported to see the impact of embedded feature selection methods of XGBoost and/or Pen-LR in predictive performances obtained in this study over DS1, DS2 and DS4.
Table 5.5 Summary of performance results obtained by application of SVM over the number of features (NFS) obtained after XGBoost and/or Pen-LR ranking of features in PAAM and OTU abundance count of DS1, dealing with human throat microbiome. Results are averaged over all folds and choosing the best value after resampling results across tuning parameters.

<table>
<thead>
<tr>
<th>Input</th>
<th>PAAM</th>
<th>OTU ABUNDANCE MATRIX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature Selection</td>
<td>Classifier</td>
<td>NFS</td>
</tr>
<tr>
<td>XGBoost SVM</td>
<td>20</td>
<td>0.833</td>
</tr>
<tr>
<td>XGBoost RF</td>
<td>20</td>
<td>0.767</td>
</tr>
<tr>
<td>XGBoost &amp; Pen-LR RF</td>
<td>40</td>
<td>0.750</td>
</tr>
<tr>
<td>XGBoost &amp; Pen-LR SVM</td>
<td>40</td>
<td>0.983</td>
</tr>
</tbody>
</table>

Table 5.6 Summary of performance results obtained by application of SVM over the number of features (NFS) obtained after XGBoost and/or Pen-LR ranking of features in PAAM and OTU abundance count of DS2, dealing with human body sites. Results are averaged over all folds and choosing the best value after resampling results across tuning parameters.

<table>
<thead>
<tr>
<th>Input</th>
<th>PAAM</th>
<th>OTU ABUNDANCE MATRIX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature Selection</td>
<td>Classifier</td>
<td>NFS</td>
</tr>
<tr>
<td>XGBoost SVM</td>
<td>20</td>
<td>0.903</td>
</tr>
<tr>
<td>XGBoost RF</td>
<td>20</td>
<td>0.975</td>
</tr>
<tr>
<td>XGBoost &amp; Pen-LR RF</td>
<td>40</td>
<td>0.975</td>
</tr>
<tr>
<td>XGBoost &amp; Pen-LR SVM</td>
<td>40</td>
<td>0.904</td>
</tr>
</tbody>
</table>
Table 5.7 Summary of performance results obtained by application of SVM over the number of features (NFS) obtained after XGBoost and/or Pen-LR ranking of features in PAAM and OTU abundance count of DS4, dealing with soil microbiome. Results are averaged over all folds and choosing the best value after resampling results across tuning parameters.

<table>
<thead>
<tr>
<th>Feature Selection</th>
<th>Classifier</th>
<th>NFS</th>
<th>Accuracy</th>
<th>Kappa</th>
<th>NFS</th>
<th>Accuracy</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SVM</td>
<td>20</td>
<td>0.913</td>
<td>0.871</td>
<td>20</td>
<td>0.928</td>
<td>0.892</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>20</td>
<td>0.956</td>
<td>0.935</td>
<td>20</td>
<td>0.942</td>
<td>0.913</td>
</tr>
<tr>
<td>XGBoost &amp; Pen-LR</td>
<td>RF</td>
<td>38</td>
<td>0.971</td>
<td>0.956</td>
<td>38</td>
<td>0.971</td>
<td>0.957</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>38</td>
<td>0.986</td>
<td>0.978</td>
<td>38</td>
<td>0.978</td>
<td>0.967</td>
</tr>
</tbody>
</table>

From the tables above (Table 5.5-5.7) it is evident that the application of embedded methods has the potential to improve the predictive performance over high-dimensional metagenomes in PAAM. When the features from XGBoost and Pen-LR (over DS1) were combined, the predictive performance of SVM with degree = 1, scale =0.1 and C = 0.25, was improved (Accuracy: 0.983 and Kappa: 0.966) (bold-faced value in Table 5.5) with only 40 microbial features (a mix of internal node and leaf-level node features).

Over DS2, feature set obtained by application of XGBoost attained a competitive Accuracy of 0.975 and Kappa of 0.957 with only 20 features (bold-faced value in Table 5.6) in comparison to earlier results of RF applied over RFI_10% (~1165 features from PAAM, Table 5.3). This obtained feature set of 20 entries contained all the ancestral node features. This indicates the importance of the generated feature space of PAAM.
However, over the use case of DS4, XGBoost and/or Pen-LR applied over PAAM (bold-faced in Table 5.7) provided competitive performance to RFI feature selection but with a smaller number of features (38) in comparison to 107 features selected by RFI (Table 5.4).

Overall, results in Table 5.5-5.7 achieved better results over PAAM indicating microbial signatures at different levels of phylogeny can be successfully associated with functional traits. Remarkable performance improvement with PAAM-ML was observed in the use case of DS1 (Table 5.5). Over DS2 and DS4, improvement in predictive performance was attained with relatively few numbers of features via using embedded feature selections over PAAM.

As an exploratory study, further, the use of CFS as a filter-based feature selector strategy is utilised over PAAM (in the upcoming section 5.5.4).

### 5.5.4 Classification of the microbiome using feature selection strategy of correlation-based filtering (CFS)

Another example of how the PAAM feature space can be used to determine features correlating better to metagenomic functions is explored in this section. CFS has been employed as a feature filter over the PAAM and OTU Count matrix and the results are discussed in this section as a benchmark. The obtained results are reported in Table 5.8.
Table 5.8 Summary of performance results obtained by application of RF, SVM over the number of features (NFS) obtained after CFS employed over PAAM and OTU abundance count of DS1, DS2 and DS4. Results are averaged over all folds and choosing the best value after resampling results across different tuning parameters.

<table>
<thead>
<tr>
<th>Input</th>
<th>PAAM</th>
<th>OTU ABUNDANCE MATRIX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>Feature Selection</td>
<td>Classifier</td>
</tr>
<tr>
<td>DS1</td>
<td>CFS</td>
<td>SVM</td>
</tr>
<tr>
<td></td>
<td>CFS</td>
<td>RF</td>
</tr>
<tr>
<td>DS2</td>
<td>CFS</td>
<td>SVM</td>
</tr>
<tr>
<td></td>
<td>CFS</td>
<td>RF</td>
</tr>
<tr>
<td>DS4</td>
<td>CFS</td>
<td>SVM</td>
</tr>
<tr>
<td></td>
<td>CFS</td>
<td>RF</td>
</tr>
</tbody>
</table>

The values reported in Table 5.8 are provided as an example of a performance benchmark when RF and SVM were employed over the CFS obtained features from different datasets of DS1, DS2 and DS4. Over DS1 and DS4, the performance of ML models seems to be relatively better over the CFS selected features from PAAM than their application over an OTU abundance count matrix (Table 5.8).

However, the embedded feature selection methods obtained a higher Accuracy and Kappa values in comparison to CFS (comparing results in Table 5.2-5.7 to Table 5.8). The performance of the ML classifier depends on features selected from the metagenomic data source. Embedded feature selection technique over PAAM (Table 5.2-5.7) gave improved effect to functional metagenomics in comparison to CFS filter-based selection (Table 5.8), thereby improving the performance of learning over functional metagenomes.
5.5.5 A comparison with other phylogeny-aware supervised ML Methods

Firstly, results obtained by applying the penalized regression-based MetaPhyl method (with default settings), to integrate a phylogeny tree and OTU abundances in metagenomic classification are contrasted with PAAM-ML. Table 5.9 displays the predictive power of metagenomic functions over all three data sources (DS1, DS2 and DS4) using each classification technique. These results are also presented in terms of Accuracy and Kappa values.

MetaPhyl [87] takes advantage of the similarities between only leaf-level OTUs, as encoded by the natural properties of a phylogenetic tree. On the other hand, PAAM-ML considers ancestral internal nodes as well in functional analysis of metagenomes, unlike MetaPhyl.

**TABLE 5.9 Summary of performance results obtained by application of PAAM-ML (represented as PAAM_featureselection_classifier) and MetaPhyl employed over DS1, DS2 and DS4.**

<table>
<thead>
<tr>
<th>Data Source (Validation)</th>
<th>Modelling (chosen combination at which best performance results were obtained in previous sections)</th>
<th>PAAM-ML</th>
<th>METAPHYL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Kappa</td>
<td>Default Settings</td>
</tr>
<tr>
<td>DS1 (LOOCV)</td>
<td>0.983</td>
<td>0.966</td>
<td>DS1</td>
</tr>
<tr>
<td></td>
<td>PAAM_XGBoost&amp;Pen-LR_SVM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DS2 (10-fold Cross-Validation)</td>
<td>0.975</td>
<td>0.957</td>
<td>DS2</td>
</tr>
<tr>
<td></td>
<td>PAAM_XGBoost_RF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DS4 (LOOCV)</td>
<td>0.993</td>
<td>0.989</td>
<td>DS4</td>
</tr>
<tr>
<td></td>
<td>PAAM_RFI_SVM</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
PAAM-based ML has shown better performance on the metagenomic data sources in comparison to the state-of-the-art phylogenetic method of MetaPhyl except in the case of DS2 as clearly indicated in Table 5.9.

PAAM-ML approach is further evaluated by following a benchmark comparison with another phylogeny-aware method based on the transformation of feature space using the PhILR method [133], [238]. The obtained results are indicated in Table 5.10.

**TABLE 5.10 Summary of performance results obtained by application of PAAM-ML (represented as PAAM_featureselection_classifier) and ML (RF/SVM) employed over PhILR transformed data obtained from DS1, DS2 and DS4.**

<table>
<thead>
<tr>
<th>Data Source (Validation)</th>
<th>Models</th>
<th>PAAM Feature Space</th>
<th>PhILR[133] Transformed Feature Space</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Modelling (Number of Features)</td>
<td>Accuracy</td>
</tr>
<tr>
<td>DS1 (LOOCV)</td>
<td>PAAM_XGBoost &amp; Pen-LR_SVM (40)</td>
<td>0.983</td>
<td>0.966</td>
</tr>
<tr>
<td>DS2 (10-fold Cross-Validation)</td>
<td>PAAM_XGBoost_RF (20)</td>
<td>0.975</td>
<td>0.957</td>
</tr>
<tr>
<td>DS4 (LOOCV)</td>
<td>PAAM_RFI_SVM (215)</td>
<td>0.993</td>
<td>0.989</td>
</tr>
</tbody>
</table>

Overall, the results (Table 5.10) indicate that PAAM-based ML produced relatively high-performance values in comparison to PhILR when integrating phylogeny and abundance counts as inputs to the prediction models in DS1, DS2, and DS4. Significant differences were observed between the classification models applied on PAAM and PhILR, in terms of their predictive performance ($p$-value < 0.05).
Furthermore, by performing ANOVA analysis, significant differences were observed between the phylogeny-aware classification models in terms of their predictive Accuracy and Kappa based on different datasets of DS1, DS2 and DS4 (ANOVA, \( p \)-value < 0.05). The post-hoc tests indicated a significant difference between PAAM-ML and PhILR (Tukey’s HSD test attained a value of 0.367 > threshold of 0.253) and; the marginal significant difference between PAAM-ML and MetaPhyl with Least Significant Difference (LSD) of 0.209 > 0.207.

The variation in predictive performance (Accuracy and Kappa) of different subgroups of phylogeny-aware models is shown in Figure 5.6. Figure 5.6 shows the effectiveness of PAAM-ML modelling over other phylogenetic models.

![Figure 5.6 An illustration of the comparison of variation in performance of three phylogeny-aware models.](image)

In summary of section 5.4, PAAM-based feature modelling indicates the potential to either improve or provide competent metagenomic functional
Experiments and results

predictions to modelling over the OTU abundances. Additionally, by performing the $t$-test, a significant difference was observed between PAAM-based ML modelling and OTU abundance matrix-based ML modelling over DS1 ($p$-value < 0.05).

It was also additionally observed that that combinatorial features in PAAM are better correlated with the outcome of interest (phenotype) and were included in the top-ranked features obtained from embedded selection strategies of RFI and XGBoost over the use case datasets of DS1, DS2 and DS4. The maximum correlation of XgBoost and Pen-LR-selected ancestral node features from PAAM was observed as 0.449, in comparison to the correlation of leaf-level features in OTU abundance count table (which was observed as 0.323) with the phenotype (Smoker or Non-Smoker) in DS1, and; over DS2, maximum correlation of internal nodes with phenotypes (body sites) was observed as (0.628) in comparison to leaf node OTUs (0.10); over DS3, maximum correlation of internal nodes with phenotypes (substrates) was observed as (0.618) in comparison to leaf node OTUs (0.466). A similar observation that internal nodes correlate better with phenotype was made with regards to the RFI feature selection. Since modelled ancestral features correlated better with the phenotypic functional class, this indicates their utility in functional analyses of metagenomes.

Embedded feature selection methods show their ability to be applied to PAAM. While the results discussed in this section, show the utility of PAAM in comparison to the OTU abundance count matrix adding credence to the proposed framework. However, it has been noted that it is not essential that PAAM-based selection demonstrates always a superior benchmark
performance in every use case (as shown in Table 5.8) discussed in this study, but an advantage lies in building links between microbiome evolution and abundances, enabling a novel evolutionary data structure and biological knowledge-driven predictions. A real advantage of PAAM-based feature space was indicated by the results reported in Table 5.2-5.7.

5.6 Discussions

To construct microbial features composed of taxa at different levels of taxonomic classification based on the biological assumption that close species to have similar effects on the phenotype, the proposed approach of PAAM-ML regularized abundances with branch lengths (distances) annotated on the phylogenetic tree as weights. Inclusion of branch lengths is potentially a better measure as fewer abundance taxa lineages might have undergone a significant evolutionary change (branch length is more) or vice versa. For this purpose, the designed data structure of PAAM in the process utilises both the tree topology as well as the evolutionary distances annotated on phylogenetic branches. The proposed method assumes that the phylogenetic tree information is known a priori, which is typically inferred from molecular sequences. Any arbitrary ordering of independent OTUs (leaf-level features), as a precursor to the application of ML models in functional metagenomics, does not exploit the biological relevance. To improve over this observation, identifying microbial features at different leaves of phylogenetic taxonomy is one of the central themes for the current study. Working along these lines, various feature selection strategies were explored by the framework. Embedded feature selection strategies proved to be efficient. Intriguing results were attained by
employing PAAM-ML over varied metagenomic environments. One of the contributory observations is the importance of ancestral nodes in the functional classification of metagenomes. Internal (ancestral) node features created as part of PAAM were better correlated with the phenotypic class. The results obtained in section 5.4 have shown that the ML over PAAM (PAAM-ML) (Table 5.2-5.7) performs better than other methods such as RF applied directly over the abundance profiles (Table 5.1) [69],[93]; and phylogeny-driven ways of MetaPhyl [87] (Table 5.9), and PhILR [133] (Table 5.10) employed in this chapter. The proposed approach also performed better than the MetaPhyl method in two of the data sources. The construction of PAAM is available as a utility to research community at a GitHub repository for any extensive analysis.

There are possible future extensions to this work by extending it to other multivariate analysis methods, such as principal component analysis (PCA), principal coordinate analysis (PCoA), and clustering of taxa based on PAAM to study microbial phylogenetic relationships or interactions further [37].

Some relevant groups of microbial species that played an important role in classification were also identified along with the proposed framework. In human microbiome (HMP) body sites (DS2), phylum: Firmicutes with Genus: Lactobacillus and Weissella; were noted as two of the top-ranked features to differentiate different body site niches. The study such as in [242] reported the existence of Lactobacillus in different body sites and its association with human health. In DS4, which relates to the soil microbiome, phylum Proteobacteria with genus Pseudomonas served as top-ranked features. The other predominant phylum of Actinobacteria served as important features in classifying soil metagenomes into sugar treated substrates. The findings in [243] reported the
involvement of *Proteobacteria* and *Actinobacterium* lineages of the soil microbiome in metabolic biochemical processes, confirming our results.

Unfortunately, with the DS1, taxonomical mapping was not found at the source of this dataset. Based on the above observations made along the generalized workflow shown in Figure 5.1, this sub-section introduces a simple generative alternative model (shown in Figure 5.7) for determining functional metagenomic predictions, integrating the phylogenetic context. It has the potential to organise functional predictions according to the natural structure of the underlying data.

![Graphical representation of the generative PAAM-ML prediction framework proposed as a useful tool in functional metagenomics.](image)

**Figure 5.7** Graphical representation of the generative PAAM-ML prediction framework proposed as a useful tool in functional metagenomics.
5.7 Summary

It is worth linking structural phylogeny of metagenomes with their functions. The inclusion of phylogenetic topology and the evolutionary history in microbiome structure (along Approach 2 as proposed in Chapter 3), has the potential to relate better to metagenomic functions and, giving better ML performance. Prioritizing features in hierarchal feature space based on ancestral of OTUs could further improve the predictive ability of phenotypic functions.

This chapter presented a novel comprehensive framework (PAAM-ML) of representative predictive ML models for inferring phenotypes in metagenomic studies with (1) the integration of phylogenetic evolution and abundance of metagenomes via construction of a novel data structure (PAAM); (2) an application of feature selection techniques; (3) use of predictive models (classifiers); and, (4) the assessment of their performance. An overview of the predictive classification performance over four data sources (DS1, DS2 and DS4) is provided along with the conclusive experiments and results. The novel application of integration of phylogeny and abundances of microbial taxa has been employed to engineer feature vectors at different levels of taxonomical hierarchy for the prediction metagenomic functions. This indicated an impact on performance values obtained by state-of-the-art classifiers. In this work, we implemented a novel 2D matrix PAAM, which combines level-by-level evolutionary information obtained from a phylogenetic tree with microbiome abundance profiles. However, PAAM has almost twice as many features as the OTU abundance count table, and hence the data has higher dimensionality. Therefore, applying feature selection approaches across the columns of PAAM obtained improved ML modelling in predicting metagenomic functions.
Along with the proposed framework PAAM-ML, embedded feature selection of RFI, XGBoost and Pen-LR, were implemented to rank or select informative features from the PAAM along with a benchmark of CFS (filter-based technique). The output of these as a rank ordering of significant predictors was worthy of further investigation. The different classifiers of RF and SVM were applied over the engineered features. The measures of Accuracy and Kappa were used to evaluate the classification performance the proposed methodology provided better or competitive performance in comparison to state-of-the-art ML models applied over the raw abundances. The method delivered significantly better performance in contrast to the state-of-the-art process of the phylogeny-aware model of PhILR [133].

The proposed approach improved appreciably over the MetaPhyl [87] method. The section on Experiments and Results shows that the phylogeny-aware modelling of features in microbiome studies could improve the classification performance. Furthermore, the likelihood of biologically relevant predictions increases with the incorporation of biological domain knowledge. It was observed that more than half of the essential features selected were not original OTUs, indicating the importance of hierarchical combinations formed by combining OTUs based on phylogeny (i.e., the internal nodes) in the proposed framework. A piece of evidence is indicated from PAAM-ML that the integrated quantitative profiles obtained from phylogeny and the OTUs are useful in classifying microbiome datasets. The construction of PAAM has provided a valuable facility to the researchers for applying predictive techniques to infer metagenomic functions.
However, there exist further possibilities to work towards the development of a novel classification method for 16S rRNA data by taking advantage of the natural structuring of microbial phylogeny in the functioning of ML classifier itself. In chapter 6, a study involving the prediction of phenotypes using a novel method of Phylogeny-aware RF model is presented. Figure 5.8 highlights the summarized features of the current study.
**Key Scientific Discovery**

- A phylogeny-driven framework (Figure 5.7) is proposed which is driven by the construction of a novel phylogeny-feature space ‘PAAM’ combining abundances and phylogenetic knowledge (from a phylogenetic tree) for classifying high-dimensional metagenomes into their functional phenotypes. The feature space contains both leaf-level OTUs as well as the ancestral internal nodes (taxa). The key to the formulation of this novel idea lies in an observation that the internal nodes (ancestors) could help in better classification of microbial genes into their functional phenotypes.
- The proposed approach considers dependencies between different microbial features and their relationship with the functional class.
- It was discovered that the functional analysis of metagenomes was benefitted from including the internal nodes in the analysis. Application of embedded feature selection methods of RFI and XGBoost over PAAM resulted in smaller feature space (dominated by the internal nodes) and supported better functional classification.
- As an example, the use of EM methods as feature selectors over PAAM produced relatively better classification performance. 98% Accuracy was obtained when SVM was applied over the 40 microbial features selected from PAAM of DS1 via XGBoost and Pen-LR methods in comparison to the application of the classical method of SVM over the OTU abundances of DS1 (which attained only 72 % accurate results). Similarly, over DS2, RF applied over only 20 features obtained from embedded feature selection from PAAM attained relatively better performance results with just 20 human microbial features (all of which were reported as internal nodes) differentiating body sites in comparison to the application of RF over the S831 OTU features. Over DS4, the competitive performance was attained with just 38 features from PAAM involving internal nodes in comparison to the application of classifiers over 1072 features in the OTU abundance count table.
- The most discriminative microbial features were backed by the better correlation of community features involving internal (ancestral nodes) to the functional phenotype.
- *phylum: Firmicutes* was selected as a predominant feature for distinguishing human body sites.
- *Proteobacteria* and *Actinobacterium* lineages of soil microbiome were noticed as top taxa for distinguishing between different glucose treatments of soil.

**Strengths**

- The framework proposed in this chapter is efficient in terms of predictive performance (Accuracy, Kappa) over the high-dimensional metagenomes and integrates the biological domain knowledge of phylogeny to classify functional metagenomes according to natural characteristics. It attained better performance when benchmarked with state-of-the-art abundance-driven and phylogeny-driven methods.
- ML classifiers are applied over the microbial signatures spanning the different levels of the taxonomy.

**Limitations**

- The PAAM feature space created is high-dimensional and much dependent on the application of feature selection strategies before the application of ML classifiers. The crux of the problem is coping with the complexity and high-dimensionality of PAAM.
- Currently, PAAM feature-space is utilised only in the functional classification of metagenomes; however in future, its usage in visualizing the microbiome, determining network or interaction or similarity between microbial features, etc. could be explored.
- The proposed approach does not consider the compositional nature of metagenomes.
- In future, integration of phylogeny at the ML classifier level instead of data engineering is plausible.

Figure 5.8 A summary of key observations, strengths and limitations of this study.
CHAPTER 6

A NEW PHYLOGENY-DRIVEN RANDOM FOREST CLASSIFIER FOR FUNCTIONAL METAGENOMICS

Classifying microbiome into its functional repertoire is an important task for metagenomic studies (functional metagenomics) and an active area of research, where the research community is trying to develop ML-based computational methods. RF is a well-known method for such supervised learning when applied over the abundance profiles of microbial taxa [88], [93], [98], [99], [127], [128], [244]. To further explore and make optimization in RF-based on the biological relationships between microbial features, a new classification method based on RF as guided by the evolutionary ancestry of microbial phylogeny is developed in this chapter. It is termed as “Phylogeny-RF.” This method facilitates to capture domain knowledge of phylogeny in an ML classifier itself. Closely related microbes by phylogeny are highly correlated and tend to have similar genetic and phenotypic traits [42], [83], [84], [132], [133]. Such microbes behave similarly; and hence tend to be selected together, or one of these could be dropped from the computational analysis, to reduce the redundancy and make ML better.
The proposed algorithm Phylogeny-RF works on the similar principle and is compared to state-of-the-art classification methods, including RF over the abundance-only profiles and the phylogeny distance-aware methods of MetaPhyl[87] and PhILR[133], using 3 real-world 16S rRNA metagenomic datasets (DS1, DS3, and DS4). Experimental results have shown that the RF can further be enhanced in terms of predictive performance by using phylogenetic modelling.

### 6.1 Introduction

Research in this area of functional metagenomics has been expanding, seeking to make sense of high-dimensional, heterogeneous, and complex 16S rRNA genes present in metagenomic datasets. One of the most widely used classifications approaches for the classifying genotype into phenotype is RF [88], [93], [98], [99], [127], [128], [244]. The progress of using RF for metagenomic data has been observed, including the prediction of functional roles associated with microbial features of OTUs or ASVs, and; the ranking or selection of OTUs or ASVs [69], [88], [98],[127], [128], [228]. The application of RF in studies [88], [93], [98], [99], [127], [128], [244] for predicting metagenomic functions assumed OTUs as independent features. However, OTUs or ASVs are related by their evolutionary taxonomic hierarchy of phylogeny [40]. A current challenge in metagenomic studies is to jointly model the phylogenetic effects with high-dimensional abundance profiles of microbial genes in the ML models. Very limited research [87] has been performed in modelling phylogeny in an ML classifier itself. To incorporate the underlying phylogenetic information among the features in supervised learning problems,
regularization methods have been recommended in classification models for functional metagenomics in [85],[87]. However, research in literature lacks in exploring and modelling phylogeny in decision tree-based models. In this work, for the first time, it has been demonstrated how the phylogenetic information could be used to guide and model the RF approach (and is termed as Phylogeny-RF), accounting for functional classification of metagenomes based on the biological domain knowledge. Importantly, the proposed approach addresses the modelling of phylogeny aware OTU features or taxa within the RF model, which could prove useful in its application to metagenomic datasets. This sets up the stage to identify biologically relevant OTUs that enable this good prediction via RF and; contain more diverse biological information. The authors in [88], proposed that the traditional RF model of classification progresses by using bagging strategy over multiple decision trees.

Each decision tree in RF chooses $m$ number of predictors to make decisions, and $m = \sqrt{\text{Number of features}}$ in classification problems [88],[245],[246],[247],[248]. Phylogeny-RF particularly regularizes $m$ to be guided by a phylogenetic measure assuming that phylogenetic similar OTU features tend to share the same functional responses [42]. The phylogenetic refinement learns the decision nodes of all trees under a global objective function so that microbiome information within multiple trees of RF is biologically diverse. The proposed new approach progresses by further clustering (grouping) original OTUs based on their phylogenetic similarity into $m$ groups and, thereafter, choosing an OTU element randomly from each group as a potential feature to be chosen as a decision node in a constituting decision tree of RF model.
The performance of an RF classifier is highly dependent on the accuracy of each component decision tree. In RF, randomization could cause the occurrence of redundant and correlated trees as this could include phylogenetically correlated features. This may lead to an inefficient ensemble classification decision. Better RF modelling could be achieved through the random selection of uncorrelated and diverse microbial features. The purpose of this research is to model RF over the phylogenetically diverse features to improve the quality of functional phenotypic predictions. As when two OTU features have high phylogenetic similarity, they will behave similarly; and henceforth, one of the two features could be chosen to minimise the redundancy and make learning over the more biologically relevant signatures in RF modelling [42], [81]. The key idea behind the proposed classification approach is to model RF over the OTU features with minimum redundancy and maximal biological relevance; to enhance its application in metagenomic studies. To summarise, this research aims to optimise over the high-dimensional feature set fed to RF in metagenomic studies by the selection of only biologically diverse and uncorrelated features to be modelled in constituting decision trees with attaining good classification performance.

### 6.2 Methodology and implementation

The proposed approach uses biological relatedness of constituting microbial taxa (OTUs) from the phylogenetic tree (structure) into the RF model to classify microbiome samples into phenotypes (functions) with regards to the three data sources DS1, DS3 and DS4 in the current chapter. The datasets are summarised in Table 6.1 for a quick reference. DS5 and DS6 have no associated
phylogenetic information available at the source. Over DS2, the approach seemed to be time-intensive. Henceforth, DS1, DS3 and DS4 are chosen in the current study.

TABLE 6.1 Summary of datasets used in the current chapter for functional analyses.

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Samples</th>
<th>OTUs</th>
<th>Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS1</td>
<td>60</td>
<td>856 (absolute abundances)</td>
<td>Smoking (28) vs. Non-smoking individuals (32)</td>
</tr>
<tr>
<td>DS3</td>
<td>622</td>
<td>2683 (relative abundances)</td>
<td>Body habitats of External Auditory Canal (EAC) (44), Gut (45), Hair (14), Nostril (46), Oral cavity (54), and Skin (419)</td>
</tr>
<tr>
<td>DS4</td>
<td>139</td>
<td>1072 (absolute abundances)</td>
<td>Substrates treatment of glucose (47) or cellulose (46) or controls (46)</td>
</tr>
</tbody>
</table>

The inclusion of tree structure would help in classifying microbes into phenotypic functions, according to the natural structure and properties of a microbial community. The stepwise procedure of the proposed method Phylogeny-RF is discussed as follows.

A. Inputs

A source phylogeny tree that is in the parenthetical Newick format was read and further processed using the read.tree() function in the ape R package.

An OTU abundance count table and meta-data are other inputs to the pipeline.

B. Generating Similarity Matrix Structure from Biological Domain Knowledge of Phylogeny

Important knowledge that can be extracted from a phylogenetic tree is a matrix depicting the phylogenetic distances between each pair of leaf nodes (OTUs). It is referred to as a Phylogenetic Distance Matrix (PDM) in the
current research, and it serves as background information for studying the phylogenetic diversity [181], [249], [250]. PDM has the OTUs as the row and column names, and the values in the corresponding matrix cells are the sum of the branch lengths separating each pair of OTUs on a phylogenetic tree. Thus, if two OTUs are close relatives, their intersection cell contains a smaller value in comparison to OTUs that are far apart on the phylogeny tree. The PDM has been generated using the cophenetic() function in the ape package in R [181], [249], [250].

C. Clustering OTUs based on PDM

Clustering plays a crucial role in bioinformatics to detect interesting patterns in genetic data based on the similarity of the constituting elements [251]. In our approach, clustering has been applied to the PDM obtained in the previous step (B) to group microbial OTUs further based on their shared phylogenetic variations. This application aims to facilitate further an RF model for classifying microbial samples by removing highly correlated OTUs, and for developing an RF model taking advantage of microbial ancestry. Intuitively, the optimal choice of the number of clusters in the proposed approach is chosen equivalent to the number of features to be chosen by RF modelling to strike a balance of selecting a microbial feature (OTU) from each cluster. A well-known k-means clustering approach [252] is chosen along the workflow of Phylogeny-RF as a precursor step to supervised learning. The parameter commonly referred as ‘k’ in k-means specifies the number of clusters to be generated; however, no such explicit choice of some clusters exists in another commonly used strategy of hierarchal clustering over distance-based matrices [253].
The points in dimension coordinates (along with the two principal directions) are derived from the similarity distance matrix (PDM) to be input to the $k$-means clustering, as the $k$-means algorithm [252], primarily deals with calculating Euclidean distances between data points in a cartesian coordinate space for finding groups of similar data features. This also aided in reducing the dimensions while dealing with the high-dimensional metagenomes. With this objective, points-in-dimensions coordinates were obtained from the similarity distance matrix of PDM by calculating eigenvectors via eigendecomposition as follows: PDM (input) → Covariance Matrix → Eigen-Decomposition → 2D Coordinate System [254]. Centering of PDM would make it a covariance matrix. The calculation of eigenvectors (corresponding to eigenvalues) over the covariance matrix is equivalent to fitting straight principal-component lines in a 2D coordinate system following the variance in data. The procedure of centering PDM is indicated in Algorithm 6.1a.

<table>
<thead>
<tr>
<th>Algorithm 6.1a Perform &quot;Centering&quot; of PDM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input.</strong> Phylogenetic Distance Matrix (PDM) of size (Number of data features $X$ Number of data features)</td>
</tr>
<tr>
<td><strong>Method.</strong></td>
</tr>
<tr>
<td>$n = \text{nrow} \left( \text{PDM} \right)$ where nrow () calculates the number of rows of PDM</td>
</tr>
<tr>
<td>$P = \text{diag}(n) - 1/n$ where diag () constructs a diagonal matrix</td>
</tr>
<tr>
<td>$PDM_c = -0.5 \times P \times PDM \times P$, where $\times$ represents matrix multiplication operator and $X$ represents multiplication operator.</td>
</tr>
<tr>
<td><strong>Output.</strong> $PDM_c$ (i.e., centred PDM)</td>
</tr>
</tbody>
</table>

Algorithm 6.1b has been used to transform the centred PDM data into a cartesian coordinate system with 2 dimensions (components) in the scope of this research by calculating eigenvectors corresponding to the eigenvalues obtained over the covariance matrix [254]. Eigenvalues are the coefficients
associated with eigenvectors and provide the measure of the data’s covariance. The two-component axes are determined by ranking eigenvectors in order of their eigenvalues (highest to lowest) (Algorithm 6.1b). The data points in this created system are further used for grouping based on \( k \)-means (Algorithm 6.2). \( k \)-means has been applied over the data points in obtained 2DCS-OUT (Algorithm 6.1b). The function of \textit{kmeans()} in the \textit{vegan} R package [180], has been used to obtain clustering. The general workflow of the \( k \)-means algorithm utilised in this chapter is shown in Algorithm 6.2.

<table>
<thead>
<tr>
<th>Algorithm 6.1b Construction of Cartesian Coordinates</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input.</strong> Centred Phylogenetic Distance Matrix (( PDM_c ))</td>
</tr>
<tr>
<td><strong>Method.</strong> Eigen = Eigenvalues of ( PDM_c ) are calculated by solving determinant ( (PDM_c - \lambda I) = 0 ); and finding value of ( \lambda ) where ‘I’ is an identity matrix [254]. ( \lambda ) serves as eigenvalues. twoDEValue = Eigen[1:2] # a set of first two eigenvalues is taken twoDEValue is a matrix of size 2X2.</td>
</tr>
<tr>
<td>Eigen Vectors = Two Eigenvectors are calculated for corresponding twoDEValue by following ( (PDM_c - \lambda I) * x = 0 ); where ( x ) values are considered as corresponding eigenvectors of ( \lambda ) eigenvalues. [254] 2DCS-OTU = EigenVectors * diag (( \sqrt{\text{twoDEValue}} )) where diag () constructs a diagonal matrix; and here * represents matrix multiplication operator. Eigen Vectors is a matrix of size \textit{Number of data features} \times 2.</td>
</tr>
<tr>
<td><strong>Output.</strong> 2DCS-OTU (i.e., a reduced two-dimensional coordinate system)</td>
</tr>
</tbody>
</table>
Obtained clusters are inputted to RF implementation in subsequent steps of the proposed method.

D. RF guided by Clustering Obtained in Step (C)

In a traditional RF classifier, a forest is constructed by building multiple decision trees [88]. The RF classifier selects a random subset of the predictors from the entire feature set to determine the best split markers based on a splitting criterion [88] to build a decision tree. These predictors are usually chosen as equivalent to $\sqrt{\text{Number of input features}}$ in the classification framework [88],[246]. However, it has been proposed in this chapter that the microbial features can be randomly chosen one from each cluster obtained in step (C), where the number of clusters $k$ is chosen equal to the $\sqrt{\text{Number of input features}}$. This would reduce phylogeny-based redundancy between OTU features while the construction of decision trees.
in RF. The rationale behind this is that the similarity and correlation patterns are present between the OTUs based on their phylogeny and similar OTUs tend to behave similarly to functions[42]. The Phylogeny-RF is shown in Algorithm 6.3.

Algorithm 6.3 A Stepwise Procedure for Phylogeny-RF (RF tuned with Phylogeny)

| **Input.** The input set consists of $N$ metagenomic samples; the dimension of the feature space is equal to the total number of OTU features ($F$) |
| Method. |
| 1. OTU features ($F$) obtained from the input are clustered into $m = \sqrt{F}$ groups using the $k$-Means algorithm (shown in Algorithm 6.2) [252] based on their phylogenetic similarity (PDM). |
| 2. A random sub-sample is generated from $N$ that serves as the training set (particularly in this research $k/f$-fold with $k/f = 5$ is implemented keeping $4$-fold over $N$ for training, and $1$-fold of data serves as the test set. |
| 3. Amongst the $m$ groups of clustered OTU features (obtained in step 1), the algorithm randomly chooses a predicting node from each clustered group to construct a decision tree. Hence, the number of predicting nodes in a decision tree would be equal to the number of clusters. It serves an alternative way to choose $\sqrt{F}$ predicting nodes in a decision tree based on the phylogenetic clusters. Chosen predictors intend to cover maximal phylogenetic diversity in this case of RF. A node is chosen from the predictors as a split point in a decision using the traditional split criteria of the Gini Impurity criterion [88]. |
| 4. Build forest by repeating steps 2 to 3 for “$n$” number times to create “$n$” number of trees.; with the application of bagging strategy [105]. |

The built forest is used to predict output class for the test set by calculating the votes for each predicted class by each constituting tree. The highly voted class is considered as the final prediction obtained from the RF algorithm

| **Output.** Phylogeny-based Phenotypic class predictions for test set samples. |

E. **Performance Evaluation**

The performance has been evaluated in this chapter using the average performance of popular assessment techniques of Accuracy, Kappa, and AUC-ROC over different phenotypic classes in DS1,3,4 [200]. Kappa and AUC-ROC serve as a logical extension to the interpretation of classification
performance by Accuracy [177], [226], especially in the case of imbalanced classes.

The framework based on the above steps (A-E) is summarised in Figure 6.1. Phylogeny-RF is implemented as a modification of a python script to code RF functionality [256]. The settings of RF parameters with maximum_depth = 6, num_folds = 5 (chosen 5-fold to uniformly evaluate all data sources with cross-validation settings) and num_trees in range of 1,10,20,40,64,100,128,164,200,225,300,500 are considered for experiments in the current research.
Figure 6.1 A graphical representation of the stepwise procedure followed in the construction of the proposed new Phylogeny-RF classifier for classifying microbiome into functional phenotype.
6.3 Experimental results and discussions

This section discusses the different sets of results obtained as part of this study.

6.3.1 Experimental results

The proposed approach is evaluated by following comparison of Phylogeny-RF, tradition RF [69], MetaPhyl [87], and RF applied over PhILR [133] transformed data. The chapter attempts to employ these methods to the same three reference datasets of DS1, DS3, and DS4 that are used for the benchmark comparison. Based on the three trial runs of 5-fold cross-validation, the average evaluation metrics of Accuracy, Kappa, and AUC-ROC of the models using RF-based classification were obtained over DS4 and DS1 (both having samples < 140). However, due to the time-intensive nature of the developed algorithm, only a trial run of 5-fold cross-validation was attempted over DS3 having 622 samples and 2683 OTUs. Table 6.2–6.4 depicts the results of the performance obtained by newly constructed Phylogeny-RF and classical RF over three microbiome data sources of DS4, DS1, and DS3, respectively.

Table 6.2 Summary of average performances obtained by application of Phylogeny-RF and RF with 5-fold (3 trial runs) cross-validation dealing with the different number of trees in RF over DS4 relating to soil microbiome (best values attained in each column are highlighted in bold in each category).

<table>
<thead>
<tr>
<th>Number of Trees</th>
<th>Phylogeny-RF</th>
<th>RF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Kappa</td>
</tr>
<tr>
<td>1 (Decision Tree)</td>
<td>0.768</td>
<td>0.666</td>
</tr>
<tr>
<td>10</td>
<td>0.787</td>
<td>0.684</td>
</tr>
<tr>
<td>20</td>
<td>0.871</td>
<td>0.812</td>
</tr>
</tbody>
</table>
### Table 6.3 Summary of average performances obtained by application of Phylogeny-RF and RF with 5-fold (3 trial runs) cross-validation dealing with a different number of trees in RF over DS1 relating to human throat microbiome (best values attained in each column are highlighted in bold in each category).

<table>
<thead>
<tr>
<th>Number of Trees</th>
<th>Phylogeny-RF</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>RF</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Kappa</td>
<td>AUC-ROC</td>
<td>Accuracy</td>
<td>Kappa</td>
<td>AUC-ROC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 (Decision Tree)</td>
<td>0.559</td>
<td>0.140</td>
<td>0.570</td>
<td>0.498</td>
<td>0.080</td>
<td>0.499</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.594</td>
<td>0.239</td>
<td>0.635</td>
<td>0.545</td>
<td>0.119</td>
<td>0.555</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.628</td>
<td>0.364</td>
<td>0.647</td>
<td>0.589</td>
<td>0.176</td>
<td>0.591</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>0.689</td>
<td>0.381</td>
<td>0.696</td>
<td>0.583</td>
<td>0.223</td>
<td>0.632</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>0.700</td>
<td>0.394</td>
<td>0.712</td>
<td>0.616</td>
<td>0.252</td>
<td>0.637</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.616</td>
<td>0.232</td>
<td>0.622</td>
<td>0.600</td>
<td>0.233</td>
<td>0.62</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>128</td>
<td>0.639</td>
<td>0.323</td>
<td>0.674</td>
<td>0.639</td>
<td>0.315</td>
<td>0.669</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>164</td>
<td>0.672</td>
<td>0.365</td>
<td>0.689</td>
<td><strong>0.667</strong></td>
<td><strong>0.317</strong></td>
<td><strong>0.678</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.672</td>
<td>0.396</td>
<td>0.711</td>
<td>0.589</td>
<td>0.243</td>
<td>0.643</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>225</td>
<td>0.661</td>
<td>0.344</td>
<td>0.694</td>
<td>0.633</td>
<td>0.268</td>
<td>0.638</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>300</td>
<td>0.694</td>
<td>0.370</td>
<td>0.696</td>
<td>0.628</td>
<td>0.302</td>
<td>0.672</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>0.655</td>
<td>0.326</td>
<td>0.671</td>
<td>0.561</td>
<td>0.229</td>
<td>0.627</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Experimental results and discussions

Table 6.4 Summary of average performances obtained by application of Phylogeny-RF and RF with 5-fold cross-validation dealing with a different number of trees in RF over DS3 relating to human body sites (Costello body habitats) microbiome (best values attained in each column are highlighted in bold in each category).

<table>
<thead>
<tr>
<th>Number of Trees</th>
<th>Accuracy</th>
<th>Kappa</th>
<th>AUC-ROC</th>
<th>Accuracy</th>
<th>Kappa</th>
<th>AUC-ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Decision Tree)</td>
<td>0.745</td>
<td>0.352</td>
<td>0.613</td>
<td>0.729</td>
<td>0.287</td>
<td>0.576</td>
</tr>
<tr>
<td>10</td>
<td>0.737</td>
<td>0.287</td>
<td>0.591</td>
<td>0.729</td>
<td>0.252</td>
<td>0.577</td>
</tr>
<tr>
<td>20</td>
<td>0.745</td>
<td>0.312</td>
<td>0.603</td>
<td>0.729</td>
<td>0.246</td>
<td>0.575</td>
</tr>
<tr>
<td>40</td>
<td>0.738</td>
<td>0.295</td>
<td>0.593</td>
<td>0.718</td>
<td>0.213</td>
<td>0.564</td>
</tr>
<tr>
<td>64</td>
<td>0.743</td>
<td>0.307</td>
<td>0.595</td>
<td>0.721</td>
<td>0.217</td>
<td>0.564</td>
</tr>
<tr>
<td>100</td>
<td>0.727</td>
<td>0.250</td>
<td>0.575</td>
<td>0.716</td>
<td>0.210</td>
<td>0.567</td>
</tr>
<tr>
<td>128</td>
<td><strong>0.753</strong></td>
<td><strong>0.348</strong></td>
<td><strong>0.609</strong></td>
<td>0.726</td>
<td>0.242</td>
<td>0.574</td>
</tr>
<tr>
<td>164</td>
<td>0.740</td>
<td>0.301</td>
<td>0.589</td>
<td>0.724</td>
<td>0.239</td>
<td>0.573</td>
</tr>
<tr>
<td>200</td>
<td>0.735</td>
<td>0.279</td>
<td>0.584</td>
<td>0.727</td>
<td>0.257</td>
<td>0.577</td>
</tr>
<tr>
<td>225</td>
<td>0.748</td>
<td>0.337</td>
<td>0.602</td>
<td><strong>0.735</strong></td>
<td><strong>0.280</strong></td>
<td><strong>0.585</strong></td>
</tr>
<tr>
<td>300</td>
<td>0.737</td>
<td>0.283</td>
<td>0.585</td>
<td>0.734</td>
<td>0.259</td>
<td>0.577</td>
</tr>
<tr>
<td>500</td>
<td>0.735</td>
<td>0.274</td>
<td>0.581</td>
<td>0.732</td>
<td>0.268</td>
<td>0.579</td>
</tr>
</tbody>
</table>

Tables 6.2–6.4 indicate that the performance of Phylogeny-RF concerning functional classification in terms of Accuracy, Kappa, and AUC-ROC, has achieved an overall improvement in comparison to the traditional RF[88] for identifying phenotypes. Phylogeny-RF over a different number of trees (ranging from 1-500) reported this improvement as significant ($p$-value < 0.05) with regards to predictive metrics (AUC-ROC and Kappa values) over the three sources DS4, DS1 and DS3 in comparison to the traditional RF method proposed by Breiman [88]. It has also been observed that the performance tends to follow a U-shaped curve, attaining maximum performance in the middle range of the number of trees (Table 6.2-6.4). However, in the case of Phylogeny-RF, better performance is being achieved in a smaller number of trees in
comparison to the traditional RF. For example, in the case of DS4 (Table 6.2), Phylogeny-RF attained the highest performance at 164 trees whilst 225 trees in traditional RF. Furthermore, the variation in AUC-ROC was analysed in the three trials of 5-fold cross-validation employed over the sources DS1 and DS4. The related Boxplot graphs are represented in Figure 6.2a,6.2b. The results demonstrated a comparatively less variation in the results of Phylogeny-RF than traditional RF[88] over DS1 and DS4 in the comprehensive analysis and hence indicated the effectiveness of proposed Phylogeny-RF.
Figure 6.2a A graphical representation with Box Plots indicating variation in AUC-ROC obtained over 3 trial runs of 5-fold cross-validation with (i) Phylogeny-RF over DS4 (ii) traditional RF over DS4.
Figure 6.2b A graphical representation with Box Plots indicating variation in AUC-ROC obtained over 3 trial runs of 5-fold cross-validation with (i) Phylogenetic RF over DS1 and (ii) traditional RF applied over DS1.
MetaPhyl [87] is a supervised classifier that involves regularization of LR by taking advantage of the natural characteristics as encoded in the phylogenetic tree. The next steps of analysis in current research is benchmarking of developed Phylogeny-RF method with the phylogeny-aware classifier of MetaPhyl (with its default settings recommended in [87]). The comparative results are reported in Table 6.5.

**Table 6.5 Summary of performance results obtained by application of Phylogeny-RF with 5-fold cross-validation and MetaPhyl with default settings over DS1, DS3, DS4.**

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Number of Trees at which best performance is attained</th>
<th>Accuracy</th>
<th>Kappa</th>
<th>AUC-ROC</th>
<th>Accuracy</th>
<th>Kappa</th>
<th>AUC-ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phylogeny-RF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DS1</td>
<td>64</td>
<td>0.700</td>
<td>0.394</td>
<td>0.712</td>
<td>0.733</td>
<td>0.471</td>
<td>0.739</td>
</tr>
<tr>
<td>DS3</td>
<td>128</td>
<td><strong>0.753</strong></td>
<td><strong>0.348</strong></td>
<td><strong>0.609</strong></td>
<td>0.553</td>
<td>0.132</td>
<td>0.557</td>
</tr>
<tr>
<td>DS4</td>
<td>164</td>
<td><strong>0.929</strong></td>
<td><strong>0.891</strong></td>
<td><strong>0.949</strong></td>
<td>0.755</td>
<td>0.633</td>
<td>0.816</td>
</tr>
<tr>
<td>MetaPhyl</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The newly developed classifier of Phylogeny-RF performed relatively better than MetaPhyl for classification performance metrics over DS3 and DS4 (as shown in Table 6.5). However, comparing Phylogeny-RF and MetaPhyl over all three data sources, reported marginal difference in the significance of results with ($p$-value = 0.042 < 0.05). The chapter progressed with comparing the traditional RF model applied to PhILR [133] transformed microbiota data (a phylogeny-based transform) for a further benchmark analysis. The newly constructed Phylogeny-RF was applied to abundance datasets and compared with RF applied over the PhILR-transformed data. The results (averaged over all 3 trial runs over all the 5-fold) are reported in Table 6.6-6.7 for sources DS4.
and DS1, respectively. Additionally, the results averaged over 5-fold cross-validation employed on DS3 are reported in Table 6.8. The Phylogeny-RF significantly improved classification performance ($p$-value $< 0.01$) in 2 of the 3 benchmarks (i.e., DS1, DS4) relative to the PhILR transform over a different number of trees (1 – 500 trees) (Table 6.6-6.7).

**Table 6.6 A comparison of average performance results obtained by application of Phylogeny-RF and classical RF applied over PhILR transformed data with 5-fold cross-validation (3 trials) on DS4 (soil microbiome) (best value in each column is bold-faced).**

<table>
<thead>
<tr>
<th>Number of Trees</th>
<th>Phylogeny-RF</th>
<th>RF applied over PhILR Transformed Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Kappa</td>
</tr>
<tr>
<td>1 (Decision Tree)</td>
<td>0.768</td>
<td>0.666</td>
</tr>
<tr>
<td>10</td>
<td>0.787</td>
<td>0.684</td>
</tr>
<tr>
<td>20</td>
<td>0.871</td>
<td>0.812</td>
</tr>
<tr>
<td>40</td>
<td>0.895</td>
<td>0.841</td>
</tr>
<tr>
<td>64</td>
<td>0.906</td>
<td>0.861</td>
</tr>
<tr>
<td>100</td>
<td>0.919</td>
<td>0.874</td>
</tr>
<tr>
<td>128</td>
<td>0.916</td>
<td>0.872</td>
</tr>
<tr>
<td>164</td>
<td><strong>0.929</strong></td>
<td><strong>0.891</strong></td>
</tr>
<tr>
<td>200</td>
<td>0.901</td>
<td>0.851</td>
</tr>
<tr>
<td>225</td>
<td>0.921</td>
<td>0.88</td>
</tr>
<tr>
<td>300</td>
<td>0.906</td>
<td>0.862</td>
</tr>
<tr>
<td>500</td>
<td>0.923</td>
<td>0.883</td>
</tr>
</tbody>
</table>
Table 6.7 A comparison of average performance results obtained by application of Phylogeny-RF and classical RF applied over PhILR transformed data with 5-fold cross-validation (3 trials) on DS1 (Throat microbiome) (best value in each column is bold-faced).

<table>
<thead>
<tr>
<th>Number of Trees</th>
<th>Phylogeny-RF</th>
<th>RF applied over PhILR Transformed Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Kappa</td>
</tr>
<tr>
<td>1 (Decision Tree)</td>
<td>0.559</td>
<td>0.14</td>
</tr>
<tr>
<td>10</td>
<td>0.594</td>
<td>0.239</td>
</tr>
<tr>
<td>20</td>
<td>0.628</td>
<td>0.364</td>
</tr>
<tr>
<td>40</td>
<td>0.689</td>
<td>0.381</td>
</tr>
<tr>
<td>64</td>
<td><strong>0.700</strong></td>
<td><strong>0.394</strong></td>
</tr>
<tr>
<td>100</td>
<td>0.616</td>
<td>0.232</td>
</tr>
<tr>
<td>128</td>
<td>0.639</td>
<td>0.323</td>
</tr>
<tr>
<td>164</td>
<td>0.672</td>
<td>0.365</td>
</tr>
<tr>
<td>200</td>
<td>0.672</td>
<td>0.396</td>
</tr>
<tr>
<td>225</td>
<td>0.661</td>
<td>0.344</td>
</tr>
<tr>
<td>300</td>
<td>0.694</td>
<td>0.37</td>
</tr>
<tr>
<td>500</td>
<td>0.655</td>
<td>0.326</td>
</tr>
</tbody>
</table>

However, an exception is the application of Phylogeny-RF over DS3, where RF applied over PhILR transformed data of DS3 performed significantly better over 5-fold cross-validation (Table 6.8). The advantage of PhILR [133] transform lies in handling the compositional nature of microbiome and it aids in the research where microbial features are measured by relative abundance counts and are related by a phylogenetic tree. Over relative abundances of DS3 dataset, PhILR [133] weighting scheme seems better matched to the analysis task of differentiating human body sites (Table 6.8).
Table 6.8 A comparison of average performance results obtained by application of Phylogeny-RF and classical RF applied over PhILR transformed data with 5-fold cross-validation on DS3 (Costello Body Habitats) (best value in each column is bold-faced).

<table>
<thead>
<tr>
<th>Number of Trees</th>
<th>Phylogeny-RF</th>
<th>RF applied over PhILR Transformed Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Kappa</td>
</tr>
<tr>
<td>1 (Decision Tree)</td>
<td>0.745</td>
<td>0.352</td>
</tr>
<tr>
<td>10</td>
<td>0.717</td>
<td>0.207</td>
</tr>
<tr>
<td>20</td>
<td>0.745</td>
<td>0.312</td>
</tr>
<tr>
<td>40</td>
<td>0.738</td>
<td>0.295</td>
</tr>
<tr>
<td>64</td>
<td>0.743</td>
<td>0.307</td>
</tr>
<tr>
<td>100</td>
<td>0.727</td>
<td>0.250</td>
</tr>
<tr>
<td>128</td>
<td><strong>0.753</strong></td>
<td><strong>0.348</strong></td>
</tr>
<tr>
<td>164</td>
<td>0.740</td>
<td>0.301</td>
</tr>
<tr>
<td>200</td>
<td>0.735</td>
<td>0.279</td>
</tr>
<tr>
<td>225</td>
<td>0.753</td>
<td>0.337</td>
</tr>
<tr>
<td>300</td>
<td>0.737</td>
<td>0.283</td>
</tr>
<tr>
<td>500</td>
<td>0.735</td>
<td>0.274</td>
</tr>
</tbody>
</table>

To summarize, results reported in Table 6.6-6.8 provide a direct comparison of Phylogeny-RF and state-of-the-art PhILR method, indicating the effectiveness of Phylogeny-RF over PhILR in two of the data sources.

A series of experiments were conducted on three high-dimensional data sources. For each dataset, it is concluded that the proposed Phylogeny-RF overall performed consistently better than traditional RF; performed better than PhILR in two of the data sources (Figure 6.3).
Experimental results and discussions

Figure 6.3 A comparison of Phylogeny-RF, RF, and RF applied over PhILR transformed data over different numbers of trees (1-500).

The significance of the differences between the three methods (Phylogeny-RF, traditional-RF, and RF applied over PhILR) in terms of their overall AUC-ROC (obtained over a different number of trees by running 5-fold cross-validation) is established by a one-way ANOVA and the post hoc analysis. AUC-ROC is chosen as it has its own advantages over Accuracy [177], [226],[257],[258]. By performing the ANOVA analysis (as described in Chapter 3), a significant difference between the predictive performance of these three methods (p < 0.05) was observed over three data sources DS1, DS3 and DS4. Using the LSD t-test, a significant difference was observed between Phylogeny-RF, traditional-RF (0.046> critical threshold of 0.039) and using Scheffe test significant differences were observed between Phylogeny-RF and RF applied over PhILR (0.211 > critical threshold of 0.115); on DS1 (TABLE 6.9).
Table 6.9 Results of ANOVA and post hoc analysis employed over AUC-ROC values obtained over a different number of trees in RF (applied over the abundance counts), Phylogeny-RF (PhyRF) and RF employed over PhILR (PHILRF) over DS1.

<table>
<thead>
<tr>
<th>DS1</th>
<th>ANOVA ANALYSIS</th>
<th>α</th>
<th>LSD</th>
<th>HSD</th>
<th>Scheffe</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SUMMARY</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Groups</td>
<td>Count</td>
<td>Sum</td>
<td>Average</td>
<td>Variance</td>
</tr>
<tr>
<td>PhyRF</td>
<td>12</td>
<td>8.017</td>
<td>0.668</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>RF</td>
<td>12</td>
<td>7.461</td>
<td>0.622</td>
<td>1.002</td>
<td></td>
</tr>
<tr>
<td>PhiILRF</td>
<td>12</td>
<td>5.488</td>
<td>0.457</td>
<td>0.002</td>
<td></td>
</tr>
</tbody>
</table>

Over DS3, Using the Tuckey’s HSD, a marginally significant difference was observed between Phylogeny-RF, traditional-RF (0.016> critical threshold of 0.012), and; using Scheffe test RF applied over PhILR proved significantly better than Phylogeny-RF in this use case (TABLE 6.10). Over DS4, using the Tuckey’s HSD, a highly significant difference was observed between Phylogeny-RF, traditional-RF (0.087> critical threshold of 0.045). Scheffe test indicated that Phylogeny-RF performed significantly better than RF applied over PhILR transform (Table 6.11).
### Table 6.10 Results of ANOVA and post hoc analysis employed over AUC-ROC values obtained over the different number of trees in RF (applied over the abundance counts), Phylogeny-RF (PhyRF) and RF employed over PhILR (PHILRF) over DS3.

<table>
<thead>
<tr>
<th>Groups</th>
<th>Count</th>
<th>Sum</th>
<th>Average</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>PhyRF</td>
<td>12</td>
<td>7.09</td>
<td>0.591</td>
<td>0.0002</td>
</tr>
<tr>
<td>RF</td>
<td>12</td>
<td>6.88</td>
<td>0.574</td>
<td>3.89E-05</td>
</tr>
<tr>
<td>PhILRF</td>
<td>12</td>
<td>8.04</td>
<td>0.670</td>
<td>0.000158</td>
</tr>
</tbody>
</table>

**ANOVA**

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>P-Value</th>
<th>F crit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>0.0631</td>
<td>2</td>
<td>0.0315</td>
<td>226.7</td>
<td>0.000</td>
<td>3.285</td>
</tr>
<tr>
<td>Within Groups</td>
<td>0.0045</td>
<td>33</td>
<td>0.0001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>0.067782</td>
<td>35</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Reject Null Hypothesis because p < 0.05 (Means are different)

Coloured cells have significant mean differences

### Table 6.11 Results of ANOVA and post hoc analysis employed over AUC-ROC values obtained over the different number of trees in RF (applied over the abundance counts), Phylogeny-RF (PhyRF) and RF employed over PhILR (PHILRF) over DS4.

<table>
<thead>
<tr>
<th>Groups</th>
<th>Count</th>
<th>Sum</th>
<th>Average</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>PhyRF</td>
<td>12</td>
<td>11.045</td>
<td>0.920</td>
<td>0.00158</td>
</tr>
<tr>
<td>RF</td>
<td>12</td>
<td>9.999</td>
<td>0.833</td>
<td>3.14E-03</td>
</tr>
<tr>
<td>PhILRF</td>
<td>12</td>
<td>8.004</td>
<td>0.667</td>
<td>0.00127</td>
</tr>
</tbody>
</table>

**ANOVA**

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>P-Value</th>
<th>F crit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>0.398</td>
<td>2</td>
<td>0.198</td>
<td>99.60</td>
<td>0.000</td>
<td>3.285</td>
</tr>
<tr>
<td>Within Groups</td>
<td>0.066</td>
<td>33</td>
<td>0.001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>0.067782</td>
<td>35</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Reject Null Hypothesis because p < 0.05 (Means are different)

Coloured cells have significant mean differences

Overall, the analysis in this section shows the effectiveness of a newly developed method of the Phylogeny-RF method.
6.3.2 Discussions

There exists an association between phylogenetic understanding of microbial evolution and the host of microbial communities [86],[259]. The current research continued constructing a computational model for linking microbiome evolution with a host by designing a tree-based ML model that uses phylogenetic distances to tune feature modelling within the popular RF classifier enabling a novel evolutionary supervised analysis. In this chapter, the construction of an evolutionary-driven model over human and soil microbiomes with different phenotypes is undertaken. The preceding section has shown that the proposed Phylogeny-RF performs better than traditional RF over the three metagenomic data sources. The Accuracy of traditional RF is improved by integrating phylogenetic knowledge by minimizing the phylogenetic correlation between microbial features of trees in the RF and maximizing the predictive ability. A decision tree-based modelling of Phylogeny-RF seemed to perform better (in terms of AUC-ROC, Kappa) than MetaPhyl, which regularized the LR model with a phylogenetic measure driven penalty to present an optimization [87], over two of the data sources. This indicates the potential of integrating phylogeny in decision tree-based methods in addition to regularization methods, as suggested in [85], [87].

The RF applied over PhILR transformed data performed better than Phylogeny-RF over DS3. However, Phylogeny-RF performed significantly better than RF applied over PhILR over DS1 and DS4 (p-value < 0.05). Phylogeny-RF performed better than RF applied over the abundance profiles over DS3,4 and provided competitive performance over DS1.
Nonetheless, what distinguishes the Phylogeny-RF from classical RF applied over the abundances, is the interpretability of microbial features (attributes) corresponding to phylogenetic similarity and reducing the redundancy between microbial features while modelling, which can be a source for biological insight. Through the methods presented in this chapter, a coherent framework is designed for performing a functional analysis of microbiome gaining insights from phylogenetic artefacts.

The clusters of microbial features obtained as a pre-cursor to Phylogeny-RF could also be used for further ML analysis. Specifically, some limitations in current research exist relating to the choice of a phylogenetic distance, the clustering scheme used, and the modelling parameters used for Phylogeny-RF. However, these could be viewed as preliminary heuristics in the current research. Additionally, if the Phylogeny-RF must have meaningful features participating in making class decisions, that should not complicate the interpretation of the traditional RF model. In terms of clustering scheme utilized as a precursor to Phylogeny-RF, current research further worked towards calculating the clustering quality. The chapter proposed the use of $k$-means as a precursor to RF as it helps in achieving good clustering quality and subsequently aids RF in attaining good quality predictions. To investigate and validate the cluster quality obtained by $k$-means more objectively, a benchmark was carried out by generating clusters over PDM with two other popular clustering techniques of Partition Around Medoids (PAM) [253],[260] and hierarchical clustering (hclust) [261] by using validation index of silhouette coefficient [262]. The value of the silhouette coefficient varies from -1 to 1, and; higher value represents better clustering quality [262]. Experiments performed on three
datasets (DS1, DS3, and DS4) in the current study verified the effectiveness of $k$-means in the proposed computational method. Experiments are carried out with the help of the cluster [263] and factoextra [264] R packages. The results of cluster validation indicated higher value of silhouette coefficient (> 0.60) with regards to $k$-means in the scope of the current study, in comparison to the other methods of PAM [253], [260] and hclust [261].

Nonetheless, the proposed implementation of the Phylogeny-RF scheme is time intensive over the high-dimensional metagenomes with many samples (such as samples > 500). In the future, efforts could be made in the direction of making it more scalable and efficient in terms of computational cost.

This exploratory study found phylogenetic diversity of metagenomes to be useful in each functional classification task capturing the relevant functional adaptations of microbial communities. For this reason, phylogenetic relationships of the input 16S rRNA microbial genes was used to regularize the popular state-of-the-art classifier RF. To meet the needs of the proposed new classifier in the current study, OTUs (microbial features) were clustered based on their phylogenetic similarity as observed in PDM which reports the pair-wise phylogenetic distances between OTUs.

PDM serves as a simplified similarity model in which OTUs that are phylogenetically close have been grouped into a cluster. RF model selecting microbial features (i.e. OTUs) from different clusters provided the phylogenetic diverse features. There are alternatives in literature for measuring the phylogenetic diversity with UniFrac [50] and Phylogenetic Interaction-Adjusted index (PINA) [91] similarity models. However, these alternatives
calculate *sample-sample* similarity based on the microbial phylogenetic diversity, however, the current study focused on calculating *OTU-OTU* similarity based on their phylogenetic diversity. Henceforth, the current study involved the usage of PDM as part of the analysis. Nonetheless, an attempt is made in Chapter 7 to utilise UniFrac and PINA over a use case of cattle microbiome for ranking of microbial features in a different approach (according to their phylogenetic diversity) and this plays an important role in classifying microbial samples which achieves one of the objectives of the MetaPlat project [144].

In the current approach, the phylogenetic similarity between only the leaf-level OTUs was considered to regularize RF; however, in future work different ways could also be explored to incorporate phylogenetic similarity between internal (ancestral nodes of OTUs) as well in addition to the leaf nodes in the functional analysis. This would pave the way towards combining Approach 2 (Figure 3.5b) and Approach 3 (Figure 3.5c).

### 6.4 Summary

Prediction of metagenomic functions is an important task as the microbiome and its related phylogeny could play a vital role in inferring biological systems [52]. A new phylogeny-driven classifier is proposed in this research for determining metagenomic functions with balanced datasets of DS 1 and DS 4 and; an imbalanced dataset of DS3. The new classifier is evaluated by comparing the performance with three metrics: overall Accuracy, AUC-ROC, and the Kappa coefficient. Experimental results demonstrate a promising performance with the preliminary investigation of Phylogeny-RF into the
prediction of the microbiome associated with different functions in human and soil samples (over three data sources). It has been demonstrated that the phylogeny distance-driven ML techniques can successfully be applied to the problem area of functional metagenomics and has the potential to outperform state-of-the-art. Furthermore, the newly constructed classification method Phylogeny-RF produced high AUC-ROC values in comparison to classical RF; and obtained higher AUC-ROC values in comparison to other phylogeny-aware methods over two of the data sources. The construction of clusters based on phylogeny provided a useful facility to researchers when applying any kind of analysis in the future to infer metagenomic functions.

This chapter presented an initial analysis of the functional prediction of metagenomes when employing the Phylogeny-RF approach. The approach is based on grouping phylogenetically close features and select one from each group to be used as a potential feature in RF subsequently. For clustering as a precursor step to the group and select features for RF, this framework finds the use of $k$-means as a suitable and popular algorithm. In future work, the regularization of tuning RF with more sets of parameters targeting optimum performance could be explored. The current approach poses limitations as it is time intensive for many samples. It is useful to extend upon the analysis by using the phylogeny-aware context in a more scalable manner or designing a scalable method. Furthermore, this initial analysis could be extended to the utilization of other different distance metrics [249] to explore the area further. Figure 6.4 summarizes the key observations made in the current study.
Key Scientific Discovery

• A new classifier Phylogeny-RF (Figure 6.1) is proposed which is driven by the observation from the literature that closely related microbes by phylogeny are highly correlated and tend to have similar genetic and phenotypic traits. The proposed approach considers phylogenetic dependencies (similarity) between microbial OTUs. It was discovered that the functional analysis of metagenomes with RF benefitted from including the phylogenetic similarity.

• RF proposed by Breiman [88] selects randomly the subset of features ($\sqrt{\text{Number of features}}$) from the original feature space for classification. However, the newly proposed approach in this study (i.e. Phylogeny-RF) also selects randomly the $\sqrt{\text{Number of features}}$ but with an intention to select phylogenetically diverse features as closely related microbes by phylogeny tend to have similar genetic and phenotypic traits.

• The similarity between microbial features (OTUs) is calculated from a phylogenetic tree in the form of a matrix (PDM) depicting the phylogenetic distances between each pair of leaf nodes (OTUs) which is clustered to find similar microbial features by phylogeny and selecting only one amongst these similar features (supporting Minimum redundancy and Maximum relevance principle of involved microbial features). However, an interesting point to observe here is phylogeny has been used to regularize RF model which is applied to the abundance count table which still supports variety in trees as per the quantitative profiles of microbial features.

• Phylogeny-RF achieves better performance in lesser number of constituent trees in comparison to the RF.

Strengths

• Attempts to integrate phylogeny in a tree-based classifier RF for the first time. In literature to date, only LR has been explored to tune its penalty with biological knowledge of phylogeny.

• The new classifier Phylogeny-RF attained significantly better performance over metagenomic use cases when benchmarked with state-of-the-art RF model and other phylogeny-driven methods of MetaPhyl and PhiLR.

Limitations

• The approach is time intensive over large datasets.

• The proposed approach does not consider the compositional nature of metagenomes.

• The approach considers only leaf-level OTUs in predictive modelling and not the internal nodes; hence in future, it might be developed further to include ancestral nodes in feature space in a scalable manner.

Figure 6.4 A summary of key observations, strengths and limitations of this study.
CHAPTER 7

PREDICTION OF METAGENOMIC FUNCTIONS IN CATTLE USING THE ABUNDANCE-DRIVEN AND PHYLOGENY-DRIVEN ANALYSIS

This chapter aims to study the functions of cattle (Bos Taurus) rumen microbial community through ML-based computational methods. It presents a use case whereby abundance-driven (based on the quantitative measure) and integrative phylogeny-driven (based on a qualitative measure) models are employed to infer the cattle functions form its microbial composition. The study illustrates the use of microbial features being modelled at different phylogenetic levels related to sampled cattle microbiome. An empirical analysis of the ML approaches is provided when investigating the effect of the structure of cattle microbiome upon the predictive performance of cattle functions, following the three Approaches (proposed as Approach 1, Approach 2, and Approach 3 in Chapter 3, Figure 3.5a-c). The current chapter considers multiple methods following these Approaches to the application or development of computational models for analysing cattle microbiome. A variety of developed methods are discussed in this chapter of the thesis as an applicative study of the MetaPlat project¹ [144] to attain its objectives.
7.1 Introduction

Various metagenomic research studies have investigated whether the microbial community present in cattle rumen is influenced by functions of diet, feeding behaviour, stress-control, or geography [58], [72], [73], [75], [77], [265]. Studying rumen microbiome is the hot topic of cattle nutrition research [70]. The core objectives of MetaPlat[144] are defined to study diet-related effects in cattle rumen with the development of accurate classification algorithms to support microbiome analysis.

In the current study, knowledge extracted from the quantitative and qualitative measures has been utilised to study cattle microbiome and their functions. Quantitative profiles are analysed using Approach 1 (an abundance-aware analysis). Qualitative knowledge in terms of biological domain knowledge (obtained from phylogeny) has been utilised to model relationships between various microbiome features (an application of Approach 2 and 3). Understanding the topological evolution of microbial community composition is crucial to provide knowledge on the functions of each metagenome and its association with the host genetics. The metagenomic studies over diverse metagenomic environments in this thesis (as conducted in Chapters 5 and 6) indicated an improvement in predictive ability with the integration of phylogenetic relationships. Henceforth, the current study aims to extend analysis based on the phylogenetic measure (phylogeny-aware analysis). Such phylogenetic knowledge is incorporated either at (i) data-modelling level (ii) feature-selection level and (iii) ML-modelling level; in the current use case related to cattle microbiome. However, no such phylogenetic dependencies are
A use case: Determine functional phenotype of the cattle rumen microbiome

drawn between the features in abundance-only driven approaches. ML approaches employed directly over microbial abundances assume feature independence.

In this chapter, computational models are employed to a case study related to the analysis of cattle rumen microbiome as part of the MetaPlat project\textsuperscript{1} [144], and their predictive performance is discussed along with the biological insights. To study the association of cattle microbiome with its environmental phenotypes, the use of abundance-driven and phylogeny-driven models are useful in such exploratory analysis. In the current analysis, the effect of diet and cortisol treatment over cattle rumen microbiome profiles is studied. Cattle diet plays a contributory role in controlling feed efficiency and which may further be linked with methane (CH\textsubscript{4}) production in ruminants [71].

7.2 A use case: Determine functional phenotype of the cattle rumen microbiome

Previous work such as in [58], [72], [73], [75], [77], [265] have been conducted to detail and predict diverse cattle functions. The authors in [76] recently reviewed the application of metagenomics to cattle functions comprehensively. Furthermore, the research conducted in [75], [78], [121], [266], [267] discussed the use of supplement-controlled feeding of cattle, such as with nitrate or oil treated diet, and; studied their effect on cattle rumen metabolism. MetaPlat project [144] aimed to serve a mixture of innovative research focusing on:- sample collection from cattle rumen, for sequencing and analysis; the development of accurate and/or new classification methods using ML algorithms; and production of statistical or visual representations, conveying
more useful information regarding the structure and function of cattle microbiome. The detailed description of the current use case of MetaPlat project [144]; different ML models employed to analyse such data, and the predictive performance of the employed ML approaches are detailed in the following subsections.

### 7.2.1 MetaPlat project

The MetaPlat project aimed to develop a metagenomic platform for agricultural sciences that could handle cattle rumen metagenomics data sequences well and could produce in-depth analysis to facilitate researchers in their understanding of data generated from cattle rumen samples. The project involved the collection of cattle gut microbiota measuring of diet supplements feed conversion efficiency and methane emissions at Scotland’s Rural College (SRUC) Beef and Sheep Research Centre⁴ in Edinburgh, per the requirements of the UK Animals (Scientific Procedures) Act, 1986 [268], and; under the approval of SRUC’s Animal Welfare and Ethical Review Body (AWERB) and the Animal Experiments Committee. Illumina TruSeq libraries were prepared from cattle rumen genomic DNA and sequenced on an Illumina HiSeq 2500 instrument by Edinburgh Genomics⁵ [269]. Paired end reads (2 × 100 bp) were generated, resulting in between 8.6 and 14.5 GB per sample (between 43.4 and 72.7 million paired reads). The genomic reads were aligned to the KEGG genes database [270]. All the raw data can be found under accession PRJEB10338 (http://www.ebi.ac.uk/ena/data/search?query=PRJEB10338). QIIME [20],

---

⁴ [https://www.sruc.ac.uk/info/120194/beef_and_sheep_research_centre](https://www.sruc.ac.uk/info/120194/beef_and_sheep_research_centre)

⁵ [https://www.ed.ac.uk/roslin/facilities-resources/ark-genomics](https://www.ed.ac.uk/roslin/facilities-resources/ark-genomics)
QIIME2 [25] were employed over the cattle microbiome 16 S rRNA sequences to obtain OTU or ASV tables and the related phylogenetic trees. The metagenomics dataset contained abundances of cattle microbial genes with metadata related to the different samples included such as feed (diet), and cortisol treatment along with a related phylogenetic tree. These data are utilised in the abundance or phylogeny-driven data models in this Chapter to determine the functional repertoire of cattle microbial genes.

### 7.2.2 MetaPlat data sources

*Bos taurus* rumen microbiota samples were sequenced for MetaPlat project obtaining new datasets. The following three main datasets were generated and utilised in current research.

(i) A dataset with 40 rumen samples was obtained from Aberdeen Angus (AAx) and Limousine (LIMx) cattle breeds and treated with two diets, an oil-based and a nitrate-based feeding.

(ii) The MetaPlat experiments were extended to 80 animals. 80 cattle rumen samples were categorized into four different diet supplement categories: - an Oil-based supplemented diet, a Nitrate supplemented diet, a combined dietary intake (Oil-Nitrate), and Controls. These dietary supplement strategies of nitrate addition are expected to reduce emissions of important greenhouse gas – methane (CH4) [271]. The dataset consisted of 20 samples from each of the dietary treatments and 20 from controls. The control diet contained rapeseed meal as the main protein source which was replaced either with nitrate (21.5 g nitrate/kg DM) or oil distillers' dark grains (37 g
A use case: Determine functional phenotype of the cattle rumen microbiome lipid/kg diet DM) to increase diet lipid concentration or both nitrate and oil distillers’ dark grains. The corresponding phylogenetic tree was generated.

(iii) MetaPlat was later extended further to analyse the effect on cattle productivity that may be mediated by the circulation of exogenous Dexamethasone (DEX), synthetic cortisol to probe in a controlled manner on the cattle microbial environment. A total of 118 crossbred Limousine steers were used in this dataset, generating 678 ASVs and were mapped to the metadata of DEX treatment or controls. These data were further allocated to either a concentrate-based diet (forage: concentrate ratio of 8:92 dry matter basis) or a forage diet (50:50 dry matter basis). The steers had an adaptation period of 4 weeks to habituate to the diet and cortisol treatments. The corresponding phylogenetic tree was also generated.

The datasets are summarised in Table 7.1 highlighting the number of samples, OTUs/ASVs and the phenotype.

**Table 7.1 Summary of datasets sampled as part of MetaPlat project and used in current analyses.**

<table>
<thead>
<tr>
<th>S.No.</th>
<th>Samples</th>
<th>OTUs or ASVs</th>
<th>Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>40</td>
<td>386 OTUs at the <em>genus</em> level of study</td>
<td>Nitrate (20) or Oil (20)-treated Diet</td>
</tr>
<tr>
<td>M2</td>
<td>80</td>
<td>727 ASVs at the <em>species</em> level of study</td>
<td>Controls (20), Nitrate (20), Oil-treated (20) or combined with both oil &amp; Nitrate-Diet (20)</td>
</tr>
<tr>
<td>M3</td>
<td>118</td>
<td>679 ASVs at the <em>species</em> level of study</td>
<td>DEX treatment (70) and controls Diet with forage (58) or concentrated (60) categories</td>
</tr>
</tbody>
</table>
7.2.3 Integrative analysis of MetaPlat data using proposed methods of PAAM-ML and Phylogeny-RF

An empirical analysis with the use of integrative computational ML methods to infer phenotypes related to cattle rumen (sampled from MetaPlat with related phylogenetic trees) is presented in this section. The analysis is based on the methods proposed in Chapter 5 (PAAM-ML) and Chapter 6 (Phylogeny-RF), which demonstrated that the quantitative profiles of microbial features built upon phylogeny and abundances are informative for predicting the functional traits.

7.2.3.1 Analysis based on the application of PAAM-ML to differentiate supplemented diet phenotype

The work presented in this section is based on published work in [266]. PAAM was generated for cattle rumen data (indicated as M2 in Table 7.1) using the methods described in Chapter 5 of this thesis. The following ML methods were trained over PAAM with LOOCV.

i. The feature engineering technique of ranking taxonomic features using RF importance (RFI) was applied over the ASV input, using R package `randomForest` [248]. The top 10, 30, 50 % of features were derived with RFI modelling.

ii. Pen-LR, XGBoost, and RF models were applied over RFI selected features from PAAM to differentiate constituting samples into the functional phenotype of diet. The models were implemented with the help of the `caret` package in R [234]. The models are estimated based on
Accuracy and Kappa, being default metrics in *the caret* package in R [234].

The models described above have also been employed over the RFI selected features from abundance count table of ASVs (instead of PAAM), for benchmark comparison. Furthermore, RF and Pen-LR have been applied over the PhILR[133] transformed data. The phylogenetic method of MetaPhyl [87] was also employed over the cattle microbiome data for further comparative analysis.

**Table 7.2 Predictive performance of ML models employed over PAAM and ASV abundance profiles of cattle rumen dataset using LOOCV settings.** The dataset was characterised into 4 functional categories of diet (oil, nitrate, oil-nitrate combined, and controls). The best performance in each column is bold-faced.

<table>
<thead>
<tr>
<th>Ensemble ML Models (with LOOCV)</th>
<th>PAAM</th>
<th>ASV Abundance Count Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pen-LR employed over Top 10 % features as ranked by RFI</td>
<td>0.887</td>
<td>0.875 0.833</td>
</tr>
<tr>
<td>XGBoost employed over Top 10 % features as ranked by RFI</td>
<td>0.912</td>
<td>0.862 0.81</td>
</tr>
<tr>
<td>RF employed over Top 10 % features as ranked by RFI</td>
<td>0.900</td>
<td>0.850 0.800</td>
</tr>
<tr>
<td>Pen-LR employed over Top 30 % features as ranked by RFI</td>
<td>0.937</td>
<td>0.917 0.816</td>
</tr>
<tr>
<td>XGBoost employed over Top 30 % features as ranked by RFI</td>
<td>0.900</td>
<td>0.837 0.783</td>
</tr>
<tr>
<td>RF employed over Top 30 % features as ranked by RFI</td>
<td>0.875</td>
<td>0.825 0.766</td>
</tr>
<tr>
<td>Pen-LR employed over Top 50 % features as ranked by RFI</td>
<td>0.900</td>
<td>0.862 0.816</td>
</tr>
<tr>
<td>XGBoost employed over Top 50 % features as ranked by RFI</td>
<td>0.912</td>
<td>0.837 0.783</td>
</tr>
<tr>
<td>RF employed over Top 50 % features as ranked by RFI</td>
<td>0.862</td>
<td>0.788 0.717</td>
</tr>
</tbody>
</table>

Estimation of predictive performance has been evaluated to infer diet functions related to the cattle rumen in this study and is indicated in Table 7.2.
The final parametric values used by the Pen-LR model attaining best predictive value with an Accuracy of 0.937 and Kappa of 0.917 (Table 7.2) were chosen as cost = 1, loss = L1(LASSO) and epsilon = 0.1 in its implementation with caret package [237]. It was noticed that out of 219 features in Top 30 % of PAAM built data, 150 were ancestral nodes. Overall, it is observed that modelling with RF, XGBoost, and Pen-LR over PAAM produced relatively better results (in terms of Accuracy and Kappa) than their application over the abundance count profiles. The percentages of internal nodes were found relatively more in comparison to leaf-level OTUs whilst modelled by RFI in the current study.

A summary of the few of the topmost important taxa characterised by the application of RFI over high-dimensional PAAM in the current study is presented in Table 7.3 identified at different levels of taxonomy till genus. Studies in [272],[273] have indicated the role of Lachnospiraceae, Ruminococcaceae, Bifidobacteriacea, Veillonellaceae families in studying nutritional outlook in cattle. The phyla of Firmicutes, Actinobacteria, and Proteobacteria have been identified as key features to differentiate cattle rumen into diet phenotype.
A use case: Determine functional phenotype of the cattle rumen microbiome

Table 7.3 Summary of predominant features reported by RFI method in current analyses.

<table>
<thead>
<tr>
<th>Kingdom</th>
<th>Phylum</th>
<th>Class</th>
<th>Order</th>
<th>Family</th>
<th>Genus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bacteria</td>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Clostridiales</td>
<td>Veillonellaceae</td>
<td>Veillonella</td>
</tr>
<tr>
<td>Bacteria</td>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Clostridiales</td>
<td>Lachnospiraceae</td>
<td>Oribacterium</td>
</tr>
<tr>
<td>Bacteria</td>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Clostridiales</td>
<td>Ruminococcaecae</td>
<td>Ruminococcus</td>
</tr>
<tr>
<td>Bacteria</td>
<td>Actinobacteria</td>
<td>Actinobacteriida</td>
<td>Actinomycetales</td>
<td>Bifidobacteriaceae</td>
<td>Bifidobacterium</td>
</tr>
<tr>
<td>Bacteria</td>
<td>Actinobacteria</td>
<td>Actinobacteriida</td>
<td>Bifidobacteriaceae</td>
<td>Bifidobacteriaceae</td>
<td>Bifidobacterium</td>
</tr>
<tr>
<td>Bacteria</td>
<td>Proteobacteria</td>
<td>Betaproteobacteria</td>
<td>Burkholderiales</td>
<td>Alcaligenaceae</td>
<td>Sutterella</td>
</tr>
<tr>
<td>Bacteria</td>
<td>Proteobacteria</td>
<td>Alphaproteobacteria</td>
<td>Rhodospirillales</td>
<td>Rhodospirillales</td>
<td>Azospirillum</td>
</tr>
<tr>
<td>Bacteria</td>
<td>Proteobacteria</td>
<td>Alphaproteobacteria</td>
<td>Rhodospirillales</td>
<td>Acetobacteraceae</td>
<td>Rhodovarius</td>
</tr>
</tbody>
</table>

The application of PAAM-ML incorporates the phylogenetic tree information node-by-node. The predictive models over PAAM were comprised of microbial taxa at different taxonomic levels. The results indicated that including microbial feature nodes at all levels of taxonomy, has the potential to increase the predictive performance over the sampled cattle microbiome. The intermediate ancestral nodes of the phylogenetic tree as candidate features providing more accurate predictions. RFI did play an important role in selecting candidate features important for differentiating diet phenotypes related to cattle microbiome. The ensemble of XGBoost or Pen-LR methods with RFI ranking of features over PAAM provided good classification performance over the high-dimensional metagenomes. ASVs were grouped into key phyla in the cattle rumen microbiome: *Firmicutes, Actinobacteria, and Proteobacteria*. The idea of combining quantitative and qualitative profiles along the tree was based on combining natural biological characteristics of cattle microbial taxa. This section further lists the key predictive performance differences between the PAAM-ML and other phylogeny-aware classification techniques of
A use case: Determine functional phenotype of the cattle rumen microbiome

PhILR[133] employed with ML models (LOOCV settings) and MetaPhyl (with its default settings) [87]. The results are shown in Figure 7.1 and indicate the relatively superior performance of PAAM-ML modelling over other benchmarks of PhILR[133] and MetaPhyl [87] integrative models.

Figure 7.1 Results obtained from PAAM-ML, PhILR and MetaPhyl over 80 cattle rumen samples relating to diet phenotype.

7.2.3.2 Analysis based on the application of Phylogeny-RF to differentiate supplemented diet phenotype

Phylogeny-RF (the method proposed in Chapter 6) is employed hereby to study cattle microbiome (obtained from MetaPlat as M2 data source (Table 7.1)) by capturing the phylogenetic effects in the most popular RF classifier model. Encouraged by the assumption that hierarchically close taxa to have similar phenotypic responses, clustering over the cophenetic similarity matrix (obtained from branch lengths of a phylogenetic tree) [250], [274], was conducted. The key idea behind the use of Phylogeny-RF is to model RF over the ASV features
with minimum phylogenetic redundancy and maximal biological relevance; to improve its functionality in the current cattle rumen functional study.

Table 7.4 Predictive performance of Phylogeny-RF (with 5-fold cross-validation) over cattle microbiome rumen to differentiate samples (best performance is bold-faced in each column).

<table>
<thead>
<tr>
<th>Number of Trees</th>
<th>RF</th>
<th></th>
<th>Phylogeny-RF</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Kappa</td>
<td>Accuracy</td>
<td>Kappa</td>
</tr>
<tr>
<td>10</td>
<td>0.637</td>
<td>0.525</td>
<td>0.588</td>
<td>0.454</td>
</tr>
<tr>
<td>20</td>
<td>0.600</td>
<td>0.471</td>
<td>0.675</td>
<td>0.574</td>
</tr>
<tr>
<td>40</td>
<td>0.662</td>
<td>0.540</td>
<td>0.662</td>
<td>0.545</td>
</tr>
<tr>
<td>64</td>
<td>0.712</td>
<td>0.606</td>
<td>0.713</td>
<td>0.613</td>
</tr>
<tr>
<td>100</td>
<td>0.687</td>
<td>0.580</td>
<td>0.838</td>
<td>0.777</td>
</tr>
<tr>
<td>128</td>
<td>0.737</td>
<td>0.638</td>
<td>0.775</td>
<td>0.694</td>
</tr>
<tr>
<td>164</td>
<td>0.762</td>
<td>0.671</td>
<td>0.738</td>
<td>0.632</td>
</tr>
<tr>
<td>200</td>
<td>0.762</td>
<td>0.685</td>
<td><strong>0.850</strong></td>
<td><strong>0.791</strong></td>
</tr>
<tr>
<td>225</td>
<td>0.762</td>
<td>0.685</td>
<td>0.800</td>
<td>0.731</td>
</tr>
<tr>
<td>300</td>
<td><strong>0.800</strong></td>
<td><strong>0.730</strong></td>
<td>0.750</td>
<td>0.655</td>
</tr>
<tr>
<td>500</td>
<td>0.787</td>
<td>0.703</td>
<td>0.762</td>
<td>0.662</td>
</tr>
</tbody>
</table>

Overall, results (shown in Table 7.4) indicate that the integration of biological domain knowledge is useful in the classification of metagenomes as phylogeny-guided RF attained better performance in relatively less number of trees than RF applied over the ASV abundance profiles. The results indicate increasing the reliability of microbiome analysis using phylogeny along the current workflow. These suggest that using a Phylogeny-RF approach achieved relatively high Accuracy with 200 number of trees.

7.2.3.3 Phylogeny-aware analysis of circulating glucocorticoid on the cattle rumen microbiome with forage or concentrated diet

Dexamethasone (DEX), synthetic cortisol, has been used in this analysis to probe the effect of a circulating glucocorticoid on the cattle rumen microbial behaviour. DEX treatment has shown linkages to cattle behavioural aspects such as stress, body weight, and glucose levels in literature [275]. In the current
A use case: Determine functional phenotype of the cattle rumen microbiome

context, ruminal content samples were taken before as well as post to the DEX treatment for 3 consecutive days. The impact of the fed diet (concentrated or forage) to cattle rumen was also studied under the course of this dataset (indicated as M3 in Table 7.1). The contribution of this study is the use of integrative computational techniques to identify the effect of the phenotype (a) DEX and (b) diet over the microbial profiles discriminative of cattle microbiome samples. As discussed earlier, the methods of PAAM-ML and Phylogeny-RF has been developed and implemented in this thesis to perform the phylogeny-aware analysis of microbiome. These have been used here in this analysis for DEX and diet prediction in cattle microbiome. A statistical test to check the significance of how DEX and diet are related to cattle microbiome is also conducted to confirm the results. PAAM-ML was applied over the cattle rumen samples to differentiate them in the phenotype of (a) Diet and (b) the DEX treatment. The obtained results are listed in Table 7.5. Results (Table 7.5) indicate that diet phenotype can ultimately influence host productivity, which may either depend on the concentrated or forage diet. A highly predictive feature subset, capable of achieving near-perfect Accuracy (0.983) on this dataset, was attained with XGBoost employed over Top 40% features derived from PAAM as ranked by RFI. However, the classification of cattle microbiome samples into the phenotype of DEX treatment attained maximum subject to 78.8% Accuracy only with SVM applied over 10% of RFI ranked features (Table 7.5).
Table 7.5 Predictive performance of PAAM-ML (with LOOCV settings) over cattle microbiome rumen to differentiate 118 cattle rumen samples (best performance is bold-faced).

<table>
<thead>
<tr>
<th>Ensemble ML Models (with LOOCV)</th>
<th>Diet</th>
<th></th>
<th>DEX Treatment</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Kappa</td>
<td>Accuracy</td>
<td>Kappa</td>
</tr>
<tr>
<td>Pen-LR employed over Top 10 % derived from PAAM as ranked by RFI</td>
<td>0.932</td>
<td>0.864</td>
<td>0.771</td>
<td>0.527</td>
</tr>
<tr>
<td>XGBoost employed over Top 10 % derived from PAAM as ranked by RFI</td>
<td>0.974</td>
<td>0.949</td>
<td>0.771</td>
<td>0.511</td>
</tr>
<tr>
<td>SVM employed over Top 10 % derived from PAAM as ranked by RFI</td>
<td>0.957</td>
<td>0.915</td>
<td><strong>0.788</strong></td>
<td><strong>0.556</strong></td>
</tr>
<tr>
<td>RF employed over Top 10 % derived from PAAM as ranked by RFI</td>
<td>0.949</td>
<td>0.898</td>
<td>0.627</td>
<td>0.275</td>
</tr>
<tr>
<td>Pen-LR employed over Top 20 % derived from PAAM as ranked by RFI</td>
<td>0.949</td>
<td>0.893</td>
<td>0.720</td>
<td>0.426</td>
</tr>
<tr>
<td>XGBoost employed over Top 20 % derived from PAAM as ranked by RFI</td>
<td>0.974</td>
<td>0.949</td>
<td>0.737</td>
<td>0.446</td>
</tr>
<tr>
<td>SVM employed over Top 20 % derived from PAAM as ranked by RFI</td>
<td>0.974</td>
<td>0.949</td>
<td>0.788</td>
<td>0.547</td>
</tr>
<tr>
<td>RF employed over Top 20 % derived from PAAM as ranked by RFI</td>
<td>0.966</td>
<td>0.932</td>
<td>0.618</td>
<td>0.260</td>
</tr>
<tr>
<td>Pen-LR employed over Top 40 % derived from PAAM as ranked by RFI</td>
<td>0.957</td>
<td>0.915</td>
<td>0.700</td>
<td>0.348</td>
</tr>
<tr>
<td>XGBoost employed over Top 40 % derived from PAAM as ranked by RFI</td>
<td><strong>0.983</strong></td>
<td><strong>0.966</strong></td>
<td>0.686</td>
<td>0.326</td>
</tr>
<tr>
<td>SVM employed over Top 40 % derived from PAAM as ranked by RFI</td>
<td>0.974</td>
<td>0.949</td>
<td>0.754</td>
<td>0.475</td>
</tr>
<tr>
<td>RF employed over Top 40 % derived from PAAM as ranked by RFI</td>
<td>0.975</td>
<td>0.949</td>
<td>0.593</td>
<td>0.168</td>
</tr>
</tbody>
</table>

This indicates that the type of diet (forage or concentrated) plays a better role in classifying cattle microbiome samples than the DEX treatment. Some of the microbial features playing an important role in their linkage to diet phenotype are identified by PAAM-ML and are listed in Table 7.6. The results also confirm to studies in [272],[273] indicating their importance in studying the effect of different diets on cattle rumen microbiome.
A use case: Determine functional phenotype of the cattle rumen microbiome

Table 7.6 A summary of important features selected by PAAM-ML method for classifying diet phenotype associated with cattle microbiome.

<table>
<thead>
<tr>
<th>Kingdom</th>
<th>Phylum</th>
<th>Class</th>
<th>Order</th>
<th>Family</th>
<th>Genus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bacteria</td>
<td>Proteobacteria</td>
<td>Gammaproteobacteria</td>
<td>Aeromonadales</td>
<td>Succinivibrionaceae</td>
<td>Anaerobiospirillum</td>
</tr>
<tr>
<td>Bacteria</td>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Clostridiales</td>
<td>Lachnospiraceae</td>
<td>Ruminococcus</td>
</tr>
<tr>
<td>Bacteria</td>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Clostridiales</td>
<td>Lachnospiraceae</td>
<td>Clostridium</td>
</tr>
<tr>
<td>Bacteria</td>
<td>Actinobacteria</td>
<td>Actinobacteria</td>
<td>Bifidobacteriaces</td>
<td>Bifidobacteriaceae</td>
<td>Bifidobacterium</td>
</tr>
<tr>
<td>Bacteria</td>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Clostridiales</td>
<td>Clostridaceae</td>
<td>Clostridium</td>
</tr>
<tr>
<td>Bacteria</td>
<td>Firmicutes</td>
<td>Erysipelotrichi</td>
<td>Erysipelotrichales</td>
<td>Erysipelotrichaceae</td>
<td>Anaerorhabdus</td>
</tr>
<tr>
<td>Bacteria</td>
<td>Bacteroidetes</td>
<td>Bacteroides</td>
<td>Bacteroidales</td>
<td>Prevotellaceae</td>
<td>Prevotella</td>
</tr>
</tbody>
</table>

Phylogeny-RF achieved better performance with regards to the linkage of cattle microbiome to diet phenotype in comparison to the DEX treatment, for every benchmark of the number of trees ranging from 1-500 (Table 7.7).

Table 7.7 Predictive performance of Phylogeny-RF with 5-fold cross-validation over cattle microbiome rumen to differentiate samples (best performance in each column is bold-faced).

<table>
<thead>
<tr>
<th>Number of Trees</th>
<th>Phenotype of Diet</th>
<th>Phenotype of DEX</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Kappa</td>
</tr>
<tr>
<td>10</td>
<td>0.947</td>
<td>0.893</td>
</tr>
<tr>
<td>20</td>
<td>0.947</td>
<td>0.889</td>
</tr>
<tr>
<td>40</td>
<td>0.982</td>
<td>0.965</td>
</tr>
<tr>
<td>64</td>
<td>0.982</td>
<td>0.963</td>
</tr>
<tr>
<td>100</td>
<td>0.973</td>
<td>0.948</td>
</tr>
<tr>
<td>128</td>
<td>0.965</td>
<td>0.930</td>
</tr>
<tr>
<td>164</td>
<td>0.973</td>
<td>0.945</td>
</tr>
<tr>
<td>200</td>
<td>0.982</td>
<td>0.963</td>
</tr>
<tr>
<td>225</td>
<td>0.982</td>
<td>0.959</td>
</tr>
<tr>
<td>300</td>
<td>0.982</td>
<td>0.965</td>
</tr>
<tr>
<td>500</td>
<td>0.982</td>
<td>0.963</td>
</tr>
</tbody>
</table>

Furthermore, aMiSPU [149] (a phylogeny-aware statistical test assuming the null hypothesis of “there exists no correlation between ASVs and diet as phenotype”), was conducted for detecting differences in the metadata
related to the cattle microbiome to study the impact of DEX and diet-related functions over microbial composition. aMiSPU reported a statistically significant relationship between cattle microbiome and diet ($p$-value < 0.01); however, it did not show any such significant correlation between the cattle microbiome and DEX treatment ($p = 0.80$). aMiSPU utilised UniFrac distance [50] to determine statistical significance. Applying the ordination of Principle Coordinate Analysis (PCoA) over UniFrac [50] provides some more insights into the mechanism of differentiating cattle microbial samples based on the functional phenotype of DEX and diet (Figure 7.2). Figure 7.2 shows that the diet of the cattle animals (forage or concentrate) had a much higher impact on phylogenetic diversity (as driven by the UniFrac distance), in comparison to DEX treatment.

![Figure 7.2 Cattle rumen samples differentiation as coloured by the (a) DEX treatment and (b) Diet (PCoA plot, as shown in this Figure).](image)

Microbial profile comparisons by multivariate phylogeny-aware statistical analysis with aMiSPU showed their correlation with diet phenotype. This also confirms the study in [267] supporting that diet is a major influence phenotype on cattle rumen microbial composition. Nonetheless, the DEX
A use case: Determine functional phenotype of the cattle rumen microbiome

treatment didn’t seem to have much impact on cattle rumen microbial communities in general. The current research indicates the lack of a direct effect of glucocorticoids in affecting cattle productivity by changes in ruminal microbial populations. Henceforth, the microbiome in cattle rumen seems resistant to DEX exposure.

7.2.4 Additional contributions towards a comprehensive analysis of MetaPlat generated cattle rumen data

This section outlines the additional contributory functional studies conducted using Approach 1 and 2 (Figure 3.5a, Figure 3.5b) employed over cattle rumen microbiome samples to achieve MetaPlat project's outcomes. These relate to additional contributions made by developing ML-based models to analyse functional metagenomes.

7.2.4.1 A study conducted using Approach 1 (abundance-driven) over 40 samples of cattle rumen

This study was conducted as part of MetaPlat Project to classify 40 cattle rumen samples (indicated as M1 in Table 7.1) into diet phenotype. The related work is published in [121]. The data consisted of 20 samples from an oil-based treatment and 20 samples from a nitrate-based treatment to study the effect of oil or nitrate supplemented diet on taxa abundance counts at the genus level of study. An essential, rigorous assessment of different data pre-processing, and classification techniques for inferring phenotype from this dataset was conducted. This study involved the investigation of wrappers (WFS) and filters (FFS) based methods of feature selection [34]; with the popular range of
A use case: Determine functional phenotype of the cattle rumen microbiome

classifiers of RF, SVM, and LR in the current metagenomics research [69],[93], [98]. This study explored the application of WFS based on LR over the microbiome for the first time. The experimental set-up related to this study is shown in Figure 7.3.

![Figure 7.3](image)

**Figure 7.3** A Computational approach followed to conduct a case study for functional analysis of cattle rumen microbiome associated with 40 samples and diet phenotype.

The following methods were employed to maximise the performance of the experimental design in the current study at the *genus* level of the taxonomy.

(a) Feature selection was conducted over OTU abundance count data in this preliminary study by using : - (i) Correlation-based feature selection (CFS) filter [192], selecting OTU features that are highly correlated with the class but uncorrelated with each other, and; (ii) WFS evaluating attribute sets by estimating their performance using a supervised learning method of LR and RF [232].

(b) This analysis involved the classification algorithms of RF [88], SVM [108], and LR [112]. These techniques have been selected as they have previously been employed successfully for the predictive task of inferring metagenomic functions [69], [98], [128].
A use case: Determine functional phenotype of the cattle rumen microbiome

(c) LOOCV validation procedure was carried out to assess the performance of each prediction model. The representative performance assessment metrics used for evaluating classification models were Accuracy and Kappa [200]. The results from each fold in the LOOCV process were averaged to produce the resultant performance values.

The related experiments to employ CFS, WFS with LR, WFS with RF as data pre-processing techniques and; classifiers of RF, SVM, and LR from literature [69],[98],[128],[276], were performed in WEKA 3.8 tool [277] (using the default settings). The obtained results are graphically displayed in Figure 7.4. In Figure 7.4, each panel represents the predictive performance obtained of each ML method used for classifying the 40 cattle samples. The findings report that the application of feature subset selection plays a crucial role in obtaining better classification performance in functional metagenomics. Interestingly, from Figure 7.4, it was observed that the proposed combination of feature selection with WFS including LR and, classification with LR as a classifier; provided best average classification performance in comparison to other benchmarked classifiers of RF, SVM, and LR from literature [69],[98],[128]. It was observed that the combination attained a high Accuracy of 0.950, and Kappa of 0.900 with only 12 out of 386 OTU features over the 40 cattle microbiome samples obtained as part of the MetaPlat project [144]. The experimental results over this sample case study indicated that WFS-based methods paved the path for getting a chance of finding an optimal feature subset and increased the predictive performance of differentiating between diet phenotype.
A use case: Determine functional phenotype of the cattle rumen microbiome

Figure 7.4 Results obtained from the preliminary case study over 40 cattle rumen samples.

The predominant biological OTU features obtained from WFS methods useful in the classification of phenotypes associated with the current dataset are reported in Table 7.8. The bacterial sequences were majorly dominated by *Firmicutes* in ruminants. The results in Table 7.8. also, confirm to the study conducted in [278]. The diet supplements could majorly affect the configuration of Clostridiales (Order) [278].

---

**Figure 7.4 Results obtained from the preliminary case study over 40 cattle rumen samples.**

The predominant biological OTU features obtained from WFS methods useful in the classification of phenotypes associated with the current dataset are reported in Table 7.8. The bacterial sequences were majorly dominated by *Firmicutes* in ruminants. The results in Table 7.8. also, confirm to the study conducted in [278]. The diet supplements could majorly affect the configuration of Clostridiales (Order) [278].
Table 7.8 A summary of predominant features selected by WFS method for classifying diet phenotype associated with cattle microbiome.

<table>
<thead>
<tr>
<th>Kingdom</th>
<th>Phylum</th>
<th>Class</th>
<th>Order</th>
<th>Family</th>
<th>Genus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bacteria</td>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Clostridales</td>
<td>Veillonellaceae</td>
<td>Veillonella</td>
</tr>
<tr>
<td>Bacteria</td>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Clostridales</td>
<td>Incertae Sedis XI</td>
<td>Anaerococcus</td>
</tr>
<tr>
<td>Bacteria</td>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Clostridales</td>
<td>Tissierellaceae</td>
<td>Tepidimicrobium</td>
</tr>
<tr>
<td>Bacteria</td>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Clostridales</td>
<td>Lachnospiraceae</td>
<td>Butyrovivrio, Orthobacterium,</td>
</tr>
<tr>
<td>Bacteria</td>
<td>Firmicutes</td>
<td>Bacilli</td>
<td>Bacillales</td>
<td>Staphylococaceae</td>
<td>Staphylococcus</td>
</tr>
<tr>
<td>Bacteria</td>
<td>Firmicutes</td>
<td>Bacilli</td>
<td>Lactobacillale</td>
<td>Carnobacteriaceae</td>
<td>Carnobacteriaceae, Trichococcus</td>
</tr>
<tr>
<td>Bacteria</td>
<td>Actinobacteria</td>
<td>Actinobacteria</td>
<td>Actinobacteridiae</td>
<td>Actinomycetales</td>
<td>Corynebacteriae</td>
</tr>
<tr>
<td>Bacteria</td>
<td>Bacteroidetes</td>
<td>Bacteroidetes</td>
<td>Bacteroidales</td>
<td>Bacteroidaceae</td>
<td>Bacteroides</td>
</tr>
<tr>
<td>Bacteria</td>
<td>Bacteroidetes</td>
<td>Bacteroidetes</td>
<td>Bacteroidales</td>
<td>Porphyromonadaceae</td>
<td>Porphyromonas</td>
</tr>
<tr>
<td>Bacteria</td>
<td>Actinobacteria</td>
<td>Actinobacteria</td>
<td>Actinomycetales</td>
<td>Beutenbergiaceae</td>
<td>Salana</td>
</tr>
<tr>
<td>Bacteria</td>
<td>Proteobacteria</td>
<td>Gammaproteobacteria</td>
<td>Enterobacterial</td>
<td>Enterobacteriaceae</td>
<td>Erwini</td>
</tr>
</tbody>
</table>

In this study, a useful computational approach is characterised to determine OTUs that are useful in identifying functional roles associated with the cattle microbiome. The study indicates that feature selection methods play an important role in handling high-dimensional OTU datasets. In general, optimum feature subset selection provides better insights to study linkages between OTUs and their related environmental traits, and subsequently to classify them into the functional classes. The proposed novel combination of WFS with LR as a feature selector and LR as the classifier is suitable and effective for this dataset with small sample-size (40 samples in this use case). The study also reported significant biological features obtained from WFS methods for classifying phenotypes.

However, WFS-based methods are computationally intensive for relatively large-scale data. Also, this preliminary study conducted was limited
A use case: Determine functional phenotype of the cattle rumen microbiome

to the OTU abundance count profiles and did not consider the relationships between OTUs based on their phylogeny in predictive modelling.

**7.2.4.2 A study conducted using Approach 2 over 80 samples of cattle rumen**

In an attempt to assess the topological relevance of microbial features (ancestral nodes in a phylogenetic tree) at higher taxonomical levels and; to identify their role in biological mechanisms of cattle rumen microbiome analysis, use of new phylogeny-driven Relief-based measure [90] is proposed in this analysis. The work was submitted and accepted by *Data Analytics in Metagenomics workshop* in *IEEE BIBM 2019* conference. In the proposed approach, the Relief measure [90] gained insights from a phylogenetic metric of the Phylogenetic Interaction-Adjusted index (PINA) [91] to rank the taxonomical features. Henceforth, the approach is termed as “Phylogeny-PINA Relief”.

A study in [92], utilised the UniFrac distance metric [50] to calculate the phylogeny-based similarity between metagenomic samples and to determine the nearest hit or miss and thereby applying Relief-based [90] measure for ranking of microbial clades (corresponding to ancestral node features on a phylogenetic tree). It served as a useful benchmark to study the use of phylogenetic indices [50],[91] instead of using phylogenetic weights directly in the regularization of microbial profiles. This proposed approach is unlike PAAM-ML approach proposed in Chapter 5 which introduced phylogenetic weights in the formulation of input data to ML models.

In this study, the microbial profiles of metagenomic samples are modelled firstly by merging the abundance profiles of ASVs into ancestral nodes, creating
a data structure of a Phylogeny Topology and Abundance-aware Matrix (PTAM) along the taxonomical hierarchy (topology). Here only the topology of a phylogenetic tree is considered to model the relationships between OTUs and the ancestral nodes in the input data modelling. However, the phylogenetic distances are introduced in the feature selection criteria later. The abundances of leaf-level ASVs remained same in PTAM. The profiles of ancestral nodes in the PTAM matrix were computed by combining abundances of its constituting ASVs, forming a hierarchal topology. The detailed procedure of PTAM construction is described in Algorithm 7.1.
A use case: Determine functional phenotype of the cattle rumen microbiome

Algorithm 7.1 Construction of PTAM Feature Space

**Input:** OTU Abundance Count Matrix $C_{n \times m}$, Phylogeny Tree ‘T’

**Output:** PTAM denoted as $P_{n \times 2m-1}$

**procedure** PTAMConstruction (OTU Abundance Count Matrix ‘C,’ Phylogeny Tree ‘T’)

# Quantitative profiles of leaf-level features remain intact

for each sample row, ‘$i$’ in original Abundance Count Matrix $C_{n \times m}$ with $n$ samples and $m$ OTU features do

  for each feature $j$ in $C$ and $T$

    $P_{i,j} \leftarrow C_{i,j}$

  end for

end for

# Quantitative Profile of Ancestral Nodes

for each sample row, ‘$i$’ in original Abundance Count Matrix $C_{n \times m}$ with $n$ samples and $m$ OTUs do

  $j \leftarrow m + 1$

  for each ancestral node $v$ in a phylogeny tree $T$ do

    $P(i, j) \leftarrow 0$

    for each OTU node ‘$u$’ in $C_{n \times m}$ do

      if OTU ‘$u$’ in the sample ‘$i$,’ is descendant of node ‘$v$’ in $T$ then

        $X_{u,i} \leftarrow$ abundance count of OTU ‘$u$’ in the sample ‘$i$’

        $P_{i,j} \leftarrow P_{i,j} + X_{u,i}$

      end if

    end for

  end for

  $j \leftarrow j + 1$

end for

**return** PTAM Feature Space, $P_{n \times 2m-1}$

Feature selection over such high-dimensional feature-set obtained from PTAM is expected to yield advantage in computational performance. In the proposed approach, ancestral elements were ranked according to their contribution to separating microbial samples into phenotype classes, as suggested in [92]. The study in [92] suggested the use of the phylogenetic distance matrix of UniFrac[50] to be utilised in Relief measure of feature selection. But in the current study, a recently developed index of phylogenetic
A use case: Determine functional phenotype of the cattle rumen microbiome

similarity termed as Phylogenetic INteraction-Adjusted index (PINA) [91] is utilised in place of UniFrac in the analysis. In the process of feature selection, a microbial sample is randomly selected and; the nearest sample of the same class (i.e., Hit ‘H’), and the nearest sample of a different class (i.e., miss ‘M’), in accordance to the PINA similarity[91] are utilised in determining weighted ranking of features in a Relief-based measure (Eq.7.1) [90].

for each subtree in the phylogenetic tree,

\[
W_f = W_f - \frac{\text{diff}(f, S, H)}{i} + \frac{\text{diff}(f, S, M)}{i}
\]

(7.1)

where \(W_f\) is, the weight associated with a feature \(f\) which is a combinatorial (ancestral) node in the phylogenetic Tree, ‘\(S\)’ is the randomly sampled instance, \(‘H’\) is the nearest hit and ‘\(M\)’ is the nearest miss, \(\text{diff}\) calculates the PINA distance between two samples considering sub-tree for a given feature \(f\), and \(i\) is the number of samples. PINA finds the phylogenetic similarity between samples, according to Eq.7.2 [91].

\[
\text{PINA} = \frac{\sum_{i \in A} \sum_{j \in B} n_A n_B \phi_{ij}}{\sqrt{\left(\sum_{i \in A} \sum_{j \in A} n_A n_{ij} \phi_{ij}\right)\left(\sum_{i \in B} \sum_{j \in B} n_B n_{ij} \phi_{ij}\right)}}
\]

(7.2)

where \(\phi_{ij}\) represents cophenetic distance matrix (i.e. PDM) [250] serving as a phylogenetic association matrix; the entries in the matrix are the sum of the length of the branches separating each pair of microbial species on a phylogenetic tree. Thus, if two species are far apart on the phylogeny, their distance will be larger than that for two closely related species.; \(A\) and \(B\) represent microbial samples with \(n_A\) and \(n_B\) microbial species that have been sampled. The steps involved in Phylogeny-PINA Relief are summarised in Figure 7.5. Firstly (i) PTAM is created ; (ii) Relief based on the phylogenetic
distance of PINA is applied for feature selection; (iii) classifier is applied over selected feature space, and; (iv) performance is analysed.

**Figure 7.5 A summary of the stepwise procedure of Phylogeny-PINA Relief approach.**

The approach is further benchmarked with the Relief driven by UniFrac-based similarity [50] between samples to rank internal nodes as proposed in [92]. The Accuracy and Kappa assessment metrics were employed for evaluating the predictive models in this study.

For a benchmark comparison (Table 7.9), the feature engineering was employed with - (A) application of Relief measure [90] directly over the data vectors in PTAM; (B) application of proposed PINA-driven Relief method over PTAM; (C) UniFrac-driven Relief measure [92] over PTAM; and (D) Relief-measure over PAAM [228]. The phylogeny-aware feature engineering with PINA-driven Relief was implemented with the help of Perl Script. Top n % of ancestral features by Relief measure (denoted as Top n % Imp_Fea_Relief,
where $n = 10, 30, 50, 70$) were selected to be input to the popular classifier method of RF in microbiome studies (as suggested in Chapter 2). RF method is implemented in the *caret* R package settings.

Table 7.9 Predictive performance of state-of-the-art RF model applied over (a) relief-selected features from PTAM, (b) PINA-driven relief selected features from PTAM, (c) UniFrac-driven relief selected features from PTAM and (d) relief-selected features from PAAM; with LOOCV settings. the dataset was characterised into 4 functional categories of diet (oil, nitrate, oil-nitrate combined, and controls). The best performance in each column is bold-faced.

<table>
<thead>
<tr>
<th>Feature Selection by Relief Measure</th>
<th>(A) Relief over PTAM Data Structure</th>
<th>(B) PINA-driven Relief over PTAM (Phylogeny-PINA Relief)</th>
<th>(C) UniFrac-driven Relief over PTAM</th>
<th>(D) Relief over PAAM Data Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Kappa</td>
<td>Accuracy</td>
<td>Kappa</td>
</tr>
<tr>
<td>Top 10% Imp_Fea_Relief</td>
<td>0.825</td>
<td>0.766</td>
<td>0.950</td>
<td>0.930</td>
</tr>
<tr>
<td>Top 30% Imp_Fea_Relief</td>
<td>0.862</td>
<td>0.816</td>
<td>0.950</td>
<td>0.933</td>
</tr>
<tr>
<td>Top 50% Imp_Fea_Relief</td>
<td>0.850</td>
<td>0.800</td>
<td><strong>0.962</strong></td>
<td><strong>0.950</strong></td>
</tr>
<tr>
<td>Top 70% Imp_Fea_Relief</td>
<td><strong>0.875</strong></td>
<td><strong>0.833</strong></td>
<td>0.962</td>
<td>0.950</td>
</tr>
</tbody>
</table>

The feature selection with PINA attained better performance than the Relief measure applied directly over PAAM (Phylogeny-PINA Relief), PTAM, and; the UniFrac-driven [92] Relief measure applied over PTAM (Table 7.9). This illustrates the advantage of applying the Phylogeny-PINA Relief approach when inferring functions of cattle microbiome in comparison to the application of Relief over PTAM, Relief over PAAM, or Relief-driven by UniFrac over PTAM [92]. The attraction of using phylogeny-driven Relief measure lies in its
capability to rank microbial taxa according to their biological characteristics. In the current study, class Clostridia found to be a good example, where cattle microbiome seemed to be related and playing an important role concerning the cattle dietary supplements.

### 7.2.5 Contributory scripts

The motivation of this work was to work towards MetaPlat-based outcomes. The study explores the use of a combination of abundance and phylogeny-driven methods for classifying 3 datasets (as mentioned in Table 7.1). The three key integrative methods were successfully employed with 3 generative scripts: (i) integration of phylogeny in data modelling (PAAM-ML); (ii) integration of phylogeny in microbial feature ranking (Phylogeny-PINA Relief); and (iii) integration of phylogeny in ML model (Phylogeny-RF) (Table 7.10). These methods developed as programming scripts in this thesis are made available to the research community for any kind of extensive analysis at GitHub repository.

**Table 7.10 Summary of integrative methods employed over the MetaPlat data.**

<table>
<thead>
<tr>
<th></th>
<th>Abbreviation</th>
<th>Level of Inclusion of Phylogeny</th>
<th>Implementation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ML employed over the Phylogeny and Abundance-aware data modelling (PAAM-ML)</td>
<td>Data Pre-processing/Modelling</td>
<td>*Perl Script to generate PAAM [227],[228] and R package script to apply ML using caret package.</td>
</tr>
<tr>
<td>2</td>
<td>Phylogenetic Ranking of Features at higher taxonomical level (Phylogeny-PINA Relief)</td>
<td>Feature Ranking</td>
<td>*Perl Script to generate phylogeny-aware feature space and it’s ranking, and; R package to apply ML (Work accepted in IEEE BIBM 19’) [279]</td>
</tr>
<tr>
<td>3</td>
<td>RF is driven by the Phylogenetic measure (Phylogeny-RF)</td>
<td>ML model Itself</td>
<td>*Python Script to conceive RF driven by phylogeny</td>
</tr>
</tbody>
</table>

7.3 Key findings

This chapter evaluated ML, and feature ranking approaches for functional classification of cattle metagenomes. The experiments presented in this chapter highlight the performance capability of abundance and phylogeny-aware approaches when used for cattle rumen classification. Some of the key contributory observations from this chapter are summarised below.

- In the preliminary study conducted over 40 cattle rumen samples (following Approach 1), it was observed that the reduced set of taxa obtained by WFS (i.e., wrapper) methods proved its utility to attain best performance results. However, WFS methods are computationally intensive to applied to relatively large datasets.

- In another study of the effect of diet supplements on cattle microbial taxa, supervised learning was used to demonstrate a generalizable difference between phylogeny-aware features (taxa) and non-phylogenetic absolute abundances of taxa. This aided in identifying highly discriminative taxa to be further used in other supervised models following the PAAM-ML framework (Approach 2). PAAM-ML proved effective in its application to associate cattle microbiome samples with diet (attaining high Accuracy: 0.937, Kappa: 0.917).

- A new method of ‘Phylogeny-PINA Relief’ (along Approach 2) was proposed and employed to predictive modelling of cattle microbial communities in 80 samples. Phylogeny-PINA Relief used phylogenetic interaction based on the PINA similarity matrix[91] to estimate the weights of important taxa. It thereby applies a Relief-based feature engineering
algorithm to rank the features based on estimating the weight of each feature (obtained from PINA) in the microbial community contributed by the source metagenomic environment. Previous work focused on the estimation of Relief weights by using the UniFrac phylogenetic similarity [92]. A benchmark comparison with the previous work is provided in this chapter. It was observed that the ‘Phylogeny-PINA Relief’ approach to estimating the predictive performance is driven by phylogenetic similarity measures and improves the performance. High performance with Accuracy: 0.962, Kappa: 0.950 was achieved with Phylogeny-PINA Relief.

- The chapter also contributes to a prominent study of the phylogenetically diverse microbiome features in RF modelling, showing that the functions of a microbial community can be predicted efficiently from its phylogenetic profiling as part of the ML model. This was achieved using the pipeline of Phylogeny-RF (proposed in Chapter 6 following Approach 3).

- The work in this chapter was further driven by the interest in using a phylogeny-aware statistical model to confirm how diet and DEX treatment as an environmental factor affects microbial communities from varied assemblages in cattle rumen. The result indicates a significant relation of the diet with the cattle microbiome ($p$-value < 0.05).

- Internal Nodes modelled in PAAM-ML and Phylogeny-PINA Relief served as important features in classification of phenotypic functions.

- Overall, phylum Firmicutes and Class Clostridia seems to play an important role in cattle related functions.
7.4 Limitations

The phylogenetic tree was not available for M1 (Table 7.1) having 40 samples of cattle rumen. Only a preliminary study was conducted over this dataset. However, the study employed over M1 indicated the potential use of the wrapper method in such kinds of datasets dealing with microbial abundances. Wrappers were not utilised for other datasets being time intensive. The novel approach of ‘Phylogeny-PINA Relief’ proposed in this chapter is also time-consuming, due to iterative tasks involved in calculating feature importance based on the phylogenetic similarity between the microbial features. During the application of Phylogeny-RF approach over M2, M3, modelling was employed over the feature space of leaf-level OTUs to be input to ML classifier and not the ancestral nodes, unlike PAAM-ML and Phylogeny-PINA Relief. For all of these concerned areas, future work is possible. Furthermore, the phenotypes could of feeding residual efficiency and methane emission could be explored in future for further analysis.

7.5 Summary

The chapter describes the execution of computational models while highlighting the use of methods proposed in this thesis to analyse cattle microbiome with some additional contributions; resulting in a knowledge-base for the project consortium. The work presented in this chapter focused on classifying cattle rumen microbiome extracted from three different datasets, M1, M2 and M3 (indicated in Table 7.1). A variety of integrative methods were used for each
dataset due to each datasets’ composition. Experiments were completed for each dataset using the phylogeny-aware feature set except M1 for which only abundance-driven analysis was conducted. Results showed that phylogeny-driven approaches achieved a higher Accuracy for both M2 and M3. Analysis over M1 used the conventional feature extraction approaches to classify small scale microbiome samples and attained a high Accuracy. The statistical results also indicated the relation of cattle nutritional diet with its microbial composition. Some of the important taxa were identified from phylum to genus level supporting cattle functions of a nutritional diet. The proposed approaches are extensible to study the cattle microbiome composition further. Figure 7.6 summarizes the key discoveries of this case-study over the cattle rumen microbiome.
Key Scientific Discovery

• A new approach of LR classifier applied over WBS (with LR) selected features reported high Accuracy of 95%, and Kappa of 0.900 with only 12 out of 386 OTU features over the 40 cattle microbiome samples obtained as part of the MetaPlat project.
• PAAM-ML and Phylogeny-RF achieved better performance in comparison to state-of-the-art over the MetaPlat obtained datasets.
• The phenotype of Diet had a much higher impact on cattle microbiome in comparison to the phenotype of DEX treatment.
• A new proposed approach of Phylogeny-PINA Relief attained promising results in the functional classification of cattle metagenomes indicating the utility of PINA similarity matrix over the state-of-the-art of UniFrac and the application of Relief over OTU abundances.
• The diet supplements could majorly affect the configuration of Clostridiales (Order) in cattle.
• Various other important cattle microbiome were identified for classifying cattle functions (as indicated in Tables 7.3,7.6,7.8).

Strengths

• Attempt to use PINA phylogeny-based similarity matrix in a Relief-based feature selection is made for the first time. In literature to date, UniFrac has been explored much to integrate biological knowledge of phylogeny in comparison to PINA.
• Phylogeny-aware modelling achieved a higher Accuracy in the use case related to cattle microbiome, resulting in a knowledge-base for the MetaPlat project consortium.

Limitations

• The new approach of LR applied over WFS is time-intensive over large datasets.
• Phylogeny-PINA Relief proposed in this case study is also time-consuming, due to iterative tasks involved in calculating feature importance based on the phylogenetic similarity between the microbial features.
• The study could also be further extended to study the phenotypes such as feeding residual efficiency and methane emissions.

Figure 7.6 A summary of key observations, strengths and limitations of this study.
CHAPTER 8

SUMMARY AND FUTURE WORK

The work presented in this thesis focuses primarily on exploring methods to use for functional classification of the microbiome. The central objective of this research has been to improve upon functional predictions of metagenomes considering the quantitative (abundances of microbial genomes) with qualitative aspects (evolutionary knowledge) in an integrative way. This chapter discusses a summary of the contributions made in this thesis along with limitations and the plausible extensions as future work of this research. Finally, an overall summary is presented.

8.1 Introduction

Microbiome sourced from an environment actively participate in various phenotypic functions, for instance: (1) role of the human microbiome in the diseased state; (2) in signalling association of microbiome with different environmental sites and (3) in response to supplements added to the host environment. This thesis provides a comprehensive evaluation of the sample microbiome datasets in defining their phenotypic roles. Studies [28], [32], [93], [98], [128], have suggested that supervised ML techniques can be employed to analyse and infer the functional roles of the microbiome. In Chapter 2 on
Literature Review, various computational methods were discussed that have been applied in researching functional metagenomics. Research on using functional metagenomes for deciding environmental roles is ongoing. Chapter 2 reported RF as the most popular supervised learner to classify metagenomes into functional phenotypes. However, analysis of metagenomic datasets face computational challenges due to their key characteristics and in the integration of biological domain knowledge in functional metagenomics [31],[41],[79],[80]. This thesis has aimed to address challenges of computationally inferring metagenomic functions overall by: (1) pre-processing or modelling over microbiome data to predict functional phenotype; (2) constructing new data structure incorporating both abundances and phylogeny of microbiome; (3) engineering of feature vectors by integrating phylogenetic domain knowledge; (4) developing and implementing new integrative ML model for the prediction of metagenomic functions; (5) employing different assessment methods to evaluate the functional metagenomic predictions. Additionally, the thesis has applied a variety of computational methods to categorise cattle rumen into their functions to achieve new insights from the real-world dataset sampled as an integral part of the MetaPlat project[144]. The current research has been disseminated in the form of various publications as listed in Chapter 1 (section 1.10). The main contributions of this research are outlined in the next section.

8.2 Contribution to the knowledge

It has been demonstrated that ML techniques could be employed successfully to pre-process and classify metagenomic data into phenotypic functions. The
Contribution to the knowledge

adaption of advanced, accurate and efficient methods could prove more useful in modelling over quantitative abundances of metagenomes that are high-dimensional and sparse. Furthermore, biological knowledge relating to microbial taxa could be integrated into abundance profiles to transform the microbial feature vector into more biologically relevant information for the subsequent data modelling. The computational methods relying only on abundance counts would ignore any biologically relevant change that occurs during evolution in the less abundant microbial taxa. The methods based only on phylogeny would ignore the quantitative profile counts. Nonetheless, the integration of both abundance and phylogenetic profiles would support analysis based on both the quantitative and qualitative (biological) aspects. Based on a combination of phylogeny and abundance profiles, this research provides the research community with the new computational methods to support functional metagenomics, which is an active area of research. The computational methods are developed for microbial profiles obtained from chosen metagenomic sources containing information on human, cattle, or soil microbiomes as part of the current research. These methods circumvent the predictive modelling workflow from data-pre-processing and supervised ML to performance evaluation. The performance of predictive modelling for functional classification of the microbiome is evaluated in terms of different predictive performance measures (Accuracy, Kappa and/or AUC-ROC). The computational models developed as part of this research followed either Approach 1, 2, or 3 (as proposed in Chapter 3 on Research Methodology and Materials). Approach 1 deals with ML analysis applied over the abundance-driven profiles of the microbiome. Approach 2 integrates phylogenetic tree and
Contribution to the knowledge abundance profile at the data pre-processing (modelling) stage. Approach 3 proposes integrating phylogeny in an ML model to use it further for the functional classification of the microbiome. Overall, the work presented in this thesis is contributed to specific sections within the high-level pipeline shown in Figure 8.1. The factors driving research objectives along this pipeline (Figure 8.1) are defined in the form of 4D’s (Figure 8.2) to answer the related research questions. The remainder of this section will discuss each contribution as an answer to the research question and how they relate to a specific Chapter.

**Figure 8.1 High-level pipeline design to classify microbiome (genome structure) into functional phenotypes (genome function).**

The thesis is developed upon answering the following research questions, to provide the research community with advanced methods to support the functional classification of metagenomes.
1. How to design an efficient computational approach to extend beyond the conventional methods of downstream functional analysis in metagenomics, for attaining better outcomes while dealing with high-dimensional metagenomic data? (Chapter 4)

2. How to include biological domain knowledge (i.e., taxonomical structure or phylogeny or depth covering lineages from phylum to species levels or the natural characteristics of the microbiome) into the perspective of computational metagenomics analysis for predicting functions? (Chapter 5, Chapter 7)

3. How to integrate diverse and heterogenous metagenomic data sources in a structured way for the downstream analysis of metagenomes? (Chapter 5)

4. How to develop a novel phylogenetic knowledge-driven ML Method, which could govern the functional derivations from metagenomics profiling data? (Chapter 6)

5. To perform a case-study evaluation to investigate essential questions (Who? How? and What?) of the microbiome research. (Chapter 7)
Figure 8.2 4D’s highlighting major contributory factors in the context of the current thesis.

The key contributions are listed below.

8.2.1 A comprehensive assessment of predictive models for inferring phenotypes related to two use cases of the human microbiome (DS5,6) [Chapter 4]

The work presented in this contribution is based on published work [223]. The choice of ML technique is an important task in predicting functions of sparse and high-dimensional metagenomes. The prediction of metagenomic functions can be viewed as either a binary or multi-class classification problem. The
embedded feature techniques are applied to encoded information related to the human microbiome and its relationship with the phenotype of diet (a use case of binary classification) and IBD (a use case of multi-class classification). The analysis followed Approach 1 (Figure 3.5a), as proposed in Chapter 3. A systematic comparison of different ML methods to predict functional phenotypes was performed.

The embedded techniques, such as XGBoost has been popularised in various areas but has not been explored much in the area of functional metagenomics in literature. Henceforth, these are applied in the current thesis for the functional metagenomic analysis. XGBoost seems suitable to fit the key characteristics of metagenomic data. Another such embedded model is Pen-LR (glmnet), which was applied over the sparse and high-dimensional metagenomic data. The use of embedded approaches has suggested that they can improve over the representative methods of RF and SVM [93],[128] to infer metagenomic functions in terms of computational performance. It was observed that an ensemble of embedded methods of XGBoost and/or Pen-LR (glmnet) with classical methods of RF or SVM proved to be more proficient at predicting phenotypes. Features extracted from the embedded methods of XGBoost and glmnet provided good predictive power in less computational time.

The embedded methods performed significantly better than SVM (in terms of predictive performance with AUC-ROC and the computational time in seconds). Their use also greatly improved over the execution time of most popular state-of-the-art RF and provided competitive or better predictive performance in a comparative analysis. The predictive ability to combine unique feature sets obtained from XGBoost, and glmnet was investigated along
the proposed abundance-driven workflow (termed as XGB-pen-LR-Classification) for a comprehensive assessment. Some of the representative taxa were identified using the abundance-driven framework built upon the mentioned-above embedded methods, relating to differentiating phenotypes in human microbiome samples. The proposed experimental design suggested that overall feature engineering plays an important role in analysing the high-dimensional metagenomic data and improving overall performance in human microbiome use cases under study. The study also indicated the use of the recursive feature elimination (as a WFS based approach), is computationally intensive for the large-scale and high-dimensional metagenomic data.

The proposed embedded ML-based ensembled framework of XGB-pen-LR-Classification (Chapter 4) indicated promising results in comprehensive set experiments over the use cases of high dimensionality (p features >> n samples). Hence, the study provides important methodological contributions for the use of embedded ML models in the field of functional metagenomics.

8.2.2 Proposal and application of a novel PAAM-ML framework for the prediction of metagenomic functions [Chapter 5]

The second and major contribution of this thesis is the development of a novel PAAM-ML framework that captures the evolutionary relationships between microbial features rather than considering them as independent features. The work presented in this contribution is based on published work [227],[228]. PAAM-ML framework has been proposed for the prediction of metagenomic functions in human and soil microbiome (following Approach 2). The RF
Contribution to the knowledge classification method has commonly been employed for the prediction of high-dimensional metagenomic data [93],[98],[128]. However, RF assumes independence between features [93],[128]. As the metagenomic pipelines could derive the phylogenetic knowledge with the quantitative profiles, the prediction of metagenomic functions built upon considering these relationships would inform metagenomic studies in a better way. Therefore, the RF classifier over the abundances may not always be the optimal approach employed for the metagenomic predictions. Furthermore, there is limited literature [87],[133] on involving phylogenetic distances directly in computational modelling of metagenomic functions. Therefore, PAAM-ML has been developed to address these limitations.

The PAAM-ML has been designed to incorporate prior knowledge of phylogeny into abundance counts of the microbiome with a novel weighting scheme, facilitating the construction of a new matrix structure (PAAM). A contribution in this thesis has been the incorporation of both quantitative (abundances) and qualitative (biological domain knowledge of phylogenetic relationships) aspects into the construction of new feature space for the prediction of metagenomic phenotypes. A novel framework was employed to determine metagenomic functions involving (i) construction of new integrative feature space, (ii) application of feature ranking strategies over the engineered space, and; (iii) application of the classical predictive models over the engineered feature space. By employing this approach, an improvement in predictive performance was observed in comparison to previously employed approaches [93],[98],[128]. The key benefit is that, by providing both phylogeny and abundance quantitative profiles at the data engineering (pre-
Contribution to the knowledge

processing) time, the subsequent ML classifiers microbial features gain insights from multiple granularity levels of taxonomy for their functional classification, taking advantage of natural characteristics of the microbial community. The proposed approach considered the classification task as the joint learning on tree leaves and a subset of intermediate nodes of the phylogenetic tree for a metagenomic inference. Henceforth, it is benefitted from learning informative features by traversing nodes from leaf-level to the root on a populated phylogenetic tree. However, embedding the phylogenetic tree in abundance profiles created a high-dimensional feature space due to which feature selection or ranking became important as a precursor step to an ML classifier. Using metagenomic data sources of DS1, DS2 and DS4, it has been shown that the proposed framework provided better performance than the conventional models [87],[133]. It was discovered that the structure of the PAAM has an impact on the predictive performance of the ML classifier. For the predictive task of inferring functional roles of soil and the human microbiome, the feature engineering over PAAM indicated promising results.

The construction of PAAM (a 2D matrix) along the framework would help researchers in this field to achieve a better understanding of the microbial feature space. It represents a promising way to represent metagenomic data in an integrative way. The generation of PAAM has been developed as a Perl-based application embedded in a Web-Page, which has been made available to researchers at the GitHub repository (https://github.com/jyotsnawassan/ Functional-Metagenomics). PAAM generator based on the tree parser would facilitate functional metagenomic applications. The interface of PAAM construction provides options to researchers for selecting an OTU or ASV table,
and the corresponding phylogenetic tree and to determine the structure of the PAAM, integrating both inputs. The resultant PAAM generated can be downloaded and could be evaluated with different computational models by the research fraternity for any futuristic functional metagenomic studies.

8.2.3 Phylogeny-aware RF model-based metagenomic predictions [Chapter 6]

A preliminary study on tuning the most-popular ML classifier of RF with phylogeny for the inference of metagenomic functions in soil and the human microbiome has been performed in this thesis. The study followed Approach 3, as proposed in this thesis. Initial research in [87] suggested the regularization of the LR model with phylogeny in predicting metagenomic functions. However, relatively little research has been performed within this area of incorporating biological domain knowledge of phylogeny in a supervised learner itself [85],[87] rather than at the precursor stage of data pre-processing. This thesis proposed the novel application of tuning RF model with phylogeny for functional metagenomic predictions. Results reported in this thesis have suggested that a phylogeny-aware supervised ML technique of RF can be successfully applied to the problem area of metagenomic functional predictions. The aim is to expand integrative research at the supervised learning stage. A traditional RF model constructs a constituting decision tree by selecting a random subset of the predictors from the entire feature set to determine the best split markers based on a Gini Index splitting criterion [88]. These predictors are usually chosen as equivalent to $\sqrt{\text{Number of input features}}$. However, this
thesis worked on an idea of random selection of microbial features from phylogeny-based created clusters equivalent to $\sqrt{\text{Number of input features}}$ whereby randomly choosing features from each cluster to model a decision tree by covering maximal phylogenetic diversity with minimum redundancy. The idea gains insights from the assumption that the similarity patterns are present between the microbial features based on their phylogeny and similar features tend to behave similarly to functions. The proposed method improved over the phylogeny-driven state-of-the-art of MetaPhyl and PhILR. The aim is to expand research presented in this thesis and explore the incorporation of phylogeny in ML modelling.

8.2.4 Application of data analysis models for the prediction of cattle microbiome functions in MetaPlat project [Chapter 7]

Biological information related to cattle rumen microbiome was extracted as part of the MetaPlat project [144]. Various predictive models were explored and employed over the sampled cattle rumen microbiome use cases under MetaPlat [144], following Approach 1, 2, and 3. The work presented in this contribution is based on published work [78], [121],[266].

The predominant biological OTU features obtained from WFS methods proved highly accurate in the classification of diet phenotype (supplemented with oil or nitrate) associated with 40 cattle rumen samples. However, due to the time-intensive nature of WFS methods, these are not much scalable. Another use case study as part of MetaPlat focussed on 80 samples that were treated with oil, nitrate, or oil-nitrate (combined) diet. The predictive modelling over these
80 samples involved integrative study with (i) application of PAAM-ML; (ii) a novel method developed based on Relief-based [90] ranking of microbial features based on the phylogenetic interaction index (termed as Phylogeny-PINA Relief) and; (iii) the Phylogeny-RF. These three approaches applied over cattle microbiome corresponds to the integration of phylogeny in computational modelling at three different levels: (i) data pre-processing, (ii) feature engineering, and (iii) ML classifier, respectively. These approaches provided promising results in comparison to predictive modelling over the abundance profiles of cattle microbiome.

With gradual developments in the MetaPlat project, data samples related to DEX treatment and forage diet were collected. With the phylogenetic modelling, it was observed that diet phenotype is significantly associated with cattle microbiome. However, no such significant relationship of cattle microbiome was observed with the DEX phenotype. The state-of-the-art statistical pipeline of aMiSPU [149] (based on phylogeny) reported a significant relationship of cattle microbiome composition (p-value < 0.01) with diet supplements confirming the results presented in this thesis. Furthermore, PAAM-ML and Phylogeny-RF were applied to analyse cattle microbiome regarding DEX and Diet. Some important microbial features were identified to characterise microbial samples into related phenotypes. Overall, *Bacteroidetes*, *Firmicutes*, *Actinobacteria*, and *Proteobacteria* served as important microbiome phylum in all cattle ruminants for differentiating the phenotype of supplemented diets. The case study indicates phylogenetic estimates are capable of improving the classification performance over cattle rumen microbiome.
In summary, the thesis proposed integrative computational methods based on predictive modelling with ML for analysing and evaluating microbiome and linking it to the microbial functions. These methods have been applied to a variety of metagenomic environments producing promising results.

### 8.2.5 Development of contributory tools

The research has provided solutions for representation and classification of metagenomic data in an integrative (knowledge-driven) way. These are made available on a public repository domain of GitHub (linked to https://github.com/jyotsnawassan/Functional-Metagenomics) and are listed below.

- A contributory tool of PAAM-generator involving phylogenetic tree parsing as its integral part.
- The newly developed classifier of ‘Phylogeny-RF’ for classifying metagenomic data.
- A script for phylogenetic ranking of the ancestral nodes in the integrative microbial feature space (approach termed as ‘Phylogeny-PINA Relief’).

### 8.3 Limitations and future work

Next-generation sequencing (NGS)-based-metagenomics analysis has attracted great attention due to its capability to generate microbial sequences at a great
pace. The computational inference of phenotypic functions of the microbiome using ML models is still an active area of the research area.

Research in inferring metagenomic functions can potentially be employed in the areas including:

(1) developing scalable or integrative models over microbial features;

(2) identification of groups of microbial taxa to explore and model their relationships in functional analysis;

(3) defining and modelling co-occurrence networks in the microbiome;

(4) investigating diversity patterns in the microbiome and analysing them for any kind of association with phenotypes;

(5) visualisation of the microbiome with different ordination techniques to study their compositions and their relationship with the host environment.

This thesis has contributed computational approaches for (1) and (2) applied over human, cattle, and soil microbiome use cases. The other areas of (3), (4) and (5) could be explored in future as an extension of schemes presented in this thesis. The limitations and potential future work are outlined below.

### 8.3.1 Limitations

The choice of computational methods in functional metagenomics depends on the factors, such as metagenomic dataset, the number of samples available, the nature of the phenotype classes, and the aim, scope, and resources available for the functional analysis of the microbiome. The main problem discussed in this thesis is to improve modelling over the high-dimensional and integrative
Limitations and future work

microbial feature space for determining functional predictions. However, the research did not fully address the issue of handling incomplete information generated from metagenomic pipelines and the challenge for microbiome compositional data analysis. Following subsections discuss some of the limitations presented by integrative ML methods developed within this thesis.

8.3.1.1 Compositional nature of metagenomic data that was not addressed by this research

The main problem discussed in this thesis is to improve the functional classification of metagenomes by developing more efficient and integrative methods. The proposed models in this thesis have been designed to process microbiome profiles described by the numerical quantities in feature vectors. They were not designed taking into the compositional characteristics of metagenomes. The methods developed did not consider the unique compositional variation of metagenomes, but rather considered majorly the structural variation along with the absolute abundance counts and the phylogenetic distances except over DS3 and DS5 where the relative abundance counts were considered as inputs. Absolute abundance counts were obtained from DS1, DS2, DS4, DS6 and DS7. The authors in [280] highlighted the limitation of using relative abundance counts. Also, there could be a risk if compositional metagenomic data are analysed inappropriately.
8.3.1.2 Constructing signatures at multiple levels in an integrative way in an ML model

Incorporating the phylogenetic tree information node-by-node in an ML model has been introduced in [85], using a tree-guided fused LASSO (Pen-LR). However, incorporating such information in a tree-based classifier is lacking in the literature. An attempt has been made to incorporate tree-based similarity between leaf-level nodes by traversing through a phylogenetic tree in an RF classifier in this thesis. But the features are still defined as individual OTUs/ASVs (the leaves of the phylogenetic tree) in the phylogeny-driven modelling of RF classifier. Nevertheless, features space modelling has been characterised by different levels of varying phylogenetic depth in this thesis with PAAM-ML, but similar integration in the Phylogeny-RF classifier is missing. In future, instead of defining features as tree leaves (OTUs/ASVs), modelling subset of intermediate nodes of the phylogenetic tree in a tree-based classifier itself (such as RF) may lead to better classification.

8.3.1.3 Some other limitations embraced by the methods developed in this thesis.

- All methods developed in this thesis were implemented with the different tuning of parameters of respective ML models aiming to attain better predictive performance. However, there is no standard way to define optimal parameter settings a priori. The combination of best parameters is the constraint exhibited by the developed models.
• The analysis of the quantitative abundances of the human microbiome to associate with the phenotype of diet and IBD states (in Chapter 4) did not incorporate the phylogenetic measure. Also, the scalable EM methods applied to such microbiome could be extended with advances in EM methods such as recently proposed in [281].

• The comparative results presented in PAAM-ML related research focussed on the predictive performance obtained from the application of the popular ML methods. It may be interesting to extend this comparative analysis with emerging ML models of CDF [282] which have not been explored in the current thesis. PAAM-ML models for studying dynamics have been built using the phylogenetic distances and abundances of microbiome data directly without to account for any use of distance-based similarity/dissimilarity between microbial taxa. PAAM could be extended to calculate such similarity and dissimilarity metric based on the microbial profiles and thereafter using the similarity structure in functional metagenomic analysis, and; in ordination to study the diversity of compositional patterns. Also, regularization of feature space with phylogenetic weights as introduced in Chapter 5, may be viewed as one of the heuristics; future work is likely possible to utilise additional weighting schemes to engineering the feature space and analysis.

• Phylogeny-RF particularly deals with integrating the biological knowledge in a commonly used method of RF in microbiome studies. Exhaustive tuning of the hyper-parameters of Phylogeny-RF is possible to explore the possible improvements in the predictive performance. In the future, the thesis could be extended in terms of comparing the developed framework to other
advanced phylogeny structure-driven supervised learning methods (such as incorporating internal nodes and leaf nodes of a phylogenetic tree in the tree-based classifier of RF and/or XGBoost) for an extensive analysis.

- A preliminary study has been presented in this thesis to infer diet-related cattle functions using the newly developed Phylogeny-PINA Relief approach (as part of the methods developed for the MetaPlat project). PINA served as a potentially useful phylogenetic index in the functional metagenomic analysis. However, the calculation of PINA similarity between microbial taxa posed as time-intensive in nature as it calculates the correlation between each pair of OTUs or ASVs. With the development of large-scale high-dimensional metagenomes in the future, there is a need to enhance the proposed approach to make it scalable.

- Application of visualisation-based methods could further increase the readability of the classification results in an emphasising way for human users.

The current approaches could be extended to be applied over a variety of other datasets with large-scale high-dimensional datasets and with different phenotypes. Also, the approaches in this thesis have been applied to 16S rRNA sequences; however, they can be applied to shotgun whole metagenomic sequences in the future.

8.3.2 Future work

Integrative analysis in the current thesis is centred around relationships between microorganisms that are discovered through phylogenetic inference model of
the evolution of microbial genes. This has proved useful in understanding the relationship between genetic and functional traits of microbial communities present in an environment. The two phylogeny-based inference models designed in current research are Approach 2 (Figure 3.5b) and Approach 3 (Figure 3.5c). Approach 2 combined all evolutionary lineages from a phylogenetic tree (i.e. from leaf to root node) into the abundance profiles of microbial genes to heuristically create a feature space having integrative information and subsequently, an ML classifier has been applied over the integrative space. On the other hand, in Approach 3, phylogenetic diversity (pair-wise evolutionary distances) between leaf level nodes (OTUs) was used to regularize an RF classifier. There are possible extensions to combining these Approaches. In future, the prospective study could be designed involving the development of novel approaches that account for regularizing an ML classifier with level-by-level evolutionary information obtained from a phylogenetic tree, combining Approach 2 and Approach 3. By combining these approaches (as shown in Figure 8.3), a new similarity model incorporating pairwise distances not just between the leaf-level OTUs but also including the internal nodes could be designed. This is possible by reading the evolutionary distances (i.e. branch lengths) from a phylogenetic tree. Phylogenetic diversity thereby could be calculated from groups of microbial signatures created at all levels of the taxonomy from root to leaf nodes. Intuitively, modelling with phylogeny-based similarity (from root to leaf nodes) could lead to more accurate and interpretable results by considering intermediate nodes of the phylogenetic tree as candidate microbial features. The potential advantage of using branch lengths between various nodes of the tree is to analyse in a more biologically relevant way.
Such analysis could also utilize the idea of the fusion of similar OTUs along the phylogenetic tree under a very natural assumption that closely related taxa in a tree may function similarly. A tree-guided variable fusion can also be extended to be input to other multivariate analysis methods, such as calculating principal components and applying ML classifier subsequently for estimating the functional profiles of metagenomes.

Figure 8.3 A proposed path for combining Approaches 2 and 3.

Furthermore, based on the limitations presented in the previous section (8.3.1), plausible future work on integrative modelling of functional metagenomes is discussed as follows.
8.3.2.1 Extension of study based on deep ML models over phylogeny and abundance-aware analysis

An exploratory direction to work upon is the use of recently developed methods such as Cascade Forest or Deep Forest over PAAM [162]. These methods are being adapted in different data across different domains to gain performance improvement over large-scale high-dimensional datasets. In future, these deep-forest based models could be explored in an integrative way over the large-scale high-dimensional metagenomic datasets with a phylogenetic tree. It would be interesting to explore whether integrative analysis with Deep ML models produce superior results?

8.3.2.2 An application of dimensionality reduction and visualisation techniques to an enhanced study of the microbiome

Another interesting research direction of work could be driven by distance-based visualisation analysis with techniques such as Principal coordinate analysis (PCoA), Principal component analysis(PCA), T-distributed Stochastic Neighbour Embedding (t-SNE) [37], [283] to demonstrate and analysis the separation of phenotypic classes related to microbial compositions and their similarity patterns based on different phylogenetic and abundance-driven distances such as obtained from PAAM, PINA or UniFrac. Various visualisation strategies could be explored to study the distribution of microbiome and its relation to metagenomic functions in a user-friendly way.

The analysis of the integration of metagenomics data and network-based approaches is also an active area of research [77],[284]. The evolution of the
microbiome could be explored to see if it could influence the microbial structure and the associated networks, and the functional responses. Future work could also be performed in designing computational methods to model the microbiome networks in metagenomic analysis, based on domain knowledge of both qualitative (biological) and the quantitative (abundance profiles) such as in PAAM.

8.3.2.3 Incorporation of node-by-node information in a tree-based ML modelling

In the current thesis, an attempt has been made to incorporate modelling over leaf-level OTUs/ASVs in an RF model to do integral feature selection in “Phylogeny-RF”. However, in future, adding node-by-node phylogenetic information to RF could better support decision making strategy in RF modelling by analysing whether internal nodes serve as better splitting candidate features in the constituting decision trees of an RF model. Modelling of Gini Index involved in RF with an intention that a node closer to the root in a phylogenetic tree could bring more phylogenetic effect on regularization of Gini index rather than just leaf-level OTUs/ASVs could be explored for further potential analysis.

8.3.2.4 Compositional effect of metagenomes

A potential direction of future work is to normalise and transform phylogeny and abundance-aware feature space which could consider the compositional nature of microbiome into account. Based on compositional analysis techniques
such as suggested in [285], these can be explored further to develop an extension of approaches presented in this thesis to generating a feature space with much information on inter-relationships of the microbial features in the environmental sample.

8.4 Conclusions

Microbes are abundantly present in almost every ecological niche. Due to advancements in NGS techniques, Metagenomics (known as a study of microbes) has been developed to study and analyse the microbial communities present in different ecosystems, based on a non-cultured approach. It is important to study microbes to know “who they are,” “how they are structured,” and “what they are doing” in an ecosystem. The results of the metagenomics study mainly rely on computational tools and techniques that can extract useful information and analyse metagenomic datasets. Based on ML, several computational methods are emerging to infer the functional roles of microbes (functional metagenomics). It is an active area of research to diminish the gap between the fast generation of metagenomic data via the use of NGS techniques and its analysis. This thesis has contributed to the development of new computational methods to support the identification of relevant microbial features in functional metagenomics through the integration of qualitative (biological knowledge), and quantitative (abundance-driven) approaches. This thesis has focused on developing these models for the prediction of metagenomic functions in humans, cattle, and soil microbiome. This has involved the development of various classification methods employed for the inference of metagenomic functions using either or both (a) abundance-driven
profiles and (b) integrative phylogenetic profiles. The development and novel application of phylogeny-driven frameworks of PAAM-ML, Phylogeny-RF, and Phylogeny-PINA Relief were proposed. By considering the qualitative and quantitative knowledge-based dependencies between the microbial features, an improvement in prediction performance was achieved in comparison to the classical classifiers when inferring metagenomic functions. Moreover, EM methods such as XGBoost proved useful over the abundance-driven quantitative profiles. As part of this research, diverse data sources are selected, pre-processed, and integrated using ML tools or techniques to infer metagenomic functions.

In this thesis, four major data analysis scripts have been developed and implemented for the prediction of metagenomic functions: (1) R-based script to classify high-dimensional metagenomes; (2) PAAM construction is a Perl-based GUI script embedded in Webpage. The data obtained from this script which could further be used by any tool supporting ML classification; (3) Another script was generated in Perl to implement PINA-based feature modelling of feature space for cattle rumen microbiome; (4) Furthermore, the Phylogeny-RF classifier has been extended and developed as a Python [286] script to implement RF gaining information from clusters generated using $k$-means over phylogenetic distance metric implemented in the $R$ platform [181],[250] to infer the metagenomic functions. The Perl scripts are available at a common repository GitHub repository linked to https://github.com/jyotsnawassan/Functional-Metagenomics.
References


References


[145] Q. Zhang, M. Zhang, T. Chen, Z. Sun, Y. Ma and B. Yu, "Recent advances in convolutional neural network acceleration",


References


