A Multiomics Investigation of Endocrine Resistance in Breast Cancer

by

Michael Francis O'Dea

A thesis submitted as partial fulfilment

of the requirement for the degree of

Bachelor of Advanced Science (Honours)

School of Biotechnology and Biomolecular Sciences

University of New South Wales

May, 2022

ORIGINALITY STATEMENT

'I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.'

Signed: MDen

Date: 10/5/22

ABSTRACT

Cancer is one of the leading causes of premature mortality worldwide. The majority of breast cancers are Estrogen Receptor positive (ER+) and are thus candidates for treatment using endocrine therapy. Unfortunately, 30-50% of people exhibit either intrinsic or acquired endocrine resistance, with the mechanisms behind this still largely unknown.

In this thesis I set out to explore endocrine resistance through a multiomics lens. There were three core parts to my approach. First, a custom proof-of-concept multiomics model was designed and trained on post-treatment DNA methylation, oral and stool metagenomics data to try and predict patient's endocrine response. This model performed remarkably well, achieving an AUC of 0.99 and an accuracy of 95%. On an individual classifier level, methylation data proved to be the most informative, with gut microbiota the worst individual performer. This was surprising given the well documented link between gut microbiota and estrogen through the estrobolome.

Secondly, models were constructed on the pre-treatment metagenomics data alongside age matched control samples to try and identify biomarkers for ER+ breast cancer. While these models performed only moderately well, they identified a number of bacteria that were differentially abundant between cancer and control with the majority being elevated in control patients.

Finally, a clinical models were constructed on a minimal set of pre-treatment features to attempt to identify endocrine resistance early. This model performed impressively as well, reporting an AUC of 0.97 and an accuracy of 0.95 using an SVM trained on just five prognostic markers. The power of these features to predict long-term outcomes was also investigated, with the same five markers predicting breast cancer death with an AUC of 0.85 and an accuracy of 0.79

TABLE OF CONTENTS

Page No.

TABLE OF TABLES

TABLE OF FIGURES

ACKNOWLEDGMENTS

Firstly, I want to thank my Honours supervisor Dr Fatemeh Vafaee for giving me the opportunity to pursue this uniquely challenging project in her lab. While not always easy, it's been very rewarding getting to be hands on with such a variety of data and I've learnt a lot from the experience. I'm so thankful for your advice and support.

I'd also like to thank my collaborators A/Prof. Ewan Millar and Dr David Gallego-Ortega. The biological expertise you both brought to this project was invaluable and I'm grateful to have had the chance to work on your study. I want to especially thank Ewan for his enviable email ethic, I've never met someone who responds more quickly and helpfully to emails than you. I aspire to your levels of communication prowess.

Thanks also to Braydon Meyer for his work pre-processing the methylation data and Azadeh Safarchi for her advice on working with metagenomics data. Thanks to the whole Vafaee lab for your support throughout my project, especially Seb and James for sharing this tumultuous honours year with me. Thanks to the Waters lab for treating me like an honorary member of the lab and letting me join your trivia team, with special thanks to my good friend Ashley Milton for all your guidance and support through the honours process.

Thank you to my amazing parents for everything you do. I couldn't have gotten to this point without your ever-present love and support. I hope I've made you proud.

Lastly, I'd like to thank my wonderful partner Sam. You have been my rock throughout this project and have constantly gone above and beyond to support me. Thank you for all the walks, hugs and smiles. Thank you for everything thing you do. I could not have done this without you. And I promise that you will never, ever, have to listen to me talk about my project ever again.

ABBREVIATIONS

ER+	Estrogen Receptor Positive
ML	Machine Learning
CpG	Cytosine and guanine nucleotides linked by a phosphate group
TILs	Tumour-infiltrating Lymphocytes
SNP	Single Nucleotide Polymorphism
IQR	Inter-Quartile Range
PCA	Principal Component Analysis
RF	Random Forest
SVM	Support Vector Machine
LR	Logistic Regression
OOB	Out of Bag
XGBoost	eXtreme Gradient Boosting
LGBM	Light Gradient Boosting Machine
LEfSe	Linear discriminant analysis Effect Size
SMOTE	Synthetic Minority Over-Sampling Technique
AUC	Area Under the Curve
ER	Estrogen Receptor
PR	Progesterone Receptor
HER2	Human Epidermal Growth factor 2
GUS	β-glucuronidase
gmGUS	gut microbial β-glucuronidase

1 INTRODUCTION

1.1 What is breast cancer?

Cancer has emerged as the defining medical challenge of the 21st century, with it now serving as the leading cause of premature death in 57 countries including Australia (1). This number will only continue to increase in coming years as cardiovascular mortality continues to decline, with the total number of new cases globally projected to increase more than 60% over the next two decades, from 18.1 million new cases in 2018 to approximately 29.4 million by the year 2040 (2). In 2020, female breast cancer surpassed lung cancer to become the most commonly diagnosed cancer worldwide, with an estimated 2.3 million new cases that year alongside 685,000 deaths (3). Despite this, breast cancer's mortality rate is gradually being reduced thanks to progress in both early diagnosis and treatment options, although there is still a lot of room for improvement (4).

Currently, recommended treatment options for breast cancer patients are dependent on the patients cancer subtype (5). In 2011 the St. Gallen International Expert Consensus proposed a classification system for breast cancer that divided cases into five subgroups depending on their levels of three hormone receptors as well as the nuclear protein Ki67 (Table 1) (6).

Table 1 ***Clinical definitions of breast cancer subtypes***

Breast cancer subtypes	ER	PR	HER2	Ki67
Luminal A	*Positive*	*High*	*Negative*	*Low*
Luminal B	*Positive*	*Low*	*Negative*	*High*
Luminal B-like	*Positive*	*Any*	*Positive*	*Any*
HER2 Positive	*Negative*	*Negative*	*Positive*	*Any*
Triple Negative	*Negative*	*Negative*	*Negative*	*Any*

The hormone receptors of interest are estrogen (ER), progesterone (PR) and human epidermal growth factor 2 (HER2). HER2 positive tumour cells are particularly aggressive due to possessing more than one copy of HER2, a gene which is responsible for breast cell growth (7). Ki67 has been the focus of intense study over the last two decades due to its own association with cell proliferation (8), and has been shown to be a valuable prognostic marker in a range of cancers including breast, lung and prostate cancer (9-11). Meanwhile, tumour cells which are positive for ER are able to use estrogen to fuel tumour growth (12). These cancers are collectively known as luminal or estrogen positive (ER+) cancers and comprise around 80% of all cases (13).

1.2 Endocrine therapy

Endocrine therapy is a class of treatment which prevents breast tumours from utilising estrogen for growth and as such is a primary therapeutic option for ER+ cancers. It is most commonly used in combination with chemotherapy to reduce the risk of long term recurrence, or as a means to shrink the tumour prior to surgical removal (14). However, it can also be used on its own in cases where surgery is not an option. Tamoxifen is one of the most utilised endocrine drugs and acts by inhibiting estrogen binding to the tumour's receptors (15). The other main class of endocrine drugs are aromatase inhibitors (AIs) such as anastrozole and letrozole. Instead of preventing binding to the cancer's receptors, these drugs attempt to lower the levels of estrogen in circulation by inhibiting production of the aromatase enzyme (16). Aromatase is responsible for converting androgen into estrogen, which is the primary means of estrogen production post-menopause. Unlike tamoxifen, which can be used in all ER+ cancers, aromatase inhibitors are only effective in post-menopausal women or when taken in combination with ovarian

suppression drugs (17). Importantly, AIs have been shown to be more effective than tamoxifen as both adjuvant (18) and neoadjuvant therapy (19), as well as in advanced cases of disease (20).

1.3 The problem of endocrine resistance

While aromatase inhibitors are an effective treatment option for many post-menopausal patients, 30-50% of cases will not respond to endocrine therapy (21). For half of these patients the resistance is intrinsic, while the other half acquire resistance over the course of the drugs administration (21). The mechanisms behind both intrinsic and acquired endocrine resistance are currently unknown. Early prediction of endocrine resistance is highly desirable as it would enable personalised treatment plans. For example, patients predicted as poor responders could instead be administered an alternate therapy such as the CDK4 inhibitor palbociclib, which works by interrupting breast cancer cell division (22). Currently, the main biomarker used to assess patient responsiveness is Ki67, with the percentage reduction between pre and post-treatment being the strongest predictor of intrinsic endocrine resistance (23-25). Ki67 has also been shown to be indicative of long-term patient outcomes such as recurrence free survival, which is associated with low levels of Ki67 post-treatment (26).

1.4 Existing 'omics' breast cancer research

As the most common cancer worldwide, a large number of studies have been performed investigating associations of various omics datatypes, with some datatypes receiving much more focus than others (27-30).

1.4.1 Gut microbiome and the 'estrobolome'

The gut microbiome is the omics datatype with the clearest link to estrogen abundance, with the two interacting through what was coined the 'estrobolome' (31) by

Plottel and Blaser in 2011. The estrobolome refers to the group of bacterial genes that are capable of metabolising estrogen, with the primary enzyme responsible for this being β-glucuronidase. One leading theory as to why some patients respond poorly to endocrine therapy is through estrogen reactivation by gut microbial β-glucuronidase (gmGUS) (32). The suggested pathway through which reactivation occurs is that estrogen glucuronides produced by estrogen's hepatic phase II metabolism are hydrolysed by gmGUS in the gut via bile excretion. Thus, aromatase inhibitors, which work by lowering the amount of free estrogen in the body, are less effective if estrogen is being refreshed by gmGUS. While promising, investigation of this relationship is still in the earliest stages.

1.4.2 Oral microbiota: the missing link?

Very little research has been conducted into potential links between the oral microbiome and breast cancer. The most substantial connection identified so far is the significant association of periodontitis with an increased risk of breast cancer (33), one of several cancers periodontal disease is associated with (34,35). It's theorised that this association may be caused by the chronic inflammation which characterises periodontitis, as inflammation has also been implicated as a primary driver of breast cancer (36). Additionally, while not directly related to breast cancer, there has been a recent rise in papers highlighting the influence of oral microbes on distal cancers such as colorectal cancer (37), suggesting that oral bacteria may exert similar indirect influence on breast tumour cells.

1.5 Tumour-infiltrating lymphocytes (TILs) and other clinical markers

While Ki67 is the primary prognostic marker used to determine response to endocrine therapy, it is far from the only prognostic factor undergoing research.

Tumour-infiltrating lymphocytes (TILs) are viewed as surrogate markers of adaptive immune response, with it theorised that the intensity of tumour immune response influences the effectiveness of cancer therapies. However, the clinical importance of TILs seems to vary among breast cancer subtypes. In triple negative breast cancer, high levels of TILs are associated with improved disease-free and overall survival rates (38). However, in the case of aromatase inhibitors, specifically letrozole administration, higher TILs are significantly associated with a poor treatment response (39).

1.6 Thesis aims and hypothesis

The central aim of this thesis is to explore innate endocrine resistance in ER+ breast cancer, with a focus on building predictive models and identifying potential biomarkers of patient response. To achieve this aim, I will complete the following:

1. Design and build a modular multiomics model for classifying endocrine response that integrates the predictions of individual classifiers for each of DNA methylation, oral and stool metagenomics datatypes.

My hypothesis is that, due to the heterogenous nature of breast cancer, such a multiomics approach will outperform any individual classifier. When considering the datatypes individually, I predict that the gut microbiome will have the greatest importance given the well-established link between gut microbiota and estrogen metabolism described as the ‘estrobolome’ (Section 1.4.1)

2. Construct classifiers of pre-treatment cancer metagenomics data against age matched controls to investigate microbial biomarkers of ER+ breast cancer

While extensive research has been done into biomarkers of breast cancer in the gut, relatively little research has looked on a subtype specific level at ER+ breast cancer or at oral microbes. I theorised that, while little is currently known about it, the oral microbiome was likely to have at least a few taxa differentially abundant between cancer and control patients.

3. Create a lightweight pre-treatment clinical model which can predict patient response early and offer alternative treatments to poor responders, improving overall patient outcomes.

While change in Ki67 is a good prognostic for endocrine resistance post therapy, predicting patient response prior to treatment is highly desirable as it allows patients to avoid undergoing unnecessary treatments. TILs have recently emerged as a prognostic biomarker in other subtypes of cancer. I hypothesise that a clinical model which combined TILs and Ki67 alongside other pre-treatment markers could predict endocrine resistance just as good as change in Ki67 can.

2 MATERIALS AND METHODS

The core work of this thesis, the multiomics model and its constituent classifiers, is described in Section 2.2.1, with an accompanying overview of the model's design presented in Figure 2. Comparisons of the pre-treatment cancer metagenomics samples against controls, aimed at identifying subtype-specific biomarkers, are described in Section 2.2.2. Models built on clinical pre-treatment data to enable early identification of endocrine resistance as well as long-term patient outcomes are described in Section 2.2.3.

Due to the high degree of methodological overlap between many of these models, individual techniques are described in detail at their first chronological occurrence, with the relevant section referenced in every subsequent use. All code referenced can be found in the public GitHub located here: https://github.com/michaelodear/Endocrine-resistance-breast-cancer-Honours-project.

2.1 Materials

2.1.1 St George neoadjuvant endocrine study

Most of the data analysed in this thesis came from a cross-institutional research project led by St George hospital with A/Prof. Ewan Millar as its lead investigator. This project aimed to look for biomarkers of therapeutic response in neoadjuvant treated breast cancer across a range of 'omics' datatypes. While the project recruited >200 post-menopausal women overall, this thesis focuses just on those with ER+ breast cancer. Patient biopsies were collected at St George hospital immediately prior to endocrine therapy (patients received either Anastrozole or Letrozole) as well as 1-2 weeks post treatment. Ki67 levels were assessed pre and post treatment using image analysis and

automated scoring, with patients' endocrine response type classified by their percentage reduction in Ki67 using the thresholds described in Gellert et al. (40).

These samples were used to generate the following five datatypes: histopathological images, spatial transcriptomics, DNA methylation, oral metagenomics, and stool metagenomics. Pre and post-treatment data was collected for all datatypes except DNA methylation, for which the pre-treatment core biopsies were too small to use. The images were collected at St George and are currently being explored by the Meijering lab in the UNSW School of Computer Science, while the spatial transcriptomics data was only finished at the Garvan Institute in late November. As such, both datatypes were outside the scope of this project. The methylation data was extracted at the Victorian Comprehensive Cancer Centre before undergoing preliminary analysis at the Garvan Institute, while the metagenomics DNA was extracted at St George and sequenced by the Ramaciotti Centre for Genomics. These three primary datatypes were used to build the multiomics model and are summarised in Table 2.

Table 2. ***Multiomics model datatypes***.

Patient numbers presented are for post-treatment data only. Features is the number and type of columns in the input data after pre-processing.

Data Type	Sequencing technology	Total patients	Good Responders	Poor Responders	Features
DNA methylation	*Illumina EPIC Microarray*	*26*	*13*	*13*	*802001 CpG sites*
Oral metagenomics	*Illumina NovaSeq*	*61*	*39*	*22*	*463 taxa*
Stool metagenomics	*Illumina NovaSeq*	*60*	*39*	*21*	*627 taxa*
Matched patients	-	***20***	***13***	***7***	-

In addition to the metagenomics cancer samples, 18 age matched controls were available for the stool metagenomics data, with oral metagenomics also available for nine of these patients.

The methylation data imposed two major limitations on the multiomics model. Firstly, because only post-treatment data could be obtained for it, the integrated model had to be built using post-treatment samples only. The other unfortunate limitation of the methylation data was its comparatively small sample size. Only 26 samples were available total, and only 20 of these patients had corresponding oral and stool metagenomics data due to delays in receiving ethics approval. The multiomics approach taken enabled all of the available data to be used in training the models, but only the predictions on the 20 patients with all datatypes available could be used to make the final integrated prediction. These patients are referred to as the matched patient cohort for the rest of this thesis.

2.1.2 Clinical features of matched patients

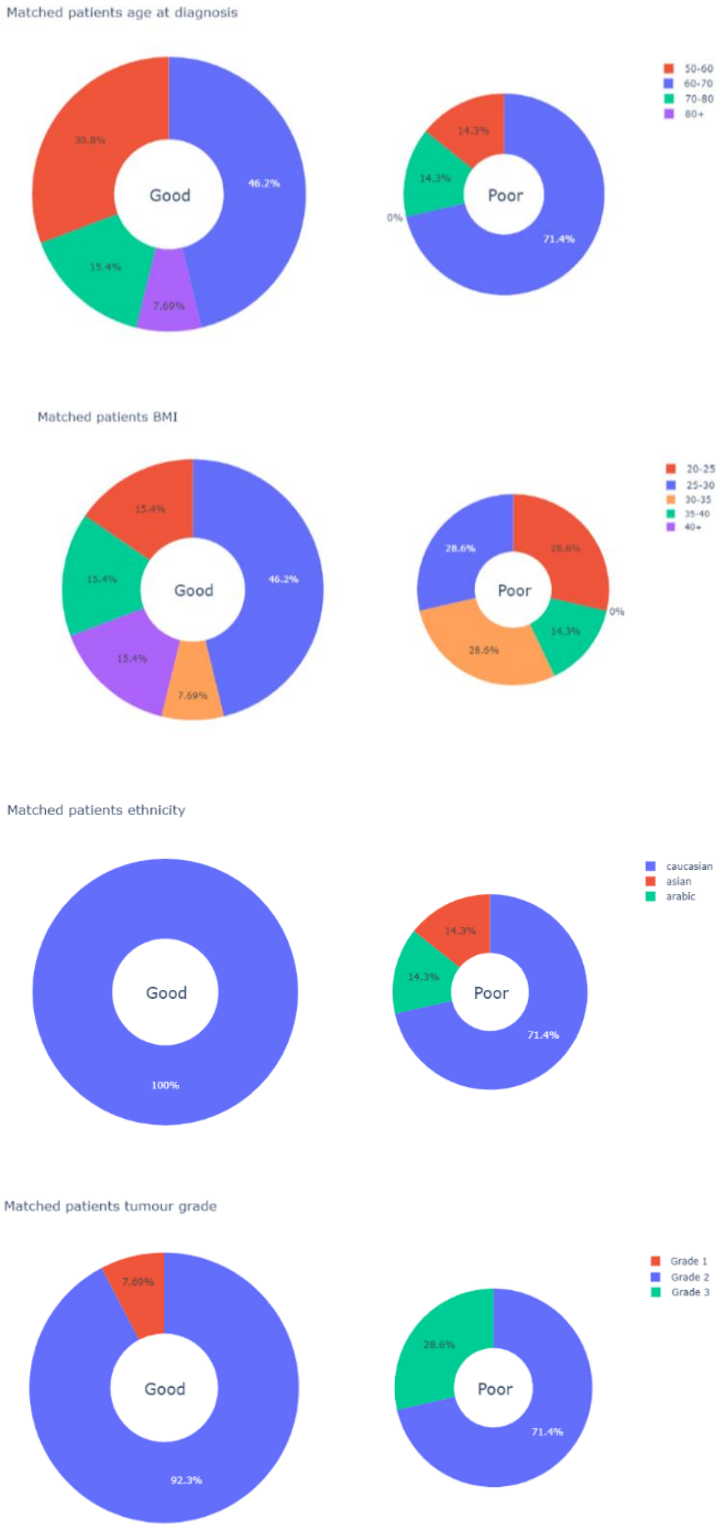

Figure 1. ***Selected clinical features for matched patient cohort.***

Patients are separated into 'good' and 'poor' responders. The size of each pie chart corresponds to the relative number of patients in that response class.

A selection of clinical features of the matched patient cohort are summarised in Figure 1, separated by responder group. While both good and poor responders have relatively similar age and BMI distributions, there are notable differences in their ethnicity and tumour grade. The matched patient cohort is overwhelmingly Caucasian, with the only non-white patients being present in the poor responder group. Meanwhile, although most tumours in both groups are Grade 2, the only low grade tumour belongs to a good responder while the only high grade tumours are found in poor responders.

In addition to these clinical factors and patient Ki67 levels, St George clinicians recorded a number of other clinical features for the patient cohort, including their BMI, tumour size and lymph nodes (both number of nodes positive for breast cancer as well as total present). Other clinical features include patient TILs % and levels of estrogen and progesterone receptors pre and post treatment.

2.1.3 St George Breast BOOST study clinical data

The St George Breast Boost study originally ran from 1998-2003 and involved the collection of samples from 688 patients with breast cancer. These patients were followed up for the next 16 years with survival data most recently updated in 2019. Findings based on this patient cohort have been published a number of times, most recently showing that TILs in breast cancer predict local failure and overall survival (41-43). Due to the length of follow-up, this dataset was used to determine how well ML models could predict long term patient outcomes based on purely pre-treatment features in breast cancer.

2.2 Methods

2.2.1 Multiomics model

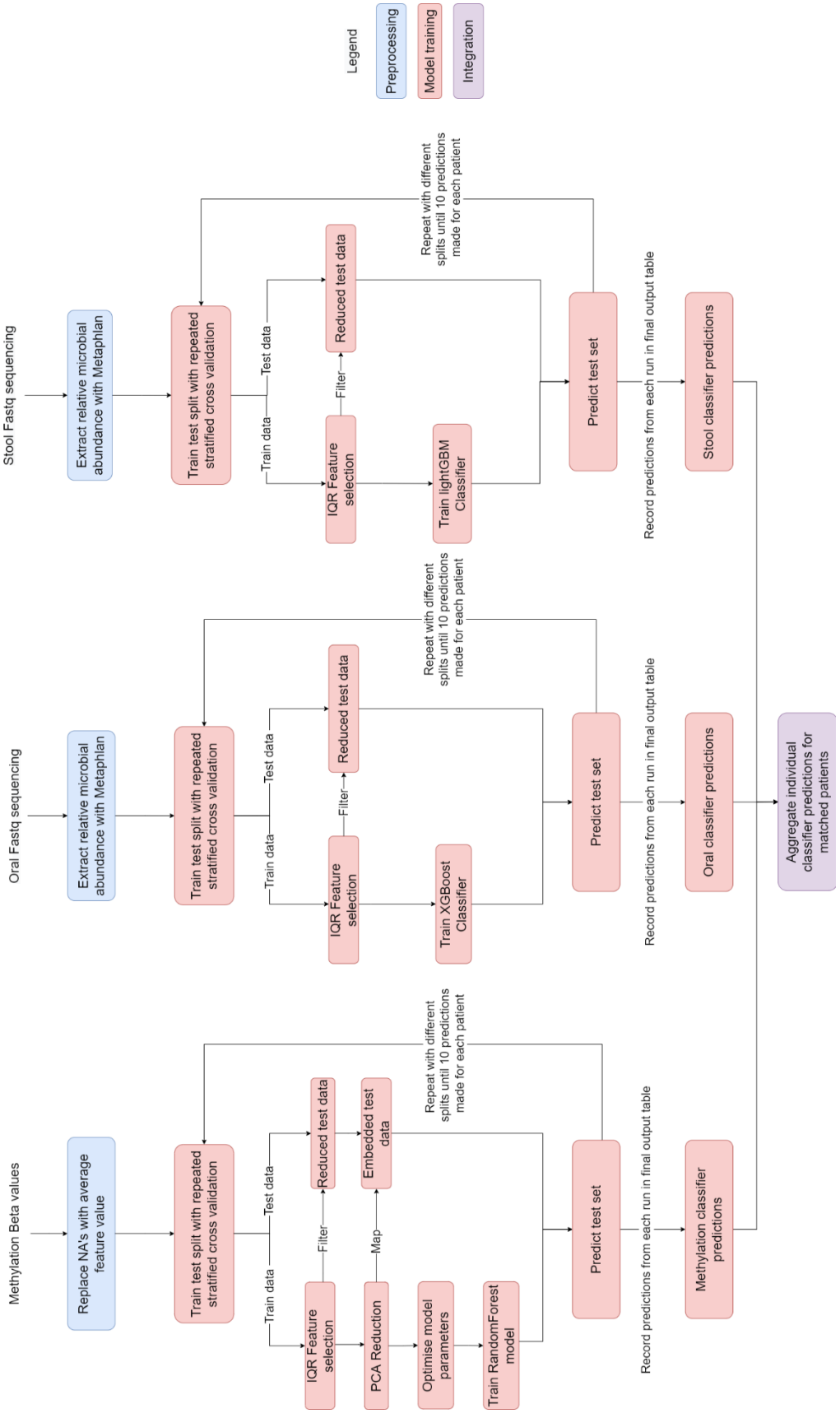

Figure 2. ***Overview of multiomics model design.***

2.2.1.1 Methylation data

2.2.1.1.1 Pre-processing

Initial pre-processing was performed by Braydon Meyer at the Garvan Institute, using the minfi R package to perform quality control and normalisation. Of 865 859 initial probes, 63 858 were identified either as SNPs, cross-reactive or bad probes and were thus removed from further analysis as part of minfi's standard Illumina EPIC workflow (44). The remaining probes were then normalised using the preprocessFunNorm function (45) which removes variability explained by the control probes in the array. It was in this format that the data was received, with beta values for 802 001 CpG sites for each of 26 patients. Beta values are scores between 0 and 1 which correspond to the percentage of reads in which that CpG site was methylated. NA values, which represented just 0.18% of the array, were then replaced by the mean value for the feature they belonged to.

2.2.1.1.2 Train-test-split

A five-fold repeated stratified cross-validation approach was taken to splitting the data into train and test sets in order to maximise the amount of training data available to the model and allow for predictions to be made for every sample in the dataset. The original 26 patients were randomly shuffled before partitioning them into five subsets as close to equal in size and patient response balance as possible. Models were trained five times, each time using a different group as the test set, with the remaining four subsets forming the training set. This process was repeated ten times in total to increase reliability of the results.

2.2.1.1.3 Inter-Quartile Range (IQR) variance for feature selection

ML models notoriously suffer from the 'Curse of Dimensionality', a phrase which refers to how such models are prone to overfit when the number of features is much larger than the number of samples (46). This is true with all three primary datatypes, but especially in the DNA methylation data where there are over 800 000 features for just 26 patients. To address this issue, models will generally apply some form of feature selection technique to reduce the dimensions of the input data. In this study, IQR was used to filter out features with variance of their inter-quartile range below a certain threshold, since features with low variance are of low discriminative value to the classifier. This threshold was determined experimentally and varied between models, with the chosen values for each described in the results. This thresholding was implemented using a custom function built around SciPy's IQR package (47), which calculated IQR on the training data. This variance was then used to reduce both the training and test data to only the features in the training data with IQR above the chosen threshold.

2.2.1.1.4 Principal Component Analysis (PCA) dimensionality reduction

Since individual CpG sites were likely to have low discriminative values on their own, the methylation data was further reduced to just ten dimensions by performing dimensionality reduction on the data using PCA. PCA creates new features equal to the number of dimensions specified which aim to maximise variance and minimise information loss. Sci-Kit Learn's implementation of PCA was fitted on the training data, with the resultant embedding used to transform the train and test data to the reduced feature space separately (48).

2.2.1.1.5 Model training and out of bag parameter optimisation

A Random Forest (RF) model was trained on the PCA transformed training data using the implementation available from Sci-Kit Learn (48). RFs are desirable options for smaller datasets since by design they only use a subset of the total data to build each decision tree, with these many individual trees combined to create a meta model that generates the final predictions. This approach also allows the samples not in a given tree, referred to as out-of-bag (OOB), to be used to estimate the model's ability to generalise.

Many ML models have a number of parameters. While the default parameters typically serve well for most use-cases, better model performance can be achieved by fine tuning these parameters. The RFs OOB score was used in this way to determine the optimal number of trees for the model to construct as well as the best method to use for the max_features parameter.

This was done by training RF classifiers on the training data with number of trees ranging from 15 to 500, in steps of five at a time. For each number of trees tested, classifiers were constructed with three different options for max_features, namely all features, square root of all features, and log of all features. The implementation for this is an open source function written by Kian Ho, Gilles Louppe and Andres Mueller available in the Sci-Kit Learn documentation (49). The OOB score was recorded for each of these configurations, with the optimal number of trees and best max_features option being used to train the RF model for that fold.

The trained model was then used to predict the probability that each test sample belonged to either the good or poor class, with the prediction for the good class being stored until ten predictions were made for every sample in the dataset. A value of less than 0.5 indicated that the model predicted the sample to be a poor responder, while a value greater than 0.5 indicated prediction of a good responder.

2.2.1.1.6 Determination of model feature importances

One unfortunate downside of PCA is that, by creating new features to describe the data, models which provide feature importances such as RF can only give the importances of the PCA components, and not the original features. However, PCA does give the relative importance of each feature in each component. As such, an indirect measure of feature importance can be made by collecting the ten features with the largest contribution for each of the top three most important PCA features according to the trained RF model. These top 30 features were saved for each fold, with the features then ranked by the percentage of folds they appeared in.

2.2.1.1.7 Final classifier prediction

Two different approaches were taken to determine the final model's prediction from the ten predictions available for each sample.

The first of these approaches simply took the mean value of the ten probabilities. The second, referred to as the majority class approach, instead considered the percentage of predictions above a given threshold, namely $\frac{number\ of\ patients\ in\ majority\ class}{total\ patient\ number}$. This threshold was used to account for models with imbalanced input data being prone to overpredicting the majority class. In the case of the methylation classifier where the input data is balanced, this threshold was simply 0.5.

2.2.1.2 Oral Metagenomics

2.2.1.2.1 Pre-processing

In order to build a classifier on the oral metagenomics data, the relative microbial abundances first had to be extracted. This was done using Metaphlan3 (50). Unlike its main competitor Kraken, Metaphlan estimates microbial abundance using clade specific markers

instead of just the total number of reads which map to a given bacteria. This ensures that the abundance of bacteria with larger genomes are not overrepresented in the data. It also means that the data does not have to be cleaned of human contamination, as any such reads will be ignored since they won't map to any microbial specific markers. A pair of scripts (submit_jobs.sh & make_qsub.py) were written to automate the profiling of the raw fastq files using Metaphlan3 on UNSW's computational cluster Katana. The output of this profiling was then parsed (using process.sh & all_abundance.py) to extract just the relative abundances for each sample, with these abundances finally aggregated (prep_metagenomics_data.py) to produce the desired input data frame, in this case comprised of post-treatment oral samples.

2.2.1.2.2 Feature selection and model training

Stratified cross validation was handled as described in 2.2.1.1.2, except with number of folds equal to the minority class, namely 22. As in the methylation classifier, IQR (2.2.1.1.3) was used as the primary means of feature selection. However, since different taxonomic ranks have different relative weights, another form of feature selection that was performed was to filter the data to only include genera.

An eXtreme Gradient Boosting (XGBoost) model with default parameters was then trained on the input data (51,52). XGBoost is generally considered to be one of the most powerful ML algorithms in use today, and uses a tree based approach like RF. However, where RF constructs multiple trees with different subsets of features and samples at once in an attempt to capture different aspects of variation in the data, gradient boosting approaches build trees that attempt to explain the errors made by previous trees, essentially giving greater weight to the trickiest samples.

The trained model was then used to predict the associated test set for the given fold, with the probability predictions being stored as in the methylation classifier. However, when it came to producing the oral classifier's final prediction for each sample from the ten instances recorded, a threshold of 0.67 was used due to the 2:1 imbalance between the good and poor responder classes.

2.2.1.2.3 Extracting feature importances

The features and relative importances assigned to them were stored in their own array. Once all folds were completed, the importances assigned to each feature were summed, with the features then ranked by this net importance. This is slightly different to the approach taken with the methylation data, where features were ranked by the percentage of folds they appeared in.

2.2.1.3 Stool Metagenomics

Pre-processing of the stool metagenomics samples was handled in an identical manner to the oral ones (2.2.1.2.1). As with the oral classifier, stratified cross-validation was performed with folds equal to the size of the minority class, in this case 21. The features were then filtered using IQR (2.2.1.1.3).

2.2.1.3.1 Model training

A Light Gradient Boosting Machine (LGBM) was then trained on the data with default parameters. Like XGBoost, LGBMs are gradient boosting models, with the key difference being that trees in LGBM are grown vertically rather than horizontally, which results in different features being selected and much faster overall performance. As in the case of the oral classifier, the trained model then predicted the test samples, with the final ten predictions being integrated using a 0.67 threshold to account for the 2:1 class

imbalance in responders as was the case for the oral dataset. Feature importances were extracted as described in 2.2.1.2.3.

2.2.1.4 Metagenomics feature importances with LEfSe

Linear discriminant analysis Effect Size (LEfSe) is a differential microbial abundance tool designed to work with the output of Metaphlan (53). It does so by performing the Kruskal-Wallis rank sum test to identify features which are statistically different among the biological class of interest. Unpaired Wilcoxon rank-sum tests can then be used to verify that these features are consistent with nominated biological subclasses, although none were used in this case. Finally, Linear Discriminant Analysis is performed to estimate the effect size of each differentially abundant feature.

To apply LEfSe to our data, Metaphlan's output first had to be transformed into the desired input format, which was done using a custom bash script (lefse.sh). The LEfSe Galaxy module was then used with the default parameters to identify the differentially abundant features in the post-treatment oral and stool data separately.

2.2.1.5 Multiomics integration at the prediction stage

All ten predictions of each of the three individual classifiers on the matched patient cohort were combined into a single array. Since the range of probabilities which different models predict varies significantly, simply averaging these 30 predictions would give undue weight to models such as XGBoost which make predictions with very high degrees of confidence. As such, integration was performed using the majority class approach described in 2.2.1.1.7, with the threshold for a prediction being classified as a good responder determined by the balance of the underlying classes, which in the case of the methylation, oral and stool classifiers was 0.5, 0.67 and 0.67 respectively. The percentage of predictions above these thresholds determined the final integrated model's prediction for

each patient expressed as a value from 0 to 1, with 0 indicating a poor responder and 1 a good responder.

2.2.2 Control vs. Cancer

2.2.2.1 Control vs cancer stool classifier

The relative microbial abundance of the control and pre-treatment cancer samples were extracted using Metaphlan as described in 2.2.1.2.1. As with the individual metagenomics classifiers, stratified cross-validation was performed with 10 repeats and folds equal to the size of the minority class. Features were then filtered using IQR (2.2.1.1.3)

2.2.2.1.1 Under sampling to address class imbalance

The most important difference between the control vs cancer workflows and the original metagenomics classifiers comes in the use of sampling techniques to account for the large imbalance in sample number between the two classes, with only 18 control patients with stool samples verses 60 cancer stool samples. Imbalances become an issue when the ratio between classes outweighs the ability of the underlying data to separate the classes. When large imbalances exist in the input data, models are incentivised to simply guess the majority class, as this will be right most of the time and thus produce a high accuracy score, even if the model isn't learning anything. Sampling techniques attempt to address this issue either by artificially increasing the number of samples belonging to the minority class or by randomly discarding samples of the majority class until an equal class balance is achieved.

The latter approach was taken in the control vs cancer stool classifier, with the RandomUnderSampler function from the Imbalanced-learn package being applied to the training data to achieve this (54). A default XGBoost model was then trained on the re-

balanced training data, with the predictions stored and handled as described in 2.2.1.2.2, except with a threshold of 0.77 used for the final prediction due to the larger imbalance.

2.2.2.2 Control vs cancer oral classifier

The control vs cancer oral classifier followed a near identical workflow to that described in 2.2.2.1, except with no IQR and a different sampling approach.

2.2.2.2.1 Oversampling to address class imbalance

Synthetic Minority Over-Sampling Technique, better known as SMOTE, is an approach to oversampling which tries to avoid overfitting by generating synthetic samples for the minority class using a nearest neighbours approach, implemented by the Imbalanced-learn package (54,55). Specifically, the algorithm picks a sample from the minority class and creates a synthetic sample at a random point in the feature space between this sample and one of the other real samples most similar to it. A LGBM model was then trained on the oversampled training data, with its predictions stored as described in 2.2.2.1.1.

2.2.3 Clinical models

2.2.3.1 Neoadjuvant clinical models

While a total of 21 different clinical features were available for 63 different patients, pre-treatment TILs were only collected for 37 of them, restricting the pre-treatment model's sample size. Sci-kit learn's implementation of a Support Vector Machine (SVM) with the sigmoid kernel was trained on these 37 patients using the five most predictive clinical features, namely pre-treatment TILs %, pre-treatment Ki67 %, tumour grade, size and total number of lymph nodes (48). SVMs attempt to separate the classes of interest by constructing a decision boundary, a line whose equation is determined by the kernel method

used. As opposed to the earlier tree methods which benefit from many features being informative of the difference between classes, SVMs thrive when just a few features account for most of the separation, as is the case in the clinical data. The results of these predictions were handled in the same manner as other individual classifiers (2.2.1.1.7)

2.2.3.2 BOOST clinical models

To ensure the comparison between the follow-up dataset and original patients was as close as possible, the 485 total patients were filtered to only include post-menopausal women with luminal breast cancer who had received endocrine therapy, reducing the dataset to just 164 patients. Logistic Regression (LR) models (with the 'l2' penalty and 'sag' solver parameters) were built on these patients to predict three long term outcomes of breast cancer, namely patient death due to breast cancer, disease free survival and overall patient death. LR models are well suited for datasets where most of the variation can be explained by just a few features, as is the case in the clinical datasets. The Sci-kit learn implementation of LR was used in this case (48).

2.2.3.2.1 Recursive feature elimination using RFECV

Recursive feature elimination is a brute force approach to determining the optimal input features for a model. A model is first trained on all available features, with the performance of the model evaluated on a validation set and the least important feature discarded. A new model is then trained on the remaining features, with the process repeating until only one feature remains. The algorithm then returns the minimal number of features which produced the best model performance. RFECV (Sci-kit Learn's implementation of the method) using a LR model was applied to the training data of each BOOST clinical model, with only the optimal features passed on to the final LR classifier (48).

2.2.4 Evaluation metrics

Classification accuracy (the percentage of correct predictions) is the most common metric for evaluating ML models. While both straightforward to calculate and interpret, it is not ideal in highly imbalanced datasets, as is the case for most of the models in this thesis. If the imbalance is severe enough, high accuracy scores can be achieved by simply always guessing the majority class, even though such a model isn't learning anything. A more useful metric is Area Under the Curve (AUC). The C in AUC refers to the Receiver-Operating Characteristic curve, which plots the true and false positive rate of a model as its classification threshold is changed, a visualisation of which is presented in Figure 3. AUC is used as the primary evaluation metric for all models in this thesis, with accuracy also reported.

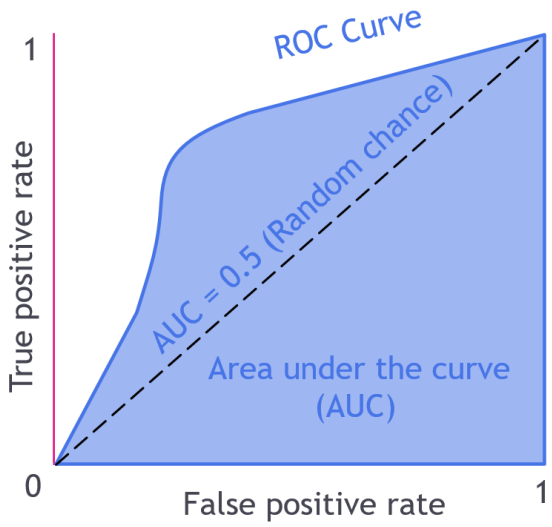

Figure 3. ***An example of a Receiver Operating Characteristic Curve.***

3 RESULTS

3.1 Multiomics model

To assess the comparative performance of the multiomics model against individual datatypes, the performance of each of the three individual classifiers was investigated and is presented in the following subsections. The most important features of each model were also explored to provide insight into potential mechanisms behind endocrine resistance.

3.1.1 Methylation classifier

IQR was a necessary first step to filter the input data to just the most variant features as described in 2.2.1.1.3 in order to reduce the features down to a size models could be trained on. In this case, the chosen threshold of 0.55 reduced the 802 001 input features down to just 144-398, a reduction of approximately 99.96%. A range is provided rather than an exact number because the chosen features varied slightly based on the samples which comprised the training dataset.

3.1.1.1 Determining methylation methodology

Seven of the most popular ML classifiers were evaluated to determine the best performing model for the methylation datatype. Since the number of features after IQR was still an order of magnitude larger than the number of samples, a further feature reduction technique was experimented with in the form of PCA transformation of the data prior to model training. The results for all models, with and without prior PCA, are presented in Table 3 below.

Table 3. ***Comparative model performances on methylation data.***

The best performance for each metric is bolded.

Model Configuration	*PCA AUC*	*PCA Accuracy*	*AUC*	*Accuracy*
		Tree methods		
RF	***0.86***	***0.73***	*0.58*	*0.54*
XGBoost	*0.82*	*0.73*	***0.66***	***0.58***
LGBM	*0.20*	*0.27*	*0.20*	*0.27*
		Linear methods		
LR	*0.78*	*0.65*	*0.73*	*0.73*
Linear SVM	*0.73*	*0.69*	*0.60*	*0.58*
RBF SVM	*0.57*	*0.46*	*0.26*	*0.31*
Sigmoid SVM	*0.49*	*0.5*	*0.68*	*0.54*

While XGBoost was observed to have the best performance on the data without PCA, RF was the best performing model with PCA as well as the best scoring model overall. As such PCA reduction followed by RF was chosen to be used for the final model. PCA is central to the final workflow, with its inclusion leading to substantial increases in both AUC and accuracy for the two best performing models. Tree based models tended to outperform linear approaches on the methylation data overall, with the exception of LGBM, which tends to struggle on small sample sizes.

3.1.1.2 Scores for best performing model

In the final methylation classifier, OOB parameter optimisation was performed for the RF model (see Section 2.2.1.1.5). This optimisation improved the performance of the final model by a couple of points in both AUC and accuracy, although this came at the cost of increasing model training time 30 fold. As described in Section 2.2.1.1.7, four separate comparisons are considered to evaluate the performance of the final classifier. These

predictions are presented in **Error! Reference source not found.**, along with a heatmap of the individual predictions recorded for the matched patient cohort.

Data Type	AUC	Accuracy
All samples mean	*0.89*	*0.77*
All samples majority class	*0.88*	*0.77*
Matched patients mean	*0.90*	*0.75*
Matched patients majority class	***0.91***	***0.80***

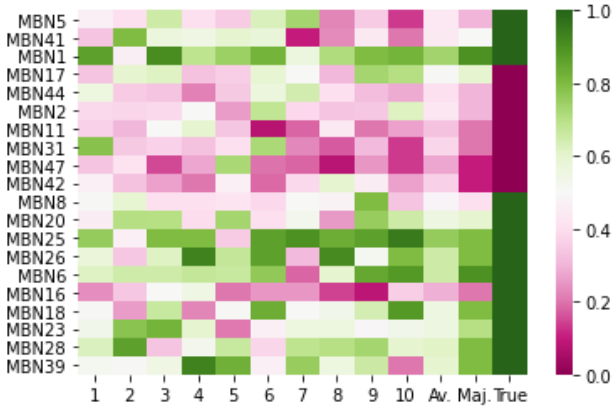

Figure 4. ***Comparative scores for the final methylation classifier [left], and heatmap of individual predictions [right].***

The colours in the heatmap correspond to whether a patient was predicted to be a good responder (green) or a poor responder (pink). The intensity of the colour corresponds to the probability of the predicted class, with darker colours indicative of more confident predictions.

The performance of the model remains relatively consistent across all four comparisons, with the best results corresponding to the matched patients cohort using the majority class method. The four matched patients incorrectly classified by this model were the poor responder MBN17 as well as the good responders MBN5, 8, and 16, as shown in the heatmap in Figure 6. Two of the six patients without matching metagenomics data, C00214 and C00219, were also misclassified as good responders. All misclassified patients were Caucasian. Notably, only two of the 26 patients were HER2 positive (MBN17 and C00214), with both being incorrectly classified as good responders.

3.1.1.3 Model feature importances

Model feature importances were extracted for the best performing classifier as described in 2.2.1.1.6, with features being ranked according to the percentage of cross

validation folds they appeared in. The top 10 features produced by the model are shown in Figure 5. Feature bars are coloured based on which endocrine response group they had higher median relative abundance in and labelled with the gene the site belongs to.

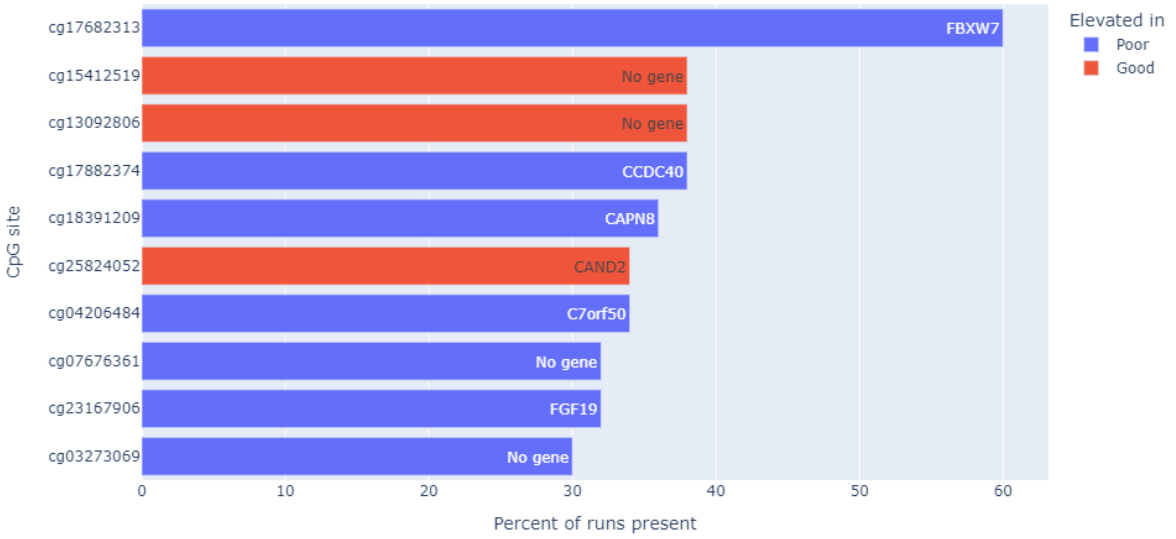

Figure 5. ***Most important CpG sites***

Features are the top 10 most important CpG sites for the top three most important PCA components in each run.

The top feature identified, cg17682313, was a clear outlier in importance, appearing in 60% of runs compared to just 38% of runs for the next most important feature. The site maps to the gene FBXW7, an important tumour suppressor gene, and was hypermethylated in poor responders. Conversely, two of the next most important features do not map to any gene and were hypermethylated in good responders. 70% of features were hypermethylated in poor responders and 60% map to a known gene. Interestingly, aside from the top two features, all other important features map to CpG islands, with five mapping to CpG shores (CpGs < 2kb either side of the island), two to CpG shelves (CpGs between 2 and 4kb either side of a CpG island) and one (cg13092806) an island itself.

Taking a closer look at the top feature for each response class provides a partial explanation for why the model misclassified four of the matched patients, with the top two CpG sites presented in **Error! Reference source not found.**.

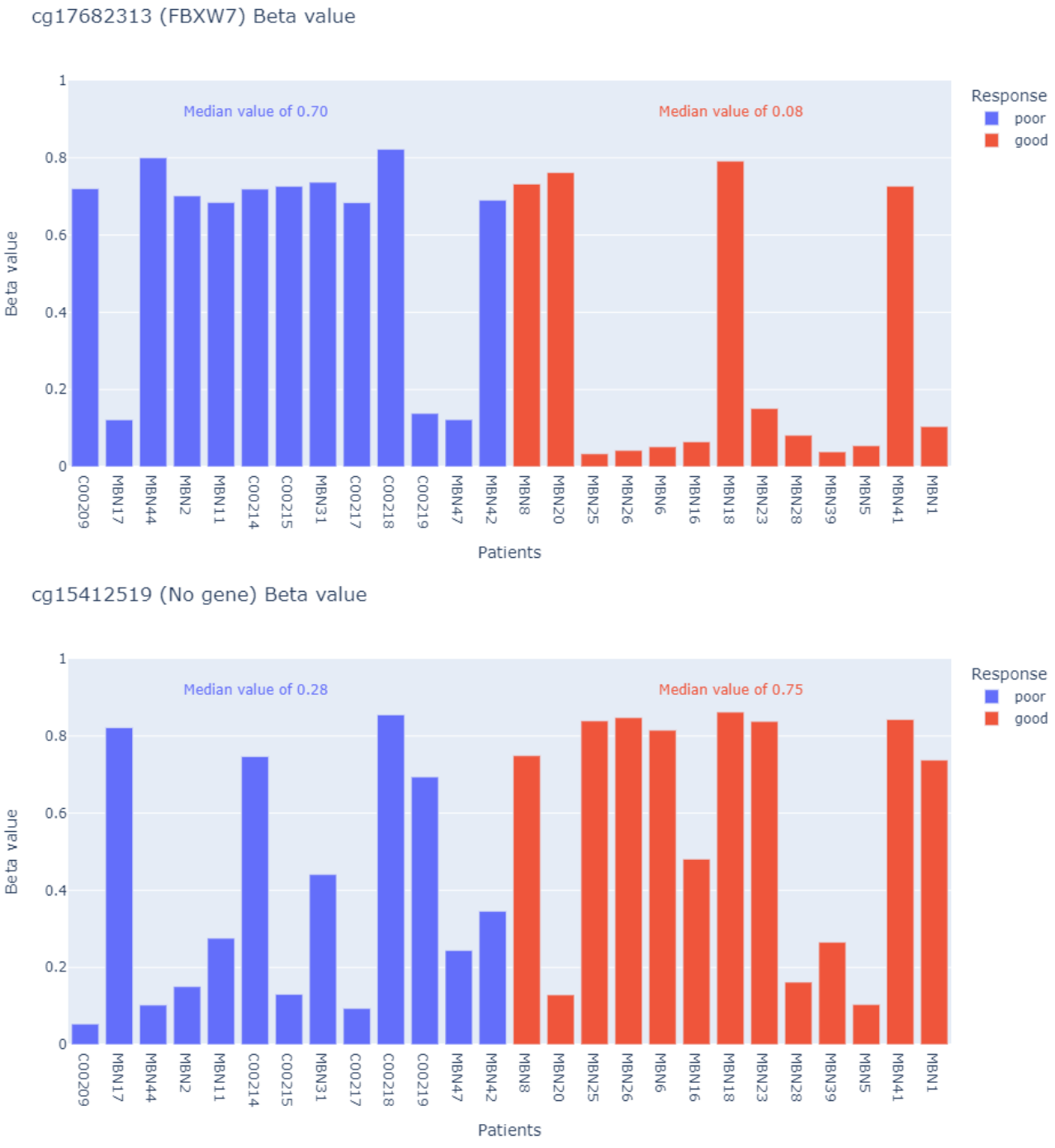

Figure 6. ***Most important CpG site for each responder class***

Patients are labelled with their metagenomics sample ID (when available), or their methylation ID otherwise.

The most important feature identified, cg17682313, displays a very binary methylation pattern, with patients either having > 70% or < 10% of their reads methylated. This distribution maps closely to patients' endocrine response, with poor responders nearly always hypermethylated. Three of the six misclassified patients (MBN17, C00219 and MBN8) display methylation values significantly different from the mean of their class.

Meanwhile, the methylation distribution of cg15412519 appears to be more varied, with three patients displaying intermediate methylation values. Five of the six misclassified patients exhibit methylation values significantly different from the median for their class, with MBN8 as the only exception.

3.1.2 Oral classifier

3.1.2.1 Determining oral methodology

As with methylation, a thorough comparison of ML model performances was conducted to decide the methodology for the final oral classifier. These models were evaluated both on all taxonomic levels as well as just the genus level. The latter approach was investigated as a means of further feature reduction following the success achieved with PCA for the methylation classifier. The IQR threshold for the all-taxa models was 0.65, which reduced the number of features from an original 463 down to approximately 80, an 82% reduction. Meanwhile, a threshold of 0.05 was used in the genera-only models to reduce the original 79 features down to 34, a reduction of 57%. The results of these experimentations are shown in Table 4.

Table 4. ***Comparative model performance for oral classifier***

The best performance for each metric is bolded.

Model Configuration	*Genera AUC*	*Genera Accuracy*	*All AUC*	*All Accuracy*
		Tree methods		
XGBoost	***0.78***	***0.84***	*0.68*	*0.70*
RF	*0.69*	*0.74*	*0.62*	*0.66*
LGBM	*0.69*	*0.70*	***0.70***	***0.75***
		Linear methods		
LR	*0.45*	*0.56*	*0.52*	*0.61*
Sigmoid SVM	*0.40*	*0.63*	*0.29*	*0.64*
RBF SVM	*0.39*	*0.64*	*0.49*	*0.62*
Linear SVM	*0.17*	*0.62*	*0.28*	*0.64*

While LGBM produced the best performance on all oral taxonomic levels, the best model performance overall was achieved by XGBoost on the genera-only oral data. As with the methylation data, tree based models strongly outperformed more linear approaches, with none of the linear models recording an AUC better than random chance.

3.1.2.2 Scores for best performing model

The full comparative results for the best oral classifier (XGBoost on genera-only) are presented in

Figure 7 below.

Data Type	AUC	Accuracy
All samples mean	*0.78*	*0.84*
All samples majority class	*0.79*	*0.84*
Matched patients mean	*0.74*	*0.90*
Matched patients majority class	***0.82***	***0.90***

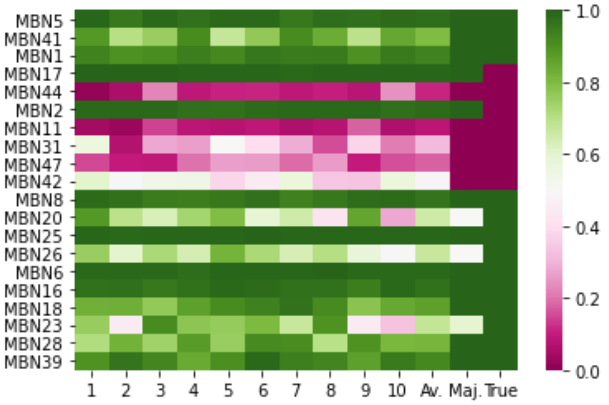

Figure 7. ***Comparative scores for the final oral classifier [left], and heatmap of individual predictions [right].***

The colours in the heatmap correspond to whether a patient was predicted to be a good responder (green) or a poor responder (pink). The intensity of the colour corresponds to the probability of the predicted class, with darker colours indicative of more confident predictions.

The model scores very high on accuracy, making only two errors in its matched patients comparisons. These mistakes correspond to two poor responders, patient MBN17 and MBN2. However, since poor responders are the minority class, misclassifying two of the seven in the matched patient cohort results in the oral classifier recording a worse AUC than the methylation classifier, where most of the misses were good responders. A similar imbalance in missed patients is observed in the full oral cohort, with three good responders and seven poor responders being misclassified from 61 total patients. Two of these missed patients, both poor responders, were of Asian ethnicity (MBN2 and MBN3), out of six total Asian patients in the oral metagenomics dataset.

3.1.2.3 Model feature importances

Feature importances for the best oral classifier were extracted as described in 2.2.1.2.2, with the features ranked by the sum of their XGBoost feature importances across all runs. A secondary means of investigating feature importances was performed using

LEfSe to identify differential taxa as described in section 2.2.1.4. The two approaches are integrated in Figure 8, which depicts the important genera as scored by XGBoost, with features labelled with their LDA score if they were identified as differentially abundant by LEfSe.

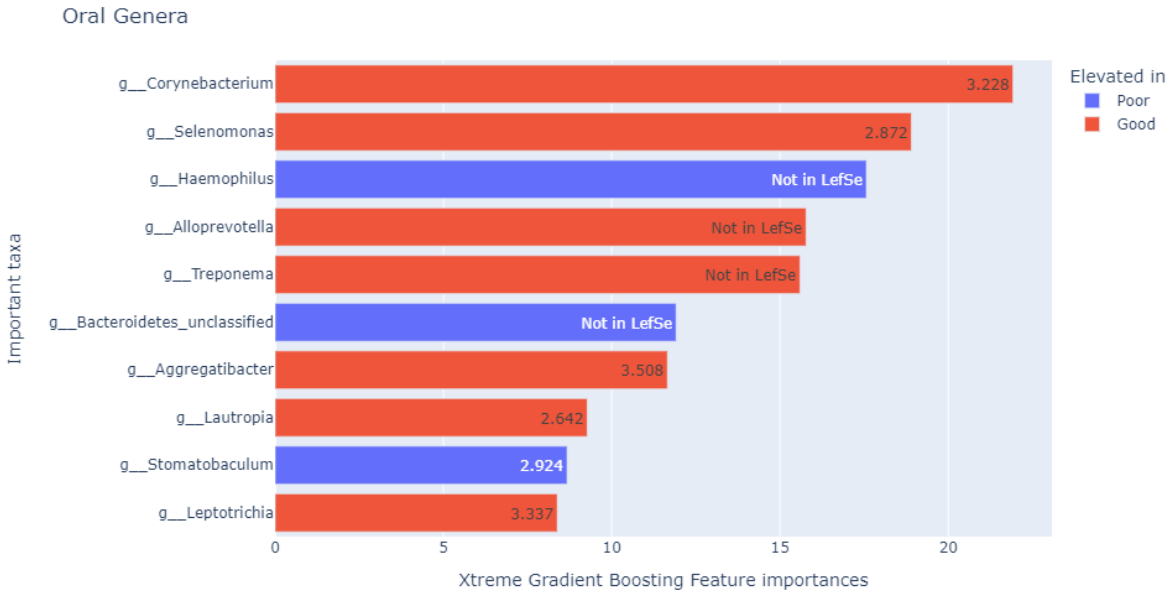

Figure 8. ***Most important oral genera***

Features were coloured based off the response class they were more abundant in and labelled with their LDA score if they received one.

The majority of genera were elevated in good responders, including both the most important XGBoost genera *Corynebacterium* as well as the most important genera according to LEfSe, *Aggregatibacter*. Only one of the genera elevated in poor responders, *Stomatobaculum*, was also identified as differentially abundant by LEfSe. 10 genera were identified as differentially abundant by LEfSe overall, with six of them ranked in the top 10 XGBoost features as well.

3.1.3 Stool classifier

3.1.3.1 Determining stool methodology

As was done for the oral classifiers, the comparative performance of the most popular ML classifiers was evaluated, both on all taxa and genera only (Table 5).

Table 5. ***Comparative model performance for stool classifier.***
The best performance for each metric is bolded.

Model Configuration	*Genera AUC*	*Genera Accuracy*	*All AUC*	*All Accuracy*
		Tree methods		
LGBM	*0.54*	*0.55*	***0.74***	***0.65***
XGBoost	*0.62*	*0.60*	*0.67*	*0.60*
RF	*0.60*	*0.55*	*0.59*	*0.60*
		Linear methods		
Linear SVM	***0.72***	***0.73***	*0.59*	*0.67*
Sigmoid SVM	*0.67*	*0.7*	*0.36*	*0.65*
LR	*0.64*	*0.63*	*0.58*	*0.58*
RBF SVM	*0.21*	*0.65*	*0.46*	*0.65*

This comparison found LGBM on all taxa to achieve the best AUC, while a Linear SVM on the genera only data recorded the highest accuracy but a slightly worse AUC. This was the only case in the multiomics model datatypes where a linear model scored the best for a given comparison. While the accuracy of the Linear SVM was higher, LGBM on all taxa was selected to be used in the final model due to its higher AUC, with this metric carrying more weight on the roughly 2:1 imbalanced stool dataset.

3.1.3.2 Scores for best performing model

The full results for the best stool classifier are presented in Figure 9.

Data Type	AUC	Accuracy
All samples mean	*0.74*	*0.65*
All samples majority class	*0.75*	*0.65*
Matched patients mean	*0.73*	*0.65*
Matched patients majority class	***0.81***	***0.70***

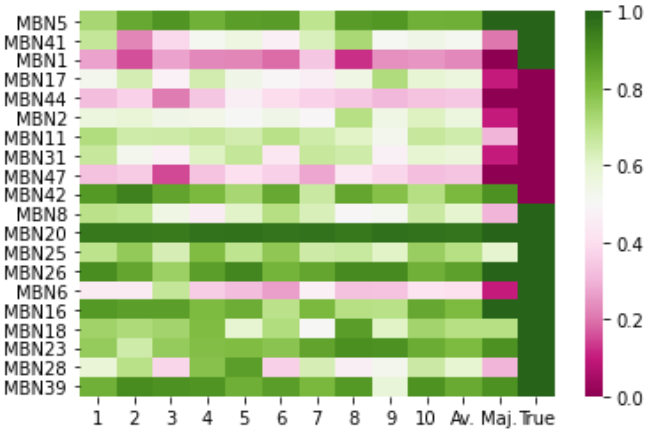

Figure 9. ***Comparative scores for the final stool classifier [left], and heatmap of individual predictions [right].***

The colours in the heatmap correspond to whether a patient was predicted to be a good responder (green) or a poor responder (pink). The intensity of the colour corresponds to the probability of the predicted class, with darker colours indicative of more confident predictions.

While the AUC for the matched patients majority class approach was quite good, the accuracy was noticeably lower than the other two classifiers, with six patients from the matched cohort misclassified. Five of these missed patients were good responders (MBN41, MBN1, MBN8, MBN6 and MBN28) while one (MBN42) was a poor responder.

3.1.3.3 Model feature importances

As with the oral classifier, the feature importances presented in Figure 10 are a combination of two approaches, with features ranked based on their importance to the LGBM model and labelled with their LDA score if they were identified as differentially abundant by LEfSe.

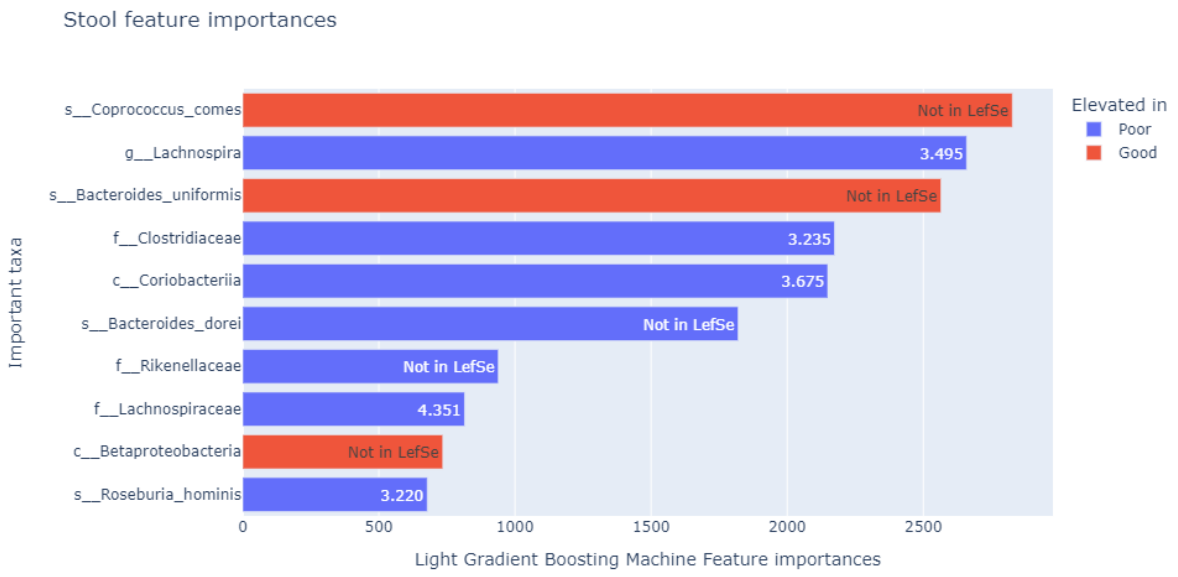

Figure 10. ***Most important stool taxa***

Features were coloured based off the response class they were more abundant in and labelled with their LDA score if they received one.

While the majority of important stool features were elevated in poor responders (including all 13 of the differential taxa identified by LEfSe), the top feature according to the LGBM model was the species *Coprococcus comes*, which was instead elevated in good responders. Meanwhile, the most important feature according to LEfSe was the *Lachnospiraceae* family, with an LDA score of 4.351. The majority of important stool features were higher order taxa, especially among those elevated in poor responders.

3.1.4 Integrated classifier

3.1.4.1 Multiomics prediction scores and importance of individual classifiers

The predictions from each of the individual classifiers were integrated using the majority class approach described in 2.2.1.1.7 across 30 repeats, with 10 repeats per datatype. These final predictions were then evaluated, with the results presented in Figure 11.

Data Type	AUC	Accuracy
DNA Methylation	*0.91*	*0.80*
Oral Metagenomics	*0.82*	*0.90*
Stool Metagenomics	*0.81*	*0.7*
Integrated model	***0.99***	***0.95***

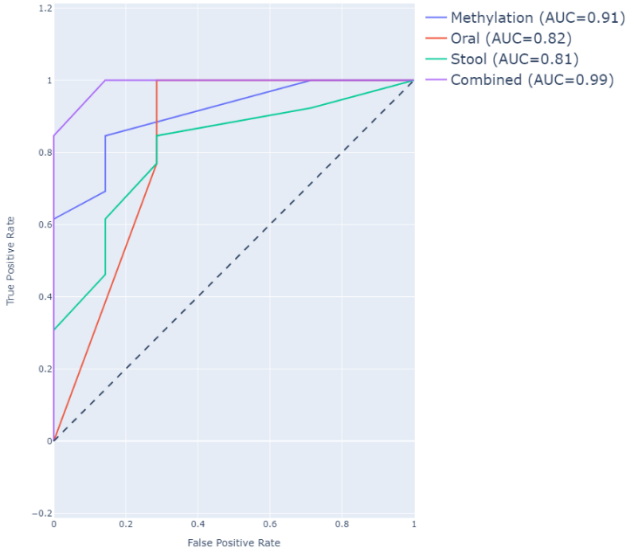

Figure 11. ***Performance comparison of the individual classifiers with the integrated multi-omics model***

The curves on the right are the ROC curves each model's AUC is derived from.

The integrated model achieved substantial improvements in both AUC and accuracy over the individual classifiers, with the final model achieving near perfect performance, only misclassifying a single one of the 20 matched patients: patient MBN17. A comparison of the ROC curves for each approach is presented in Figure 11, demonstrating that the combined model is equal to or better than each of the individual classifiers at every classification threshold.

3.1.4.2 Misclassified patients

The performance of the individual models and combined approach is visualised on an individual patient level in the heatmap in Figure 12. It presents the majority class predictions for the matched patient cohort for each individual classifier, along with the final combined prediction and the true patient response types.

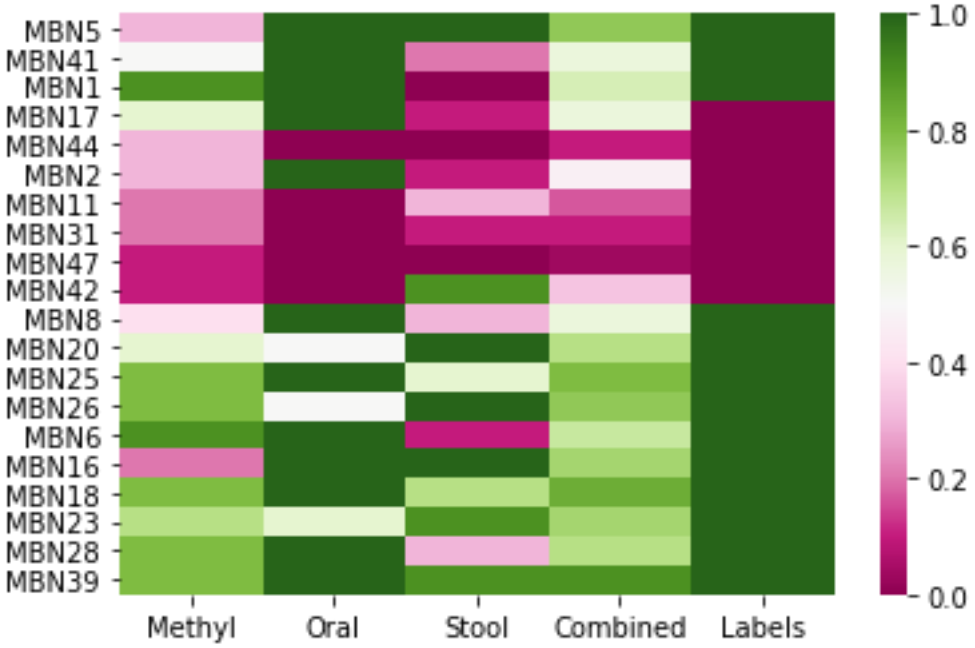

Figure 12. ***Individual and majority class predictions.***
Individual patient predictions for the methylation (methyl), oral, and stool models are compared to the combined (majority class) prediction and the true value (labels). The intensity of the colour corresponds to the probability of the predicted class, with darker colours indicative of more confident predictions.

The missed patients of the methylation classifier are relatively balanced between response classes. However, the oral model only misclassified poor responders whereas the

stool model was much more likely to misclassify good responders. Perhaps as a result, there is relatively little overlap of misclassified patients between the individual classifiers, with only patients MBN17 and MBN8 being misclassified by more than one model and no patients being incorrectly predicted by both metagenomics datatypes.

While the combined approach achieves 95% accuracy by only misclassifying patient MBN17, three other patients are correctly predicted with a low degree of confidence. MBN17's misclassification is also a low confidence prediction, for a total of 20% of the matched patient cohort with integrated predictions in the range of 0.4-0.6 (Table 6). A score of ≥ 0.5 indicated a patient was predicted as a good responder, while a score < 0.5 meant they were predicted to be a poor responder.

Table 6. ***Breakdown of low confidence patient predictions from the integrated model for each individual model.***

For 'true class', a value of 0 indicates a poor responder and 1 indicates a good responder.

Low confidence predictions	*Methylation prediction*	*Oral prediction*	*Stool prediction*	*Combined prediction*	*True class*
MBN17	*0.6*	*1.0*	*0.1*	*0.57*	*0*
MBN8	*0.4*	*1.0*	*0.3*	*0.57*	*1*
MBN41	*0.5*	*1.0*	*0.2*	*0.57*	*1*
MBN2	*0.3*	*1.0*	*0.1*	*0.47*	*0*

If we consider only high confidence predictions, the combined model achieves 100% accuracy at the cost of returning inconclusive results for 20% of samples. Interestingly, MBN8 is still predicted correctly as a good responder despite being incorrectly predicted by the majority of datatypes, illustrating the potential for high confidence predictions from a single model to outweigh low confidence predictions from multiple models.

3.2 Control vs. Cancer

3.2.1 Control vs. oral

3.2.1.1 Comparative model performances

Given the dominance of tree based models on the post-treatment metagenomics data, the performance of these models was compared on the control vs pre-treatment oral data. The class imbalance of this comparison was quite severe, with just nine control patients compared to 61 cancer patients. The oversampling technique SMOTE was experimented with as a means to address this imbalance, as described in 2.2.2.2.1. The results of these comparisons are recorded in Table 7. Under sampling was also experimented with but is not presented as no model achieved better than random chance.

Table 7. ***Comparative model performance for control vs. oral classifier.***
The best performance for each metric is bolded.

Model Configuration	*SMOTE AUC*	*SMOTE Accuracy*	*AUC*	*Accuracy*
		Tree methods		
LGBM	***0.73***	***0.87***	*0.51*	*0.87*
XGBoost	*0.61*	*0.83*	*0.50*	*0.84*
RF	*0.57*	*0.87*	*0.44*	*0.87*

LGBM proved to be the best performing model on this comparison, achieving a modest 0.73 AUC. Oversampling with SMOTE was shown to be central to achieving this, with no model trained on the unsampled data achieving better AUC than random chance.

3.2.1.2 Model feature importances

Important features for the SMOTE LGBM model were handled similarly to the multiomics metagenomics models, with features evaluated through a combination of model feature importance and LEfSe score. The combined importances are presented in Figure 13.

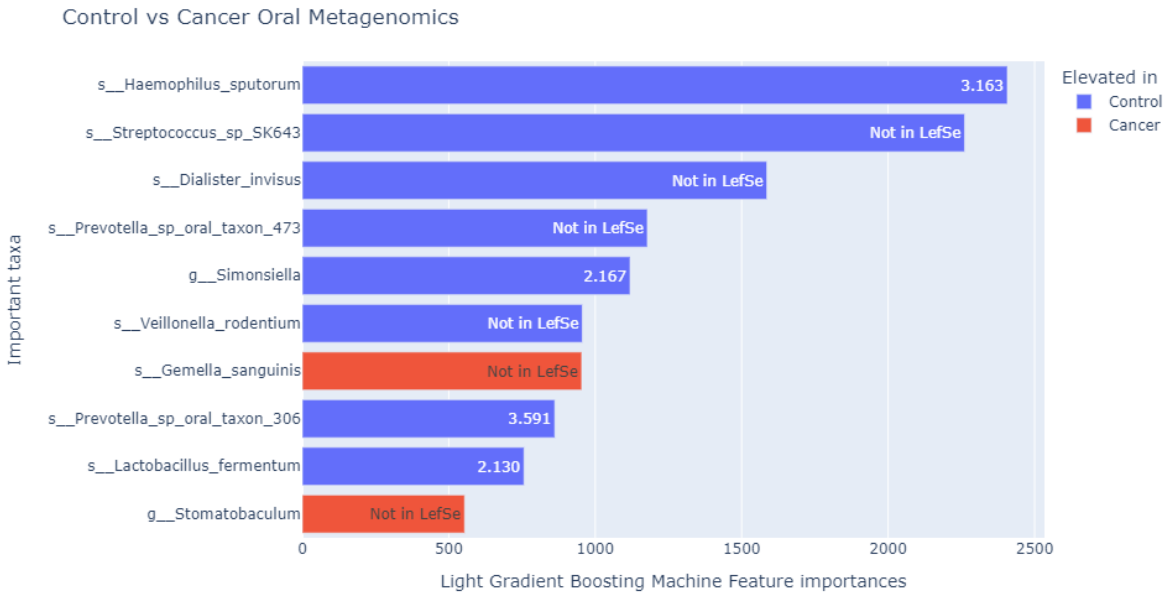

Figure 13. ***Most important features of control vs. cancer in oral microbiome***

Features were coloured based off the response class they were more abundant in and labelled with their LDA score if they received one.

The majority of top features are observed to be elevated in the control patient cohort, with three of these features belonging to unnamed taxa. While the LEfSe analysis identified a total of 21 differential taxa, there was relatively limited overlap with the important model features, with only four such differential taxa appearing in both methodologies. Notably, this does include the most important feature according to the LGBM model, *Haemophilus sputorum.* The presence of *Stomatobaculum* is also worth

highlighting, as it is also an important feature of the post-treatment oral classifier where it is elevated in poor responders (Figure 14).

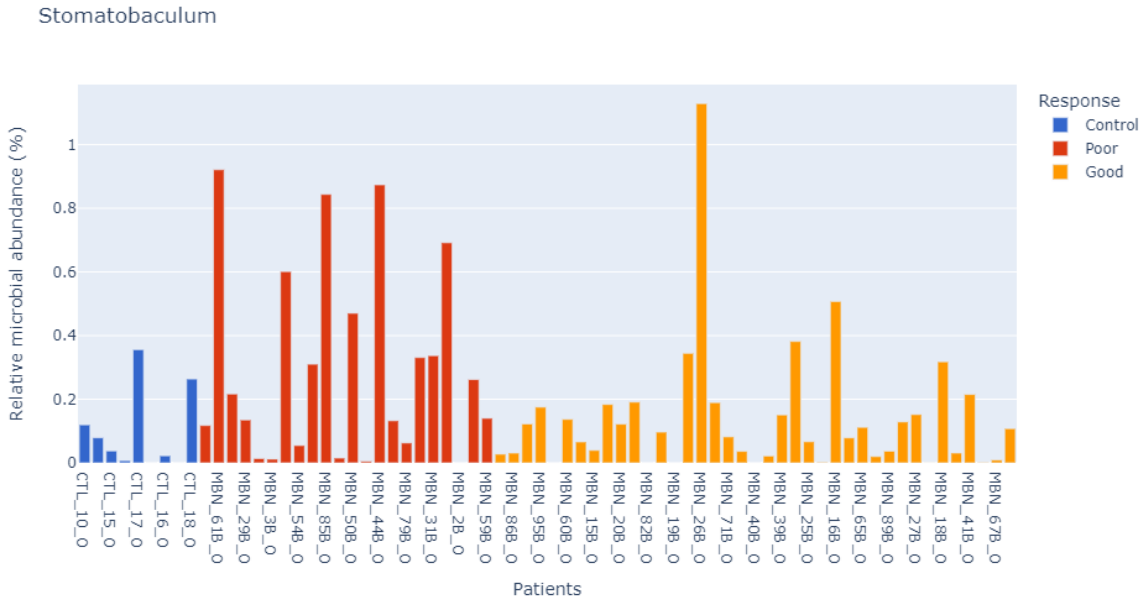

Figure 14. ***Relative abundance of Stomatobaculum genera in the oral microbiome***

Stomatobaculum displays three distinct levels of abundance, lowest in control samples and highest in poor responders, with good responders in between.

One other feature of the control oral model which displays a noticeable distribution is that of *Dialister invisus.* Whilst not identified as differential by LEfSe for the control vs cancer oral comparison, it is present in every control patient's oral microbiome. Conversely, although it was not highlighted as an important feature in the post-treatment oral classifier due to filtering to genera only, *D. invisus* was identified by LEfSe as a significantly differential bacteria elevated in good responders. Finally, *D. invisus* also has a noticeable presence in the gut microbiome, but this is exclusively among cancer patients, with none of the 18 stool control samples having any *D. invisus* present (Figure 15b).

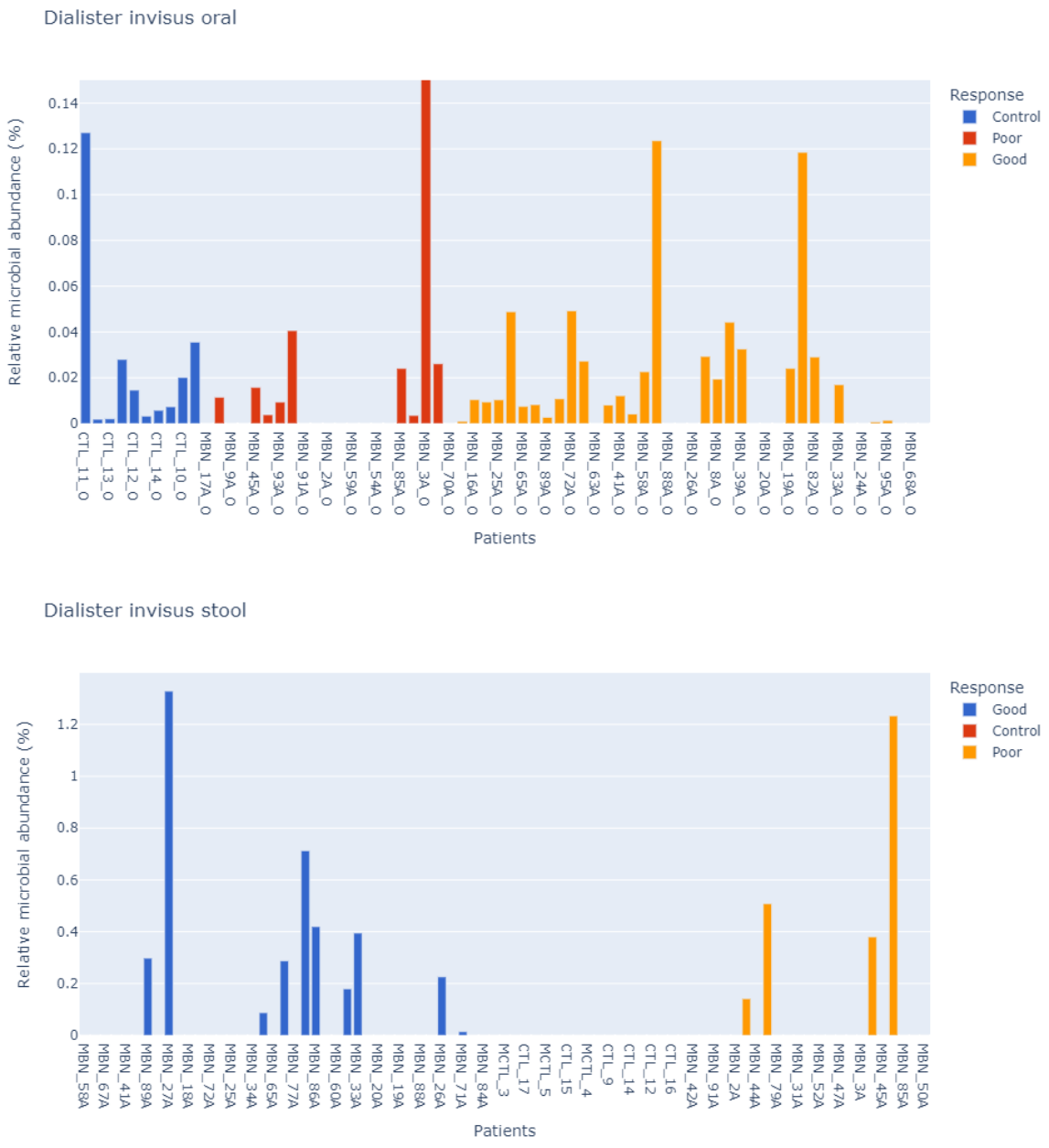

Figure 15. ***Relative abundance of D. invisus in the oral and gut microbiomes***

D. invisus was present in the oral microbiomes of all nine control patients but not in the gut microbiome of any of the 18 gut control samples.

3.2.2 Control vs. stool

3.2.2.1 Comparative model performances

As with the control vs cancer oral classifier, the performance of tree-based ML models with and without re-sampling was evaluated (Table 8). IQR with a threshold of 0.1 was used to filter the input data from 688 features down to approximately 148, a ~78% reduction.

Table 8. ***Comparative model performance for control vs. stool classifier.***
The best performance for each metric is bolded.

Model configuration	*Under sampling AUC*	*Under sampling Accuracy*	*SMOTE AUC*	*SMOTE Accuracy*	*AUC*	*Accuracy*
			Tree methods			
XGBoost	***0.70***	***0.67***	*0.62*	*0.70*	*0.53*	*0.74*
RF	*0.63*	*0.60*	***0.69***	***0.74***	***0.60***	***0.77***
LGBM	*0.50*	*0.78*	*0.59*	*0.72*	*0.57*	*0.69*

The model with the highest AUC by a small margin was XGBoost using under sampling to counter class imbalance (see 2.2.2.1.1). This is an acceptable, if not outstanding, performance. While other models achieved higher accuracy, this is irrelevant as none of the scores were higher than 0.78, the resulting accuracy for simply guessing cancer for all patients thanks to the class imbalance.

3.2.2.2 Model feature importances

Feature importances for the control vs cancer pre-treatment stool comparison were derived in the standard manner for metagenomics classifiers, with features ranked by model

importance and labelled with their LEfSe LDA score where they were identified as differential (Figure 16).

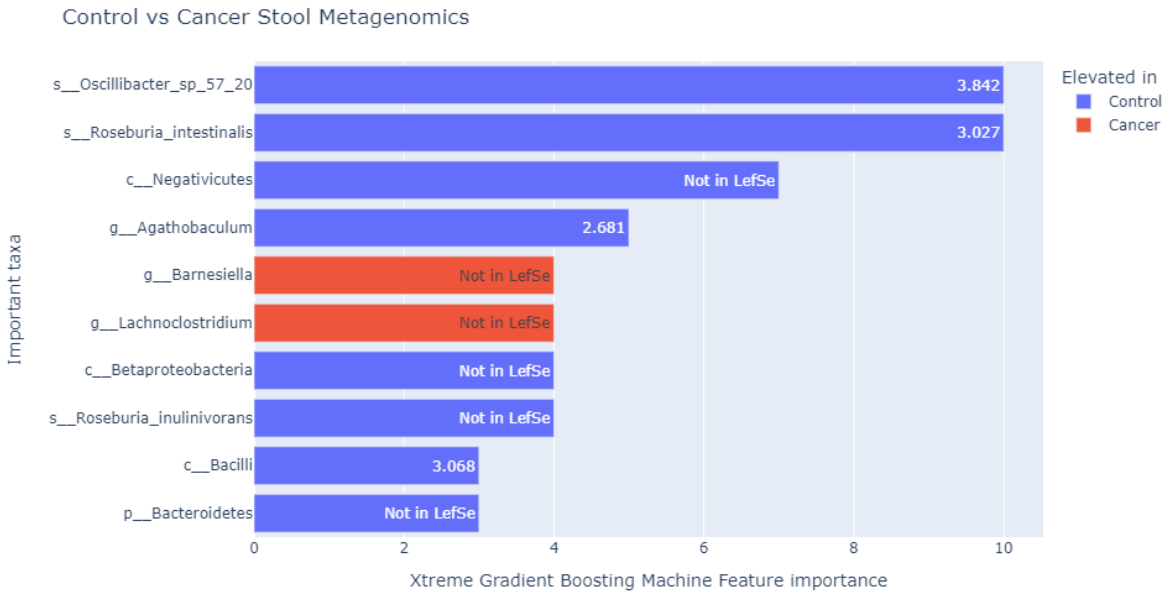

Figure 16. ***Most important control vs. cancer stool bacteria***

Features were coloured based off the response class they were more abundant in and labelled with their LDA score if they received one.

As with control cancer comparisons in the oral microbiome, the majority of important features are elevated in control patients, with the tied most important features also being identified as differential by LEfSe. 45 differential taxa were identified by LEfSe, with just four featuring in the top 10 most important features according to XGBoost. The elevation of *Oscillibacter_sp_57_20* in control patients was particularly prominent, as can be seen in Figure 17.

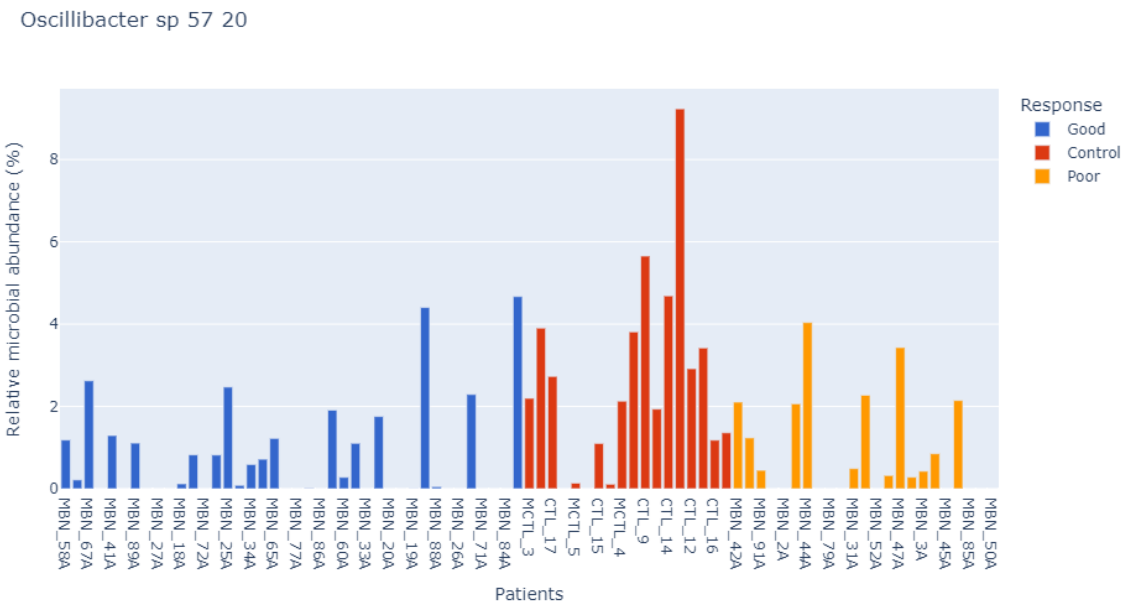

Figure 17. ***Relative microbial abundance of Oscillibacter sp 57 20 in the stool microbiome***

Oscillibacter sp 57 20 is significantly elevated in control patients compared to either of the cancer groups

3.3 Clinical model

3.3.1 Clinical dataset

Classifiers were constructed on the pre-treatment clinical data to evaluate the feasibility of ML models for predicting endocrine resistance early. While 21 total features were present in the clinical data (13 of which were pre-treatment), models were built using only the five most prognostic features. This was done to reduce the clinical burden of the model, with patient Ki67 and TIL % pre-treatment, along with tumour grade, size and number of lymph nodes being the five features selected for investigation.

3.3.1.1 Model comparison

The performance of all seven models experimented with so far in this thesis was evaluated on the five clinical features chosen for investigation (Table 9).

Table 9. ***Comparative model performance for clinical data classifier.***

The best performance for each metric is bolded.

Model Configuration	*AUC*	*Accuracy*
	Tree methods	
RF	*0.63*	*0.69*
XGBoost	*0.39*	*0.51*
Light GBM	*0.38*	*0.66*
	Linear methods	
Sigmoid SVM	***0.93***	***0.86***
LR	*0.87*	*0.80*
Linear SVM	*0.80*	*0.80*
RBF SVM	*0.70*	*0.83*

Unlike the metagenomics and methylation models, linear approaches significantly outperformed tree based models on the clinical data, with only one tree model (RF) achieving a performance better than chance. Meanwhile, all three SVMs and LR achieved good to great AUC scores, with sigmoid SVM having the best results.

3.3.1.2 Scores for best model

Comparative scores for the SVM model are presented in Figure 18.

Data Type	**AUC**	**Accuracy**
All samples mean	*0.93*	*0.86*
All samples majority class	*0.91*	*0.89*
Matched patients mean	*0.97*	*0.85*
Matched patients majority class	***0.97***	***0.95***

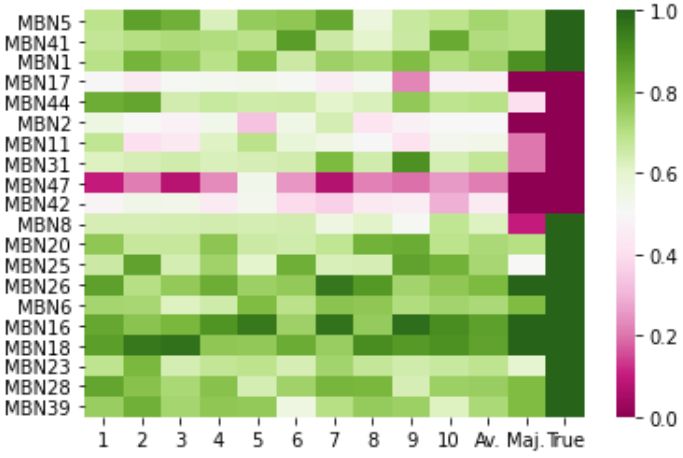

Figure 18. ***Comparative scores for the final clinical data classifier [left], and heatmap of individual predictions [right].***

The colours in the heatmap correspond to whether a patient was predicted to be a good responder (green) or a poor responder (pink). The intensity of the colour corresponds to the probability of the predicted class, with darker colours indicative of more confident predictions.

While the AUC of 0.93 on all samples was already a great performance, the matched patient cohort with majority class prediction was a further improvement, with 0.97 AUC and 95% accuracy. Unlike previous classifiers, this accuracy was noticeably higher than the other comparisons and so could be the result of chance given the small size of the matched patient cohort.

3.3.1.3 Relative contribution of each feature

Since SVMs do not naturally generate feature importances for their input data, a range of sigmoid SVMs were constructed on different subsets of the five clinical features to establish their relative importance to the final model (Table 10).

Table 10. Relative contribution of clinical features

Scores of model's with different features removed to estimate the importance of each individual feature

Clinical feature subset	AUC	Accuracy
All features	*0.93*	*0.86*
No Grade	*0.86*	*0.83*
No nodes	*0.86*	*0.77*
No Size	*0.70*	*0.74*
Ki67 & TILs only	*0.66*	*0.71*
No TILs	*0.50*	*0.69*
No Ki67	*0.35*	*0.66*

Removal of any single feature reduced the performance of the model noticeably, with even the least important feature (tumour grade) resulting in a 0.07 AUC drop upon

removal. The two most important features are Ki67 and TILs, which were able to generate an acceptable model on their own, and the removal of either of these features resulted in a model with AUC no better than random chance.

3.3.1.4 Comparing clinical predictions to multiomics predictions

For an individual classifier, the clinical model performed exceptionally well, with a near perfect AUC and accuracy. The only patient misclassified by the clinical model was patient MBN8. This patient was also misclassified by the methylation and stool classifiers, although the integrated multiomics model predicted the patient correctly with a low degree of confidence. A heatmap of the clinical and multiomics models predictions on all individual patients is presented in Figure 19.

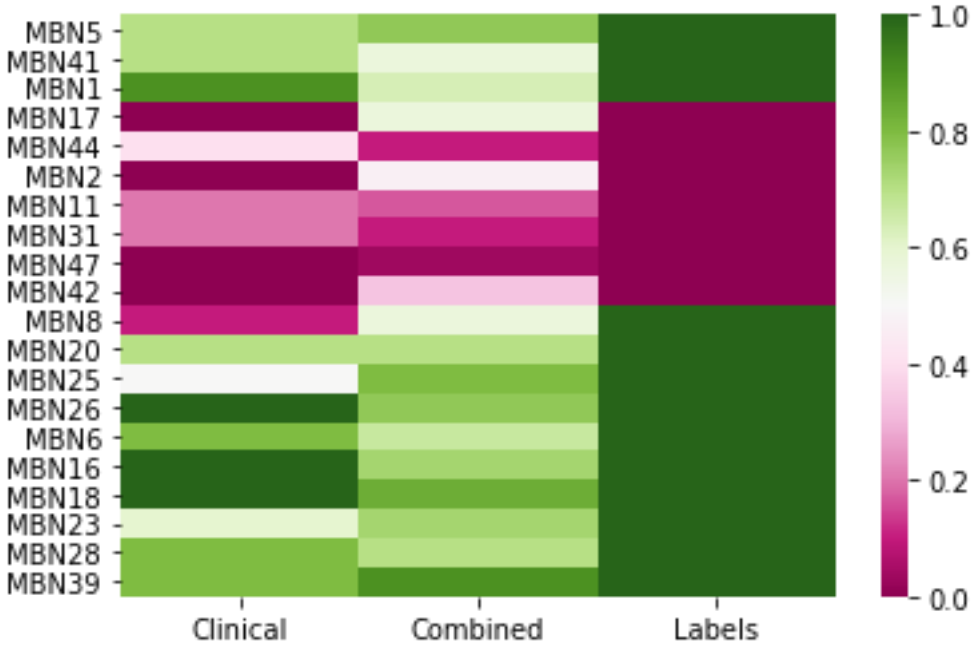

Figure 19. ***Comparison of predictions for each patient from the clinical data model (Clinical) and the integrated multiomics model (Combined).***
Labels is the correct value for each patient. Red and green again correspond to a predictions of poor and good responders, respectively, with intensity of colour increasing with increasing prediction confidence.

Overall, the clinical model's predictions were much more confident than those of the combined model due to not having to integrate the results of three separate classifiers. Patients MBN17 and MBN2, which were both low confidence predictions in the combined model, were conversely among the most confidently correct predictions made by the clinical model. Interestingly, when considering the mean approach to integrating predictions (see 2.2.1.1.7 for details), the accuracy not only dropped from 95% to 85% but MBN8 was no longer misclassified. Instead, poor responders MBN44, MBN31 and MBN11 were missed, neither of which were ever incorrectly classified in the methylation or metagenomics models

3.3.2 Clinical validation

The long term prognostic value of the five clinical features investigated in 3.3.1 was evaluated using data from a 20 year follow-up study. Once patients were filtered to just luminal post-menopausal women who had received endocrine therapy (164 of the original 485 total patients), the closest matching subset to the patient cohort studied in the rest of this thesis, the imbalance in some of the long term outcomes was severe. For instance, in terms of the outcome of breast cancer death, only 13 patients were recorded as having died specifically of breast cancer, with the remaining 151 still alive or died of non-breast cancer specific causes. As such, under sampling was applied to the training data for all comparisons on this dataset.

The next step following under sampling was to determine the best model for the new dataset. Given the success of linear approaches on the original clinical data, the top two such models were evaluated on their ability to predict breast cancer death status. Recursive feature elimination (2.2.3.2.1) was experimented with as a means of parameter optimisation for the models, with only the optimal feature subset being trained on. The results of these comparisons are presented in Table 11.

Table 11. ***Comparison of linear model performance on BOOST follow-up dataset.***

Model Configuration	*RFECV AUC*	*RFECV Accuracy*	*AUC*	*Accuracy*
LR	***0.85***	***0.76***	*0.66*	*0.55*
SVC	*0.78*	*0.62*	***0.63***	***0.60***

The best performing model by a significant margin was LR with preliminary RFECV. This workflow was then used to predict to other long-term outcomes of interest - disease-free survival status and death from any cause, with the results presented in

Patient Outcome	**AUC**	**Accuracy**
Breast cancer death mean	*0.85*	*0.76*
Breast cancer death Majority class	*0.84*	*0.79*
Disease free status mean	*0.77*	*0.73*
Disease free status Majority class	*0.79*	*0.72*
Death mean	*0.64*	*0.58*
Death Majority class	*0.59*	*0.59*

.

Table 12. ***Performance of RFECV LR at predicting long-term outcomes of interest***

Patient Outcome	**AUC**	**Accuracy**
Breast cancer death mean	*0.85*	*0.76*
Breast cancer death Majority class	*0.84*	*0.79*
Disease free status mean	*0.77*	*0.73*
Disease free status Majority class	*0.79*	*0.72*
Death mean	*0.64*	*0.58*
Death Majority class	*0.59*	*0.59*

This workflow performed very well at predicting both breast cancer death and disease free survival, but only achieved moderate performance when predicting death from

any cause. There was no noticeable difference between mean and majority class prediction methods for any of the three long term outcomes.

4 DISCUSSION

In this thesis multiomics and clinical models were constructed to predict endocrine resistance in ER+ breast cancer patients. Both of these approaches resulted in excellent AUC scores according to the thresholds set out in the original ROC paper by C Metz (56). To explore the potential mechanisms underlying endocrine resistance, as well as ER+ breast cancer more broadly, the top features were extracted for each of the individual classifiers that comprise the multiomics model as well as for model's trained to separate pre-treatment ER+ cancer patients from healthy controls. In this section, the clinical characteristics of these features will be explored in relation to how they might influence cancer diagnosis or patient response to endocrine therapy.

4.1 Diversity of model methodologies

The methodologies of each model presented in this thesis were determined based off of exhaustive experimentation, with the end result being that none of the final models share the exact same workflow structure. This suggests that there is no one size fits all approach to use when applying ML to your problem of interest. But a few clear trends did emerge in the process of determining the final model approaches. A clear distinction was evident in the performance of tree based models and linear models on each datatype, with tree based models performing best on the omics data while linear approaches performed much better in the clinical models. Furthermore, feature selection with IQR proved essential to training successful ML models in all omics models except the control oral model, which only achieved moderate performance.

4.2 Performance comparison of multiomics and individual models

While all three individual classifiers demonstrated quite good performances in terms of AUC, the multiomics model outperformed them all on both AUC and accuracy. Despite the gut microbiota having the most obvious link to endocrine resistance through its involvement in the estrobolome it was actually the least valuable of the three omics datatypes investigated, recording both the lowest AUC and accuracy of any datatype. The comparatively under-studied oral microbiome, meanwhile, achieved the highest accuracy of any of the individual classifiers, suggesting its interaction with endocrine resistance may be closer than was previously believed.

4.2.1 Advantages of integration approach

The approach used to perform integration in this thesis is very simple as far as integration attempts go, with the final prediction simply the number of predictions where the patient was predicted as a good responder over the total number of predictions. This majority class approach was favoured over simply taking the mean of the ten predictions to account for the fact that different ML model types predict with different levels of confidence, as can be seen visually by simply comparing the heatmap for the oral classifier in Figure 7 with the much paler heatmap for the methylation classifier in Figure 4.

By performing integration at the prediction stage this enables models to be trained on patients which don't have data available for ever omics type being integrated, which was a key concern for this study in which two thirds of the metagenomics patients don't have matched methylation data available. The other main benefit this kind of integration enables is for the model to be flexible and modular, with individual classifiers able to be replaced with models trained on different omics types without affecting the multiomics model at all.

4.2.2 Misclassified patients

The only patient misclassified by the multi-omics model was patient MBN17. Notably, this patient was the only HER2 positive case in the matched patient cohort, meaning it is the only example of the Luminal B-like subtype in this group. The five breast cancer subtypes have been shown to each have their own specific methylation profiles (57). As such, it's possible that this patient was not misclassified by the methylation classifier but instead that the methylation profile of HER2 positive poor responders is closer to that of good responders than it is to other poor responders. In fact, the only other HER2 positive patient in the methylation data was also misclassified as a good rather than a poor responder, lending credence to this theory. A similar rationale could potentially explain why MBN17 was also missed by the oral classifier. However, one other HER2 positive patient is present in the oral metagenomics dataset and was correctly classified, although they were a good responder, so it is not a perfect equivalent.

4.3 Methylation feature importances

By far the most important CpG site identified by the classifier was cg17682313, which maps to the gene FBXW7. FBXW7 is a major tumour suppressor in human cancers, especially breast cancer, with the gene being methylated in 51% of primary breast tumours (58). Methylation of FBXW7 reduces its overall expression, but it's effect on breast cancer prognosis is somewhat unclear. One study (59) observed that reduced expression of FBXW7 (and in turn, high methylation of the gene) was associated with poor prognosis. However, a second group observed that high expression of FBXW7 was associated with decreased overall survival in normal type cancers (60). Regardless, given the stark divide in methylation levels of good and poor responders at cg17682313, FBXW7 clearly plays an important role in ER+ endocrine resistance. CAPN8 and FGF19 are both also linked to

cancer in some way. CAPN8 has been shown to form a complex with CAPN9, with the complex termed G-calpain involved in supressing tumorigenesis (61). FGF19's link is more direct with high FGF19 expression predicting worse prognosis in invasive ductal carcinoma of the breast (62). FGF19 was also shown to act as a oncogenic signalling pathway in breast cancer via interaction with its receptor, FGFR4 (63).

4.4 Stool feature importances

4.4.1 Biomarkers of endocrine response

As covered in Section 1.4.2, the 'estrobolome' is currently one of the leading theories to explain why some patients respond poorly to endocrine therapy, with bacteria which encode gmGUS possessing the potential to reactivate estrogen and thus weaken the impact of aromatase inhibitors. All species present in the top stool features (*Coprococcus comes, Bacteroides uniformis, Bacteroides dorei* and *Roseburia hominis)* encode GUS or a GUS candidate in the Human Microbiome Project Database (64), with *Bacteroides uniformis* and *Roseburia hominis* having their GUS estrogen reactivating abilities confirmed via in vitro assay (65). While you would initially expect species which code for an estrogen reactivating enzyme to be elevated in poor responders, the two highest ranked species are both more abundant in good responders. Whilst *C.comes* has been established in another study as being associated with good prognosis (66), the very same study links *B. uniformis* to overall poor prognosis. This discrepancy needs further investigation to establish whether *B.uniformis* is positively or negatively associated with patient outcomes. Should later research demonstrate gmGUS to be at least in part responsible for endocrine resistance, one option would be to target drugs specifically to inhibit this enzyme, with current popular options targeting the "bacterial loop", a bacteria-specific structure missing in the orthologous mammalian GUS (32).

4.4.2 Biomarkers of healthy patients

The important features of the control vs stool model are dominated by short chain fatty acid (SCFA) producing bacteria, with *Roseburia intestinalis* and *Roseburia inulinivorans* in particular producing butyrate. Butyrate is a SCFA which exhibits a number of anticancer activities, including inducing cancer cell apoptosis, anti-inflammatory effects, inhibition of histone deacetylation and suppression of angiogenesis (67). *R.inulinovorans* has also been identified previously (68) as one of only a few bacteria which decreased in abundance in postmenopausal breast cancer patients compared to controls. Perhaps the most interesting feature however is *Oscillibacter sp 57 20*, an uncultivated bacterium which was identified last year in a study of 1098 individuals as one of 15 bacterium indicative of good overall health (69).

4.5 Oral feature importances

4.5.1 Biomarkers of endocrine response

As discussed in Section 1.4.2, relatively few links between oral microbiota and breast cancer have been established. The only major exception to this is periodontitis, which is significantly associated with breast cancer risk. It is thus notable that the Aggregatibacter genera is identified as a significant feature (with the highest LefSe score) given that Aggregatibacter aphrophilus is one of the main causes of periodontal infection.

Meanwhile, although model feature importances were generated for ER+ breast cancer vs. control patients in the oral microbiome, the extremely small size of the control patient set as well as the abundance of uncultivated bacteria leaves questions over how reliable these results can be, with only one species have any documented association with breast cancer, *Dialister invisus.*

4.6 *Dialister invisus:* a unique indicator of cancer dysbiosis

While the gut and oral microbiomes are treated as independent in this thesis, there is an ongoing debate as to whether oral microbes are able to colonise the gut, given that the two are connected via the passage of saliva. Most studies have found almost no evidence of colonisation of oral bacteria in the gut with just one notable exception: *Dialister invisus* (70). *D.invisus* is a prolific oral pathogen which found at elevated levels in half a dozen different cancers (71). In general, it's believed that oral microbiota are unable to colonise the gut due to the physiochemical differences of the two habitats as well as the variety of biological barriers they must cross along the way, such as gastric acid, bile salts and antimicrobial peptides. Yet *D.invisus* bucks this trend, as can be observed in Figure 15, with noticeable relative abundance in both the stool and oral microbiomes of cancer patients. However, while *D.invisus* is present in every single oral control sample, it is not present in a single one of the stool controls. This suggests that invasion of the gut by *D.invisus* is a sign of dysbiosis related to patients cancer diagnosis, with *D.invisus* in the oral microbiome alone indicative that a patient is healthy. Interestingly, amongst cancer patients *D.invisus* is significantly more abundant in good responders, suggesting that its association with cancer is independent of estrogen.

4.7 Limitations

This study is subject to some major limitations which greatly impact the scope and significance of its results.

4.7.1 Post treatment only

Probably the biggest restriction of this study was the lack of pre-treatment methylation data forcing the multiomics model to be built on post-treatment data in order to include all omics types. As a result, the multiomics model must be treated as a 'proof-of-

concept' predictive model demonstrating that multiomics integration can substantially improve on the performance of ML models on any individual datatypes, since you can't predict patient response to endocrine therapy early using post treatment data. This motivated the decision to construct models on the pre-treatment clinical data to satisfy the aim of identifying poor endocrine responders early and offering them alternative treatments.

4.7.2 Small sample size

Another difficulty with this project was the relatively small sample sizes available for building predictive models on, with the methylation data posing a particularly great restriction. While good performance was achieved for the models regardless, the limited sample size restricts the conclusions we can draw from these models, with the control vs oral classifier being an extreme example of this. It also limits what techniques can be attempted. While deep learning methods have shown great promise especially in the realm of image classification, the small sample sizes of biological datasets such as this are simply inappropriate for deep learning and neural network approaches, which are reliant on large numbers of samples to work.

4.7.3 Lack of wet lab validation

A final important word of caution about these results is the highly speculative nature of any feature importance investigations, as only so much can be gleaned about their activity from the literature and data provided. While the differential bacteria identified in the results section can serve as a starting point for further investigation into the mechanisms underlying endocrine resistance, their suggested associations with endocrine resistance need substantial experimental validation before any judgements can be made.

4.8 Future Directions

Future work on this project should start off attempting to replicate the performance of the multiomics model on pre-treatment, either using methylation data should more samples be collected or one of the other datatypes available to the larger project but outside of the scope of this thesis, such as spatial transcriptomics data or histological images. A pre-treatment multiomics model could also consider integrating the clinical features given their impressive standalone performance. Another avenue worth exploring would be profiling the metabolic pathways of the metagenomics data with HumanN.

5 REFERENCES

1. Bray, F., Laversanne, M., Weiderpass, E., and Soerjomataram, I. (2021) The ever-increasing importance of cancer as a leading cause of premature death worldwide. *Cancer* **127**, 3029-3030
2. Jemal, A., Torre, L., Soerjomataram, I., and Bray, F. (2019) *The Cancer Atlas*, Third ed., American Cancer Society, Atlanta, GA
3. Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021) Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: a cancer journal for clinicians* **71**, 209-249
4. Marmot, M. G., Altman, D. G., Cameron, D. A., Dewar, J. A., Thompson, S. G., and Wilcox, M. (2013) The benefits and harms of breast cancer screening: an independent review. *British journal of cancer* **108**, 2205-2240
5. Waks, A. G., and Winer, E. P. (2019) Breast Cancer Treatment: A Review. *Jama* **321**, 288-300
6. Inic, Z., Zegarac, M., Inic, M., Markovic, I., Kozomara, Z., Djurisic, I., Inic, I., Pupic, G., and Jancic, S. (2014) Difference between Luminal A and Luminal B Subtypes According to Ki-67, Tumor Size, and Progesterone Receptor Negativity Providing Prognostic Information. *Clin Med Insights Oncol* **8**, 107-111
7. Loibl, S., and Gianni, L. (2017) HER2-positive breast cancer. *The Lancet* **389**, 2415-2429
8. Brown, D. C., and Gatter, K. C. (2002) Ki67 protein: the immaculate deception? *Histopathology* **40**, 2-11
9. Ishihara, M., Mukai, H., Nagai, S., Onozawa, M., Nihei, K., Shimada, T., and Wada, N. (2013) Retrospective Analysis of Risk Factors for Central Nervous System Metastases in Operable Breast Cancer: Effects of Biologic Subtype and Ki67 Overexpression on Survival. *Oncology* **84**, 135-140
10. Ciancio, N., Galasso, M. G., Campisi, R., Bivona, L., Migliore, M., and Di Maria, G. U. (2012) Prognostic value of p53 and Ki67 expression in fiberoptic bronchial biopsies of patients with non small cell lung cancer. *Multidisciplinary respiratory medicine* **7**, 29
11. Josefsson, A., Wikström, P., Egevad, L., Granfors, T., Karlberg, L., Stattin, P., and Bergh, A. (2012) Low endoglin vascular density and Ki67 index in Gleason score 6 tumours may identify prostate cancer patients suitable for surveillance. *Scandinavian journal of urology and nephrology* **46**, 247-257
12. Yamashita, H. (2015) Tumor biology in estrogen receptor-positive, human epidermal growth factor receptor type 2-negative breast cancer: Mind the menopausal status. *World J Clin Oncol* **6**, 220-224
13. Bulut, N., and Altundag, K. (2015) Does estrogen receptor determination affect prognosis in early stage breast cancers? *Int J Clin Exp Med* **8**, 21454-21459
14. Awan, A., and Esfahani, K. (2018) Endocrine therapy for breast cancer in the primary care setting. *Curr Oncol* **25**, 285-291
15. Smith, G. L. (2014) The Long and Short of Tamoxifen Therapy: A Review of the ATLAS Trial. *J Adv Pract Oncol* **5**, 57-60

16. Riemsma, R., Forbes, C. A., Kessels, A., Lykopoulos, K., Amonkar, M. M., Rea, D. W., and Kleijnen, J. (2010) Systematic review of aromatase inhibitors in the first-line treatment for hormone sensitive advanced or metastatic breast cancer. *Breast cancer research and treatment* **123**, 9-24
17. Bradley, R., Braybrooke, J., Gray, R., Hills, R. K., Liu, Z., Pan, H., Peto, R., Dodwell, D., McGale, P., Taylor, C., Francis, P. A., Gnant, M., Perrone, F., Regan, M. M., Berry, R., Boddington, C., Clarke, M., Davies, C., Davies, L., Duane, F., Evans, V., Gay, J., Gettins, L., Godwin, J., James, S., Liu, H., MacKinnon, E., Mannu, G., McHugh, T., Morris, P., Read, S., Straiton, E., Jakesz, R., Fesl, C., Pagani, O., Gelber, R., De Laurentiis, M., De Placido, S., Gallo, C., Albain, K., Anderson, S., Arriagada, R., Bartlett, J., Bergsten-Nordström, E., Bliss, J., Brain, E., Carey, L., Coleman, R., Cuzick, J., Davidson, N., Del Mastro, L., Di Leo, A., Dignam, J., Dowsett, M., Ejlertsen, B., Goetz, M., Goodwin, P., Halpin-Murphy, P., Hayes, D., Hill, C., Jagsi, R., Janni, W., Loibl, S., Mamounas, E. P., Martín, M., Mukai, H., Nekljudova, V., Norton, L., Ohashi, Y., Pierce, L., Poortmans, P., Pritchard, K. I., Raina, V., Rea, D., Robertson, J., Rutgers, E., Spanic, T., Sparano, J., Steger, G., Tang, G., Toi, M., Tutt, A., Viale, G., Wang, X., Whelan, T., Wilcken, N., Wolmark, N., Cameron, D., Bergh, J., and Swain, S. M. (2022) Aromatase inhibitors versus tamoxifen in premenopausal women with oestrogen receptor-positive early-stage breast cancer treated with ovarian suppression: a patient-level meta-analysis of 7030 women from four randomised trials. *The Lancet Oncology* **23**, 382-392
18. Baum, M., Budzar, A. U., Cuzick, J., Forbes, J., Houghton, J. H., Klijn, J. G., and Sahmoud, T. (2002) Anastrozole alone or in combination with tamoxifen versus tamoxifen alone for adjuvant treatment of postmenopausal women with early breast cancer: first results of the ATAC randomised trial. *Lancet (London, England)* **359**, 2131-2139
19. Eiermann, W., Paepke, S., Appfelstaedt, J., Llombart-Cussac, A., Eremin, J., Vinholes, J., Mauriac, L., Ellis, M., Lassus, M., Chaudri-Ross, H. A., Dugan, M., and Borgs, M. (2001) Preoperative treatment of postmenopausal breast cancer patients with letrozole: A randomized double-blind multicenter study. *Annals of oncology : official journal of the European Society for Medical Oncology* **12**, 1527-1532
20. Wong, Z. W., and Ellis, M. J. (2004) First-line endocrine treatment of breast cancer: aromatase inhibitor or antioestrogen? *British journal of cancer* **90**, 20-25
21. Dowsett, M., Smith, I., Robertson, J., Robison, L., Pinhel, I., Johnson, L., Salter, J., Dunbier, A., Anderson, H., Ghazoui, Z., Skene, T., Evans, A., A'Hern, R., Iskender, A., Wilcox, M., and Bliss, J. (2011) Endocrine therapy, new biologicals, and new study designs for presurgical studies in breast cancer. *Journal of the National Cancer Institute. Monographs* **2011**, 120-123
22. Martin, M., Zielinski, C., Ruiz-Borrego, M., Carrasco, E., Turner, N., Ciruelos, E. M., Muñoz, M., Bermejo, B., Margeli, M., Anton, A., Kahan, Z., Csöszi, T., Casas, M. I., Murillo, L., Morales, S., Alba, E., Gal-Yam, E., Guerrero-Zotano, A., Calvo, L., de la Haba-Rodriguez, J., Ramos, M., Alvarez, I., Garcia-Palomo, A., Huang Bartlett, C., Koehler, M., Caballero, R., Corsaro, M., Huang, X., Garcia-Sáenz, J. A., Chacón, J. I., Swift, C., Thallinger, C., and Gil-Gil, M. (2021) Palbociclib in combination with endocrine therapy versus capecitabine in hormonal receptor-positive, human epidermal growth factor 2-negative, aromatase inhibitor-resistant metastatic

breast cancer: a phase III randomised controlled trial-PEARL. *Annals of oncology : official journal of the European Society for Medical Oncology* **32**, 488-499

23. Ellis, M. J., Tao, Y., Luo, J., A'Hern, R., Evans, D. B., Bhatnagar, A. S., Chaudri Ross, H. A., von Kameke, A., Miller, W. R., Smith, I., Eiermann, W., and Dowsett, M. (2008) Outcome prediction for estrogen receptor-positive breast cancer based on postneoadjuvant endocrine therapy tumor characteristics. *Journal of the National Cancer Institute* **100**, 1380-1388
24. Dowsett, M., Smith, I. E., Ebbs, S. R., Dixon, J. M., Skene, A., A'Hern, R., Salter, J., Detre, S., Hills, M., and Walsh, G. (2007) Prognostic value of Ki67 expression after short-term presurgical endocrine therapy for primary breast cancer. *Journal of the National Cancer Institute* **99**, 167-170
25. Turnbull, A. K., Arthur, L. M., Renshaw, L., Larionov, A. A., Kay, C., Dunbier, A. K., Thomas, J. S., Dowsett, M., Sims, A. H., and Dixon, J. M. (2015) Accurate Prediction and Validation of Response to Endocrine Therapy in Breast Cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **33**, 2270-2278
26. Sheri, A., Smith, I. E., Johnston, S. R., A'Hern, R., Nerurkar, A., Jones, R. L., Hills, M., Detre, S., Pinder, S. E., Symmans, W. F., and Dowsett, M. (2015) Residual proliferative cancer burden to predict long-term outcome following neoadjuvant chemotherapy. *Annals of oncology : official journal of the European Society for Medical Oncology* **26**, 75-80
27. Zhang, M., Wang, Y., Wang, Y., Jiang, L., Li, X., Gao, H., Wei, M., and Zhao, L. (2020) Integrative Analysis of DNA Methylation and Gene Expression to Determine Specific Diagnostic Biomarkers and Prognostic Biomarkers of Breast Cancer. **8**
28. Ennour-Idrissi, K., Dragic, D., Issa, E., Michaud, A., Chang, S.-L., Provencher, L., Durocher, F., and Diorio, C. (2020) DNA Methylation and Breast Cancer Risk: An Epigenome-Wide Study of Normal Breast Tissue and Blood. *Cancers* **12**, 3088
29. Soleimani Dodaran, M., Borgoni, S., Sofyalı, E., Verschure, P. J., Wiemann, S., Moerland, P. D., and van Kampen, A. H. C. (2020) Candidate methylation sites associated with endocrine therapy resistance in ER+/HER2- breast cancer. *BMC Cancer* **20**, 676
30. Batra, R. N., Lifshitz, A., Vidakovic, A. T., Chin, S.-F., Sati-Batra, A., Sammut, S.-J., Provenzano, E., Ali, H. R., Dariush, A., Bruna, A., Murphy, L., Purushotham, A., Ellis, I., Green, A., Garrett-Bakelman, F. E., Mason, C., Melnick, A., Aparicio, S. A. J. R., Rueda, O. M., Tanay, A., and Caldas, C. (2021) DNA methylation landscapes of 1538 breast cancers reveal a replication-linked clock, epigenomic instability and cis-regulation. *Nature Communications* **12**, 5406
31. Plottel, C. S., and Blaser, M. J. (2011) Microbiome and malignancy. *Cell Host Microbe* **10**, 324-335
32. Sui, Y., Wu, J., and Chen, J. (2021) The Role of Gut Microbial β-Glucuronidase in Estrogen Reactivation and Breast Cancer. **9**
33. Shao, J., Wu, L., Leng, W.-D., Fang, C., Zhu, Y.-J., Jin, Y.-H., and Zeng, X.-T. (2018) Periodontal Disease and Breast Cancer: A Meta-Analysis of 1,73,162 Participants. *Front Oncol* **8**, 601-601
34. Dizdar, O., Hayran, M., Guven, D. C., Yılmaz, T. B., Taheri, S., Akman, A. C., Bilgin, E., Hüseyin, B., and Berker, E. (2017) Increased cancer risk in patients with periodontitis. *Current medical research and opinion* **33**, 2195-2200

35. Michaud, D. S., Fu, Z., Shi, J., and Chung, M. (2017) Periodontal Disease, Tooth Loss, and Cancer Risk. *Epidemiologic reviews* **39**, 49-58
36. Jiang, X., and Shapiro, D. J. (2014) The immune system and inflammation in breast cancer. *Molecular and cellular endocrinology* **382**, 673-682
37. (!!! INVALID CITATION !!! {}).
38. García-Teijido, P., Cabal, M. L., Fernández, I. P., and Pérez, Y. F. (2016) Tumor-Infiltrating Lymphocytes in Triple Negative Breast Cancer: The Future of Immune Targeting. *Clin Med Insights Oncol* **10**, 31-39
39. Skriver, S. K., Jensen, M. B., Knoop, A. S., Ejlertsen, B., and Laenkholm, A. V. (2020) Tumour-infiltrating lymphocytes and response to neoadjuvant letrozole in patients with early oestrogen receptor-positive breast cancer: analysis from a nationwide phase II DBCG trial. *Breast cancer research : BCR* **22**, 46
40. Gellert, P., Segal, C. V., Gao, Q., López-Knowles, E., Martin, L.-A., Dodson, A., Li, T., Miller, C. A., Lu, C., Mardis, E. R., Gillman, A., Morden, J., Graf, M., Sidhu, K., Evans, A., Shere, M., Holcombe, C., McIntosh, S. A., Bundred, N., Skene, A., Maxwell, W., Robertson, J., Bliss, J. M., Smith, I., Dowsett, M., Johnston, S., Todd, R., Horgan, K., Chan, S., Holt, S. D. H., Parton, M., Laidlaw, I., Vaidya, J. S., Irvine, T., Hoar, F., Khattak, I., Kothari, A., Brazil, L., Gallegos, N., Wheatley, D., Johnson, T., Sparrow, G., Ledwidge, S., Mortimer, C., Ornstein, M., Ferguson, D., Adamson, D., Cutress, R., Johnson, R., Crowley, C., Winters, Z., Hamed, H., Burcombe, R., Cleator, S., Kelleher, M., Roberts, J., Vesty, S., Hadaki, M., Quigley, M., Doughty, J., Laws, S., Seetharam, S., Thorne, A., Donnelly, P., Group, P. T. M., and Trialists. (2016) Impact of mutational profiles on response of primary oestrogen receptor-positive breast cancers to oestrogen deprivation. *Nature Communications* **7**, 13294
41. Millar, E. K., Graham, P. H., O'Toole, S. A., McNeil, C. M., Browne, L., Morey, A. L., Eggleton, S., Beretov, J., Theocharous, C., Capp, A., Nasser, E., Kearsley, J. H., Delaney, G., Papadatos, G., Fox, C., and Sutherland, R. L. (2009) Prediction of local recurrence, distant metastases, and death after breast-conserving therapy in early-stage invasive breast cancer using a five-biomarker panel. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **27**, 4701-4708
42. Millar, E., Browne, L., Slapetova, I., Shang, F., Ren, Y., Bradshaw, R., Ann Brauer, H., O'Toole, S., Beretov, J., Whan, R., and Graham, P. H. (2020) TILs Immunophenotype in Breast Cancer Predicts Local Failure and Overall Survival: Analysis in a Large Radiotherapy Trial with Long-Term Follow-Up. *Cancers* **12**
43. Millar, E. K., Graham, P. H., McNeil, C. M., Browne, L., O'Toole, S. A., Boulghourjian, A., Kearsley, J. H., Papadatos, G., Delaney, G., Fox, C., Nasser, E., Capp, A., and Sutherland, R. L. (2011) Prediction of outcome of early ER+ breast cancer is improved using a biomarker panel, which includes Ki-67 and p53. *British journal of cancer* **105**, 272-280
44. Fortin, J.-P., Triche, T. J., Jr., and Hansen, K. D. (2017) Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics* **33**, 558-560
45. Fortin, J.-P., Labbe, A., Lemire, M., Zanke, B. W., Hudson, T. J., Fertig, E. J., Greenwood, C. M. T., and Hansen, K. D. (2014) Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biology* **15**, 503

46. Keogh, E., and Mueen, A. (2017) Curse of Dimensionality. in *Encyclopedia of Machine Learning and Data Mining* (Sammut, C., and Webb, G. I. eds.), Springer US, Boston, MA. pp 314-315
47. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., Vijaykumar, A., Bardelli, A. P., Rothberg, A., Hilboll, A., Kloeckner, A., Scopatz, A., Lee, A., Rokem, A., Woods, C. N., Fulton, C., Masson, C., Häggström, C., Fitzgerald, C., Nicholson, D. A., Hagen, D. R., Pasechnik, D. V., Olivetti, E., Martin, E., Wieser, E., Silva, F., Lenders, F., Wilhelm, F., Young, G., Price, G. A., Ingold, G.-L., Allen, G. E., Lee, G. R., Audren, H., Probst, I., Dietrich, J. P., Silterra, J., Webber, J. T., Slavič, J., Nothman, J., Buchner, J., Kulick, J., Schönberger, J. L., de Miranda Cardoso, J. V., Reimer, J., Harrington, J., Rodríguez, J. L. C., Nunez-Iglesias, J., Kuczynski, J., Tritz, K., Thoma, M., Newville, M., Kümmerer, M., Bolingbroke, M., Tartre, M., Pak, M., Smith, N. J., Nowaczyk, N., Shebanov, N., Pavlyk, O., Brodtkorb, P. A., Lee, P., McGibbon, R. T., Feldbauer, R., Lewis, S., Tygier, S., Sievert, S., Vigna, S., Peterson, S., More, S., Pudlik, T., Oshima, T., Pingel, T. J., Robitaille, T. P., Spura, T., Jones, T. R., Cera, T., Leslie, T., Zito, T., Krauss, T., Upadhyay, U., Halchenko, Y. O., Vázquez-Baeza, Y., and SciPy, C. (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **17**, 261-272
48. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. (2011) Scikit-learn: Machine Learning in Python. **12**, 2825–2830
49. Hastie, T., Tibshirani, R., and Friedman, J. (2009) *Elements of Statistical Learning Ed. 2*, Springer
50. Beghini, F., McIver, L. J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Manghi, P., Scholz, M., Thomas, A. M., Valles-Colomer, M., Weingart, G., Zhang, Y., Zolfo, M., Huttenhower, C., Franzosa, E. A., and Segata, N. (2021) Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife* **10**
51. Friedman, J. H. (2001) Greedy function approximation: A gradient boosting machine. **The Annals of Statistics**, 1189-1232, 1144
52. Chen, T., and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, San Francisco, California, USA
53. Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., and Huttenhower, C. (2011) Metagenomic biomarker discovery and explanation. *Genome Biol* **12**, R60
54. Lemaître, G., Nogueira, F., and Aridas, C. K. (2017) Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. **18**, 559–563

55. Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002) SMOTE: synthetic minority over-sampling technique. **16**, 321–357
56. Metz, C. E. (1978) Basic principles of ROC analysis. *Seminars in nuclear medicine* **8**, 283-298
57. Holm, K., Hegardt, C., Staaf, J., Vallon-Christersson, J., Jönsson, G., Olsson, H., Borg, A., and Ringnér, M. (2010) Molecular subtypes of breast cancer are associated with characteristic DNA methylation patterns. *Breast cancer research : BCR* **12**, R36
58. Akhoondi, S., Lindström, L., Widschwendter, M., Corcoran, M., Bergh, J., Spruck, C., Grandér, D., and Sangfelt, O. (2010) Inactivation of FBXW7/hCDC4-β expression by promoter hypermethylation is associated with favorable prognosis in primary breast cancer. *Breast cancer research : BCR* **12**, R105
59. Ibusuki, M., Yamamoto, Y., Shinriki, S., Ando, Y., and Iwase, H. (2011) Reduced expression of ubiquitin ligase FBXW7 mRNA is associated with poor prognosis in breast cancer patients. **102**, 439-445
60. Wei, G., Wang, Y., Zhang, P., Lu, J., and Mao, J. H. (2012) Evaluating the prognostic significance of FBXW7 expression level in human breast cancer by a meta-analysis of transcriptional profiles. *Journal of cancer science & therapy* **4**, 299-305
61. Hata, S., Abe, M., Suzuki, H., Kitamura, F., Toyama-Sorimachi, N., Abe, K., Sakimura, K., and Sorimachi, H. (2010) Calpain 8/nCL-2 and calpain 9/nCL-4 constitute an active protease complex, G-calpain, involved in gastric mucosal defense. *PLoS genetics* **6**, e1001040
62. Buhmeida, A., Dallol, A., Merdad, A., Al-Maghrabi, J., Gari, M. A., Abu-Elmagd, M. M., Chaudhary, A. G., Abuzenadah, A. M., Nedjadi, T., Ermiah, E., Al-Thubaity, F., and Al-Qahtani, M. H. (2014) High fibroblast growth factor 19 (FGF19) expression predicts worse prognosis in invasive ductal carcinoma of breast. *Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine* **35**, 2817-2824
63. Liu, Y., Cao, M., Cai, Y., Li, X., Zhao, C., and Cui, R. (2020) Dissecting the Role of the FGF19-FGFR4 Signaling Pathway in Cancer Development and Progression. **8**
64. Wallace, B. D., Wang, H., Lane, K. T., Scott, J. E., Orans, J., Koo, J. S., Venkatesh, M., Jobin, C., Yeh, L. A., Mani, S., and Redinbo, M. R. (2010) Alleviating cancer drug toxicity by inhibiting a bacterial enzyme. *Science (New York, N.Y.)* **330**, 831-835
65. Ervin, S. M., Li, H., Lim, L., Roberts, L. R., Liang, X., Mani, S., and Redinbo, M. R. (2019) Gut microbial β-glucuronidases reactivate estrogens as components of the estrobolome that reactivate estrogens. *The Journal of biological chemistry* **294**, 18586-18599
66. Terrisse, S., Derosa, L., Iebba, V., Ghiringhelli, F., Vaz-Luis, I., Kroemer, G., Fidelle, M., Christodoulidis, S., Segata, N., Thomas, A. M., Martin, A.-L., Sirven, A., Everhard, S., Aprahamian, F., Nirmalathasan, N., Aarnoutse, R., Smidt, M., Ziemons, J., Caldas, C., Loibl, S., Denkert, C., Durand, S., Iglesias, C., Pietrantonio, F., Routy, B., André, F., Pasolli, E., Delaloge, S., and Zitvogel, L. (2021) Intestinal microbiota influences clinical outcome and side effects of early breast cancer treatment. *Cell Death & Differentiation* **28**, 2778-2796
67. Ruo, S. W., Alkayyali, T., Win, M., Tara, A., Joseph, C., Kannan, A., Srivastava, K., Ochuba, O., Sandhu, J. K., Went, T. R., Sultan, W., Kantamaneni, K., and Poudel, S. (2021) Role of Gut Microbiota Dysbiosis in Breast Cancer and Novel Approaches in Prevention, Diagnosis, and Treatment. *Cureus* **13**, e17472-e17472

68. Zhu, J., Liao, M., Yao, Z., Liang, W., Li, Q., Liu, J., Yang, H., Ji, Y., Wei, W., Tan, A., Liang, S., Chen, Y., Lin, H., Zhu, X., Huang, S., Tian, J., Tang, R., Wang, Q., and Mo, Z. (2018) Breast cancer in postmenopausal women is associated with an altered gut metagenome. *Microbiome* **6**, 136-136
69. Asnicar, F., Berry, S. E., Valdes, A. M., Nguyen, L. H., Piccinno, G., Drew, D. A., Leeming, E., Gibson, R., Le Roy, C., Khatib, H. A., Francis, L., Mazidi, M., Mompeo, O., Valles-Colomer, M., Tett, A., Beghini, F., Dubois, L., Bazzani, D., Thomas, A. M., Mirzayi, C., Khleborodova, A., Oh, S., Hine, R., Bonnett, C., Capdevila, J., Danzanvilliers, S., Giordano, F., Geistlinger, L., Waldron, L., Davies, R., Hadjigeorgiou, G., Wolf, J., Ordovás, J. M., Gardner, C., Franks, P. W., Chan, A. T., Huttenhower, C., Spector, T. D., and Segata, N. (2021) Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. *Nat Med* **27**, 321-332
70. Rashidi, A., Ebadi, M., Weisdorf, D. J., Costalonga, M., and Staley, C. (2021) No evidence for colonization of oral bacteria in the distal gut in healthy adults. **118**, e2114152118
71. Demirci, M. J. E. J. o. M., and Oncology. (2021) Dialister in Microbiome of Cancer Patients: A Systematic Review and Meta-Analysis.