Host- Microbiome Omics Integration for Cancer Analysis And Diagnostics

# Todo

[pay more attention to the sentence structure, begin with most important message in paragraphs]

# Abstract

# Introduction

[

why is cancer important?

Increasing data facilitates cancer diagnostic research

host omics integration

* TCGA data set
* is used for cancer diagnostics but comes with challenges

microbial omics integration

* what is microbiota?
* What data sets are there?
  + Raw data sets
  + Mined from TCGA
* microbiota differs per person, cancer is environmentally affected
* it has an effect on health
* types of analysis done in research (e.g. taxonomical analysis, meta proteomics)

the need for a holistic view

* studies use one or the other but not both
* both are required
* what have studies so far done?
* Available methods are lacking and challenges for such methods
* benefit a new study could bring

Towards a holistic view

* what we aim to do?
  + What question are we answering?
* How do we aim to do it?

Cancer is one of the leading causes of death and is responsible for millions of deaths every year. The aetiology, morphology and progression of different cancers depends on a complex interplay of various biological and environmental factors. Recently, it is becoming increasingly easy to investigate this complex interplay thanks to the development of more modern sequencing technologies and availability of biological data. This data availability has made it more accessible for researchers to use various omics data to perform various tasks related to cancer diagnostics. This relates to the analysis and integration of both host and microbial omics data.

## Host multi-omics integration

The availability of host omics data of various data layers, such as gene expression, and DNA methylation or copy number variation has enabled researchers to derive useful insights on the aetiology and morphology of different cancers. While there are many such data sources that have been made available, one of the most impactful sources has been The Cancer Genome Atlas (TCGA), a repository of genomic profiles of over 30 types of cancer that can be used for cancer diagnostics 1.

While using individual omic types can derive useful insights, an important development has been the usage of so-called multi-omics analyses methods, where data from multiple omics layers are integrated to deliver additional insights over single omics methods for cancer diagnostics. While more powerful, these methods deal with additional challenges owing to the heterogeneity of the data, noise, high dimensionality and sparsity of the multi-omics data 2.

## Microbiome multi-omics integration

Next to host omics data, a promising field of research relates to the analysis of microbial omics data. There are many microorganisms which live in communities on different human tissues, called the human microbiota. Namely, an ecosystem of 10 to 100 trillion microorganisms encompassing 500 to 1000 unique species for each individual 3. Due to the aforementioned advances in sequencing technology, it is becoming increasingly easy to measure the identity, metabolic potential and expression of this microbiota. This is leading to various data sets on the human microbiota which can be exploited.

Some popular data sets include MetaHIT, the human microbiome project, and TCMA, a microbial data set derived from TCGA. The Metagenomics of the Human Intestinal Tract (MetaHIT) data set contains sequenced intestinal and stool microbial data from healthy patients and those with certain noncancer diseases 4. The integrative human microbiome project (iHMP) is the second stage of the human microbiome project and contains data on both host and microbial omics layers and establishes a baseline omics composition across varying populations, and across a population with specific (mostly non cancer-related) disease states 5.

While the above mentioned data sets are crucial, they often contain data from tissue swabs and stool samples, which are not necessarily representative of the microbiome of internal organs 3. Next-generation sequencing data sets, such as TCGA, contain, next to host sequencing data, microbial sequencing data. This aspect of TCGA is mostly unexplored yet can be mined to obtain data on, for example, viromes and bacteriomes of different cancers using different tissues, such as blood, which is most likely to contain useful microbial contamination 6. However, the microbial reads in this data set are often a result of contamination 3. One such data set which makes this data available, is the cancer microbiome Atlas (TCMA).

Using this data, various aspects of the human microbiome are being investigated. It has become clear that the microbiome exhibits significant person-to-person variation 7 and can be affected by many factors such as diet, environmental exposure and lifestyle choices 8. Additionally, microbial communities are unique to each cancer type 6. This has provided a motivation for the analyses of the impact of microbial composition on disease phenotypes and progression.

In light of this, recent research has been elucidating the effect that the microbiota can have on a host organism 9. It has become clear that the microbiota can have a clear effect on the phenotype of humans. Many microorganisms are enriched in certain cancers or differentially expressed among healthy and unhealthy individuals 10 and between tumor versus normal samples 11. Even further, there are indications that composition and changes in the microbiota has a direct influence on oncogenesis 12. As an example, patients with Parkinson’s have a relatively lower incidence of multiple cancers, possibly through mechanisms involving the microbiota 13. It has also been revealed that there are significantly different abundances of microbes, which are also associated with different functional categories, among different disease groups for cervical cancer, even when controlling for multiple other clinical variables 10. As an example of a mechanism through which the microbiome can affect a patient, it has been discovered that certain bacteria can bind to and alter the function of immune system cells which infiltrate tumors, thereby affecting carcinogenesis and resistance to chemotherapy 8. However, they can also positively affect patient health, for example, by positively influencing immune system cells to promote antitumor immunity.

After obtaining the data, analysis of microbial data is often done through taxonomic analysis, which can use 16S rRNA sequencing data to analyze the types of microbes present in a tissue, or functional analysis, which studies the function of present microbes by identifying and characterizing exons or analyzing metabolites. Here as well, analyses has to deal with multiple challenges, one of the biggest being batch effects, as the same tools can give different results.

Taxonomical analysis often investigates whether certain taxa or species of bacteria are overexpressed in different cancer or healthy samples using hypotheses tests 3,6,14,15, possibly also by using predictor models (e.g. regression) 6,15. It is also possible to analyze whether certain subgroups of microbiota are more likely to be found together in one tissue versus another or analyze if they are associated with certain symptoms of a certain cancer 3. Additionally, hypothesis tests can be used to see if microbial abundance is predictive of overall survival 3,16 or progression 16. Studies often also investigate the relation with clinical factors such as gender or age 14–16, which can often be confounders. This can be done by using predictor models which use these clinical factors as features 15. Finally, it is also possible to combine metaproteomics with metagenomics to investigate functional, as well as taxonomical differences of microbes between healthy and diseased samples 17.

## The need for a holistic view

It is clear that both host omics and microbial omics data can be used to obtain useful biological insights into the aetiology of different cancers. As shown, many studies use one or the other to understand different biological processes without considering their interplay 9. However, it has become clear that the host can alter the human microbiota and vice versa 18. Thus, the integration of host and microbiota omics data could help to better understand the aetiology and physiology of different cancers and provide new insights 12. This field, where a holistic approach is taken to biological data, is known as hologenomics. It is based on the assumption underlying the hologenome theory, which posits that the host and microbial genome are biologically dependent and must be analyzed together in order to investigate the phenotype of an organism 19.

In terms of hologenomics, previous studies have attempted to combine microbial and host omics data to investigate correlations between bacterial co-abundance groups and host gene expression patterns 3,16, host mutations 14 or proteins 3. Greathouse et al. examines the interaction between microbiota and TP53 in lung cancer by investigating the abundance and diversity of specific microbial species in lung tumors with TP53 mutations with TCGA(abundance) and NCI-MD data using statistical tests 14. Similarly, Chakladar et al. combine rRNA sequencing data from TCGA and intra-pancreatic microbe abundance data mined from TCGA to investigate cancer associated genes and pathways for pancreatic adenocarcinoma 16. Specifically, they pair abundance data with clinical variables and cancer and immune associated gene expression to determine if the up or down regulation of certain pathways is correlated with certain microbes using GSEA. Finally, Dohlman et al. use TCMA and various TCGA omics data to investigate correlations between bacterial co-abundance groups and gene expression patterns of certain genes, including through the use of GSEA of these correlated genes with 3.

There are a lack of specialized methods which are able to perform this kind of integrated analysis even though the development of such tools could provide helpful new insights 9. Such a method would have to deal with multiple challenges, namely the nonuniformity or linearity of this interaction, the high dimensionality compared to low amount of samples and missing values 9. Next to this, it would also have to deal with problems typical to multi-omics integration such as high heterogeneity of data and noise. Finally, care would have to be taken as microbial samples can be plagued with batch effects and contaminants 3.

The development of such methods could help power multiple diagnostic goals, such as predicting cancer response to therapy by elucidating why certain immunotherapies work or fail in patients, providing insights into how cancers develop, aid in the development of microbial consortia to push out disease associated microorganisms from a gut or tumor, identify targets for vaccines or targets for therapies which reduce the amount of certain microorganisms, such as antibiotics or phage-based therapy 8. Finally, such methods could also be used to identify biomarkers and predict cancer versus normal samples.

As stated, there are not a lot of studies which directly identify the relation between host and microbial omics data, and certainly not how both datatypes relate to patient phenotype. As host multi-omics integration for cancer diagnostics is already a thriving field, one promising direction could be to use host multi-omics integration methods for holo- omics data. As holo- omics data generates highly complex data sets which require feature reduction 18, a possible method which can be investigated is the use of autoencoders to reduce and extract features which could then possibly be used in a predictor model.

## Towards a holistic view

This paper aims to integrate host and metagenomics data for cancer patients in order to investigate whether a holistic view provides additional insights for cancer diagnostics versus simply using the layers individually. Specifically, it aims to investigate the question:

Does integrating host and microbial omics data provide additional power over using the individual layers?

In this case, power refers to prediction performance.

To this end, we leveraged the powerful TCGA and TCMA data sets and integrated gene expression and microbial genus abundance data for colon adenocarcinoma (COAD), esophageal carcinoma (ESCA), head and neck squamous carcinoma (HNSC) and stomach adenocarcinoma (STAD). We then investigated whether the integration of each layer provides additional prediction performance for tumor versus normal prediction and stage prediction then when simply using the individual layers separately.

# Materials and methods

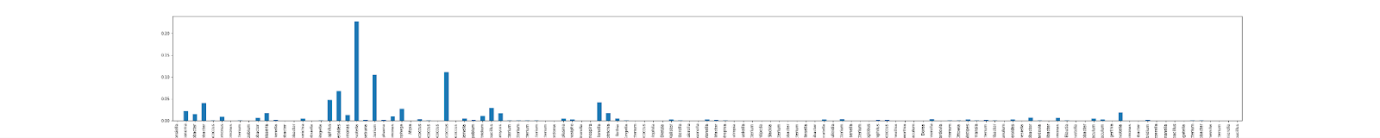
## Data

In order to explore the performance of cancer diagnostics in a holistic context, we integrated microbial data mined from patient tumor samples and host omics data mined from these same samples. We then only keep samples for which there is both microbial, as well as host omics data.

### Microbial data

This study used a statistical model to isolate tissue embedded microbial species present in TCGA samples from contaminants to obtain this data set, which was then subsequently validated using 16S rRNA amplicon sequencing on the original tissue samples. The resulting TCMA database accessible through the TCMA portal [[1]](#footnote-1), contains tissue resident microbial relative abundance data for 3689 unique samples and 1772 patients across 21 anatomic sites and 5 TCGA projects (HNSC, ESCA, STAD, COAD and READ).

The highest specification that this data set reaches is the genus level. As the phylum level and everything above genus is less specific than the genus level, we decided to continue with only the genus taxonomical abundance data, which contains 221 taxa. The following is a visual exploration of the abundance data. There are 119 nonzero features among the overlapped samples, which can be seen below.



### Host omics data

The host omics data consisted of a TCGA data set which was extracted and processed by a prior study 20. This data set consists of level 3 RNA-seq gene expression data for 9732 tumors and 727 tumor adjacent normal samples encompassing 33 total different cancer types. The TCGA RNA-seq was obtained using the UCSC Xena data browser on March 8, 2016. The expression values of these genes consists of a pre-processed and batch-corrected gene abundance x sample matrix with RSEM values normalized using a log2(FPKM + 1) transformation. The gene expression values of the 5000 genes with the highest variability were used, as evaluated using Median Absolute Deviation (MAD). Finally, the expression values are also min – max scaled.

In further experiments involving the gene expression-only modality, we only used those patients for which both host omics, as well as microbial data was available.

### Clinical data

The clinical data was accessed using the Snaptron web server. We match the clinical data with the corresponding patient samples in order to obtain details for the tumor and stage endpoints. To determine whether a sample is tumor or normal, the sample type code [[2]](#footnote-2) is used, where codes in the range 01 – 09 are tumors and those in the range 10 – 19 are normal samples [[3]](#footnote-3). The stage clinical data is used to determine the tumor stage of each sample. Normal samples are considered as Stage 0.

### Overlapped data

In order to investigate the effects of microbial data on cancer diagnostics, we create an overlapped set samples for which there is both host- and microbial omics data in the above-described data sets.

In the first step, the TCGA gene expression data is joined with the clinical data. To do this, the “portion\_id” field of each row in the clinical data set, which contains the code for the project, the tissue source site (TSS), participant ID, sample type, vial, and portion id is used. This field is stripped of the portion ID and the vial, the duplicates are dropped (to remove samples with the same ID but different measuring technologies), and each row is then joined with the samples in the GE data set, which contains the same attributes up to and including the sample type.

To join the data set with GE + clinical data to the microbial dataset, the sample barcode for the microbial samples is stripped of the all samples with a vial type of “B”, and then the vial type attribute is removed altogether. This data set is then joined with the GE + clinical data set.

The table below displays the class balance of tumor versus normal samples for each modality and the integrated set.

Table

Description automatically generated

The next table displays the balance for the stage samples. Normal samples are classified as stage 0.

Table

Description automatically generated

Due to the lack of samples for READ, we use data for the cancers COAD, ESCA, HNSC and STAD.

## Methods

### Exploration of holistic view

Feature exploration was performed using PCA and t-SNE. PCA was performed using the *scikit-learn* PCA function with 2 components. t-SNE was performed using the *scikit-learn* t-SNE function using the default settings, including 1000 maximum iterations. A static random seed was used for all experiments.

Feature reduction was performed using the chi2 test Anthony regression

### Complex modality integration

for the integration

#### Nonnegative Matrix Factorization

For the integration using nonnegative matrix factorization, the NMF function from the *scikit-learn* decomposition package was used. This was set to use 30 and order to allow comparison with the autoencoder integration method. We selected the initialization to be random, and supplied a static random seed to allow for reproducibility between experiments.

#### Autoencoder

For the integration using autoencoder, we used *PyTorch* to define the autoencoder architecture and train the model. We use the MSE loss function and Adam as the optimizer.

For hyper parameter tuning, we used the *Skorch* package to wrap the Autoencoder module into an *Scikit-learn* compatible model. The *scikit-learn* GridSearchCV package was then used to perform hyper parameter tuning with an inner cross-validation of 5 folds.

### Predictive performance

For the stage endpoint, the prediction of the different stages 0-4 it is modeled as a regression, rather than a classification problem. The prediction of this endpoint was performed using the *scikit-learn* Random Forest Regressor model and the Elastic Net model, both of which were initialized with a random seed of 0.

For the tumor endpoint, the prediction was modeled as a binary classification problem, and the *scikit-learn* Support Vector Machine model was used with a random seed of 0.

#### Prediction pipeline

For the prediction pipeline, each experiment was performed for each combination of cancers (i.e. COAD, ESCA, HNSC, STAD), for each modality (i.e. gene expression, genus and the concatenation of gene expression and genus).

We used a random sampling approach to obtain an estimation of how a prediction model performed when used on each data modality separately, compared to how the model performed when using a combination of the modalities using various integration methods.

For each of these combinations, a random stratified split was performed to split the data into 80% training and 20% testing. This split was performed 200 hundred times by using the *scikit-learn* train\_test\_split function. Each split iteration was assigned a custom seed to ensure consistency between experiment runs.

Using 80% of the data, this data set portion was used to performed feature selection for various feature amounts (i.e. 6, 10, 26, 50, 100, 200), up to the maximum amount of features present within the modality. Each experiment was also carried out when using no feature selection. The selected feature amounts of the top selected features and thus are subsets of each other. For example, the top 6 features are contained within the top 10 features which are in turn contained within the top 26 features etc.

The feature selection was performed using linear regression and the chi-squared test. For the linear regression, a model was trained using the feature values of the samples along with the targets. The magnitude of the coefficients for each feature value was then used to obtain a feature ranking.

For the chi-squared test, the chi-squared test is simply used to obtain the feature rankings.

After obtaining these top features, a prediction model is trained on the 80% training set using only the top features that were selected.

If hyper parameter tuning is used, then a stratified 5-fold cross validation split is used on the 80% training set to tune hyperparameters using the *scikit-learn* RandomizedSearchCV package the RMSE score. A random seed is used which is equal to the random seed used for the current random sampling iteration which initiated the hyper parameter tuning. The optimal hyper parameters are then found using 100 random search iterations and the hyper parameter set which obtained the best performance during the cross validation procedure is then used to train on the entire 80% training set.

After obtaining this final model trained on the 80% training set, it is tested on the 20% testing set.

#### Modality enforcement

Modality enforcement was performed by separating the GE and Genus modality and performing feature selection for each feature amount separately for each modality before concatenating the selected features again and then training and evaluating a model. If

# Results

## Characterization of data

To investigate the effects of the host-omics approach, we used preprocessed samples from the Cancer Genome Atlas (TCGA). It turns out that gene expression and DNA methylation data works well with multiple cancers 21. [Insert more evidence of why GE without other stuff is good] Thus, in order to examine the benefits of a holistic view on omics data for cancer diagnostics, gene expression features from TCGA are used. This set contains 5000 features for each sample.

To investigate the effects of the microbial omics approach, we used the Cancer Microbiome Atlas (TCMA). This is a microbial database which is based on data that is mined and processed from reads contained in TCGA, which attempts to identify microbial species in tissue and blood samples while dealing with contaminants 3. TCMA contains microbial data available from multiple cancers allowing for cross cancer analyses using the same source with common methodologies.

Finally, for the holo-omic approach, we overlapped samples for which there is both host omics and microbial omics data available.

## There is little quantitative performance improvement with the holo-omic approach

To investigate the possible benefits of a holistic view of omics integration, we used the different omics layers and microbial data in a predictive model for numerous cancer diagnostic endpoints. The performance of the predictive model in different settings was used as a measure of the usefulness of each modality.

We used random sampling to randomly split the data set with a stratified split into 80% training and 20% testing 200 times. In each iteration, a feature selection method is first performed on the training set. These features are then used to train a model on this same training set. The performance of the model is then tested on the testing set for the available endpoints. This is repeated for each cancer, layer and feature amount. The figure below displays the combinatorial options of the pipeline (Figure 1).

Diagram

Description automatically generated

Figure : the different combinatorial options of the predictions pipeline

The prediction models are then evaluated using the precision, recall and f1-score. This pipeline thus generates n different sets of values for these evaluation metrics. For the evaluation, we consider the average of the metric across each class value. For each combination of the value types above, the average of across the n random sampling iterations are then plotted along with their standard deviation across these iterations.

This was done for stage prediction and tumor versus normal prediction. The predictive performance of the GE model separately, and the microbial data model were used as a baseline. The integration of both of these data sets were then compared to this baseline.

For the tumor prediction, we chose an SVM model as it is a well-established method for classification that has shown effectiveness across multiple prior studies 22–24. For stage prediction, we model it as an ordinal categorical variable and chose an elastic net model for the same reason 22.

Prior to generating a prediction model, we performed feature selection in order to reduce the effect of noisy variables and only include the most powerful predictors, which has been shown to improve prediction performance 25.

For tumor prediction, we used the Chi-square test, which is a commonly used feature selection method for categorical variables which has previously been successfully used to select important gene features 26.

For dual class and multiclass prediction targets, partial least squares regression based methods have also seen some successful use for gene feature selection 24. Penalized regression methods work well for biological data sets, where there are more features than samples 25. In particular LASSO seems to work particularly well 25, as well as LASSO-based methods 27

### Holo-omic approach does not lead to improved tumor prediction performance

The results indicate that integrating gene expression with microbial taxonomic data does not lead to a significant improvement in prediction performance over using gene expression data alone (Figure 4).

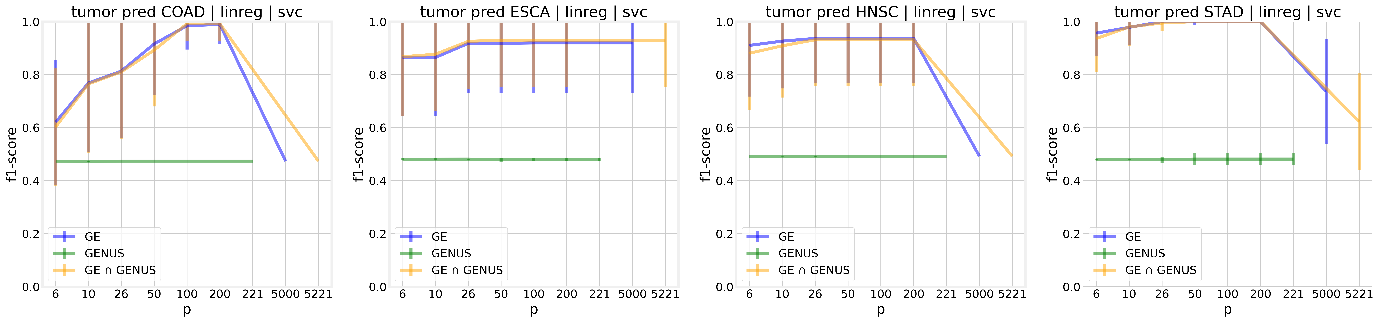


Figure : f1-score for the tumor versus normal prediction endpoint for COAD (colon adenocarcinoma), ESCA (esophageal carcinoma), HNSC (head and neck squamous carcinoma) and STAD (stomach adenocarcinoma), respectively. Each graph contains the f1-score for each modality (i.e. genus abundance data (GENUS), gene expression data (GE) and the concatenated genus + gene expression features (GE ∩ GENUS)). The endpoints of each horizontal line segment indicate the average f1-score across every random sample iteration, while the vertical line segments indicates the standard deviation of the f1-score across these iterations.

This result is consistent for tumor versus normal prediction, but also stage prediction Figure 5.

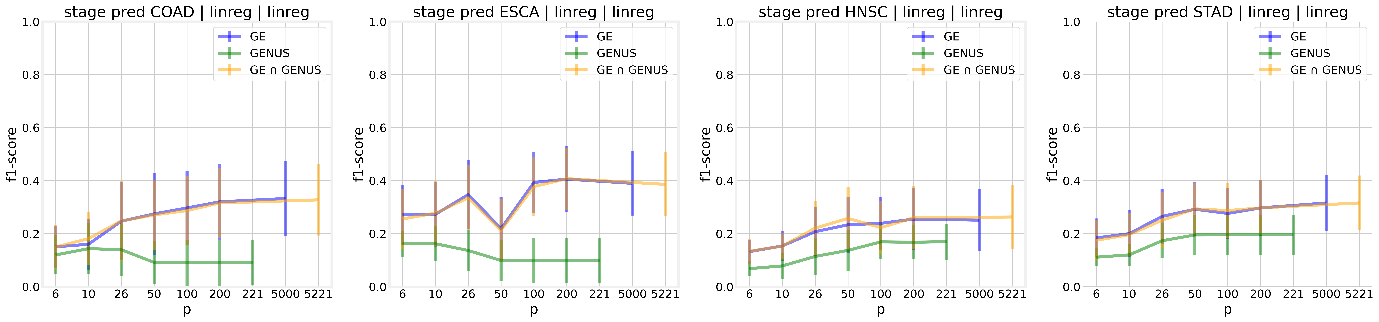


Figure : f1-score for the stage prediction endpoint for COAD (colon adenocarcinoma), ESCA (esophageal carcinoma), HNSC (head and neck squamous carcinoma) and STAD (stomach adenocarcinoma), respectively. Each graph contains the f1-score for each modality (i.e. genus abundance data (GENUS), gene expression data (GE) and the concatenated genus + gene expression features (GE ∩ GENUS)). The endpoints of each horizontal line segment indicate the average f1-score across every random sample iteration, which in turn consists of the average f1-score for each stage class. The vertical line segments indicates the standard deviation of the f1-score across these iterations.

These results also seem to be consistent across the different amount of selected features.

Throughout this paper, we chose to only continue displaying the following experiments for STAD (stomach adenocarcinoma), as it is the cancer within the available data set with the most simple aetiology and the highest amount of samples and balance between classes. However, the same experiments have also been run for other cancers, and the results of these can be found in the appendix.

### Genus layer provides worst performance

The genus layer seems to provide the worst performance. This is possibly because the taxonomic genus data is mined from the tumor samples and might not be directly related to the aetiology of the disease. Thus, it might not contain enough predictive power by itself. It is worth noting that this does not necessarily mean that microbial data might not be enough to predict these endpoints, but rather that the specific microbial data collected might need additional non-microbial features to be powerful.

Interestingly, this difference is smaller for stage prediction. This is possibly because of the extreme imbalance of classes in the data set. The more classes there are, the harder it is to maintain a balance. Thus, for stage prediction, all the models achieve relatively low performance.

This is relatively consistent with previous results of prediction models using TCGA microbial data 6.

### Results are independent of prediction model

To investigate whether the results are due to the prediction model not being able to properly capture the information contained within the individual and overlapped layers and the interaction between these layers, we ran the prediction pipeline using multiple prediction models and hyper parameter tuning.

The results indicate that the lack of performance is not due to the model not being able to properly capture complex nonlinear relationships.

We ran the main experimental setup using a hyper parameter tuned elastic net model (Figure 6). As can be seen, this leads to similar results, namely that there is no improvement when integrating the two modalities.

Chart, line chart

Description automatically generated

Figure : RMSE for the stage prediction endpoint for STAD using an elastic net model.

We have also repeated the same experiment using a random forest model with similar results (Figure 7).

Chart, line chart

Description automatically generated

Figure : RMSE for the stage prediction endpoint for STAD using a random forest model.

It appears that the random forest model achieves a similar root mean squared error as the elastic net model. Yet, neither of the models offer additional improvement when integrating the two modalities.

## Lower performance is not due to feature selection

The feature selection method and process has a significant impact on the performance and results of prediction models [CITATION HERE]. To investigate whether the previous results were due to the feature selection method or process chosen, we investigated different types of feature selection methods and validated the selected features to ensure that the previous results were not due to the feature selection pipeline.

### Results independent of feature selection method

To investigate whether the lower performance was due to the feature selection method, we attempted the experiments while using the chi2 test as well (Figure 7). Using the different feature selection method offers no additional improvement.

Chart

Description automatically generated

Figure : Cancer predictions for stage using chi2

### Feature selection is dominated by one modality

To investigate the distribution of the types of features selected for each feature selection method and feature amounts, the proportion for which each layer is represented in the final feature sets is investigated by saving the selected features for each combinatorial state of the pipeline in each iteration. The proportion is then plotted for all the different cancers for each feature amount.

It seems that when performing feature selection on the integrated data set, most of the features selected originate from the gene expression set (Figure 7).

One possible reason for the domination of gene expression features is that the original data set contains much more gene expression features than taxonomic genus features (more than 20 times as much). Thus, there are simply more gene expression features to select from. It is clear that the microbial features are still useful, given that they are still reasonably well represented when selecting the top 5 or 10 features.

The chi-square test is not model for dependency between features.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9580915/> there is no best feature selection and linear regression doesn’t automatically model dependencies ()

#### GE modality dominance is smaller in stage prediction

The dominance of the gene expression modality seems to be smaller for stage prediction, possibly because there is a wider variety of endpoint values and thus relationships between input and output to consider during the feature selection process.

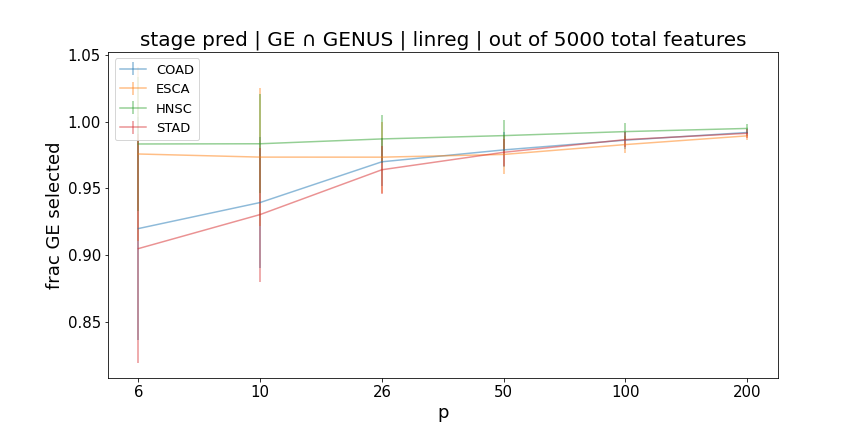


Figure : the fraction of GE features (vertical axis) selected from the total amount of features for each feature amount (horizontal axis) when doing stage prediction with the genus + gene expression modality (GE ∩ GENUS). Each line displays the results for a different cancer, namely (COAD (colon adenocarcinoma), ESCA (esophageal carcinoma), HNSC (head and neck squamous carcinoma) and STAD (stomach adenocarcinoma). These results are for the linreg feature selection method.

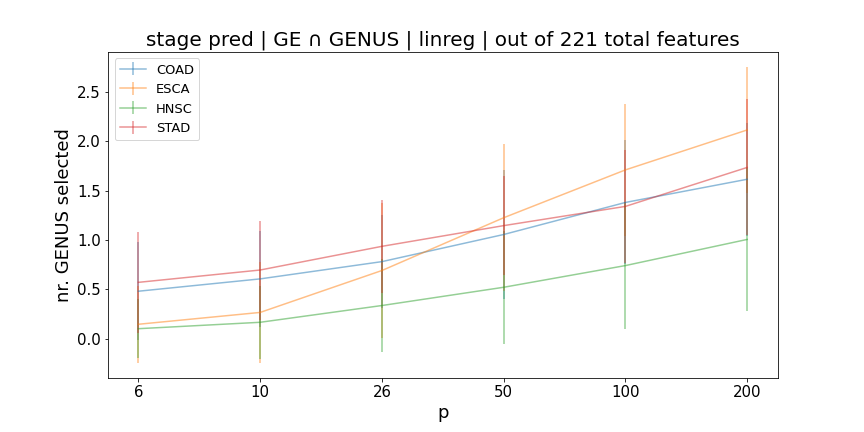


Figure : the absolute amount of GENUS features (vertical axis) selected from the total amount of features for each feature amount (horizontal axis) when doing stage prediction with the genus + gene expression modality (GE ∩ GENUS). Each line displays the results for a different cancer, namely (COAD (colon adenocarcinoma), ESCA (esophageal carcinoma), HNSC (head and neck squamous carcinoma) and STAD (stomach adenocarcinoma). These results are for the linreg feature selection method.

One possible way to deal with this is to correct for the amount of base features in each modality. Additionally, it is possible to use feature selection on the original modalities separately prior to concatenating.

When selecting features with linreg for HNSC for tumor versus normal prediction, one microbial genus which is consistently selected is Prevotella. Across iterations, even when only 6 features are selected, this one is selected.

For linreg/HNSC/Stage prediction, in certain iterations with 6 or 10 features selected, the microbial Genus which is consistently selected (when one is selected) is Fusobacterium.

This is consistent with previous studies confirming the association between Fusobacterium and oral squamous cell carcinoma 28.

The same counts for linreg for COAD for stage prediction with 6 features but him with Bacteroides.

[Maybe make a table with all the combinatorial combinations along with distribution of selected or the most frequently selected genus] maybe try univariate for PERMANOVA tests? (

### Enforcing modality parity during feature selection does not improve performance

To determine whether the domination of one modality in the features selected has an effect on the added performance of the holo-omic approach, we repeated the above experiments while enforcing parity in the amount of features selected of each modality. To do this, we performed feature selection prior to integrating the modalities to ensure that the same amount of features are selected of each modality.

As can be seen, this leads to similar or worse performance than only using GE or the non-enforced-parity feature selection approach with the overlapped layer (Figure 4).

[Display performance of integrated modality versus individual modalities when enforcing modality parity]

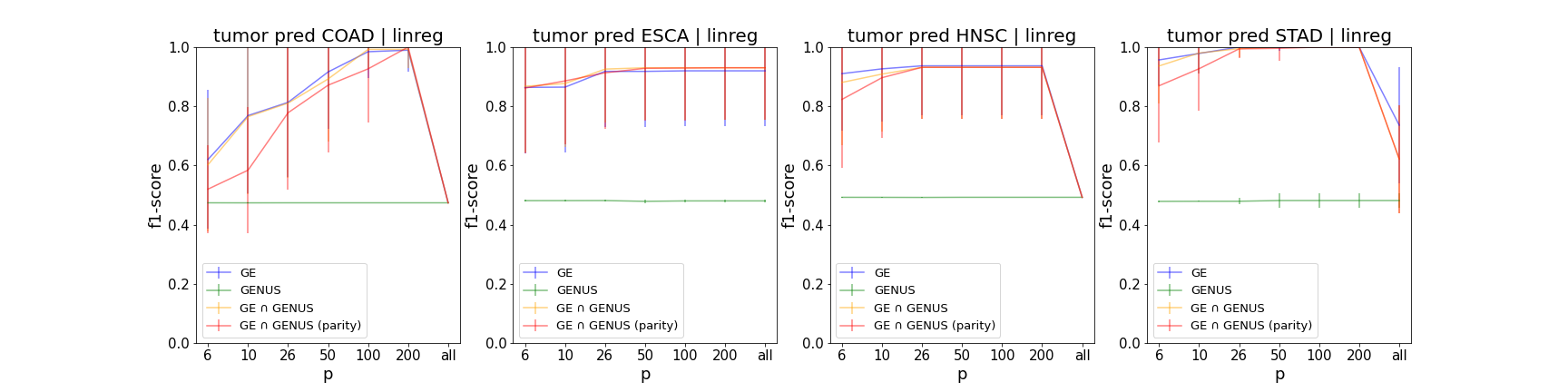


Figure : f1-score for the tumor versus normal prediction endpoint for COAD (colon adenocarcinoma), ESCA (esophageal carcinoma), HNSC (head and neck squamous carcinoma) and STAD (stomach adenocarcinoma), respectively. Each graph contains the f1-score for each modality (i.e. genus abundance data (GENUS), gene expression data (GE) and the concatenated genus + gene expression features (GE ∩ GENUS), and concatenated genus + gene expression features with enforced parity (GE ∩ GENUS (parity))). The endpoints of each horizontal line segment indicate the average f1-score across every random sample iteration, while the vertical line segments indicates the standard deviation of the f1-score across these iterations.

### Selected genus features are supported by research

To investigate whether the lack of performance was due to the features selected, we investigated the top features selected by the feature selection algorithm across the different cancers (Figure 15).

One genus species which was frequently selected in the overlap set for stomach adenocarcinoma, is *Helicobacter*. This seems to be validated by previous studies linking this cancer to *Helicobacter pylori* 29–31. This bacteria can induce gastritis, which can then lead to stomach adenocarcinoma. Interestingly, this bacteria has a possible protective effect against esophageal adenocarcinoma 29.

For colorectal adenocarcinoma, a frequently selected genus was *Bacteroides*. A previous study has found that *Bacteroides fragilis* was disproportionately present in tumor and adjacent non-tumor tissues of colorectal cancer patients compared to other investigated bacteria and significantly higher in tumor tissues than normal samples 32.

For esophageal cancer and head and neck squamous carcinoma, the most frequently selected genus was *fusobacterium*. Previous research has found that *Fusobacterium Nucleatum* is significantly associated with tumor samples and with tumor stage in esophageal cancer, while controlling for clinical confounders 33. Similar results have held for head and neck squamous carcinoma, with *Fusobacterial* populations showing an increased abundance in tumor versus normal samples 34.

Application

Description automatically generated with low confidence

Figure : top genus features selected of the GE ∩ GENUS data set with linear regression feature selection with a feature selection number of 10. The table rows denote the name of the top selected genus features, while the Frequency column denotes how many times the feature was selected across 200 random sampling iterations.

## No improvement from complex integration

In order to determine whether the lack of performance when integrating the two modalities was due to the simple, concatenation-based integration method, we attempted to integrate the two modalities using a more advanced and proven integration method. Namely, an autoencoder and nonnegative matrix factorization.

### Autoencoder

Recently, autoencoders have seen much use in bioinformatics. Chaudhary et al. integrated host multi-omics data using an autoenoder for liver cancer survival rate prediction and subtype classification 23 as well as other prediction tasks 22. As it showed effectiveness, we based our autoencoder architecture and strategy on the paper by Chaudhary et al.

To investigate whoever the lack of integration performance was due to the integration method chosen, we used the autoenoder to integrate the genus and gene expression features. It seems that the model flat lines and does not improve across feature selection amounts (Figure 15).

Chart, line chart

Description automatically generated

Figure : RMSE of stage prediction for STAD when using an elastic net prediction model with linear regression feature selection and the features extracted from an autoencoder (red line). As there are only 30 total features extracted, the prediction performance corresponding to the feature selection amount of 30 is the prediction performance when no features selection is performed.

To investigate whether the lack of performance is due to the integration method or the combination of both modalities, we also build a model on each modality integrated with the autoencoder separately.

### Nonnegative matrix factorization

To further rule out whether the lack of improvement when integrating modalities is due to the feature extraction method, we also integrated the different modalities using nonnegative matrix factorization. It is a method which does not assume noncorrelation between components, which might be more in line with biological data than other commonly used methods such as PCA and ICA, and provides easy to interpret results 35. It is also found some success when being used on gene expression data to identify disease clusters 35.

This is a method which extracts meaningful features from a high dimensional space by decomposing the feature matrix into a coefficients matrix and a components matrix , where is the number of bases components. These bases components are then used to obtain a lower dimensional representation of the features data. While, unlike the prediction model using the AE integrated features, the model using the NMF features does not flatline, it still does not provide additional improvement over using the gene expression layer separately (Figure 16).

Chart, line chart

Description automatically generated

Figure : RMSE of stage prediction for STAD when using an elastic net prediction model with linear regression feature selection and the features extracted using nonnegative matrix factorization (red line). As there are only 30 total features extracted, the prediction performance corresponding to the feature selection amount of 30 is the prediction performance when no features selection is performed.

## Discussion

### Genus data selection

Microbial data used only goes up to the genus level and we don’t have of viruses or other types of data.

Changes within individuals and populations are often smaller than the variation between individuals. Repeated daily or weekly measurements might be needed to capture the most important the most specific host- microbiome interactions 5. Thus, the available data might not be enough to properly capture the relevant variation between individuals which lead to differing disease states. Additionally, there might also be differences across the clinical domain. For example, tumor samples for a certain cancer might differ across genders 11.

The data might not be perfect, as care often has to be taken to clear the data of contaminants when microbial data is mined from TCGA 16.

Poore et al. 6 investigated microbial reads from TCGA whole-genome sequencing and RNA-sequencing to identify microbial signatures and discriminate within and across cancers. Using a prediction model, the authors achieved a good performance discriminating between early and late stage cancer for certain cancers, such as COAD and STAD, but not for discriminating intermediates stages. Thus, the microbial structure might not correlate with cancer stages for all types of cancers. The lack of performance might be due to microbial heterogeneity. The authors also achieve good performance in tumor versus normal prediction for COAD, HNSC and STAD.

Hermida et al. 36 build a prediction model combining gene expression and microbial abundance data. It was found that gene expression data was a much more powerful predictor than microbial abundance data and that integrating the two modalities offered little to no improvement when predicting drug response and patient prognosis.

In terms of feature selection, penalized regression methods might not properly capture grouping of information 25.

# Conclusion

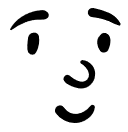
In conclusion, it appears that integrating genus and ge data does not help improve prediction performance. It is clear that the human microbiome has an effect on cancer aetiology, however, the prediction performance of just ge alone might be enough to capture the underlying patterns relevant for cancer diagnostics. This is similar to the redundancy of other omics types in human cancer diagnostics using multi-omic data.,

One problem might also be the data set. Future work could explore different data sets, such as the human microbiome project 2.

# Future works

It could be worth exploring other modalities as well besides only gene expression. Why didn’t you use AUC? Does it make sense to use f1 score for tumor prediction? What about balanced accuracy for hyper parameter tuning?

# Acknowledgments

I would like to thank my supervisor Prof. Thomas and daily supervisor Aakash for always being there to supervise me and give me advice. While it was not always easy, I learned a lot.

# Appendix

## Data Exploration

## There is little qualitative difference in performance with holo-omic approach

We performed a preliminary exploration of the available data sets in order to examine whether a holistic view could have benefits for cancer diagnostics. To this end, we used numerous dimensionality reduction techniques in order to examine the separation between classes for multiple diagnostic endpoints as an indicator of the possible predictive value of integrating omics and microbial data. The diagnostic endpoints were tumor versus normal prediction, and tumor stage prediction. We investigated the class separation using the PCA and t-SNE dimensionality reduction techniques.

For each dimensionality reduction technique and for all 4 cancers, we performed dimensionality reduction on the gene expression data set separately, the genus taxonomical abundance data, and then on the concatenation of these datasets. Additionally, this was done for increasing amounts of selected features, and when using all features. Feature selection was performed using the chi-square test and using all data.

#### Feature selection robustness

Devise experiment to examine feature selection robustness across iterations. The features are relatively varied

[picture here preferably of linear regression selection and stomach cancer in stage]a

### Tumor PCA does not show additional class separation

The PCA for the tumor classification endpoint shows that there is no meaningful additional separation between the classes when integrating genus taxonomic data with gene expression versus using only gene expression data (Figure 1). Additionally, the genus layer does not seem to provide much by itself separating power. This is possibly because the microbial information is not directly related to the phenotype and thus does not provide enough explaining power by itself.

Chart, scatter chart

Description automatically generated

Figure : PCA of STAD (stomach adenocarcinoma) for all modalities when there is no feature selection. The first graph contains the PCA for the genus abundance data (GENUS), the second graph for the gene expression data (GE) and the third graph for the concatenated genus + gene expression features (GE ∩ GENUS). The horizontal axis displays the first principal component, while the vertical axis displays the second principal component of the PCA. Finally, samples in red denote tumor samples while those in blue denote normal samples.

This result is consistent across cancer types (i.e. for COAD, ESCA and HNSC). The results for these additional cancer types can be found in the appendix.

### Results are consistent across feature selection amounts

For PCA, performing feature selection leads to increased separation.

For tumor versus normal prediction, features (0,5,10) exhibit about the same behavior. GE tends to for show much better separation than genus, and GE+Genus is almost exactly the same as just GE, possibly because GE has much more features, and when features are selected, the selected features are probably GE features.

### Using different dimensionality reduction techniques does not offer additional separation

To determine whether the lack of separation of the holo-omic approach was due to the dimensionality reduction method chosen, we repeated the above experiment using t-SNE.

Chart, scatter chart

Description automatically generated

Figure : t-SNE of STAD (stomach adenocarcinoma) for all modalities when there is no feature selection. The first graph contains the PCA for the genus abundance data (GENUS), the second graph for the gene expression data (GE) and the third graph for the concatenated genus + gene expression features (GE ∩ GENUS). The horizontal axis displays the first t-SNE component, while the vertical axis displays the second t-SNE component. Finally, samples in red denote tumor samples while those in blue denote normal samples.

With t-SNE, using the overlapping gene expression with genus data again provides a nearly identical amount of separation as when only using gene expression data (Figure 2). Similar to PCA, using only the genus modality provides the worst separation between classes .

### There is unclear class separation for the stage class

In order to determine whether the above results were due to the classification endpoint chosen, we repeated the results using a different endpoint, the tumor stage. For the stage class endpoint, there is not much separation to be seen across any of the modalities (Figure 3). This is likely because of the class imbalance.

Chart, scatter chart

Description automatically generated

Figure : PCA of STAD (stomach adenocarcinoma) for all modalities when there is no feature selection. The first graph contains the PCA for the genus abundance data (GENUS), the second graph for the gene expression data (GE) and the third graph for the concatenated genus + gene expression features (GE ∩ GENUS). The horizontal axis displays the first principal component, while the vertical axis displays the second principal component of the PCA. Finally, the different colored points represent the cancer stage of the different samples, with stage one being a normal non-tumor sample.

Again, there is not much difference between the GE and the overlapped set. These results are similar when using t-SNE.

### Exploring the modalities balance of integrated features

One problem with the above is that the integration occurs via simple concatenation, and if the feature selection process simply selects features from only one data set, then the integration performance will approximate that of the individual data set.

## Predictive performance

This section contains experiments for all cancers.

### Tumor prediction for all cancers

These are the performances for tumor prediction using SVC and chi-squared for all the cancers.

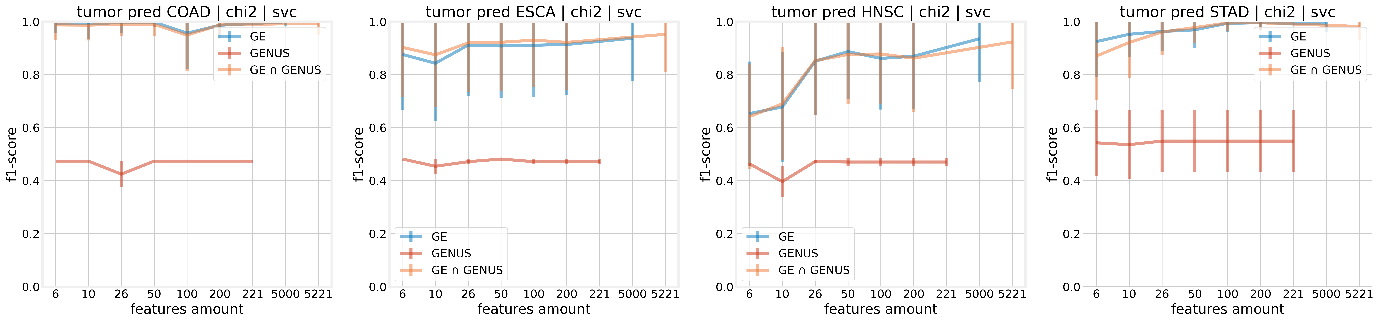


Figure : f1-score for the tumor versus normal prediction endpoint for COAD (colon adenocarcinoma), ESCA (esophageal carcinoma), HNSC (head and neck squamous carcinoma) and STAD (stomach adenocarcinoma), respectively. Each graph contains the f1-score for each modality (i.e. genus abundance data (GENUS), gene expression data (GE) and the concatenated genus + gene expression features (GE ∩ GENUS)). The endpoints of each horizontal line segment indicate the average f1-score across every random sample iteration, while the vertical line segments indicates the standard deviation of the f1-score across these iterations.

### Modality parity enforcement

These are the results when using modality enforcement for all cancers.

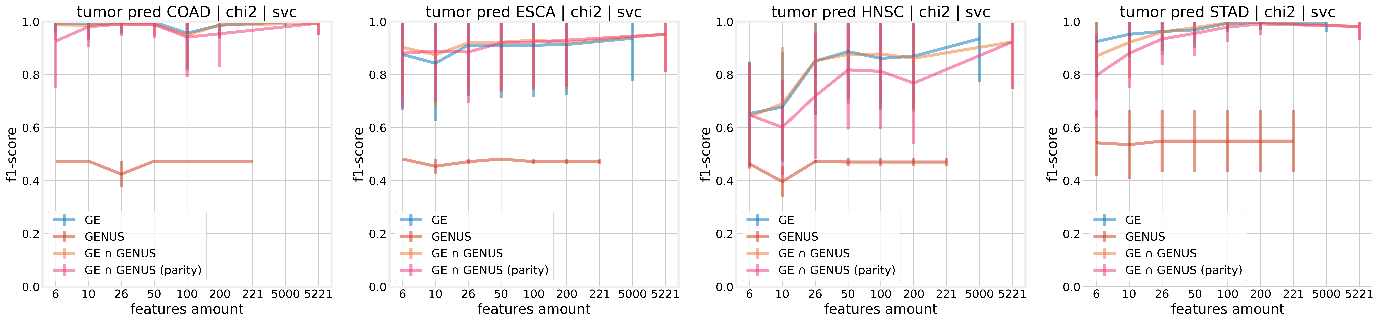


Figure : f1-score for the tumor versus normal prediction endpoint for COAD (colon adenocarcinoma), ESCA (esophageal carcinoma), HNSC (head and neck squamous carcinoma) and STAD (stomach adenocarcinoma), respectively. Each graph contains the f1-score for each modality (i.e. genus abundance data (GENUS), gene expression data (GE) and the concatenated genus + gene expression features (GE ∩ GENUS), and concatenated genus + gene expression features with enforced parity (GE ∩ GENUS (parity))). The endpoints of each horizontal line segment indicate the average f1-score across every random sample iteration, while the vertical line segments indicates the standard deviation of the f1-score across these iterations.

Chart, line chart

Description automatically generated

Figure : the fraction of GE features (vertical axis) selected from the total amount of features for each feature amount (horizontal axis) when doing tumor versus normal prediction with the genus + gene expression modality (GE ∩ GENUS). Each line displays the results for a different cancer, namely (COAD (colon adenocarcinoma), ESCA (esophageal carcinoma), HNSC (head and neck squamous carcinoma) and STAD (stomach adenocarcinoma). These results are for the linreg feature selection method.

Furthermore, out of the 221 total GENUS features, less than 1% is selected in the feature selection process (Figure 8).

Chart, line chart

Description automatically generated

Figure : the absolute amount of GENUS features (vertical axis) selected from the total amount of features for each feature amount (horizontal axis) when doing tumor versus normal prediction with the genus + gene expression modality (GE ∩ GENUS). Each line displays the results for a different cancer, namely (COAD (colon adenocarcinoma), ESCA (esophageal carcinoma), HNSC (head and neck squamous carcinoma) and STAD (stomach adenocarcinoma). These results are for the linreg feature selection method.

### Random Forest Regression

We have also conducted the random Forest Regression Model with hyper parameter tuning for all the different cancers.

### Results independent of feature selection method

To investigate whether the lower performance was due to the feature selection method, we attempted the experiments while using the chi2 test as well.

Chart, box and whisker chart

Description automatically generated

Figure : Same

Chart

Description automatically generated

Figure : Cancer predictions for stage using chi2

# Purgatory

The first step for the analysis of microbial omics data is choosing which type of data to use. Microbial data is often obtained through either amplicon sequencing or shotgun sequencing. Amplicon sequencing amplifies specific regions of the microbial 16S rRNA gene while shotgun sequencing reads all genomic DNA in a sample. There are numerous crucial differences between these types of sequencing methods, which can affect the type of downstream analyses which can be performed.

Shotgun sequencing makes it easier to identify species, and occasionally even strains. It can identify all taxa, including fungi and viruses, instead of just bacteria and archaea. However, it might lead to sequencing of host DNA which can obscure results, contains more complex data and might have less reference genomes available to match to a certain taxa. Either way, it has provided valuable insight into the phylogeny, biodiversity, metabolic abilities and functional diversity of many organisms.

Result sectioning:

**there is little qualitative difference in performance**

tumor PCA does not show additional class separation

results are consistent across feature selection amounts

using different dimensionality reduction techniques does not help

there is unclear class separation for stage endpoint

**there is little quantitative difference in performance**

holo- omics approach does not lead to improvement

genus layer provides worse performance

results are independent of prediction model

feature selection is dominated by one modality

results consistent across prediction targets

enforcing class balance does not improve performance

**lack of­­ performance improvement is due to data**

genus data selection

feature selection is sensible

**unsectioned**

smarter integration method does not improve results (AE and NMF)

no improvement for different feature selection technique

# References

1. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. Review<br>The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol. Onkol.* **2015**, 68–77 (2015).

2. Machiraju, G., Amar, D. & Ashley, E. Multi-omics factorization illustrates the added value of deep learning approaches. 7.

3. Dohlman, A. B. *et al.* The cancer microbiome atlas: a pan-cancer comparative analysis to distinguish tissue-resident microbiota from contaminants. *Cell Host Microbe* **29**, 281-298.e5 (2021).

4. Ehrlich, S. D. MetaHIT: The European Union Project on Metagenomics of the Human Intestinal Tract. in *Metagenomics of the Human Body* (ed. Nelson, K. E.) 307–316 (Springer, 2011). doi:10.1007/978-1-4419-7089-3\_15.

5. Proctor, L. M. *et al.* The Integrative Human Microbiome Project. *Nature* **569**, 641–648 (2019).

6. Poore, G. D. *et al.* Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature* **579**, 567–574 (2020).

7. Knippel, R. J., Drewes, J. L. & Sears, C. L. The Cancer Microbiome: Recent Highlights and Knowledge Gaps. *Cancer Discov.* **11**, 2378–2395 (2021).

8. Garrett, W. S. The gut microbiota and colon cancer. *Science* **364**, 1133–1135 (2019).

9. Nyholm, L. *et al.* Holo-Omics: Integrated Host-Microbiota Multi-omics for Basic and Applied Biological Research. *iScience* **23**, 101414 (2020).

10. Kwon, M., Seo, S.-S., Kim, M. K., Lee, D. O. & Lim, M. C. Compositional and Functional Differences between Microbiota and Cervical Carcinogenesis as Identified by Shotgun Metagenomic Sequencing. *Cancers* **11**, 309 (2019).

11. Gnanasekar, A. *et al.* The intratumor microbiome predicts prognosis across gender and subtypes in papillary thyroid carcinoma. *Comput. Struct. Biotechnol. J.* **19**, 1986–1997 (2021).

12. Elinav, E., Garrett, W. S., Trinchieri, G. & Wargo, J. The cancer microbiome. *Nat. Rev. Cancer* **19**, 371–376 (2019).

13. Fang, H., Du, Y., Pan, S., Zhong, M. & Tang, J. Patients with Parkinson’s disease predict a lower incidence of colorectal cancer. *BMC Geriatr.* **21**, 564 (2021).

14. Greathouse, K. L. *et al.* Interaction between the microbiome and TP53 in human lung cancer. *Genome Biol.* **19**, 123 (2018).

15. Wang, Y., Wang, Y. & Wang, J. A comprehensive analysis of intratumor microbiome in head and neck squamous cell carcinoma. *Eur. Arch. Otorhinolaryngol.* (2022) doi:10.1007/s00405-022-07284-z.

16. Chakladar, J. *et al.* The Pancreatic Microbiome is Associated with Carcinogenesis and Worse Prognosis in Males and Smokers. *Cancers* **12**, 2672 (2020).

17. Erickson, A. R. *et al.* Integrated Metagenomics/Metaproteomics Reveals Human Host-Microbiota Signatures of Crohn’s Disease. *PLoS ONE* **7**, e49138 (2012).

18. Alberdi, A., Andersen, S. B., Limborg, M. T., Dunn, R. R. & Gilbert, M. T. P. Disentangling host–microbiota complexity through hologenomics. *Nat. Rev. Genet.* 1–17 (2021) doi:10.1038/s41576-021-00421-0.

19. Limborg, M. T. *et al.* Applied Hologenomics: Feasibility and Potential in Aquaculture. *Trends Biotechnol.* **36**, 252–264 (2018).

20. Way, G. P. & Greene, C. S. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. *Pac. Symp. Biocomput. Pac. Symp. Biocomput.* **23**, 80–91 (2018).

21. Duan, R. *et al.* Evaluation and comparison of multi-omics data integration methods for cancer subtyping. *PLOS Comput. Biol.* **17**, e1009224 (2021).

22. Ding, M. Q., Chen, L., Cooper, G. F., Young, J. D. & Lu, X. Precision Oncology beyond Targeted Therapy: Combining Omics Data with Machine Learning Matches the Majority of Cancer Cells to Effective Therapeutics. *Mol. Cancer Res.* **16**, 269–278 (2018).

23. Chaudhary, K., Poirion, O. B., Lu, L. & Garmire, L. X. Deep Learning–Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. *Clin. Cancer Res.* **24**, 1248–1259 (2018).

24. Koul, N. & Manvi, S. S. Feature Selection From Gene Expression Data Using Simulated Annealing and Partial Least Squares Regression Coefficients. *Glob. Transit. Proc.* **3**, 251–256 (2022).

25. Sen Puliparambil, B., Tomal, J. H. & Yan, Y. A Novel Algorithm for Feature Selection Using Penalized Regression with Applications to Single-Cell RNA Sequencing Data. *Biology* **11**, 1495 (2022).

26. Rupapara, V. *et al.* Blood cancer prediction using leukemia microarray gene data and hybrid logistic vector trees model. *Sci. Rep.* **12**, 1000 (2022).

27. Abdelnaby, M., Alfonse, M. & Roushdy, M. *A Hybrid Mutual Information-LASSO-Genetic Algorithm Selection Approach for Classifying Breast Cancer*. (2021). doi:10.1007/978-981-16-2275-5\_36.

28. Al-hebshi, N. N. *et al.* Inflammatory bacteriome featuring Fusobacterium nucleatum and Pseudomonas aeruginosa identified in association with oral squamous cell carcinoma. *Sci. Rep.* **7**, 1834 (2017).

29. Peek, R. M. & Blaser, M. J. Helicobacter pylori and gastrointestinal tract adenocarcinomas. *Nat. Rev. Cancer* **2**, 28–37 (2002).

30. Blaser, M. J. *et al.* Infection with Helicobacter pylori Strains Possessing cagA Is Associated with an Increased Risk of Developing Adenocarcinoma of the Stomach1. *Cancer Res.* **55**, 2111–2115 (1995).

31. Parsonnet, J. *et al.* Helicobacter pylori Infection and the Risk of Gastric Carcinoma. *N. Engl. J. Med.* **325**, 1127–1131 (1991).

32. Shariati, A. *et al.* Association between colorectal cancer and Fusobacterium nucleatum and Bacteroides fragilis bacteria in Iranian patients: a preliminary study. *Infect. Agent. Cancer* **16**, 41 (2021).

33. Yamamura, K. *et al.* Human Microbiome *Fusobacterium Nucleatum* in Esophageal Cancer Tissue Is Associated with Prognosis. *Clin. Cancer Res.* **22**, 5574–5581 (2016).

34. Shin, J. M. *et al.* Microbial Communities Associated with Primary and Metastatic Head and Neck Squamous Cell Carcinoma – A High Fusobacterial and Low Streptococcal Signature. *Sci. Rep.* **7**, 9934 (2017).

35. Hamamoto, R. *et al.* Application of non-negative matrix factorization in oncology: one approach for establishing precision medicine. *Brief. Bioinform.* **23**, bbac246 (2022).

36. Hermida, L. C., Gertz, E. M. & Ruppin, E. Predicting cancer prognosis and drug response from the tumor microbiome. *Nat. Commun.* **13**, 2896 (2022).

1. <https://tcma.pratt.duke.edu/> [↑](#footnote-ref-1)
2. https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/sample-type-codes [↑](#footnote-ref-2)
3. https://docs.gdc.cancer.gov/Encyclopedia/pages/images/TCGA-TCGAbarcode-080518-1750-4378.pdf [↑](#footnote-ref-3)