Host- Microbiome Omics Integration for Cancer Analysis And Diagnostics

# Abstract

Cancer is one of the leading causes of death in the world. While there have been many studies investigating the progression and prevalence of cancer in tissues using host omics data or microbial data, there is a lack of research using a holo- omics approach combining both types of data. Such an approach could offer additional insights as those microbiota has been shown to have an effect on cancer morphology and aetiology.

To investigate whether such an approach offers additional information, we built prediction models using individual gene expression layer, genus abundance data as a baseline and then compared it to the prediction performance when using both modalities together integrated in differing ways, using different prediction models and predicting for different prediction targets.

The results indicate that there is no additional improvement. This might be because of the data or the power of gene expression.

# Introduction

Cancer is among the leading causes of death worldwide, being 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 the availability of biological data. Such data availability has made it more accessible for researchers to use various omics data to perform various tasks for cancer diagnostics which relate 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 (i.e. genomics), DNA methylation (i.e. epigenomics) 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. Using host omics data from this database, researchers have been able to derive insights such as finding biomarkers, determining differences in biological composition of tumor samples versus normal samples and examining features related to survival. While using individual omic types has led to useful insights, an important development has been the usage of so-called multi-omics analyses methods, where data from multiple omics layers are integrated. This has been shown to lead to additional insights and model performance over single omics methods for cancer diagnostics 2,3, although it is worth noting that these methods deal with additional challenges owing to the heterogeneity of the data, noise, high dimensionality and sparsity of the multi-omics data 4.

## Microbiome-based analysis

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 microbiome. Namely, an ecosystem of 10 to 100 trillion microorganisms encompassing 500 to 1000 unique species for each individual 5. 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 microbiome which can be exploited.

Such data sets include MetaHIT 6 and iHMP 7, which contain microbial data from healthy and diseased patients. However, a problem with these data sets is that they often contain data from tissue swabs and stool samples, which are not necessarily representative of the microbiome of internal organs 5. 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 tumor tissues and blood. However, the microbial reads in this data set are often a result of contamination 5,8 and thus extensive care needs to be taken to properly decontaminate the data. There have been a number of studies that have managed to do this and obtained useful insights on the relation between the tumor microbiome and certain cancers 9,10. One major problem is that these data sets are often not readily available. A data set which tries to combat this, is the Cancer Microbiome Atlas (TCMA), which contains batch-corrected and decontaminated microbial data mined from TCGA whole-genome sequencing (WGS) and whole-exome sequencing (WXS) TCGA experiments

Research has been done on this data BLA [also motivated use of predictor models]

Regardless of the source, microbial data has been used to study various aspects of the human microbiome and its association with diseases. Taxonomical data can be used to investigate bacterial abundance differences between cancer and healthy samples using hypotheses tests 5,9,11,12 or predictor models (e.g. regression) 9,12, examine co-abundance of microbiota in certain tissues 5, and examine the association with overall survival 5,13 and clinical factors such as gender or age 11–13, which can often be confoundersm, possibly using predictor models 12. It is also possible to combine metaproteomics with metagenomics to also investigate functional differences of microbes between healthy and diseased samples 14. Such investigations have shown that the microbiome exhibits significant variation between individuals 15, cancer types 9,16,17, cancer subtypes 17, healthy and unhealthy individuals 17 and tumor versus normal samples 18, tumor tissues across patient survival rates 19. This relates not only to microbial abundance data, but also functional categories 17. And can be affected by many factors such as diet, environmental exposure and lifestyle choices 20.

Furthermore, it has become clear that the microbiome is not only associated with human phenotypes but that composition and changes in the microbiome has a direct influence on oncogenesis 21–23 and tumor immunotherapy response 22,23, whether positive or negative. As an example of a mechanism through which the microbiome can affect a patient, certain bacteria can bind to and alter the function of immune system cells which infiltrate tumors, thereby affecting carcinogenesis and resistance to chemotherapy 20.

## 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 24. However, it has become clear that the host can alter the human microbiota and vice versa 25. Thus, the integration of host and microbial data could help to better understand the aetiology and physiology of different cancers and provide new insights 21. 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 26.

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 5,13, host mutations 11 or proteins 5. However, these studies often use data sets which are not made available, making cross analysis more difficult.

More analysis into hologenomics 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 20. Finally, such methods could also be used to identify biomarkers and predict cancer versus normal samples. They can also be used for prognostic assessment 19. Due to the open availability of the TCMA dataset, this has opened an avenue for new explorations into the intricacies of a hologenomic approach to cancer diagnostics.

## Towards a holistic view

In this paper we aimed to investigate whether a holistic view provides additional insights for cancer diagnostics versus simply using each modality individually by integrating host genomic and microbiome data for cancer patients. Specifically, we aimed to investigate the question:

**Does integrating host and microbial omics data provide additional power over using the modalities individually?**

In this case, power refers to prediction performance, as prediction models have shown much effectiveness in dealing with the challenges present in such an integrated data set and identifying important features and relationships for both microbial 9 and gene expression 27 features.

To this end, we leveraged the TCGA and TCMA data sets and integrated gene expression and microbial genus abundance data for tumor tissues and tumor adjacent normal tissues 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 modality provided additional prediction performance for tumor versus normal prediction and stage prediction than when simply using the modalities separately. While we only showcase the results for STAD due to the relative simplicity of the cancer and high quantity and class balance in the dataset, much of the experiments have also been performed for the other three cancers and can reasonably be generalized to these cancers.

We found that using the integrated modality did not provide additional performance over using the gene expression (GE) modality separately, and that the prediction models that used the genus modality performed the worst. This is likely due to the high information density of the gene expression layer and low amount of information conveyed by the genus layer, due to the specific data source used for the genus data.

# Materials and methods

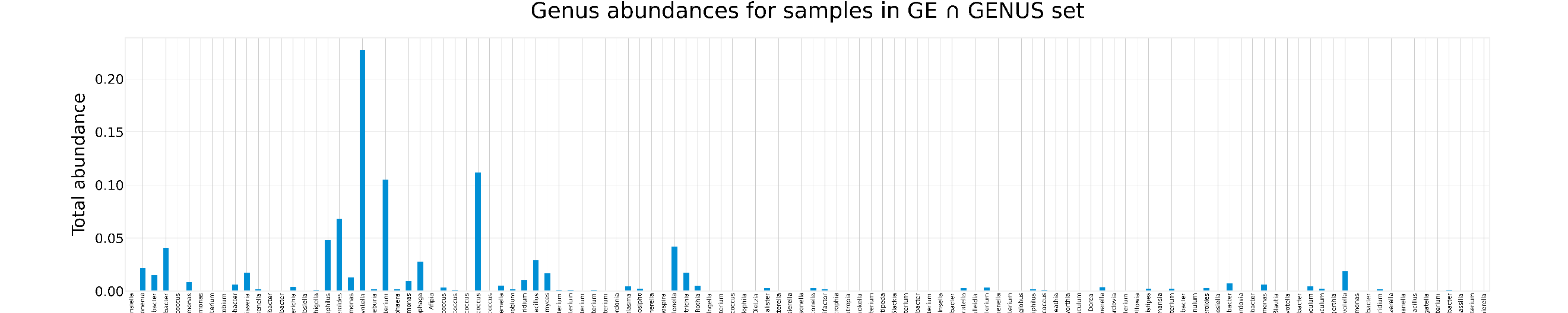
## Data

In order to explore the performance of a holo-omic approach for cancer diagnostics in a holistic context, we used GE features from TCGA and microbial genus relative abundance features from TCMA. We then only kept samples for which there was 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. [insert better picture]



### Host omics data

The host omics data consisted of a TCGA data set which was extracted and processed by a prior study 28. 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 was used to determine the tumor stage of each sample. We grouped every substage together to obtain a final stage. For example, samples that were stage IIA and IIB were grouped together under the bin of stage II. finally, we modelled the normal tumor adjacent samples 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.

|  |  |  |  |
| --- | --- | --- | --- |
| Cancer | Normal samples | Tumor samples | Total |
|  |  |  |  |
| STAD | 9 | 113 | 122 |
| COAD | 3 | 45 | 48 |
| ESCA | 7 | 59 | 66 |
| HNSC | 7 | 154 | 161 |
| READ | 0 | 3 | 3 |
|  |  |  |  |

Table : Number of tumor and normal adjacent tissue (NAT) samples in the overlap set for each cancer

The next table displays the balance for the stage samples. There are less total samples per cancer than the tumor versus normal categorization due to the stage for certain samples being absent. Normal samples are classified as stage 0.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Cancer | Stage 0 | Stage 1 | Stage 2 | Stage 3 | Stage 4 | Total |
|  |  |  |  |  |  |  |
| STAD | 9 | 19 | 36 | 27 | 16 | 107 |
| COAD | 3 | 12 | 15 | 11 | 6 | 47 |
| ESCA | 7 | 8 | 30 | 15 | 2 | 62 |
| HNSC | 7 | 10 | 23 | 23 | 71 | 134 |
| READ | 0 | 0 | 2 | 1 | 0 | 3 |
|  |  |  |  |  |  |  |

Table : Number of samples for each tumor stage in the overlap set for each cancer. Stage 0 corresponds to normal adjacent tissue (NAT) samples.

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

## 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.

## Prediction pipeline

We evaluated the usefulness of each data modality by evaluating the performance of a prediction model when using the modality to train the prediction model for a prediction task. For the stage prediction endpoint, the prediction of the different stages 0-4 was modeled as a regression problem and was performed using the *scikit-learn* random forest regressor model and the elastic net model, while the tumor versus normal prediction endpoint was modeled as a binary classification problem and was performed using the *scikit-learn* support vector machine model. All the prediction models were initialized with a random seed of 0.

For the prediction pipeline, each experiment was performed for each combination of cancers (i.e. COAD, ESCA, HNSC, STAD) and for each modality (i.e. gene expression, genus and the combination of gene expression and genus). To train and evaluate the model, we used a random-sampling based approach.

### Baseline model

To provide a reference to help interpret the performance of each prediction model, we also determined what the predictive performance would be for a model which always predicts the majority target value for each prediction task. For example, for STAD and the tumor versus normal classification endpoint, the baseline model is one which always predicts that a sample is a tumor sample and would have an F1 score of 0.96. For stage, the baseline model would be one which always predicts that a sample has a stage of II (i.e. a target value of 2), and would have an RMSE of 1.168.

### Data splitting

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. This was done by performing a random stratified split of the relevant dataset into 80% training and 20% testing using the *scikit-learn* train\_test\_split function. This split was performed 200 times to obtain a reliable estimate of the model performance and each split iteration was assigned a custom seed based on the iteration count to ensure consistency between experiment runs.

### Feature selection

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.

Diagram

Description automatically generated

Figure : The general prediction pipeline

Diagram

Description automatically generated

Figure : Data spitting and prediction pipeline

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. To do this, we first performed feature selection to create an ordered list of the top 200 features, ranked according to the weights assigned by the relevant feature selection method. From this pool, we then ran the experiments using the top 100, top 50, top 25, top 10 and top 6 features. Thus, the sets of selected features for each feature selection amount 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.

To perform the modality parity enforcement experiments, we performed feature selection using the GE and genus modality separately for each feature amount. For example, when selecting 200 features, we first selected 100 features using only the GE modality separately, then selected 100 from the genus modality, and then combined these together when training and evaluating the model. 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

For the stage prediction endpoint, the feature selection was performed using the Pearson correlation coefficient, an elastic net model, and a random forest regressor model. The Pearson correlation coefficient-based feature selection was performed by using the *scikit-learn* SelectKBest function with the r\_regression method as the scoring function, while the latter two methods were performed by using the ElasticNet and RandomForestRegressor packages, respectively. Both were initialized with a random seed of zero. For the latter two model-based feature selection methods, we also first performed hyper parameter tuning to find the best parameters for the model, and then trained the model on the 80% training set. For the elastic net-based feature selection, we obtained a feature ranking using the magnitude of the coefficients for each feature value after training the model while for the random forest regressor-based model, we used a feature ranking based on the mean decrease in variance among the target values after using a certain feature as a tree splitting node. Finally, for the tumor vs normal prediction endpoint, feature selection was performed using the ANOVA F-test, implemented with the *scikit-learn* SelectKBest function with thef\_classif method as the scoring function.

### Model training

After performing feature selection, we trained a prediction model using the same 80% of the data with each feature selection amount, and also with all the features for the relevant modality (i.e. no feature selection). After finding the best-performing hyperparameters, we trained the relevant model on the 80% portion of the data using these hyperparameters.

### Hyperparameter tuning

Whenever we used a prediction model for either feature selection or to evaluate the predictive performance of a modality, we first performed hyperparameter tuning to find the optimal parameters. This was done using randomized search with the *scikit-learn* RandomizedSearchCV package. We used a stratified 5-fold cross validation split to evaluate the performance of each hyperparameter set and evaluated 100 different randomly sampled hyperparameter sets.

In terms of the searched hyperparameter space, for the elastic net we explored alpha : {1e-5, 1e-4 … 1e, 1e2} ∪ {0}, l1\_ratio : {0, 0.1, 0.2 … 0.9}. For the random forest regressor, we explored n\_estimators : {5, 20, 50, 100, 200, 400}, max\_features : {‘auto’, ‘sqrt’}, max\_depth : {10, 30, 60, 100}, min\_samples\_split : {2, 5, 10}, min\_samples\_leaf : {1, 2, 4}, bootstrap : {True, False}. For the support vector machine, we explored C : {1e-10, 1e-9 … 1e9, 1e10}, class\_weight : { ‘balanced’, None}, kernel : {‘linear’, ‘rbf’} and gamma : {1e-10, 1e-9 … 1e9, 1e10} ∪ {‘scale’, ‘auto’}.

The performance of each sampled hyperparameter set was evaluated using the root mean squared error as a scoring function for the stage prediction endpoint and balanced accuracy for the tumor versus normal endpoint. Finally, to ensure reproducibility of the results, we also used a static random seed which only differed across random sampling iterations.

### Testing and evaluation

After training a model, we then used the 20% testing set of the current random sampling iteration to evaluate the model. For the binary tumor versus normal prediction endpoint, we used the f1-score, while for the continuous stage prediction endpoint, we used the RMSE. For the latter, since the range of stage targets only spans the interval [0,4], we also clamp the prediction values of the prediction model to always be within this range.

As there are 200 random sampling iterations, the prediction pipeline thus generates 200 different sets of values for these evaluation metrics. For the evaluation, we consider the average of each metric across the 200 random sampling iterations, and the standard deviation.

[Statistical significance test, maybe 2 sided Wilcoxon signed rank https://www.nature.com/articles/s41467-022-30512-3#Abs1]

# 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 3. [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 5. It contains genus-level relative abundance data available from multiple cancers allowing for cross cancer analyses using the same source with common methodologies. The authors have made this data openly available for the cancers STAD, COAD, ESCA, HNSC and READ, as these were the cancers with the most microbial reads. We omitted the READ data from our analysis due to a lack of matching data in the TCGA set, and focus our analysis on the STAD data, 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.

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 gene expression omics data and microbial abundance data of various cancers in a predictive model for the cancer diagnostics prediction tasks of tumor versus normal prediction and tumor stage classification. The tumor versus normal prediction endpoint was modeled as a binary classification, while the tumor stage prediction was modeled as a regression task with target values ranging from 0 (i.e. normal samples) to 4 (i.e. Stage IV tumor samples). To establish a baseline, we built prediction models on the gene expression (GE) data set and the microbial abundance data (GENUS) separately and evaluated the prediction performance. We then built prediction models on the concatenation of both of these data sets (GE ∩ GENUS) and compared it to the established baseline. Finally, we also compared the performance to that of a model which always predicts the majority class.

We used random sampling to perform 200 random stratified splits of each modality data set into 80% training and 20% testing. In each random sampling iteration, we also used a feature selection method on the training set beforehand to select the most important features. These features were then used to train a model on this same training set and the performance of the model was then tested on the testing set for the available prediction endpoints. This was repeated for each cancer, modality and feature selection amount, including when performing no feature selection. After this procedure, we obtained 200 scores for each evaluation metric and plotted the mean and standard deviation of these scores. For tumor vs normal prediction, we used the f1-score, as it can handle imbalanced data sets, while for tumor stage prediction, we used the root-mean-squared error (RMSE). Additional details are described in the materials and methods section (Figure 1).

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, the results of which can be found in the appendix.

### Holo-omic approach does not lead to improved prediction performance for either prediction target

To investigate the utility of a holo-omic approach, we built a prediction model using data of each modality for the prediction targets tumor versus normal prediction and stage prediction. For tumor versus normal prediction, we used a support vector machine (SVM) due to its ability to capture nonlinear relationships, deal with imbalanced data sets and previously demonstrated performance 2,29,30 and the ANOVA f-test for feature selection, which has seen some success in selecting genetic features 2. For stage prediction, we used an elastic net model as the predictor model due to its interpretability and previously demonstrated success 29,31 and also as the feature selector due to the demonstrated ability of penalized linear regression-based methods to select genetic features 19,32.

The results indicate that integrating gene expression with microbial abundance data does not lead to a significant improvement in prediction performance (P < 0.?) over using gene expression data alone for tumor versus normal prediction (Figure 2A). These results are consistent across all cancers investigated (Figure S1). Part of the reason is that using gene expression alone already performs quite well, which does not leave much opportunity for a performance improvement with the integrated modality. It can be seen from the standard deviation of the GE performance that the model simply defaults to predicting every sample as a tumor in the worst case ( i.e. the same behavior as the baseline). Still, this high-performance of gene expression is likely because gene expression is among the most causative host-omic layers for tumor development and thus contains the most informative features. Additionally, due to the possible interaction between gene with expression and genus abundance, namely that it affects the tumor environment which in turn cultivates different microbiota, it might already contain much of the discriminatory information that the genus good contain.

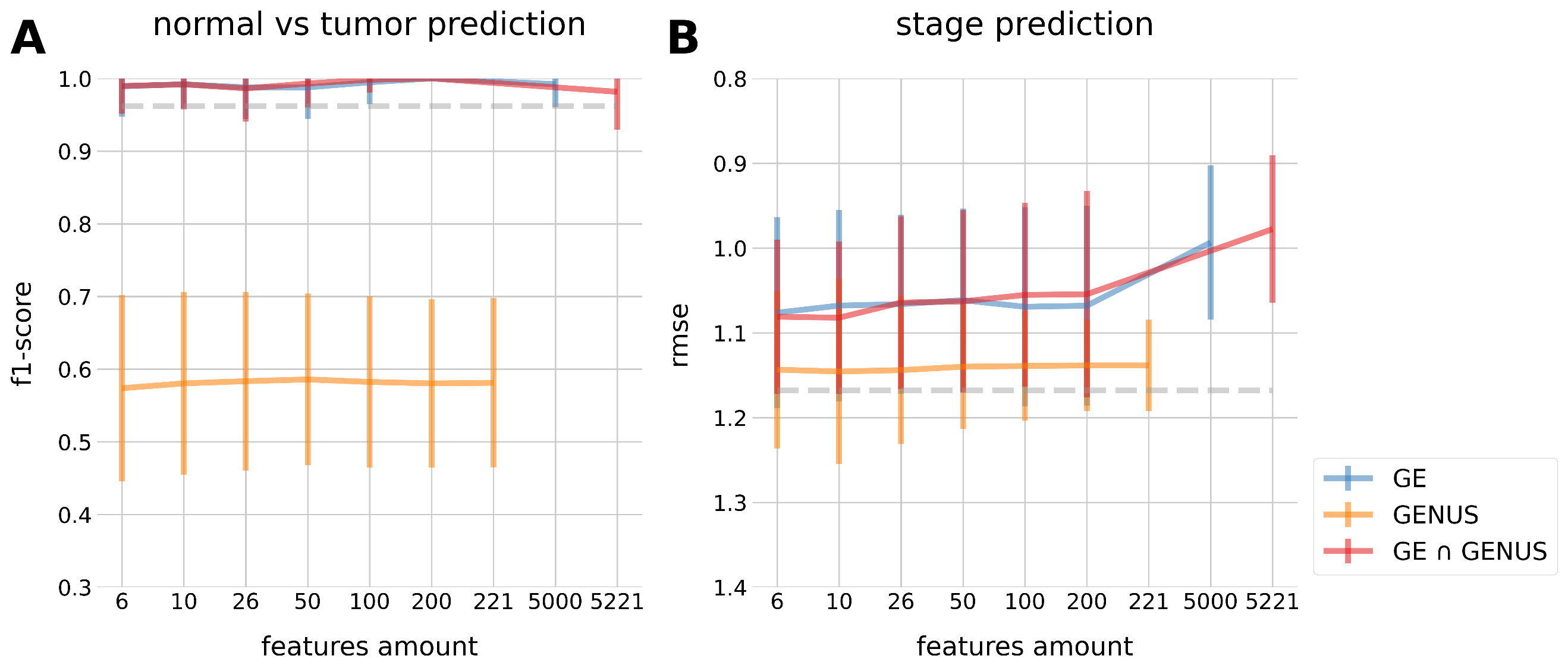


Figure : Predictive performance of prediction endpoints for STAD (stomach adenocarcinoma). **A**, f1-score for the tumor versus normal prediction endpoint. Each line contains the f1-score for a support vector machine trained and tested on a different modality, namely on the genus abundance data (GENUS), the gene expression data (GE) and the concatenated genus + gene expression data (GE ∩ GENUS). The endpoints of each horizontal line segment indicate the average f1-score across every random sampling iteration, while the vertical line segments indicates the standard deviation of the f1-score across these iterations. The last point of each line indicates the prediction performance when all the features of the relevant modality are included (i.e. when there is no feature selection). **B,** Root-mean-squared error (RMSE) for the stage prediction endpoint. **C,** RMSE for the stage prediction endpoint for STAD using a random forest model.

The genus layer seems to provide the worst performance and is even outperformed by the baseline model. This is partly due to the baseline model (f1-score = 0.96) already being able to perform quite well due to the heavy class imbalance of 113 tumor samples versus 9 tumor adjacent normal samples. The taxonomic genus data alone does not seem to allow the model to discriminate well between the sample types, indicating that it does not contain enough discriminatory information. This can also be seen by the higher standard deviation in performance for the Genus modality. One possible reason for this is that the high microbiota variation between individuals and similarity between sample types due to microbiota transferring between the tumor and adjacent normal samples could reduce the discriminatory ability of the abundance features.

We investigated whether a prediction endpoint with a better class balance might alter the results by performing the experiments for the stage prediction endpoint. Again, we observed no statistically significant difference (p < 0.??) between the prediction model when using the gene expression data (GE) compared to the overlapped data (GE ∩ GENUS) and the genus layer performs the worst (p = ??) (Figure 2B). Furthermore, these results were consistent across all cancers (Figure S3).

Interestingly, the difference between modalities is smaller for stage prediction and the models all perform below the baseline. This is partly expected as the prediction task is harder, with there being more classes. At the same time, the baseline is also easier to beat, as the model which always predicts the majority class will deviate from the actual class more often than if there are only two classes. The genus modality seems to predict mostly stage II, and thus perform similarly to the baseline. The imbalance of the data set combined with the lack of samples might also make it harder for the model to learn to discriminate between stages. This could be the reason that even GE and GE ∩ GENUS perform relatively poor. Additionally, the high interperson variation of microbiota and movement between tumor and NAT might also play a role in the performance. These results are consistent across all cancers and prediction targets. Due to the availability of data and higher class balance, we will continue showing results only for STAD stage prediction. And elastic net for feature selection and model training.

### Using nonlinear prediction model does not improve results

To investigate whether the results were due to the linear elastic net 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 a random forest regressor, which is able to capture nonlinear relationships between features and has seen some success in prediction models with gene-based features.

Again, the holo-omic approach does not offer additional improvement over the individual gene expression layer, indicating that the lack of performance improvement with the holo-omics approach is not due to the model not being able to properly capture complex nonlinear relationships (Figure 2C).

## Lack of performance improvement with integrated set is not due to feature selection

The feature selection method and process has a significant impact on the performance and results of prediction models She 32. 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 ran the prediction pipeline while using the Pearson correlation coefficient and random forest as feature selection methods. The Pearson correlation coefficient has previously successfully been used to find important genetic features and is model agnostic, while random forest based feature selection has also been successful and can capture relationships between features. As can be seen, for neither the Pearson correlation coefficient (Figure 5A) or the random forest feature selection (Figure 5B) is there any improvement when integrating the modalities. It is worth noting that the performance of the Genus model flat lines beyond the feature amount of 50. This is because there are only 52 nonzero features in the Genus data set, which could be informative.

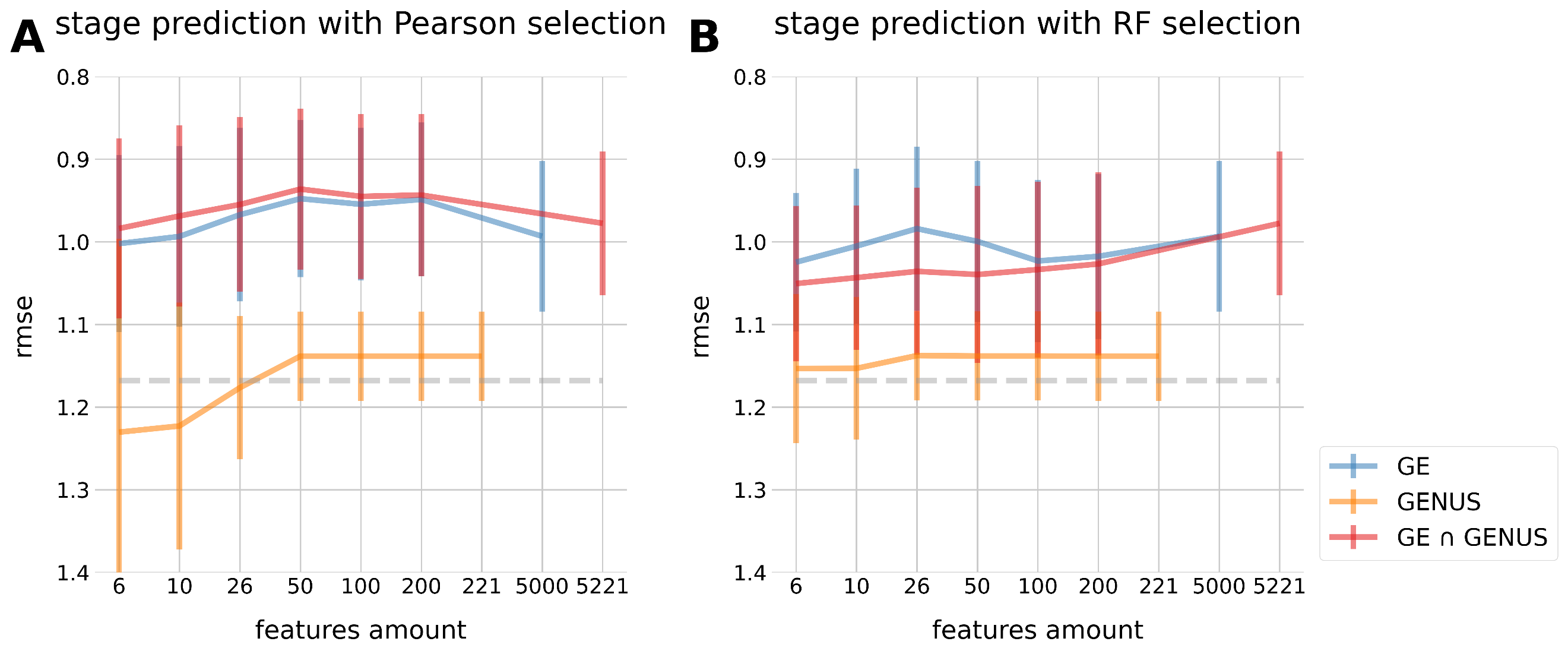


Figure : Cancer predictions for stage using pearson as feature selector

### Feature selection is dominated by gene expression features

To investigate whether the prediction models trained on the integrated set were making use of both with GE and genus features, we investigated what fraction of the features selected for different feature selection amounts consisted of GE features. Thus, we plotted the proportion of selected GE features for stage prediction for the three different feature selection methods seen so far (elastic net, random forest and Pearson correlation) and displayed the fraction for different feature selection amounts (Figure 6A). It seems that when performing feature selection on the integrated dataset, almost all of the features selected originate from the GE set. Furthermore, when investigating the absolute amount of genus features selected, this corresponds to approximately 1 genus feature being selected across the random sampling iterations and selected feature amounts (Figure 6B).

This result is partly expected as there are roughly 25 times more GE features than genus features, (5000 vs 221). If one assumes that both modalities are as predictive of the target, we would naturally expect more 25x more GE features to be selected. Thus, for the top 6 and top 10 features, the amount of genus features selected is within expectations. However, for the higher amount of selected features, the genus features represent a disproportionately low fraction of the total selected feature set.

Additionally, the difference in features selected is also due to the preprocessing of the GE data, and the lack of informative features in the genus data. As previously described, the GE features used were a collection of features selected from a larger pool of GE features from TCGA that had the largest variability, as measured by the mean absolute deviation. Thus, it is more likely that features from this data set would also contain more features which vary with the target class. Additionally, the genus data only contains 52 non-zero features for the shown STAD cancer, leading to the other 169 genus features not having any variability. This, combined with the previously seen lack of informative information in the genus features needs to almost no genus features being selected when compared to the GE features.

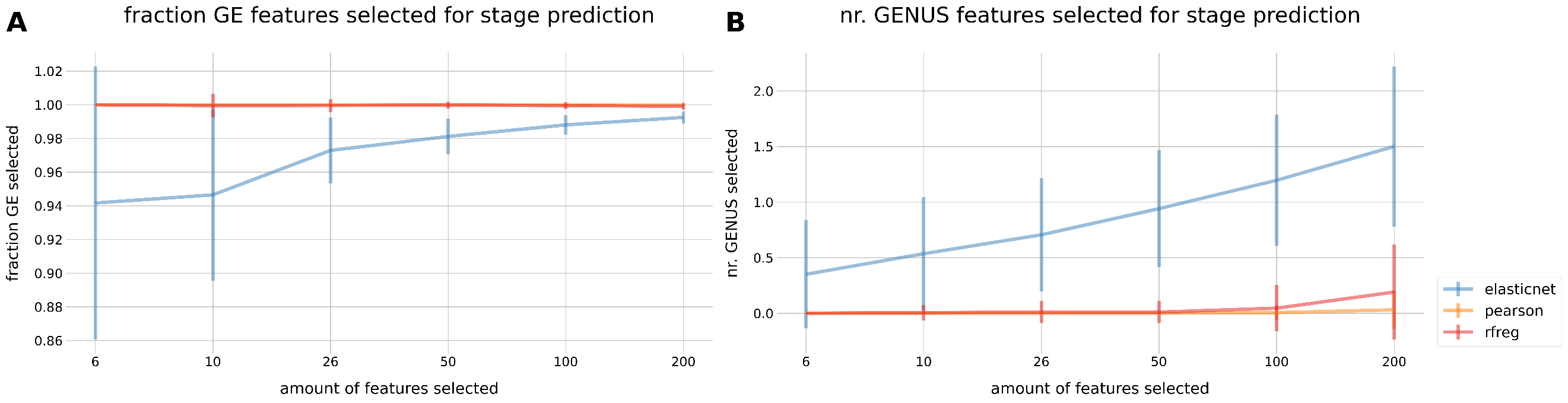


Figure : Features selected from each modality. **A**, The fraction of GE features (vertical axis) selected from the total amount of features for each feature amount (horizontal axis) when doing stage prediction STAD with the genus + gene expression modality (GE ∩ GENUS). These features were selected out of 5221 total GE ∩ GENUS features using an elastic net model. The endpoints of each horizontal line segment indicate the average fraction of GE features selected across the 200 random sampling iterations, while the error bars indicate the standard deviation across these iterations. **B**, 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 for STAD with the genus + gene expression modality (GE ∩ GENUS). These features were selected out of 5221 total GE ∩ GENUS features using an elastic net model.

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

As the domination of gene expression features during the selection process could prevent the prediction model from properly capturing the information of both the gene expression and the genus data, we attempted to avoid this by repeating the prediction experiments while enforcing parity in the amount of features selected from each modality. To do this, we performed feature selection prior to integrating the modalities and ensured that for each feature selection amount, half of the features were from the GE modality while the other half was from the Genus set.

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 8). Comparing the lowest scores for each modality indicates that there is no statistically significant difference between the model trained with modality enforcement and the one trained without modality enforcement (P = 0.5). This is likely because the genus modality is not offering additional information over the GE features already in the feature selection set. This also follows from the fact that few genus features are selected without parity enforcement, indicating that these features offer less discriminatory information. Thus, even when enforcing parity and ensuring there are genus features during the model training, it achieves similar performance to the GE model as the prediction model still assigns the most importance to the GE features and uses them to discriminate between the target values.

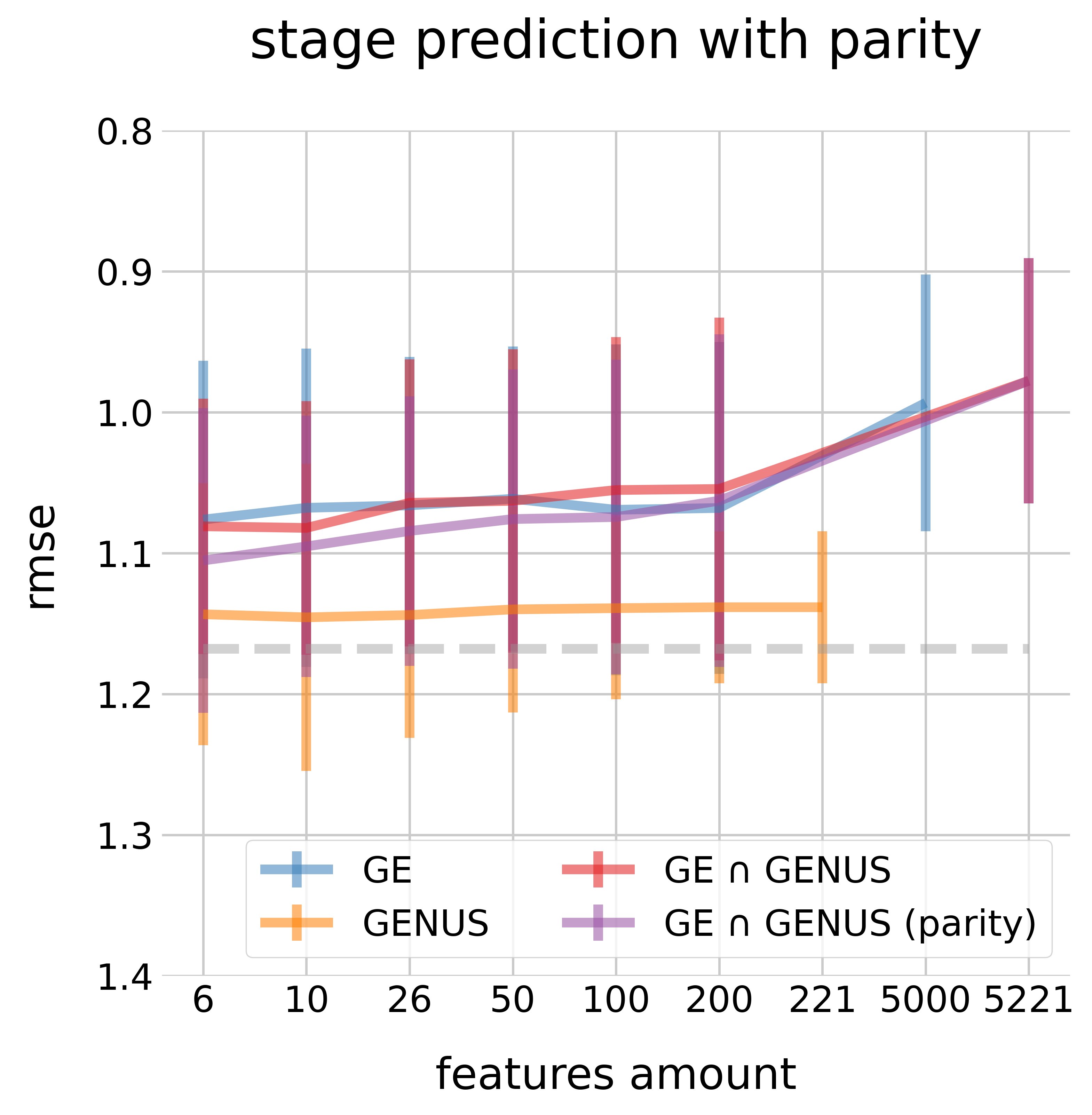


Figure : root-mean-squared error (RMSE) for the stage prediction endpoint for STAD (stomach adenocarcinoma). Each line contains the RMSE for an elastic net model trained and tested on a different modality, namely on the genus abundance data (GENUS), the gene expression data (GE), the concatenated genus + gene expression data (GE ∩ GENUS) and the concatenated genus + gene expression features with enforced parity (GE ∩ GENUS (parity)). The endpoints of each horizontal line segment indicate the average RMSE across every random sampling iteration, while the vertical line segments indicates the standard deviation of the RMSE across these iterations.

### Selected 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.

One genus species which was frequently selected in the overlap set for stomach adenocarcinoma, is *Helicobacter* (Table 3). This seems to be validated by previous studies linking this cancer to *Helicobacter pylori* 33–35. This bacteria can induce gastritis, which can then lead to stomach adenocarcinoma. The genus *Lactobacillus* is also linked to stomach cancer and possibly interacts with *H. pylori* 36.

The second most frequently selected genus feature was the HOXC10 gene, which has been found to be differentially expressed in stomach cancer tissues versus normal tissues and significantly promote tumor development 37. The third most selected feature was the PRSS21 gene, which was previously found to be among the most important biomarkers in a gene signature set for detecting metastasis in stomach cancer patients 27. This indicates that the feature selection step is selecting biologically relevant features.

|  |  |  |  |
| --- | --- | --- | --- |
| Rank | Feature name | Frequency selected | Feature type |
| 1 | TDRD9 | 84% | GE |
| 2 | HOXC10 | 83% | GE |
| 3 | PRSS21 | 78.5% | GE |
| 4 | HOXA13 | 67% | GE |
| 5 | HOXC9 | 62% | GE |
| 6 | *Helicobacter* | 53% | Genus |
| 136 | *Lactobacillus* | 0.5% | Genus |

Table : 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.

## Using complex integration method does not improve performance of integrated set

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.

### No improvement with holo-omic approach with autoencoder integrated features

As autoencoders have successfully been used to integrate multi-omics host features, partially due to its ability to capture nonlinear relationships between features, we used an autoencoder to integrate the GE and GENUS features and then trained a prediction model on the integrated data. We based our model on a deep autoencoder architecture successfully used by Chaudhary et al. to integrate host multi-omics data for liver cancer. Using an autoencoder for liver cancer survival rate prediction and subtype classification 2 as well as other prediction tasks 29.

Again, there is no additional improvement with the holo-omics approach when using AE integrated features over only using GE features (P = 0.5) (Figure 10). As can be seen, the model does converge to the same RMSE error with fewer amounts of features, which indicates that it is able to capture a latent representation of the integrated features. However, it does not appear to offer additional information over simply using the gene expression data alone.

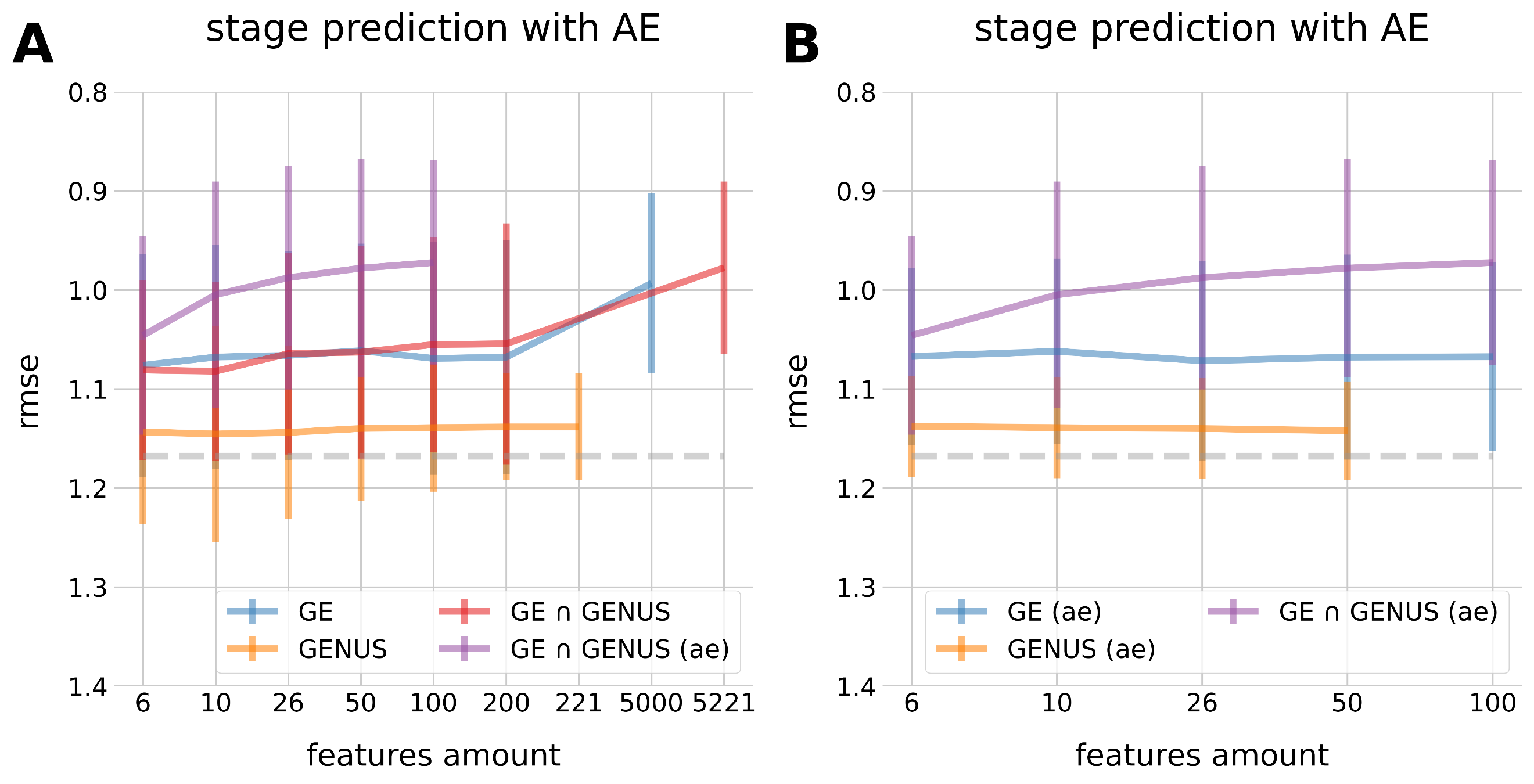


Figure : root-mean-squared error (RMSE) for the stage prediction endpoint for STAD (stomach adenocarcinoma). Each line contains the RMSE for an elastic net model trained and tested on a different modality, namely on the genus abundance data (GENUS), the gene expression data (GE), the concatenated genus + gene expression data (GE ∩ GENUS) and the genus + gene expression features integrated with an autoencoder (GE ∩ GENUS (ae)). The endpoints of each horizontal line segment indicate the average RMSE across every random sampling iteration, while the vertical line segments indicates the standard deviation of the RMSE across these iterations.

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.

### No improvement with holo-omic approach with nonnegative matrix factorization integrated features

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 38. It is also found some success when being used on gene expression data to identify disease clusters 38.

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 11).

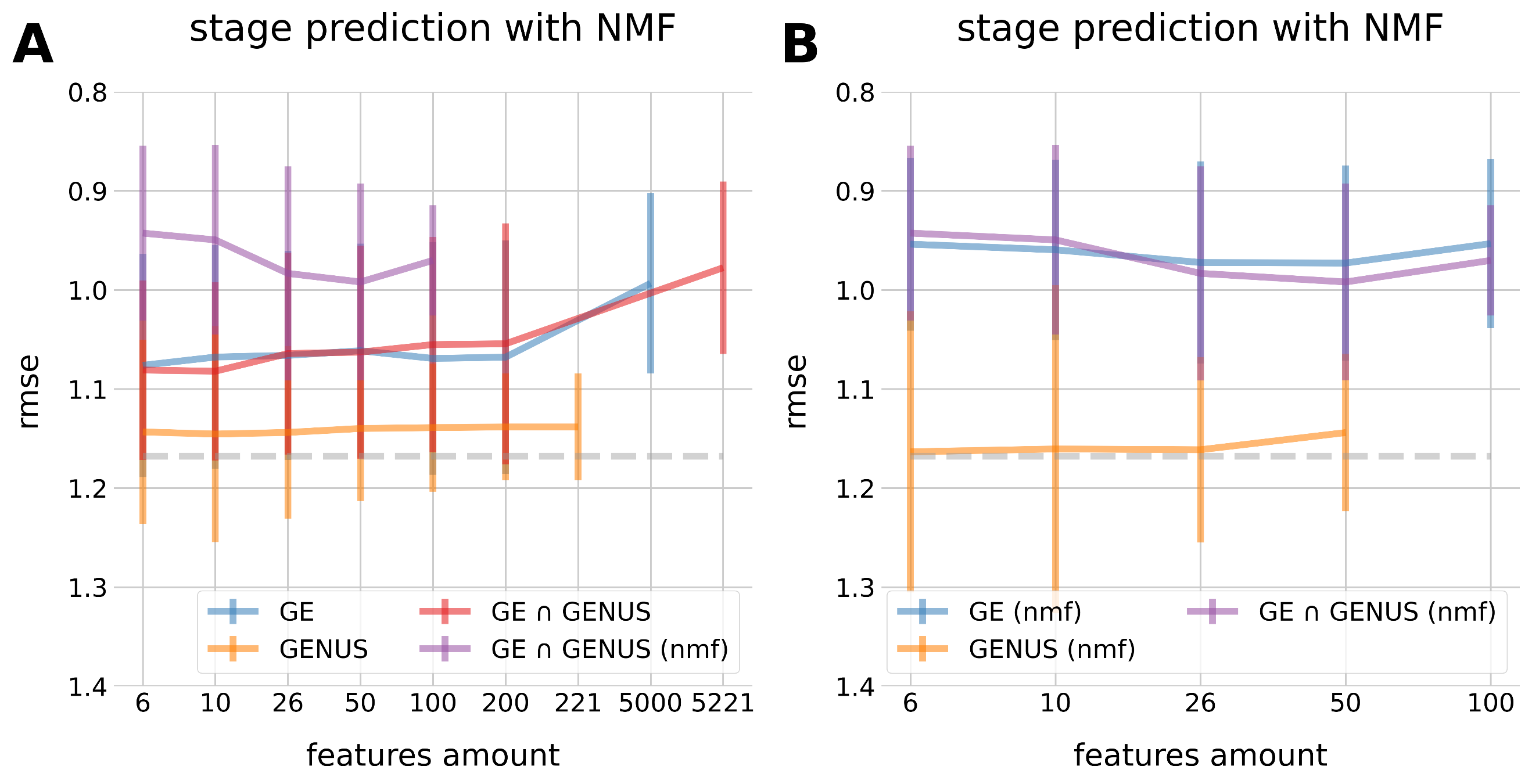
She

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

It seems that across prediction targets, prediction models, feature selection methods, integration methods and cancers, integrating the GE modality with the genus modality does not offer additional predictive power over simply using the GE modality, and that the genus modality provides the least predictive power when used alone. This does not necessarily mean that the holo-omics approach never leads to an improvement in performance, but that it is not the case with the specific genus and host-omic data set used.

The first possible reason for the results is that the GE data is information-dense and recapitulates much of the information of other host omics layers and even the genus abundance data. Namely, gene expression data is known to recapitulate information of more ‘upstream’ datatypes, such as gene mutation and methylation data 3,31. In models that integrate GE with other host omic features, GE features can end up dominating in prediction models 31. Thus, using a less informative datatype might have shown additional improvement when using the genus data. Additionally, gene expression can be correlated with and be affected by certain microbial abundance groups 5. [insert evidence of this for STAD] Thus, certain information given by the microbial abundance data might already be contained within the gene expression data. Furthermore, while there is evidence of a causal link between the microbiome and certain cancers (e.g. *Helicobacter pylori* causing gastric cancer and hepatitis B or C causing liver cancer 39), this is not necessarily the case for all cancers. Especially for intratumoral bacteria, the presence of microbial communities might simply be due to an infection of existing tumors 39. The interactions between cancer and microbiota are bidirectional, because the cancer can also lead to an environment which fosters certain microbiota which in turn affects the cancer 39. In general, the human shapes the microbiota and vice versa 40 . For intratumoral microbes (tumor microbiome), this makes it difficult to determine whether the microbial composition is defined by the tumor or a transient stochastic composition caused by traveling microbes 41. It is unknown whether differential abundance of taxa is a cause or consequence of developing SCC 11.

For the genus data, the data collection process might have removed valuable information. This data was previously mined from existing TCGA whole genome sequencing data. During this process of mining microbial data from TCGA, care has to be taken to clear the data of contaminants 5,13, especially for low biomass samples such as the human tumor microbiome, as contamination can arise during sample collection, DNA extraction and laboratory environment 5,16. During this decontamination step, the original authors used a statistical technique that analyzes microbial abundance data within and across tissues and eliminates those which it finds likely to be contamination. While the authors validated their approach by comparing the mined microbial abundance distribution with that of the original matched TCGA samples, this was only done with 8 samples, and only with CRC data. Even for these samples, differences remained between the mined tumor microbiome data and the microbiome of the original samples. In the end, the decontamination process could have removed valid and informative microbial information, which could happen when mining TCGA data 8.

Another complicating matter is that the lack of samples combined with the high amount of variation in the microbiome can make the prediction tasks harder. The microbiome exhibits significant person-to-person variation 15,42, and is variable across multiple axes such as age, geography, diet 42 and time 40 for the gut 43 and gender 18. Furthermore, there is less variation between individuals on the genus level than the species and strain level 43, this variation can make it harder for a model to properly capture the relevant variation between individuals which lead to differing disease states. As an example, one genus could contain species that are correlated with tumor tissues, but also those correlated with normal tissues 1. To combat this, a higher quantity, but also more specific data might be needed, such as repeated measurements at different times, which might be needed to capture the most important host- microbiome interactions 7. Additionally, rather than only abundance data, it might be necessary to also look at the active expression of microbial genes answer can be less person-to-person variation on the functional level 42 as microbial data mined from TCGA cannot determine whether microbial reads were intra- or extracellular or from alive or dead bacteria 10.

There can be distinct intratumoral bacteria across different subtypes of the same cancer and a tumor sample with its NAT microbiome 16, however this is not always the case 13. Additionally, bacteria from the NAT might be transferred to tumor tissues, leading to a similarity between the microbiomes which might not be conducive for discrimination between the two tissues 16. When investigating tumor, tumor adjacent normal and normal samples, [study] found high similarities in [in-] diversity between the tumor and NAT and dissimilarity of the two tissue types with normal samples 11. Thus, it might be needed to also use normal samples. For CRC as well, there is microbial dysbyiosis between tumor and NAT for the same patient and also across stages 39. Furthermore, microbiota in tissues such as in the gut can interact with other tissues and organs through varying pathways 19,43.

There have been other studies which also use TCGA whole-genome sequencing for cancer diagnostics. Poore et al. 9 created a microbial abundance data set from TCGA data. The results for tumor versus normal prediction are similar to below. They also did not achieve good performance discriminating between stages with microbial abundance data. Using this data set, Hermida et al. 10 build a prediction model combining gene expression and microbial abundance data. They also found that tumor microbial abundance data with clinical variables was only marginally predictive of patient survival rates, and that gene expression data was a much more powerful predictor than microbial abundance data. Additionally, integrating the two modalities did not offer any statistically significant improvement in survival rate prediction. [Contains other studies which use the TCGA microbes in discussion]

[include information about lack of overlapped samples, results could be different for certain endpoints if you use more TCGA data especially normal samples but probably not]

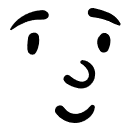
To obtain more meaningful results and truly capture the predictive power of microbial abundance data and its relation to host omics data, it might be necessary to have more specific microbial information, such as species-level information, metagenomic/transcriptonomic (functional) data, which could give a more direct measure of microbial activity and function 41, data on the virome and mycobiome, which are also associated with cancer 41 and the location and organization in the tissue of the microbial data which could also impact tumorigenesis 39. We would also require a decontamination pipeline which removes less valid microbial information, although it is unclear whether it is possible from mining TCGA data alone and this data set would have to be validated with matched tissue samples. Additionally, we would need normal healthy samples rather than tumor adjacent samples.

# Conclusion

It appears that integrating TCMA genus abundance and TCGA gene expression data, whether through simple concatenation or complex integration, does not help improve prediction performance over only using the gene expression modality, and that the genus modality performs the worst. This result was consistent for the cancers STAD, COAD, HNSC and ESCA, which were the cancers with the most microbial reads in the TCMA data set. It is clear that the human microbiome has an effect on cancer aetiology, however, for certain data sets, 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 host multi-omic data.

These results are most likely due to the gene expression set already containing much information, possibly even recapitulating certain information contained within the genus set. Furthermore, the size and quality of the genus set also hinders and complicates the drawing of conclusions from the result. Thus, while the conclusions drawn from this study can help gain insight into the usefulness of a holo-omics approach with TCMA data, it cannot be generalized to microbial abundance data from other data sets. To properly explore the usefulness of a holo-omic approach, it might be necessary to gather more specific data, such as the expressed genes, species and strand, other types of microbes and a more expressive decontamination method which is validated.

# Acknowledgments

I would like to thank my supervisor Prof. Thomas and daily supervisor Akash for always being there to supervise me and give me advice. While it was not always easy, I learned a lot. I required extra resources, but they did not hesitate to help me along and assist me. She



# 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.

### 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

,

#### 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 12). 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 13). 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 14). 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. [give stat significance table]

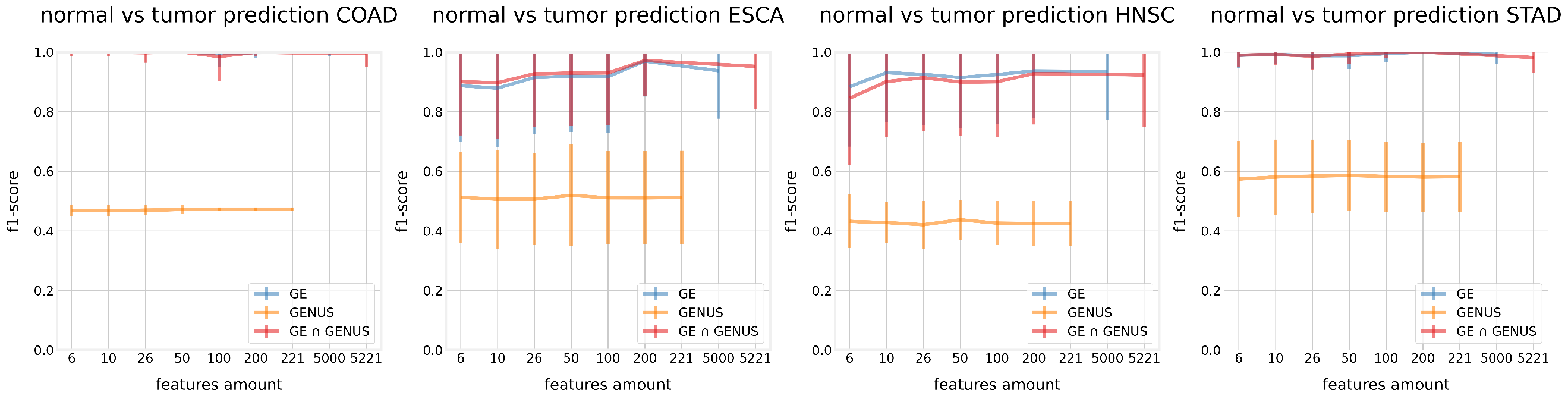


Figure : f1-score for the tumor versus normal prediction endpoint for COAD (colon adenocarcinoma), ESCA (esophageal carcinoma), HNSC (head and neck squamous carcinoma) and STAD (stomach adenocarcinoma), respectively. Each graph contains the f1-score for each modality (i.e. genus abundance data (GENUS), 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.

### Stage prediction for all cancers

These are the performances for stage prediction for all cancers

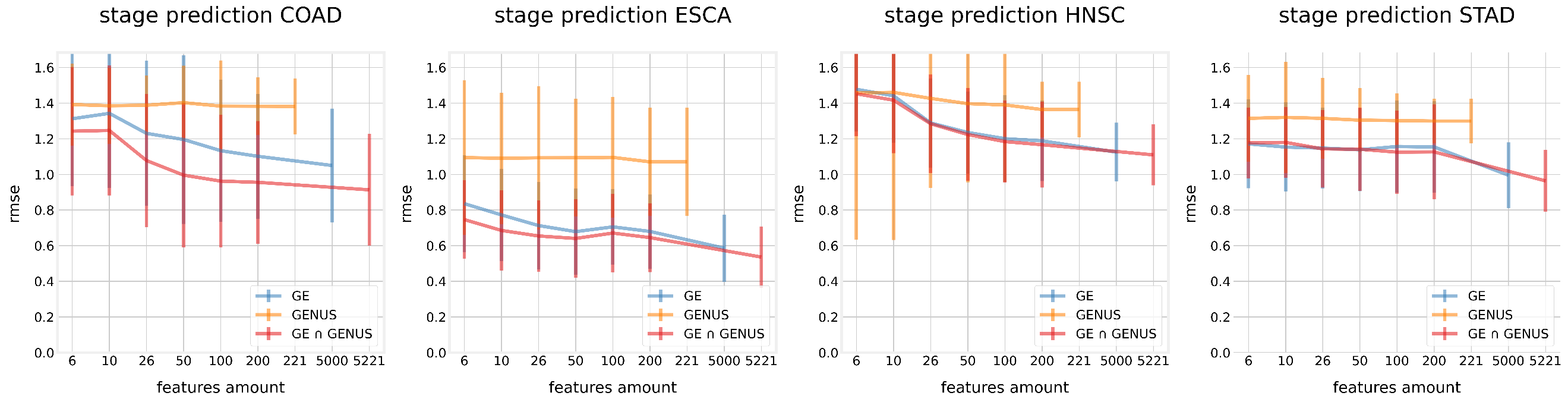
if

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.

### Modality parity enforcement

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

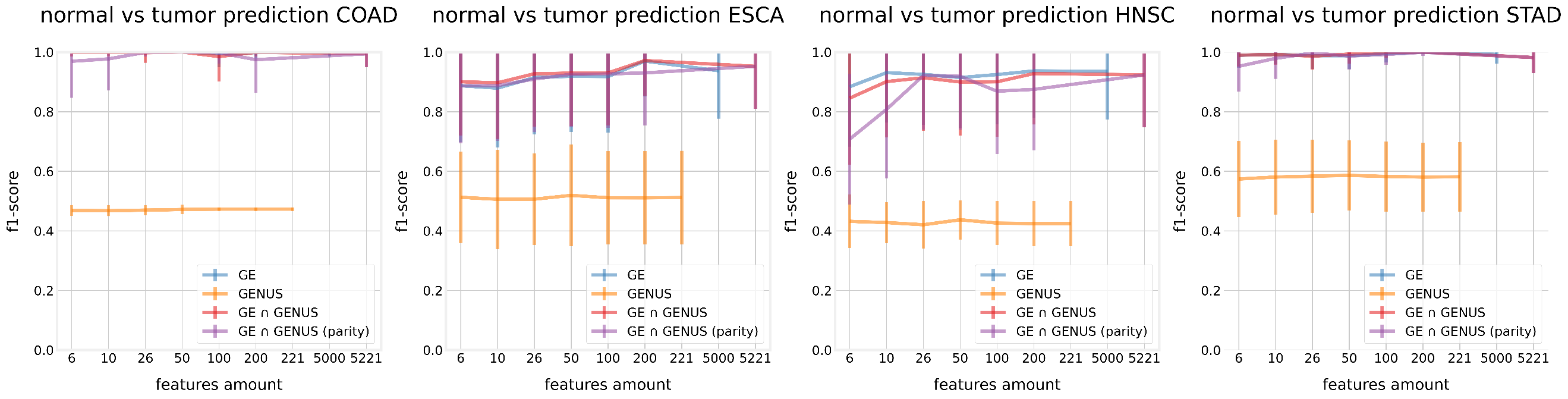


Figure : f1-score for the tumor versus normal prediction endpoint for COAD (colon adenocarcinoma), ESCA (esophageal carcinoma), HNSC (head and neck squamous carcinoma) and STAD (stomach adenocarcinoma), respectively. Each graph contains the f1-score for each modality (i.e. genus abundance data (GENUS), 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 19).

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.

### Complex integration with different models

Complex integration with random Forest

### 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

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

### 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.

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 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.

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 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 44.

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? (

# 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.

Such data sets include the Metagenomics of the Human Intestinal Tract (MetaHIT) data set, containing intestinal and stool microbial data from healthy patients and those with certain noncancer diseases 6, the integrative human microbiome project (iHMP) which 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 7. A problem with these data sets is that they often contain data from tissue swabs and stool samples, which are not necessarily representative of the microbiome of internal organs 5. 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 9. However, the microbial reads in this data set are often a result of contamination 5. These data sets are often not readily available. A data set which tries to combat this, 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 15 and can be affected by many factors such as diet, environmental exposure and lifestyle choices 20. Additionally, microbial communities are unique to each cancer type 9,16. 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 24. 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 17 and between tumor versus normal samples 18. Even further, there are indications that composition and changes in the microbiota has a direct influence on oncogenesis 21–23 and tumor immunotherapy response 22,23. As an example, patients with Parkinson’s have a relatively lower incidence of multiple cancers, possibly through mechanisms involving the microbiota 45. 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 17. 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 20. 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 5,9,11,12, possibly also by using predictor models (e.g. regression) 9,12. 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 5. Additionally, hypothesis tests can be used to see if microbial abundance is predictive of overall survival 5,13 or progression 13. Studies often also investigate the relation with clinical factors such as gender or age 11–13, which can often be confounders. This can be done by using predictor models which use these clinical factors as features 12. 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 14.

However, they can also positively affect patient health, for example, by positively influencing immune system cells to promote antitumor immunity.

This has provided a motivation for the analyses of the impact of microbial composition on disease phenotypes and progression.

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 25, 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.

Dohlman et al. use TCMA and TCGA gene expression, protein expression and methylation data to investigate correlations between bacterial co-abundance groups and expression patterns of certain genes. This is only for colorectal cancer and did not directly use tumor versus normal data or stage and without prediction model but they do link to survival 5. They also use GSEA.

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 13. 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. They correlate microbial abundance with survival, metastasis and tumor grade, . They did not have much validation and validation was lackluster.

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 11. Essentially compare abundance in tumor (T) tissues and non-tumor adjacent (NT) tissues. They validate with TCGA

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 24. 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 24. 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 5.

Table

Description automatically generated

Table

Description automatically generated

For STAD and stage, a model which always predicts the majority class would have a RMSE of 4\*9 + 19 + 0 + 27 + 4\*16 = 146. 146/107 = 1.364 . root = 1.168

the total is 236 and the average stage is 2.206. Let’s call it 2.

For tumor, Precision: 113 / (113 + 9) = 0.926 . Recall = 113 / (113 + 0) = 1.

F1-score = 2 \* (0.926 \* 1 ) / (0.926 + 1) = 0.96

|  |  |  |  |
| --- | --- | --- | --- |
| Rank | Feature name | Frequency selected | Feature type |
| 1 | TDRD9 | 84% | GE |
| 2 | HOXC10 | 83% | GE |
| 3 | PRSS21 | 78.5% | GE |
| 6 | *Helicobacter* | 53% | Genus |
| 136 | *Lactobacillus* | 0.5% | Genus |

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

[

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?

As an example, the amount of diversity within a tumor sample has previously been found to be predictive of survival rates in pancreatic cancer 19. However, these results were found after stratifying patients across clinical variables such as age, gender and stage. This is not possible if there are too few samples, the diversity didn’t differ across these variables actually.

Abundance differences within the gut have a clear effect on tumor progression and treatment response 23. It is not clear whether this is the case for every tissue

Another reason for the lack of results for the stage investigation is that the stage label is based on the TNM standard, where tumors are classified based on the morphology, location and distance with which it has spread. These stages are clinically determined and do not necessarily correlate with expression patterns, leading to difficulties for a model to predict based on these features. Although, in CRC, they correlate 2 different biological entities and contain many differentially expressed genes 46.

Poore et al. 9 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. 10 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.

[ [w Brendan predictor being slightly above]hat selected features of genus making sense and possibl]

microbial abundance was not correlated with tumor grade, but was for metastasis PAAD 13 they use the species level as well

there is a study showing no difference between tumor and normal (same paper)

it is unknown whether differential abundance of taxa is a cause or consequence of developing SCC 11

there was an increase in abundance richness and alpha diversity in tumor and non-tumor tissue compared to immediate autopsy control tissues.(same paper)

In terms of feature selection, penalized regression methods might not properly capture grouping of information 32.

Interestingly, this bacteria has a possible protective effect against esophageal adenocarcinoma 33.

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 47.

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 48. Similar results have held for head and neck squamous carcinoma, with *Fusobacterial* populations showing an increased abundance in tumor versus normal samples 49.

Example:

abbreviation and title and table: <https://www.sciencedirect.com/science/article/pii/S2352396419300635>

Statistical test, contingency table and data and figures

<https://www.nature.com/articles/s41467-022-30512-3#Abs1>

also TCM a paper

problem with genus abundance? https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002687

# References

Automatic citation updates are disabled. To see the bibliography, click Refresh in the Zotero tab.

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)