Self-Prescribed Prescriptions: Personalized Radiation Treatment Using Genomic Biomarkers

Abstract

The goal of this project was to build a predictive model that could correlate genetic features to radiation sensitivity in cancer cell lines. This would ultimately work toward the creation of personalized medicine for cancer patients to help them make more informed decisions while convalescing. The project focused on the incorporation of biological domain knowledge including using tSNE dimensionality reduction and stratification to explore the addition of site histology. Furthermore, we collaborated with cancer biology experts on the curation of biologically informed datasets focused on apoptosis and cell division regulation. Modern machine learning models were tested with random forest outperforming and tuned using Monte-Carlo cross validation. These results were summarized in the creation of an API to handle the machine learning analysis and facilitate further research into personalized medicine development.

Executive Summary

This research project focuses on healthcare machine learning, especially applying properties of biology to create more informed datasets and accurate models. Even though there has been considerable past machine learning research on genomic datasets and correlations with cancer, scientific advancement in this field has proven to be a difficult task.

We studied the impact of radiation therapy, often used with chemotherapy, as a treatment for various cancers. The treatment process is long and arduous, demanding a significant amount of energy and commitment from the patients. Furthermore, radiation therapy works with high variability. Some patients may not experience any shrinkage in tumor size while others may undergo complete recovery. We believe the varying degrees of success of radiation treatment is strongly correlated with genetics. The ultimate goal of our research is to personalize medicine so that patients can feed in their genome and our algorithm reports back whether radiation therapy is worth it.

The bulk of our research consists of using biology to build more informative predictive models. This includes using tSNE dimensionality reduction by cell line histology (ex. lung versus pancreas cells) to stratify gene expression data and Monte-Carlo cross validation to tune sensitive hyperparameters given the interconnectedness of the dataset. We also collaborate directly with cancer biologists at MGH to apply prior biological knowledge. Since genomic data consists of tens of thousands of features, there is a high chance of false positive correlations given the sheer size. We believe understanding the molecular pathways of cancer can help reduce noise and yield more informed datasets for analysis. Utlimately, an API that handles the machine learning side (including data processing) was produced so we can focus on the fundamental question of biology.

Acknowledgements

The process of writing a thesis has been a long and demanding process. I have many people to thank for their constant support and encouragement throughout this journey. I would like to start off thanking my advisers Assistant Professor David Craft of Harvard Medical School/Massachusetts General Hospital and Assistant Professor Yaron Singer of Harvard School of Engineering and Applied Sciences. Their constant pursuit of excellence inspired me to delve deeply into my research. I am especially grateful to David Craft for scheduling regular meetings, responding incredibly quickly to my emails, and connecting me to other talented researchers in this field. I am extremely fortunate to work alongside his lab and benefit from his dedication and constant drive.

Next, I would like to thank Timo Deist for his help with the machine learning tuning algorithm, which is incredibly complex and detailed. I appreciate his words of encouragement as a fellow researcher tackling healthcare machine learning questions and understanding of the complexity of the problem.

I would like to thank Christie Eyler, a full time cancer biologist, for taking the time to help with the creation of the Expert Gene List and going above and beyond by seeking the help of a librarian to generate a comprehensive list with minimal bias. I appreciate her initiative and hope to embody her drive for thoroughness in future research projects.

I am grateful toward Andrew Piatti for his help in assembling the framework and collaborating on the analysis for the cell line analyzer API. I benefitted tremendously from his software engineering expertise and appreciate his constant support, from answering my questions on Slack to video calling to discuss broader concepts. His flexibility and dedication to the project enabled the API to take on a larger role and greater adaptability than initially designed.

Lastly, I am extremely thankful for my parents for supporting my interest in math and science from a young age and investing in my education. I hope to make them proud every day and acknowledge the sacrifices they endured to make my four undergraduate years at Harvard possible. I would also like to thank my friends for patiently listening to me talk about my thesis for over a year and offering never ending support.

To my parents.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Cancer Treatments

Chemotherapy is a long and arduous treatment option with varying levels of success in combating cancer. More than 650,000 patients are undergo chemotherapy every year [1]. The process works by targeting and removing fast-dividing cancer cells across the human body. In most cases, patients undergo chemotherapy to reduce painful symptoms and control the spread of cancerous cells. In the best case, chemotherapy can completely cure an individual of cancer with no chance of remission [14]. Like other cancer treatments, chemotherapy is not without side effects and risks. Since the drugs travel throughout the entire body, they can inflict unintended consequences on healthy, fast-dividing cells such as hair follicles and bone marrow cells. This yields the common side effects of hair loss and anemia from blood cell reduction. The drugs could also hurt vital organs such as the liver or heart, producing potentially life-threatening damage.

An increasingly common treatment for cancer patients is a hybrid of chemotherapy and radiation treatment [6]. Unlike chemotherapy, which administers a drug cocktail through injections, radiation treatment works by utilizing large machinery

to administer dosages of high-energy beams. The machine can focus the beam to selectively target cancer cells within a set perimeter of the body. Although there is always a risk, the side effects tend to be lighter such as fatigue and nausea. Due to the high variability of radiation therapy success, researchers hypothesize that an individual's genome may influence the success of a given treatment. Certain genetic biomarkers may increase cancer cell sensitivity or resilience to radiation.

1.2 Personalized Medicine

This research project aims to explore the growing field of personalized medicine, which generally stratifies patients into categories based on biological characteristics such as gender or age. This allows medical professionals to capitalized on the natural variance occurring in genomes to tailor medicines with higher chances of a cure [5]. Our experiment delves deeply into genomic biomarkers, which are detailed features on the level of DNA bases. Applying machine learning algorithms can help identify the individual features of a cell line that most correlate to sensitivity during radiation treatment. The future goal would be to sequence a patient's genome and report back the success likelihood of radiation therapy. Since the risks of radiation therapy are high, such an algorithm can help patients make better informed decisions while battling cancer.

Chapter 2

Previous Work

Previous research into personalized medicine has shown promise for the use of genomic information in determining cancer treatments. In particular for studying radiation, AUC data is a leading metric for measuring cell sensitivity. This measures the area under the curve of approximate cell line survival in response to varying dosages of radiation. AUC data generation is described more thoroughly in Section 4.1. The following related work summaries describe the motivation of the project including:

1. Past personalized medicine research into small molecules
2. Replication of recent copy number and mutation trend data

2.1 Personalized Medicine Research

The field of personalized medicine is rapidly growing and increasing in prominence. In 2015, Seashore-Ludlow *et al.* [10] published the largest cell line sensitivity dataset with a comprehensive table of collected mutation data across well-researched tumor-related cell lines. The dataset also integrated research on small molecules as predictors of cell line response. The paper confirmed existing relationships between

FDA-approved therapies and genomic patterns as well as identified new trends such as KRAS-mutant cancers.

A later paper by Rees *et al.* 2016 [8] delved deeper into small-molecular or genetic perturbations and their impact on basal gene expression. Rees tested 481 compounds on 823 human cell lines, a subset of which were affect by cancer. Their contributions include a comprehensive repository of gene expression and somatic copy number data. Furthermore, Rees conducted novel discoveries on small molecule behavior, including inhibitor ML239. The molecule was previously regarded as a early harbinger of breast cancer stem cell but was later determined to be the result of fatty acid FADS2 activation. Rees's paper determined how many small molecules relate to each others' mechanisms of action and while individual molecules could be correlated to basal gene expression, no holistic portrayal was offered.

Neither paper was able to identify any overarching trends relating genomic patterns to consistent cell line responses. While a few critical genes were discovered to be correlated, we want to develop a more robust holistic algorithm rather than drill into the behavior of a few genes out of tens of thousands. We view such a behavior as dangerous because it encourages cherry-picking genes with high correlation and retrospectively justifying their inclusion.

2.2 Using Genomic Biomarkers

Using the genomic dataset provided in Yard *et al.* 2016, many of the biomarker-cell line correlations featured in the paper were successfully replicated. Figure 2.1 plots integral survival against cell histology, cancerous tumor type, and cell line, origin in the body [17]. Research suggests a strong correlation between a cell's origination to the integral survival rate, measured by AUC. The black column scatter plots correspond to individual cell lines while the red point represents the average. Although

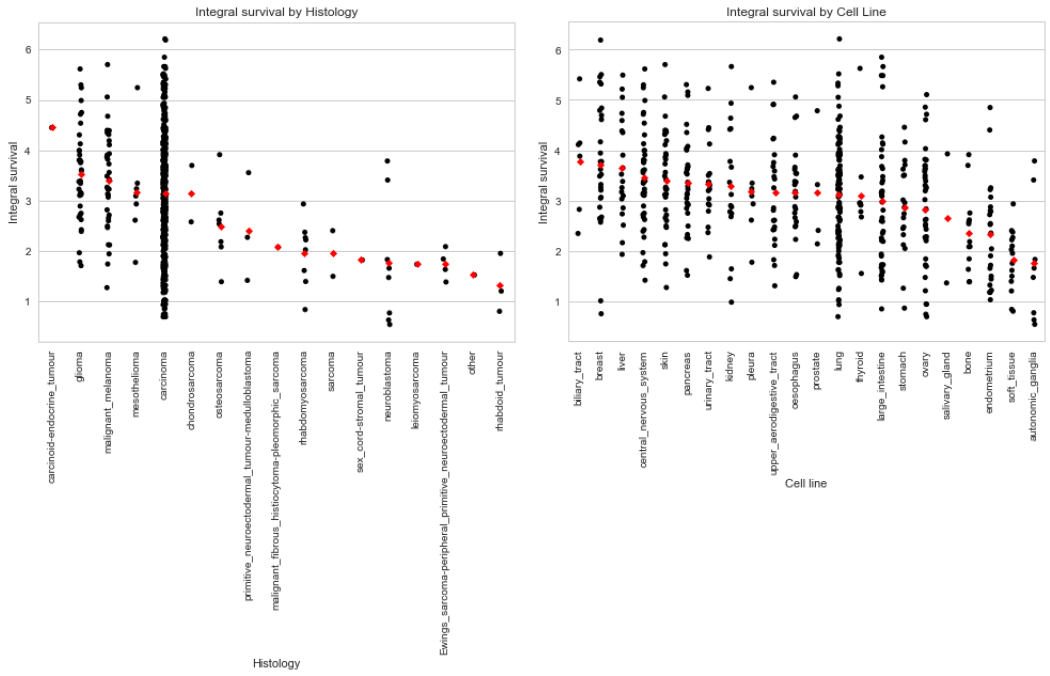

Figure 2.1: AUC survival varies based on cell histology (L) and cell line (R)

additional research is necessary to have a more complete dataset, we observe sharp discrepancies such as the average AUC for glicoma is 3× that of rhabdoid tumours. These promising observations prompt scientists to investigate genomic features such as histology and their relation to radiation sensitivity. Since research in this field is still undergoing rapidly development, there has been no highly successful use of cell line histology or site to inform personalized medicine decisions.

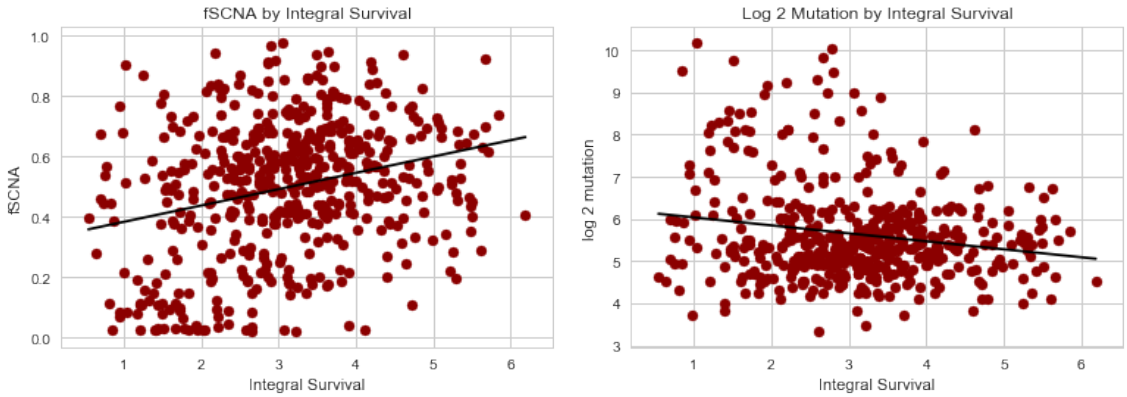

Figure 2.2: AUC sensitivity to copy number (L) and mutation (R)

Furthermore, Yard was able to create a preliminary analysis of more general trends in the genomic biomarker dataset. In Figure 2.2, AUC data is plotted on the x-axis against fSCNA (somatic copy number) and $\log_2$ number of mutations per sample. The data provided exhibits slight evidence of trends relating genomic biomarkers to the cell's sensitivity to radiation. However, this this correlation is weak and subject to arbitrary trends by chance due to the high dimensionality of data points. This prompts us to approach dimensional reduction through not just traditional algorithms but also domain knowledge by using biologically informed datasets. Since these biomarkers are well studied, there are known cancer pathway correlations between features that may help cultivate a smaller and more rigorous dataset.

Chapter 3

Road Map

Our research utilizes modern machine learning techniques to analyze genomic data. The paper is divided into two main sections: 1) machine learning model development and result analysis and 2) the creation of an API to facilitate machine learning analysis given any set of feature matrices and model parameters. Given that the API is a tool to encourage further research, the first segment is more extensive and results-orientated.

3.1 Machine Learning Summary

Using fundamental domain knowledge, we selected biomarker categories with potential correlations to radiation sensitivity. Prior to beginning the model analysis, the biomarker data was carefully cleaned and reformatted to a consistent number of cell lines. After finalizing the datasets, we began with naive model testing including various dimensional analysis techniques to reduce the size of the feature dataset. Next, we tuned our algorithms with the help of the Harvard Odyssey to refine the sensitivity of our models to improve prediction accuracy. Finally, we curated biologically informed datasets as a means of reducing noise by incorporating known domain knowledge. These datasets selected for genes with critical roles in the cancer biological pathway.

The hope is that selecting more informative genes can reduce the rate of false positives and derive more meaningful results.

3.2 API Overview

The API was created to facilitate further research into personalized radiation treatment and make machine learning analysis more approachable to scientists with less computer science background. The API begins by taking in a set of input datasets with additional optional parameters to specify the model analysis. Since other researchers may incorporate domain knowledge differently, producing their own gene list, we want to simplify the analytical barriers for healthcare machine learning research. Additional API toggles such as dataset splits and model specifications allow researchers to tailor the algorithm to best fit their dataset.

Chapter 4

Data Preparation

4.1 Data Overview and Collection

Our research concentrated on four primary datasets that address radiation-impacted cancer cell lines (CCLs). A final 498 CCLs were selected for experimentation comprising 26 cancer types ranging from Lung (most represented) to Carcinoid (least represented). The diversity of CCLs captures site-specific variation and provides a more holistic understanding of cancer sensitivity across the body. The AUC data (1) served as the prediction results and $(2-4)$ represent biomarkers that served as prediction parameters.

1. AUC Data from Yard *et al.* 2016 [17]

   Our research utilizes AUC as a measure of radiation sensitivity and thus, a proxy for the probability of success of cancer radiation treatment. AUC data is calculated as the area under the curve by approximating CCL cell survival in response to increasing dosages of radiation. The area is estimated as a trapezoid and multiplied by the dose interval:

$$\frac{f(X_1) + f(X_2)}{2} \cdot \triangle X$$

Radition dosages were administered in intervals of $1, 2, 3, 4, 5, 6, 8, 10$ Gy, recorded in $\log_2$ transformations, and normalized on a scale of 0 to 7 by the constant $\frac{7}{\log_2 10}$.

2. Mutation Data from Seashore-Ludlow *et al.* 2015 [10]

   The mutation dataset from Seashore-Ludlow consists of general mutations as well as nucleotid specific changes including base substitutions and frameshift mutations such as base deletions. The dataset is substantially smaller than either the copy number or gene expression datasets at 15% feature size. The data was collected by sequencing more than $1,600$ genes and applying mass spectometry genotyping. The process identified 381 critical mutations in 33 known cancer-related cell lines. Mutation data was processed through binary one-hot encoding to address cells with any coding mutation. One limitation is that various mutations (ex. frameshift versus base substitution) were treated identically in our model due to the small dataset size. Separating out specific mutations could yield as low as 1-3 instances for each feature.

3. Copy Number Data from Rees *et al.* 2016 [8]

   Somatic copy number data represents the frequency of a gene being represented in an individual's genome. Duplication of genes occur often as some translocations shift promoters to magnify transcription of select genes across chromosomes. Copy number was calculated using the GenePattern pipeline copy number probe and hg18 Affymetrix probe annotations. The copy numbers were approximated using the probe's set-specific linear calibration curves and normalized to zero. The Rees dataset stores copy number data on a $\log_2$ scale to amplify the variability of copy number data and give greater weight to more extreme values.

4. Gene Expression Data from Rees *et al.* 2016 [2] [3]

   Gene expression data corresponds to the intensity of mRNA gene translation, which can vary widely as a result of transcription factors and epigenetics. mRNA expression can be measured through a probe set using Robust Multi-array Average (RMA) and normalized by quantile. Gene expression values were supplemented with ssGSEA enrichment scores, which measure the degree of co-ordination between genes in a gene set, to enhance the dataset with higher-level interpretability.

4.2 Data Compilation

The data compilation process was extensive to ensure the various biomarker datasets would be compatible for machine learning analysis. Mutation values were especially processed to extract corresponding index CCLs, later converted to master CCLs, for each cell line feature mutation. All feature datasets were transformed to a standardized format of CCL IDs along the rows and genetic features as columns. For naive testing, biomarker datasets were analyzed both independently and jointly, in which the features were intersected together to form a larger dataset.

Biomarker	# Features	# CCL IDs
Mutation	3,235	473
Copy Number	23,174	498
Gene Expression	18,543	498
Total	44,952	473

Table 4.1: Biomarket Dataset Sizes with Limited CCL IDs

We observe in table 4.1 that copy number and gene expression data both have the max nubmer of CCL IDs at 498 while mutation data is short 25 cell lines, which is a 5% reduction in the number of samples. Since the number of features is already significantly higher than the number of samples, mutation data was dropped from

Biomarker	# Features	# CCL IDs
Copy Number	23,174	498
Gene Expression	18,543	498
Total	41,717	498

Table 4.2: Biomarket Dataset Sizes without Mutation Data

our model. For a dataset with $p >> n$, it is crucial to preserve as many cell lines as possible to increase model accuracy. Furthermore, mutation data was generated through one-hot encoding on any coding mutation, which fails to capture the intricacies between different mutations. Thus, the decision is justified due to their limited informability and low number of features compared to the other biomarkers. After dropping mutation data, the number of CCL IDs is consistent across all biomarkers at 498 to maximize the dataset size while enabling the inclusion of informative biomarkers.

Chapter 5

Model Analysis

5.1 tSNE

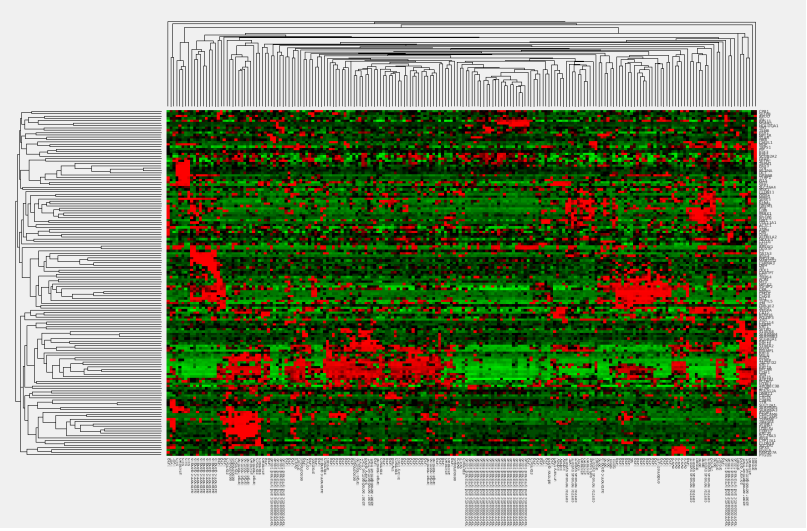

Figure 5.1: Summarized heat map of gene expression exhibits clear patterns when clustered by histology (Figure courtesy of David Craft)

The biomarker data was first analyzed for feature engineering purposes using genomic properties. In Figure 5.1, the cell lines were clustered by histology in a dendrogram and a heat map was generated using relative gene expression values. The red coloring indicates higher levels of gene expression correlation between the cell lines. There are evident clusters of red within the heat map, particularly in the

lower left quadrant, suggesting strong influence of histology on gene expression. These trends were further explored using tSNE dimensionality reduction.

tSNE is a non-linear dimensional reduction technique that clusters points based on a probability distribution of pairwise selection. The probability distribution is defined by minimizing the Kullback-Leibler divergence, which is a measure of relative entropy [13]. tSNE was applied after using PCA to reduce the biomarker dataset to 100 dimensions. This captured a comprehensive total 98.98% variance within the dataset.

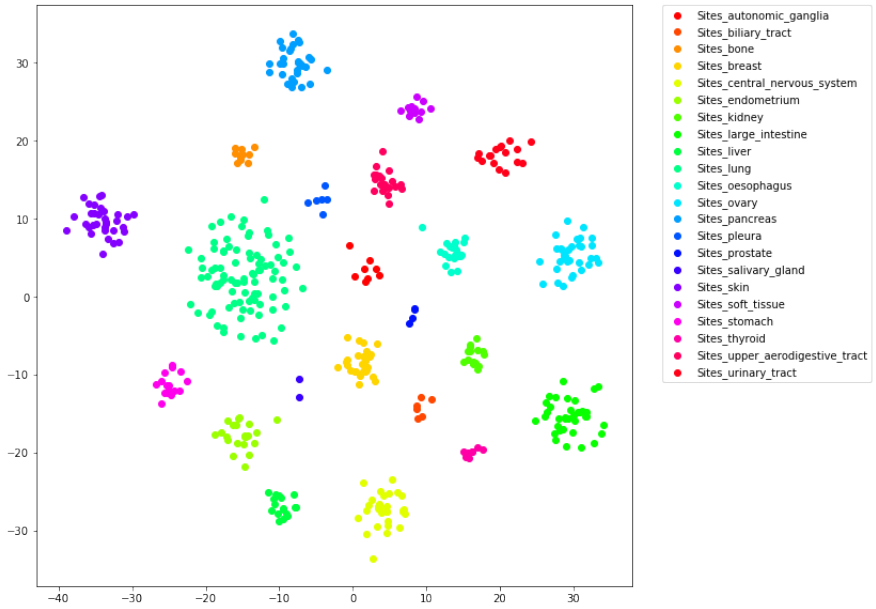

Figure 5.2: tSNE reduction applied to gene expression data

After analysis, the 2-dimensional tSNE plot shows clear clustering of data points based on histology and emphasizes how genomic features can be useful in building informed machine learning models. Although there may be a few points further out from the center of the clusters, no two sites overlap and a clear boundary can be drawn with a nonlinear algorithm. The results of tSNE emphasize the need for data stratification [7]; the clear separation of data points show that different histologies exhibit inherent, individual properties.

5.2 Stratification

While conducting the training/testing split, the dataset should be approximately evenly split along CCL histology, as informed by the clear clustering in tSNE. If not stratified, it is possible that all prostate cell lines end up in the testing dataset with no reliable prostate training set cell lines to accurately predict radiation AUC. Given the inherent differences across cell lines, conflating histologies could result in additional noise and perform poorly in the model. Histology was captured with in-place one-hot encoding for each cell line. The testing and training sets were divided to maintain a equal proportion of all types of cell histologies within the 80/20 split. Overall, the addition of stratification agrees with our intuition because the training dataset is now better representative of the entire datset.

5.3 Naive Models

Following data preparation and cleaning, we proceeded with Naive model testing to visualize the data and understand the problem specifics. A combination of linear and non-linear higher dimension models were selected for full coverage of the dataset. The straightforward nature of linear models would prevent overfitting to the dataset while the nonlinear techniques could encode additional complexity.

Linear Models:

- Linear Regression
- Elastic Net
- Lasso
- SVR Linear

Nonlinear Models:

- SVR RBF
- Random Forest

Datasets were split on a 80/20 training testing split for analysis. Lasso and Elastic Net were fit with their default parameters of $\alpha = 0.1$. The naive models were poor predictors of radiation AUC with the non-linear models significantly outperforming the linear models due to their ability to encode additional complexity. Random Forest yielded the highest accuracy with r-squared of 0.22 while elastic net had the lowest r-squared of near 0, suggesting that the correlation is as arbitrary as guessing the median. These low r-squared values indicate that using naive models on all feature data exhibit weak correlation.

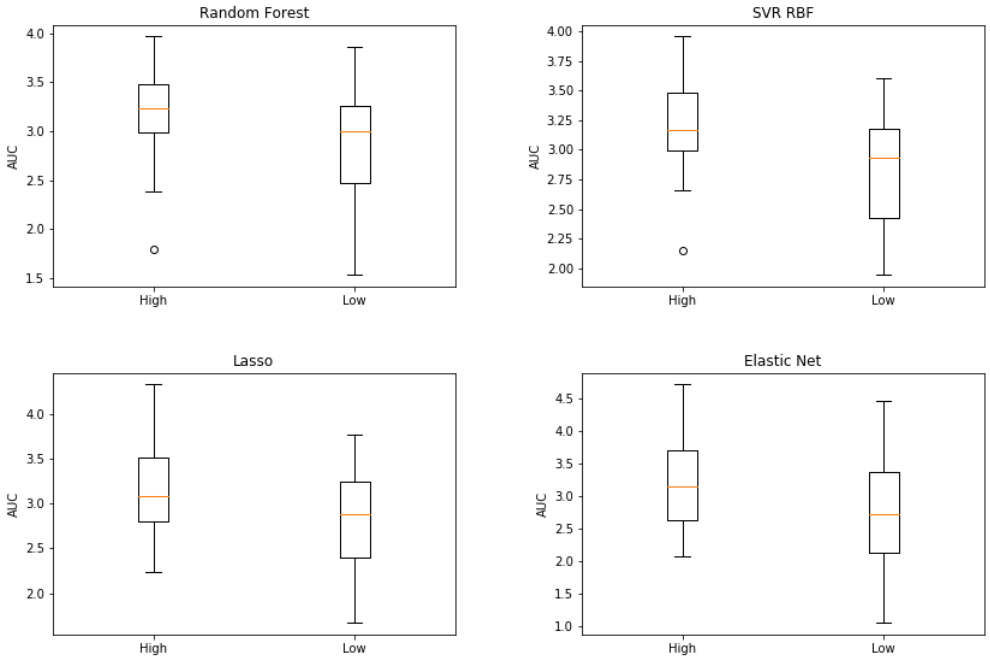

Figure 5.3: AUC can be distinguished using non-linear and linear model output

Howver, in the box and whiskers plot in Figure 5.3, we see evidence that the models have potential in their ability to distinguish between radiation AUC values. The AUC values were split 50/50 into high and low categories based on the dataset mean and correlate with model output results. While the means are not statistically significantly separated, we see clear trends that the high category also yields higher model predicted AUC values. Overall, the high category also has lower variance than the low category, suggesting that there may be more in common between cell lines that are highly sensitive to radiation. Additional feature selection steps are necessary to improve the predictive accuracy of our models.

5.4 Principal Components Analysis

Since the dataset is feature-heavy, with nearly 100 times as many features as CCL IDs, principle component analysis (PCA) was applied to reduce dimensionality by finding the axes of maximal variance. This would decrease noise within the dataset by reducing the chance of arbitrary correlations and strengthen the true informative parameters of the dataset. PCA was applied to reduce down to 50 dimensions, capturing 85.3% of all variance.

After applying PCA, the overall accuracy of the models increased. Linear regression, elastic net, Lasso, and SVR linear all reached r-squared values of 0.23. In particular, the linear models performed well, indicating that feature tuning may be necessary to boost the performance of nonlinear models. We hypothesize that following PCA reduction, linear models succeeded in their simplicity. In closer examination of the model outputs, SVR RBF filtered the model too deeply and ended up predicting the same AUC for all inputs, yielding a poor r-square value of essentially 0. These test results help inform the direction of future research to attempt to build more accurate models.

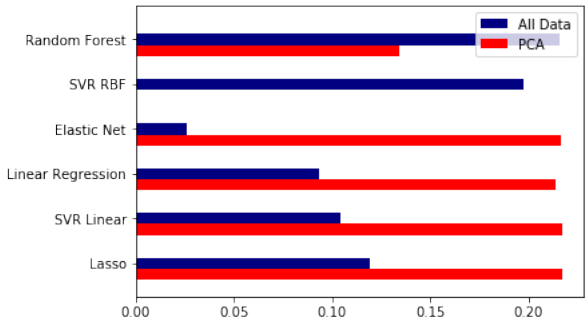

Figure 5.4: Naive model accuracy for all biomarkers and PCA

Overall, PCA dimensionality reduction is not commonly used in genomic datasets because the improvements to model performance are minimal and not intuitive. As evident in prior biological research, diseases can often be attributed to a specific mutation or chain of transcription malfunctions. Applying PCA, which reduces the entire genome to 50 (or some number of) dimensions does not intuitively align with prior genetic discoveries and the 50 dimensions are rendered uninterpretable both for computer scientists and for biologists, who often study the influence of particular sets of genes.

Another common dimensionality reduction technique was not applicable due to the nature of genetic data. Despite generating correlation matrices and removing highly correlated duplicates being a common strategy [4], it was not useful due to the high number of features. It is likely that two biomarkers may be correlated purely by chance so we are unable to distinguish which one is actually informative. Thus, relying on correlation matrices for a high number of features is not just computationally expensive but may actually result in the dropping of informative features. Other dimensionality reduction techniques that incorporate domain knowledge may be more successful.

Chapter 6

Parameter Tuning

Our research proceeded using the most successful model from naive testing: random forest. While the linear models outperformed random forest on PCA compressed data, we chose to proceed with random forest because it was more accurate on all biomarker data, has more precise tuning ability, and is more widespread applicable to other biological datasets.

6.1 Tuning Algorithm

Due to the relative success of linear models against deep learning models seen in section 4.2 PCA, we hypothesize that the nonlinear models are not accurately tuned. Because of the additional complexity these models can encode, we expect their performance to surpass the more straightforward linear regressions.

Figure 5.1 provides a pseudocode overview of the Monte-Carlo cross validation procedure taken to tune the nonlinear models. Monte-Carlo cross validation randomly and independently samples data, allowing for the same data points to show up across multiple samples. In contrast, k-fold cross-validation stratifies the dataset into different tiers with no overlap such that each data point is utilized once. Although both methodologies have their strengths, Monte-Carlo cross validation was selected

as a lower variance sampling process and a more representative way of handling a large and highly varied dataset [16].

The algorithm is divided into an outer and inner loop with the MCCV process occurring twice. The outer runs the initial 80/20 training-testing split on the feature dataset. The inner loop then further divides the outer-training data for tuning/validation and testing purposes on an 80/20 split. Each algorithm is tuned to optimize hyperparameter configurations and select optimal features based on performance of the inner-testing data. For random forest, the number of trees was tuned through the hyperparamter `n_estimators`. Following hyperparameter tuning, r-squared values were calculated based on outer-testing data. After the completion of the outer loop, these r-square values were averaged and compared to determine the best performing model.

The tuning algorithm does require prior knowledge of the hyperparamters including suggested ranges for testing and default values. If more than one hyperparameter requires tuning, a comprehensive grid search will be conducted. Furthermore, given the large feature dataset size, the algorithm becomes computationally intensive to run with a large set of tuning parameters.

6.2 Harvard Odyssey

Due to the heavy computational intensity required by the tuning algorithm, the Harvard Odyssey was employed to increase processing time. The Odyssey is a large computing cluster that enables parallel processing computation with Intel 64 core units and AMD x86_64 architectures. In 2015, over 25.7 million jobs were completed, requiring 240 million hours of computation [12]. Access was granted through the FAS Research Computing Group at Harvard University and has facilitated the data collection process. The algorithmic procedure was divided into batches, made pos-

Figure 6.1: Tuning Algorithm Pseudocode

Algorithm 1: Experimental design

```
for outer MCCV repetition i = 1 : 10 do
    randomly sample 80% of all rows as outer-training data;
    assign the remaining rows as outer-holdout data;
    for inner MCCV repetition j = 1 : 10 do
        randomly subsample 80% of all outer-training rows as inner-training
         data;
        assign the remaining rows as inner-holdout data;
        foreach algorithm a ∈ A do
            foreach hyperparameter configuration h_a ∈ H_a do
                train algorithm a with hyperparameter configuration h on
                 inner-training data;
                predict outcomes for inner-holdout data;
                compute performance metric p_{a,h_a,j} on inner-holdout data
                 predictions;
            end
        end
    end
    compute p_{a,h_a}, the average of the 10 performance metrics p_{a,h,j} for each
     algorithm a and hyperparemeter configuration h_a;
    foreach algorithm a ∈ A do
        select hyperparameter configuration h*_a with best average inner-holdout
         performance metric p_{a,h_a};
        train algorithm a with hyperparameter configuration h*_a on
         outer-training data;
        predict outcomes for outer-holdout data;
        compute performance metric P_{a,i} on outer-holdout data predictions;
    end
end
compute P_a, the average of the 10 performance metrics P_{a,i} for each algorithm
 a;
compare P_a
```

A is the set of algorithms a: random forest, linear SVM, RBF SVM, etc.

H_a is the set of possible hyperparameters combinations h_a for algorithm a.

MCCV is Monte-Carlo Cross Validation

sible by the iterative nature of MCCV, and the results were averaged together post runs. As seen in Figure 6.2, random forest performance was optimized at $n = 46$ for `n_estimators`. For lower values, the model is under tuned and for higher values, random forest starts to overfit to the dataset. Although there is high variation within an individual run, when averaged over iterations, the r-square values start to converge. The blue line signifies a single run and the orange line corresponds to the averaged final values. The r-squared values are not wholly compatible with naive analysis due to using a different methodology, MCCV, but allow us to optimize parameters that can be expended for future analysis.

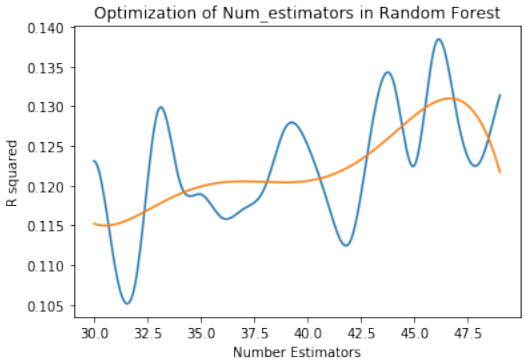

Figure 6.2: Random forest number of estimators parameter optimization

Chapter 7

Biologically Informed Datasets

7.1 Motivation

Another approach of feature engineering is to investigate the inherent nature of the dataset and incorporate domain knowledge. As we are using genetic biomarker data, a significant portion of these genes have been previously researched and their functionality is known. By curating a biologically informed dataset, we can reduce the number of false positives and cut down on noise by simplifying the dataset to relevant genes [9]. However, a gene's functionality may change when influenced by radiation. It is possible that a carcinogenic gene mutation may become harmless or a benign gene variant becomes carcinogenic and resilient to cancer treatments. Thus the outputs of using prior knowledge are highly variable depending on the gene list. Furthermore, the result would be worse if researchers fish around for relevant genes with high correlation scores and produce a justification in retrospect. Thus, using these gene lists should be done holistically and generated through thorough library searches to ensure the ethics of the process. Using biologically informed datasets is a powerful tool but one that must be implemented carefully for truly objective analysis.

In collaborating with multiple professors at Harvard Medical School and Massachusetts General Hospital in the Radiology Oncology department, a list of 200 genes relevant to the onset and severity of cancer was compiled. This is by no means a comprehensive or finalized gene list but merely our attempt to control for inherent exposure and biases to genes while representing the human cancer pathway. Machine learning analysis was then conducted on our curated gene list in hopes of constructing more accurate predictive models.

7.2 Gene List

Feature	Value Set
Expert #1 Inclusion	$\{0, 1\}$
Expert #2 Inclusion	$\{0, 1\}$
Expert #3 Inclusion	$\{0, 1\}$
Number of Citation References	$\{0, 1, 2, 3\}$
Total Score	$\{0, 1, 2, 3, 4, 5, 6\}$

Table 7.1: Library search to generate biologically informed dataset

The full 200 gene list was determine through a variety of sources to control for any scientist's individual biases and exposure. Three cancer biology experts were consulted from Massachusetts General Hospital who independently generated their own gene lists of $\sim 100-300$ genes each. Next, PubMed was scrapped for references to each gene and sorted by frequency of observation. This library wide search aimed to counteract the potential bias from an individual researcher and shed light on potential forgotten genes. The top 60 of genes were prescribed a citation score of 3. Subsequent groups of 60 were prescribed a citation scores of 2 and 1. Any remaining genes were given score 0 for a dearth of citation references and relevance. As summarized in Table 7.1, genes with a total score of 2 or higher out of a maximum score of 6 were added to the curated dataset. It is worth noting that genes overlooked by experts but highly present in the cancer pathway through citations will be included. Furthermore,

for more obscure genes, at least two experts need to independently agree to include the gene for its addition into the dataset.

An example of five genes with the highest score of 6 are profiled below in Table 7.2 with a short description of their functionality in relation to cancer pathways. These genes have known mutants and variations in gene copy number that may influence radiation sensitivity. The reason for selection covers the current academic understanding of that gene's role in the cancer pathway and how it might react when exposed to radiation treatment, if known. A majority of the highest scoring genes regulate apoptosis and cell death, intuitively making sense with the reproductive mechanism of fast-dividing cancer cells. However, the functionality of these genes after undergoing radiation treatment has not been documented and it is uncertain if their behavior remains the same.

HUGO ID	**Gene ID**	**Reason for Selection**
PTEN	5728	PTEN is a common gene found across the body in almost all tissues. The gene regulates cell division by controlling division time and thus, functions as a tumor supressor. PTEN removes phosphate groups on proteins and fats. If cell division is disrupted, PTEN will trigger the cell to enter apoptosis. Outside of cell division, PTEN also functions to maintain cell stability through controlling movement, adhesion, and angiogenesis, and formation of blood vessels.

Table 7.2: Select Carcinogenic Genes

HUGO ID	Gene ID	Reason for Selection
TP53	7157	TP53 or tumor protein p53 is a well-studied tumor suppressor that regulates cell division. Cancer cells are able to proliferate through rapid and unregulated cell division so mutations in TP53 are commonly linked to breast, lung, ovarian, and other types of cancer. The protein can be found in the cell nucleus and binds directly to DNA to repair DNA if damaged. If the DNA is unrepairable, the protein stops cell division and undergoes apoptosis. This helps ensure all body cells remain healthy and the damaged cells are unable to reproduce.
KRAS	3845	The KRAS gene is responsible for the RAS/MAPK signaling pathway that connects the outside of the cell to the cell nucleus. Select signals can trigger cell division or differentiation. When KRAS is mutated, it can yield cancerous results in which normal cell division and apoptosis functionality are disrupted.
BCL2	596	BCL2 controls apoptosis by producing an outer mitochondrial membrane protein that controls mitochondrial membrane permeability. BCL2 functions in a feedback loop with caspases to influence the cell death pathway. There are known carcinogenic BCL2 mutants formed through alternative splicing of chromosome 18.
ATM	472	The ATM gene controls the making of cell nucleus proteins that influence the rate of cell division. The protein impacts the regulation and normal functional development of prominent systems in the human body including the nervous system and immune system. Furthermore, the ATM protein can aid in DNA repair of damaged strands from toxic chemicals or natural errors during chromosomal cell division. ATM's role in DNA repair helps ensure the stability of the DNA strand and the cell's genetic information.

Table 7.3: Select Carcinogenic Genes (continued)

7.3 Data Analysis

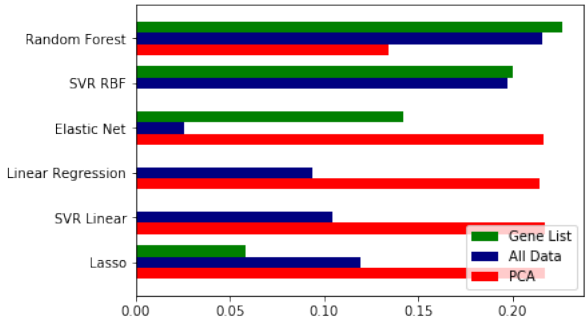

Figure 7.1: Comparative Accuracy with Domain Knowledge Addition

Inclusion of biologically informed knowledge was most successful on the non-linear models, boosting the accuracy slightly of random forest and SVR RBF over the all data results. For linear models, the reduction in the number of genetic data points decreased the accuracy of linear models tremendously, suggesting that either 1) the previous trends may have been arbitrary due to the high volume of past data or 2) the selected gene list is a poor representation of the informative parameters under a linear model. However, we note that these improvements are marginal and overall, the model still requires much improvement. The increase in the accuracy does signal the potential of domain knowledge with feature engineering. We are able to generate competitive model results with a much smaller and intuitive dataset. This would make it more time and cost efficient to generate relevant genetic datasets and facilitate collaboration across biology and computer science departments. Sharing this research with scientists can encourage a range of methodologies to incorporate domain knowledge.

Chapter 8

API Development

The quest to build informative biological algorithms using genomic data has been in progress for decades. With the billion dollar price tag for a cure, many computer scientists have fruitlessly built deep-learning algorithms in attempts to solve this problem. This lack of success may be the result of 1) limitations of current computer science knowledge and model sophistication or 2) the realization that domain knowledge is crucial to increase model accuracy. I believe the future of healthcare machine learning lies in biological feature engineering and selection. To encourage more biologists to take active roles in machine learning research, even with less technical backgrounds, we built an API to facilitate healthcare machine learning research on bioinformed datasets.

8.1 API Overview

The Cell Line Analyzer API is comprised of three separate channels:

1. **Argument Processing Service**: data validation and process initialization
2. **Data Formatting Service**: data pre-processing and formatting

3. **Machine Learning Service**: applications of machine learning models and hyperparameter tuning

8.2 Argument Processing Service

Argument Processing Service validates the given input matrices or otherwise instructs the user on the requirements for pre-processed data. Since we are focusing on applying biological domain knowledge to feature selection, the efforts were focused on machine learning theory rather than data cleaning, which is highly dependent on the dataset given and difficult to generalize. Thus, we expect the data to be formatted strictly with no missing values (must be pre-imputed).

For each X-value dataset representing features:

$$n = \text{number of samples (CCL IDs)}$$
$$m_p = \text{number of features of type } p \text{ where}$$
$$p \in \{\text{mutation, gene expression, copy number, etc.}\}$$

Thus, X-value datasets which represent a parameter p such as mutation must take dimensions $n \times m_p$. X-value datasets must contain a column header with the HUGO gene label. This agrees with academic cancer biology notation and facilitates the process of identifying important genes with a more interpretable nomenclature. All X-value datasets are assumed to be of the same type, either categorical or numerical.

Y-value datasets which represent results such as AUC values much take dimensions $n \times 2$. The first column of the Y-value dataset corresponds to a list of CCL IDs and the second column contains the result values for analysis. All values within the Y dataset are assumed to be numerical.

To run the API, the `arguments.txt` file must be updated with the correct input parameters. `results` takes in a string with the file name of the CSV. `data_split` accepts an array of length three with percentages corresponding to the desired train, validation, and test set divisions. The three percentages must sum to 100%. `important_features` takes in a comma separated list of feature names from designated files. These features will be given additional weights in the machine learning analysis and allows for insertion of domain knowledge to strengthen the model. `is_classifier` toggles between regression with value `0` and classification with value `1`. A completed `arguments.txt` file with correct notation can be found below.

```
results=results.csv
data_split=[80,10,10]
important_features=features.feature_two, categorical.feature_cat
is_classifier=1
```

8.3 Data Formatting Service

Data Formatting Service handles the pre-processing of the datasets. In handling categorical variables, data formatting service allows for in-place one-hot encoding or expansion of one-hot encoding in binary features. The in-place encoding enables the option of stratification of the dataset during the train/test split on a categorical variable. The binary one-hot encoding is recommended for machine learning analysis of categorical variables as a more interpretable result.

If there is reason to believe a given variable strongly influences the remaining features, the train/test split should be conducted with stratification to ensure a comprehensive training set. The option for proceeding without stratification is also available. The output of the data formatting service returns six outputs. `features` captures a list of genetic features such as gene mutations and copy number values.

`is_classifier` outputs `True` or `False` for whether the program desirse classification analysis. If `False`, the program will interpret the dataset as a regression problem. `results` returns the array of Y values, which was determined in the `arguments.txt` folder. `trainingMatrix` represents each cell line as a labeled list with the cell line CCL ID as the label and an array of parameter values corresponding to the order of features set by `features`. `testingMatrix` and `validationMatrix` are formatted identically with their respective datasets. This notation, while involved, makes it easy to examine the data and select for individual cell lines or features that influence the dataset.

8.4 Machine Learning Analysis

In the machine learning analysis segment, random forest is tuned and tested on the given datasets. Tuning is done through multiple iterations of Monte-Carlo cross validation in accordance to Figure 5.1: Tuning Algorithm Pseudocode.

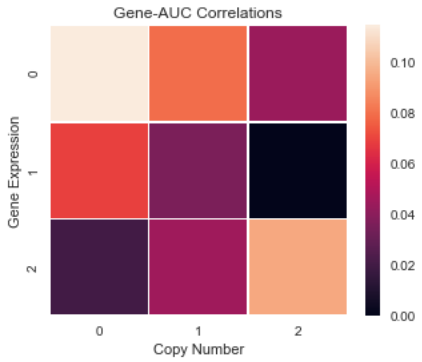

Figure 8.1: Comparative Heat Map of Various Gene List

The idea behind the analysis is to facilitate the study of select gene lists to easily incorporate domain knowledge. Multiple gene lists can be inserted simultaneously to compare how pairwise analysis of datasets performs. A preliminary analysis was conducted on Expert Gene List (described in section 6.2), the L1000 List (a widely accepted list of $\sim$ 1000 genes that represent the larger genetic library [11]), and a

tissue differentiation list from Xu *et al.* 2016 [15]. Unfortunately, this combination of gene lists still leaves much to be desired. The best result surprisingly comes from utilizing the Expert Gene List on both the copy number and gene expression dataset. We hypothesize that this observation may be because the Expert Gene List was tailor constructed to address the cancer pathway, unlike the other two lists which only broadly acknowledge the disease. The purpose of the API directly addresses this issue head on. By facilitating the machine learning process, biologists can test out their own gene lists or otherwise incorporate domain knowledge in a way previously blocked by computational barriers.

8.5 Running the Code

The code can be downloaded here on GitHub and run with `python __main__.py`.

Enter `0` to access the Cell Line Analyzer and the path to the desired folder containing `arugments.txt`, feature datasets, and the output dataset. The results will be printed in the terminal or can be saved to a CSV by passing in the target folder as an argument, `python __main__.py PATH/TO/FOLDER`.

Chapter 9

Conclusion

9.1 Research Summary

The research in this paper was divided into two segments: 1) model analysis and tuning and 2) API development. During the feature engineering process, using tSNE dimensionality reduction proved that different cell line histologies share inherent properties that make domain knowledge valuable to include. Furthermore, dataset stratification was built in to evenly divide the data between sites. After testing a combination of linear and non-linear models, random forest performed the best. However, it is clear based on the r-squared accuracies that significant tuning and model development is still required. After implementing Monte-Carlo cross validation, random forest was tuned to optimize the number of trees of $n = 46$ on these specific datasets. Combined with with an expert curated gene list of the cancer pathway focused on apoptosis and cell replication, r-squared accuracy reached 0.23. The lessons learned and procedures in this model analysis were all included in the API. Publically available on GitHub, the API can perform machine learning analysis on any combination of features and output parameters.

9.2 Future Work

Over the course of conducting this research, I have been convinced that biological domain knowledge is necessary to tackle genetic datasets. Despite attempts to tune the datasets and utilize creative feature engineering with cell line histologies, the model accuracies were capped. Addressing these issues will either involve significant machine learning model developments, particularly models more equipped to handle cases with a larger number of features than samples ($p >> n$), or the involvement of domain knowledge to reduce the problem dimensionality and scope.

The creation of the API opens the door for future research in personalized radiation therapy. Researchers can test out their own genes in various permutations to incorporate domain knowledge how they see fit. The greatest advantage of the API is its flexibility; it is possible to test various combinations of features and output results. Other than measuring radiation AUC, the API could be utilized, for example, to reflect diabetes insulin levels or blood pressure measurements in response to genetics. The computational barriers which once only enabled computer scientists to investigate these issues are lowered and more scientists can now dedicate their efforts to investigating personalized medicine.

Appendix A

Code Source

The code written to perform machine learning analysis and utilize the API are publically available on Github for download and future research. The API includes a comprehensive README with instructions on set up and input requirements. Full access of the sample and feature datasets can be found from the original cited papers.

1. The analysis for chapters 2 through 6 can be found here.

2. The API generated can be found here.

3. The Expert Gene List can be made available upon request.

Appendix B

Supplemental Figures

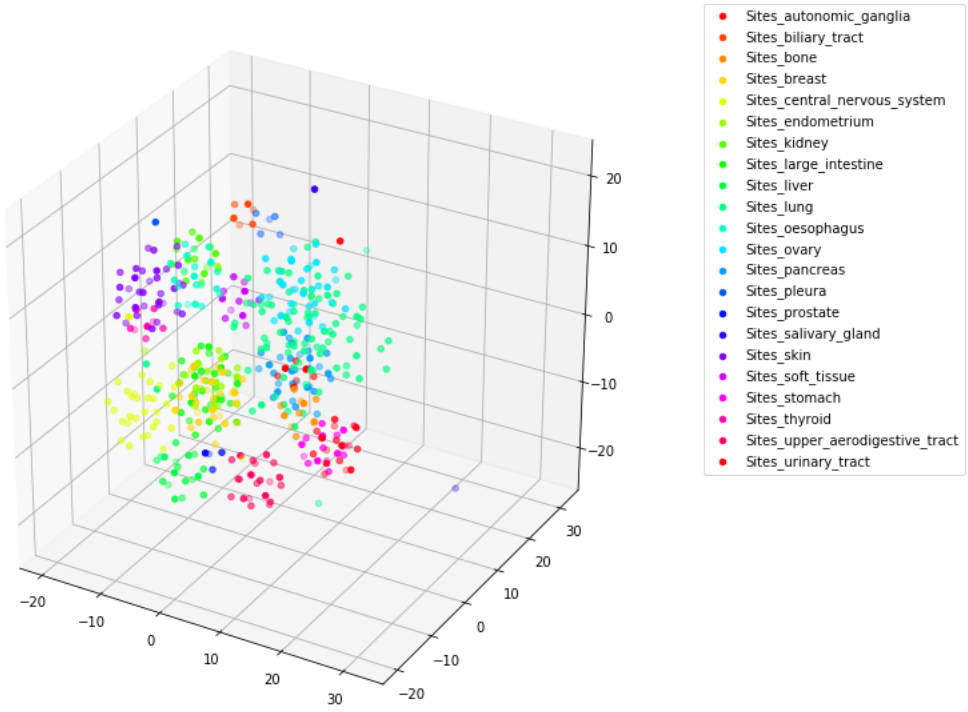

Figure B.1: 3D dimensionality reduction using tSNE

3D dimensionality reduction was also conducted and encodes more complexity than 2D tSNE. While there is clear clustering by cell line histology, there are also cases of overlap between sites. 2D tSNE was included in the paper for having more distinct and interpretable boundaries.

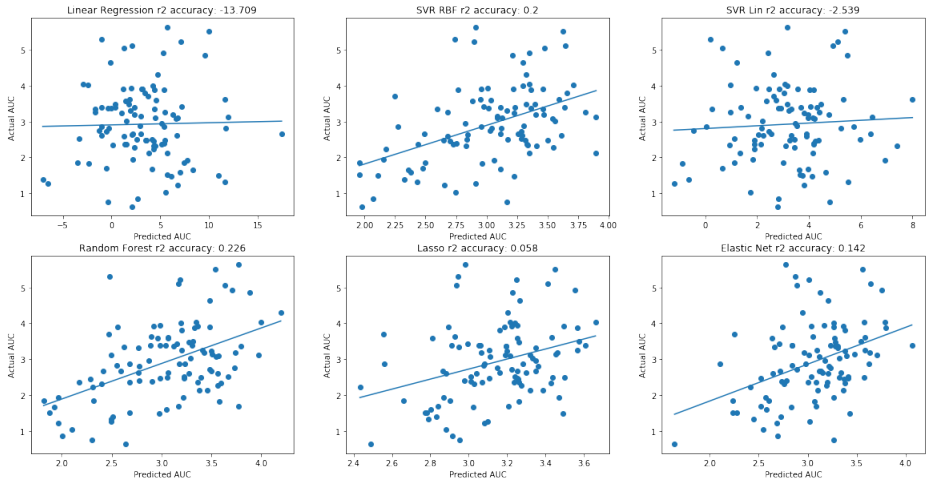

Figure B.2: Estimates following model prediction using Expert Gene List feature selection

Random forest has a clear trend of predicting higher model AUC values for higher true AUC values. However, it is clear that a majority of the linear model predictions are close to random, signified by the flat line.

	Lasso	Lin Reg	Net	RF	SVM Linear	SVM RBF
Naive	0.119339	0.093567	0.02597	0.215765	0.104139	0.197470
PCA	0.213642	0.206254	0.21080	0.133938	0.217092	<0
Gene List	0.058	<0	0.142	0.226	<0	0.2

Table B.1: Raw table of r-square model prediction accuracies

Of the possible raw values, random forest outperforms the other models with the highest r-squared value of 0.23 while incorporating gene selection with the Expert Gene List. A below zero value signifies that the model performed worse than guessing a horizontal line, signaling of a poor and inappropriate fit.

Bibliography

[1] Centers for Disease Control and Prevention. Information for Health Care Providers. https://www.cdc.gov/cancer/preventinfections/providers.htm, 2017.

[2] Freije, W. A.; Castro-Vargas, F. E.; Fang, Z.; Horvath, S.; Cloughesy, T.; Liau, L. M.; Mischel, P. S.; Nelson, S. F. Gene Expression Profiling of Gliomas Strongly Predicts Survival. http://cancerres.aacrjournals.org/content/64/18/6503.short, Sept 2004.

[3] Gravendeel, L. A. M.; Kouwenhoven, M. C. M.; Gevaert, O.; de Rooi, J. J.; Stubbs, A. P.; Duijm, J. E.; Daemen, A.; Bleeker, F. E.; Bralten, L. B. C.; Kloosterhof, N. K.; De Moor, B.; Eilers, P. H. C.; van der Spek, P. J.; Kros, J. M.; Sillevis Smitt, P. A. E., van den Bent, M. J.; French, P. J. Intrinsic Gene Expression Profiles of Gliomas Are a Better Predictor of Survival than Histology.

[4] Hall, M. A. Correlation-based Feature Selection for Machine Learning. https://www.lri.fr/~pierres/donn%E9es/save/these/articles/lpr-queue/hall99correlationbased.pdf, April 1999.

[5] Hamburg, M. A.; Collins, F. S. The Path to Personalized Medicine. http://www.nejm.org/doi/full/10.1056/nejmp1006304#article_citing_articles, July 2010.

[6] Peters, W. A. III; Liu, P. Y.; Barrett, R. J. II; Stock, R. J.; Monk, B. J.; Berek, J. S.; Souhami, L.; Grigsby, P.; Gordon, W. Jr.; Alberts, D. S. Concurrent Chemotherapy and Pelvic Radiation

Therapy Compared With Pelvic Radiation Therapy Alone as Adjuvant Therapy After Radical Surgery in High-Risk Early-Stage Cancer of the Cervix. https://journals.lww.com/obgynsurvey/Abstract/2000/08000/Concurrent_Chemotherapy_and_Pelvic_Radiation.17.aspx, August 2000.

[7] Provost, F. Machine Learning from Imbalanced Data Sets 101.

[8] Rees, M. G., Seashore-Ludlow, B., Cheah, J. H., Adams, D. J., Price, E. V., Gill, S., Javaid, S., Coletti, M. E., Jones, V. L., Bodycombe, N. E., Soule, C. K., Alexander, B., Li, A., Montgomery, P., Kotz, J. D., Hon, C. S., Munoz, B., Liefeld, T., Dancik, V., Haber, D. A., Clish, C. B., Bittker, J. A., Palmer, M., Wagner, B. K., Clemons, P. A., Shamji, A. F., Schreiber, S. L. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. https://www.nature.com/articles/nchembio.1986, Jun 2016.

[9] Saeys, Y.; Inza, I.; Larrañaga, P. A review of feature selection techniques in bioinformatics. https://doi.org/10.1093/bioinformatics/btm344, Oct 2007.

[10] Seashore-Ludlow, B., Rees, M. G., Cheah, J. H., Cokol, M., Price, E. V., Coletti, M. E., Jones, V., Bodycombe, N. E., Soule, C. K., Gould, J., Alexander, B., Li, A., Montgomery, P., Wawer, M. J., Kuru, N., Kotz, J. D., Hon, C. S., Munoz, B., Liefeld, T, Dančík, V, Bittker, J. A., Palmer, M., Bradner, J. E., Shamji, A. F., Clemons, P. A., Schreiber, S. L. Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. http://cancerdiscovery.aacrjournals.org/content/5/11/1210.long, Nov 2015.

[11] Subramanian, A.; Narayan, R.; Corsello, S. M.; Peck, D. D.; Natoli, T. E.; Lu, X.; Gould, J.; Davis, J. F.; Tubelli, A. A.; Asiedu, J. K.; Lahr, D. L.; Hirschman, J. E.; Liu, Z.; Donahue, M.; Julian, B.; Khan, M.; Wadden, D.; Smith, I. C.; Lam, D.; Liberzon, A.; Toder, C.; Bagul, M.; Orzechowski, M.; Enache, O. M.; Piccioni, F.; Johnson, S. A.; Lyons, N. J.; Berger, A. H.; Shamji, A. F.; Brooks, A. N.; Vrcic, A.; Flynn, C.; Rosains, J.; Takeda, D. Y.; Hu, R.; Davison, D.; Lamb, J.; Ardlie, K.; Hogstrom, L.; Greenside, P.; Gray, N. S.; Clemons, P.

A.; Silver, S.; Wu, X.; Zhao, W. N.; Read-Button, W.; Wu, X.; Haggarty, S. J.; Ronco, L. V.; Boehm, J., Schreiber, S. L.; Doench, J. G.; Bittker, J. A.; Root, D. E.; Wong, B.; Golub, T. R. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. http://www.cell.com/cell/abstract/S0092-8674(17)31309-0, Nov 2017.

[12] Harvard University. Odyssey Architecture. https://www.rc.fas.harvard.edu/resources/odyssey-architecture/, 2013.

[13] Van der Maaten, L.J.P.; Hinton, G.E. Visualizing High-Dimensional Data Using t-SNE. http://jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf, Nov 2008.

[14] WebMD. Chemotherapy: How It Works and How You'll Feel. https://www.webmd.com/cancer/chemotherapy-what-to-expect#1, 2018.

[15] Xu, Q., Chen, J., Ni, S., Tan, C., Xu, M., Dong, Lei, Yuan, L., Wang Q., Du, X. Pan-cancer transcriptome analysis reveals a gene expression signature for the identification of tumor tissue origin. https://www.nature.com/articles/modpathol201660, March 2016.

[16] Xu, Q. S.; Liang, Y. Z. Monte Carlo cross validation. https://www.sciencedirect.com/science/article/abs/pii/S0169743900001222, April 2001.

[17] Yard, B., Adams, D. J., Chie, E. K., Tamayo, P., Battaglia, J. S., Gopal, P., Rogacki, K., Pearson, B. E., Phillips, J., Raymond, D. P., Pennell, N. A., Almeida, F., Cheah, J. H., Clemons, P. A., Shamji, A., Peacock, C. D., Schreiber, S. L., Hammerman, P. S., & Abazeed, M. E. A genetic basis for the variation in the vulnerability of cancer to DNA damage. https://www.nature.com/articles/ncomms11428.pdf, Apr 2016.