Final report:

Predicting molecular biomarkers for cancer therapeutic response

1.0 Problem statement:

1.1 Introduction:

The fight against cancer continues to be a global health priority, and personalized cancer therapeutics represent a promising avenue to improve treatment outcomes. In the field of precision medicine, identifying predictive biomarkers for therapeutic response is of paramount importance. Transcriptomic, proteomic, and genomic (OMICS) data offer a wealth of information about the molecular characteristics of patients, which can be leveraged to uncover crucial features that indicate whether a particular therapy will be effective for an individual. In this project, we aim to leverage machine learning and statistical techniques to identify and analyze potential biomarkers that can aid in predicting biomarkers- both enrolment and causative. These biomarkers play a pivotal role in understanding disease progression, patient enrolment in clinical trials, and, most importantly, drug targeting strategies.

1.2 Objective:

The primary objective of this computational project is to develop robust predictive models using state-of-the-art algorithms to identify and assess potential features or regressors associated with enrolment biomarkers and causative biomarkers. In short, the hypothesis we want to test is:

‘What transcriptomic, proteomic, genomic features predict response to a therapy?’

We aim to harness the power of advanced bioinformatics and machine learning techniques to identify predictive transcriptomic, proteomic, and genomic features that are associated with responses to novel and specific cancer chemo-therapeutic agents. By achieving this objective, we strive to enhance our understanding of mechanism of actions (MOA) for drug candidates.

1.3 Criteria for success:

Prediction Accuracy: A well-performing model will have a high precision and recall, as well as a high area under the receiver operating characteristic curve (AUC-ROC). However, often times biological datasets for novel small molecules lack validation datasets/ gold standard. Hence, in absence of such gold standards, precision and recall calculations are not possible and such a dataset only serves as a discovery tool.

Generalizability: The predictive models should exhibit generalizability and perform well when applied to independent and real-world clinical datasets. This means that the models should not be overfit to the training data and can effectively predict treatment responses in diverse patient populations and across various cancer types.

Robust Biomarkers: Success will be measured by the ability to identify robust transcriptomic, proteomic, and genomic biomarkers that consistently and significantly correlate with positive treatment responses. These biomarkers should demonstrate statistical significance and biological relevance, contributing valuable insights into cancer therapeutic mechanisms.

Integration of Multi-Omics Data: The successful integration of diverse omics data (transcriptomic, proteomic, and genomic) should provide a comprehensive understanding of the molecular landscape of cancer and its relationship to treatment responses. The project should demonstrate the effectiveness of combining multiple data types to achieve more accurate predictive models.

Clinical and mechanistic Relevance: The identified predictive biomarkers should have clinical and mechanistic relevance and be associated with well-established cancer therapies or potential drug targets. Success will be achieved by revealing actionable insights that can guide treatment decisions and lead to improved patient outcomes.

Scope of solution space:

The scope of the solution space in this computational project encompasses establishment of the machine learning algorithm for identification of predictive biomarkers for cancer therapeutics using OMICS data.

Constraints of Solution Space:

One potential limitation is the availability of high-quality and comprehensive multi-omics datasets from diverse patient populations, which could impact the generalizability of the predictive models. Additionally, N/P ratio issue (small sample, high feature datasets) causes overfitting in the biologic datasets. Appropriate model needs to be chosen to circumvent that.

Stakeholders:

Potential stakeholders in this computational project are both clinicians and drug developing biotech companies. In the long term, this analysis can directly be applied for bench to bedside research to help improve patient outcome.

2.0 Key Data Sources and acquisition:

Key data sources for cancer diverse and large-scale cancer OMICS datasets include public repositories such as The Cancer Genome Atlas (TCGA), Genomic Data Commons (GDC), and the European Bioinformatics Institute (EBI). However, this databases lack companion chemical and small molecule response information (LogAC50: Logarithm of concentration at which 50% of maximum activity is observed). Hence, we are working with a chemi-genomic dataset outlined in a published manuscript (PMID: 28455392). The dataset has OMICS data paired with compound response data for large amount of patient-derived cell lines. Normalized and analyzed data was available in the form of two csv files- one file corresponds to the cell line OMICS (transcriptomics, copy-number variations, metabolomics, genomics and proteomics) data and the other one enlists drug response data. A subset of 12 novel compounds were tested in a range of ~200 cell lines. From this the transcriptomic dataset was more informative due to technical ease of data acquisition as well as due to enrichment and complexity of transcript level dataset itself. This study was hence limited to only transcriptomic or gene expression analysis.

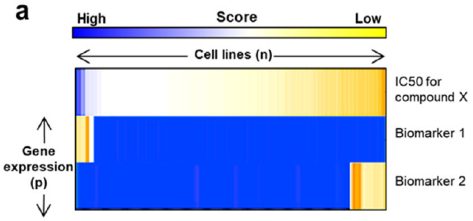

Figure 1: Cartoon illustrating the chemi-genomic datasets. The compound response AC50/IC50 of a cell line (horizontal axis) is correlated with the biomarker level of that cell line. This is final form data.

3.0 Data wrangling:

Discovery, structuring and cleanup:

The csv files were merged into a single file. The gene expression dataset was transposed and drug response data was merged (inner) to generate an intersection of dataset where each rows indicate a gene (n= 54,491) and each column represents a cell line (n= 29). Data was converted to numeric and cleaned up for missing values (NaN and NA) by filling with zero (0).

Exploratory data analysis (EDA):

Mean LogAC50 value was 5.7µM (± 0.6 std.). This concentration is pretty high for an approved drug. However, this compounds are of novel origin and has not yet been improved potency-wise. Expression of the genes varied from zero to few thousands. Gene expressions were already normalized and hence no normalization was performed.

4.0 Modeling:

Since the variables to be compared are of quantitative nature, and not of categorical nature, the modeling is not suitable for classification based algorithms. For this quantitative dataset, from existing literature, there are a few modeling options for this existed-

1) Principal component analysis:
- Pros: Reliable dimensionality reduction; computationally efficient
- Cons: Unable to explicitly highlight specific features (genes, proteins, or genomic variations) driving treatment response, making it less suitable for feature selection tasks; Assumption of linearity; lack of biological meaningfulness

2) Non-zero matrix factorization (NMF):
- Pros: Suitable for capturing non-negative gene expression patterns; provides a more biologically interpretable representation of the data compared to traditional matrix factorization methods like PCA, individual feature selection is possible
- Cons: Sensitive to the choice of the number of factors or components; non-convex nature may lead to multiple local minima; potential overfitting, individual feature selection is computationally intensive

3) Regression
- Pros: Allows easy inclusion of clinical covariates and confounding factors; both positive and negative regulators are identifiable; quantitative
- Cons: Overfitting; may not capture non-linear relationship

For our study, we were focusing on linear relationships. And hence, both NMF and regression were of interest. Regularization (Lasso, Ridge or Elastic net) deals with the overfitting issues better for regression. Hence, consistent with previous literature, a regularized regression was the best model applicable here.

Although, unlike NMF and regression, PCA is not suitable for individual feature isolation, in our analysis, we found that PCA components required to explain variations in the dataset is almost linearly proportional up to 25 dimensions (Figure 1). This indicate that the dataset complexity is not dominated by a handful of outlier components. On the other hand, NMF was highly dependent on the chosen component number, which means an astronomical component numbers need to be investigated to get to gene level information.

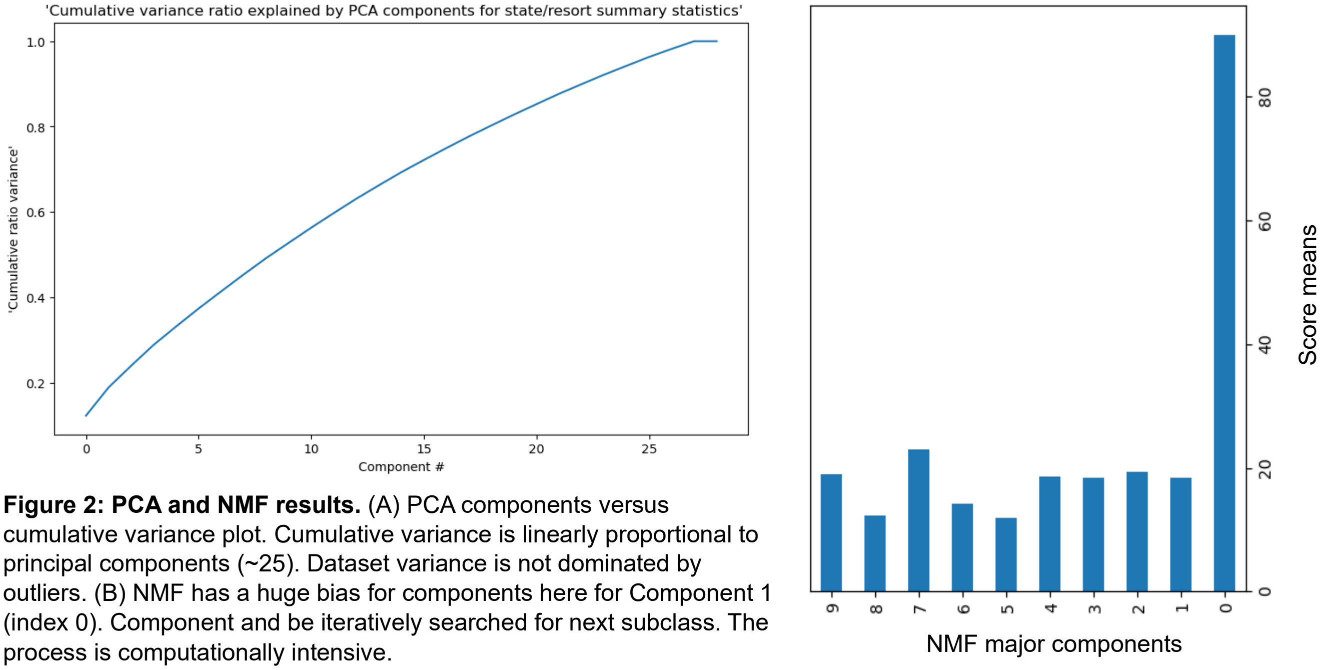

Figure 2: PCA and NMF results. (A) PCA components versus cumulative variance plot. Cumulative variance is linearly proportional to principal components (~25). Dataset variance is not dominated by outliers. (B) NMF has a huge bias for components here for Component 1 (index 0). Component and be iteratively searched for next subclass. The process is computationally intensive.

5.0 Results:

Regularized regression (lasso, L1 and ridge, L2) algorithm Elastic net was picked for the model. A 75-25 train-test split and 10-fold cross validation was chosen. For some of the other parameters, a hyperparameter tune up Gridsearch was performed, where "max_iter" (Maximum number of iterations taken for the solvers to converge) variations of 1, 5, 10, 100, 500, 1000, 5000, 10000, "alpha" (α is the regularization strength) variations of 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, and "l1_ratio" (lasso and ridge penalty ratio) variations of 0.0, 1.0, 0.1 were tested. The results indicated an alpha value drop at 0.0001, with the rest being at default (l1_ratio= 0.5 and max_iter -1000) was the best tune-up parameters.

Each compounds were tested individually and the dataset generated feature importance for the regression curve fitting. From this analysis, we were interested in finding exemplar/ outliers that had disproportionately contributed in terms of feature importance.

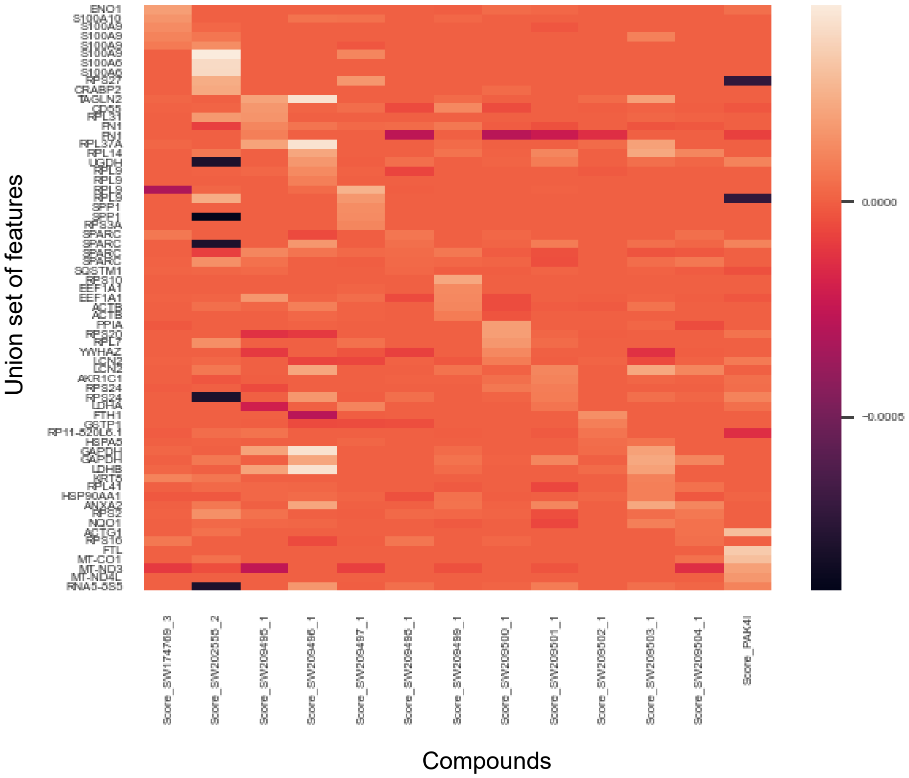

Figure 3: Heatmap for positively correlated genes. X axis represents 13 compounds and Y axis represents the features. The intensity represents the feature importance score.

From this heatmap (Figure 3), on positive side of the spectrum (positive correlator of the response), a compound with known mechanism of action (PAK4), the best biomarker was ACTG1. ACTG1 was also specific since it did not correlate with other compounds as a biomarker. ACTG1 codes for the gamma subunit of the Actine cytoskeleton in the cell. PAK4 inhibitor targets the protein PAK4, a protein that regulates cytoskeleton rearrangement. Hence, there are biologically meaningful connection to be made here. Some of the other biomarkers from this plot were mitochondria related genes. Mitochondrial activity can be associated to the compound associated death and hence potentially non-specific. The second compound (from left) was one of the most promising one. It identified S100 proteins as a major biomarkers. These proteins are calcium binding and hence there may be a calcium binding structural function being affected by this compound. S100 proteins regulate many pathway functions including cellular growth in the body and hence is a good target to start with.

The fourth (from left) compound identified GAPDH as a biomarker. This gene is ubiquitously expressed in the cell and often used as a non-specific hit. The higher the expression of this in cells the higher metabolically active (such as proliferation and etc.). GAPDH being pulled as a candidate most likely indicates that the biomarker is predicting cellular metabolism/growth as a whole and not selective to any particular biological process. This is also indicative of the fact that

3rd, 11th and 12th (from right) compounds also were predicting this as somewhat of a candidate biomarker. These observations indicate that these compounds were mostly not very specific for a particular hit (like chemotherapy) and hence could be globally cytotoxic.

On negative side of the spectrum (negative regulator of the response) (Figure 5), for second compound (from left) UDP-Glucose 6-Dehydrogenase was predicted. It is an enzyme that regulates particular membrane protein such as glycan biology as well as many other functions. It also found secreted phospho proteins (SPP1) as a negative biomarker. These observations indicate that protein ligand binding process (which S100 proteins are also involved in), may be a potential mechanism of action for this protein.

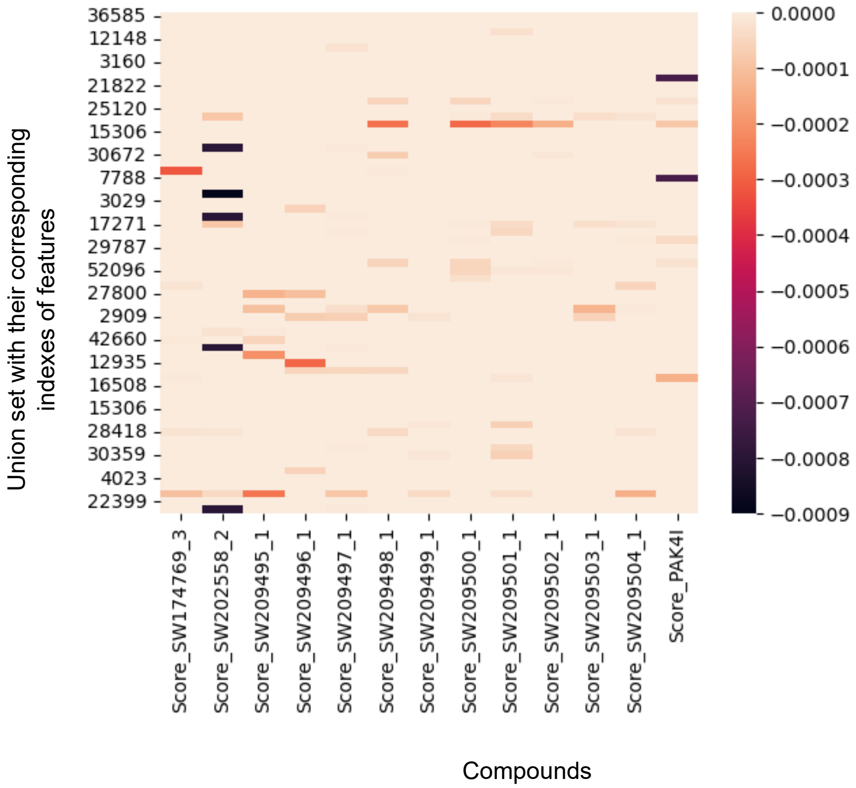

Figure 4: Heatmap for negatively correlated genes. X axis represents 13 compounds and Y axis represents the features. The intensity represents the feature importance score.

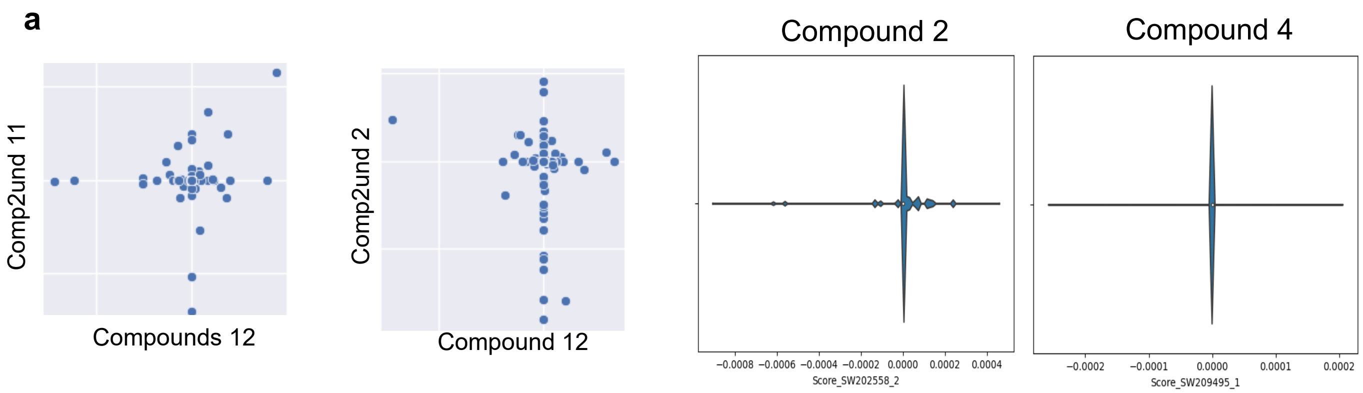

Figure 5: Knowledge generation from the study. (a) Correlation between 'non-specific' (Compound 11) vs 'non-specific' (Compound 12) and 'specific (Compound 2)' vs 'non-specific' (Compound 12) features using pairs plot. (b) Violin plot distribution of specific (Compound 2) and non-specific (Compound 4).

6.0 Future directions:

A range of improvements for the algorithm exist, such as bigger dataset with validation/ gold standard would be the obvious next step. With the validation set, a range of different accuracy metric can be checked (such as ROC, precision, recall etc.). Furthermore, dividing the cell lines based on their mutational status (such as oncogenic mutations KRAS, EGFR, BRAF etc.) would be way to classify the bigger cohorts. A range of machine learning based classifications (SVM, random forest, decision tree, neural network) can then be applied on the data. A hybrid model would also be something to look at.

Additionally, even then, incorporating SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) to gain insights into the contribution of each feature/regressor towards the prediction of biomarkers would be interesting.