11-18-2018

Classifying Aneugens Using Functional Data Analysis Techniques

Daniel Gremmell
dg3778@rit.edu

RIT

Classifying Aneugens Using Functional Data Analysis Techniques

by

Daniel Gremmell

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Applied Statistics

School of Mathematical Sciences

College of Science

Thesis Committee: Peter Bajorski Minh Pham Stephen Dertinger

Rochester Institute of Technology

Rochester, NY

November 18, 2018

Abstract

In the field of genetic toxicology the term aneugen is used to indicate chemical or physical agents that cause chromosomes to malsegregate during division, thereby resulting in altered DNA content in daughter cells. This form of chromosome damage can be detected in certain mammalian cell-based assays, however the molecular mechanism(s) responsible for aneugenic effects are not apparent from these conventional tests. However, the responsible molecular initiating event (MIE) is of interest to pharmaceutical, chemical, and agro-chemical industries, because this knowledge can assist their efforts to design out such liabilities and/or avoid similar chemical structures altogether. This study evaluated the ability of several experimental biomarkers to identify the MIE of aneugens from the functional curves that originate from human TK6 cells exposed to fluorescent Taxol (Taxol 488) for four hours and co-treated with known aneugens over a range of concentrations.

A large functional space of classifiers were evaluated using two stages of cross validation. First, a wide space was searched using a variety of depth, area under the curve (AUC) summarized and kernel methods to identify the top performing models. The top models were then evaluated in a second stage of cross-validation to establish a mean error rate and log loss that approached their theoretical distributions.

In searching the large space of non-parametric and functional classifiers, it was found that a K-Nearest Neighbors Model (KNN) using a single neighbor on an H-Modal Depth calculation of the functional curve could properly classify MIE of aneugens with cross-validated error rates close to zero and well below other methods such as AUC summary methods and other depth based methods. Similar to the KNN model, a Kernel Support Vector Machine with an ANOVAdot kernel could classify aneugens from the raw functional curve data not requiring a depth based calculation.

While those models are best, they have the benefit of having more data observations in the form of replicate data. If the data are summarized to remove replicates, the linear discriminant analysis model with AUC summarized data is the best model.

This study shows that it is possible to use the raw functional curves from an experiment of aneugens to identify their MIE to an accurate degree using machine learning methods.

Acknowledgements

Thanks to Peter Bajorski, PhD, Minh Pham Stephen Dertinger, PhD and Derek Bernacki for his strong support on bench tests and data.

Introduction

In the field of genetic toxicity, aneugens and clastogens are classes of effects on genomes. According to Plant, "Aneugens induce unequal segregation of chromosomes during cell division and results in a cell with more or fewer chromosomes than normal (aneuploidy)," (Plant, 2003). Plant continues to point out that aneugens have a threshold effect and only cause genotoxicity at certain levels, a characteristic that has implications for risk assessments. Aneugenic chemicals can be detected in mammalian cell-based assays, however the molecular mechanism(s) responsible for chromosome malsegregation are not apparent from these conventional tests. Yet the responsible molecular initiating event (MIE) is of interest to pharmaceutical, chemical, and agro-chemical industries, because this knowledge can assist their efforts to design out such liabilities and/or avoid similar chemical structures in the future. This writing report evaluates and establishes the ability of several biomarkers measured in a novel, multiplexed flow cytometric assay classifiers to identify the molecular initiating event (MIE) assay of an aneugens from the functional curves that originate from cultured TK6 cells that have been exposed to fluorescent Taxol (Taxol 488) for four hours and co-treated with known aneugens over a range of concentrations.

The range of concentrations from these experiments create a functional curve that can result in different classes of response. These classes can be discriminated using the raw functional data curves and present a secondary level class of response for aneugens by:

1. Comparing functional data classifiers using statistical depth methods.
2. Evaluating area under the curve transformations and classifier strength.
3. Searching for the best generalized model between the depth, AUC and raw functional curve methods.

Methods

Functional Depth Methods

The concept of statistical depth was developed by Tukey as a way to visualize data. This concept was introduced with univariate data and later extended to multivariate analysis and is widely used in outlier detection and functional classification. In multivariate analysis, the halfspace depth is defined as:

$HD_n(x_i) = min(F_n(x_i), 1 - F_n(x_i))$

Where $x_i ... x_n$ is a sample drawn with a cumulative distribution of F (Febrero-Bande, Galeano, Gonzãlez-Manteiga 2008).

This depth function essentially looks for the minimum depth in a cloud of points by looking to the left and right of the point in the space. The measure essentially works as an indicator of centrality with the central points having a higher depth than the points on the edges. This concept drives the derivation of other forms of depth. The depth function is key to the classification methods assessed in this writing when classifying a sample of curves related to aneugens.

H-Modal Depth

The modal depth is used to calculate depth based on how close the curve is to the cluster of curves around it.

H-modal depth is defined as:

$HM(x_i, h) = \sum_{i=1}^{n} \frac{K(||x_i - x_k||)}{h}$

Where K represents a Gaussian Kernel and h represents a bandwidth matrix. The Gaussian kernel is found to be:

$K(u) = \frac{1}{\sqrt{2\pi}} exp(-\frac{u^2}{2}), u > 0$

The bandwidth matrix is the 15% quantile of the calculated distance in the functional data, which is the L2 norm in the modal depth function $||x_i - x_k||$. The bandwidth matrix does not impact the modal depth calculations since the points in the center are not sensitive to the choice of the bandwidth matrix.

Before illustrating H-modal depth, a plot of the curves can be seen in Figure 1. The position of the curve (1-20) is seen on the X-axis and the Y-axis is the fold 488-Taxol variable.

An example of H-modal depth can be illustrated using the 488-Taxol curve as seen in Figure 2. The first curve is colored to distinguish it from the others. This curves peaks above a value of 9 at the third observation. This curve would be considered to be at a lower mode depth than the curves clustered around the main group near the bottom of the plot (Febrero-Bande, Galeano, Gonzãlez-Manteiga 2008).

The mode depth in Figure 2 is found to be 0.3989423. This is compared to a curve closer to the grouping of other curves illustrated in Figure 3 where the depth being measured is in red.

The second chemical in Figure 3 has a mode depth of 5.5913002.

Since it occurs with a smaller distance to the rest of the curves, it receives a higher depth measure than the curve that lies outside of the clustered group of curves.

Calculating depth using this method for 488 Taxol and PH3-KI67 Ratio gives a plot of points as seen in Figure 4 for each curve instead of a group of functional curves. The variable 488 Taxol is on the x-axis

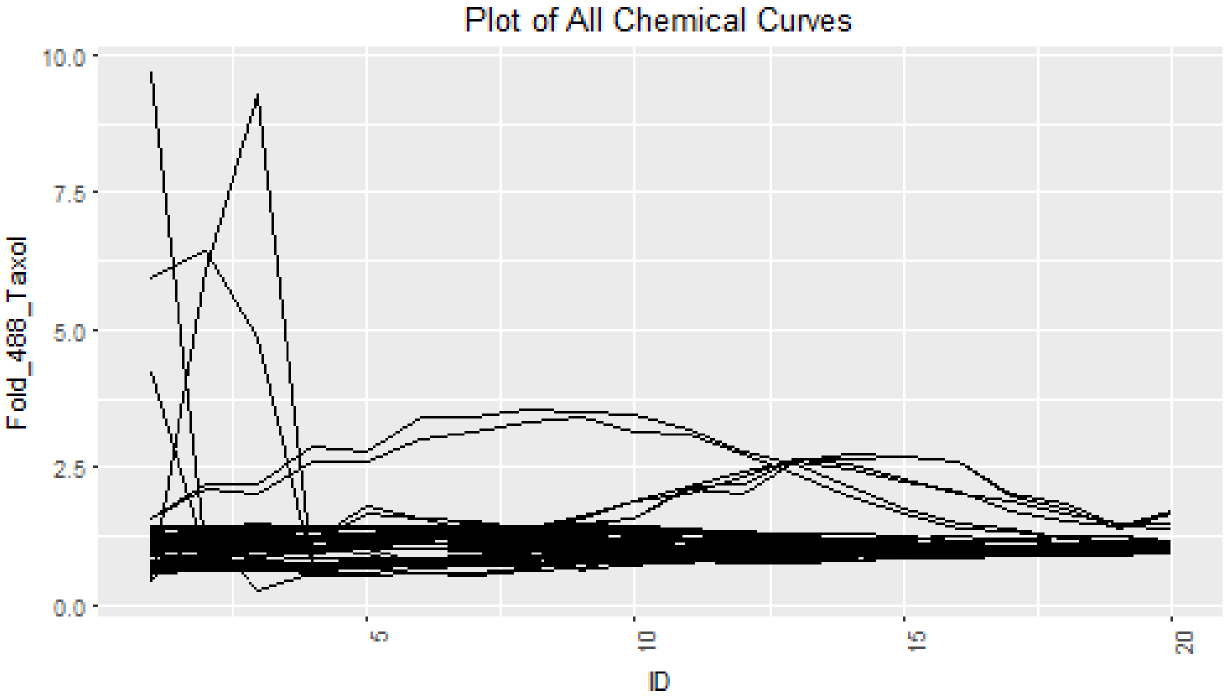

Figure 1: Plot of all chemical curves.

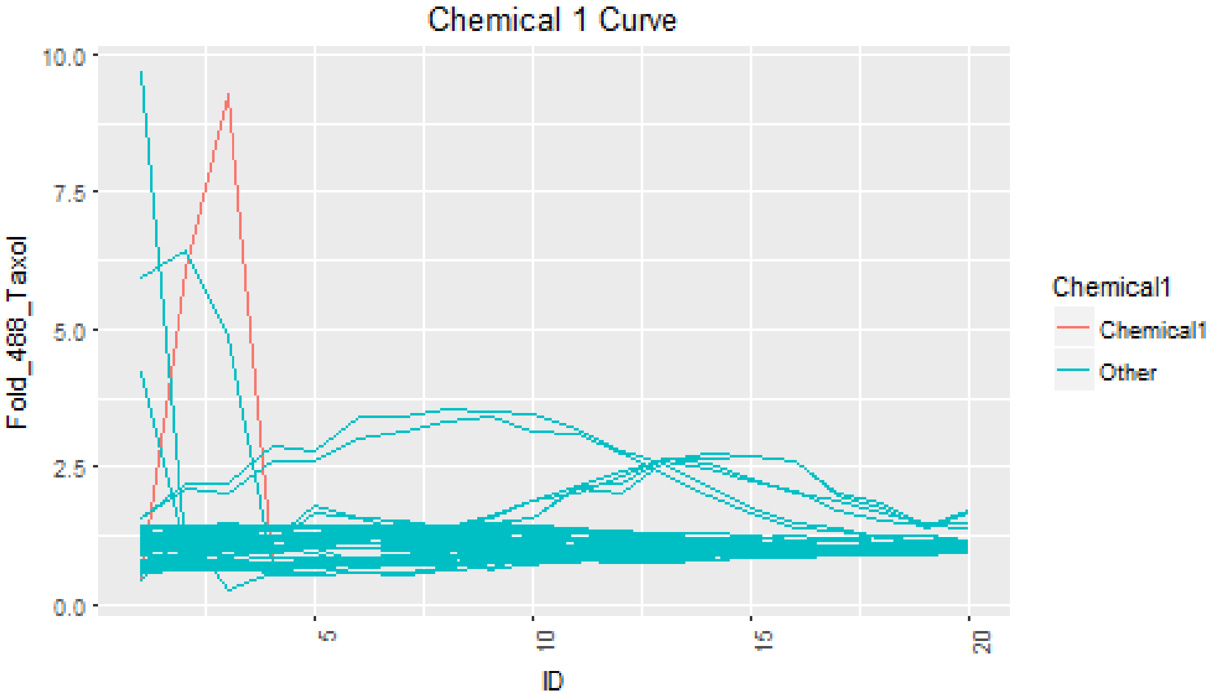

Figure 2: Chemical 1 in red differs from other curves example.

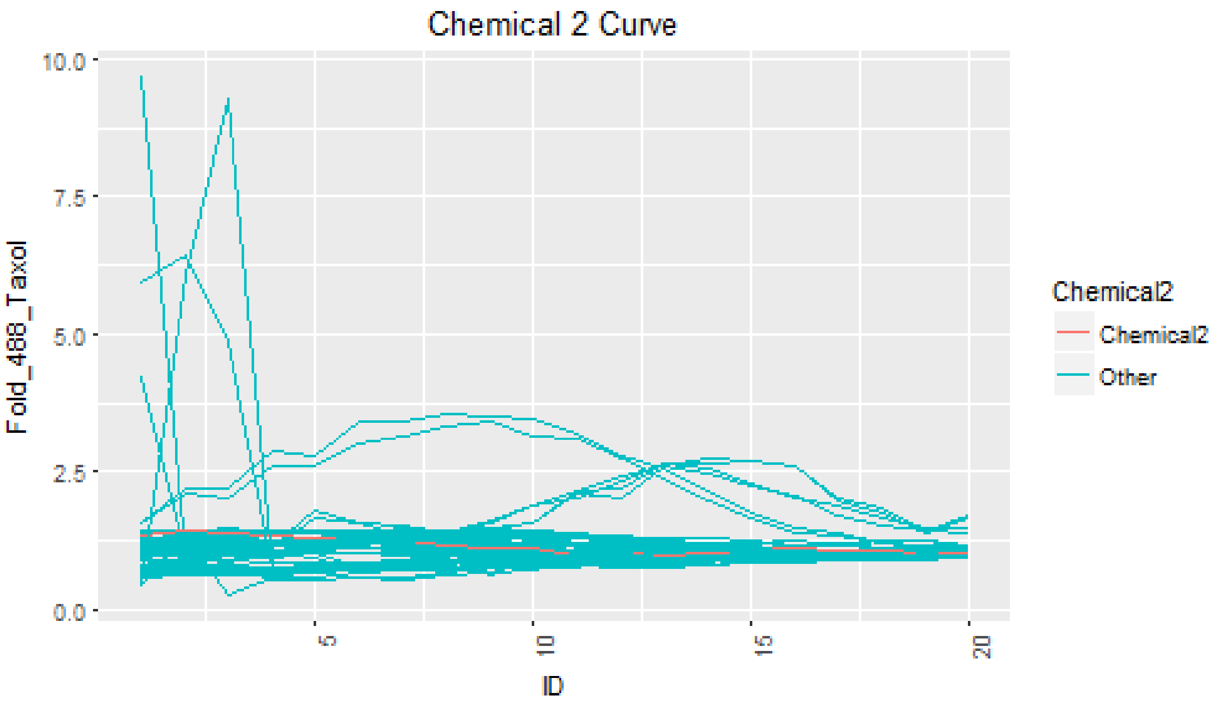

Figure 3: Chemical curve in red closer to cluster of other curves.

and the PH3-KI67 ratio variable is on the y-axis. The measurements reflect a single curve's mode depth measurement.

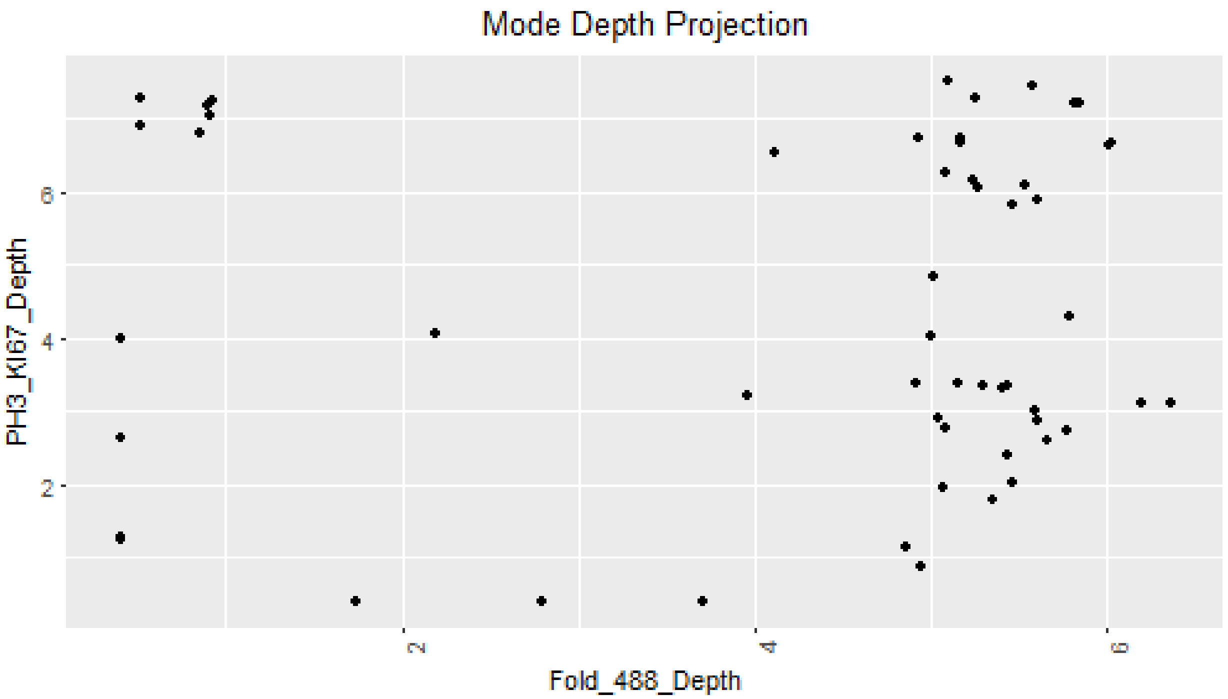

Figure 4: H-modal depth of each curve plotted in 2D space.

Fraiman and Muniz Depth

Fraiman and Muniz Depth is also known as integrated depth. It takes into account the area under the depth curve to produce an average measure of depth based on a curve.

If we assume x_i is a function at time t, then Fraiman-Muniz depth is found by:

$FM(x_i) = \int_a^b 1 - |\frac{1}{2} - F_{n,t}(x_i(t))|$

Where F represents the empirical cumulative distribution at time t where a represents the minimum of t period and b represents the maximum of t period and n represents the number of iid curves, (Flores, Lillo and Romo, 2015). Essentially, the estimate of the depth is the mean of the curves returned by the function $1 - |\frac{1}{2} - F_{n,t}(x_i(t))|$. This is represented by the sample function which is a weighted sum depths. This is shown by Flores, Lillo and Romo as:

$SFM(x_i) = \sum_{j=2}^{m} \Delta_j [1 - |\frac{1}{2} - F_{n,t_j}(x_i(t_j))|]$

Where $t \in [0,1]$ and $\Delta_j = (t_j - t_{j-1})$

This provides the equivalent weight of each point of the curve resulting in:

$SFM(x_i) = \frac{1}{m} \sum_{j=2}^{m} [1 - |\frac{1}{2} - F_{n,t_j}(x_i(t_j))|]$

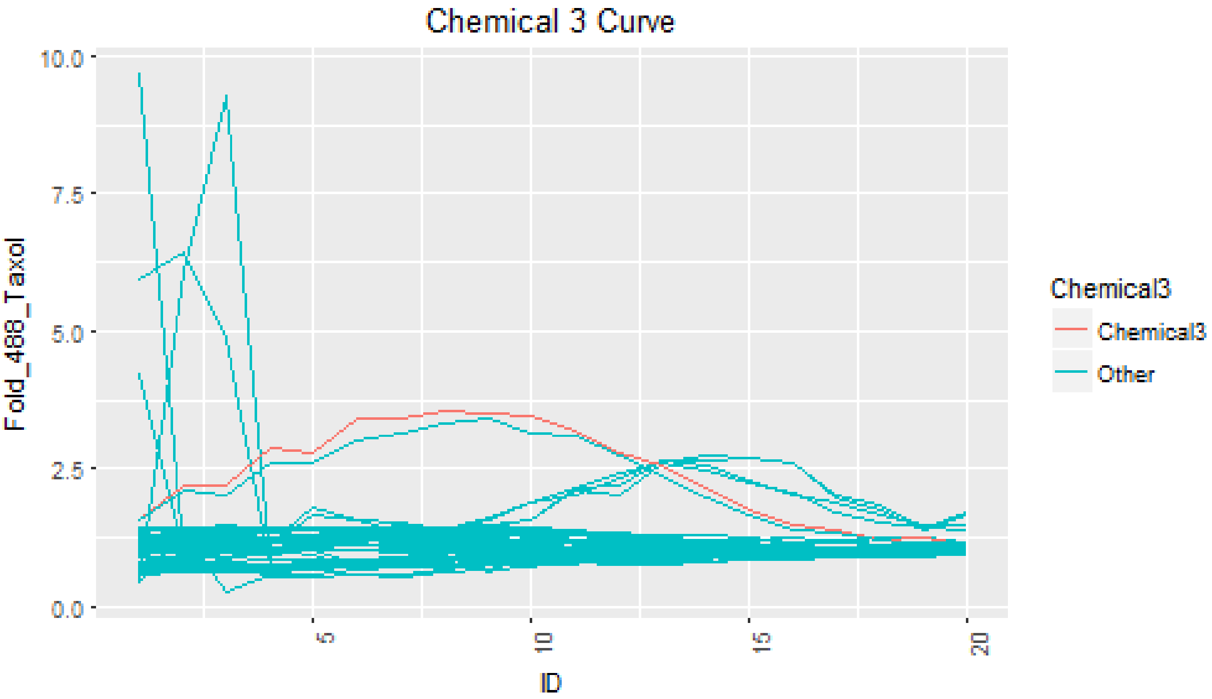

Figure 5: Curve of FM Depth chemical measured in red.

In the case of Fraiman and Muniz Depth, the behavior this drives is a lower depth measure for curves that operate in a higher space of the response and a higher depth for curves with measures closer to zero. However, it takes into account the length of the curve and is less extreme for outliers. For example, the curve colored in the plot in Figure 5 again plotted using the 488-Taxol variable operates in a higher response space than the rest of the curves and therefore has a lower depth measure.

The depth measure for the red curve in Figure 5 is 0.0851852. The cluster of curves towards the bottom of the plot ranges on the higher side of the 0-1 space the depth measure provides.

Calculating depth using this method for 488 Taxol and PH3-KI67 Ratio gives a plot of points as seen in Figure 6 with fold 488-Taxol depth on the x-axis and the y-axis being Ph3-KI67 ratio depth. This plot is noticeably different than the H-Modal plot.

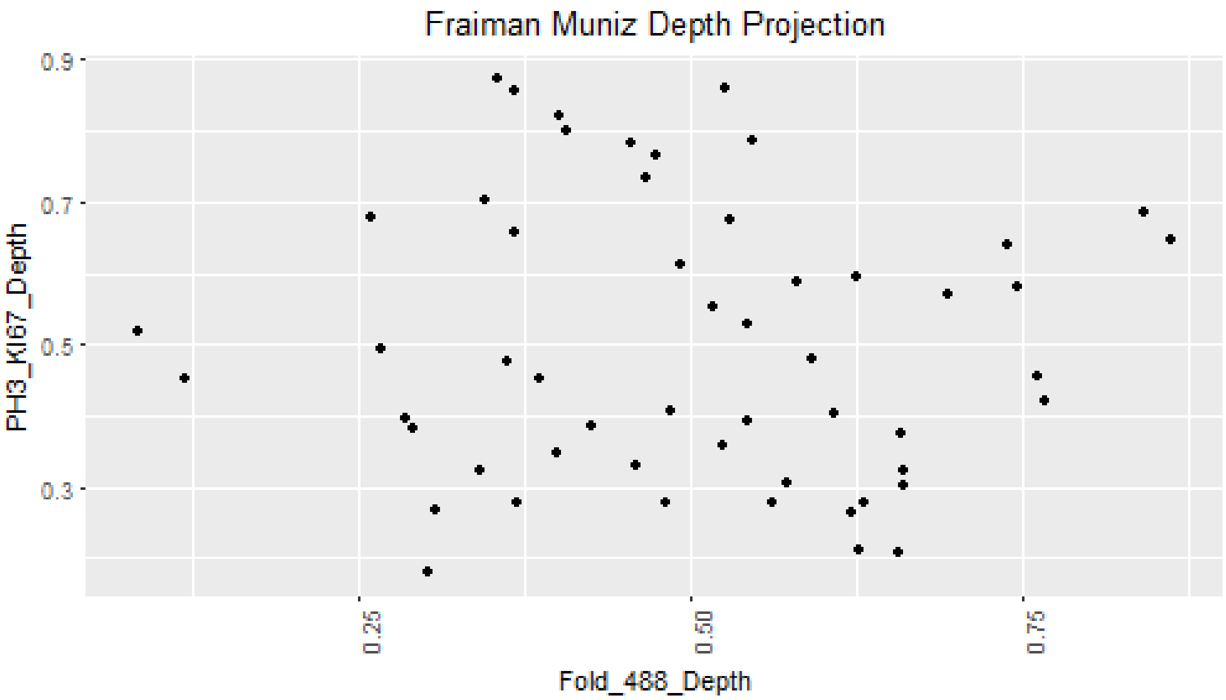

Figure 6: Plot of Fraimen Muniz Depth in 2D space.

Random Projection Depth

The random projection depth uses projections to measure the data depth. A number of depths is taken and the mean of those depths provides the depth estimate for the curve. Therefore, the random project depth process can be described as:

1. Project out a vector generated by a stochastic process whose values are random directions.
2. Find the inner product between the projection vector and the data. This returns a point.
3. This is then passed through a depth function and the points updated.
4. This is retained and the process is repeated N times.
5. Calculate the mean of the projections points to return the depths.

There are two ways of doing this. First, using only the univariate points and second using the first derivative to project a point in space (Flores, Lillo and Romo, 2015).

The random points are projected using a random Gaussian process defined by:

$\nu(t) = \sigma^2 exp(-\frac{|d|}{\theta})$

Where $\sigma^2 = 1$ and $\theta = 0.2(t_m - t_1)$ and $\mu = 0$ and $d = t - t_i$

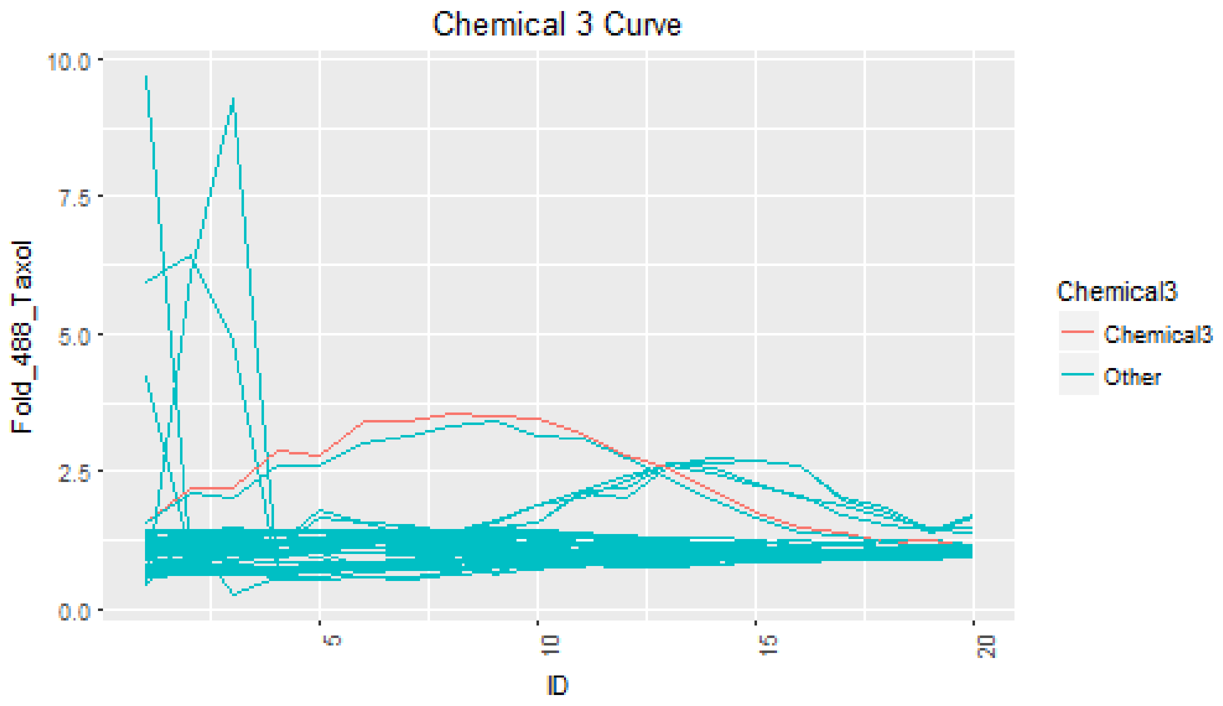

Figure 7: Plot of depth chemical being measured in red.

The univariate vector ν is found by:

$PJ_{i,\nu} = \int_0^1 \nu(t)x_i(t)$

Next, the depth method used to assess the projected vector is half depth:

$HD_n(x_i) = min(F_n(x_i), 1 - F_n(x_i))$

This leads to the estimate of depth which is found by:

$RP(x_i) = \frac{1}{P}\sum_{i=1}^{P} HD(PJ_{i,\nu}, PJ_{i,\nu})$

Using the derivative method adds projected point:

$PJ'_{i,\nu} = \int_0^1 \nu(t)x'_i(t)$

Which then updates the depth estimate function to be:

$RP(x_i) = \frac{1}{P}\sum_{i=1}^{P} HD(PJ_{i,\nu}, PJ'_{i,\nu})$

The curve of chemical three representing the depth being measured is shown in Figure 7.

The depth of this curve in Figure 7 comes out to be 0.05. Using a single derivative, the depth is 0.0895881

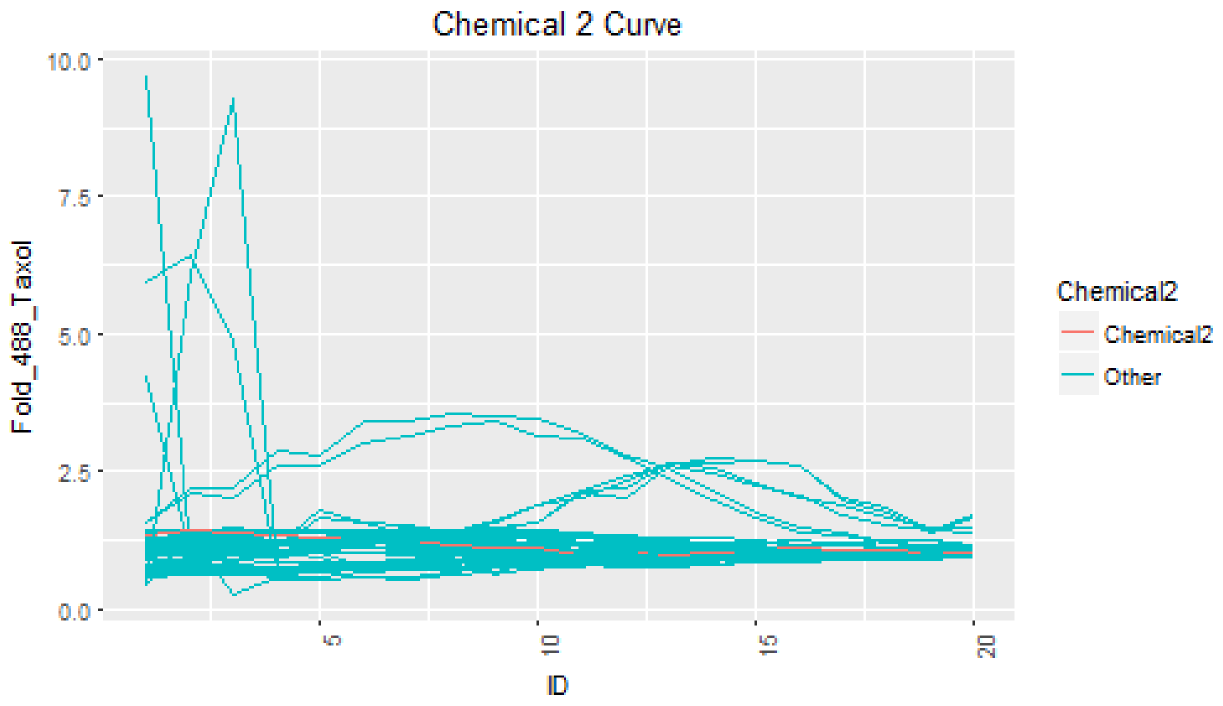

Figure 8: Plot of random projection depth chemical being measured in red.

Comparing this to a plot of another curve colored in red in Figure 8 which is heavily clustered with other curves.

The depth of the curve in Figure 8 is 0.2981481. Using a single derivative, the depth is 0.6803665

Thus, calculating the depth of 488 Taxol and PH3 - KI67 ratio gives is plotted in Figure 9, which again shows a different space than the H-Modal and Fraimen and Muniz depths.

Then using a single derivative the plot of fold 488-Taxol and PH3-KI67 depth can be shown in Figure 10 as a 2D point plot with one point representing a single chemical curve.

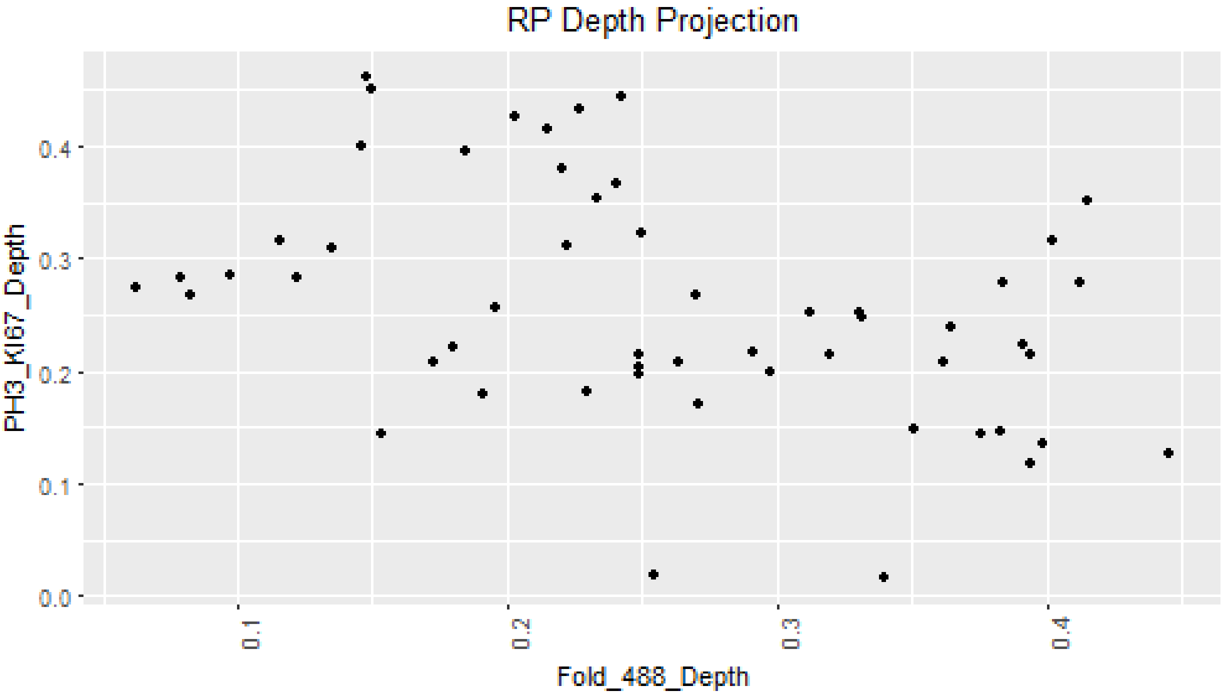

Figure 9: Plot of random projection depth in 2D space.

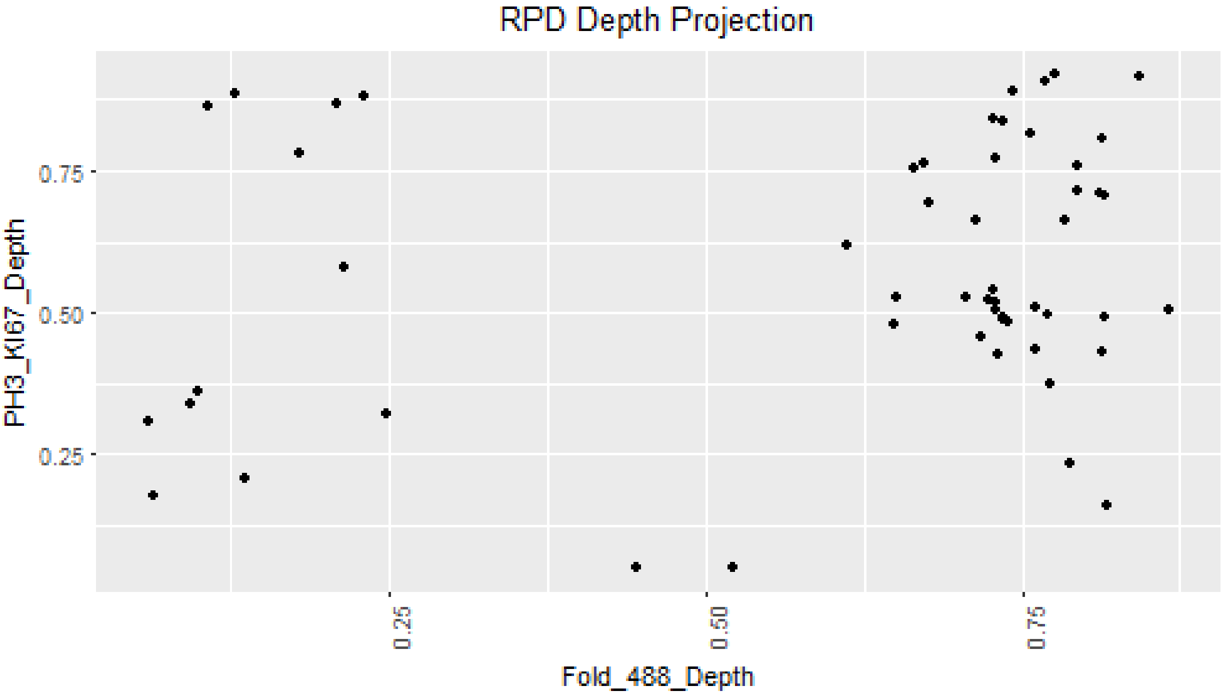

Figure 10: Plot of random projection with single derivative depth in 2D space.

Loss Functions

The loss function being used is the so called 0-1 loss function found by:

$l(Y, f(X)) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{Y_i \neq f(X_i)\}}$

Therefore, the best function would be the one that minimizes the 0-1 loss function:

$\hat{f} = argmin_{f \in F}(\frac{1}{n} \sum_{i=1}^{n} 1_{\{Y_i \neq f(X_i)})$

In addition to the 0-1 loss function, learning methods will also be evaluated by the multiple log loss function also known as multiclass cross entropy. It is found by:

$-\sum_{i=1}^{M} 1_{\{Y_i = \hat{Y}_i | X_i\}} log(Pr[Y_i = \hat{Y}_i | X_i])$

Multiclass cross entropy allows multiple functions to be assessed that may produce the same 0-1 loss rates. Multiclass cross entropy is a useful measure for helping understand the confidence at which the classification is being made by using the probability estimates.

Model Evaluation Process

Therefore the process for finding the best function by minimizing the loss function is:

1. Fit function to training data.
2. Measure training data loss using 0-1 loss.
3. Fit 10 rounds of 10 fold cross validation.
4. Measure mean 0-1 loss for each fold.
5. Calculate the mean loss across all rounds.
6. Identify the function with the smallest loss across all 10 rounds.

This process allows an estimate of error to be established.

Models Evaluated

The Unsupervised Methods assessed were:

1. K-Means Clustering
2. Hierarchical Clustering

The functional data functions assessed were:

Depth Methods:

1. Linear Discriminant Analysis
2. Decision Tree
3. Kernel Methods
4. K Nearest Neighbors

The base methods baselined:

1. Linear Discriminant Analysis
2. Decision Trees
3. K Nearest Neighbors
4. Kernel Support Vector Machines

Data and Classes

Aneugen MIE Assay

A mammalian cell culture assay for determining genotoxic mode of action is commercially available from Litron Laboratories (Rochester, NY) under the name MultiFlow® (Bryce et al., 2016). Based on several DNA damage response biomarkers, this multiplexed assay characterizes compounds as clastogenic, aneugenic or nongenotoxic. The results presented herein are based on modifications to the assay that extend its capabilities to include the determination of aneugenic agents' molecular initiating events (MIE). Specifically, scientists at Litron have designed a follow-up assay with the goal of discriminating between the two most common mechanisms of aneugenicity: tubulin binding and off-target inhibition of mitotic kinases, especially Aurora kinase B. For these experiments, TK6 cells were exposed to a group of 27 known aneugens. Each chemical test was replicated twice starting with overcytoxic concentrations and decreasing into concentrations that are not overcytoxic and in the presence of 488 Taxol. The aneugens consisted of 15 diverse tubulin binders and 12 mitotic kinase inhibitors. After 4 hrs at $37°C$ cells were added to MultiFlow base solution (detergent, nucleic acid dye, RNase and counting beads) supplemented with fluorescent antibodies against phospho-histone H3 (p-H3) and Ki-67. The resulting detergent-liberated nuclei were then analyzed via flow cytometry. Key measurements included nuclei's Taxol 488-associated fluorescence, frequency of phospho-histone H3 (p-H3) positive events, and frequency of Ki-67-positive events. These metrics were each converted to fold-change relative to concurrent mean solvent control values. Table 1, below, describes the various components used to prepare cells for analysis, and provides rationale explaining why these particular candidate biomarkers were chosen for investigation.

Data Characteristics

The data consists of a group of 27 different chemicals. Each chemical test was replicated twice starting with overcytoxic concentrations and decreasing into concentrations that are not overcytoxic. The concentration itself is a uniformly distributed space with the process being measured at a constant rate of decline in the concentration.

There 488 Taxol represents a treatment of cells for a 4 hour period while pH3 and KI67 represent reagents in the form of detergent solutions. The Fold of taxol 488 and pH3 and KI67 variables represent a transformation of the agent measurements.

The classes being used as a baseline prediction response represent a category of second level attributes related to aneugen chemicals. These attributes represent molecular initialization events (MLE). The four main classes

Table 1: Assay Reagents, Roles*

Reagent	Biology	Intended.Role
Proprietary non-ionic detergent solution	Gently strips away outer (cytoplastic) membranes, produces a suspension of nuclei and aggregates of metaphase chromosomes	Prepares cells for flow cytometric analysis; provides propidium iodide dye and antibody reagents access to chromatin
Propidium iodide and RNase	Propidium iodine is a fluorescent nuclei acid dye; in conjunction with RNase it becomes a DNA-specific dye	Labels nuclei and aggregates of metaphase chromosomes; allows flow cytometric analyses to focus on these events by distinguishing them from debris and other particles
Counting Beads	NA	A consistent number of fluorescent latex particles in each sample provides a means to calculate relative nuclei density; also serve quality control functions
Taxol 488	Taxol is a drug that tightly binds the beta-tubulin portion of cellular microtubules; Taxol 488 is a fluorescent derivative of Taxol	Treating cells with Taxol 488 leads to fluorescently-labeled microtubules; as microtubule growth and contraction is a highly dynamic process we would expect Taxol 488-treated cells to exhibit perturbed fluorescence when co-incubated with tubulin-binding aneugens
Anti-Ki-67	Ki-67 is a nuclear protein that is associated with cell proliferation; during mitosis most of the protein is found on the surface of chromosomes	When detergent-liberated nuclei and metaphase chromosome aggregates are incubated with fluorescent antibodies against Ki-67 it is cells undergoing metaphase that exhibit the most fluorescence; this provides a means for identify and counting this portion of cells
Anti-phospho-histone H3	Histone H3 is a core histone protein that helps organize the structure of chromatin; serine 10 and 28 residues are only phosphorylated during mitosis; Aurora kinase B is responsible for serine 10 phosphorylation	Fluorescent antibodies against phospho-histone H3 (p-H3) (serine 10) can be used to identify and enumerate metaphase cells; when cells are treated with sufficiently high concentrations of aneugens that act via Aurora kinase B inhibition, we would expect metaphase cells (i.e., identifiable by their Ki-67+ status) to lose their p-H3+ signal

Note:
*Information provided by Dr. Stephen Dertinger, Litron Laboratories.

are:

1. AKB - Aurora Kinase B Inhibition
2. TBD - Tubulin Destabilizer
3. TBS - Tubulin Stabilizer
4. OTH - Other

The other category is a catch all for an MIE not accounted for in the other three categories.

Tubulin destabilizing agents result in the depolymerization of microtubules by binding to tubulin polymers. Tubulin stabilizing agents results in polymerization of microtubules by binding to tubulin polymer. Despite the differences, both result in cell death by disrupting the mitotic spindle (Fanale, Bronte and Passiglia, 2015). Aurora kinase inhibitors seek to disrupt aurora kinase which is often found in human tumors (Bavetsias and Linardopoulos, 2015).

In order to functionally assess the chemicals, the data can be thought of as a continuous curve consisting of 20 time points for each independent chemical measurement. Since there are 27 chemicals with 2 observations, the data set is thus 54 x 20 observations.

Data Transformations

Certain measurements cannot be recorded due to a division by zero error, in this case those NULL values were replaced with zeroes for the purpose of this analysis.

The data concentrations are different scales across each chemical. Some range from zero to one, while others may range from 0 - 100. In order to visually represent the data, the concentration must be scaled to a 0-1 scale after the log of the chemical is taken. The log data is scaled using:

$\frac{X-min(X)}{max(X)-min(X)}$

Once scaled, the curves can be visualized on the same scale and patterns can be observed. The scaling was done by the chemical and replication.

First, 488 Taxol response plotted against the scaled concentration variable can be plotted as seen in Figure 11.

These plots show some distinct patterns by class. The AKB and OTH classes tend to rise while TBD tends to remain rather flat with some large swings in the overcytoxic observations and TBS tends to rise then fall in the form of a hump.

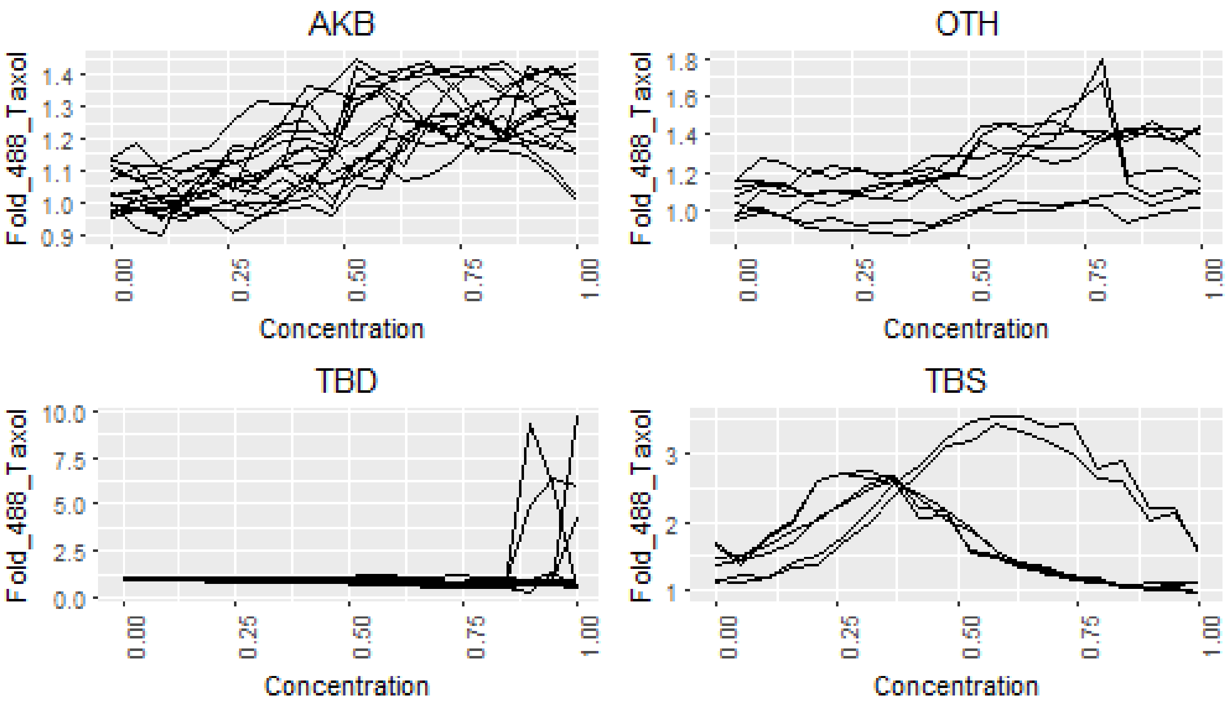

Figure 11: Plot of Fold Taxol 488 y-axis and concentration x-axis by class.

pH3 and KI67 Fold variables can also be plotted in a similar fashion as shown in Figure 12 for PH3 and Figure 13 for KI67.

These plots show similar patterns. AKB and TBD tend to be less defined than with the fold 488 taxol plot. However, OTH and TBS show two distinct patterns. TBS tends to increase then slow its increase while OTH tends to converge as concentration decreases.

Last, the pH3 and KI67 ratio can be plotted as seen in Figure 14.

Each class tends to have a specific pattern with this variable. AKB has a declining pattern while OTH and TBD tend to be flat curves with extreme values at the overcytoxic levels. TBS tends to have an increasing curve when concentration is increased.

These curves illustrate some potential discrimination between different classes of chemicals based on their raw experimental data and replications.

AUC Summarization

The area under the curve summarization method stems from using a definite integral to calculate the area that lies under the continuous curve in a Cartesian space. The values used in this data set were

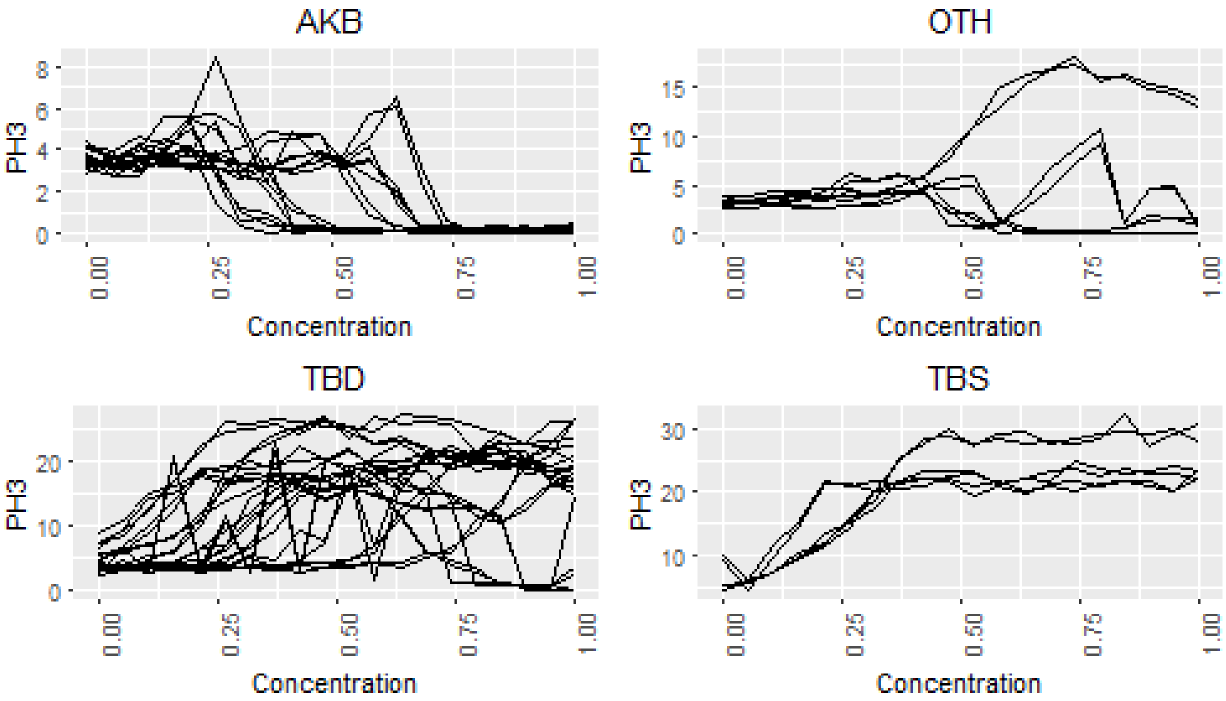

Figure 12: Plot of PH3 y-axis and concentration x-axis by class.

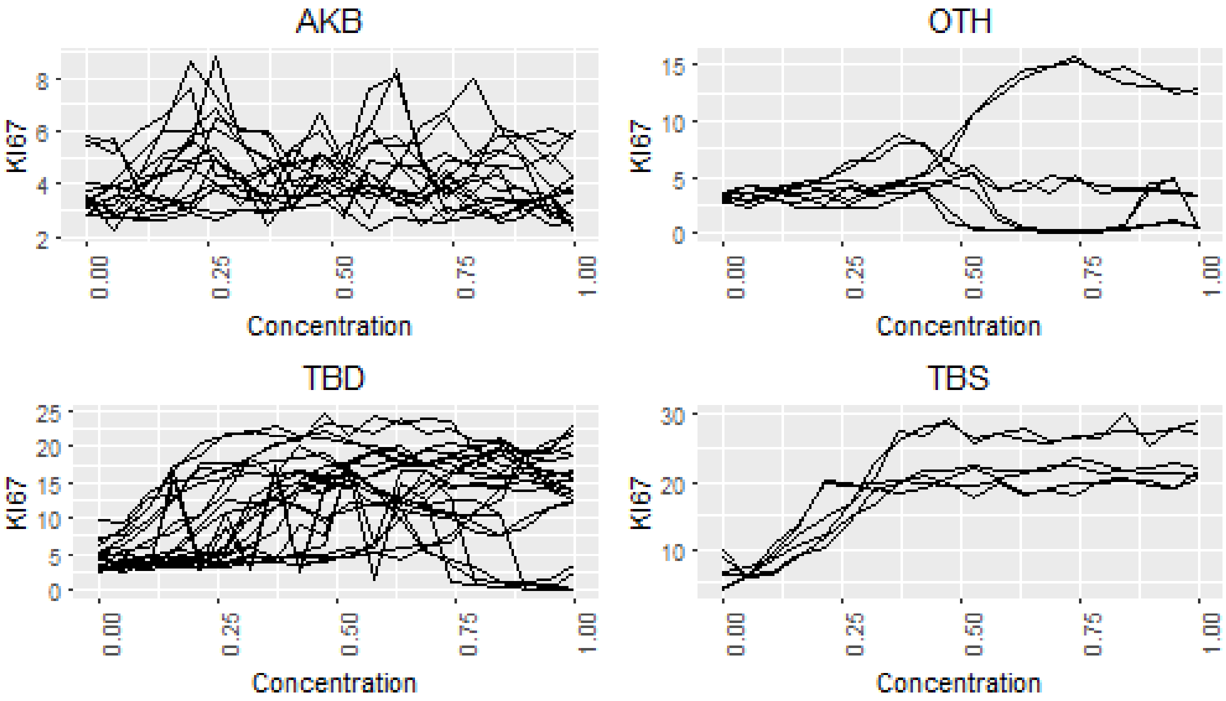

Figure 13: Plot of KI67 y-axis and concentration x-axis by class.

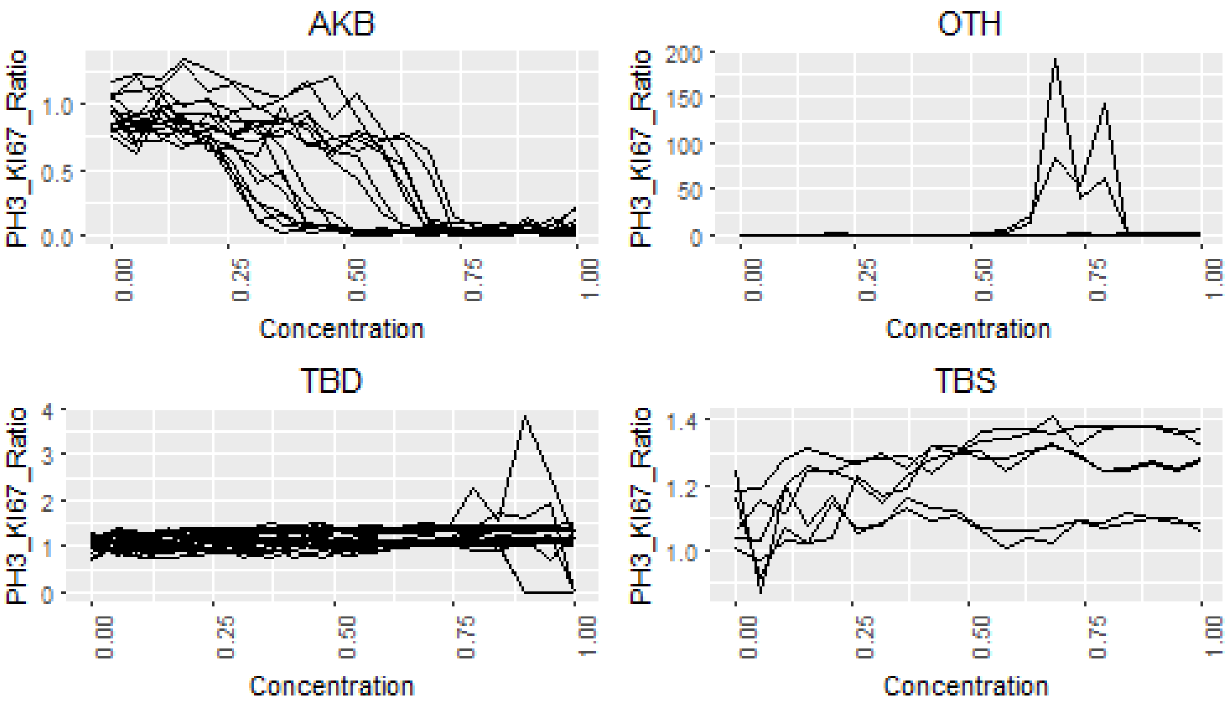

Figure 14: Plot of PH3-KI67 Ratio y-axis and concentration x-axis by class.

calculated in Microsoft Excel which lacks definite integral functionality. The procedure for calculating the AUC summarization data is:

1. Normalize the concentration space to a 0-1 range.
2. Calculate the average values of each biomarker for each chemical. There are two observations at each concentration for each biomarker. The average between these two values provides the input observation for each chemical.
3. One is subtracted from each biomarker.
4. Each curve AUC value is calculated using $\frac{1}{2}(B_i + B_{i+1})(C_{i+1} - C_i)$. Where B represents the biomarker at observation i and C represents the normalized concentration at observation i.
5. At $C_i = 1$, the AUC value is zero.
6. AUC is then computed by summing the calculated values found by $\sum_{i=1}^{n-1} \frac{1}{2}(B_i + B_{i+1})(C_{i+1} - C_i)$.

The curve must be ordered from concentrations 0-1. This process is repeated for each biomarker.

Results

Unsupervised Learning Methods

In the process of assessing methods to classify the different chemical aneugens, unsupervised models can be assessed to determine if there are any key learnings and patterns from using distance based metrics. This can be done with different forms of the data either as raw curves, AUC summarized or depth based methods. The result of each model will be the optimal number of centers, if relevant and then a table demonstrating the groups each observation is assigned to compared to the actual class. In the case of the table, four clusters are always used since there are four classes. Clustering will only be done for Taxol 488 and the pH3-KI67 ratio for the functional curves.

K-Means Functional Curves

For K-means assessment, first the elbow plot of the cluster sum of squares is created, followed by a plot illustrating the relevant Bayesian Information Criterion (BIC) as a way to objectively set the optimal cluster. The optimal cluster according to the BIC is the one that minimizes the BIC. In this case the Bayesian Information Criterion is defined as:

$BIC(M) = nlog(\frac{WCSS}{n}) + log(n)(d+1)$

Where n is the number of observations, $WCSS$ is the within cluster sum of squares error and d represents the number of centers. An elbow plot of numerous K-means fits can be seen in Figure 15 with the x-axis being the number of clusters used. The K-means fit using the BIC criterion is shown in Figure 16 with the x-axis being the number of clusters used. Last, Table 2 shows the table of clusters for each curve in the dataset along the column markers and the actual classes in the row names.

Table 2: Table of cluster groupings against actual classes for the 488 Taxol variable.

	1	**2**	**3**	**4**
AKB	0	0	16	0
OTH	0	0	6	2
TBD	2	0	2	20
TBS	0	6	0	0

The same process can be repeated for the pH3-KI67 variable. An elbow plot of numerous K-means fits can be seen in Figure 17 with the x-axis being the number of clusters used. The K-means fit using the BIC criterion is shown in Figure 18 with the x-axis being the number of clusters used. Last, Table 3 shows the table of clusters for each curve in the dataset along the column markers and the actual classes in the row names.

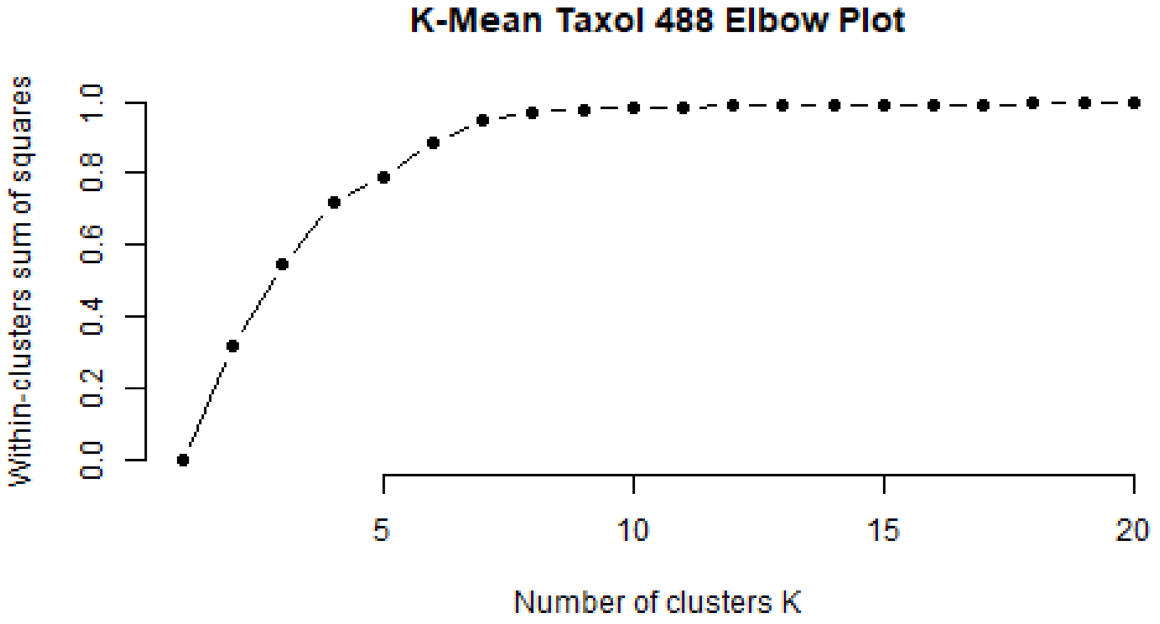

Figure 15: Elbow Plot of K-Means fit for the 488 Taxol variable.

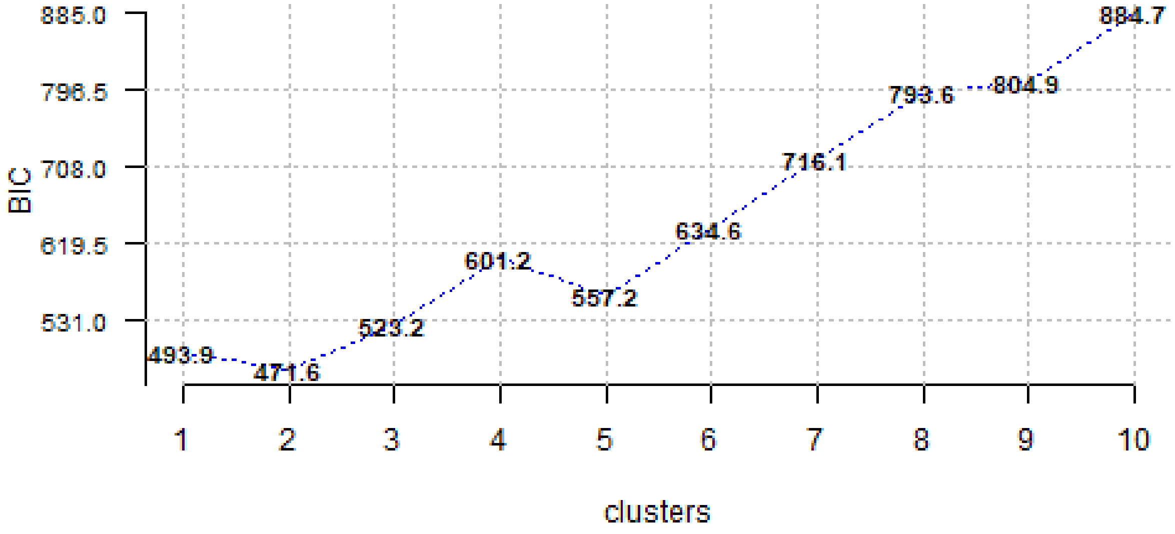

Figure 16: Plot of BIC criterion for K-Means fit for the 488 Taxol variable.

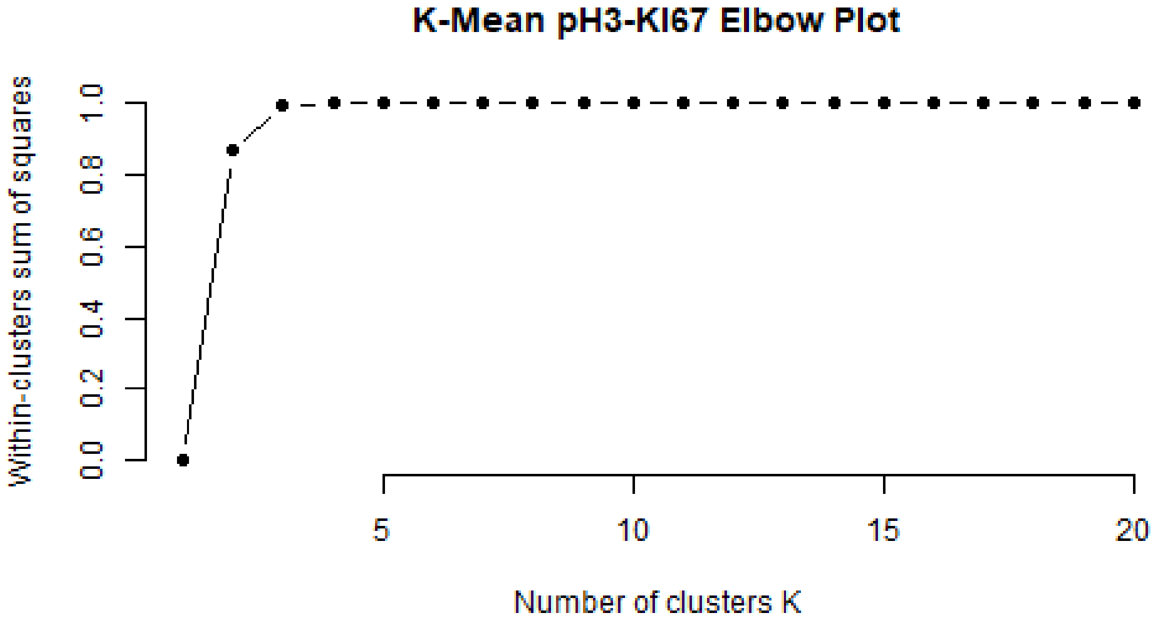

Figure 17: Elbow Plot of K-Means fit for PH3-KI67 ratio variable.

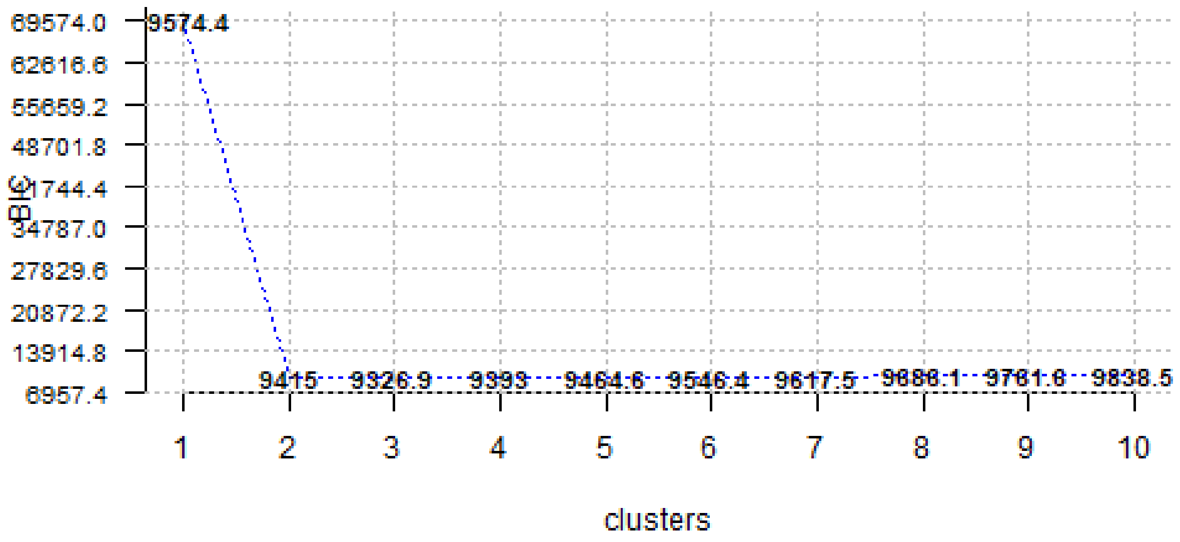

Figure 18: Plot of BIC criterion for K-Means fit for the PH3-KI67 variable.

Table 3: Table of cluster groupings against actual classes for the PH3-KI67 variable.

	1	**2**	**3**	**4**
AKB	0	16	0	0
OTH	1	2	4	1
TBD	0	0	24	0
TBS	0	0	6	0

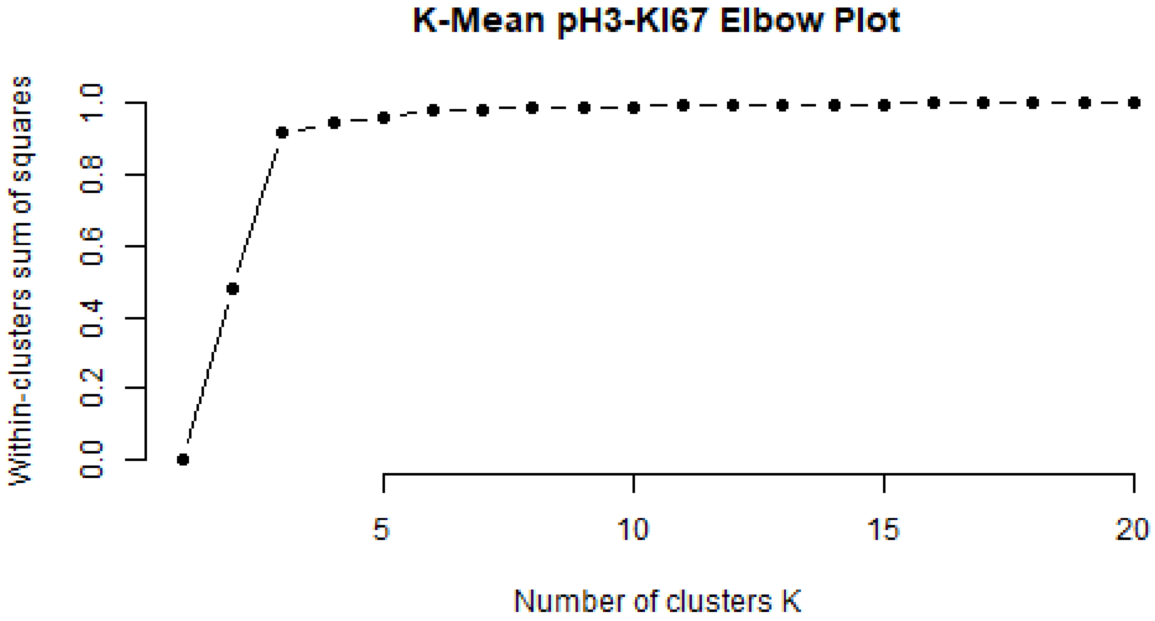

Figure 19: Elbow Plot of K-Means fit for AUC Summary.

K-Means AUC Summary Data

All variables were used in this assessment of K-means since there is a summary point for each variable and each chemical. The same plots and tables are produced to illustrate the results of the K-Means experiment. An elbow plot of numerous K-means fits can be seen in Figure 19 with the x-axis being the number of clusters used. The K-means fit using the BIC criterion is shown in Figure 20 with the x-axis being the number of clusters used. Last, Table 4 shows the table of clusters for each curve in the dataset along the column markers and the actual classes in the row names.

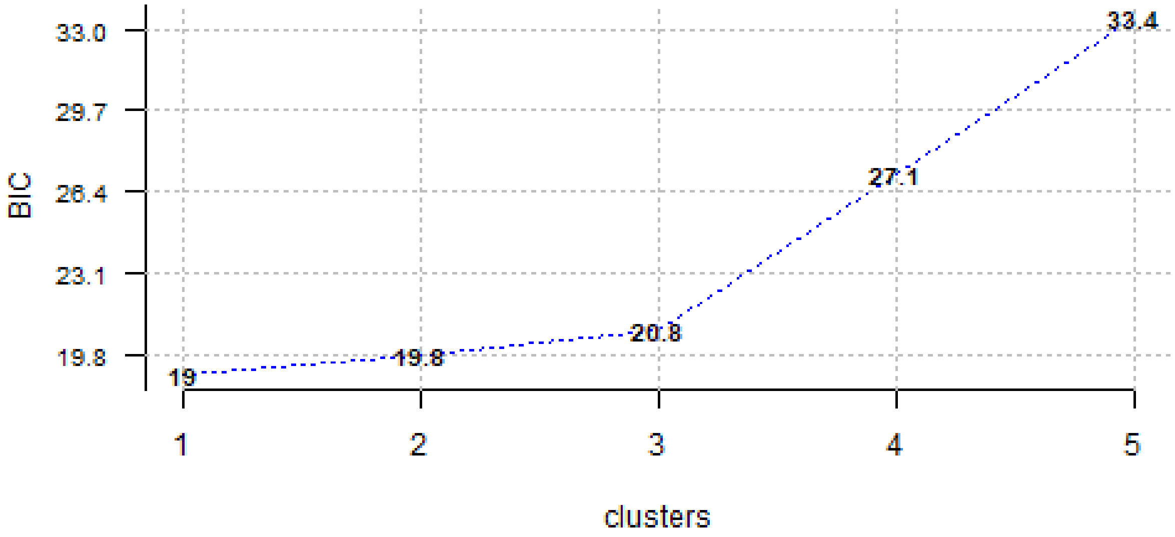

Figure 20: Plot of BIC criterion for K-Means fit for the AUC Summary.

Table 4: Table of cluster groupings against actual classes for the AUC Summary.

	1	**2**	**3**	**4**
AKB	5	0	0	3
OTH	0	3	0	1
TBD	0	12	0	0
TBS	0	0	3	0

K-Means Depth Measures

All variables are used in the K-Means Assessment of Depth measures. The results can be shown in the form of graphs similar to the AUC summary data.

H-Modal Depth

An elbow plot of numerous K-means fits can be seen in Figure 21 with the x-axis being the number of clusters used. The K-means fit using the BIC criterion is shown in Figure 22 with the x-axis being the number of clusters used. Last, Table 5 shows the table of clusters for each curve in the dataset along the column markers and the actual classes in the row names.

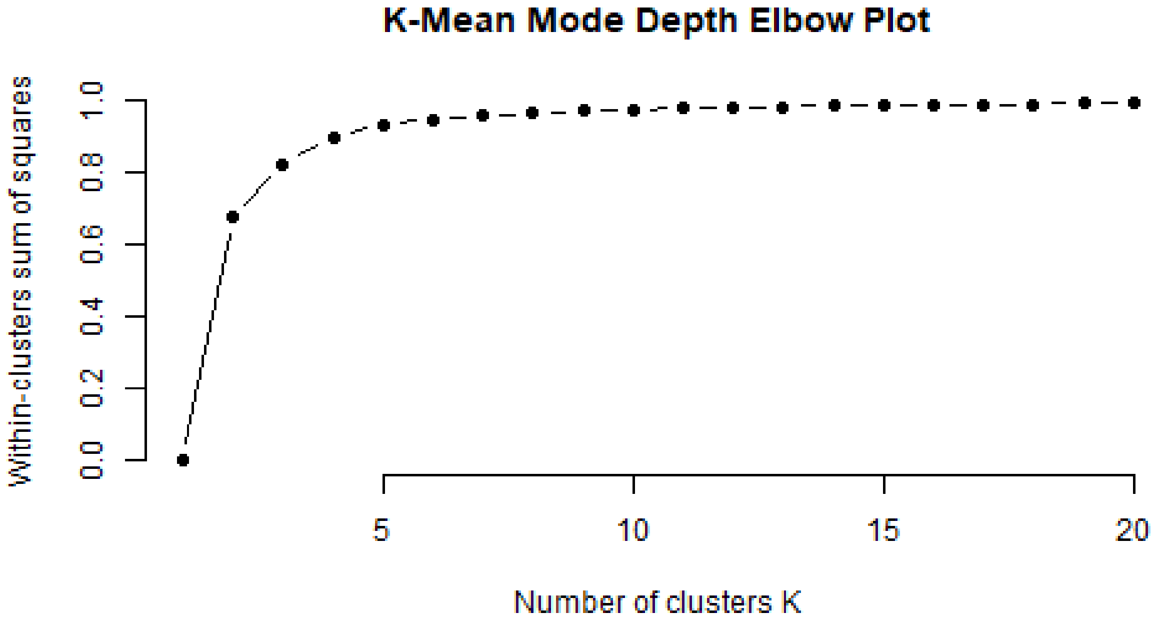

Figure 21: Elbow Plot of K-Means fit for H-Modal Depth.

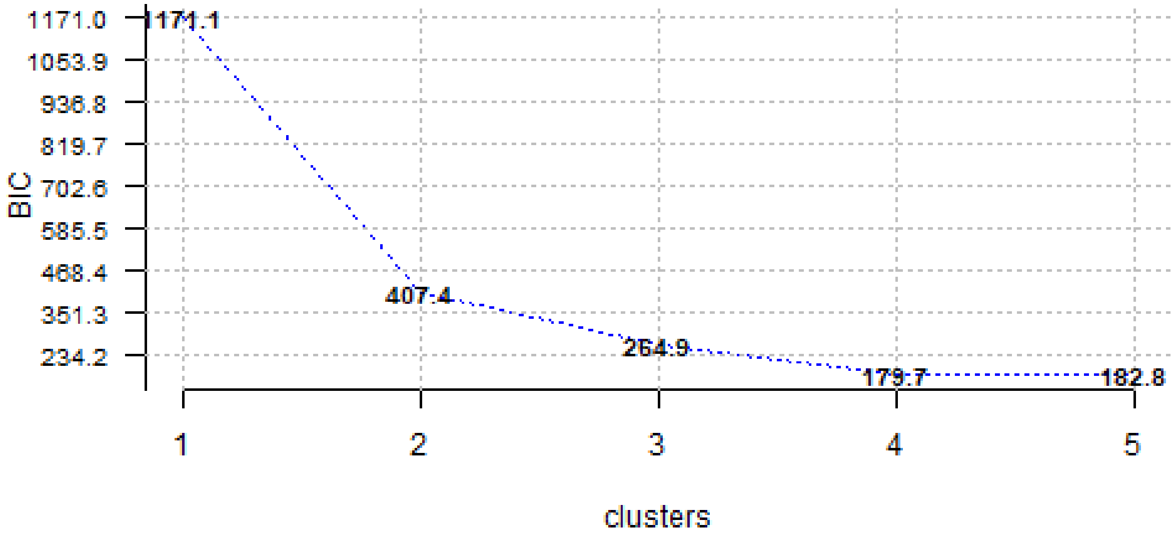

Figure 22: Plot of BIC criterion for K-Means fit for the H-Modal Depth.

Table 5: Table of cluster groupings against actual classes for the H-Modal Depth.

	1	**2**	**3**	**4**
AKB	16	0	0	0
OTH	4	2	2	0
TBD	0	18	4	2
TBS	0	0	0	6

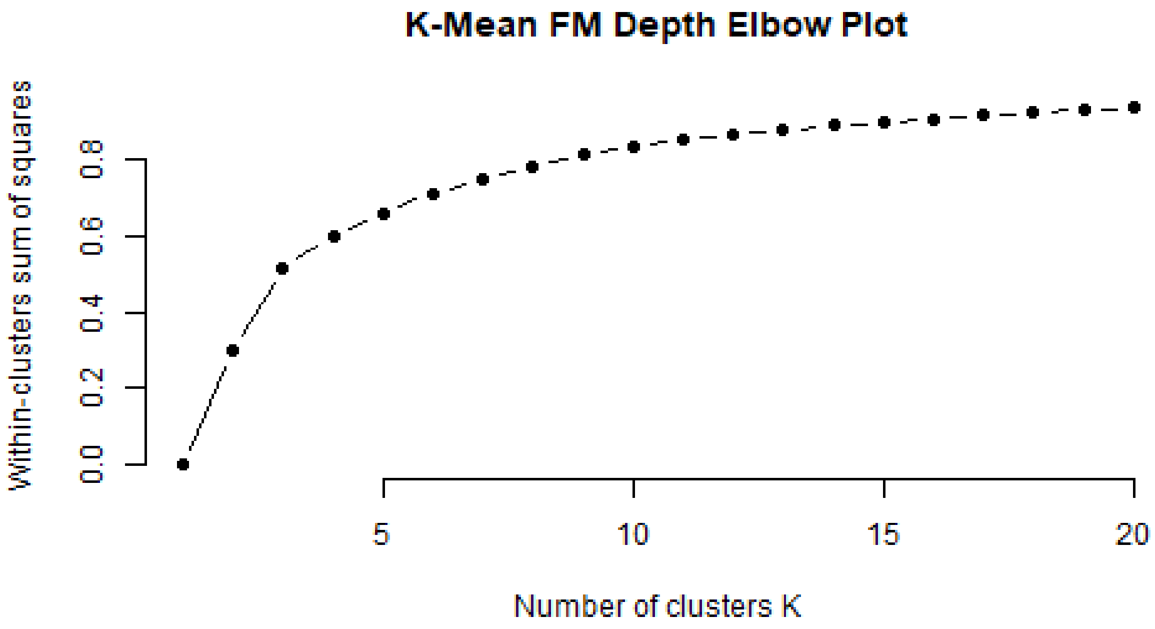

Figure 23: Elbow Plot of K-Means fit for Fraimen and Muniz Depth.

Fraimen and Muniz Depth

An elbow plot of numerous K-means fits can be seen in Figure 23 with the x-axis being the number of clusters used. The K-means fit using the BIC criterion is shown in Figure 24 with the x-axis being the number of clusters used. Last, Table 6 shows the table of clusters for each curve in the dataset along the column markers and the actual classes in the row names.

Table 6: Table of cluster groupings against actual classes for the Fraimen and Muniz Depth.

	1	**2**	**3**	**4**
AKB	4	11	1	0
OTH	2	0	6	0
TBD	11	4	6	3
TBS	0	0	0	6

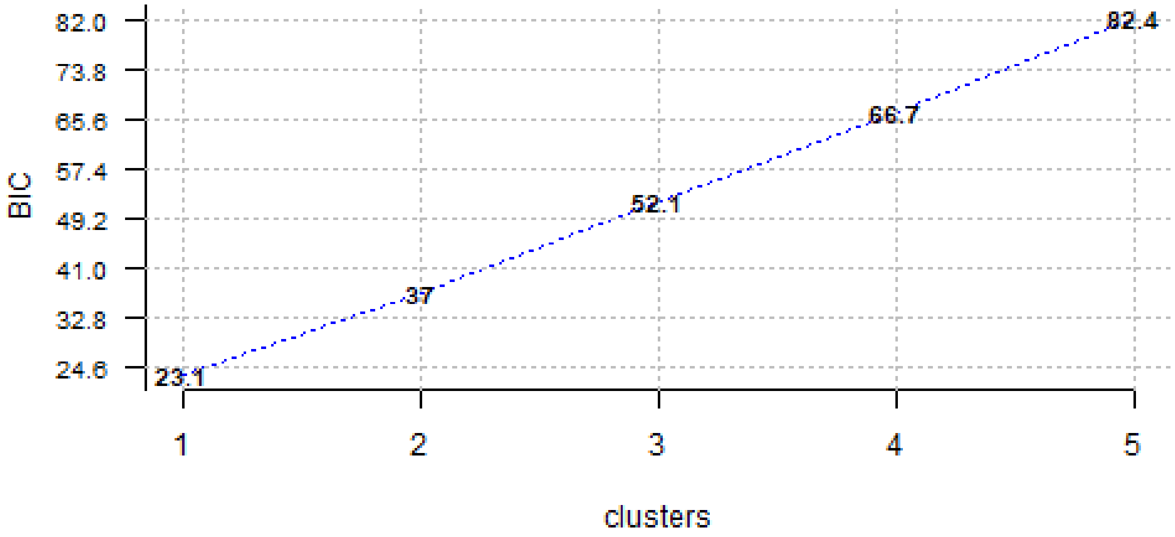

Figure 24: Plot of BIC criterion for K-Means fit for the Fraimen and Muniz Depth.

Random Projection Depths

An elbow plot of numerous K-means fits can be seen in Figure 25 with the x-axis being the number of clusters used. The K-means fit using the BIC criterion is shown in Figure 26 with the x-axis being the number of clusters used. Last, Table 7 shows the table of clusters for each curve in the dataset along the column markers and the actual classes in the row names.

Table 7: Table of cluster groupings against actual classes for the Random Projections Depth.

	1	**2**	**3**	**4**
AKB	0	2	6	8
OTH	0	4	2	2
TBD	6	10	3	5
TBS	6	0	0	0

An elbow plot of numerous K-means fits can be seen in Figure 27 with the x-axis being the number of clusters used. The K-means fit using the BIC criterion is shown in Figure 28 with the x-axis being the number of clusters used. Last, Table 8 shows the table of clusters for each curve in the dataset along the column markers and the actual classes in the row names.

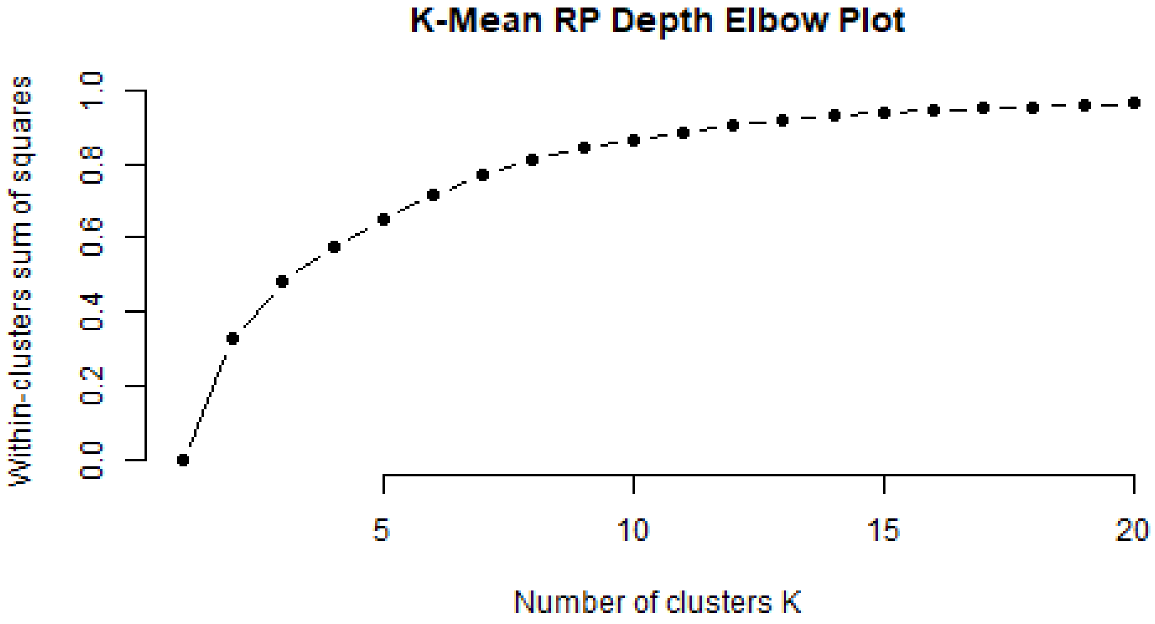

Figure 25: Elbow Plot of K-Means fit for the Random Projections Depth.

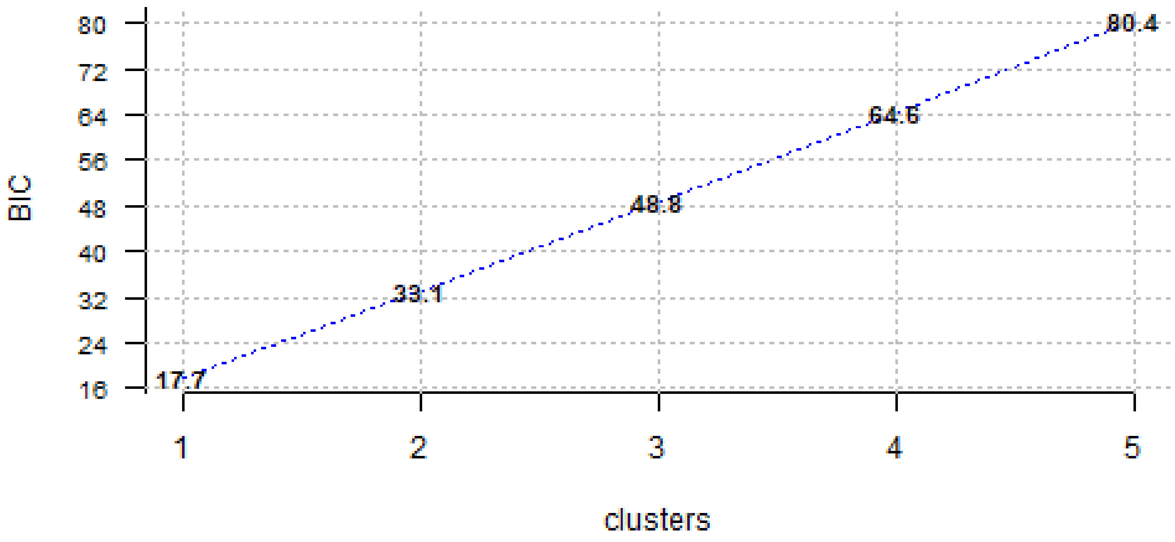

Figure 26: Plot of BIC criterion for K-Means fit for the Random Projections Depth.

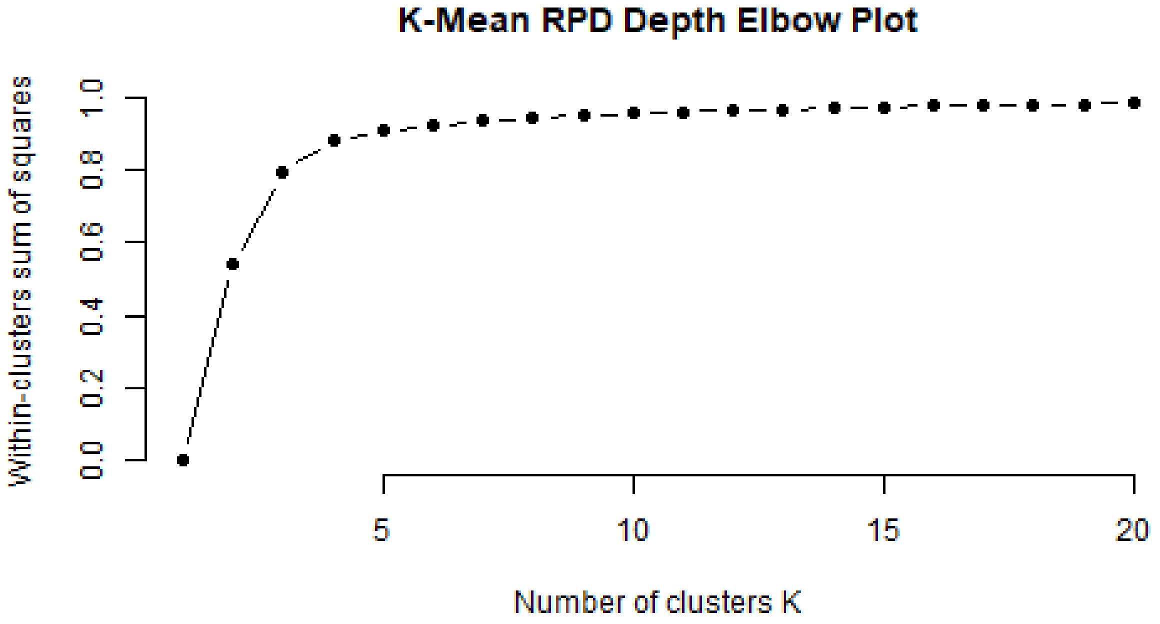

Figure 27: Elbow Plot of K-Means fit for the Random Projections Depth with a single derivative.

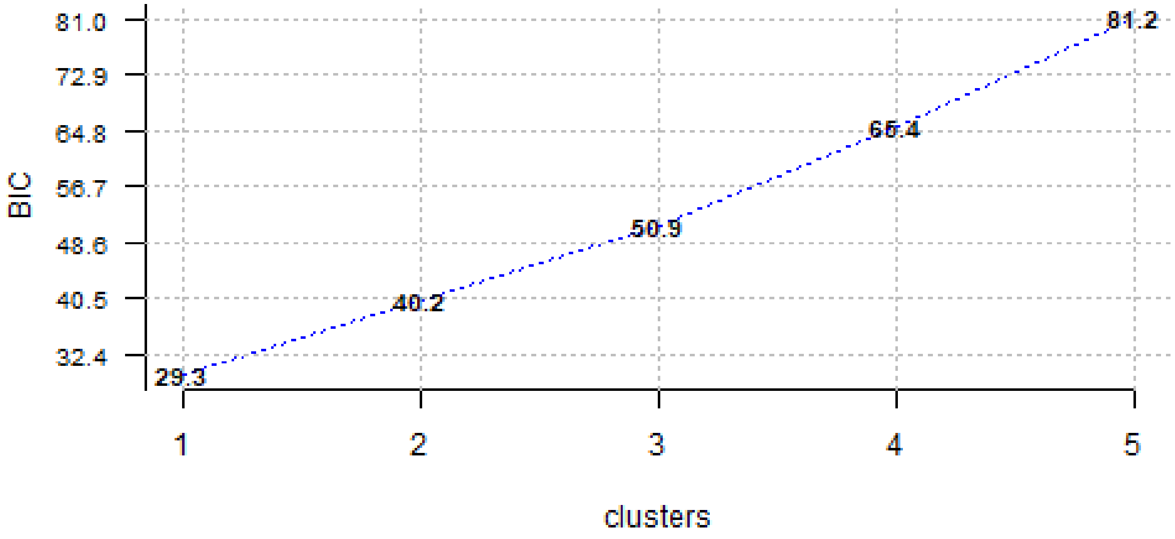

Figure 28: Plot of BIC criterion for K-Means fit for the Random Projections Depth with a single derivative.

Table 8: Table of cluster groupings against actual classes for the Random Projections Depth with a single derivative.

	1	**2**	**3**	**4**
AKB	16	0	0	0
OTH	4	2	2	0
TBD	0	18	6	0
TBS	0	0	0	6

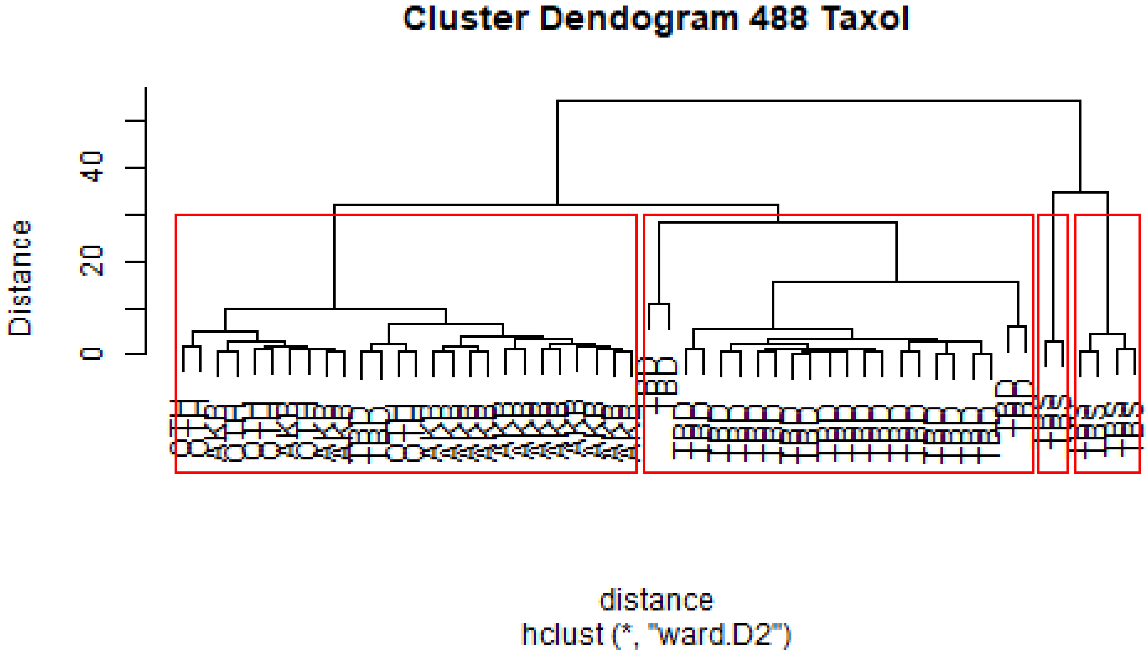

Figure 29: Dendogram of Hierarchical Clustering of 488 Taxol.

Hierarchical Clustering Functional Curves

The same assessment can be performed for hierarchical clustering. A dendogram can be produced to show the clustering of classes. This allows an assessment of the clustering and which groups tend to be clustered together based on Manhattan distance defined as:

$d(x_i, x_j) = ||x_i - x_j||_1$

This is also known as the L1 norm.

A dendogram for 488 Taxol can be shown in Figure 29 and PH3-KI67 Ratio in Figure 30 with the x-axis showing the classes in each cluster and the red boxes showing the created clusters.

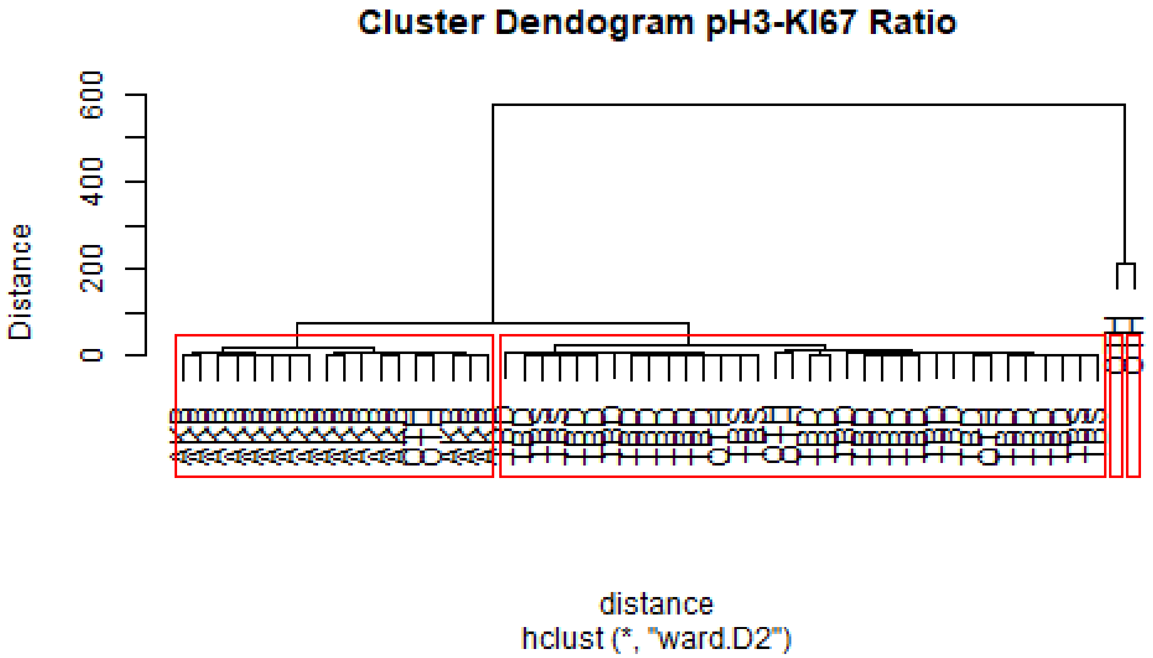

Figure 30: Dendogram of Hierarchical Clustering of PH3-KI67 Ratio.

Hierarchical Clustering AUC Summary

The same exercise can be performed with the AUC summary data can be shown in Figure 31 with the x-axis showing the classes in each cluster and the red boxes showing the created clusters.

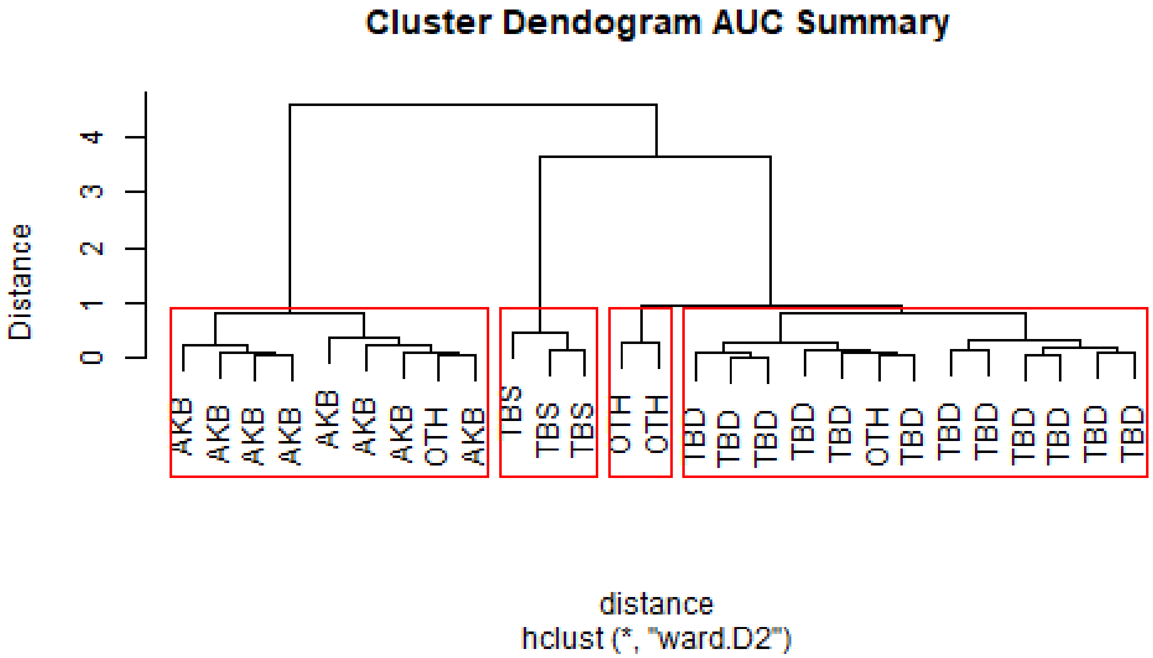

Figure 31: Dendogram of Hierarchical Clustering of AUC Summary.

Hierarchical Clustering Depth Measures

The exercise is then performed with each depth method. The Depth clustering can be shown in Figure 32 - 35 with the x-axis showing the classes in each cluster and the red boxes showing the created clusters.

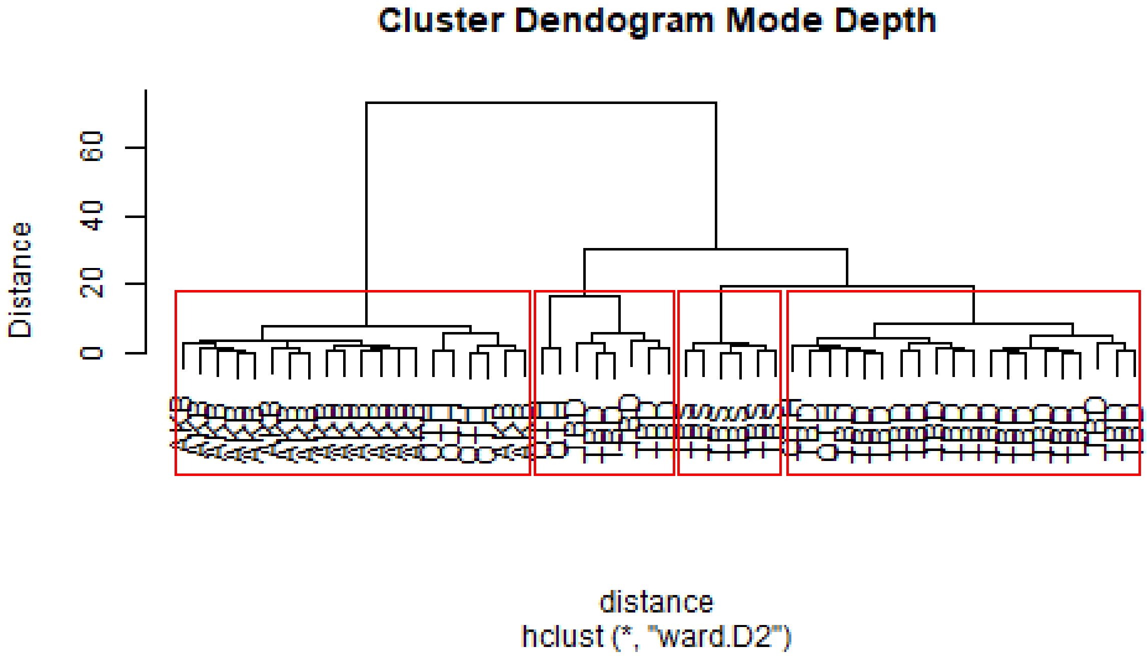

Figure 32: Dendogram of Hierarchical Clustering of H-Modal Depth.

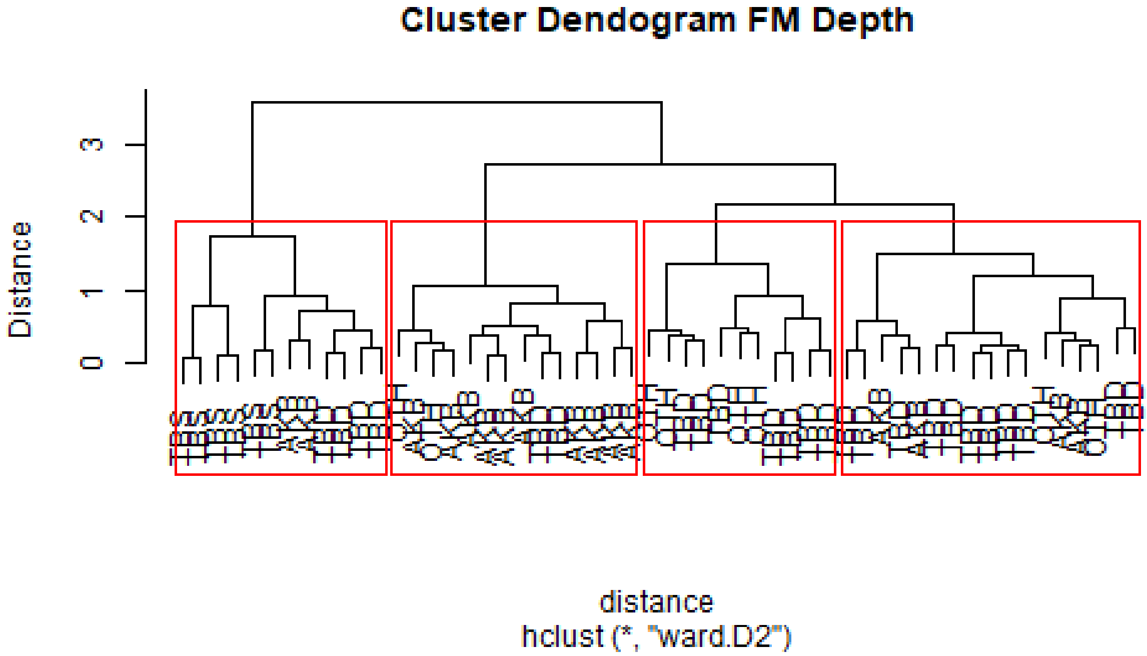

Figure 33: Dendogram of Hierarchical Clustering of Fraimen and Muniz Depth.

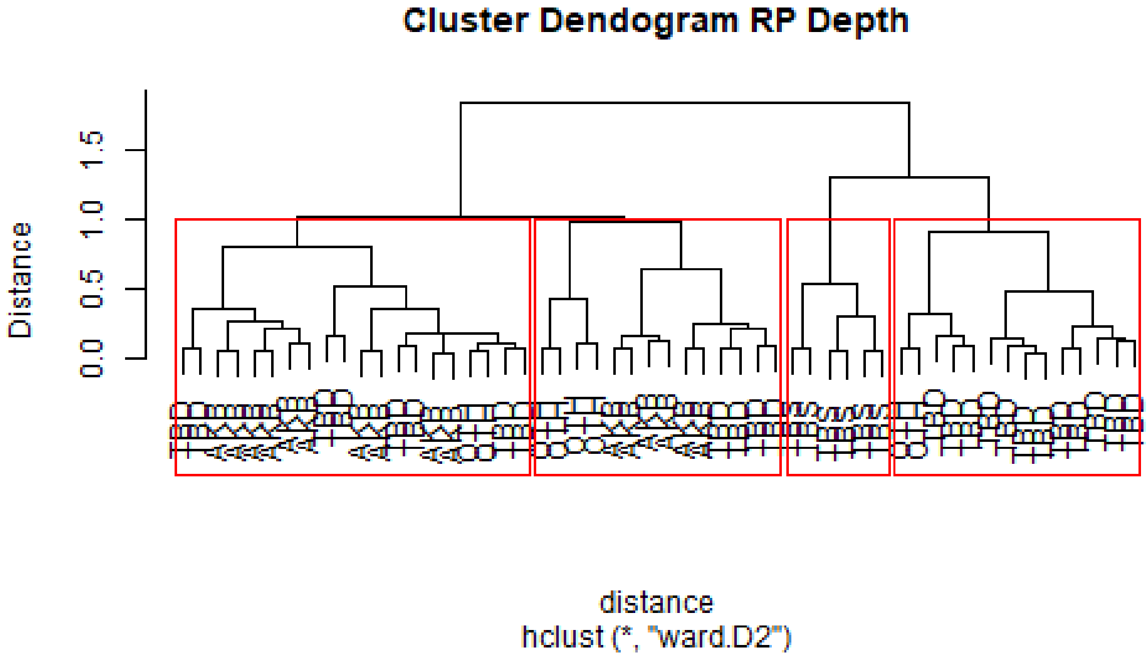

Figure 34: Dendogram of Hierarchical Clustering of Random Projections Depth.

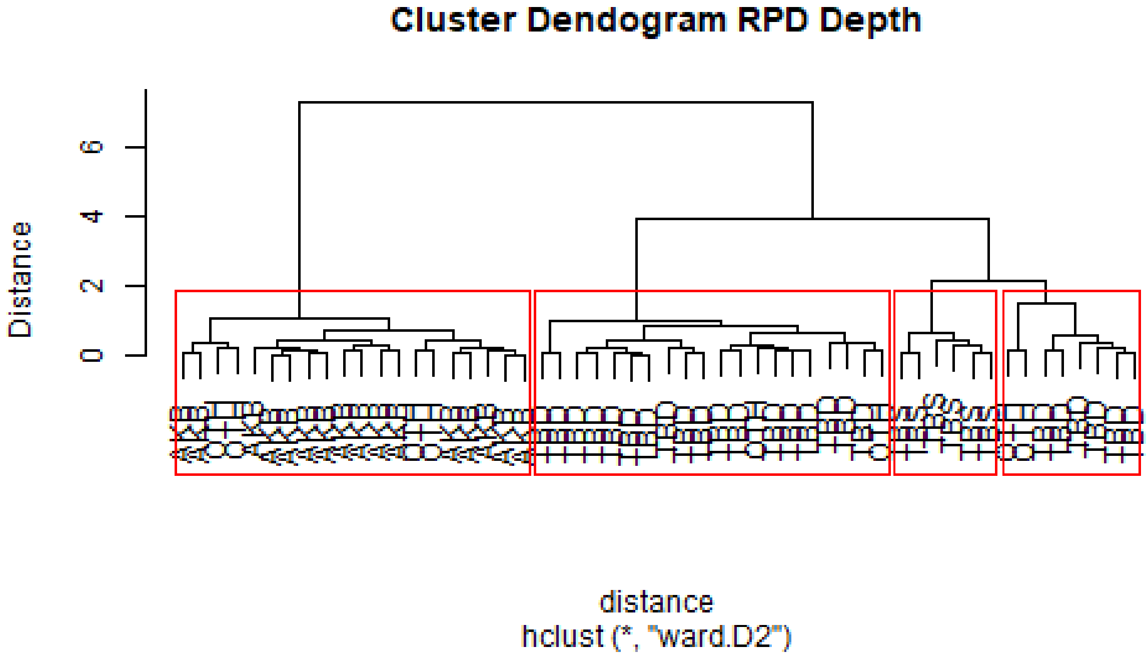

Figure 35: Dendogram of Hierarchical Clustering of Random Projections Depth with a single derivative.

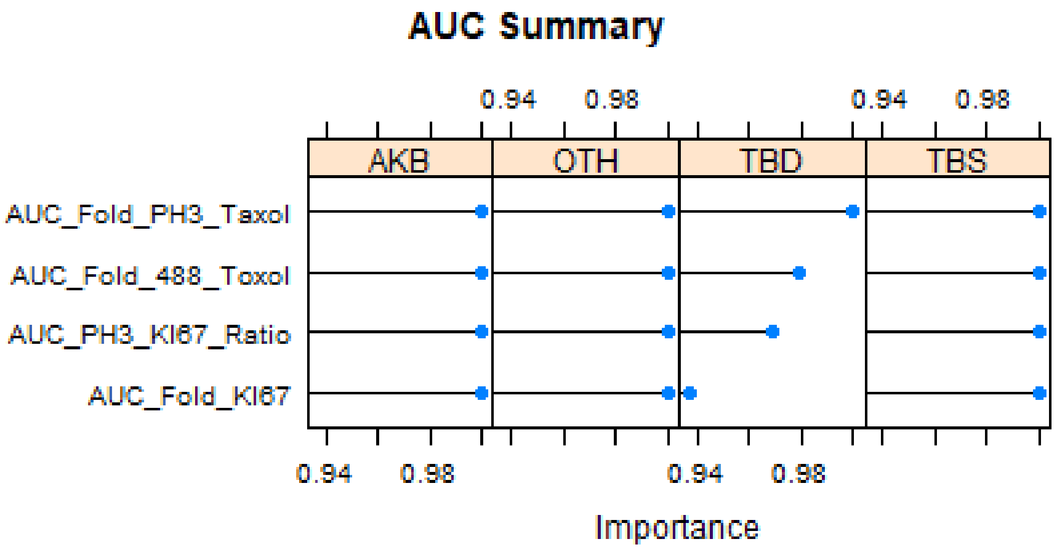

Figure 36: Plot of variable importance using LVQ Model for AUC Summary.

Supervised Learning Methods Model Evaluation

Feature Selection

To understand the impact of feature selection, a Learning Vector Quantization model can be fit to illustrate importance of each feature across each class. The plot in Figure 36 shows the importance along the x-axis and splits each class into its own plot. The variable name is on the y-axis. An importance of 1 would indicate the variable contributes strongly to the model. Cross validation is used to fit the LVQ model and assess importance of variables.

AUC summary shows all factors important except for the TBD class which seems to rely on 488 Taxol and pH3 Taxol heavily. This means all variables are most likely needed with the AUC summary data to predict the classes that are not TBD. Next, the LVQ model can be fit to the depth measures to assess importance. The results are shown in Figure 37 - 40.

On some depth measures, Fold 488 and pH3-KI67 variables contribute to the model for all classes. On other depth measures, the opposite is true in that pH3 and KI67 on their own contribute. Therefore a series of models will be fit with all factors, only Fold 488 Taxol and pH3-KI67 Ratio and the last on only pH3 and KI67 variables. What is interesting is the TBD class is heavily driven by two variables, which change based on the measure. For the most part, TBS and OTH require almost all variables.

Model Comparison

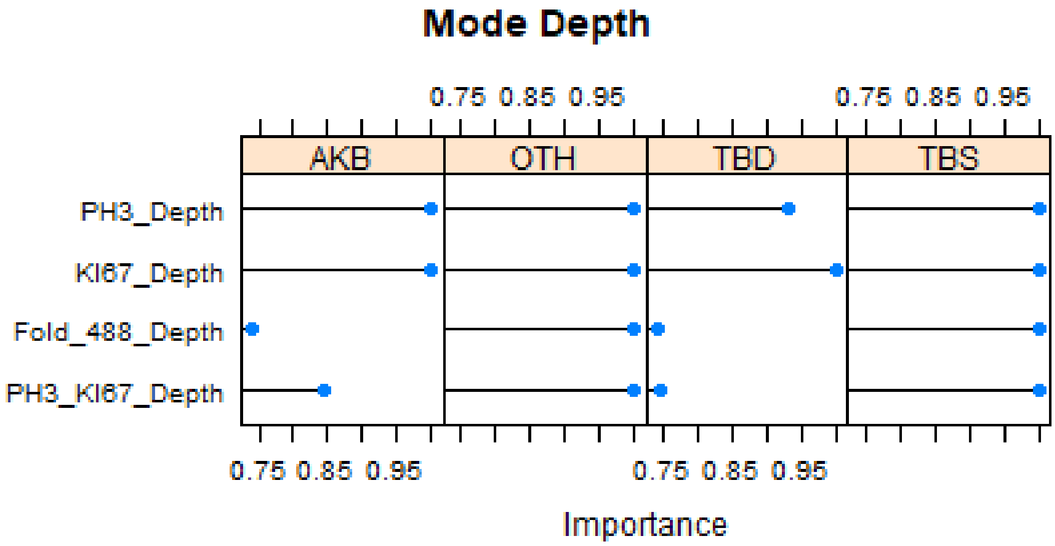

Figure 37: Plot of variable importance using LVQ Model for Depth Methods.

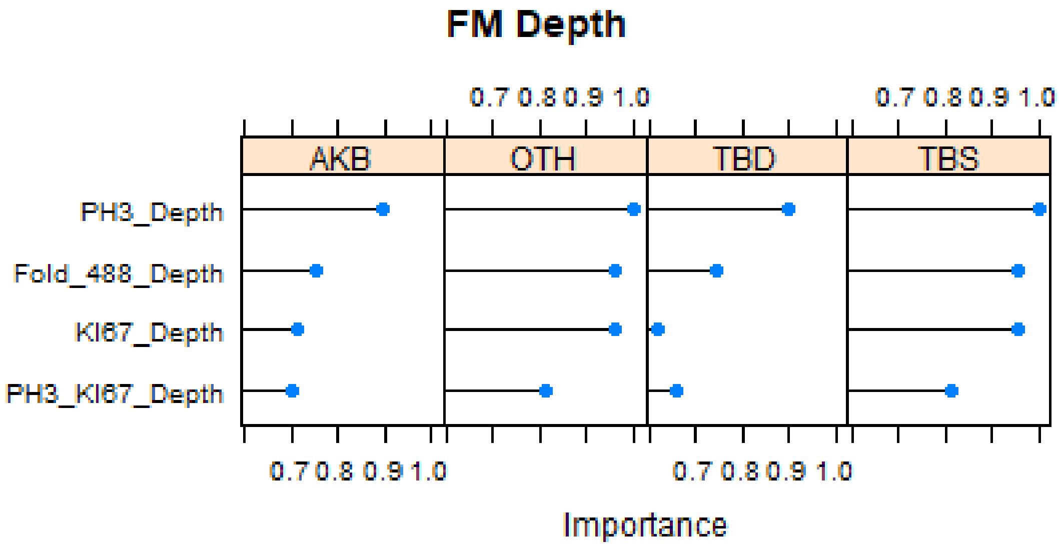

Figure 38: Plot of variable importance using LVQ Model for Depth Methods.

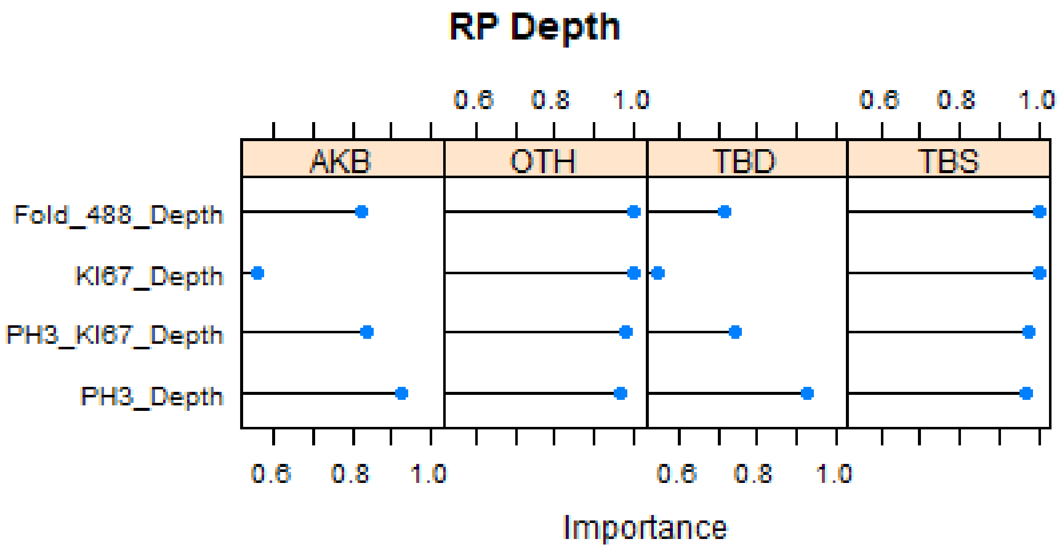

Figure 39: Plot of variable importance using LVQ Model for Depth Methods.

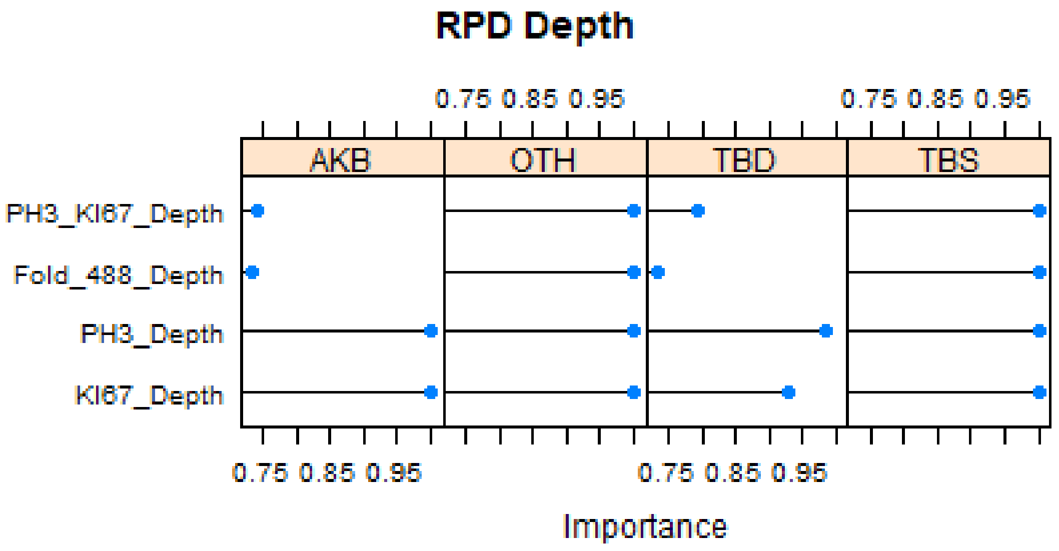

Figure 40: Plot of variable importance using LVQ Model for Depth Methods.

Initially, a large functional space is searched with numerous models and differing of what variables are actually used in the model. This space includes methods such as K-Nearest Neighbors and Linear Discriminant Analysis fit to data transformed with one of the statistical depth methods. Each model is fit with 10 rounds of 10-fold cross validation to receive an initial error estimate. A different split of variables is assessed for each model to determine the best variable and model combination through error assessment. As seen from the feature selection, each class and depth method has its variation contributed to differently from each factor. The question is then which feature and model mix provides the best empirical error estimation.

Narrowing the Model Space

After running the different variable splits, the model selection space can be narrowed down by taking the top models from each factor combination and compare them to see which should proceed for further error estimation. The models were selected as any model with an error that is less than 0.05 and a log loss that is less than one. The models chosen can be seen in the table in Table 9 and a bar plot in Figure 41. In the Figures, the word "all" represents using all variables to fit the model while "Taxol" refers to only 488 Taxol and pH3-KI67 Ratio being used. Last, pH3KI67 refers to only the pH3 and KI67 variables being used in the model.

Table 9: Table of model statistics from top models fit with 10 rounds of 10 fold cross validation.

	Mean Error	**Standard Error**	**Lower Bound**	**Upper Bound**	**Mean Log Loss**
ksvm__anovadot__all	0.0000000	0.0000000	0.0000000	0.0000000	0.2729584
knn__mode__unif__taxol	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
knn__mode__unif__all	0.0016667	0.0016667	-0.0016667	0.0050000	0.0287826
ksvm__polydot__all	0.0033333	0.0033333	-0.0033333	0.0100000	0.2729584
ksvm__vanilladot__all	0.0033333	0.0033333	-0.0033333	0.0100000	0.2729584
ksvm__anovadot__pH3KI67	0.0056667	0.0043178	-0.0029689	0.0143022	0.2289327
knn__mode__unif__pH3KI67	0.0101905	0.0058806	-0.0015708	0.0219517	0.1759853
knn__FM__unif__all	0.0149286	0.0051635	0.0046015	0.0252556	0.2578103
knn__RP__unif__all	0.0201905	0.0063524	0.0074857	0.0328953	0.3486812
lda__mode__all	0.0287143	0.0081546	0.0124050	0.0450236	0.3362988
auc__lda	0.0306667	0.0009620	0.0287427	0.0325907	0.3790810

These models are the ones to be evaluated with a larger run of cross validation to identify the best model.

Validation of Top Models

For validation of the top models, the selected models are ran with 100 rounds of 10-fold cross validation. This will provide a deeper estimate of the error associated as the empirical error distribution begins to approach its theoretical.

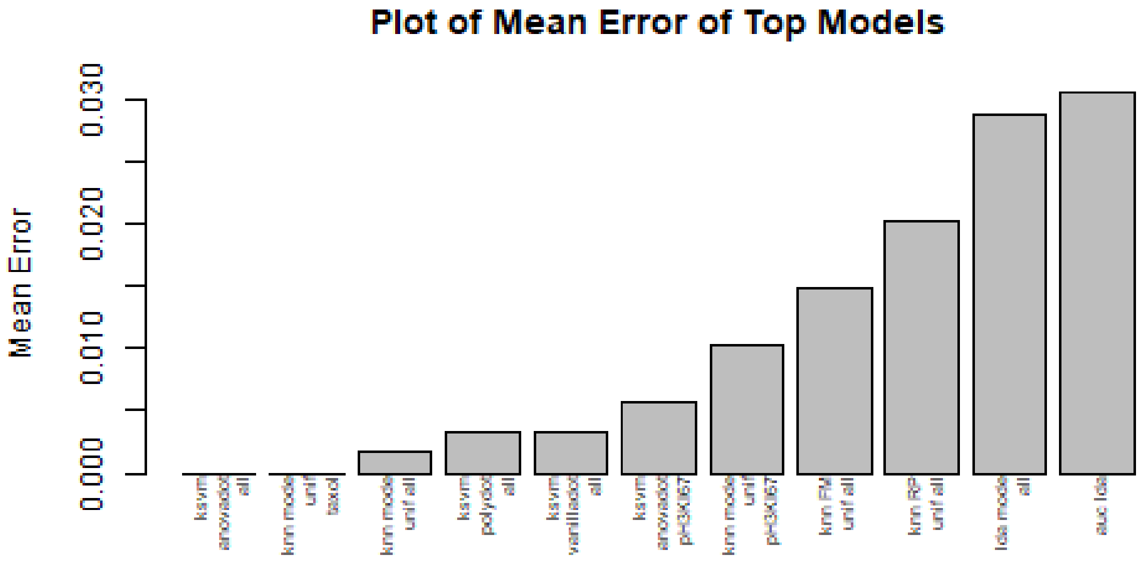

Figure 41: Barplot of error from top models fit with 10 rounds of 10 fold cross validation.

A table of the mean error, mean log loss and upper and lower bounds can be shown in Table 10.

Table 10: Table of model results from 100 rounds of 10 fold cross validation.

	Mean Error	**Standard Error**	**Lower Bound**	**Upper Bound**	**Mean Log Loss**
knn__mode__unif__all	0.0009000	1.3e-05	0.0008740	0.0009260	0.0155426
knn__mode__unif__taxol	0.0010429	1.4e-05	0.0010149	0.0010709	0.0180097
ksvm__anovadot__all	0.0015190	1.6e-05	0.0014870	0.0015510	0.2085787
knn__mode__unif__pH3KI67	0.0034571	3.3e-05	0.0033911	0.0035231	0.0597034
ksvm__polydot__all	0.0054905	4.5e-05	0.0054005	0.0055805	0.2848173
ksvm__vanilladot__all	0.0054905	4.5e-05	0.0054005	0.0055805	0.2848175
ksvm__anovadot__pH3KI67	0.0055929	3.9e-05	0.0055149	0.0056709	0.2004352
knn__RP__unif__all	0.0220929	6.3e-05	0.0219669	0.0222189	0.3815345
knn__FM__unif__all	0.0237548	6.7e-05	0.0236208	0.0238888	0.4102350
lda__mode__all	0.0325643	8.4e-05	0.0323963	0.0327323	0.2829074
auc__lda	0.0334667	9.9e-05	0.0332687	0.0336647	0.3682407

Discussion

Unsupervised Learning Methods

K Means

In reviewing the unsupervised learning method results. Some interesting conclusions can be drawn. The fold 488 Taxol and pH3-KI67 K-means results show the BIC point to an optimal K value of two and three clusters respectively as seen from Figure 16 and Figure 18. In the table of clusters versus classes, both variables and models tend to cluster the OTH category in many different classes and never in its own cluster as shown in Figure 17 and Figure 20. This is in line with the thinking of OTH being a leftover class made up of differing reactions than the stabilizers, destabilizers and inhibitors. Also, both variables allow K-means to find distinct categories for AKB and TBS. This shows that there is some distance that contributes to the separation of these classes from TBD. The TBD class is spread out in the 488 Taxol results, but the pH3-KI67 ratio results show a distinct TBD category. This is interesting as feature selection shows pH3-KI67 does not really contribute strongly to TBD's variation and relative importance in the model.

Moving on to the AUC K-means model, the elbow plot in Figure 19 is distinct at around three clusters, while the BIC plot in Figure 20 shows no value is added after a single cluster. The table in Figure 23 shows a distinct separation in the TBD and TBS classes, but struggles to separate the AKB and OTH classes from each other. This could lead to a possible theory that these two classes have less separation in distance in the AUC summary data.

The H-Modal depth method shows the optimal BIC selected cluster to be four as seen in Figure 22, which is promising. However, in the table of results in Figure 26, it does not cleanly separate the clusters to our known classes. OTH and TBD tend to be blended with other classes. The other depth methods (Fraimen and Muniz, Random Projections) perform worse than the H-Modal depth in that they only select a single cluster as optimal and the classes tend to become more blended than the H-Modal depth which can be seen in Figures 29 - 28.

This K-Means analysis leads to some initial clues of where supervised learning models could run into some trouble. Mode depth appears to perform best out of the depth methods and different variables tend to have different impacts to classes.

Hierarchical Clustering

The Manhattan distance based hierarchical model for 488 Taxol does a good job of separating the TBD and the TBS class to an extent as seen in Figure 29, but again fails to separate OTH from AKB. The pH3-KI67 ratio model in Figure 30 blends OTH with other classes and fails to separate the TBD and TBS models. Again, this could possibly mean this variable does not contribute much to TBD's variation and thus is not a good separator.

The AUC model does a decent job in segmenting the classes as seen in Figure 31. It does tend to lump OTH in with other classes, which is to be expected at this point. The depth methods do not show a clear separation between classes similar to the K-means methods. The mode depth appears to perform the best in this case, while not impressive on its own.

Unsupervised clustering in general does not provide a clear solution for segmenting the dataset just based on distance based metrics. However, in some cases, it does not perform as poorly as expected. There is clearly some contribution to the class selection from a different mix of variables and through different ways of transforming those variables. These initial results contribute to similar conclusions found when fitting the supervised models.

Supervised Learning Methods and Top Model Selection

In assessing the results of supervised learning, there is a wide model space reviewed. It becomes clear very quickly that some methods just do not perform well. The kernel methods and tree methods do not have impressive error rates with many being greater than 0.50 for kernel methods. This is worse than random guessing. The depth measure used and variable selection does not really help this situation. The best kernel method gets slightly above a 0.05 error rate. This is a triweight kernel using mode depth and all variables. However, this performance is rare for the kernel methods with the majority of them performing poorly.

As expected, the depth methods affect which models perform better than others. Very few of the random projection methods do not perform well enough to move on to deeper error estimation. In fact, the random projection methods tend to overfit quite a bit scoring highly on the training set, then exhibiting large loss during cross validation.

Overall, K-Nearest Neighbors performs well when coupled with the mode depth and Fraimen and Muniz depths. All versions of K-Nearest Neighbors with mode depth are captured in the top models as seen in Figure 41. The version with only fold 488 Taxol and pH3-KI67 variables and the version with all variables appear to perform similarly with the pH3-KI67 only models performing what appears to be noticeably worse. This would lead to the conclusion of the importance of the fold 488 Taxol and pH3-KI67 in the discrimination with pH3 and KI67 independent variables adding minimal value to the model. The random projection models do not perform as well with the nearest neighbors model with only one random projection model having error less than 0.05. The nearest neighbors base methods actually do not perform poorly either. However, they are removed due to a high amount of log loss measured across all variable combinations.

Linear discriminant analysis is also strongest when paired with a mode depth. Contrary to nearest neighbors, variable selection degrades the LDA models with mode depth. LDA performs the best when all variables are used in the model. LDA also declines rapidly when switching out the depth method away from the mode depth. Last, the LDA method performs well with the AUC summary data. There is some degree of separation with the AUC data, however it does appear to be as strong as with the depth methods.

Last, Kernel Support Vector machines also have some impressive results. These methods are fit to the data presented as a large matrix consisting of 20 columns for each variable. The data is not transformed to a depth method the for KSVM classifier. The kernel selection also makes a significant difference for the KSVM classifier with the ANOVAdot kernel outperforming every other kernel such as the Gaussian kernel, polydot and laplace dot kernels. Variable selection itself also makes a difference to the model. The ANOVAdot kernel performs best when all variables are used and slightly degrades when subsets are used. However, the degrade is less for the pH3 and KI67 variables than with the fold 488 Taxol and pH3-KI67 ratio variables. This would lead to the conclusion that the feature selection performed rings true for the KSVM model with pH3 and KI67 contributing stronger to the model than fold 488 Taxol and pH3-KI67 variables especially for the TBD class.

Overall, the models chosen to proceed based on their error rates are differing combinations of K-nearest neighbors with depth methods, linear discriminant analysis with depth methods and AUC summarization and kernel support vector machines with differing kernels fit on the raw data. Kernel methods not part of the support vector machine and the tree methods are eliminated due to their error rates being much higher than the methods selected.

Validation of Top Models

Selection of the Best Model

The best way to understand the performance of the 100 rounds of 10-fold cross validation on the top models is through box plots of the distribution of the means of rounds.

The box plots show there appears to be minimal difference in the first three models and the variation begins to increase by the fourth model. The last four models are visibly inferior to the first seven models. This

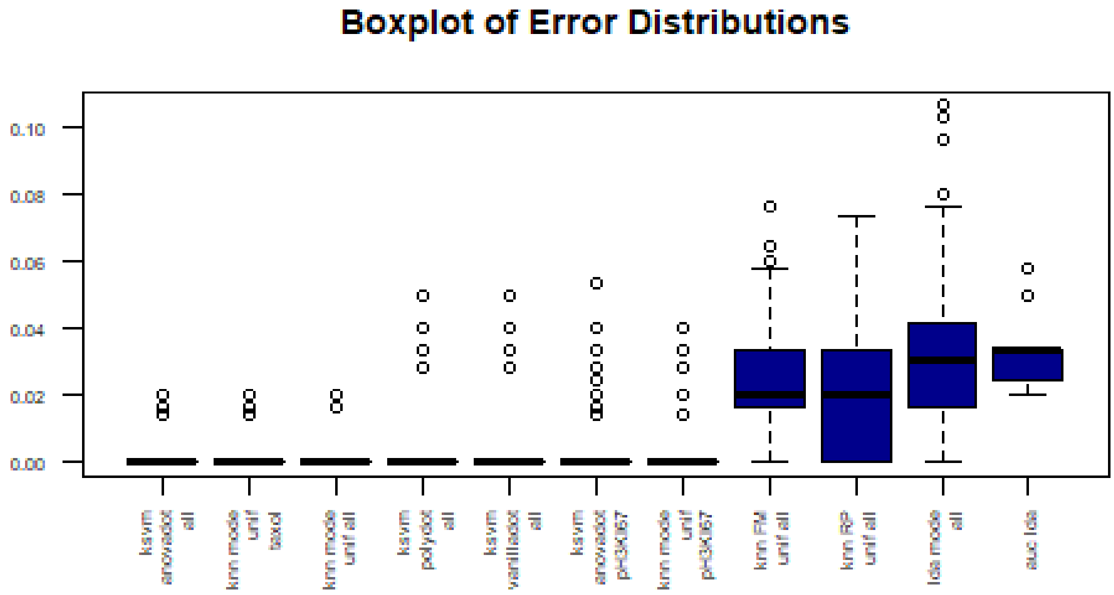

Figure 42: Boxplot of Error Distributions for top models fit with 100 rounds of 10 fold cross validation.

shows that the AUC LDA method is worse than the KSVM ANOVAdot method. These results are supported by the table of error which ranks the KNN Mode model with all variables as the model with the lowest error closely followed by the KNN mode model with only fold Taxol 488 and pH3-KI67 ratio variables and the KSVM ANOVAdot model with all variables. The question then becomes, is there a significant difference in the error rates of these models and which is the best model?

The Tukey pairwise comparison is ran across an ANOVA model fit with the error rates as a response and the model names as factors. The results can be seen in Table 11. Simply looking at one set of full pairwise comparisons for the first model confirms what is apparent from the box plots. There is no statistical difference in the comparisons of the first seven models. The differences start to become significant in the last four models. The major factor differentiating the models seems to be due to variation. However, log loss helps solve this problem to an extent. Out of the first seven models, the KNN mode models have much lower log loss than the KSVM model.

A box plot of log loss in Figure 43 shows how the models are distributed.

Table 11: Table of pairwise Tukey tests of error for first model.

	diff	lwr	upr	p adj
knn__mode__unif__taxol-ksvm__anovadot__all	-0.0004762	-0.0063465	0.0053941	1.0000000
knn__mode__unif__all-ksvm__anovadot__all	-0.0006190	-0.0064894	0.0052513	0.9999998
ksvm__polydot__all-ksvm__anovadot__all	0.0039714	-0.0018989	0.0098418	0.5180950
ksvm__vanilladot__all-ksvm__anovadot__all	0.0039714	-0.0018989	0.0098418	0.5180950
ksvm__anovadot__pH3KI67-ksvm__anovadot__all	0.0040738	-0.0017965	0.0099441	0.4777807
knn__mode__unif__pH3KI67-ksvm__anovadot__all	0.0019381	-0.0039322	0.0078084	0.9930047
knn__FM__unif__all-ksvm__anovadot__all	0.0222357	0.0163654	0.0281060	0.0000000
knn__RP__unif__all-ksvm__anovadot__all	0.0205738	0.0147035	0.0264441	0.0000000
lda__mode__all-ksvm__anovadot__all	0.0310452	0.0251749	0.0369156	0.0000000
auc__lda-ksvm__anovadot__all	0.0319476	0.0260773	0.0378179	0.0000000

The last four models have larger amounts of log loss, however, this becomes more ambiguous in the higher performing error rate models. Another Tukey pairwise comparison is ran to illustrate significant differences between log loss. Table 12 shows that the KSVM ANOVAdot model's log loss differs significantly from the KNN mode models. This changes as the comparison is performed to other KSVM models.

Table 12: Table of pairwise Tukey tests of log loss for first model.

	diff	lwr	upr	p adj
knn__mode__unif__taxol-ksvm__anovadot__all	-0.190569	-0.267622	-0.113515	0.000000
knn__mode__unif__all-ksvm__anovadot__all	-0.193036	-0.270090	-0.115983	0.000000
ksvm__polydot__all-ksvm__anovadot__all	0.076239	-0.000815	0.153292	0.055391
ksvm__vanilladot__all-ksvm__anovadot__all	0.076239	-0.000815	0.153292	0.055390
ksvm__anovadot__pH3KI67-ksvm__anovadot__all	-0.008143	-0.085197	0.068910	1.000000
knn__mode__unif__pH3KI67-ksvm__anovadot__all	-0.148875	-0.225929	-0.071822	0.000000
knn__FM__unif__all-ksvm__anovadot__all	0.201656	0.124603	0.278710	0.000000
knn__RP__unif__all-ksvm__anovadot__all	0.172956	0.095902	0.250009	0.000000
lda__mode__all-ksvm__anovadot__all	0.074329	-0.002725	0.151382	0.069972
auc__lda-ksvm__anovadot__all	0.159662	0.082609	0.236716	0.000000

However, in comparing the two KNN-mode models as seen in Table 13, the Tukey pairwise comparison finds no difference between the two models based on log loss. This is not so when compared to every other model except another KNN model with different variables selected.

Table 13: Table of pairwise Tukey tests of log loss for second model.

	diff	lwr	upr	p adj
knn__mode__unif__all-knn__mode__unif__taxol	-0.002467	-0.079521	0.074586	1.000000
ksvm__polydot__all-knn__mode__unif__taxol	0.266808	0.189754	0.343861	0.000000
ksvm__vanilladot__all-knn__mode__unif__taxol	0.266808	0.189754	0.343861	0.000000
ksvm__anovadot__pH3KI67-knn__mode__unif__taxol	0.182425	0.105372	0.259479	0.000000
knn__mode__unif__pH3KI67-knn__mode__unif__taxol	0.041694	-0.035360	0.118747	0.811537
knn__FM__unif__all-knn__mode__unif__taxol	0.392225	0.315172	0.469279	0.000000
knn__RP__unif__all-knn__mode__unif__taxol	0.363525	0.286471	0.440578	0.000000
lda__mode__all-knn__mode__unif__taxol	0.264898	0.187844	0.341951	0.000000
auc__lda-knn__mode__unif__taxol	0.350231	0.273177	0.427284	0.000000

Therefore in trying to decide on the best model, error is not the only metric to drives that decision. Using the log loss helps to break some ties and show that the KNN Mode Depth models perform significantly better

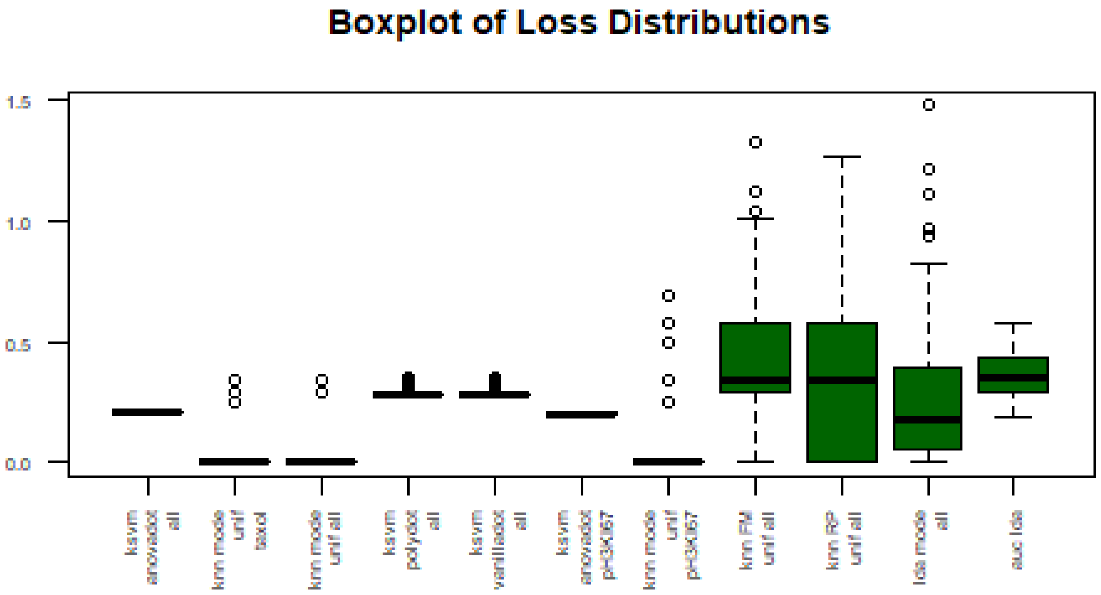

Figure 43: Boxplot of Loss Distributions for top models fit with 100 rounds of 10 fold cross validation.

than the KSVM ANOVAdot model when using log loss as the metric for comparison. In deciding between the two KNN mode models, the fact that both are relatively similar except in their variable selection shows that adding all variables to the model provides minimal value and in fact a simpler model with less variables performs as well.

Assessment of Model Function

Now that the best models have been identified, the question becomes why do these models work? The answer lies in the depth methods and the discrimination they drive. Earlier, plots of the depth methods were reviewed, however if the classes are imposed over those plots as colors, it shows some distinct patterns. It is easy to see in Figure 44 that the mode depth creates a scenario where each point essentially has a neighbor close to it even when it is separated from the cluster of its class. The AKB and TBD variables in colors black and blue respectively exhibit strong clusters and separation in a lot of the plots. The variables Fold 488 Taxol and pH3-KI67 help the TBS class in green from being distinct unto itself while the pH3 and KI67 variables help the OTH class be distinct unto itself. Recall the K-Nearest Neighbor model is found by:

$KNN = \frac{1}{k} \sum_{x_i \in \nu_k(x)} I(Y = j)$

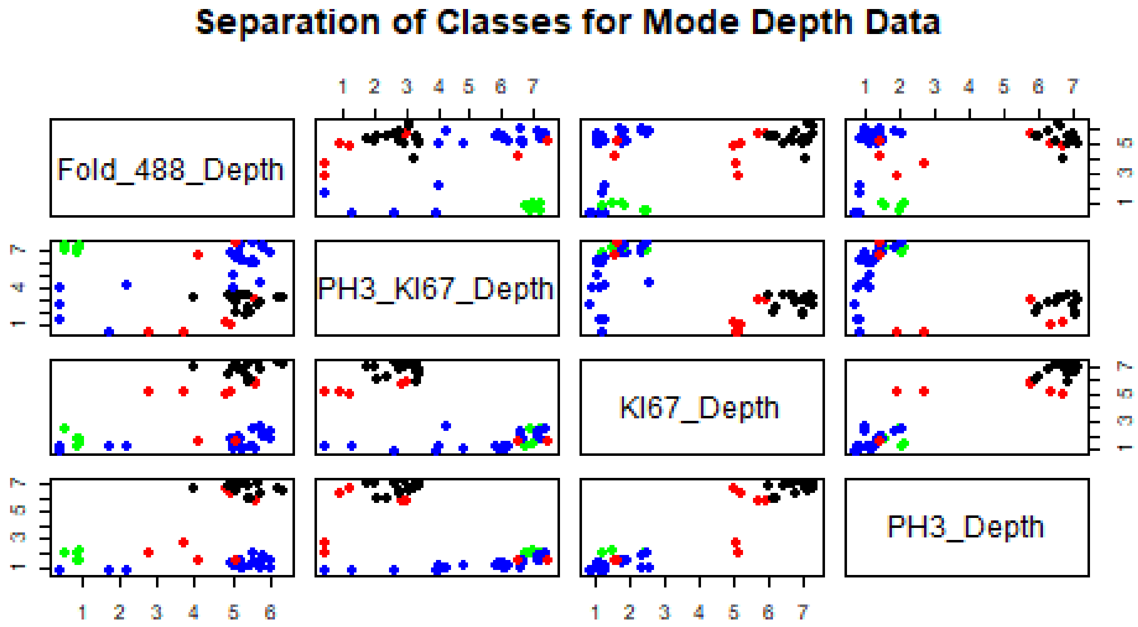

Figure 44: Pairs plot of H-Modal Depth colored by class

Where k is the number of neighbors, $\nu_k(x)$ is the neighborhood of x values in the k space and j represents class membership. Therefore, KNN works as an equidistant weighting mechanism to weight points based on their distance. With the single nearest neighbor, the closest point receives the weight and drives the class selection. This model is incredibly simple, but effective in this case.

Since only one neighbor is used, the model is able to find each class' partner close by. Very few points are isolated unto themselves. the single neighbor is also a very flexible model, so it is not bound by linear separability as LDA would be.

As a contrast, the random projection depth is plotted in Figure 45. With this depth, there is often close clustering of the TBS and TBD classes and the OTH classes with the TBD and AKB classes.

Comparing this with the AUC Summary data in Figure 46. The AUC summary data struggles with the linear bounds that LDA restricts it to and the nearest neighbors model has some instances where the OTH class and the AKB class are closely overlapped. It does an admirable job, but the mode depth is able to better segment than the AUC summary method.

The Kernel Support Vector machine's performance with an ANOVAdot kernel is interesting, but makes sense. The KSVM is defined as:

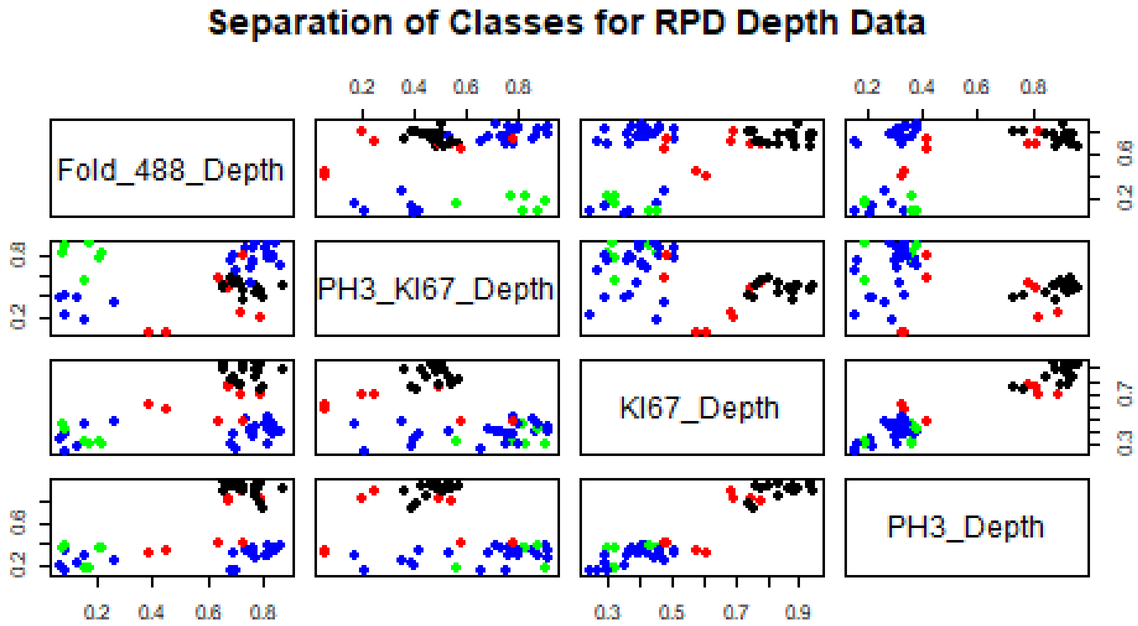

Figure 45: Pairs plot of Random Projection Depth with single derivative colored by class

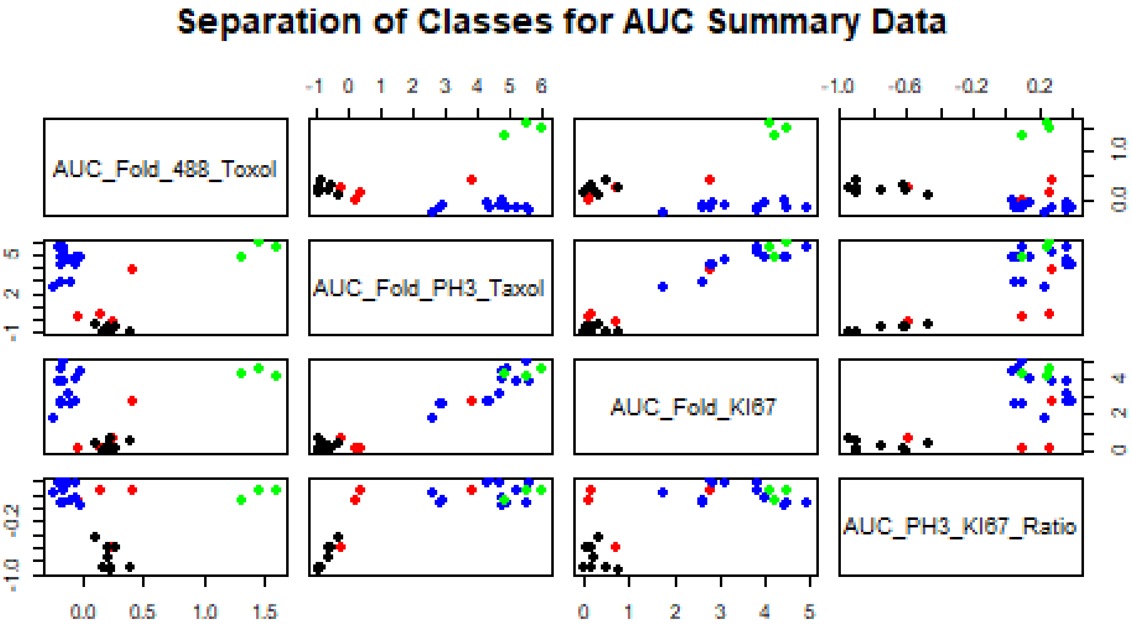

Figure 46: Pairs plot of AUC Summary colored by class

$KSVM = sign(\sum_{i=1}^{|s|} \hat{\alpha}_{s_j} Y_{s_j} K(x_{s_j}, x) + \hat{b})$

Where $K(x_{s_j}, x)$ represents the ANOVAdot kernel found by:

$K(x_{s_j}, x) = (\sum_{l=1}^{q} exp(-\delta(x_{i_l} - x_{j_l})^2))^d$

The fact the KSVM works well is due to the inner product the kernel finds. This allows a large amount of variables to be projected and variables that do not contribute essentially receive zero weight in the model and are dropped. Therefore, summarizing the functional curves into statistical depths or summaries is not necessary as the KSVM will weight the important pieces of the curve and then translate those to the KSVM model for use.

However, there is something else happening here. There are replicates in the original raw data. KNN certainly benefits from this data by always having a neighbor and KSVM benefits less, but still benefits from the extra data. In practice, having numerous similar training cases is often common practice. This just means when a new curve comes in, it will check this dimensional space, look for a match and classify. If the chemicals used are representative of aneugens, then the model will still perform well. However, as a counter balance, we can review non-replicate data to see how it performs.

Counter Balance to Raw Data

Table 14: Table of error and log loss for 100 rounds of summarized data.

	Mean Error	Standard Error	Lower Bound	Upper Bound	Mean Log Loss
auc.lda	0.033467	0.000099	0.033269	0.033665	0.368241
auc.knn	0.064000	0.001347	0.061306	0.066694	0.000000
func.mda.all.auc	0.072333	0.001392	0.069549	0.075117	1.098121
func.lda.all.avg.mode	0.132683	0.000171	0.132341	0.133025	0.617137
func.mda.all.avg.mode	0.134833	0.001746	0.131341	0.138325	1.686309
func.fda.all.avg.mode	0.138383	0.000181	0.138021	0.138745	0.671213
func.knn.all.avg.mode	0.157800	0.000187	0.157426	0.158174	0.000000
func.knn.all.avg.rp	0.231050	0.000208	0.230634	0.231466	3.990138
func.knn.all.avg.rpp	0.248617	0.000209	0.248199	0.249035	4.293507
func.lda.all.avg.fm	0.279683	0.000268	0.279147	0.280219	3.133328

In the assessment of the models, it is found that replicates help train the KNN and KSVM models, if only non-replicated data is used, methods such as mode depth and KNN do not fair as well. This is due to reducing the number of observations, which removes KSVM from being a possible model. However, adding methods such as mixture discriminant analysis and flexible data analysis add some new models in place of KSVM. The non-AUC models were fit by averaging the replicates of data. The results are that the depth methods do not perform as well as the AUC summary data. The linear discriminant analysis model with the AUC summary data is the best model easily. Adding flexibility with MDA does not help improve the error as seen

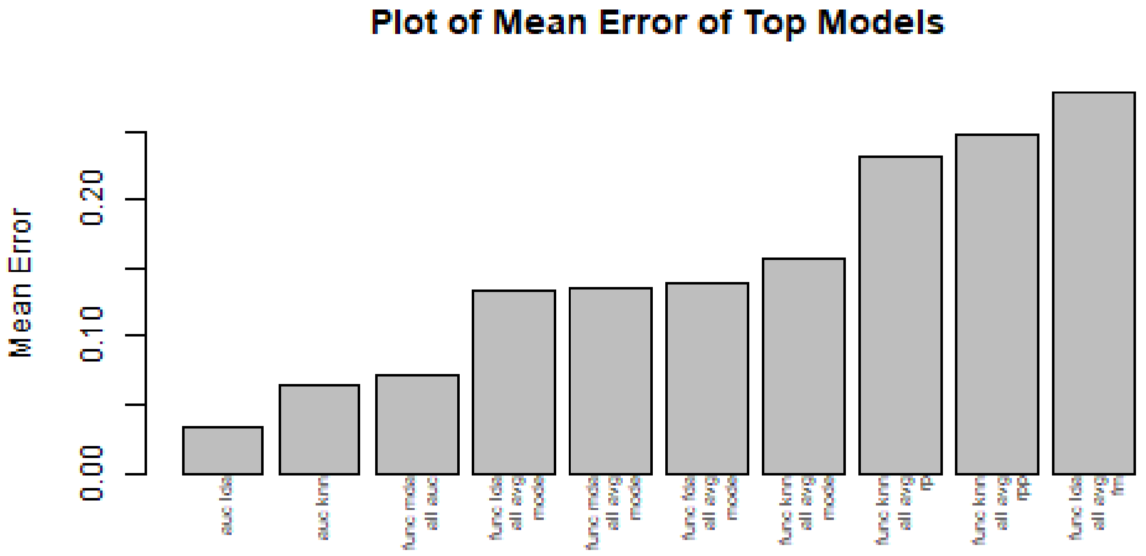

Figure 47: Barplot of error from models fit with 100 rounds of 10 fold cross validation.

from the bar plot in Figure 47.

Conclusion

In conclusion, aneugens can be classified using a K Nearest Neighbors Model on an H-Modal depth transformation of a functional curve and with kernel support vector machines. This allows similar class aneugens to become clustered together even when some of the classes are unique among themselves in their vector space. These are promising results that illustrate the impact of statistical distances to functional curves. However, the weakness of this approach is the need for replicate data. In practice, this may not be a bad thing and the model may still most likely provide good results. Removing replicates makes the LDA model with the AUC summary data the best model.

There are some further areas of research that could expand upon this assessment. For example, certain weak learners could be bagged or boosted to improve their classification performance. Trees and random projection spaces may improve using bagging or boosting ensembles. Recurrent neural networks, specifically the long short term memory network also produces some interesting results in an initial assessment. Using eight prior points in an LSTM leads to good model performance. However, this method was not included in this study due to the extreme amount of variation it drives. With more time and experimentation, the LSTM could be a valuable avenue of research for the functional curve. Also, it would be interesting to collect a larger sample of curves to see if AUC summary data discrimination can be improved and to measure how the KNN and KSVM models perform with unseen data. Another key area for investigation is performing leave one out cross validation across a chemical. This would pull a single chemical and its replicates out of model evaluation to get a better estimate of error of purely unseen chemical curves. The data used in this study were averaged and did not perform as well as AUC summarized data, but it is possible the raw functional curves could improve performance.

References

1. Bavetsias, V., and Linardopoulos, S. (2015). “Aurora Kinase Inhibitors: Current Status and Outlook”. Frontiers in oncology, 5, 278. doi:10.3389/fonc.2015.00278

2. Bryce SM, Bernacki DT, Bemis JC, Dertinger SD. 2016. Genotoxic mode of action predictions from a multiplexed flow cytometric assay and a machine learning approach. Environ Mol Mutagen 57:171-189.

3. Fanale, Daniele and Bronte, Giuseppe and Passiglia, Francesco, et al., “Stabilizing versus Destabilizing the Microtubules: A Double-Edge Sword for an Effective Cancer Treatment Option?,” Analytical Cellular Pathology, vol. 2015, Article ID 690916, 19 pages, 2015. https://doi.org/10.1155/2015/690916.

4. Febrero-Bande, Manuel and Galeano, Pedro and Gonzãlez-Manteiga, Wenceslao. (2008). “Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels”. Environmetrics. 19. 331 - 345. 10.1002/env.878.

5. Flores, Ramon and Lillo, Rosa and Juan, Romo. (2015). “Homogeneity Test for Functional Data”. Cornell University Library. https://arxiv.org/pdf/1507.01835.pdf

6. Plante, N. (2003). “Molecular toxicology”. Taylor & Francis. pp. 63-65.