Supplementary Material for *IntegratedMRF*: Random Forest based Framework for integrating prediction from different data types

1 Predictive Models

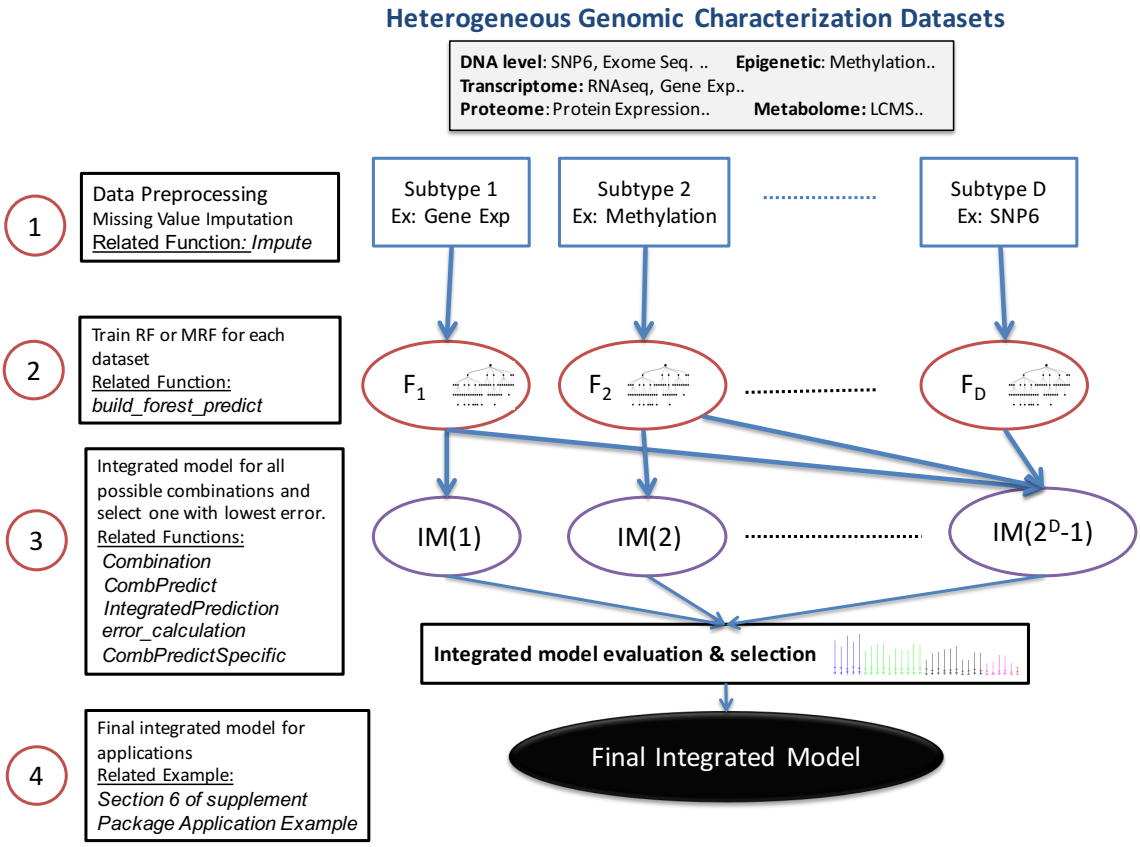

Figure 1: Overview of Integrated Prediction Methodology

1.1 Random Forest (RF)

Random Forest regression refers to ensemble of regression trees where a set of T un-pruned regression trees are generated based on bootstrap sampling from the original training data. For selecting the feature for splitting at each node, a random set of m features from the total M features are used. The inclusion of the concepts of *bagging* (Bootstrap sampling for each tree) and *random subspace sampling* (node split selected from random subset of features) increase the independence of the generated trees and thus the averaging of the prediction over multiple trees has lower variance compared to individual regression trees.

Process of splitting a node

Let, $x(i,j)$ and $y(i)(i = 1,.,n; j = 1,...,M)$ denote the input and output response samples respectively. At each node of the regression tree, the node cost is considered as the sum of squares:

$$D(\eta_P) = \sum_{i \in \eta_P} (y(i) - \mu(\eta_P))^2 \quad (1)$$

where $\mu(\eta_P)$ is the expected value $y(i)$ in node η_P.

The reduction in cost for partition γ at node η_P is

$$C(\gamma, \eta_P) = D(\eta_P) - D(\eta_L) - D(\eta_R) \tag{2}$$

where η_L (left node with samples $x(I \in \eta_p, j_s) \leq z$) and η_R (right node with samples satisfying $x_{tr}(I \in \eta_p, j_s) > z$) denotes the child nodes after the split. The partition $\gamma*$ that maximizes $C(\gamma, \eta_P)$ for all possible partitions is selected for node η_P.

Forest Prediction

If $\hat{y}(x, i)$ denotes the prediction for Tree i for input x, the prediction for the forest is given by

$$\hat{y}(x) = \frac{1}{T} \sum_{i=1}^{T} \hat{y}(x, i) \tag{3}$$

1.2 Multivariate Random Forest (MRF)

The multiple response scenario has output $y(i, k)(i = 1, ..., n; k = 1, ..., r)$. The primary difference between MRF and RF is in the generation of the trees with different node costs $D_{MRF}(\eta)$ and $D(\eta)$.

The node cost $D(\eta_P) = \sum_{i \in \eta_P} (y(i) - \mu(\psi_P))^2$ for the univariate case is the sum of squares of the differences between the output response and the mean output response for the node. For multivariate case, the difference between a sample point and the multivariate mean distribution is desirable and can be achieved as the sum of the squares of *Mahalanobis Distance* as shown next:

$$D_{MRF}(\eta_P) = \sum_{i \in \eta_P} (\mathbf{y}(i) - \mu(\eta_P))\Gamma^{-1}(\mathbf{y}(i) - \mu(\eta_P))^T \tag{4}$$

where Γ is the covariance matrix, $\mathbf{y}(i)$ is the row vector $(y(i, 1), .., (y(i, r))$ and $\mu(\eta_P)$ is the row vector denoting the mean of $\mathbf{y}(i)$ in node η_P. Note that the covariance between output responses Γ can be considered as dependent on the samples at the node (i.e. $\Gamma(\eta_P)$) but we have decided to use the covariance generated from the initial training samples.

The inverse covariance matrix (Γ^{-1}) is a precision matrix which provides a measure of conditional dependence between multiple random variables. The Mahalanobis distance square normalizes the output responses by their standard deviations and in case of Γ being diagonal, it represents the normalized Euclidean distance.

The MRF provides prediction for multiple output responses as compared to single response in case of RF. When the drugs are related due to similarity of action (such as common targets for targeted drugs) , the MRF can incorporate that correlation to improve the prediction accuracy as illustrated through examples in results section.

2 Error Estimation Techniques

2.1 Preprocessing of the Data

Dream challenge drug sensitivity data was normalized by dividing it with the maximum value. The AUC values of CCLE dataset has been normalized by dividing the AUC measures by the number of drug concentrations tested which was eight.

2.2 Training (Re-substitution) Error

Re-substitution is the most basic approach that uses the training samples as the testing samples. Since the testing samples have already been used to train the model, re-substitution estimates will provide an optimistic view of the model accuracy. In small sample scenarios (< 100 samples), re-substitution estimates does not further reduce the number of limited available samples for training and can potentially be used for model selection. However, we need to be careful about over-fitting as re-substitution estimates will keep improving when we increase the number of features in the model.

2.3 K-fold Cross Validation and Leave One Out Error Estimation

In K-fold Cross Validation accuracy estimate, we partition the data randomly into K distinct folds of approximately equivalent sizes. We train the model on $K - 1$ folds and test on the remaining fold

and repeat it K times each time selecting a different fold to hold out for testing. The final K-fold Cross Validation accuracy estimate is the average accuracy over the K testing folds. *Leave One Out* error estimate is a special form of K-fold Cross Validation where K = number of samples i.e. we leave one sample out for testing at each iteration.

2.4 Bootstrap & 0.632+ Bootstrap Error Estimation

Bootstrap [1] accuracy estimation considers sampling with replacement while generating the training samples. Consider that the samples $X_1, X_2, \cdots, X_n$ belong to an underlying distribution and we want to train a model on that distribution. Bootstrap considers sampling from the underlying empirical distribution represented by $X_1, X_2, \cdots, X_n$. Thus, n training samples are generated from the samples $X_1, X_2, \cdots, X_n$ using sampling with replacement and the samples that are not selected are used for testing the model accuracy. This procedure is repeated B number of times and average accuracy estimated. B is usually recommended to be between 25 and 200 [2].

The **bias** of an estimator is the difference between the expected value and the true value of the parameter being estimated. The Bootstrap error estimator can be upward biased i.e. the error estimate can be higher than the actual value. On the other hand, re-substitution error estimate can be downward biased i.e. the error estimate can be lower than the true value. The **0.632 Bootstrap** method [2,3] considers a combination of these two estimators (one upward biased and one downward biased) to potentially arrive at an unbiased estimator. The **0.632 Bootstrap** error estimate is given by

$$\alpha_{.632Boot} = 0.632 \cdot \alpha_{boot} + 0.368 \cdot \alpha_{resub} \tag{5}$$

Due to the contribution of re-substitution error, 0.632 Bootstrap estimator can be over fitted. 0.632+ Bootstrap error estimator [3] considers the amount of over fitting and assigns higher weight on bootstrap error estimator in the over fitted case. The scale of over fitting is calculated using no-information error rate, λ which can be estimated by permuting responses Y_i and predictors X_j

$$\lambda = \sum_{i=1}^{N}\sum_{j=1}^{N} Error[Y_i, X_j]/N^2 \tag{6}$$

The relative over fitting rate R is calculated using λ and subsequently the realtive weight (ω) of the bootstrap error estimator is derived.

$$R = \frac{Bootstrap\ Error - Resubstitution\ error}{\lambda - Resubstitution\ error} \tag{7}$$

$$\omega = \frac{0.632}{1 - 0.368R} \tag{8}$$

Thus, the **0.632+ Bootstrap** error estimate is given by

$$\alpha_{.632+Boot} = \omega \cdot \alpha_{boot} + (1 - \omega) \cdot \alpha_{resub} \tag{9}$$

2.5 True Error Estimation

The true error of a model has been estimated by using hold out data that has not been applied for any form of training or previous error estimations. For the DREAM dataset, a separate set of 18 cell line data is used for estimating true error. For the CCLE dataset, 40 or 100 cell lines out of 400 were used for training and error estimation and remaining cell lines were used for true error estimate.

2.6 Jackknife-After-Bootstrap Confidence Interval Generation

Jackknife-After-Bootstrap approach [1] is used for generating the confidence intervals of 0.632 Bootstrap errors. Let, N_k denote the set of bootstrap samples that does not contain sample X_k and the 0.632 bootstrap estimate computed from N_k is denoted by ϵ_k. The standard error can be computed as

$$s = \sqrt{\frac{n-1}{n}\sum_{k=1}^{n}(\epsilon_k - \bar{\epsilon})^2} \tag{10}$$

where $\bar{\epsilon} = (1/n)\sum_{k=1}^{n} \epsilon_k$. The $100(1-\alpha)\%$ prediction intervals for the true error can be computed as $[\bar{\epsilon} - sz_{\alpha/2}, \bar{\epsilon} + sz_{\alpha/2}]$ where z_α is the α quantile of the standard normal distribution. Since we consider the absolute error, the lower bound of the confidence interval is calculated as max $[0, \bar{\epsilon} - sz_{\alpha/2}]$.

3 Integration of Models

We approached the integration of prediction from individual models of different genetic characterizations as a linear regression problem. Let $\Omega_i(j)$ denote the prediction obtained by Random Forest approach for genomic characterization dataset G_i and cell line j. The weight α_i for each dataset G_i is obtained by minimizing

$$\sum_j (y_j - \sum_i \alpha_i \Omega_i(j))^2 \tag{11}$$

where y_j is the experimental drug response for cell line j and α_i is the corresponding weight of dataset G_i.

Following the generation of the weights of the individual datasets, the combined prediction result $\hat{Y}_C(j)$ is generated as follows [4]:

$$\hat{Y}_C(j) = \sum_i \alpha_i \Omega_i(j) \tag{12}$$

4 Datasets

4.1 Dream Challenge Dataset

For the NCI-DREAM Drug Sensitivity prediction sub-challenge 1, genomic characterizations were provided for 53 cell lines (48 breast cancer cell lines and 5 non-malignant breast cell lines) that were exposed to 31 therapeutic compounds at a concentration required to inhibit proliferation by 50% after 72 hours (GI50) and responses to these 31 drugs were provided for 35 of these 53 cell lines. The drug response of the remaining 18 cell lines were provided later and used as validation dataset. Multiple types of genomic and epigenetic data (copy number variation, methylation, gene expression through micro-array, RNA sequencing, exome sequencing and protein abundance) were generated before exposure of the cells to the drugs for each of the 53 cell lines [5]. Details of the genomic characterization datasets are provided in Table 1. From table 1, we note that the genomic characterizations were not available for all the 53 cell lines and each dataset had missing information for some of the cell lines (the number of such cell lines is denoted by Missing cell lines in table 1). The last column denotes whether the genomic dataset had some missing values for the cell lines containing that specific genomic characterization.

Table 1: Description of Genomic Datasets for NCI-DREAM drug sensitivity challenge. Out of 53 cell lines, 35 cell lines are used for training and 18 for testing the prediction accuracy

Data Type	**Dimension**	**Missing cell lines**	**Missing values(Y/N)**
Gene Expression	46×18632	7	N
Methylation	41×27551	12	N
RNa seq	44×36953	9	Y
RPPA	42×131	11	N
SNP6	47×27234	6	Y

4.2 CCLE Dataset

Cancer Cell Line Encyclopedia(CCLE) dataset was downloaded from "http://www.broadinstitute.org/ccle/home". CCLE dataset consists of two forms of genetic characterizations (i) Gene Expression and (ii) Single Nucleotide Polymorphisms (SNP6). The corresponding file for Gene expression dataset is *CCLE_Expression_Entrez_2012-09-29.gct*. In this dataset, there are 18988 gene features for 1037 cell lines with no missing values. The SNP6 data has been extracted from *CCLE_copynumber_byGene_2013-12-03.txt*. For 1043 cell lines, there are 23316 features. For our computations, we have selected 1012 cell lines that are common to both gene expression and SNP6.

Drug sensitivity data has been downloaded from the addendum published by Barretina et al [6] where 24 drug responses are recorded for 504 cell lines. The dose-response data at eight concentrations were reduced to a fitted model using a decision tree methodology based on NIH/NCGC assay guidelines (http://assay.nih.gov/assay/index.php/Table_of_Contents). For our predictions, we have considered *Area Under the Curve* (AUC) as the drug sensitivity data.

5 Results

5.1 Model Performance for different Genetic Characterization Combinations for NCI-DREAM using RF

From the NCI-Dream challenge dataset, we consider 5 genomic characterizations (Gene expression, Methylation, RNA seq, RPPA and SNP6). Based on 5 different genomic characterizations, there are $(2^5 - 1) = 31$ possible nonempty combinations. Figure 2 shows 5 different error estimates (Leave one out, 5-Fold Cross Validation, Bootstrap and 0.632+ Bootstrap) and validation errors for the 31 different combinations of datasets for a specific drug in the DREAM challenge dataset. The integration of prediction has been achieved through linear regression over Random Forest models. The different error estimates have variations across the different dataset combinations but we do **observe an overall trend of the error estimators producing smaller values when higher number of datasets are being used for prediction** (leftmost results with only one dataset and the rightmost with all five datasets).

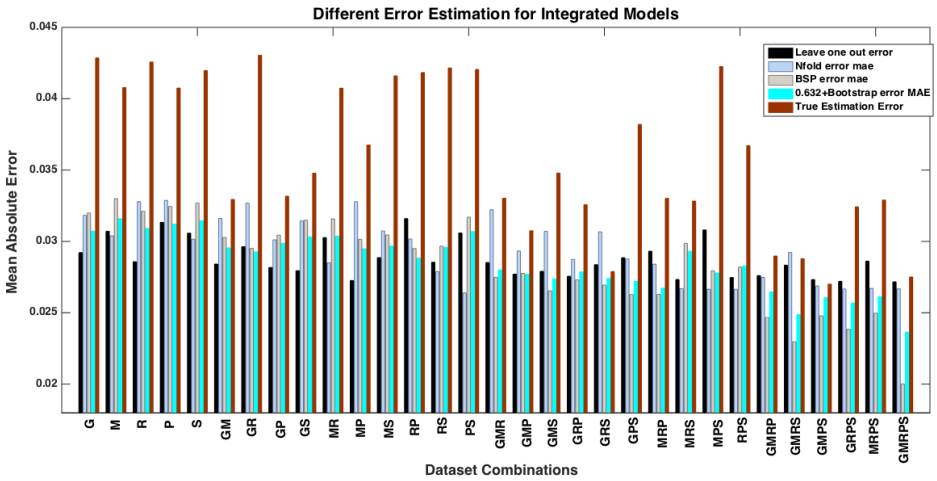

Figure 2: Leave-one-out, 5 fold Cross Validation, Bootstrap, 0.632+ Bootstrap and Validation error for Drug-10 for different dataset combinations.The datasets are denoted by: G: Gene Expression, M:Methylation, R:RNASeq, P:RPPA and S:SNP6. For instance, MR denotes Methylation and RNASeq data combination.

Figure 3 shows the 80% Confidence Interval for drug-15 along with bootstrap error estimate (squares) for 31 different dataset combinations. From figure 3, we observe that the **addition of dataset reduces the confidence interval**. The confidence interval is highest when using only one dataset and the confidence interval decreases gradually with addition of more datasets (going right). The lowest confidence interval is shown by the rightmost bar which represents the confidence interval of 0.632 bootstrap error while using all 5 datasets for prediction.

5.2 Comparing Integrated Prediction Options for NCI-DREAM dataset

In the previous section, we have used all the features for each dataset for a data type prediction and combined the predictions using linear regression. In this section, we analyze the effect of feature selection on the performance along with the effect of concatenating all the features before model generation which is in contrast to earlier results where we generated individual models for each dataset and integrated the predictions from these individual models. Figure 4 shows the Mean

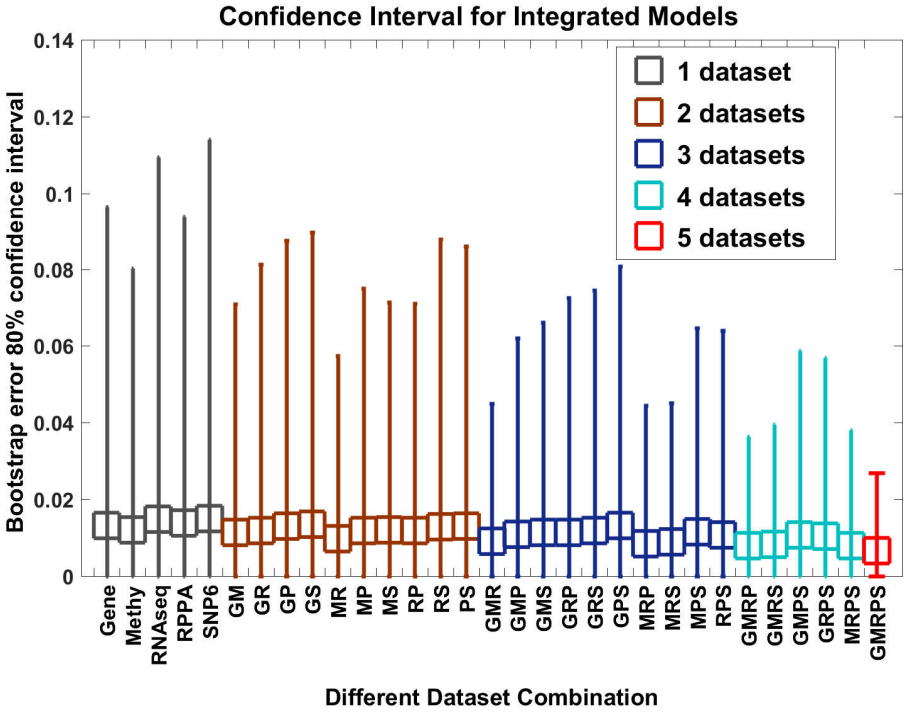

Figure 3: Mean Bootstrap error and 80% confidence intervals for Drug 15 for $31(=2^5-1)$ different dataset combinations. The datasets are denoted by: G: Gene Expression, M:Methylation, R:RNASeq, P:RPPA and S:SNP6. For instance, MR denotes Methylation and RNASeq data combination.

Absolute Validation Error for integrated model with or without feature selection and with or without dataset concatenation for Drug-28.

Based on Fig. 4, we observe that concatenating the datasets without feature selection may increase the Mean Absolute validation error (red bars). The differences between single RF over appended datasets (last three bars black, yellow, red) and integrated RFs (gray and blue) are more pronounced when larger number of datasets are used. As compared to our previous results (gray bars), a minor improvement was achieved by applying feature selection in individual datasets and generating individual models for each dataset for final integrated model (blue bars).

The reported results allude to the observation that **it is better to generate individual predictive models for each dataset and then combine them to form an integrated model rather than designing a single model from a concatenated dataset**.

5.3 Predictions for integrated models using CCLE Dataset

There are 2 genetic characterization information in CCLE, gene expression and SNP6, resulting in $2^2-1=3$ possible dataset combinations. For the three possible combinations, we have calculated the 0.632 bootstrap error and confidence interval, Leave-one-out error and validation error. Figure 5 shows the different error estimates for drug *Saracainib* when 100 random samples from the CCLE dataset were used for training and remaining 404 samples were used to generate the validation or true error. **Similar to DREAM challenge results, we observe a reduction in error when datasets are combined for sensitivity prediction**. Note that due to the availability of only two forms of genetic characterizations, the reduction in average error is significantly lower than what was observed in DREAM challenge dataset using five datasets.

5.3.1 Effect of Sample Size on Confidence Interval

The number of cell lines selected for generating the random forest can have an impact on the errors when sample size is limited. For instance, figure 6(a) shows the 0.632BSP error and confidence interval with training set of 100 cell lines for *Saracatinib*, while figure 6(b) shows that the confidence Interval reduces when 250 samples are used for training. **The results indicate a reduction in confidence interval with larger number of training samples.**

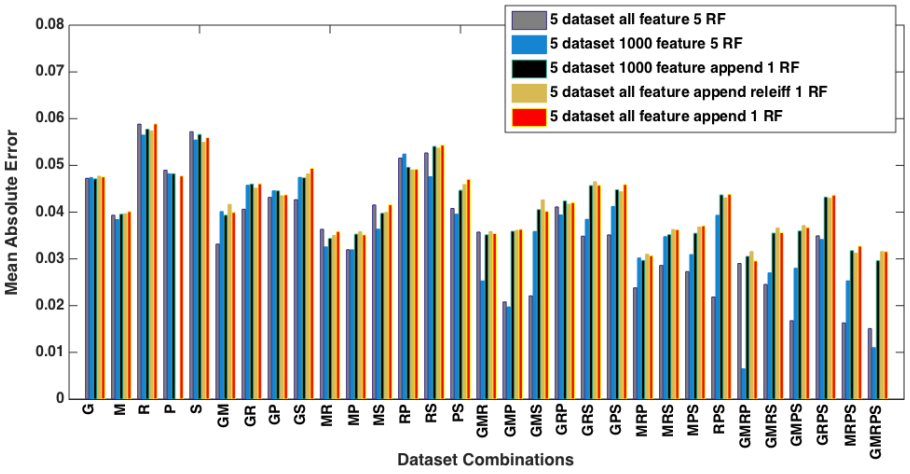

Figure 4: Mean Absolute Error for validation set for Drug-28 with different integrated models. These models were build using different dataset combinations along with following ways of feature selection and model generation (i) *5 dataset all feature 5 RF*: All features of the dataset are used and random forests designed for each individual dataset that are subsequently integrated using linear regression (ii) *5 dataset 1000 features 5 RF*: Important 1000 features of each dataset are selected using RELIEFF feature selection and RF models designed for each dataset that are subsequently integrated using linear regression (iii) *5 dataset 1000 feature append 1 RF*: Important 1000 features of each dataset selected using RELIEFF and appended and a single RF designed from the appended dataset. (iv) *5 dataset all feature append relieff 1 RF*: All features of all datasets appended and important 1000 features selected using RELIEFF from the appended dataset. A single RF designed for the features selected from the appended dataset. (v) *5 dataset all feature append 1 RF*: All features of all datasets appended and a single random forest designed from this combined dataset

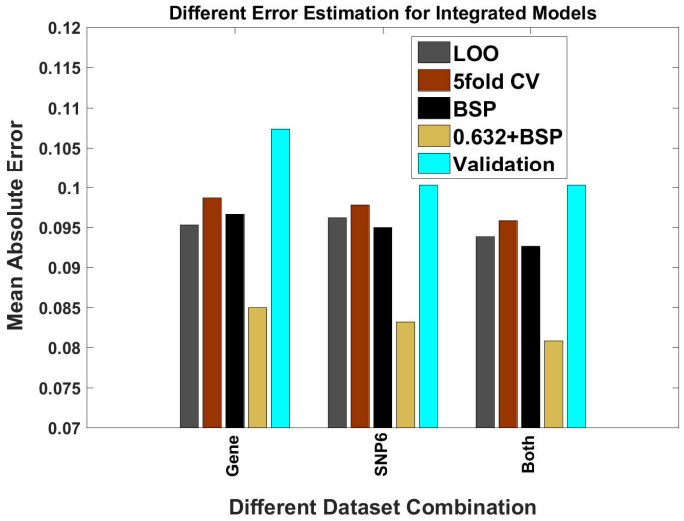

Figure 5: Leave-one-out, 5-fold cross validation, Bootstrap, 0.632 Bootstrap+ and Validation error for drug *Saracainib* (gene expression, SNP6 and combined dataset from left to right respectively).

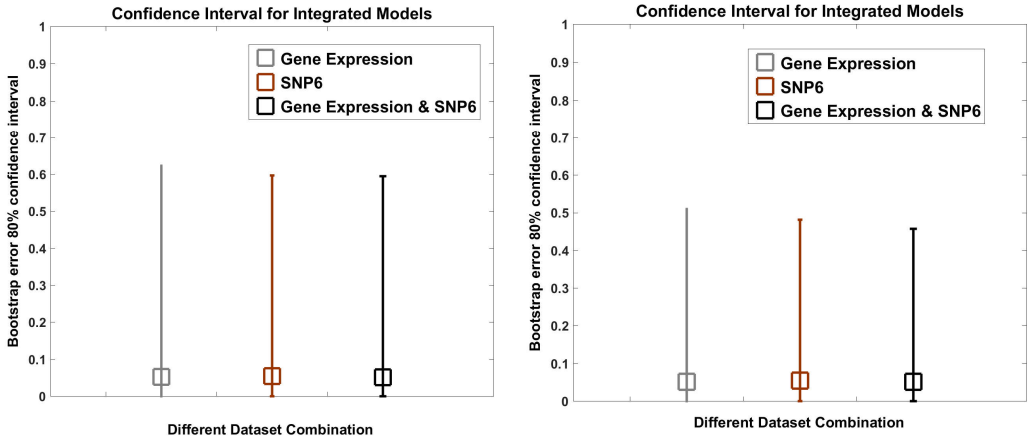

Figure 6: (a) 80% confidence interval of 0.632 Bootstrap error for drug *Saracainib* with 100 cell lines for building integrated random forest. (b) 80% confidence interval of 0.632 Bootstrap error for drug *Saracainib* with 250 cell lines for building integrated random forest (gene expression, SNP6 and combined dataset from left to right respectively)

5.3.2 Error Estimation Performance using CCLE dataset

The CCLE dataset consists of 24 drugs. For each drug, the number of cell lines with genetic characterization data and pharmacological data varies between 350 to 490. We utilize 100 cell lines to build the integrated random forest model along with estimation of Leave-one-out, 5-fold cross validation, Bootstrap and 0.632 Bootstrap+ errors. The remaining cell lines of that drug are used as holdout data to calculate the validation or true error. Figure 7 shows the Leave-one-out, 5-fold cross validation, Bootstrap, 0.632+ Bootstrap and Validation error estimates in the form of Mean Absolute Error for all the 24 drugs. The Bootstrap errors were calculated with $B = 40$. Figure 7 shows that the value of LOO error is higher than 0.632+ Bootstrap error for all the drugs but when compared to validation (true) error, **LOO is closer to true error as compared to 0.632+ bootstrap for the CCLE 100 sample training scenario**. However, Leave One Out error estimation can be extremely computationally intensive when the number of samples increases as a separate model has to be trained for each additional sample. For sample size greater than 50 or 100, 5 fold cross validation or Bootstrap error estimates provides comparable results with less computational complexity.

The 80% confidence interval of 0.632 bootstrap error for the integrated model is shown in figure 8.

5.4 Multivariate Random Forest Performance

NCI-Dream Challenge Dataset

For the Dream Challenge data, we have considered 5 genetic characterizations. There are 31 drugs resulting in $\binom{31}{2} = 465$ pairwise combinations. Drug-13 had standard deviation of zero for the common cell lines and thus we excluded these 30 combinations from the 465 possible pairwise combinations.

High Correlation among drug pairs: Table 2 shows the performance comparison for RF and MRF for drug pairs that have the highest correlations. We observe that the **performance of MRF in terms of both correlation coefficient and MAE is better than RF for drug pairs that have high correlations**. Table 3 shows the MAE for individual datasets and integrated model for RF and MRF for the pair of drugs: drug 1 and drug 2. We note that the MAE decreases when MRF is used as compared to RF.

Low Correlation among drug pairs: In this section, we consider the comparison of RF and MRF performance for drug pairs that have low correlations. We expect that the performance of MRF will be poor for the scenario where the drug sensitivities are not correlated. The considered drug pairs in this section are the ones with the lowest absolute correlation coefficients among the drug response pairs. The results for these drug pairs are shown in Table 4. We observe that the **performance of MRF is worse than the performance of RF for the drug pairs that have**

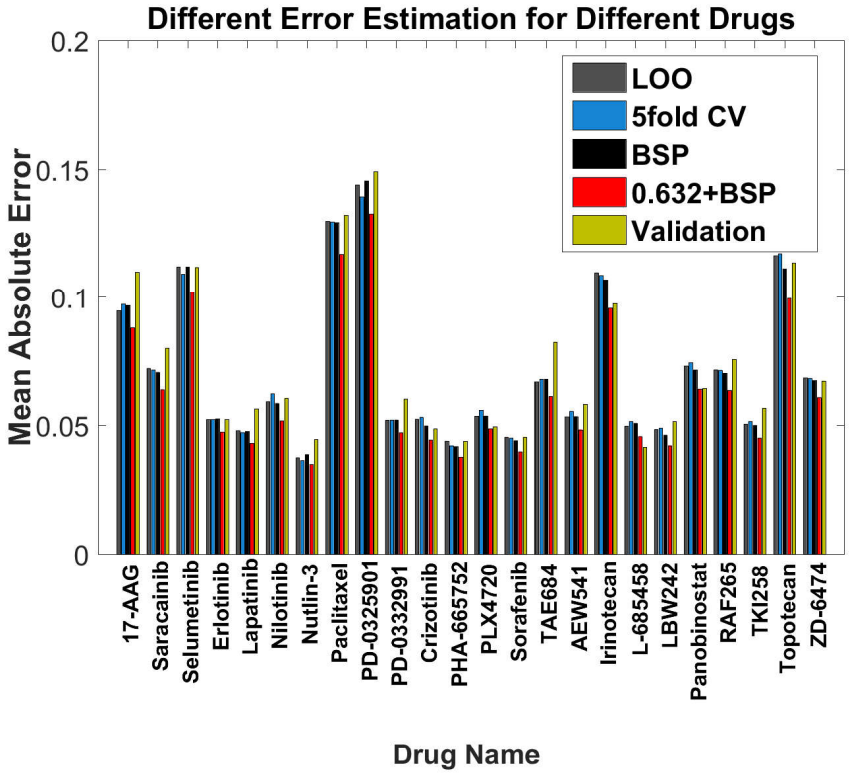

Figure 7: Leave-one-out, 5-fold cross validation, Bootstrap, 0.632 Bootstrap+ and Validation error for all 24 drugs for a certain run. For different runs, the cell lines selected for training can change significantly but we observed that the change in the errors are minimal.

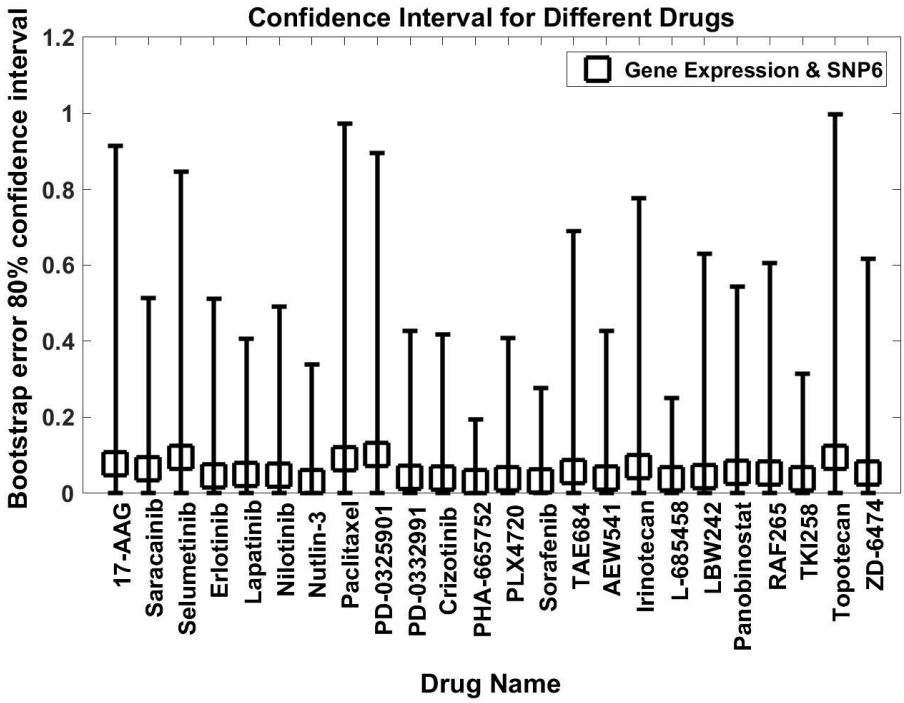

Figure 8: 80% confidence interval of 0.632 Bootstrap error for all 24 drugs of CCLE dataset. The square indicates the 0.632 bootstrap error.

Table 2: 5 fold CV results for DREAM challenge dataset drug sensitivity prediction for ten drug pairs with HIGH correlation in the form of correlation coefficients between predicted and experimental values and Mean Absolute Errors (MAE). RF and MRF denotes regular Random Forest and Multivariate Random Forest respectively. Results are for fixed number of trees = 100, feature size = 10 and leaf size = 2.

			Correlation coefficient		MAE	
Drug Set	Drug Name	Correlation	RF	MRF	RF	MRF
S_{C1}	Drug-1 Drug-2	0.8959	0.1515 0.4175	0.4000 0.5811	0.0301 0.0241	0.0295 0.0203
S_{C2}	Drug-8 Drug-18	0.7313	0.3404 0.4979	0.3724 0.5650	0.0460 0.0320	0.0445 0.0282
S_{C3}	Drug-30 Drug-31	0.7573	0.2984 0.3015	0.3342 0.3522	0.0287 0.0298	0.0250 0.0263
S_{C4}	Drug-19 Drug-23	0.6693	0.4841 0.4408	0.5435 0.5667	0.0178 0.0165	0.0162 0.0154
S_{C5}	Drug-23 Drug-28	0.6646	0.7208 0.4313	0.7649 0.4836	0.0125 0.0156	0.0122 0.0136
S_{C6}	Drug-19 Drug-28	0.6413	0.5295 0.5802	0.7328 0.7121	0.0180 0.0138	0.0127 0.0115
S_{C7}	Drug-14 Drug-16	0.6265	0.5871 0.5442	0.5916 0.7183	0.0506 0.0256	0.0456 0.0217
S_{C8}	Drug-1 Drug-21	0.6212	0.2674 0.3612	0.4759 0.4287	0.0255 0.0289	0.0246 0.0256
S_{C9}	Drug-8 Drug-22	0.6126	0.5973 0.2240	0.7056 0.5986	0.0401 0.0133	0.0323 0.0106
S_{C10}	Drug-2 Drug-21	0.6021	0.5594 0.1414	0.6476 0.2857	0.0857 0.0654	0.0836 0.0657

Table 3: Mean Absolute Error of Single Dataset and integrated model using all 5 datasets.

Drug Name	Type	GENE	Methylation	RNAseq	RPPA	SNP6	Integrated Model
Drug 1	RF	0.0346	0.0327	0.0350	0.0355	0.0318	0.0301
Drug 2	RF	0.0298	0.0285	0.0347	0.0335	0.0306	0.0241
Drug 1 Drug 2	MRF	0.0327	0.0295	0.0329	0.0345	0.0322	0.0295 0.0203

low correlations.

Table 4: 5 fold CV results for DREAM challenge drug sensitivity prediction dataset for five drug sets with LOW correlation in the form of correlation coefficients and Mean Absolute Errors (MAE). RF and MRF denotes regular Random Forest and Multivariate Random Forest respectively. Results are for fixed number of trees = 100 and feature size = 10 and leaf size= 2.

			Correlation coefficient		MAE	
Drug Set	Drug Name	Correlation	RF	MRF	RF	MRF
S_{C11}	Drug-16	1.84E-04	0.3643	0.3471	0.0239	0.0254
	Drug-28		0.6304	0.5527	0.0120	0.0130
S_{C12}	Drug-15	8.08E-04	0.3125	0.2545	0.0156	0.0172
	Drug-22		0.2693	0.2095	0.0121	0.0131
S_{C13}	Drug-9	2.30E-03	0.5553	0.2672	0.0174	0.0217
	Drug-21		0.4345	0.1639	0.0059	0.0067
S_{C14}	Drug-6	6.90E-03	0.6104	0.1919	0.0129	0.0147
	Drug-30		0.4107	0.0147	0.0223	0.0247
S_{C15}	Drug-17	8.50E-03	0.4550	0.4345	0.0477	0.0475
	Drug-29		0.7236	0.4763	0.0153	0.0188

For the implementation of RF and MRF, we have considered all features of the genomic characterizations. When we have applied feature selection or concatenated all the features of all datasets, we did not observe any improvement in performance.

CCLE Dataset

The CCLE [7] database includes genomic characterization for 1037 cell lines and drug responses over 24 drugs for around 400 to 490 cell lines. For comparison between RF, MRF, Elastic Net (EN) and Kernelized Bayesian multitask learning (KBMTL) approaches for predicting responses, 4 sets of drugs were selected. The first set $S_{C1} = \{$ *Selumetinib*, *PD-0325901* $\}$ has *MET* as a common target, the second set $S_{C2} = \{$ *Erlotinib*, *Lapatinib* $\}$ has *EGFR* as a common target, the third set $S_{C3} = \{$ *ZD6474*, *AZD0530* $\}$ has *EGFR* as a common target and the fourth set $S_U = \{$ *17-AAG*, *AEW541* $\}$ has no common target.

Each cell line had initially 18,988 features (probesets) as gene expressions. We reduced it to 500 for each drug response using RELIEFF feature selection and used a union of the 500 features in each of the four sets of drugs. We have used all the cell lines that have gene expression and drug responses for specific pairs of drugs. To report our results, we compared 5 fold cross-validated Pearson correlation coefficients between predicted and experimental responses for RF and MRF. For both RF and MRF, we set the minimum size of samples in each leaf to $n_{size} = 5$, the number of trees in the forest to $T = 200$ and the splitting in each node considers $m = 10$ random features.

The correlation coefficients using 5 fold cross validation error estimation are illustrated for each drug set in table 5. Table 5 shows that **MRF performed better than Elastic Net (EN), Kernelized Bayesian multitask learning (KBMTL) and RF for the related drug pairs** S_{C1}, S_{C2}, S_{C3}. **When there is no relationship in the drug pair as in** S_U**, univariate RF performs better than the MRF approach**.

GDSC Dataset

The Genomics of Drug Sensitivity for Cancer (GDSC) dataset [8] includes genomic characterizations for 700 cell lines and drug responses for 140 drugs. For comparing multivariate Random Forest (MRF) with Elastic Net (EN), Kernelized Bayesian multitask learning ($KBMTL$) and Random Forest (RF), we considered ten sets of drug pairs with the highest correlation coefficients between the responses and five sets of drug pairs with the lowest correlation coefficients between the responses. The correlation coefficients using 5 fold cross validation error estimation are illustrated for each drug set in tables 6 and 7. Similar to earlier results, we observe that MRF performs better than RF, EN and KBMTL for drug pairs that have high correlation coefficient between the responses (S_1 to S_{10})

Table 5: 5 fold Cross validation results for CCLE dataset drug sensitivity prediction. Correlation co-efficients between actual and predicted drug sensitivity using Elastic Net (EN), Kernelized Bayesian multitask learning ($KBMTL$), Random Forest (RF) and Multivariate Random Forest (MRF) are reported here.

			Correlation Co-efficients			
Drug Set	Common Target	Drug Name	EN	$KBMTL$	RF	MRF
S_1	MEK	Selumetinib	0.38	0.45	0.52	0.55
		PD-0325901	0.36	0.48	0.58	0.60
S_2	EGFR	Erlotinib	0.28	0.38	0.39	0.41
		Lapatinib	0.34	0.41	0.44	0.46
S_3	EGFR	ZD6474	0.23	0.29	0.32	0.34
		AZD0530	0.25	0.24	0.29	0.30
S_U	None	17-AAG	0.29	0.40	0.37	0.36
		AEW541	0.27	0.32	0.38	0.38

whereas RF performs better than MRF for drug pairs with low correlation coefficient between the responses (S_{11} to S_{15}).

Table 6: 5 fold Cross validation results for GDSC dataset drug sensitivity prediction for ten drug sets with HIGH correlation. Correlation coefficients between actual and predicted drug sensitivity using Elastic Net (EN), Kernelized Bayesian multitask learning ($KBMTL$), Random Forest (RF) and Multivariate Random Forest (MRF) are reported here.

			Correlation Co-efficients			
Drug Set	Common Target	Drug Name	EN	$KBMTL$	RF	MRF
S_1	0.8439	RDEA119	0.6161	0.5669	0.6345	0.6606
		PD-0325901	0.4815	0.4679	0.6129	0.6334
S_2	0.8410	BI-2536	0.2277	0.2289	0.2575	0.2755
		GW843682X	0.2979	0.2755	0.3144	0.3322
S_3	0.8366	CI-1040	0.4609	0.5106	0.5867	0.6020
		PD-0325901	0.5060	0.5211	0.6245	0.6495
S_4	0.8209	RDEA119	0.4942	0.5371	0.6365	0.6539
		CI-1040	0.4647	0.5168	0.5879	0.6026
S_5	0.8175	Paclitaxel	0.2038	0.2935	0.3116	0.3301
		BI-2536	0.0390	0.0365	0.2313	0.2507
S_6	0.8156	Doxorubicin	0.3524	0.3957	0.4204	0.4348
		Etoposide	0.4136	0.3649	0.4583	0.4701
S_7	0.8047	Dasatinib	0.5886	0.5546	0.6788	0.7013
		WH-4-023	0.5228	0.5825	0.5925	0.6103
S_8	0.7993	Paclitaxel	0.1962	0.3075	0.3593	0.3738
		GW843682X	0.2751	0.2718	0.3391	0.3569
S_9	0.7634	Doxorubicin	0.3690	0.3249	0.4275	0.4422
		Epothilone B	0.4101	0.3417	0.4342	0.4527
S_{10}	0.7543	Dasatinib	0.6317	0.5927	0.6753	0.6903
		A-770041	0.4582	0.4841	0.4645	0.5093

5.5 Biological Validation and Variable Importance Measure (VIM)

We have examined the variable importance measure for Random Forest and Multivariate Random Forest models trained on GDSC database in terms of protein interaction network enrichment analysis. In this section, we will primarily provide the detailed results for Drug pair AZD0530 and Erlotinib of GDSC.

For MRF, the top 100 features or probe-sets were estimated based on the frequency of their being selected in multivariate regression tree generation. For the individual RF models for two separate drugs, 100 top ranked probe-sets were estimated separately. Note that that multiple probe-set IDs can map to a single Gene Symbol of a protein. We have utilized *HG-U133A Plus 2* of Affymetrix for

Table 7: 5 fold Cross validation results for GDSC dataset drug sensitivity prediction for five drug sets with LOW correlation. Correlation coefficients between actual and predicted drug sensitivity using Elastic Net (EN), Kernelized Bayesian multitask learning ($KBMTL$), Random Forest (RF) and Multivariate Random Forest (MRF) are reported here.

			Correlation Co-efficients			
Drug Set	Common Target	Drug Name	EN	$KBMTL$	RF	MRF
S_{11}	6.59E-07	Mitomycin C	0.2808	0.2525	0.3744	0.3818
		Axitinib	0.3126	0.3270	0.3609	0.3166
S_{12}	1.76E-05	JW-7-52-1	0.1568	0.2220	0.3708	0.3605
		Methotrexate	0.6752	0.5078	0.6954	0.6890
S_{13}	8.62E-05	Lapatinib	0.4915	0.4938	0.5918	0.5547
		Shikonin	0.2329	0.3607	0.3990	0.4009
S_{14}	9.74E-05	Vinorelbine	0.3191	0.3287	0.3276	0.3606
		Lenalidomide	0.0059	0.0113	0.2826	0.2300
S_{15}	1.12E-04	AZD-2281	0.1768	0.2439	0.4202	0.4146
		PD-0325901	0.5165	0.5138	0.6010	0.6096

mapping the probe-sets into proteins. Based on this mapping, we arrived at 94 top ranked proteins for RF1, 89 top ranked proteins for RF2 and 94 top ranked proteins for MRF. The enrichment analysis was conducted using STRING-db database (http://string-db.org/) and the p-value for the network based on top 100 features of MRF model is less than the p-values of the individual networks generated from the RF top 100 features as shown in Table 8 illustrating higher connectivity between the MRF selected features as compared to RF model selected features. Table 8 also shows that similar conclusion will be obtained based on the network features of clustering coefficient (MRF clustering coefficient is higher) and ratio of observed to expected node edges (ratio for MRF is higher as compared to individual RF models).

Table 8: Biological Network Enrichment Analysis for RF and MRF model selected proteins

	RF model of AZD0530	RF model of Erlotinib	MRF model
P-value	4.44e-16	6.05e-11	1.11e-16
Clustering Coefficient	0.895	0.893	0.935
Ratio of observed to expected nodes	2.32	2.06	2.54

The protein-protein interaction (PPI) networks or network connectivity graphs for top proteins using RF1, RF2 and MRF are shown in Figures 9,10 and 11 respectively.

6 Package Application Example

To provide an overview of the *IntegratedMRF* package, a practical implementation is shown next using NCI-DREAM Challenge Dataset. The description of the dataset has been provided in earlier Dataset section and it has been attached as a demo data in the package with reduced number of predictor features. The lines in **Bold** and *Italic* represent comments and code respectively.

Set the working directory, which also contains the dataset
library(IntegratedMRF)#**Call the package**
Drug=c(1,2,10)#**Number of output responses that the user wants to model in a multivariate form**
n_tree=10#**Number of trees in the forest**
m_feature=5#**Number of randomly selected features considered for a split in each regression tree node**

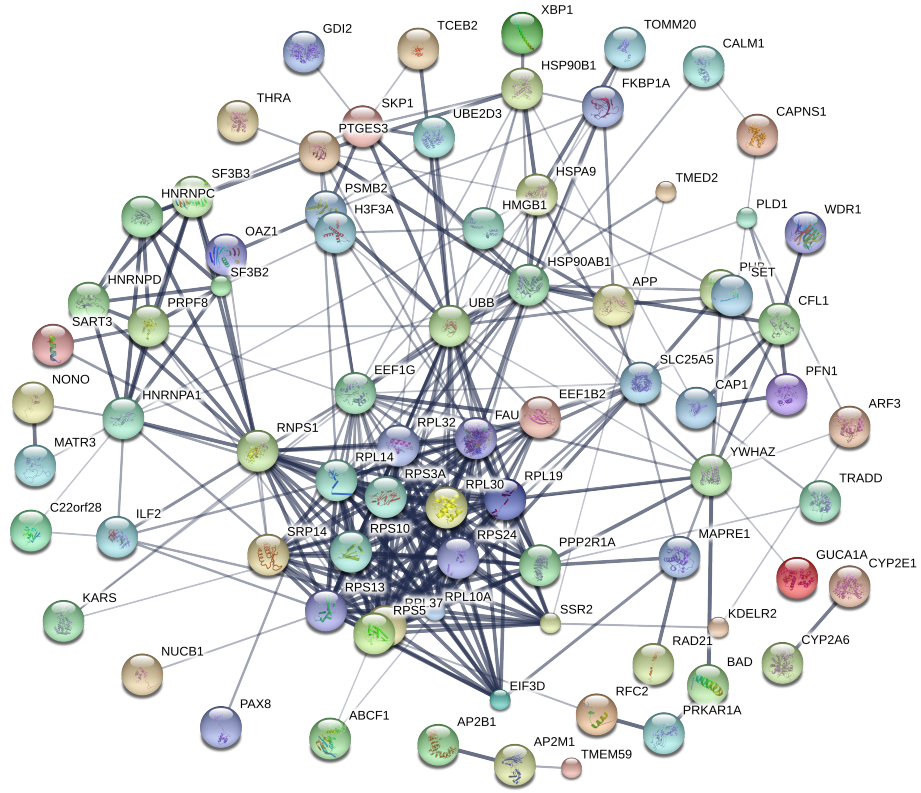

Figure 9: Protein-protein interaction network observed between top regulators generated from RF model of Drug AZD0530 in GDSC dataset

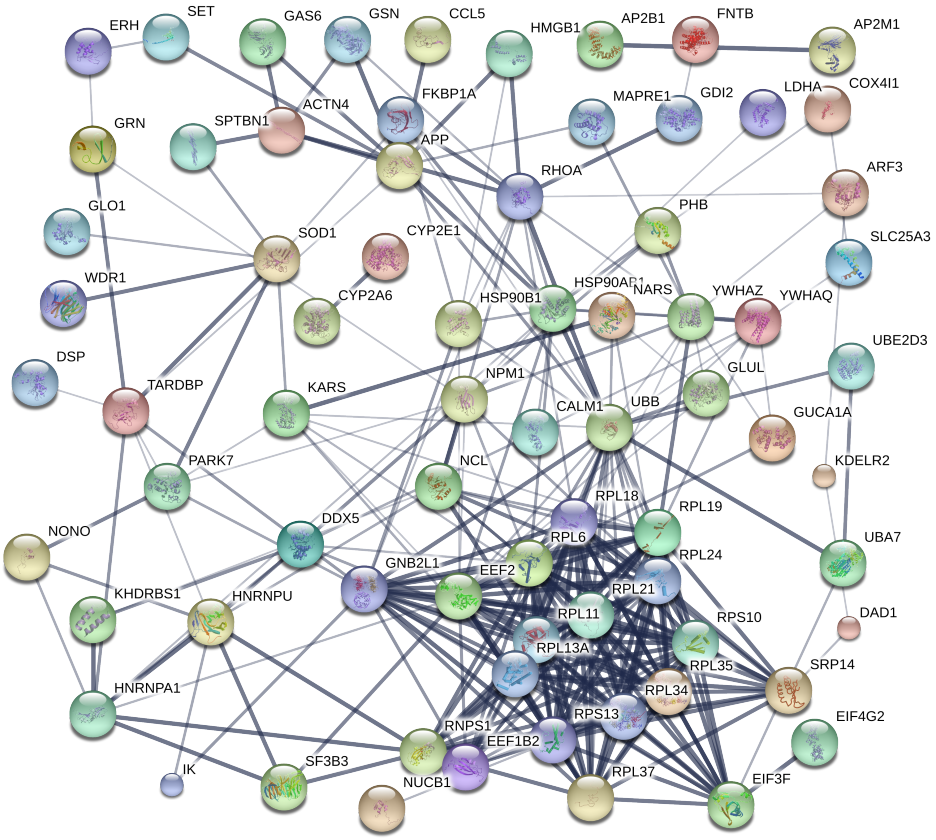

Figure 10: Protein-protein interaction network observed between top regulators generated from RF model of Drug Erlotinib in GDSC dataset

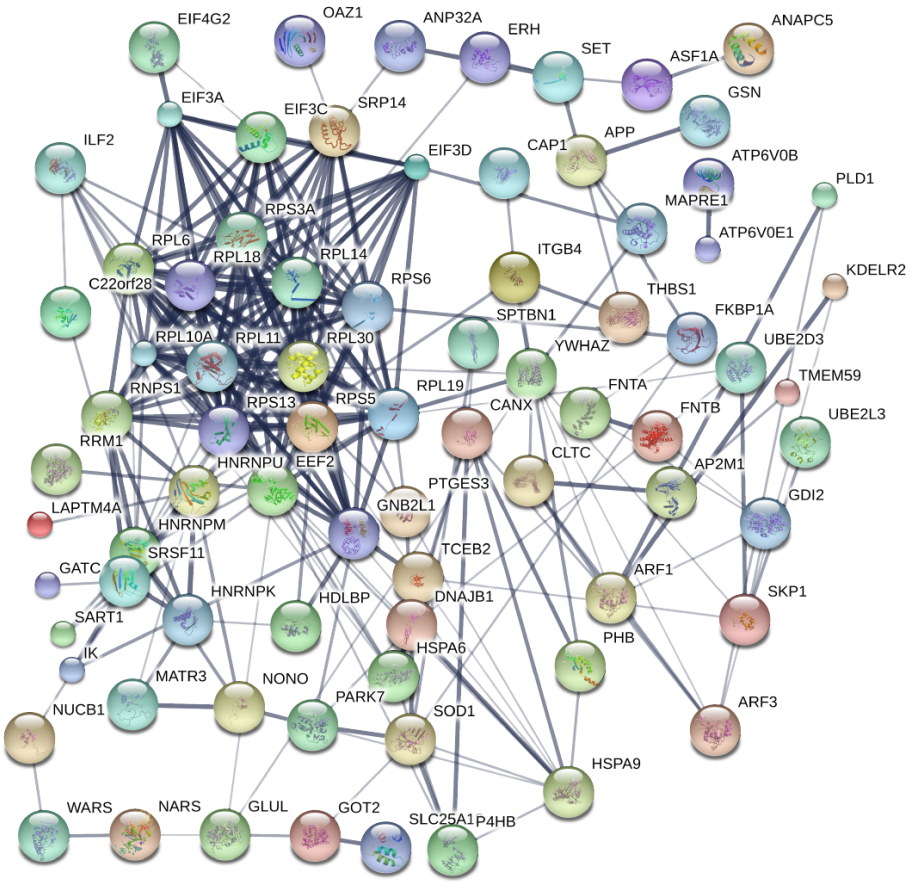

Figure 11: Protein-protein interaction network observed between top regulators generated from MRF model of Drug AZD0530 and Erlotinib in GDSC dataset

min_leaf=1#**Minimum number of samples in the leaf node**
Confidence_Level=80#**Confidence level for calculation of confidence interval**

data(Dream_Dataset) # **Loading Dream Dataset**
finalX_Dream=Dream_Dataset[[1]]#**List of matrices where each matrix represent a specific data subtype**
Cell_line_Index_Dream=Dream_Dataset[[2]]#**list of samples for each subtype of dataset**
finalY_train_Dream=Dream_Dataset[[3]]#**A 35 x 31 matrix of output features for training samples, where 35 is number of samples and 31 is the number of output features or drugs.**
finalY_train_cell_Dream=Dream_Dataset[[4]]#**Sample names of output features for training samples**
finalY_test_Dream=Dream_Dataset[[5]]#**A 18 x 31 matrix of output features for testing samples, where 18 is number of samples and 31 is the number of output features or drugs.**
finalY_test_cell_Dream=Dream_Dataset[[6]]#**Sample names of output features for testing samples**

finalY_train_Dream_Drug=matrix(finalY_train_Dream[,Drug],ncol=length(Drug))#**Taking the output responses of training samples of drugs which are user defined**
finalY_test_Dream_Drug=matrix(finalY_test_Dream[,Drug],ncol=length(Drug)) #**Taking the output responses of testing samples of drugs which are user defined**

#Combination: Calculates combination weights for different subtypes of dataset combinations to generate integrated Random Forest (RF) or Multivariate Random Forest (MRF) model based on different error estimates such as Bootstrap, N-fold cross validation, 0.632+ Bootstrap or Leave one out. Also calculates different errors for these error estimation methods and confidence interval. Refer to the package manual for function details.
Result=Combination(finalX_Dream,finalY_train_Dream_Drug,Cell_line_Index_Dream, finalY_train_cell_Dream,n_tree,m_feature,min_leaf,Confidence_Level)

#CombPredict: Generates Random Forest or Multivariate Random Forest model for each subtype of dataset and predicts testing samples using the generated models. Subsequently, the prediction for different subtypes of dataset are combined using the Combination weights generated from *Combination* **function. Refer to the package manual for function details.**
Prediction1=CombPredict(finalX_Dream,finalY_train_Dream_Drug,Cell_line_Index_Dream, finalY_train_cell_Dream,finalY_test_cell_Dream,n_tree, m_feature, min_leaf, Result[[1]])

#IntegratedPrediction: Generates Random Forest or Multivariate Random Forest model for each subtype of dataset and predicts testing samples using the generated models. Subsequently, the prediction for different subtypes of dataset are combined using the Combination weights generated from Integrated Model which is based on Bootstrap error estimate. Refer to the package manual for function details.
Prediction2=IntegratedPrediction(finalX_Dream,finalY_train_Dream_Drug,Cell_line_Index_Dream, finalY_train_cell_Dream,finalY_test_cell_Dream, n_tree, m_feature, min_leaf)

6.1 Parameter Selection Guidelines

The generation of the multivariate random forests involve several parameters that are critical to the performance and/or computation burden of the modeling approach and Table 9 provides guidelines for the selection of these important features.

6.2 Computational complexity in terms of expected time for potential scenarios

The computation time to run the package is dependent on a number of variables including number of datasets integrated, number of trees in the forest, number of features to split at each node, number of output responses, number of samples used and the type of error estimation being used. We provide the computation times for some potential scenarios in Table 10 so that a practitioner can generate

Table 9: Guidelines for selection of Parameters

Parameter	Range of valid values	Suggested values	Comment
n_tree	int $[1, \infty]$	100 *to* 500	Larger number of trees preferred for large feature set. The computational burden increases linearly with increase in the number of trees.
min_leaf	int [1, —sample size—]	3 *to* 10	For splitting, the minimum number of samples at a node has to be greater than min_leaf.
$m\_feature$	int [1, —feature size—]	5 *to* 50	Number of randomly selected features considered for a split at each node. A large value can increase correlation between trees resulting in higher error variance whereas a small value can miss selection of important features.
finalY train	float $[-\infty, \infty]$	Data Dependent	Too few or highly similar responses can cause numerical problems while inverting the covariance matrix.

an estimate of time required for another specific case. Simulation was conducted in an Intel Core i7 computer with 12GB RAM. For single output response case, the package applies random forest, whereas for more than one response scenario, the package automatically applies multivariate random forest.

Table 10: Simulation time in seconds for various error estimators for different scenarios. NCI-Dream Challenge Dataset has been used to obtain these time estimates. The number of samples used to design the models were 28 for 2 datasets and 20 for 5 datasets. Here, number of minimum leaf node used in the trees is 1.

Number of Responses	**Number of Datasets**	**Number of Trees**	**Number of features**	**Leave-one-out time (in s)**	**N fold cross validation time (in s)**	**Bootstrap time (in s)**	**0.632+ Bootstrap time (in s)**
1	2	10	10	51.12	5.23	39.75	41.68
1	2	10	100	66.9	6.7	54.58	57.37
1	5	10	10	53.26	7.07	45.07	48.12
1	5	10	100	67.19	9.46	55.40	59.17
1	2	100	10	399.14	41.81	324.53	341.33
1	2	100	100	611.02	60.91	528.54	554.67
1	5	100	10	476.37	62.37	403.59	432.34
1	5	100	100	698	104.5	519.38	556.7
3	2	10	10	51.36	5.04	39.28	41.36
3	2	10	100	114.33	10.35	85.28	89.73
3	5	10	10	55.48	7.76	46.31	49.68
3	5	10	100	107.71	13.82	88.88	95.14
3	2	100	10	530.03	52.73	410.30	430.74
3	2	100	100	1034.29	99.19	812	853.04
3	5	100	10	521.01	79.21	475.23	509.78
3	5	100	100	1063.91	141.1	823.05	882.58

References

[1] B. Efron, "Bootstrap methods: another look at jackknife," *Ann. Statist.*, vol. 7, pp. 1–26, 1979.

[2] Bradley Efron, "Estimating the error rate of a prediction rule: Improvements on cross-validation," *Journal of the American Statistical Association*, vol. 78, pp. 316–331, 1983.

[3] Robert Tibshirani Bradley Efron, "Improvements on cross-validation: The .632+ bootstrap method," *Journal of the American Statistical Association*, vol. 92, no. 438, pp. 548–560, 1997.

[4] Q. Wan and R. Pal, "An ensemble based top performing approach for nci-dream drug sensitivity prediction challenge," *PLOS One*, vol. 9, no. 6, pp. e101183, 2014.

[5] J. C. Costello and *et al.*, "A community effort to assess and improve drug sensitivity prediction algorithms," *Nature Biotechnology*, p. doi:10.1038/nbt.2877, jun 2014.

[6] Jordi Barretina and *et al.*, "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.," *Nature*, vol. 483, no. 7391, pp. 603–607, Mar. 2012.

[7] Broad-Novartis Cancer Cell Line Encyclopedia, "*http://www.broadinstitute.org/ccle/home*," Genetic and pharmacologic characterization of a large panel of human cancer cell lines.

[8] Wanjuan et al. Yang, "Genomics of drug sensitivity in cancer (gdsc): a resource for therapeutic biomarker discovery in cancer cells," *Nucleic acids research*, vol. 41, no. D1, pp. D955–D961, 2013.