# Predicting Cancer cell Line dependencies from the protein expression data of reverse-phase protein arrays

Paper: <https://pubmed.ncbi.nlm.nih.gov/32330068/>

## Abstract:

* Purpose: Using the protein expression data generated by reverse-phase protein arrays, we aim to assess their predictive power in identifying cancer cell line dependencies and to develop a related analytic tool.
* Four available expression-related features: - copy number alteration, DNA methylation, messenger RNA expression, and protein expression
* Method: Took set of genes with robust dependency data and four above features and performed same-gene predictions of the cancer dependency using different molecular features.
* Data: Multiomic data from DepMap, Cancer Cell line encyclopedia project, Consistency of cancer dependency data between CRISPR/Cas9 and short hairpin RNA-mediated perturbation platforms
* Results: Protein expression data contained substantial predictive power for cancer dependencies and best predictive feature for CRISPR/Cas9-based dependency data.
* Understanding tumor vulnerabilities and identifying therapeutic opportunities.
* Predicting cancer dependencies of a cell lines from different expression-related features of a gene.

## Important Notes:

* The response variable (model outcome) is a vector of dependency scores of each gene across cell lines.
* A score of 0 means that a gene is not essential, whereas score of – 1 corresponds to the median value of all common essential genes.

## Approach:

* First constructed a robust cancer dependency set by collecting genes and cell lines that showed high consistency between shRNA (DEMETER2) and CRISPR/Cas9 (DepMap19Q1) experiments.
* Next, this robust dependency set was overlapped with the cell lines and genes from CCLE to extract available data for model training.
* To conduct a fair comparison, only three expression-related DNA or RNA features (CNA, DNA methylation, and mRNA expression) with the RPPA-based protein expression data (total protein levels) from the same set of cell lines and performed cis-prediction (for the same gene) between the cancer dependency (response variable) and molecular features (explanatory variables) trained models for each gene dependency.

ML Schema:

* Diagram

  Description automatically generated
* Cancer cell lines samples were randomly split into 70% training set and 30% held-out testing set.
* Classifiers used: Linear Regression, Random Forest, and Conditional Random Forest.
* Baseline Model: Averaged dependency score as the predicted values.
* Model training: 10-fold cross validation using the training set and repeated procedure 10 times to avoid model overfitting. Then, applied the model to held-out testing set.
* The model performance was measured and compared using RMSE and R2.
* Trained model for each dependency.
* A dependency was flagged as predictable if it had at least one classifier that outperformed (had lower RMSE than) the baseline model in both the training and testing predictions. For the genes with predictable dependencies, we selected the best classifier (with the highest R2 ) based on the testing results and used the selected classifier to retrain the model using all samples.
* Finally, to evaluate the individual contribution of each feature, feature importance analysis was performed to identify a best predictor for each dependency. (In R v3.5.0 using the caret package with ML methods of lm, rf, and coforest. In addition, varlmp function to estimate the feature importance)

## Conclusion:

* Assessed the potential of RPPA-based protein expression data to infer cancer dependencies through ML-based feature importance analysis.
* Findings provided strong rationale for incorporating protein expression data into the prediction tasks of cancer dependencies.

## Limitations:

* The study access relatively small number of genes and proteins assessed which limits the statistical power compared with other expression-related features.
* Current RPPA platform covers only approximately 200 protein markers, and in future they aim to expand the protein list to approximately 500 proteins.

## Future Scope:

* In addition to cancer dependency to its gene, protein level likely helps to estimate the influence of other genes, and for such applications, an advanced ML strategy with sophisticated feature selection technique is required.

Takeaways:

* The expression level of a protein is a good predictor for the corresponding cancer dependency across cancer cell lines.
* Got important features for predicting the cancer cell dependencies from the protein expression data obtained from performing RPPA on a gene.
* 