Machine learning approach informs biology of cancer drug response

Alexandre Casadesús
Nikola Panajotovikj

Eliot Y. Zhu & Adam J. Dupuy

What are the questions the authors are trying to answer?

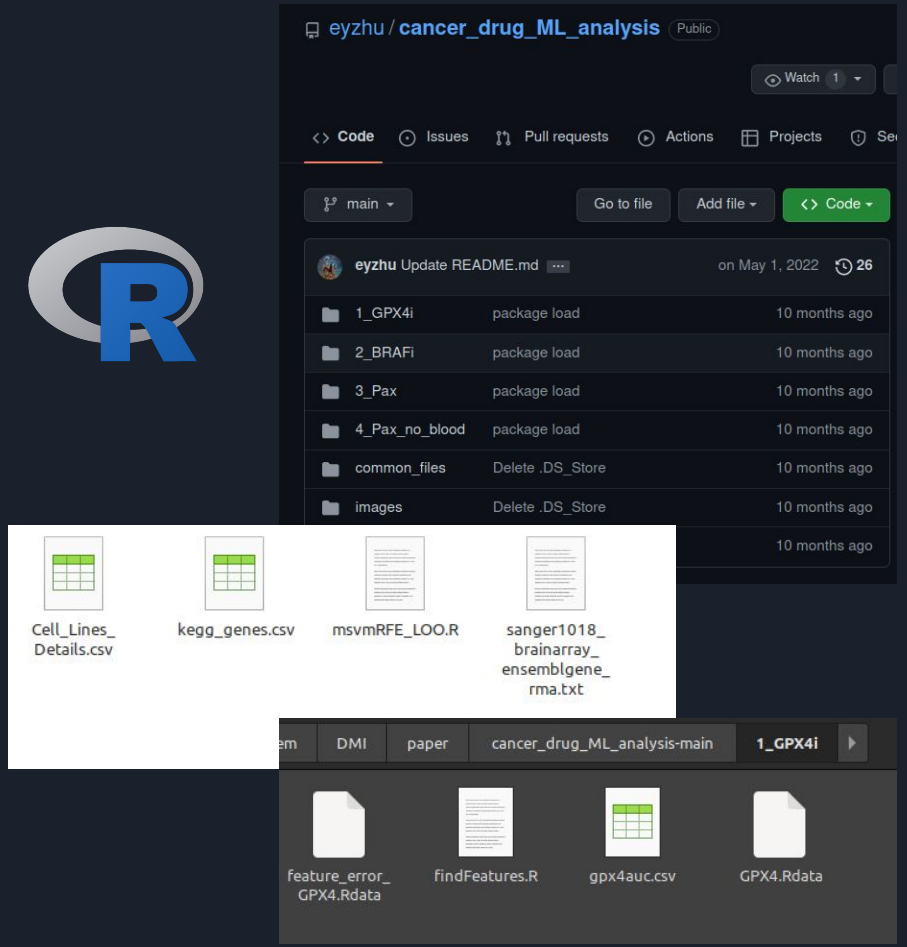
The researchers built a ***machine learning algorithm that uses knowledge of biological pathways and protein interactions to perform an analysis strategy for revealing specific pathways that contribute (i.e. to sensitivity or resistance) to drug response using publicly available pharmacogenomic cancer cell line datasets.***

DATA & CODE

Data quality and size. Is the data/code available? Is the analysis reproducible?

- GCSD RMA-Normalized microarray gene expression data.
- GCSD/CTRP v2 drug response data.
- KEGG Pathways data.
- R Scripts

Data and code **available**.
The analysis is **reproducible**



DATA

What are the preprocessing steps of the data?

How are they reported?

Few data preprocessing steps mentioned in paper:

- Synchronization of different input data sets (genes and cell lines)
- *BRAFi* (case study 2): use of 2 *BRAF* inhibitors for drug response vector

DATA - Supplementary table

Additional file 1. Fig S1:

Dotted line at AUC of 9 was the cutoff used to separate sensitive from resistant cancers.

Additional file 2. Fig S2: A)

Minimum feature ranking for each module. **B**) GO Biological Processes pathway enrichment of genes contained within modules presented in **A**). P-values shown are corrected for multiple hypothesis testing using the Holm-Bonferroni method.

Additional file 3. Supplementary Table S1.

KEGG pathways that passed pathway activity selection for ML210 analysis.

Additional file 4. Supplementary Table S2.

Top 20 enriched GO Biological Processes of genes returned by Boruta. P-values shown are corrected for multiple hypothesis testing using the Holm-Bonferroni method.

Additional file 5. Supplementary Table S3.

Top 20 Enriched GO Biological Processes of t-test derived genes for ML210 analysis. P-values shown are corrected for multiple hypothesis testing using the Holm-Bonferroni method.

Additional file 6. Supplementary Table S4.

Top 20 Enriched GO Biological Processes enrichment of elastic-net derived genes for ML210 analysis. P-values shown are uncorrected for multiple hypothesis testing.

Additional file 7. Supplementary Table S5.

KEGG pathways that passed pathway activity selection for BRAFi analysis.

Additional file 8. Supplementary Table S6.

Top 20 GO Biological Processes enriched in important modules for BRAFi analysis. P-values shown are corrected for multiple hypothesis testing using the Holm-Bonferroni method.

Additional file 9. Supplementary Table S7.

Top 20 GO Biological Processes enrichment of t-test derived genes for BRAFi analysis. P-values shown are uncorrected for multiple hypothesis testing.

Additional file 10. Supplementary Table S8.

Top 20 GO Biological Processes enrichment of elastic-net derived genes for BRAFi analysis. P-values shown are corrected for multiple hypothesis testing using the Holm-Bonferroni method.

Additional file 11. Supplementary Table S9.

Top 20 GO Biological Processes enrichment of t-test derived genes for PTX analysis. P-values shown are corrected for multiple hypothesis testing using the Holm-Bonferroni method.

Additional file 12. Supplementary Table S10.

Top 20 GO Biological Processes enrichment of elastic-net derived genes for PTX analysis. P-values shown are uncorrected for multiple hypothesis testing.

Additional file 13. Supplementary Table S11.

Top 20 GO Biological Processes enrichment of t-test derived genes for PTX analysis without blood cancers. P-values shown are uncorrected for multiple hypothesis testing.

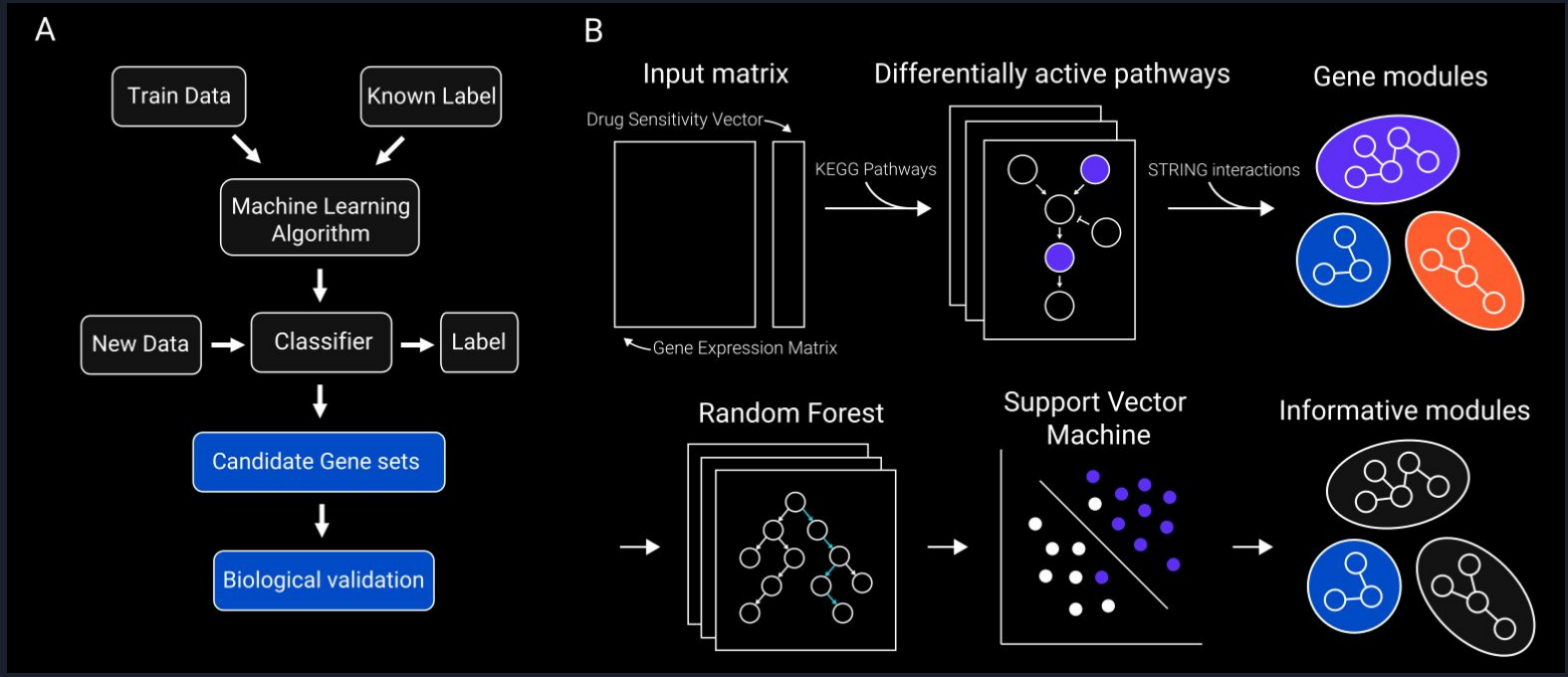
Additional file 14. Supplementary Table S12.

Top 20 GO Biological Processes enrichment of elastic-net derived genes for PTX analysis without blood cancers. P-values shown are uncorrected for multiple hypothesis testing.

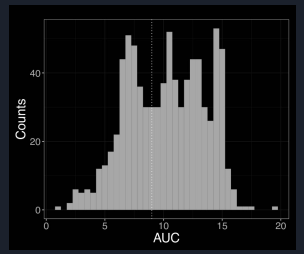
The Algorithm.

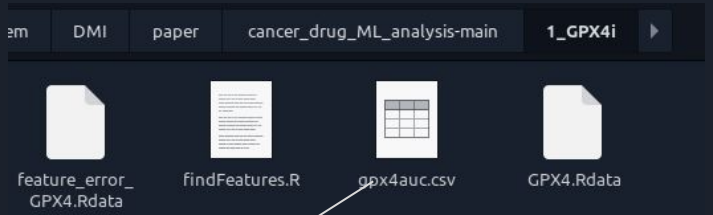
*"Our approach **emphasizes prioritization of biologically meaningful features used for classification rather than predictive performance.**"*

*"Conceptually, our approach is based on the **support vector machine learning algorithm** combined with **multiple layers of feature selection**. Additionally, we use **protein–protein interaction data to annotate important features** with pathway-level information. Ultimately, our approach returns a **ranked list of features, i.e. genes, that are grouped into mutually exclusive modules containing closely interacting genes**."*



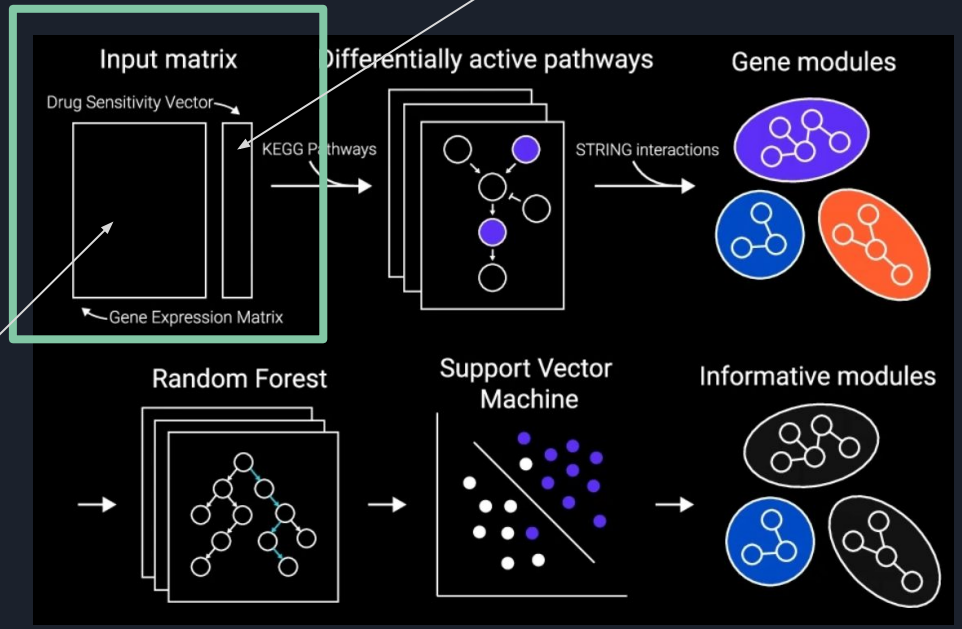
The Algorithm
- Input





1) To perform the machine learning analysis, they use ***RMA-normalized microarray gene expression from Genomics of Drug Sensitivity in Cancer (GDSC).***

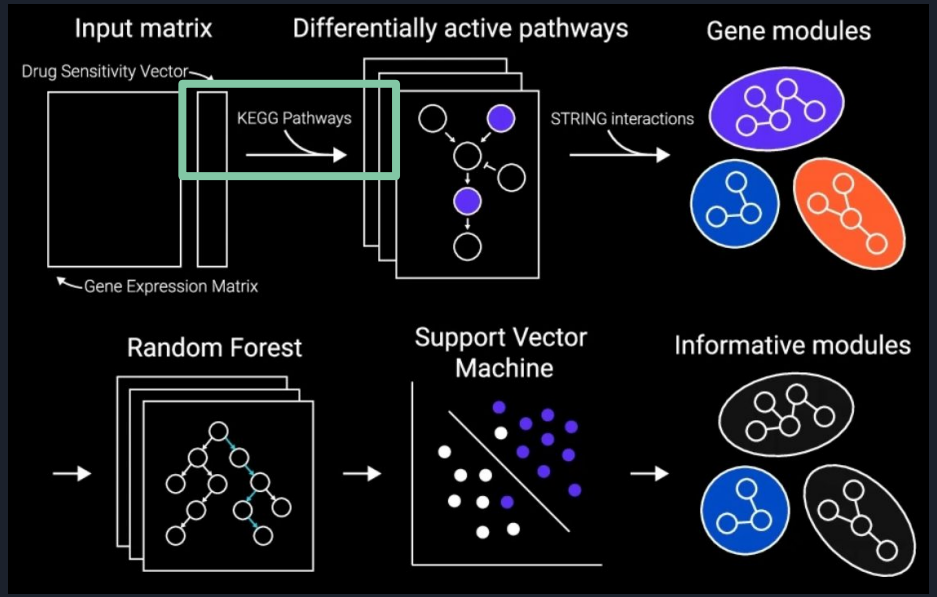
They use ***ML210 and PTX drug response data*** from the Cancer Therapeutics Response Portal V2 (CTRP v2). They used ***VEM and Dabrafenib response data*** from GDSC. They used area under the curve (AUC) as the metric for drug response.



1. First download array expression data and place this file in the common_files folder. As of 4/17/2022, There's no live link to the file I used, so I uploaded it to dropbox. The file was orginally obtained from ftp://ftp.sanger.ac.uk/pub/project/cancerrxgene/releases/current_release/sanger1018_brainarray_ensemblgene_rma.txt.gz.
    - https://www.dropbox.com/s/63d664sknfh8iv3/sanger1018_brainarray_ensemblgene_rma.txt?dl=0

The Algorithm - KEGG pathway

2) A list of KEGG pathways related to cellular processes, metabolism, genetic information processing, and environmental information processing is considered. KEGG (Kyoto Encyclopedia of Genes and Genomes) is a database that provides information on biological pathways and interactions between molecules in cells.

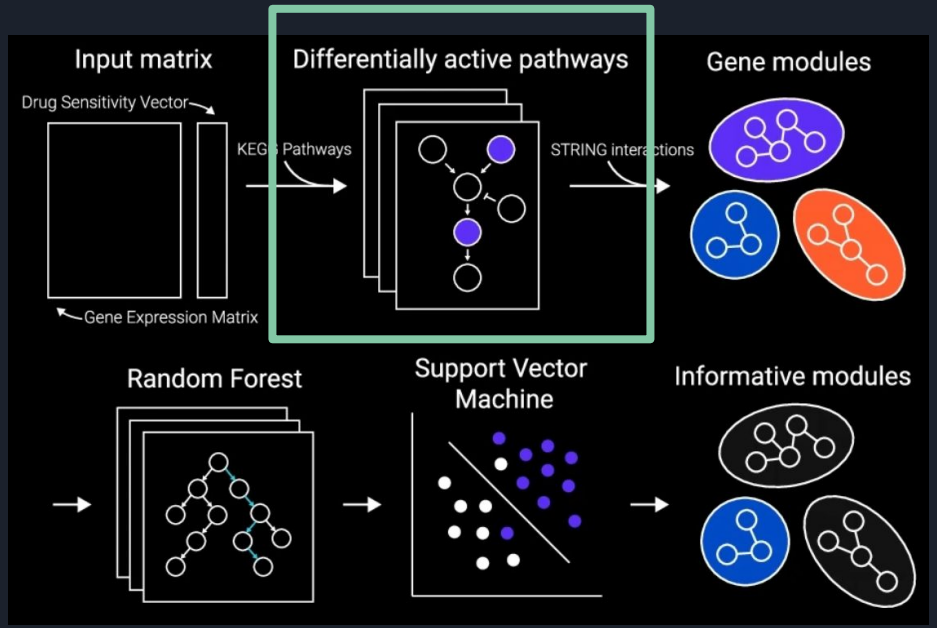


The Algorithm - pathway activity scores

3) Pathway activity scores are computed for each pathway to ***measure the difference in pathway activity between drug-sensitive and drug-resistant cells***.

Pathway activity scores are a measure of the ***overall activity of a biological pathway***, which can be determined by looking at the expression levels of genes involved in that pathway.

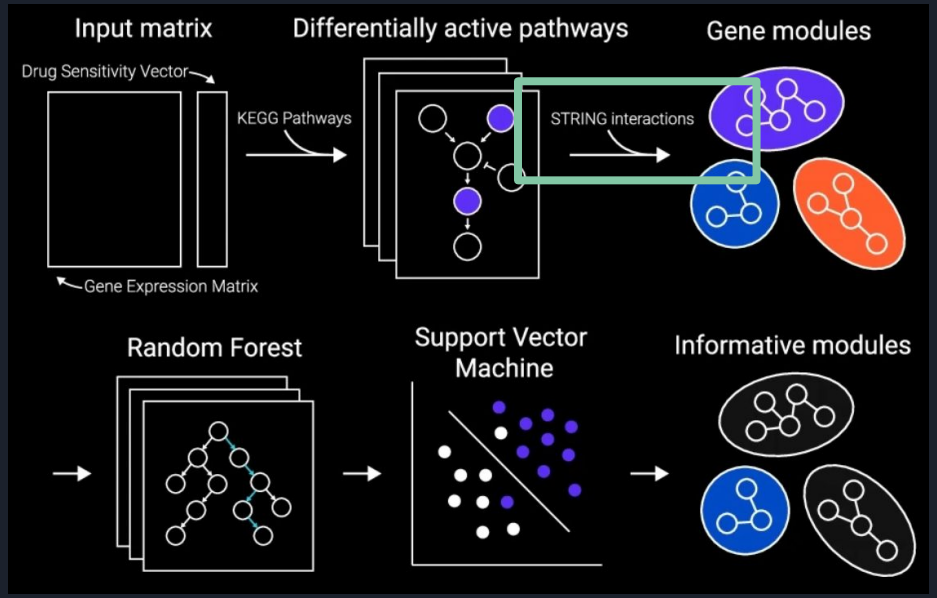
$$a_{pj} = \sum_{i=1}^{k} \frac{z_{ij}}{\sqrt{k}}$$



The Algorithm - String interactions and Gene modules

4) STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) is a database and web resource that collects and integrates information about ***protein-protein interactions***, including direct physical and indirect functional associations between proteins. STRING interactions can represent:

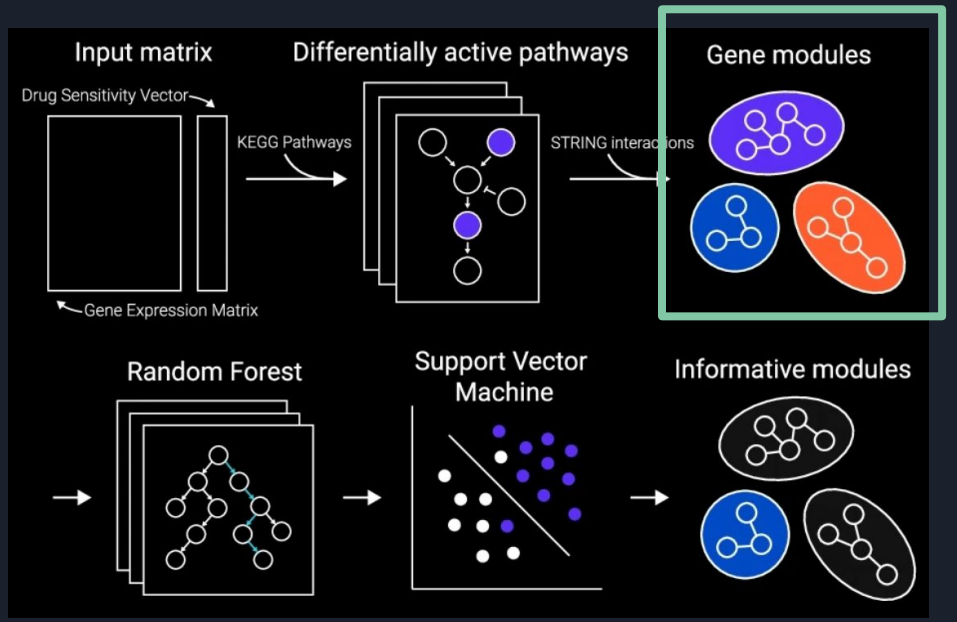
physical interactions, co-expression, co-localization, and shared pathways.



The Algorithm - String interactions and Gene modules

4) Genes from ***significant pathways are grouped into mutually exclusive network modules*** using ***hierarchical clustering*** of the dissimilarity (STRING interactions) between genes.

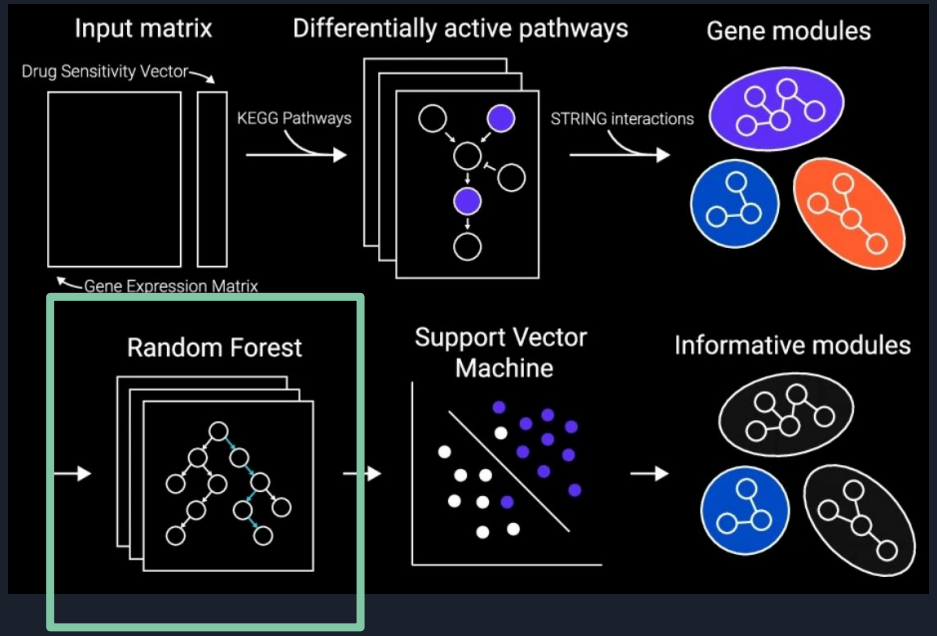
This helps to ***identify sets of genes that are functionally related and may play a role in drug response.*** The dissimilarity between genes is calculated based on the differences in their expression patterns across different cell lines.



The Algorithm - Boruta feature extraction

5) The most informative genes in each module are identified using a ***random-forest-based feature selection algorithm called Boruta***.

Boruta is a machine learning algorithm that can identify the most informative features (in this case, genes) that are relevant for drug response. It works by comparing the importance of each feature to a set of shadow features (randomly generated noise variables).

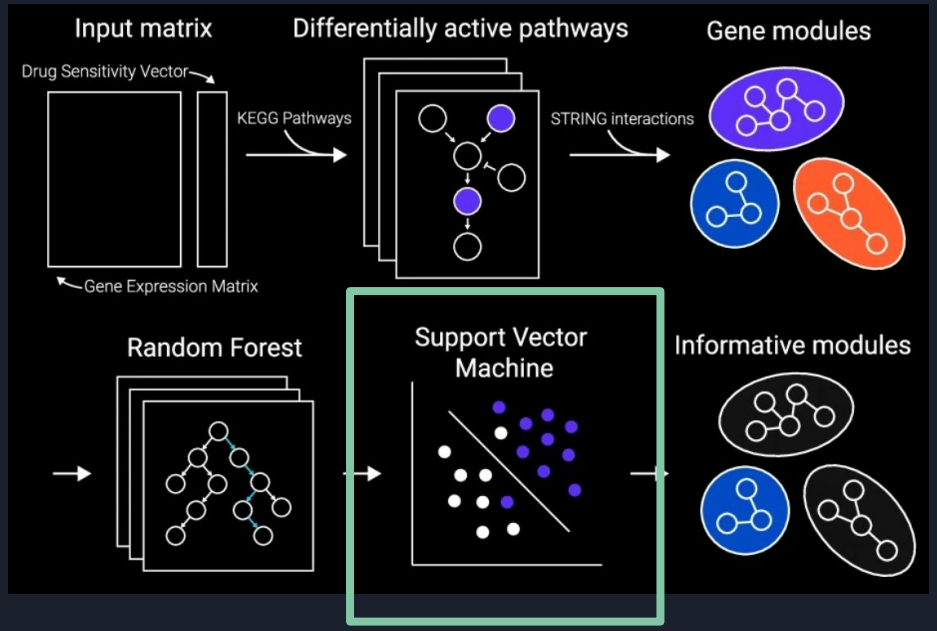


The Algorithm - RFE SVM

6) ***A classifier is built using the support vector machine (SVM) learning algorithm with recursive feature elimination (RFE) using the informative genes.***

- SVM is a **supervised** machine learning algorithm that can learn complex decision boundaries between classes.
- RFE is a technique used to ***select the most informative features (genes) for the classifier***. RFE involves running the SVM iteratively while removing the least informative feature at each iteration.

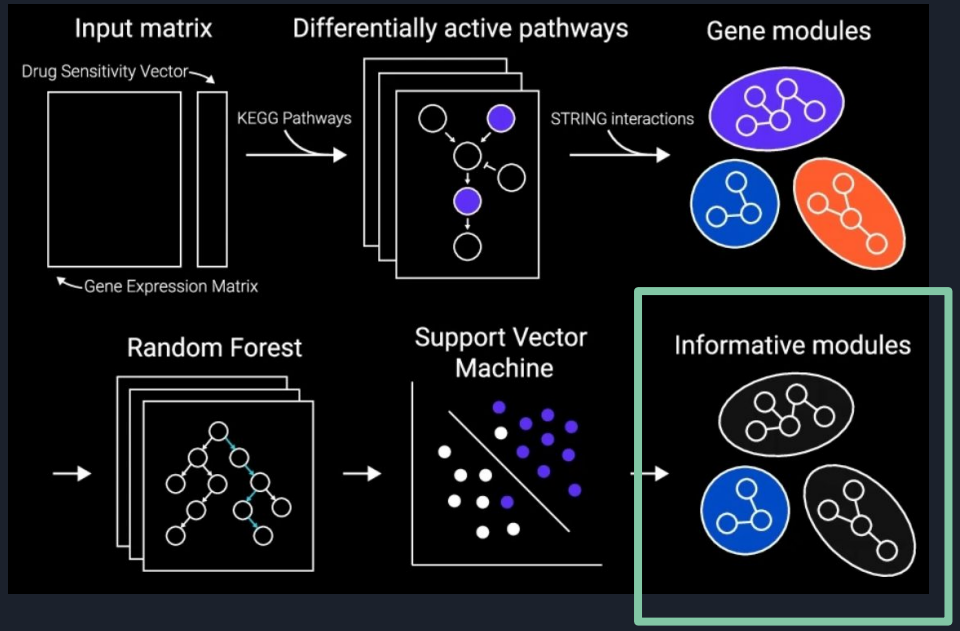
The rank for a feature is given as an average of a feature rank across leave-one-out-cross validation. ***The ranking of each feature determines the importance of the module it belongs to.***



The Algorithm - Informative Modules

7) ***The rank of each feature determines the importance of the module it belongs to, and the biological representation of each module is determined using Gene Ontology pathways enrichment analysis.*** This information is used to identify the pathways that are most relevant for drug response.

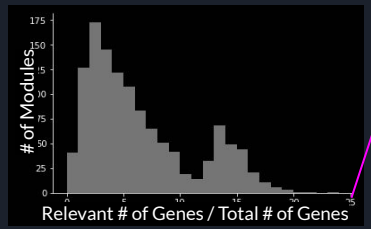
Gene Ontology (GO) is a database that provides information on the functions of genes and their roles in biological processes. This helps to provide a functional interpretation of the modules and identify potential drug targets.

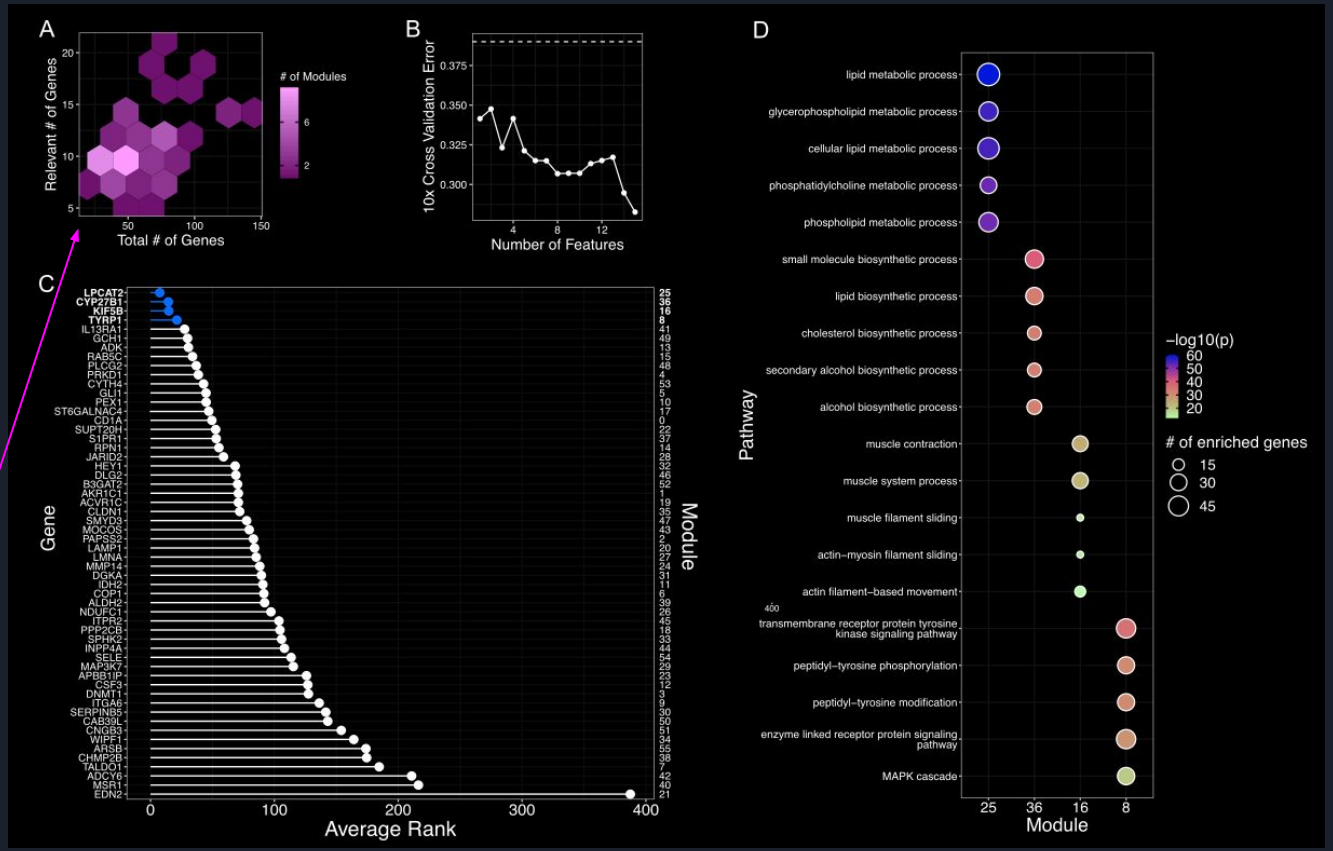


Case Study 1

Data: cancer cell lines treated with ML210.

The approach identified lipid metabolism as the top pathway that determines sensitivity to ML210, which is consistent with the knowledge that the balance of MUFAs and PUFAs determines susceptibility to ferroptosis. The authors also performed a negative control and repeated the analysis using all human KEGG pathways.

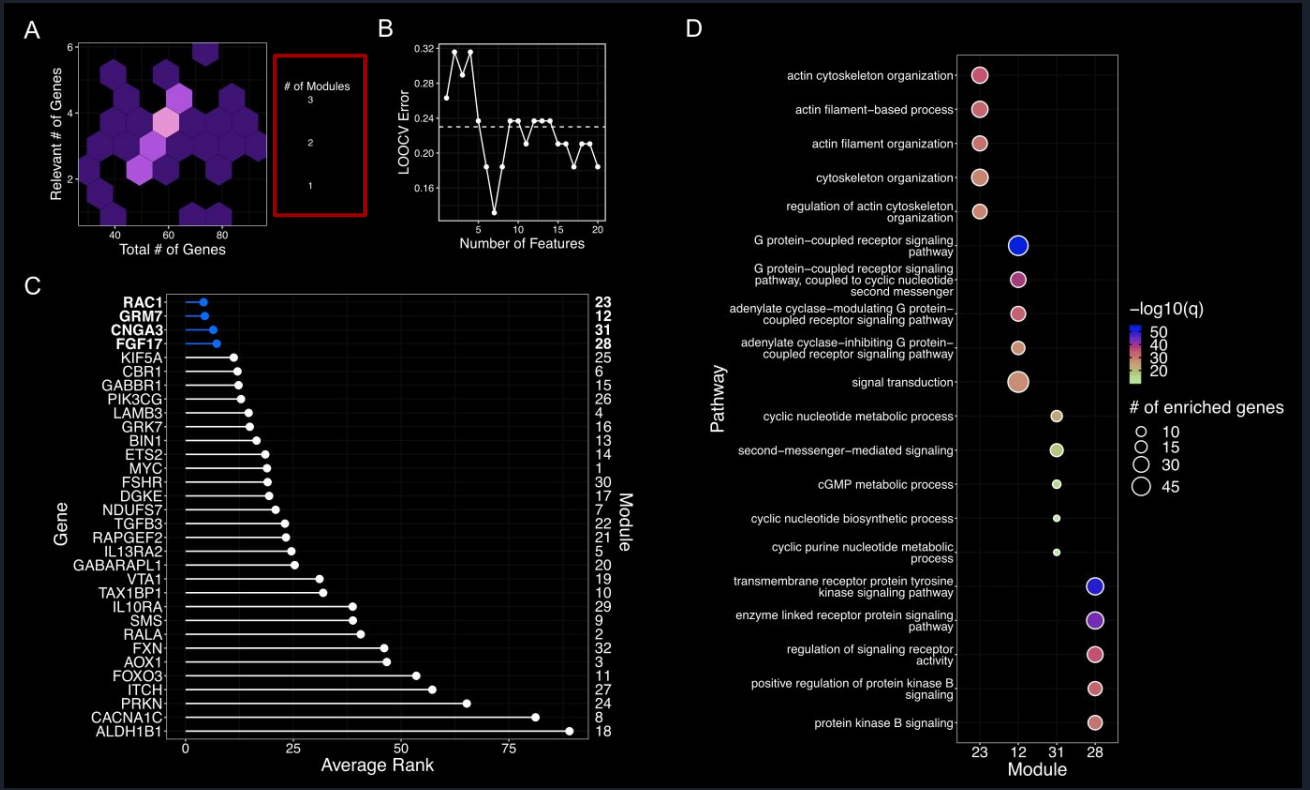




Case Study 2

Data: cancer cell lines treated with BRAFi.

Their method identified Rac1/cytoskeletal signaling as the most salient driver of drug resistance and prioritized the “transmembrane receptor protein tyrosine kinase signaling pathway” which is consistent with other studies that report certain RTKs drive intrinsic drug resistance to BRAFi in BRAFV600 cutaneous melanoma.



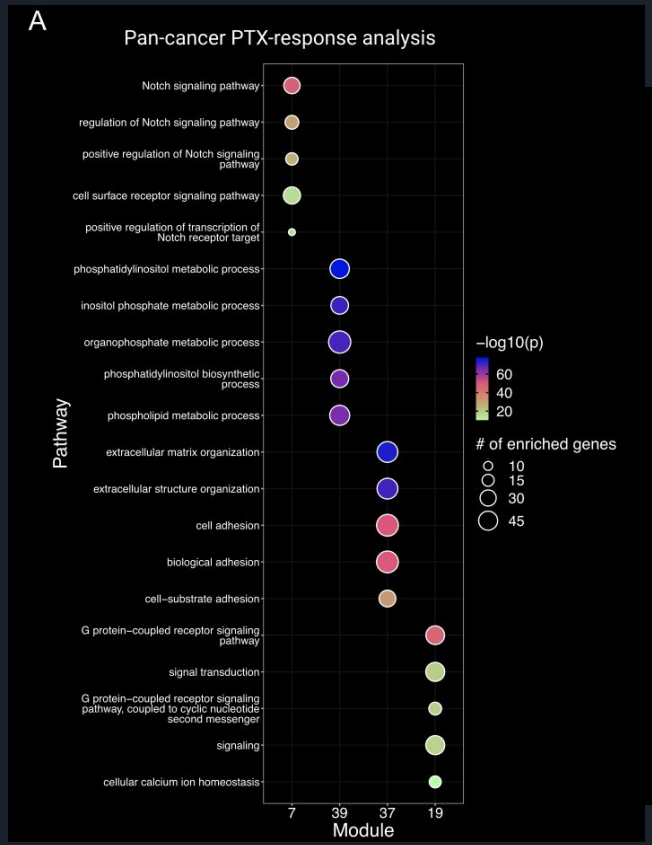
Case Study 3

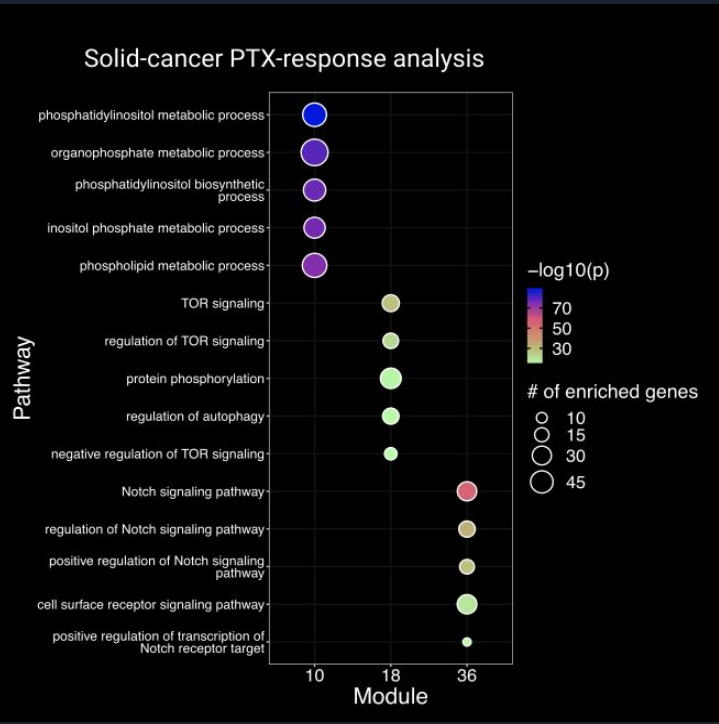
Data: cancer cell lines treated with Paclitaxel (PTX).

Approach applied to all cancers and only-solid cancers both identified:

- Notch signaling (resistance)
- Cell adhesion and Akt signaling (sensitivity)

Further bibliographic research led to discovery of previously **unreported connection between drug resistance to PTX and NOTCH3/PAX8 signaling.**





Questions?

Reference:

Zhu, E.Y., Dupuy, A.J. Machine learning approach informs biology of cancer drug response. BMC Bioinformatics 23, 184 (2022). https://doi.org/10.1186/s12859-022-04720-z

Thank you for your attention.