Hacettepe University

Artificial Intelligence Engineering Department

Research Project - 2025 Summer

SBSL Paper Review

August 30, 2025

Student:
Mohamed Yahya Mansouri

Supervisor:
Dr.Gülden Olgun

I Introduction

This document is an attempt to review the paper "Overcoming selection bias in synthetic lethality prediction" written by: Colm Seale, Yasin Tepeli, and Joana P.Gonçalves.
The goal is to understand the cause of Selection bias along with exploring the materials and methods used to address it.

Synthetic lethality is the simultaneous inactivation of two genes leading to cell death while disabling only one of them is non-lethal. [10][3] claim that this concept leverages medicine strategies in cancer treatment by selectively targeting specific genes without affecting normal genes. Identifying and validating synthetic lethal pairs remains a challenge and is under great scrutiny due to a variety of reasons such as the vast number of genes in a human cell, approximately 20000 protein-coding genes[7] leading to a combination of about 200 million possibilities. There are different methods employed to identify these pairs. Some of these methods are laboratory-based[12] which include yeast screens, drug screens and more. Some others are computational-based[13] such as statistical methods, network-based or topological-based methods, and machine learning-based.

These methods, however, ignore the problem of selection bias hindering generalization and overestimating performance as claimed by Seale et al.

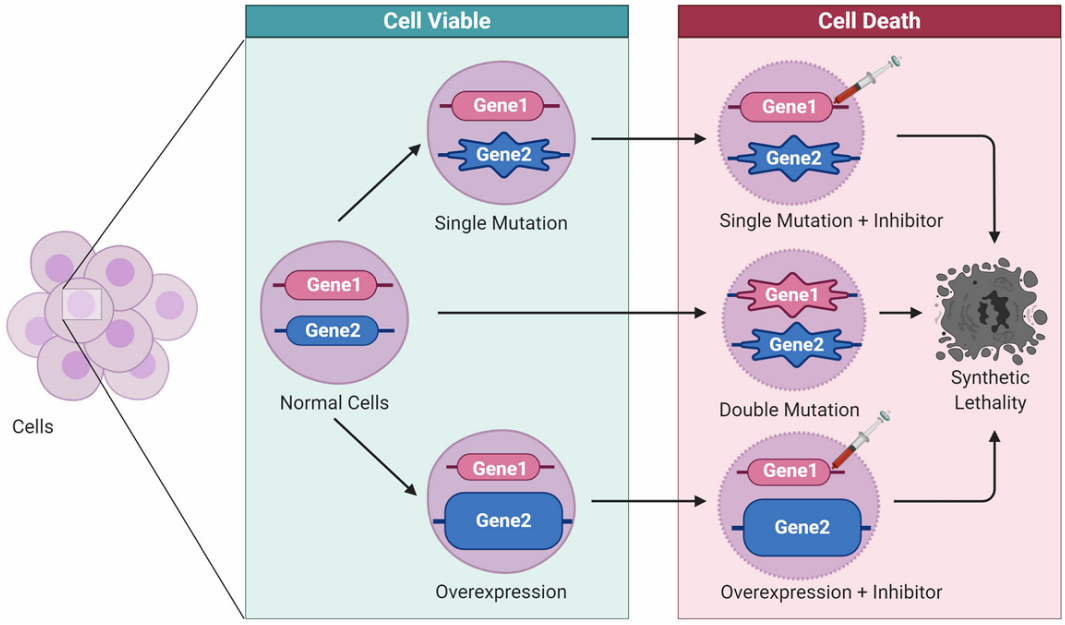


Figure 1: The principle of synthetic lethality. An individual genetic event is compatible with cell viability (left), whereas the co-occurrence of multiple genetic events causes cell death (right). The star represents a mutation; the large bubble represents genetic overexpression; the syringe represents DNA damage response inhibitor administration

II Selection Bias

[6] defines selection bias as *"an error in choosing the individuals or groups to take part in a study."* At its core, selection bias occurs when the sample data used for analysis and model training is not representative of the broader or target population. A good example of this distorting problem is trying to understand the relationship between lung-disease and the exposure to smoking. Selection bias occurs if the control group is taken from a hospital instead of randomly picked from the community since hospital patients may have other diseases affecting

lung function potentially leading to model overestimation.[1].
The paper being reviewed identifies selection bias as data imbalance found in the SL labels. The authors claim that this undermining issue is caused due to focusing on single cancer type labels or specific genes being overrepresented and dominating datasets. Furthermore, the existence of multiple pairs of the same genes is a dangerous pitfall inducing gene-specific patterns in prediction rather than SL mechanism generalization. This detrimental bias manifests itself in SL topology-based or network-based models[10][5]. The latter focuses on the structure of biological networks such as protein-protein interaction network[1] addressing questions like: How many neighbors do adjacent genes (represented as nodes) share? How close are they (shortest-distance problem)? What kind of interaction (represented as edges) a pair of genes has? Etc. The topological-based methods exploit a very limited set of SL labels since wet-lab experiments focus on a subset of gene[5] making the data used for these models skewed towards those genes rather than following a uniform distribution. In the manuscript under review, the authors propose more robust models, much less susceptible to selection bias called SBSL (selection bias-resilient synthetic lethality) prediction models based on logistic regression, a linear model, and random forest, a non-linear model.

III Solution

III.I Features

The SBSL models are feature-based models averting the SL graph topology-based methods and solely emphasizing on molecular data. Table1 shows the 31 features the dataset used for training has. Exploiting previous studies concerning synthetic lethal pairs detection using gene knockdown/knockout experiments[2], the authors used gene pair entries from ISLE and DiscoverSL as the foundation of their big dataset. Mainly, four features were taken from These two datasets encompassing the name of genes, cancer types, and SL labels. In order to merge these two datasets into one, primary preprocessing stages were followed such handling duplicate entries, feature selection through discarding any variable having near zero variance.

In addition to the four basic features, Seale et al engineered a comprehensive panel of 27 more molecular-based features making a total of 31 predictors to be included in the dataset. These variables are of paramount importance as they were taken from separate datasets derived from independent studies and observations. **This potentially protects the models from bias as these features characterized genes and gene pair relationships without relying on the recorded SL interactions.**

[1] A network based on a graph where nodes are the proteins and the edges describe the physical or functional interactions between them.

[2] Gene knockdown and knockout are two methods to understand gene function through controlling their activity. Knockout edits the DNA to completely remove or disable a gene. Usually, CRISPR method is used. Gene knockdown decreases gene expression achieved using the RNAi method through targeting mRNA.

Table 1: Overview of Features for SL Prediction

Independent Variable	**Feature Name**	**Data Source**
Gene Dependencies	CRISPR/RNAi_dept_stat CRISPR/RNAi_dep_pvalue CRISPR/RNAi_cor_stat CRISPR/RNAi_cor_pvalue CRISPR/RNAi_avg	DepMap and DEMETER2 Data v2
Mutual Exclusivity	discoversl_mutex_amp discoversl_mutex_del discoversl_mutex_mut discoversl_mutex MUTEX mutex_alt	The Cancer Genome Atlas(TCGA)
Survival	logrank_pval	Broad Institute Firehose Pipeline
Co-expression	tumour_corr/pvalue normal_corr/pvalue gtex_corr/pvalue	TCGA and GTEx
Differential expression	diff_exp_logFC diff_exp_pvalue	TCGA
Pathway co-participation	pathway_coparticipation	KEGG, Reactome, PID
Important Features	gene1 gene2 cancer_type	ISLE and DiscoverSL

Note: While the source of the data is given in the table above, some of the features are created in the project using those data, not explicitly copied and pasted.

1. Gene Dependencies

   - *CRISPR/RNAi_dept_stat*: This feature quantifies how much one gene's dependency differs depending on the other gene mutation. It runs a Wilcoxon rank-sum[3][9] test to compare the dependency scores of one gene between cell lines with and without a non-silent mutation in the second gene(context-dependent dependency).
   - *CRISPR/RNAi_dep_pvalue*: It depicts the significance of the difference of the dependency scores from the *CRIPSR/RNAi_dep_stat* feature.
   - *CRISPR/RNAi_cor_stat*: Person's correlation measures the linear relationship between two genes' dependency scores across all cell lines; a positive value indicates that they tend to be essential together, following the same behavior.
   - *CRISPR/RNAi_cor_pvalue*: A two tailed t-test p-value indicates how significant the *CRISPR/RNAi_cor_stat* linear correlation is.
   - *CRISPR/RNAi_avg*: Average of the means of the dependency scores for both genes.

   If one gene becomes more functional due to a mutation or perturbations in another gene then it hints at a synthetic lethality interaction between the pair of genes.

[3]This tests whether or not two distributions have the same median. In SL context, it is used to assess whether the effect of a gene loss alters the dependency score on the other gene through comparing the group with mutation and the group without.

2. Mutual Exclusivity

   - discoversl_mutex_amp: The p value that the genes are rarely co-amplified.
   - discoversl_mutex_del: The p value that the genes are rarely co-deleted
   - discoversl_mutex_mut: the p value that the genes are rarely co-mutated
   - discoversl_mutex: summarizing all the above alteration types
   - MUTEX: reflects the strength of mutual exclusivity between two genes using Java mutex algorithm[2].
   - mutex_alt: It captures how rarely two genes are altered together. Cancer cells avoid a combined perturbation of both genes since it is lethal. A low *mutex_alt* value indicates this anti-co-occurrence. The authors have used the hypergeometric test to calculate the p-value. A low p-value indicates that it is rarer than random to find co-occurrence indicating Mutual Exclusivity. However, [8] challenges this method, and calls it a naive approach, due to the numerous assumptions it needs to be applied. The assumptions are: "*mutations are mutually independent, every gene has the same chance to be mutated in a patient, and every patient has the same probability of harbouring a mutation.*" However, [8] claim that, due to the human nature, these assumptions do not hold true. **As a result, it is important to investigate the importance of this feature on the model after training.**

   Co-alteration is lethal for cancer cells. That's why they tend to avoid it so mutual exclusivity is an attribute to synthetic lethality.

3. Survival

   - logrank_pval: Two-tailed p-value testing whether co-alteration of two genes significantly affects patient survival. The p-value comes from the Wald test on the coefficients β_1 in a Cox proportional hazards model. S(A,B) = 1 if both genes are altered and 0 otherwise.

   If co-alteration of a pair of genes negatively impacts a patient survival, then it suggests synthetic lethality interaction due to the dual perturbation.

4. Co-expression

   - tumour_corr/pvalue, normal_corr/pvalue, gtex_corr/pvalue: Person's score between two genes across all samples. A positive score means both genes have expression levels that rise and fall together. The pvalue indicates how significant the correlation is.

5. Differential expression

   - diff_exp_logFC: This feature quantifies how an expression in one genes changes according to the presence and absence of non-silent mutations in the second gene.
   - diff_exp_pvalue: Two-tailed p-value test to indicate the significance of expression difference.

6. Pathway co-participation

   - pathway_coparticipation: quantifies how unlikely it is for genes a pair of genes to share a set of pathways. A lower value yields stronger evidence that the gene pair collaborates within common pathways.

   Genes sharing the same pathways are more likely to have a functionality interaction reflecting on a potential synthetic lethality relationship.

This is the first step to mitigate the notable Selection Bias problem. The second approach of the SBSL framework fully relies on the model architecture.

III.II Model Architecture

Seale et al used logistic regression with L0L2 regularization as a linear classification model, and Random Forest models with regularization as their non linear model. As explained in [4], Logistic regression regularization establishes sparse solutions[4] which is very instrumental for projects with a high number of independent variables mitigating the curse of dimensionality. L0 regularization penalty imposes sparsity through aiming at minimizing the number of active features through combinatorial solutions due to its non-convex nature[4] so it can be viewed as using heuristics to find an approximation to an optimal solution. L2 regularization, on the other hand, shrinks the coefficients of the features toward zero stabilizing the model. The combination of both tools decreases model complexity through feature balance along with securing generalization.
[4] describes L0L2 and L1L2(Elastic Net) regularization as a natural approach to prune features and avoid overfitting so it might be helpful to verify and use this in the future.

For non-linear models, the authors of the paper under review used Multivariate modeling with Unbiased Variable selection in R (MUVR) to achieve robust minimal variable selection. MUVR builds the classification model(or regression) while simultaneously selecting the most informative and relative variables using a repeated double cross-validation (rdCV) procedure[11]. **It is worth the examination to test for future relevant work as it scored high in the work of Seale et al when trained on BRCA ad LUAD cancer.** Last but not least, Regularized Random Forests (RRF) was also employed as a non-linear modeling approach to perform feature selection more effectively beyond the capabilities of the standard Random Forests. In this framework, two key parameters govern model complexity control: **mtry** determining how many features to look at during the tree splits promoting diversity where a high mtry value can sharpen the split but risk overfitting. The other parameter is **coefReg** as a penalty when determining the feature to use for splitting where a lower value means stricter feature selection. *After running the code of the project, the compatible mtry value found is 8 and appropriate coefReg value is 0.9*

To measure the success of their work, the authors compares the SBSL models to other five SL prediction models. The table below categorizes them.

Network-based	**Statistical based**	**ML-based**
pca-gCMF	DAISY	DiscoverSL
GRSMF		
GCATSL		

Table 2: Models used for comparison with SBSL.

Logistic Regression	**Random Forest**
L0L2	Elastic Net
MUVR	RRF

Table 3: The SBSL Models

[4]Refers to the classification model where most of the feature coefficients are zeros. It indirectly enables variable selection through setting the weights of some noisy variables to zero.

III.III Preprocessing Stage

After creating the combined dataset encompassing the 31 features, the data got downsampled to avoid data imbalance and then was split into 70% training and the rest for testing. Feature Standardization was also applied using the Z-score standardization through subtracting by the mean and dividing by the standard deviation. The paper claimed that any features having a near zero variance is discarded from the dataset.

IV Results

For the Precision-Recall Curve in Figure 2)a, Random Forest and MUVR exhibit the highest precision across all recall levels along with achieving a better trade-off between TPR and FPR than the rest of the models. Elastic Net and L0L2 are in the middle range performing less than RRF and MUVR but better than DAISY and DiscoverSL ranking SBSL models the highest across these metrics. Having the narrowest shaded regions, MUVR and RRF show lower variance compared to the other datasets which is evidence for the stability of the two models suggesting generalization and reliability. Daisy and DiscoverSL on the other hand fall behind in the FPR vs TPR curve as their plots are close to the random guess line along with drastically dropping in the precision in the Precision-Recall Curve.

According to the AUROC[5] table in Figure 2)c, the SBSL model MUVR and the network-based methods performed the best with average AUROC score above 0.80 with SBSL models scoring the highest for the COAD and LUAD cancer with little difference in the BRCA cancer. Due to the low performance on the OV cancer by the SBSL, the authors hypothesized that this low performance stems from the low Tumour Mutational Burden (TMB)[6] which decreases the insights derived from features relying on mutation data. Moreover, according to the paper, OV cancer had small number of entries, limited to 86 unique genes. This is verified after investigating the code.

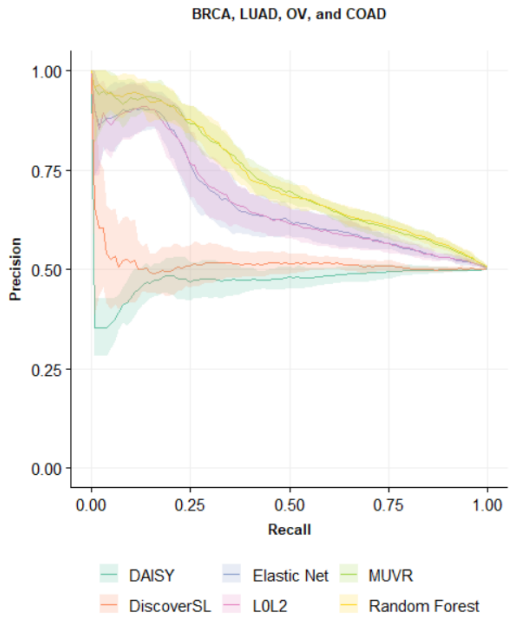
Furthermore, the authors reason that the high performance of the topological models on Ovarian(OV) cancer is an evidence of selection bias. They mentioned there was many rows of the OV entries identical looking almost like copy-pastes exhibiting the same SL patterns which hints at heavy selection bias. This is an important heads-up to test for future work.

Seale et al. performed interesting experiments to test whether the models performed well on unseen data in the training dataset. Their experimental setup consisted of training BRCA model on the ISLE dataset and testing it using DiscoverSL dataset. Any common entries between ISLE and DiscoverSL were removed from the train set. The same experimental setup was followed only changing the cancer type from BRCA to LUAD where they used DiscoverSL as the training set and ISLE as the test set. The authors found that the SBSL models generalized better, specifically the linear models, beating the topological-based methods. They even supported the robustness of their model using gene holdout experiments where they tested their model against the number of genes shared between the train set and the test set. The topological-based methods performance drastically dropped to nearly making random guesses.

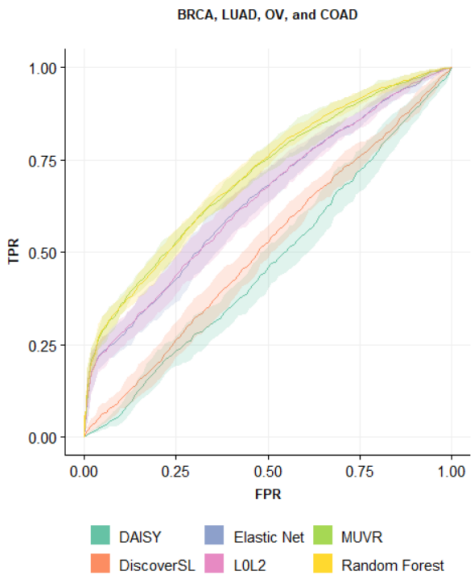
The researchers have also inferred from a set of three experiments that not all cancers are equal in SL prediction. This means that training models on specific cancers yields more promising and reliable results than training models on multiple cancers. This is observed from the table in figure 3, taken from training the L0L2 and MUVR models on one-cancer dataset and

[5]AUROC stands for Area Under the Receiver Operating Characteristic curve. It is the area under the curve that plots True Positive Rate vs False Positive Rate. Its range is bound to [0,1]. The highest score the more capable the model to classify true positives than false positives.

[6]National Cancer Institute defines TMB as the total number of mutations found in the DNA of cancer cells.



(a) Precision-Recall Curve



(b) False Positive Rate vs True Positive Rate Curve

Method	BRCA	COAD	LUAD	OV
Elastic Net	0.84 ± 0.01	0.60 ± 0.02	0.85 ± 0.02	0.59 ± 0.03
L0L2	0.84 ± 0.01	0.60 ± 0.02	0.85 ± 0.02	0.59 ± 0.03
MUVR	0.86 ± 0.01	**0.64 ± 0.01**	**0.87 ± 0.01**	0.56 ± 0.07
RRF	0.86 ± 0.01	0.63 ± 0.02	0.87 ± 0.02	0.57 ± 0.07
DAISY	0.61 ± 0.02	0.38 ± 0.02	0.44 ± 0.03	0.41 ± 0.04
DiscoverSL	0.54 ± 0.02	0.54 ± 0.02	0.54 ± 0.03	0.45 ± 0.04
GCATSL	0.59 ± 0.04	0.51 ± 0.01	0.86 ± 0.03	**0.99 ± 0.02**
GRSMF	0.82 ± 0.01	0.57 ± 0.02	0.87 ± 0.02	**0.99 ± 0.01**
pca-gCMF	**0.90 ± 0.02**	0.54 ± 0.03	0.87 ± 0.02	0.94 ± 0.05

(c) Classification performance of SL prediction models within a cancer type, denoted by the AUROC curve.

Method	BRCA	COAD	LUAD	OV
Elastic Net	0.87 ± 0.01	0.59 ± 0.01	**0.87 ± 0.02**	0.58 ± 0.03
L0L2	0.88 ± 0.01	0.59 ± 0.02	**0.87 ± 0.02**	0.58 ± 0.04
MUVR	**0.89 ± 0.01**	0.62 ± 0.02	**0.87 ± 0.02**	0.54 ± 0.06
RRF	**0.89 ± 0.01**	**0.63 ± 0.02**	**0.87 ± 0.02**	0.55 ± 0.05
DAISY	0.58 ± 0.02	0.43 ± 0.01	0.47 ± 0.02	0.48 ± 0.04
DiscoverSL	0.55 ± 0.02	0.53 ± 0.02	0.55 ± 0.03	0.49 ± 0.04
GCATSL	0.55 ± 0.02	0.50 ± 0.01	0.82 ± 0.04	**0.98 ± 0.03**
GRSMF	0.81 ± 0.01	0.59 ± 0.02	0.85 ± 0.02	0.97 ± 0.04
pca-gCMF	**0.89 ± 0.04**	0.56 ± 0.03	0.83 ± 0.03	0.93 ± 0.06

(d) Classification performance of SL prediction models within a cancer type, denoted by the AUPRC

Figure 2: Model Analysis. The plots were extracted from the code, while the last two were taken from the paper.

Method	Cancer	Pan-cancer		One-cancer
		Unbalanced	Balanced	
L0L2	BRCA	0.64 ± 0.02	0.75 ± 0.01	**0.83 ± 0.01**
	COAD	0.52 ± 0.02	0.51 ± 0.02	**0.60 ± 0.02**
	LUAD	0.73 ± 0.03	0.79 ± 0.02	**0.83 ± 0.02**
	OV	0.40 ± 0.04	0.53 ± 0.04	**0.58 ± 0.03**
MUVR	BRCA	0.76 ± 0.01	0.82 ± 0.02	**0.86 ± 0.01**
	COAD	0.62 ± 0.02	0.60 ± 0.01	**0.64 ± 0.01**
	LUAD	0.81 ± 0.02	0.83 ± 0.02	**0.86 ± 0.01**
	OV	**0.55 ± 0.06**	0.52 ± 0.04	0.54 ± 0.07

Figure 3: Performance of one-cancer and pan-cancer models(AUROC)

pan-cancer[7]. Except OV, the AUROC score is high for all cancers and both models. Yet, there was a notable degradation of the score for the Pan-cancer method especially the unbalanced set. Driven by scientific curiosity, the authors assessed whether cancer-specific models can make SL predictions on unseen cancer types in the second experiment. However the models were not

[7]A dataset that includes different samples from multiple cancer types, not limited to only one.

able to generalize well, even previous well performed models. Interestingly, the L0L2 model was able to generalize when trained on the BRCA cancer type and tested against the LUAD with a mean AUROC score of 0.79. The third experiment aimed at checking whether multiple cancer model were able to predict well on unseen cancer types. This showed positive results where for instance, a model trained on COAD, LUAD, and OV was able to generalize well on BRCA.

Lastly, feature importance analysis revealed that gene dependency-based features contributed the highest to the prediction models. Figure 4 suggests that *RNAi_dep_stat* ranked the highest suggesting that the SBSL linear models heavily rely on the dependency scores. This is further backed up by moderate to high importance of the following dependency features: *CRISPR_avg, RNAi_avg, RNAi_cor_pvalue, RNAi_cor_stat, and CRISPR_dep_pvalue*. In fact, the absence of the features lead to a marked performance drop across all models where the mean AUROC score decreased from 0.83 and 0.85 to 0.64 and 0.76 for the BRCA and LUAD cancer types respectively. The boxplots in Figure 5 exhibits how the performance of all models declines after removing the dependency features. **To explain informatively, the dependency-based features, which explains how a cell depends on a gene before and after mutating the other gene, captures critical biological relationships between the pair of genes that depicts potential Synthetic lethal interaction. This in return drives the predictive power of the SBSL models.**
The authors claimed that while CRISPR and RNAi are different knockout methods, there is a moderate to high correlation between them. This was verified from the correlation heatmap in figure 6.

Figure 4 reveals that *mutex_alt* feature is not very important, closely to having zero effect, for Logistic Regression and Random Forest. With the use of Accumulated Local Effects(ALE) to evaluate the relationship between independent and target variables, *mutex_alt* exhibited zero effect which is shown in Figure 7. It might be due to the heavy reliance on the dependency scores.

Figure 4: Feature Importance Analysis. The y-axis includes the names of the features and the x-axis. This figure was extracted from running the code.

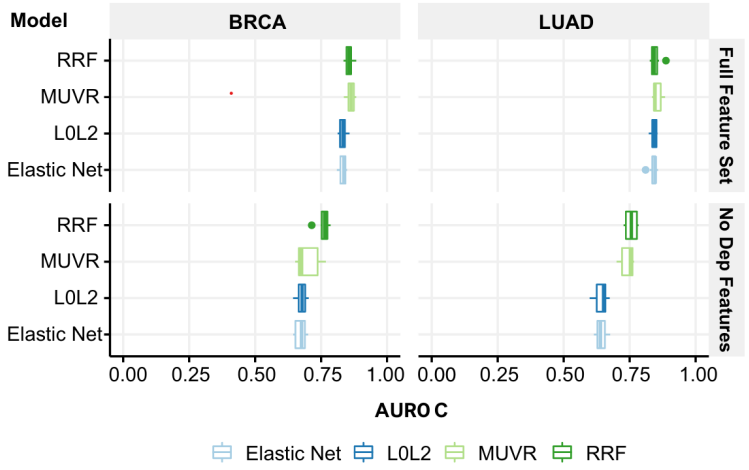


Figure 5: Performance of SBSL models with and without gene dependency-based features (AUROC over 10 runs), respectively, labeled 'Full Feature Set' and 'No Dep Features'. This figure is taken from the paper.

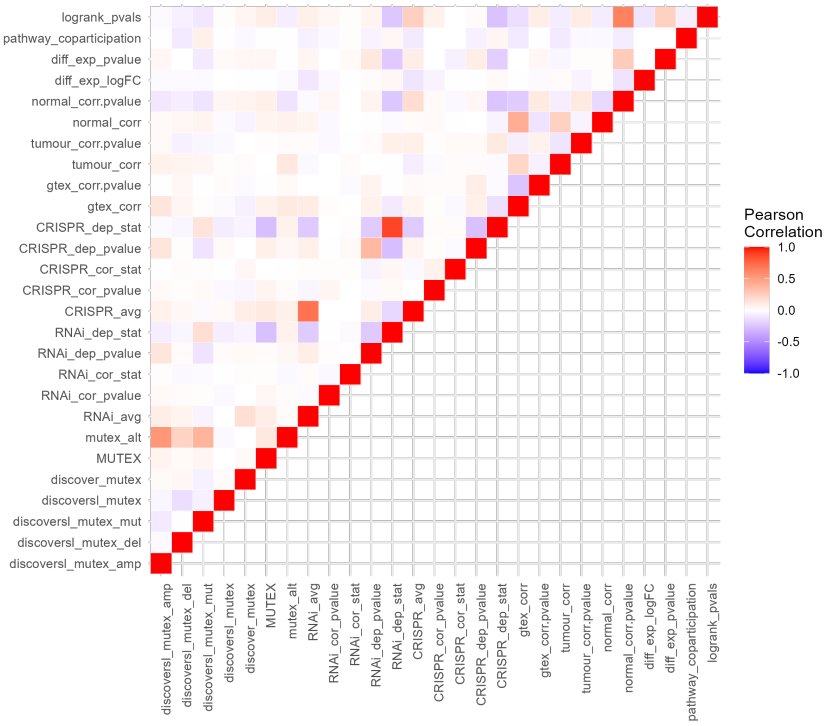


Figure 6: Classification performance of SL prediction models within a cancer type, denoted by the AUROC curve. This figure was extracted from running the code.

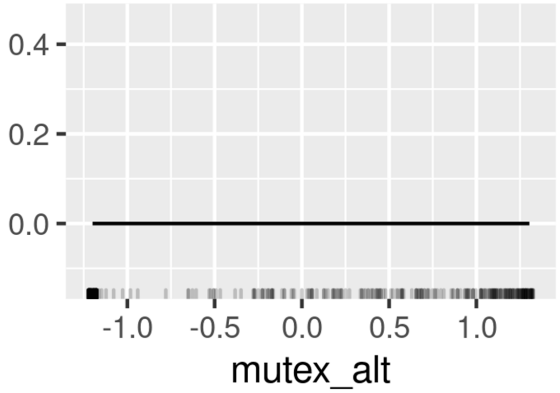


Figure 7: ALE Plot for *mutex_alt* showing minimal effect on model prediction.

V Conclusion

By deliberately selecting features and leveraging a hybrid model architecture combining L0L2 logistic regression and the MUVR model, SBSL was able to consistently outperform network-based methods across nearly all cancer types, with the ovarian cancer being the sole exception due to topological model overestimation.

Incorporating L0L2 along with MUVR was instrumental in combating selection bias through the inner removal of noisy features. Equally important is the data preprocessing phase and careful data splits to train the model and test it on unseen data. It is essential to notice that molecular-based features taken independently of SL annotations were essential to avoid selection bias in the SBSL project.

The authors admitted that one limitation of the SBSL project, however, is heavily relying on the dependency score which are not available for rarer cancer types as there are not enough representative cell lines to perform gene knockout/knockdown to learn gene dependency.

The authors put forward two recommendations to effectively evaluate SL prediction models. The first is to assess model performance on all cancer types in the dataset to ensure unbiased and generalizable SL predictors. The second recommendation advocates for a deep examination on selection bias through gene holdout experiments. A single holdout ensures that for each test pair, at most one can appear again in the training set. A double holdout ensures that no gene in the test set appear in the training set. None (baseline) ensures that no pair of genes appear in the train set if it appears in the test set, however, one of the genes of that pair may be included.

Such detailed examinations are seemingly prominent steps to ensure good learning practices for SL prediction models.

References

[1] Lorraine K. Alexander et al. *ERIC Notebook Second Edition - Selection Bias*. https://sph.unc.edu/wp-content/uploads/sites/112/2015/07/nciph_ERIC13.pdf. University of North Carolina at Chapel Hill. Accessed July 2025. 2008.

[2] Ozgun Babur et al. "Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations". In: *Genome Biology* 16 (Dec. 2015). DOI: 10.1186/s13059-015-0612-6.

[3] Graeme Benstead-Hume et al. "Predicting synthetic lethal interactions using conserved patterns in protein interaction networks". In: *PLoS Computational Biology* 15.4 (2019), e1006888. DOI: 10.1371/journal.pcbi.1006888. URL: https://doi.org/10.1371/journal.pcbi.1006888.

[4] Anna Deza and Alper Atamturk. *Safe Screening for Logistic Regression with ℓ_0-ℓ_2 Regularization*. 2022. arXiv: 2202.00467 [stat.ML]. URL: https://arxiv.org/abs/2202.00467.

[5] Joana Gon¸calves Mathijs J. de Wolf Yasin Tepeli. "Mitigating selection bias in synthetic lethality prediction using metric learning". MA thesis. Delft, The Netherlands: Delft University of Technology, June 2023. URL: https://resolver.tudelft.nl/uuid:95cb2d5b-194a-49ec-8281-98cf0f4e35c0.

[6] National Cancer Institute. *Selection Bias*. https://www.cancer.gov/publications/dictionaries/cancer-terms/expand/S. NCI Dictionary of Cancer Terms. Accessed July 2025. 2009.

[7] National Human Genome Research Institute - Eric Green. *Gene*. https://www.genome.gov/genetics-glossary/Gene. Accessed July 2025. 2025.

[8] Pietro Pinoli, Sriganesh Srihari, and Limsoon Wong. "Identifying collateral and synthetic lethal vulnerabilities within the DNA-damage response". In: *BMC Bioinformatics* 22 (May 2021). DOI: 10.1186/s12859-021-04168-7.

[9] ROBERT H. RIFFENBURGH. "Chapter 6 - Statistical Testing, Risks, and Odds in Medical Decisions". In: *Statistics in Medicine (Second Edition)*. Ed. by ROBERT H. RIFFENBURGH. Second Edition. Burlington: Academic Press, 2006, pp. 93–114. ISBN: 978-0-12-088770-5. DOI: https://doi.org/10.1016/B978-012088770-5/50045-9. URL: https://www.sciencedirect.com/science/article/pii/B9780120887705500459.

[10] Colm Seale, Yasin Tepeli, and Joana P Gonçalves. "Overcoming selection bias in synthetic lethality prediction". In: *Bioinformatics* 38.18 (July 2022), pp. 4360–4368. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btac523. eprint: https://academic.oup.com/bioinformatics/article-pdf/38/18/4360/49884644/btac523.pdf. URL: https://doi.org/10.1093/bioinformatics/btac523.

[11] Lin Shi et al. "Variable selection and validation in multivariate modelling". In: *Bioinformatics* 35.6 (Aug. 2018), pp. 972–980. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty710. eprint: https://academic.oup.com/bioinformatics/article-pdf/35/6/972/48967176/bioinformatics_35_6_972.pdf. URL: https://doi.org/10.1093/bioinformatics/bty710.

[12] Wipawee Topatana et al. "Advances in synthetic lethality for cancer therapy: cellular mechanism and clinical translation". In: *Journal of Hematology & Oncology* 13.1 (2020), p. 118. DOI: 10.1186/s13045-020-00956-5. URL: https://doi.org/10.1186/s13045-020-00956-5.

[13] Jing Wang et al. “Computational methods, databases and tools for synthetic lethality prediction”. In: *Briefings in Bioinformatics* 23.3 (Mar. 2022), bbac106. ISSN: 1477-4054. DOI: 10.1093/bib/bbac106. eprint: https://academic.oup.com/bib/article-pdf/23/3/bbac106/43745875/bbac106.pdf. URL: https://doi.org/10.1093/bib/bbac106.