MOSL: Integrating multi-omics and machine learning to predict synthetic lethality in cancer cell lines

Tan Pham*[1,2][0000−0002−6824−172X], Dang Vu[2][0009−0003−3268−024X], Tien Dang[3][0009−0001−0405−8205], Binh T. Nguyen[1,2][0000−0001−5249−9702], and Tuan-Anh Tran**[4][0000−0002−3291−2413]

[1] AISIA Research Lab, University of Science, Vietnam National University
[2] University of Science, Vietnam National University
[3] University of Information Technology, Vietnam National University
[4] Institut Pasteur, France

Abstract. Synthetic lethality (SL) is a genetic interaction in which the simultaneous perturbation of two genes leads to cell death, whereas perturbation of either gene alone is viable. Challenging in predicting synthetic lethality pairs usually comes from a highly biased source of data (only SL positive or only SL negative pairs) and insufficient data background (research data are usually single-omic). In this work, we create a large multi-omics dataset coming from diverse settings, including transcriptomics profiles, genetic perturbations (i.e., RNA interference, Clustered Regularly Interspaced Short Palindromic Repeats), and genomics data (i.e., nucleotide sequences and amino acid sequences). Here, we also propose a multimodal model containing different modules specifically designed to capture biological insights of each type of omic data. Quantitative results show that our method has a top-notch specificity in predicting synthetic lethality despite its simplicity compared to other advanced techniques like graph-based models with a specificity of 99.75%, a sensitivity of 90. 55% and precision of 97. 88%.

Keywords: Computational Biology · Machine Learning · Collective intelligence · Soft computing · Multicriteria decision making.

1 Introduction

Synthetic lethality (SL) is a genetic interaction where the simultaneous loss-of-function from two genes leads to cellular death, while the perturbation of either gene does not [36]. SL interaction was first discovered in *Drosophila melanogaster* (also known as the common fruit fly) [24] and *Saccharomyces cerevisiae* (baker's yeast) [3]. Genetic experiments on *D. melanogaster* showed that simultaneous inactivation of both gene *Bar* and gene *glass*, encoding for transcription factors,

* First author
** Corresponding author

had led to neural defects and death in the early stage of development. Meanwhile, independent mutations in either gene only resulted in irreproducibility [34].

SL researches then progressed as tools to discover novel therapeutic targets in humans, especially oncology. Hartwell et al. [13] proposed to extrapolate synthetic lethal interaction observed in *S. cerevisiae* to develop anticancer therapeutics for human. Accordingly, McManus et al. [27] illustrated the synthetic lethal interaction of two homologous genes, *RAD54* and *RAD27*, in both yeast and cancer cell lines. The discovery of synthetic lethality has revolutionised the field of drug discovery in oncology, driving cancer treatments towards precise and context-dependent medicine. Cancer cells often carry defects in at least one tumor suppression gene, which causes its synthetic lethal partner to be essential for the viability of those cancer cells. Inhibiting the partner gene prevents cancer cells from proliferating while doing no harm to the surrounding healthy cells. [7,29]. Due to its effectiveness in inhibiting cancerous cells, many targeted therapies had undergone clinical trials and later on approved by the US Food and Drug Administration (FDA) as resorts for oncology [15,23,5]. An exemplified FDA-approved targeted therapy is Imatinib [25]. Imatinib is designed to directly target and inhibit the BCR-ABL tyrosine kinase fusion that inscrease the survival of patients with acute myelogenous leukemia by more than 10 years [31].

The discovery of novel SL interactions, nevertheless, faces challenges. Finding novel synthetic lethality targets requires screening over every possible relevant gene pairs combination to identify SL candidates. Human genome is a highlysophisticated genomic structure with up to 3 billion base pairs, making up 20,000 genes in 23 different chromosomes. Manually experimenting all pairs of genes in human genome is costly and labour intensive. It is, hence, crucial design automatic computational methods to preliminarily screen for possible SL pairs. Another difficulty in the discovery process also lies in the diverse aspects from a specific disease. Human disease hardly comes from one reason but it rather comes from different factors such as errors during transcription and translation processes or genetic mutations (either indels or single nucleotide polymorphism). As a result, in addition to improving computational methods to reduce the SL discovery time, we also take into account introducing more biological contexts so that our results would be more reliable and meaningful. Our contribution to this work consists of the following:

(i) Curating a synthetic lethality interaction dataset from reliable biological experiments with every SL pair having a high statistical confident level >60%.
(ii) A multi-omics data pool for every gene in the pre-constructed interaction dataset covering five different types of omics, including transcriptomics, RNA interference, CRISPR screens, DNA sequences, and amino acid sequences. DNA sequences and amino acid sequences were preprocessed to match with their biological insights.
(iii) We benchmarked conventional machine learning and deep learning models for predicting synthetic lethality, specifically focusing on the integration of multi-omics data. Our approach utilizes a simple yet highly effective and

scalable multi-omics fusion strategy proficient at handling these potentially dissimilar data modalities. This fusion method demonstrated state-of-the-art performance with these specific multi-omics inputs, outperforming more complex architectures like Graph Neural Networks in identifying synthetic lethal interactions.

2 Literature review

2.1 Multi-omics machine learning

Multi-omics refers to the study of integrating multiple types of biological data, e.g., genomics, transcriptomics, metabolomics, or gene perturbations, to gain a multi-context understanding on a biological problem. Machine learning is a normal basis when it comes to effectively address the harmonic combination among different types of omics. This is due to their ability to capture high-level independent features and relationships among different types of data. The combination technique could be performed either data-side (by constructing a multi-omics graph network and performing a learning algorithm on it [40]) or model-side (by combining feature embeddings from multiple models [17]).

2.2 Predicting synthetic lethality

Efforts in discovering synthetic lethality pairs have been introduced in the literature, varying from biostatistical methods to modern deep learning methods. They aimed to reduce the workload of screening every possible gene pair across 20,000 human genes in the human genome. Original biostatistical methods often use gene inactivation or co-expression by Wilcoxon rank sum tests based on mutational data [16,20]. Drawbacks in those methods rely primarily on the extensive prior in-domain knowledge and massive efforts conducting sufficient data to perform these types of tests. A resorting approach for these limitations is to apply more advanced predictive methods such as machine learning. The biggest advantage of using machine learning models is that it does not require extensive computational effort but rather focuses on data preprocessing and problem formulation. The more data are collected, the better performance is acquired. Common computational biological methods for predicting synthetic lethality could be named as decision tree [41] and support vector machine [39].

3 Dataset construction

The topmost overview for the multi-omics data and interaction dataset for MOSL is illustrated in Figure 1. Our work includes a workflow to construct overview interaction dataset for MOSL and a workflow to supply multi-omics data for the given interaction dataset. There are three major omics data that are put into investigation in this work: gene perturbations (RNA interference and Clustered Regularly Interspaced Short Palindromic Repeats), genomics (DNA sequences and protein's primary structure), and transcriptomics.

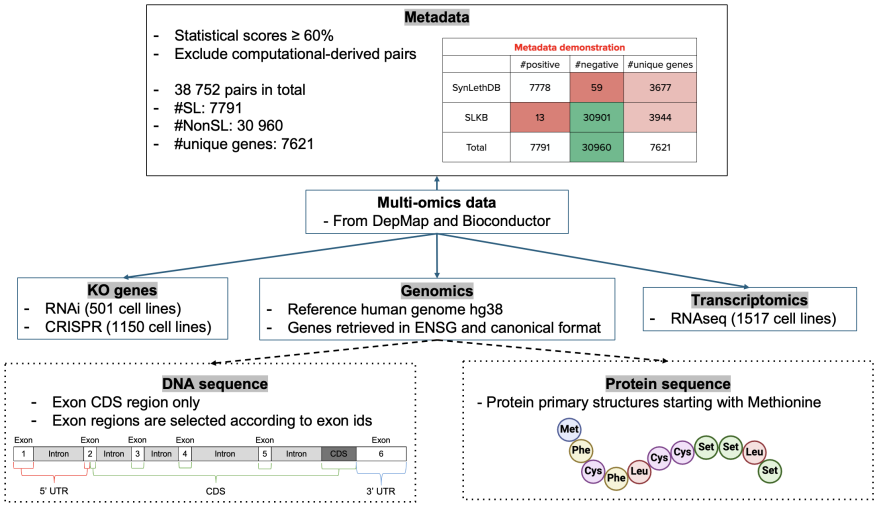

Fig. 1: Interaction dataset and multi-omics data overview for MOSL.

3.1 Synthetic lethality interaction dataset synthesis

We combine the two most popular synthetic lethality (SL) datasets: SynLethDB 2.0 and SLKB, to construct a general interaction dataset for predicting SL pairs.

SynLethDB 2.0 (SynLethDB2) [38], which is an extended version from the original SynLethDB [12], includes more than 50,000 SL pairs from 5 species (including *H. sapiens*, *M. musculus*, *D. melanogaster*, and *S. cerevisiae*). In this work, we concentrate more on human SL data in SynLethDB2 with 35,943 different SL and 2,899 non-SL pairs (a non-SL pair indicates that there is no synthetic lethality interaction between 2 genes) in human cells, encompassing relationships among 14,100 genes (9744 unique genes), 53 cancers, 1898 drugs. SynLethDB2 comes from 4 sources: (1) *in vitro* experiments (e.g., CRISPR-Cas9), (2) *in silico* experiments (computational methods), (3) published results from papers in PubMed using "synthetic lethal" as the querying keyword, and (4) public SL datasets (e.g., GenomeRNAi [32], BioGRID [30]).

Within SynLethDB2, 55.34% of all SL pairs (n=19,892/N=35,943) have a confidence score of more than 50%. Among SL pairs having an above average confidence level, 63.14% SL gene pairs (n=12,560/N=19,892) have a confidence interval in $(50\%; 75\%]$ and 36.81% SL gene pairs (n=7332/N=19,892) have a confidence interval larger than 75% with only 0.03% SL gene pairs (n=5, N=19,892) SL pairs have a confidence level of over 90%. However, more than half of the non-SL pairs (n=1623/N=2899) have a confidence score of less or equal to 10%, which is an extremely low confidence score for any biological research as in this work. Only 3.2% (n=93/N=2899) of the non-SL pairs have a confidence level of 50%. From the point of 50% confidence score, only 1.43% (n=38/N=2899) non-SL pairs are claimed to be confident for more than 70%, 0.66% (n=19/N=2899) pairs are 80% and 0.17% (n=5/N=2899) pairs are more

than 90%. The most frequent gene that appeared across SL pairs in SynLethDB2 is *KRAS* (n=2419/N=35,943; 6.73%), and 3610 unique genes (10.04%) appear only once across the dataset as a target of a certain SL pair. Gene *SNAPC1* appears the most across non-SL pairs (n=140/N=2499; 4.83%), and 94 unique genes (3.24%) appear only once across non-SL pairs in SynLethDB2.

On the other hand, SLKB dataset [11] focuses more on human non-SL data with 16,059 SL gene pairs and 264,424 non-SL gene pairs. SLKB SL data entirely comprises of results from wet-lab experiments from 11 gene Combination Double Knock Out (CDKO) studies within 6127 unique genes in 22 different cancer cell lines. A pair is considered as having a synthetic lethality interaction in SLKB only if it passes 3 per 5 predefined SL scores (e.g., GEMINI score, Horlbeck score). SLKB is a good complement for the interaction dataset. Since there are more non-SL pairs presented than SL ones, with a ratio of non-SL pairs of 94.27% (n=264,424/N=280,482). Across SLKB, up to 98.61% of SL pairs (n=15,835/N=16,059) and 69.85% (n=184,705/N=264,424) of non-SL pairs have a statistical score of less or equal to 10%. Only 0.12% SL pairs (n=20/N=16,059) and up to 14.21% (n=37,570/N=264,424) non-SL pairs have a confidence score from 50%. On the quantile from 70% to 100% confidence level, 9.23% (n=24,405/N=264,424) non-SL pairs and 0.04% (n=6/N=16,059) SL pairs have a confidence level of greater or equal to 70%. There are 17,365 (6.57%; N=280,482) non-SL gene pairs and 3 SL pairs (0.02%, N=16,059) at the statistical level of more than or equal to 80% with only 1 SL pair acquiring the 100% statistical score and 3,460 non-SL pairs achieving the same level of confidence.

To mitigate the severe class imbalance in synthetic lethality (SL) studies and to construct a high-quality dataset that ensures both robust data confidence and an adequate number of SL pairs for analysis, we implemented a multi-step curation pipeline. Initially, we combined SynLethDB2 and SLKB into a unified interaction resource. All duplicated SL and non-SL gene pairs in the fused interaction datasets were removed. Subsequently, we applied a confidence filter. We only retained interactions with a minimum confidence score of 0.6. Finally, we excluded all SL interactions that had been derived solely from computational predictions. This rigorous process ensures that our final interaction dataset consists exclusively of validated SL and non-SL relationships from wet-lab experiments, forming a reliable foundation for downstream modeling.

3.2 Transcriptomics and knock-out gene data for gene expression

Advances in genome engineering techniques such as RNA interference (RNAi) and CRISPR/Cas9 have a significant impact on SL screening in human cells. RNAi is a natural process in which RNA mediates sequence-specific gene silencing (knocking out) by inhibiting translation or transcription. Rather than operating at the post-transcriptional level in RNAi, CRISPR/Cas9 is usually designed to bring about functional knock-out genes at the nucleotide level. Both CRISPR and RNAi could be enabled to perform high-throughput SL screening. RNAseq, on the other hand, measures the transcriptomic profiles in cancer cell lines to study gene expression levels across different cancer cell lines, oncogene

activation, tumor suppressor loss, and many other aspects depending on research questions.

Cancer Dependency Map portal (DepMap), project Achilles and DRIVE [8] provides a dataset for population-based SL screening in multiple cancer cell lines. Experimental results appeared in DepMap come from diverse categories of wet-lab techniques and from a great number of cancer cell lines. We would like to investigate three types of common omics data in DepMap in this scope of work, including CRISPR, RNAi, and transcriptomics. In addition to the difference in the experimental design, each type of omics is conducted in a different number of cancer cell lines and the kinds of cancer cell lines: RNAi consists of gene expression data in 501 cancer cell lines, CRISPR gene perturbation consists of gene expression data in 1150 cancer cell lines, and transcriptomics profile consists of gene expression data in 1517 cancer cell lines.

3.3 Genomics data

Genomics is the study of the structure and function of an organism's genome. This work investigates genomic features that could enhance distinctive characteristics between SL and non-SL pairs. We employ two types of genomic features: DNA sequences and protein primary structures.

The Human Genome Project (HGP) is an international scientific research effort aiming to map and understand all the genes within the human species, collectively known as the genome. HGP sequenced entire human genomes from a predefined reference *H. sapiens* and comprises of more than 20,000 human genes. In this work, we leverage the reference human genome assembly hg38 patch 14 (GRCh38.p14) from HGP to collect gene transcripts via Bioconductor [10] and Ensembl database [26]. For every unique gene name from the aforeconstructed interaction dataset, we query its DNA sequence transcripts and amino acid transcripts, respectively. To maintain the purity of the genomics data, DNA sequences are collected solely in ENSG format, excluding gene names having more than 1 ENSG format, and amino acid sequences are collected in canonical form.

For DNA sequence data, there are four nucleotide characters: "A," "C," "T," and "G," which represent Adenine, Cytosine, Thymine, Guanine, and a marker for any unidentified nucleotides (denoted as character "N"), respectively. For the preconstructed interaction dataset, the average length for DNA sequences is 65,248 nucleotides. Gene *H4C7* has the least nucleotides with 386 base pairs, and gene *RBFOX1* has the most nucleotides with 2,473,539 base pairs. For protein primary structure data, there are 21 amino acid characters: 'N' (Asparagine), 'D' (Aspartic acid), 'Q' (Glutamine), 'P' (Proline), 'L' (Leucine), 'W' (Tryptophan), 'R' (Arginine), 'U' (Selenocysteine), 'T' (Threonine), 'M' (Methionine), 'A' (Alanine), 'C' (Cysteine), 'Y' (Tyrosine), 'E' (Glutamic acid), 'H' (Histidine), 'F' (Phenylalanine), 'G' (Glycine), 'I' (Isoleucine), 'S' (Serine), 'V' (Valine), 'K' (Lysine) and an arbitrary character '$*$' to indicate an unidentified amino acid within a gene transcript. Across all unique genes in the interaction dataset, the

mean amino acid sequence length is 649, with gene *TMSB10* having the shortest amino acid sequence with 44 amino acids, and gene *TTN* having the longest sequence length with 35,991 amino acids.

To match the physiological nature of the human body, we conduct two preprocessing steps:

1. Since all protein sequences start with the Methionine amino acid (the letter "M" in the amino acid alphabet) [21], we excluded all genes that do not meet this criterion.
2. To reduce the high dimensionality for DNA data (the longest DNA sequence in interaction dataset has up to more than 2.4 million base pairs), we only select exon coding sequences (exon CDS) inside the region from a start codon to a stop codon due to the high informative concentration toward the transcription and translation process. Figure 2 shows different components within a DNA transcript. Final DNA sequences for DNA genomics data are transcripts queried from any given gene names after removing 5' UTR segments, 3' UTR segments, and intron CDS regions.

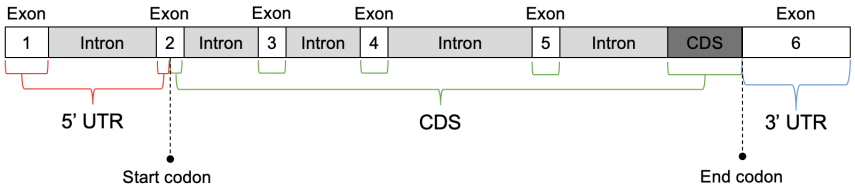

Fig. 2: A DNA transcript comprises of a 5' untranslated region (5' UTR), a coding sequence region (CDS), and a 3' untranslated region (3' UTR) [2] We only focus on exon segments in the CDS region so that we sustain important information for a gene transcript while downsampling the number of nucleotides needed to be processed.

4 Predictive methods and benchmarking

The final overall interaction dataset for predicting synthetic lethality in this work after being processed to be suitable for synthesized multi-omics data is shown in Table 1.

4.1 Problem formulation

Given a gene set $\mathcal{G} = \{g_i\}_{i=1}^{N}$ with N as the number of total unique genes appearing in any SL interaction dataset. We define a set of genetic interactions

	#SL positive	#SL negative	#unique genes
SynLethDB	7778	59	3677
SLKB	13	30901	3944
Total	7791	30960	5610

Table 1: Overall metadata for MOSL. All SL and non-SL gene pairs are collected from SynLethDB 2.0 and SLKB with confidence levels of equal to or more than 60%, excluding computational methods.

$\mathcal{R} = \{(g_i, g_j) \mid g_i, g_j \sim \mathcal{G}\} \in \{0, 1\}$, such that:

$$(g_i, g_j) = \begin{cases} 1 & \text{if there exists a SL interaction} \\ 0 & \text{otherwise} \end{cases}$$

For every gene $g \in \mathcal{G}$, we construct a multi-omics feature space Ω_g, with $\mathcal{D}_{g_{aa}}$, $\mathcal{D}_{g_{nu}}$, $\mathcal{D}_{g_{CRISPR}}$, $\mathcal{D}_{g_{RNAi}}$, $\mathcal{D}_{g_{rnaseq}}$ respectively denoting the space of amino acid sequences, nucleotide sequences, CRISPR gene expression, RNAi gene expression, and transcriptomics data. The genomic data space is a textual space; each value in that space is either a natural language character or an ambiguous character $*$.

$$\Omega_g = \{\mathcal{D}_{g_{aa}}, \mathcal{D}_{g_{nu}}, \mathcal{D}_{g_{CRISPR}}, \mathcal{D}_{g_{RNAi}}, \mathcal{D}_{g_{rnaseq}} \mid g \sim \mathcal{G}\}$$

such that:

$$\mathcal{D}_{g_{nu}} \in \{A, C, T, G, *\}^{N_{nu\_g_i}}$$

$$\mathcal{D}_{g_{aa}} \in \{N, D, Q, P, L, *, W, R, U, T, M, A, C, Y, E, H, F, G, I, S, V, K\}^{N_{aa\_g_i}}$$

$$\mathcal{D}_{g_{CRISPR}} \in \mathbb{R}^{1150}; \quad \mathcal{D}_{g_{RNAi}} \in \mathbb{R}^{501}; \quad \mathcal{D}_{g_{rnaseq}} \in \mathbb{R}^{1517}$$

and $N_{nu\_g_i}$ is the number of nucleotide characters in a gene, $N_{aa\_g_i}$ is the number of amino acid characters in that gene.

In this work, our ultimate goal is tailoring a functional mapping with a set of trainable parameters θ as the predictive model $f_\theta\left(y_i \mid g_i, g_j, \Omega_i, \Omega_j\right)$ where $y_i \in \{0; 1\}$ is the SL prediction for a given gene pair (g_i, g_j):

$$f_\theta : (g_i \times \Omega_i, g_j \times \Omega_j) \mapsto \{0; 1\}$$

4.2 MOSL: Multi-omics Synthetic Lethality

Our data includes up to five different modalities derived from only one gene. Hence, it is vital to construct a multi-component feature extractor to independently and effectively extract important information from diverse contexts. Then, we have to map all five feature spaces into one joint dimension so that our later model can learn the extracted features more efficiently.

In terms of data, MOSL consists of two kinds of data modality: tabular data (from transcriptomics, gene perturbations) and natural language data (from genomics). We could easily represent any tabular data by just converting each row

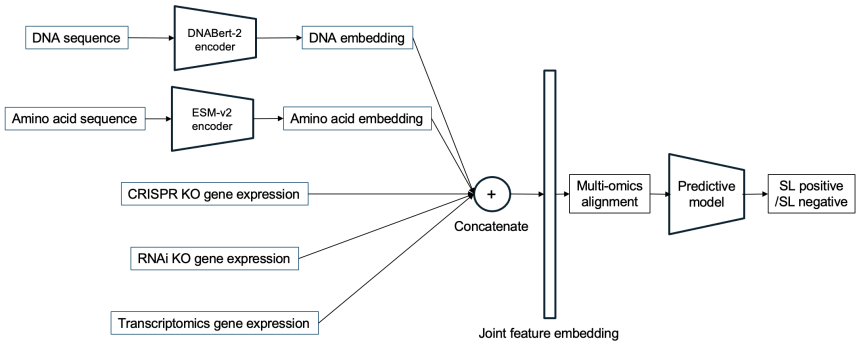

Fig. 3: MOSL framework to integrate multi-omics data. We can scale up this framework into as many omics data as possible. For the sake of simplicity, in this work, we only use normalization for the multi-omics alignment step, but we could extend it into other advanced techniques like cross-attention or optimal transport.

into a numeric vector and forwarding it to any computational models. Natural language data, on the other hand, is a completely different story. Different string-like data require different models to capture their high-level semantic meaning from the data themselves. We need to tailor suitable models to make the feature extraction process perform as best as possible, or else the extracted features might dilute even the baseline performance. The feature extractor should be both a large model and trained specifically for the respective task so that it has sufficient parameters and contexts to capture all the important knowledge from the training data. In this work, we use two language models to extract sequential genomic characteristics, which are DNABERT-2 [43] to extract DNA sequences and ESM-v2 [22] to extract amino acid sequences.

DNABERT-2 [43] is a Transformer-based [37] architecture for genomic sequence modeling. It uses Byte Pair Encoding [33] for efficient tokenization and replaces positional embeddings with attention-based mechanisms, allowing it to process longer DNA sequences and better capture long-range dependencies in genomic data. ESM-2 [22] is a protein language model trained with a masked language modeling objective [9], utilizing an attention mechanism to learn interaction patterns between amino acid pairs in the input sequence. Pre-trained on Uniref [35], ESM-2 advances protein language modeling by achieving higher performance in structure prediction and unsupervised contact prediction compared to previous SOTA methods such as Prot-T5 and ESM-1b. Their empirical studies reveal that training on large-scale datasets for unsupervised learning tasks leads to performance improvements as model size increases. Both DNABERT-2 and ESM-2 leverage transformer architectures and large-scale datasets to extract meaningful sequence embeddings, effectively scaling with longer inputs. In this work, we use ESM-2 (650M parameters version) and DNABERT-2 (117M parameters version) to balance computational efficiency and performance.

4.3 Experimental settings and evaluation metrics

We use two standard metrics used in the medical and biological domains to assess the effectiveness of MOSL: specificity and sensitivity. Specificity measures the precision in the decision-making process from our models when it comes to predicting whether a person is in the healthy cohort or not. Sensitivity, on the other hand, quantifies the number of cancer patients that are correctly predicted. These two metrics have a trade-off to each other since a high specificity often means a lower sensitivity and vice versa. It is, hence, important to construct a predictive method that could balance both specificity and sensitivity. Additionally, to make further comparison across *in silico* experiments, we also use accuracy score to indicate the total correct predictions that benchmarked models make.

We evaluate the effectiveness of integrating multi-omics data for predicting synthetic lethality by benchmarking various predictive models. Our study includes traditional machine learning methods, e.g., Logistic Regression and Support Vector Machines (SVM) [14], alongside ensemble-based techniques such as Random Forest [4] and XGBoost [6]. Additionally, we incorporate deep learning approaches, particularly TabNet [1]. TabNet directly processes raw tabular data, utilizing a sequential attention mechanism to highlight the most relevant features at each decision step. By leveraging the attention mechanism to learn from tabular data, TabNet could be used to interpret data insights and extract the level of feature importance contributing to the final prediction.

For single-omic analysis, we assess individual data types separately. For every omic data in transcriptomics, RNAi, and CRISPR, we treat each individual gene expression in a cancer cell line as a numeric value. As a result, one gene name in an omic could be represented as a feature vector. For genomics data, due to the nature of textual data, we follow a two-step prediction process. At first, we leverage language models to extract features. We use DNABERT-2 to extract nucleotide sequences into a 768-dimensional numeric vector and ESM-2 to extract amino acid sequences into a 2048-dimensional numeric vector. We treat embedded features as inputs for machine learning models to make further predictions afterwards.

With multi-omics analysis, as different feature sets originate from distinct embedding spaces, we explore two fusion strategies to combine multiple data sources together:

1. Cross-omics fusion: We perform a straightforward approach where features from different sources are concatenated row-wise and applied to a multi-omic alignment operation. For the sake of simplicity, in this work, we only use normalization for the multi-omics alignment step, but we could extend into other advanced alignment techniques like cross-attention [42], optimal transport [19], or multi-graph alignment [28].
2. Graph Neural Networks (GNNs): A more structured approach where features are treated as nodes in a graph, and relationships between them are learned using a message-passing mechanism. In this framework, we formulate a link prediction task to predict relationships between SL interaction (SL positive

or SL negative). Here, we employ a simple graph-based network including two message passing layers as inspired in [18].

5 Results

The detailed results for our method are shown in Table 2. Despite efforts to reduce the disproportion, our interaction dataset remains imbalanced with only 7,791 SL positive gene pairs, but there are 30,960 negative (the ratio between positive and negative is around 1:4). It is, consequently, still a non-trivial problem to balance between positive and negative predictions.

For single-omic results, genomics brings out the highest performance with a specificity of 98.95% and an accuracy score of 97.17%, compared to transcriptomics and gene perturbations. This is understandable since genomics data underwent a carefully curating process leveraging biological knowledge. We only kept highly meaningful genomics sequences, which led to the purity of the genomics data. Additionally, we conduct a 2-stage predicting pipeline for genomics data. The first one is to use a large in-context language model specifically designed to extract genetic features from genomics sequences (ESM-v2 for protein and DNABERT-2 for DNA). Afterwards, we used extracted feature embeddings to make SL predictions by applying machine learning models. This method is far more sophisticated than treating transcriptomics and gene perturbation's tabular data as input tensors and forwarding them directly into predictive models.

Multi-omics help our predictive models increase the specificity significantly. When we combine solely tabular data, which is transcriptomics, RNAi, and CRISPR, Logistic Regression results in the top specificity of up to 99%, compared to the best specificity in transcriptomics data of 97.84%, gene perturbation data of 98.23%, and 98.95% in genomics data. This is crucial for applying biomedical products to the public. Higher specificity usually means a low false discovery rate and leads to fewer people having false positive results. 99% specificity implies that only one person receives a false positive result in a cohort of 100 people. When it comes to combining more omics data to the original omics pool, we even improve the previous specificity up to 0.75%. This might be a low, increasing rate. However, this almost meets the absolute specificity percentage that any model could achieve. The best-performing predictive model also has the highest accuracy score of 97.88%, which is higher than the second best accuracy score of 0.71% (from 97.17% in ESM-2 as the feature extractor and XGBoost as the predictive model to 97.88% in the SVM multi-omics after combining every possible omics data in our data pool).

It is an abnormal insight on Table 2 that graph neural network achieves a lower performance compared to other predictive methods in both gene perturbations and genomics data. This could be reasoned via the graph construction process. Although node-level features have a reasonable dimension, as aforementioned, we have up to more than 30,000 SL negative pairs but only 7,791 SL positive pairs. When we construct a link-prediction graph neural network, the resulting graph is extremely sparse, with only around 20% of the graph nodes

having a connecting edge, and up to around 80% of the nodes not having any edges connected to them.

Predictive model	Omics	Feature type	SPEC (5615)	SEN (1436)	ACC (7051)
Logistic Regression	Transcriptomics	RNAseq	97.54	87.51	95.50
SVM			84.04	54.38	77.99
Random Forest			97.76	87.91	**95.75**
XGBoost			**97.84**	86.52	95.53
TabNet			93.35	**93.42**	93.36
Logistic Regression	Gene pertubation	RNAi + CRISPR	**98.23**	87.51	**96.05**
SVM			89.61	69.65	85.54
Random Forest			97.94	87.91	95.90
XGBoost			97.88	87.78	95.82
Graph Neural Network			45.87	83.04	53.44
TabNet			91.44	***95.41***	92.12
DNABERT-2 (117M) + Logistic Regression	Genomics	DNA sequence	95.62	71.83	90.77
Graph Neural Network			50.60	82.93	57.18
DNABERT-2 (117M) + XGBoost			98.53	**90.84**	96.96
ESM-2 + XGBoost		Amino acid (AA) sequence	98.82	90.72	**97.17**
ESM-2 + DNABERT-2 + Logistic Regression		DNA + AA embeddings	**98.95**	90.11	97.14
ESM-2 + DNABERT-2 + XGBoost			98.50	90.00	96.77
Logistic Regression	Multi-omics	RNAi + CRISPR + RNAseq + DNA sequences + Amino acid sequences	**99.00**	86.60	**96.47**
SVM			96.70	72.50	91.77
Random Forest			97.93	**88.18**	95.94
XGBoost			97.98	87.64	95.87
TabNet			96.90	88.10	95.11
Logistic Regression			95.74	78.72	92.27
SVM			***99.75***	90.55	***97.88***
Random Forest			95.74	75.00	91.52
XGBoost			94.32	78.96	91.19
TabNet			98.44	**92.75**	97.28

Table 2: Benchmarking results to assess the effectiveness of using multi-omics data in predicting synthetic lethality. **Bold** results indicate the highest results on a single-omic category. ***Bold and italic*** results indicate highest results across the total benchmark.

6 Conclusion

Throughout this work, we created a large synthetic lethality interaction dataset including 7781 SL positive gene pairs and 30,960 SL negative gene pairs. We also synthesized and produced biological-inspired preprocessing techniques to build multi-omics data for all genes in the pre-constructed interaction dataset.

We collected transcriptomic profiles and gene perturbation gene expression from DepMap as tabular data. We leveraged Bioconductor, Ensembl, and Human Genome Project to query nucleotide sequences and amino acid sequences from a reference human genome with a biological-inspired process to sustain only informative gene segments. We also benchmarked conventional methods (varying from machine learning, deep learning, to graph-based models) for performance comparison using different types of single-omic and multi-omics combination to produce larger contexts for our predictive models. Experimental results show that using multi-omics data leads to the most effective performance with a top-tier specificity of 99.75% and an accuracy score of 97.88% in predicting synthetic lethality interaction.

Acknowledgments. Tan Pham would like to thank Vingroup Innovation Foundation (VinIF) for Master's training scholarship program. We would like to thank AISIA Lab and the University of Science, Vietnam National University, for their support.

References

1. Arik, S.Ö., Pfister, T.: Tabnet: Attentive interpretable tabular learning. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 6679–6687 (2021)
2. Aspden, J.L., Wallace, E.W., Whiffin, N.: Not all exons are protein coding: Addressing a common misconception. Cell genomics **3**(4) (2023)
3. Bender, A., Pringle, J.R.: Use of a screen for synthetic lethal and multicopy suppressee mutants to identify two new genes involved in morphogenesis in saccharomyces cerevisiae. Molecular and cellular biology (1991)
4. Breiman, L.: Random forests. Machine learning **45**, 5–32 (2001)
5. Butler, S.K.: Niraparib (zejula®). Oncology Times **40**(14), 20 (2018)
6. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp. 785–794 (2016)
7. De Kegel, B., Quinn, N., Thompson, N.A., Adams, D.J., Ryan, C.J.: Comprehensive prediction of robust synthetic lethality between paralog pairs in cancer cell lines. Cell Systems **12**(12), 1144–1159 (2021)
8. DepMap, B.: DepMap 24Q4 Public (12 2024). `https://doi.org/10.25452/figshare.plus.27993248.v1`, `https://plus.figshare.com/articles/dataset/DepMap_24Q4_Public/27993248`
9. Devlin, J.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
10. Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al.: Bioconductor: open software development for computational biology and bioinformatics. Genome biology **5**, 1–16 (2004)
11. Gökbağ, B., Tang, S., Fan, K., Cheng, L., Li, L.: Slkb: Synthetic lethality knowledge base for gene combination double knockout experiments. Cancer Research **83**(7_Supplement), 6581–6581 (2023)
12. Guo, J., Liu, H., Zheng, J.: Synlethdb: synthetic lethality database toward discovery of selective and sensitive anticancer drug targets. Nucleic acids research **44**(D1), D1011–D1017 (2016)

13. Hartwell, L.H., Szankasi, P., Roberts, C.J., Murray, A.W., Friend, S.H.: Integrating genetic approaches into the discovery of anticancer drugs. Science **278**(5340), 1064–1068 (1997)
14. Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J., Scholkopf, B.: Support vector machines. IEEE Intelligent Systems and their applications **13**(4), 18–28 (1998)
15. Hoy, S.M.: Talazoparib: first global approval. Drugs **78**(18), 1939–1946 (2018)
16. Jerby-Arnon, L., Pfetzer, N., Waldman, Y.Y., McGarry, L., James, D., Shanks, E., Seashore-Ludlow, B., Weinstock, A., Geiger, T., Clemons, P.A., et al.: Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality. Cell **158**(5), 1199–1209 (2014)
17. Kang, M., Ko, E., Mersha, T.B.: A roadmap for multi-omics data integration using deep learning. Briefings in Bioinformatics **23**(1), bbab454 (2022)
18. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
19. Lee, J., Dabagia, M., Dyer, E., Rozell, C.: Hierarchical optimal transport for multimodal distribution alignment. Advances in neural information processing systems **32** (2019)
20. Lee, J.S., Das, A., Jerby-Arnon, L., Arafeh, R., Auslander, N., Davidson, M., McGarry, L., James, D., Amzallag, A., Park, S.G., et al.: Harnessing synthetic lethality to predict the response to cancer treatment. Nature communications **9**(1), 2546 (2018)
21. Lim, J.M., Kim, G., Levine, R.L.: Methionine in proteins: it's not just for protein initiation anymore. Neurochemical research **44**, 247–257 (2019)
22. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., et al.: Language models of protein sequences at the scale of evolution enable accurate structure prediction. bioRxiv (2022)
23. Lohr, L.: Rucaparib (rubraca®). Oncology Times **39**(6), 19 (2017)
24. Lucchesi, J.C.: Synthetic lethality and semi-lethality among functionally related mutants of drosophila melanogaster. Genetics **59**(1), 37 (1968)
25. Manley, P., Cowan-Jacob, S., Buchdunger, E., Fabbro, D., Fendrich, G., Furet, P., Meyer, T., Zimmermann, J.: Imatinib: a selective tyrosine kinase inhibitor. European journal of cancer **38**, S19–S27 (2002)
26. Martin, F.J., Amode, M.R., Aneja, A., Austine-Orimoloye, O., Azov, A.G., Barnes, I., Becker, A., Bennett, R., Berry, A., Bhai, J., et al.: Ensembl 2023. Nucleic acids research **51**(D1), D933–D941 (2023)
27. McManus, K.J., Barrett, I.J., Nouhi, Y., Hieter, P.: Specific synthetic lethal killing of rad54b-deficient human colorectal cancer cells by fen1 silencing. Proceedings of the National Academy of Sciences **106**(9), 3276–3281 (2009)
28. Nguyen, D.M.H., Diep, N.T., Nguyen, T.Q., Le, H.B., Nguyen, T., Nguyen, T., Nguyen, T., Ho, N., Xie, P., Wattenhofer, R., Zhou, J., Sonntag, D., Niepert, M.: Logra-med: Long context multi-graph alignment for medical vision-language model (2024), https://arxiv.org/abs/2410.02615
29. O'Neil, N.J., Bailey, M.L., Hieter, P.: Synthetic lethality and cancer. Nature Reviews Genetics **18**(10), 613–623 (2017)
30. Oughtred, R., Rust, J., Chang, C., Breitkreutz, B.J., Stark, C., Willems, A., Boucher, L., Leung, G., Kolas, N., Zhang, F., et al.: The biogrid database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Science **30**(1), 187–200 (2021)
31. Sacha, T.: Imatinib in chronic myeloid leukemia: an overview. Mediterranean journal of hematology and infectious diseases **6**(1) (2014)

32. Schmidt, E.E., Pelz, O., Buhlmann, S., Kerr, G., Horn, T., Boutros, M.: Genomernai: a database for cell-based and in vivo rnai phenotypes, 2013 update. Nucleic acids research **41**(D1), D1021–D1026 (2013)
33. Sennrich, R.: Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909 (2015)
34. Sturtevant, A.: A highly specific complementary lethal system in drosophila melanogaster. Genetics **41**(1), 118 (1956)
35. Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., Wu, C.H., Consortium, U.: Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics **31**(6), 926–932 (2015)
36. Tang, S., Gökbağ, B., Fan, K., Shao, S., Huo, Y., Wu, X., Cheng, L., Li, L.: Synthetic lethal gene pairs: Experimental approaches and predictive models. Frontiers in Genetics **13**, 961611 (2022)
37. Vaswani, A.: Attention is all you need. Advances in Neural Information Processing Systems (2017)
38. Wang, J., Wu, M., Huang, X., Wang, L., Zhang, S., Liu, H., Zheng, J.: Synlethdb 2.0: a web-based knowledge graph database on synthetic lethality for novel anticancer drug discovery. Database **2022**, baac030 (2022)
39. Wu, M., Li, X., Zhang, F., Li, X., Kwoh, C.K., Zheng, J.: In silico prediction of synthetic lethality by meta-analysis of genetic interactions, functions, and pathways in yeast and human cancer. Cancer informatics **13**, CIN–S14026 (2014)
40. Xiao, S., Lin, H., Wang, C., Wang, S., Rajapakse, J.C.: Graph neural networks with multiple prior knowledge for multi-omics data analysis. IEEE Journal of Biomedical and Health Informatics **27**(9), 4591–4600 (2023)
41. Yin, Z., Qian, B., Yang, G., Guo, L.: Predicting synthetic lethal genetic interactions in breast cancer using decision tree. In: Proceedings of the 2019 6th International Conference on Biomedical and Bioinformatics Engineering. pp. 1–6 (2019)
42. Zhou, X., Pappas, N., Smith, N.A.: Multilevel text alignment with cross-document attention. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 5012–5025. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.407, https://aclanthology.org/2020.emnlp-main.407/
43. Zhou, Z., Ji, Y., Li, W., Dutta, P., Davuluri, R., Liu, H.: Dnabert-2: Efficient foundation model and benchmark for multi-species genome (2023)