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

Tan Pham, Dang Vu, Tien Dang, Binh Nguyen, and Tuan-Anh Tran

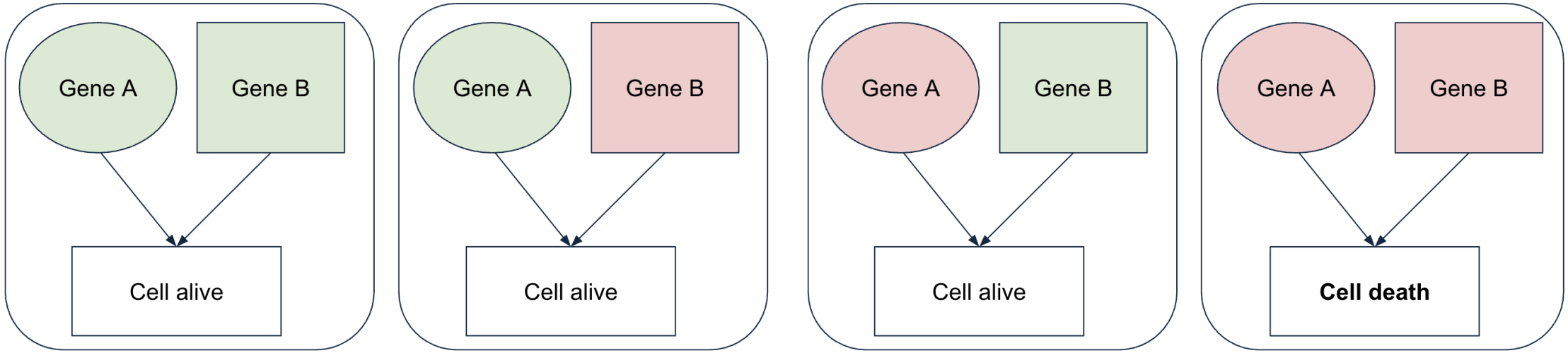
University of Science, Vietnam National University

Institut Pasteur, France

17th International Conference on Computational Collective Intelligence

Table of contents

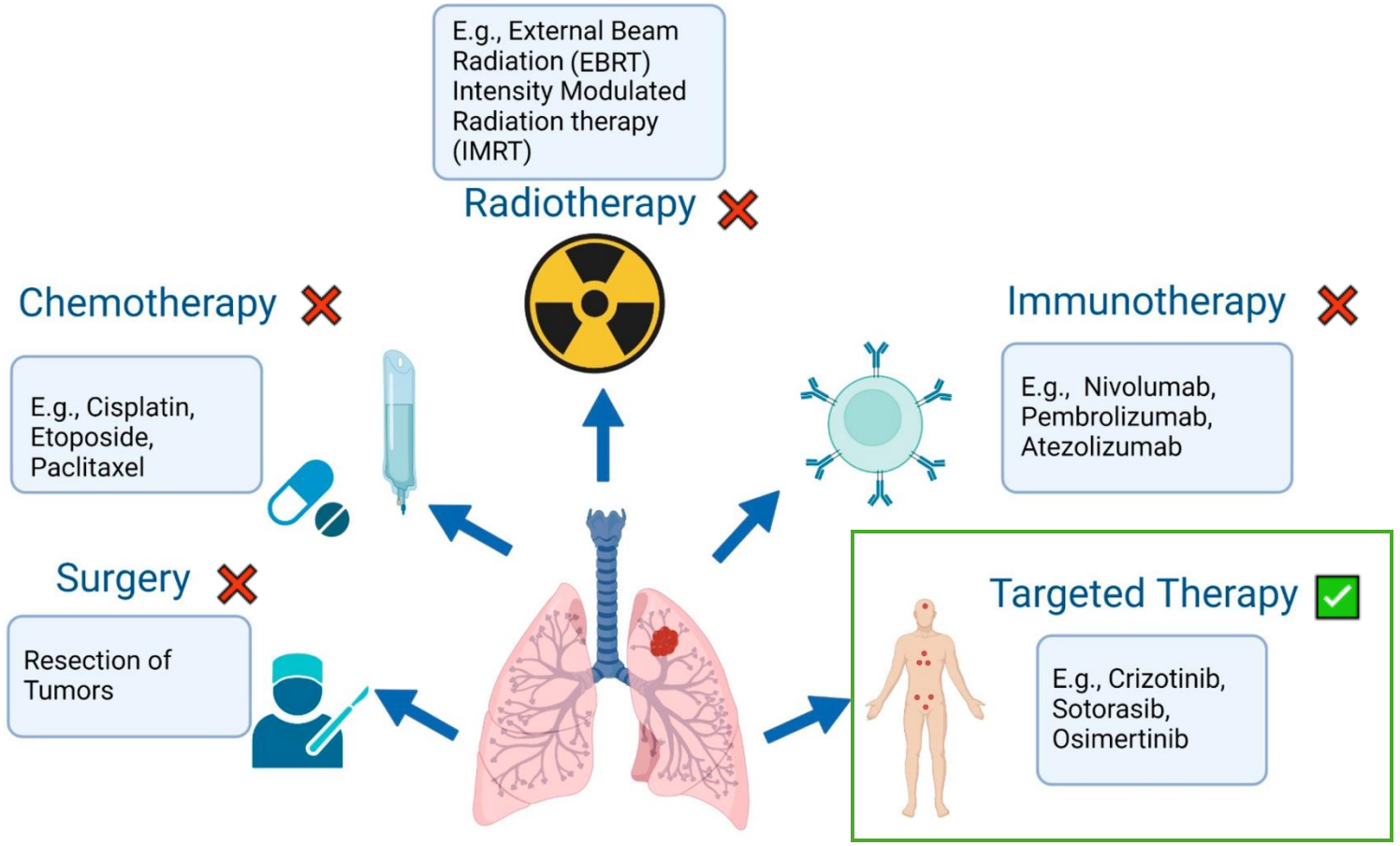
Table of contents



- Synthetic lethality (SL) refers to a genetic interaction in which:
  - The **simultaneous perturbation of two genes** leads to cell or organism death
  - Whereas viability is maintained when **only one of the pair is altered**

Tang, Shan, et al. "Synthetic lethal gene pairs: Experimental approaches and predictive models."*Frontiers in Genetics* 13 (2022): 961611.

Common therapies in oncology



Cancer treatment options

- Traditional treatments are either **costly** or **effective but damage healthy tissues** along with cancer cells
  → serious side effects and reduced quality of life.
- Genetic targeted therapies are designed to act specifically on mutations that drive tumour growth.

➔ Offering more precise and effective treatment with fewer off-target effects.

➔ Need to improve more to reduce the treatment fees.

[1] Side Effects of Cancer treatment, National Cancer Institute, National Institutes of Health (NIH), https://www.cancer.gov/about-cancer/treatment/side-effects

Current advances in targeted therapy:

- Imbalanced dataset, usually relies on 1-class classification models

  → Low performance

- Currently, recent advances only make prediction based on single-omic only.

  → Lack context of biophysics, biochemistry, physiology, metabolism in a living organism → Less reliable

Research questions

1. How to **reduce imbalance** in designing computational methods for synthetic lethality discovery ?
2. How to construct a **multi-omics data** in predicting synthetic lethality in cancer cell lines ?
3. How to **leverage** multi-omics data to **enhance** the quality for synthetic lethality interaction **prediction**.

Table of contents

- A gene set $\mathcal{G} = \{g_i\}_{i=1}^{N}$ with N is the number of genes in the synthesized dataset.
- A set $\mathcal{R} = \{(g_i, g_j) \mid g_i, g_j \sim \mathcal{G}\}$ genetic interactions.
- $\forall g \in \mathcal{G}$,

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

- Ω_g is the omics space for each gene $g \in \mathcal{G}$.
- We want to design a function mapping $f_\theta(y_i \mid g_i, g_j, \Omega_i, \Omega_j)$ to a binary space, so that:

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

- For each $g \in \mathcal{G}$, I construct a multi-omics space Ω_g:

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

  such that, $\mathcal{D}_{g_{aa}}$, $\mathcal{D}_{g_{nu}}$, $\mathcal{D}_{g_{CRISPR}}$, $\mathcal{D}_{g_{RNAi}}$, $\mathcal{D}_{g_{rnaseq}}$ respectively denotes the space of protein transcripts, DNA transcripts, CRISPR KO gene dependency, RNAi KO gene dependency and transcriptomics expressions.

- For $\mathcal{D}_{g_i} \in \Omega_g$:

$$\mathcal{D}_{g_{nu}} \in \{A, C, T, G, N\}^{N_{nu_{g_i}}}$$

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

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

Table of contents

SL genetic interaction dataset: SynLethDB 2.0 and SLKB

Table 2. Comparison of Statistics Between Two Versions of SynLethDB.

	SynLethDB 1.0	SynLethDB 2.0
# Human SLs	19 952	35 943
# Mouse SLs	366	381
# Fly SLs	423	439
# Worm SLs	105	105
# Yeast SLs	13 241	14 000
KG	No	Yes
Annotation	SLs only	Yes
Offline dataset	Yes	Yes
RESTful APIs	No	Yes

n1.name	n2.name	r.pubmed_id	r.source	r.statistic_score	SL_or_not
KRAS	USP39	19490893	Synlethality;GenomeRNAi	0.8150	1
ATM	BRCA1	3965078	Decipher;Text Mining	0.7023	0
MAPK1	MYC	22157079	GenomeRNAi;Text Mining	0.8500	1
MST1R	RPS6	20959493	Synlethality;Text Mining	0.8500	1
CSK	PAK2	29987050	High ThroughputlLow Throughput	0.9000	0
CHEK2	MTOR	31300006	CRISPR/CRISPRi	0.8500	1
KRAS	WDR83	27655641	RNAi Screen	0.7500	0
FANCD2	PARP2	28628639	Low Throughput	0.8000	0

SynLethDB-2

Table 3. Contents of synthetic lethality knowledge base (SLKB)

Category	Stored content
SLKB study reported content	
No. of SL CDKO studies	11 (currently counts available for 10)
No. of sgRNA guides + sequences	45,430 sgRNAs (1802 controls)
No. of sgRNA pair counts	3,578,017 sgRNA pairs
No. of unique genes and gene pairs	6,127 genes/280,483 gene pairs (148,040 unique)
No. of SL pairs (originally reported)	16,059 gene pairs
No. of non-SL pairs (originally reported)	264,424 gene pairs
SLKB processed content	
No. of unique genes and gene pairs (calculated scores)	6,124 genes/261,958 gene pairs (127,688 unique)
No.# of SL pairs (majority vote)	13,173 gene pairs
No. of non-SL pairs (majority vote)	248,785 gene pairs

gene_1	gene_2	SL_or_not	statistical_score
A3GALT2	ABO	Not SL	0.921199
A3GALT2	GBGT1	Not SL	0.751001
AADAC	AADACL4	Not SL	0.828146
AADACL2	AADACL4	Not SL	0.780282
AADACL3	AADACL4	Not SL	0.739161

SLKB

[1] 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 (Oxford). 2022 May 13;2022:baac030*. doi: 10.1093/database/baac030. PMID: 35562840; PMCID: PMC9216587.
[2] Gökbağ B, Tang S, Fan K, Cheng L, Yu L, Zhao Y, Li L. SLKB: synthetic lethality knowledge base. *Nucleic Acids Res. 2024 Jan 5;52(D1):D1418-D1428*. doi: 10.1093/nar/gkad806. PMID: 37889037; PMCID: PMC10767912.

Data synthesis and preprocessing

Metadata

- Statistical scores ≥ 60%
- Exclude computational-derived pairs

- 38 751 pairs in total
- #SL: 7791
- #NonSL: 30 960
- #unique genes: 7621

Metadata demonstration

	#positive	#negative	#unique genes
SynLethDB	7778	59	3677
SLKB	13	30901	3944
Total	7791	30960	7621

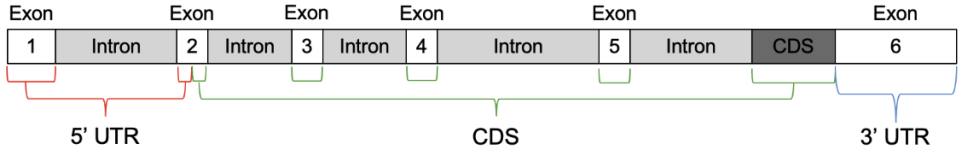
Multi-omics data
- From DepMap and Bioconductor

KO genes
- RNAi (501 cell lines)
- CRISPR (1150 cell lines)

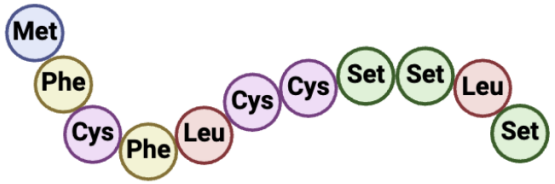
Genomics
- Reference human genome hg38
- Genes retrieved in ENSG and canonical format

Transcriptomics
- RNAseq (1517 cell lines)

DNA sequence
- Exon CDS region only
- Exon regions are selected according to exon ids



Protein sequence
- Protein primary structures starting with Methionine



Data synthesis and preprocessing

gene_name	protein_sequence	gene_sequence
KRAS	MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVI...	GGTGTTGATGATGCCTTCTATACATTAGTTCGAGAAATTCGAAAAC...
CHEK2	MSRESDVEAQQSHGSSACSQPHGSVTQSQGSSSQSQGISSSSTSTM...	CCTTCTACTAGTCGAAAGCGGCCCCGTGAAGGGGAAGCCGAGGGTG...
CNOT1	MNLDSLSLALSQISYLVDNLTKKNYRASQQEIQHIVNRHGPEADRH...	GTTATTCCAGTCGGTCGCACAGTGCTGCATGGGACAGAAGCAGGCC...
ALK	MGAIGLLWLLPLLLSTAAVGSGMGTGQRAGSPAAGPPLQPREPLSY...	GACCCGGATGTAATCAACACCGCTTTGCCGATAGAATATGGTCCAC...
KDM1A	MLSGKKAAAAAAAAAAAATGTEAGPGTAGGSENGSEVAAQPAGLSG...	GGCGCGTGCGTACGCGACGGCGGTTGGCGGCGCGCGGGCAGCGTGA...

RNAseq_CCLE	TSPAN6 (7105)	TNMD (64102)	DPM1 (8813)	SCYL3 (57147)	C1orf112 (55732)	FGR (2268)	CFH (3075)	FUCA2 (2519)	GCLC (2729)
ACH-001113	4.331992	0.000000	7.364660	2.792855	4.471187	0.028569	1.226509	3.044394	6.500005
ACH-001289	4.567424	0.584963	7.106641	2.543496	3.504620	0.000000	0.189034	3.813525	4.221877
ACH-001339	3.150560	0.000000	7.379118	2.333424	4.228049	0.056584	1.310340	6.687201	3.682573
ACH-001538	5.085340	0.000000	7.154211	2.545968	3.084064	0.000000	5.868390	6.165309	4.489928
ACH-000242	6.729417	0.000000	6.537917	2.456806	3.867896	0.799087	7.208478	5.570159	7.127117

RNAi_CCLE	143B_BONE	22RV1_PROSTATE	2313287_STOMACH	697_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
A1BG (1)	0.052466	-0.115242	-0.023172	-0.023337
NAT2 (10)	0.084173	0.000951	-0.154188	-0.079006
ADA (100)	0.207020	0.010743	-0.072102	0.045611
CDH2 (1000)	0.062192	-0.049809	0.022137	0.061709
AKT3 (10000)	0.039280	-0.076596	0.136445	0.154167

CRISPR_CCLE	A1BG (1)	A1CF (29974)	A2M (2)	A2ML1 (144568)	A3GALT2 (127550)	A4GALT (53947)	A4GNT (51146)	AAAS (8086)	AACS (65985)
ACH-000001	-0.134132	0.029103	0.016454	-0.137540	-0.047273	0.181367	-0.082437	-0.059023	0.194592
ACH-000004	-0.001436	-0.080068	-0.125263	-0.027607	-0.053838	-0.151272	0.240094	-0.038922	0.186438
ACH-000005	-0.144940	0.026541	0.160605	0.088015	-0.202605	-0.243420	0.133726	-0.034895	-0.126105
ACH-000007	-0.053334	-0.120420	0.047978	0.086984	-0.018987	-0.017309	-0.000041	-0.158419	-0.169559
ACH-000009	-0.027684	-0.144202	0.052846	0.073833	0.038823	-0.108149	0.010811	-0.088600	0.032194

Table of contents

DNABERT-2 tokenizer: BPE tokenization

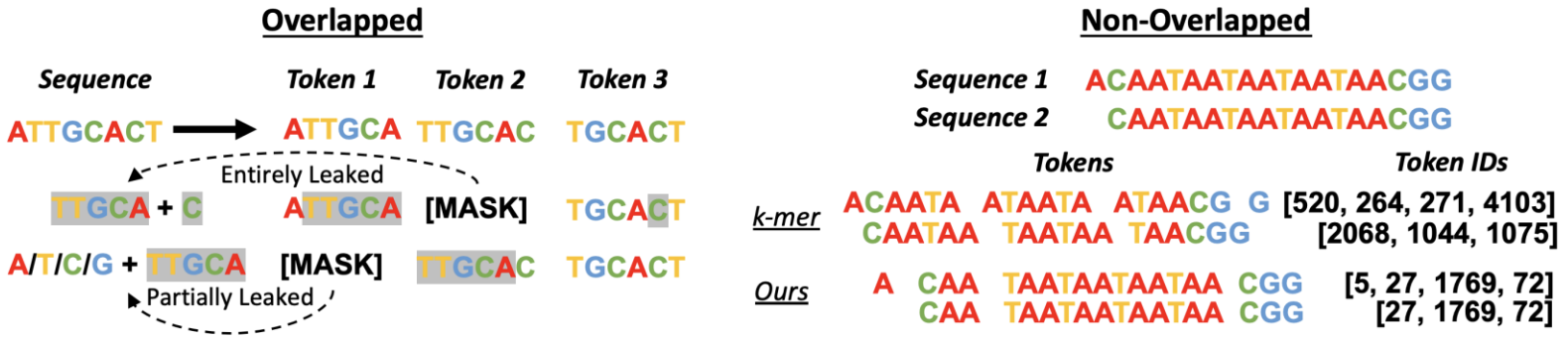


Figure 1: Illustration of the drawbacks of k-mer tokenization. In the overlapping setting, information about a masked token is leaked by its adjacent tokens, while in the non-overlapping setting, adding/deleting one nucleotide base leads to a dramatic change in the tokenized sequence.

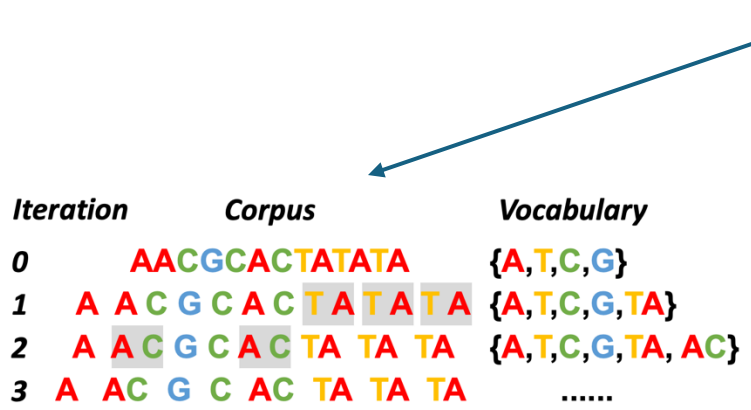


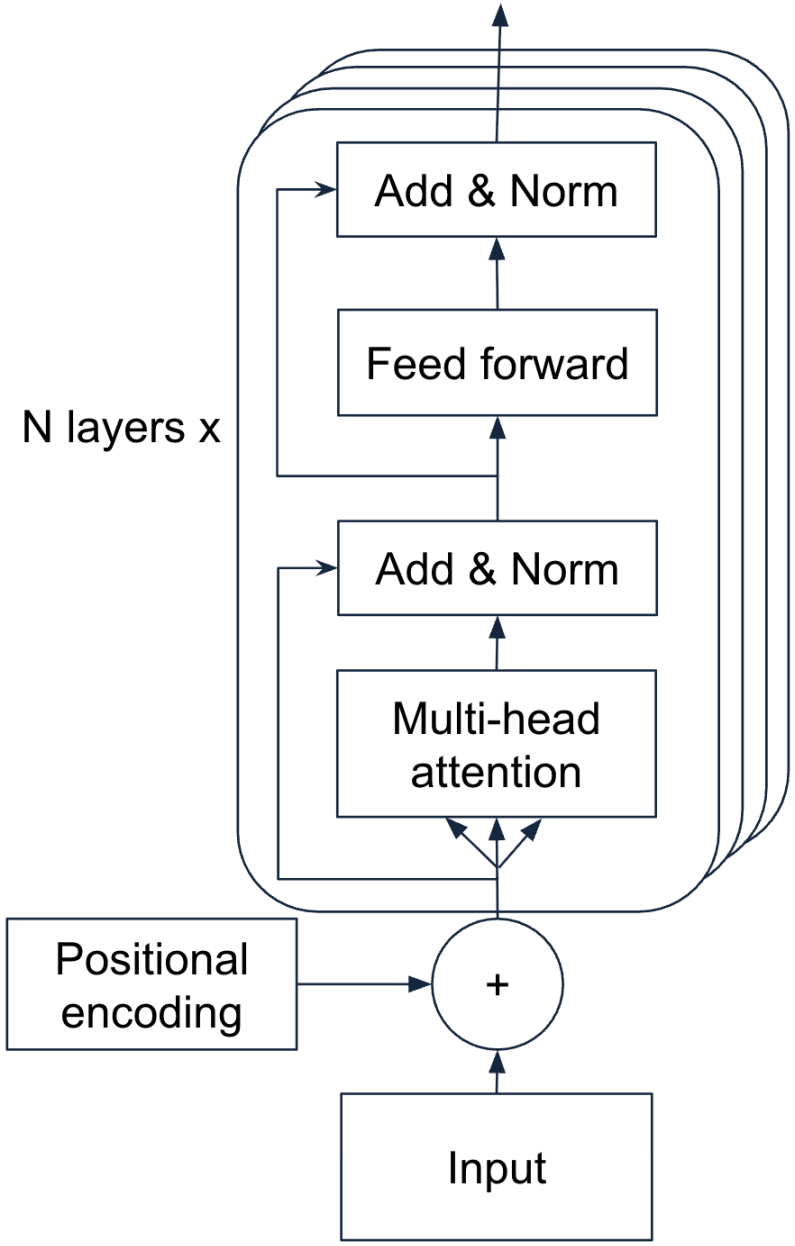
Figure 2: Illustration of the BPE vocabulary constructions.

BPE tokenization

- View the most frequent character segment as a new word
- Add to the vocab
- Update the corpus by replacing all the same segments with this new word
- Repeat

Zhou, Z., Ji, Y., Li, W., Dutta, P., Davuluri, R., & Liu, H. (2023). Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*.

DNABERT-2 as the feature extractor for DNA transcripts



- DNABERT-2 leverages BERT-like encoder only model to extract DNA features
- Pretrained on GUE and GUE+

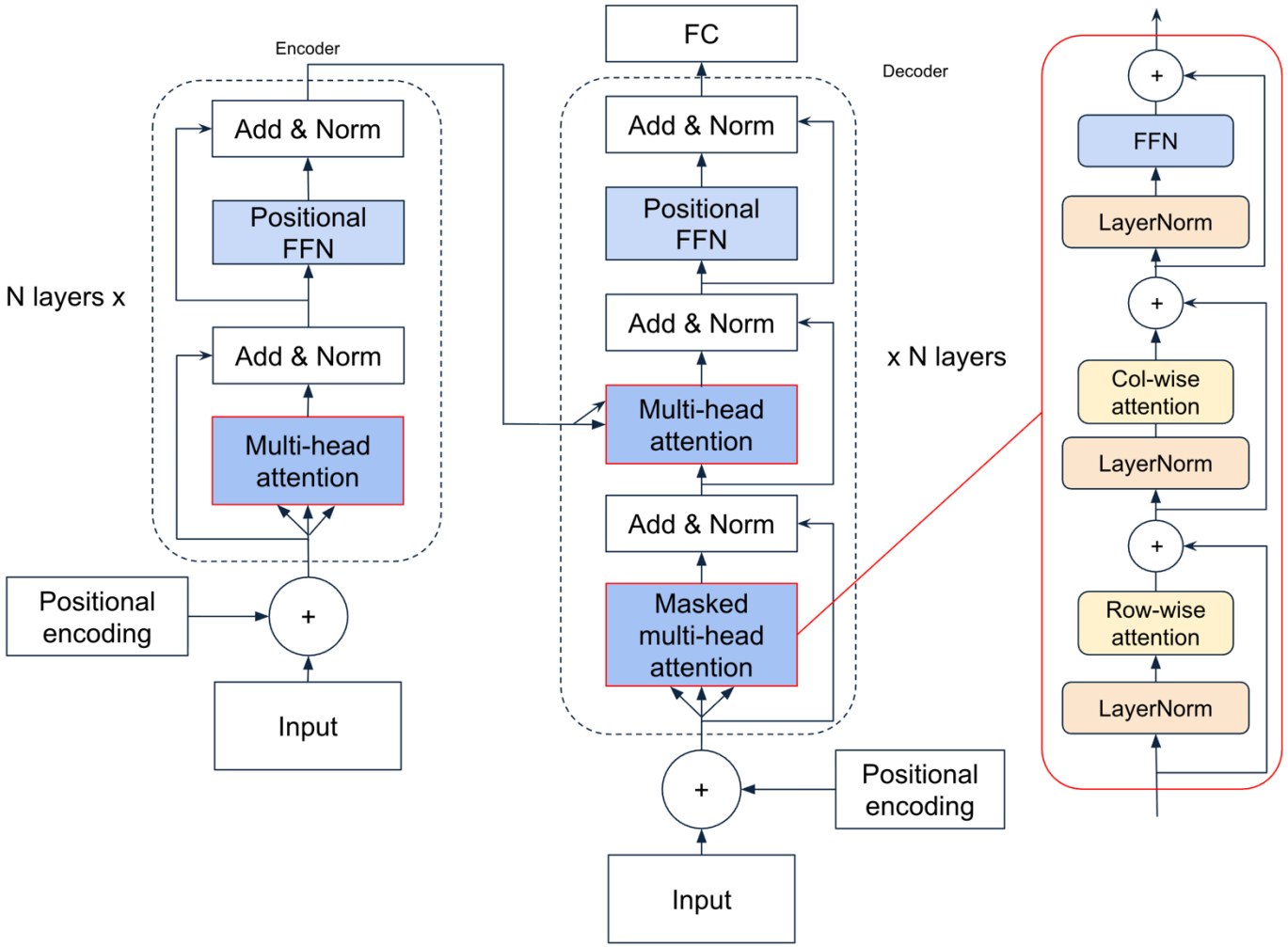
Species	Task	Num. Datasets	Num. Classes	Sequence Length
Human	Core Promoter Detection	3	2	70
	Transcription Factor Prediction	5	2	100
	Promoter Detection	3	2	300
	Splice Site Detection	1	3	400
Mouse	Transcription Factor Prediction	5	2	100
Yeast	Epigenetic Marks Prediction	10	2	500
Virus	Covid Variant Classification	1	9	1000

Table 1: Summarization of the Genome Understanding Evaluation (GUE) benchmark.

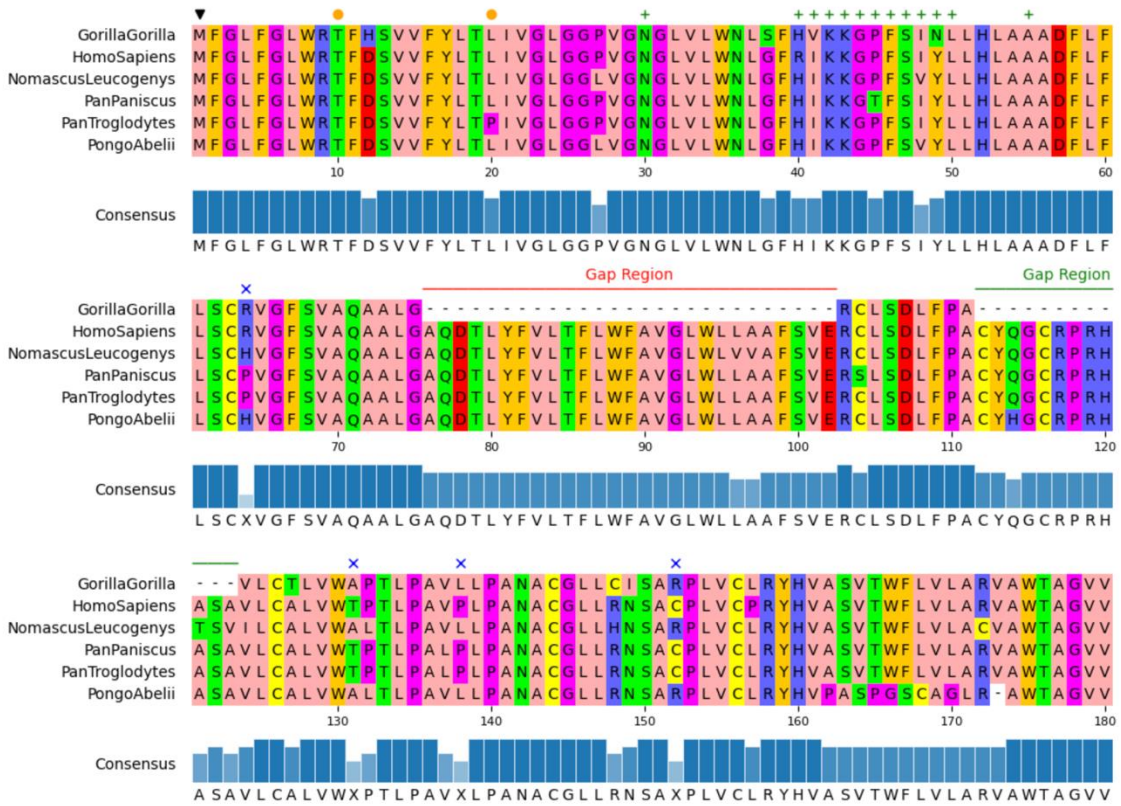
Species	Task	Num. Datasets	Num. Classes	Sequence Length
Human	Enhancer Promoter Interaction	6	2	5000
Fungi	Species Classification	1	25	5000
Virus	Species Classification	1	20	10000

Table 2: Summarization of the Genome Understanding Evaluation Plus (GUE^{+}) benchmark.

ESM-v2 as the feature extractor for protein sequences

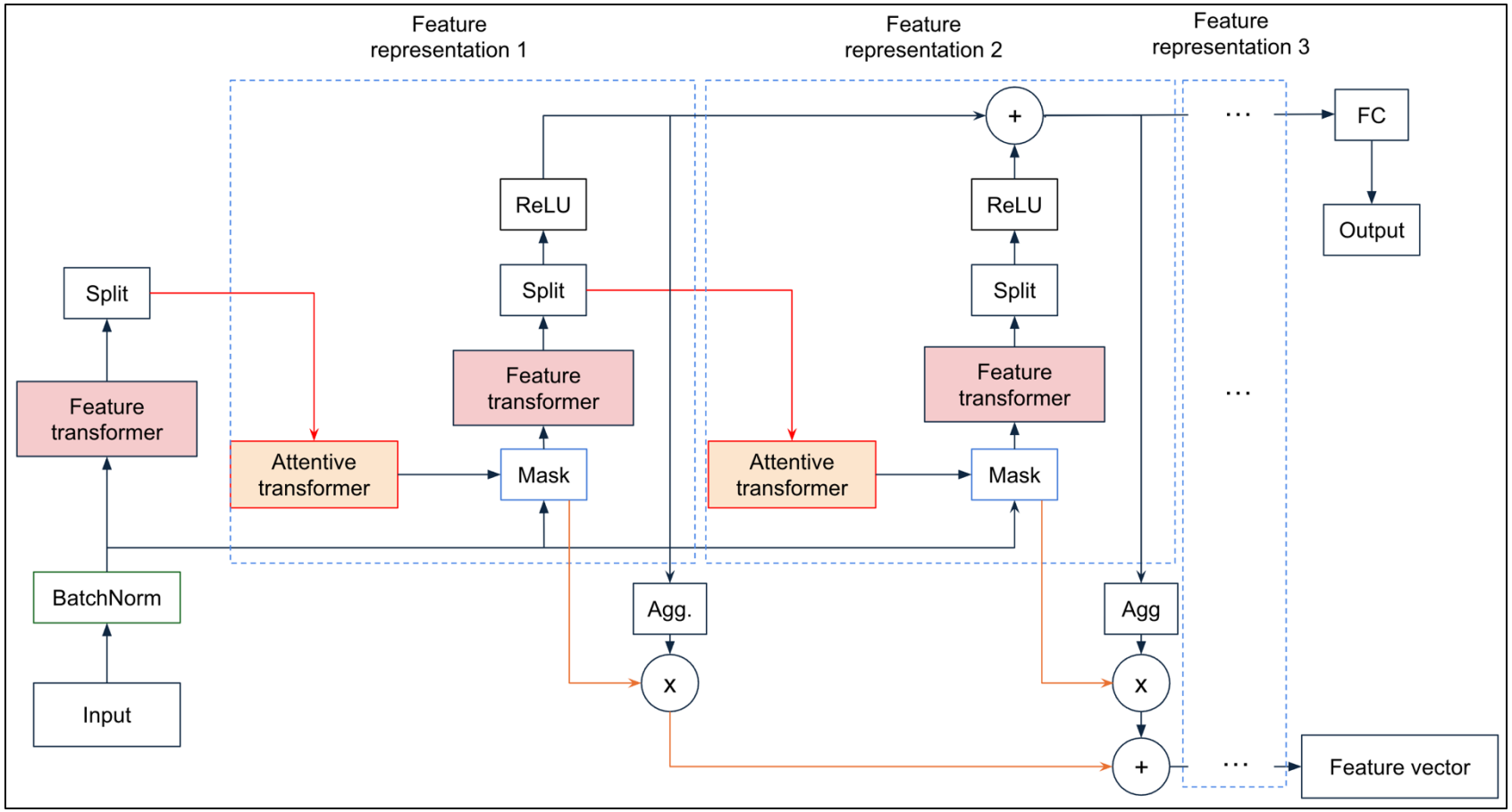


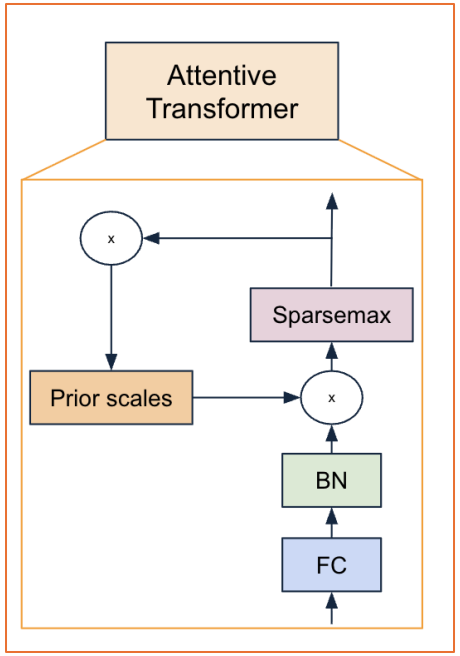
ESM-v2 architecture

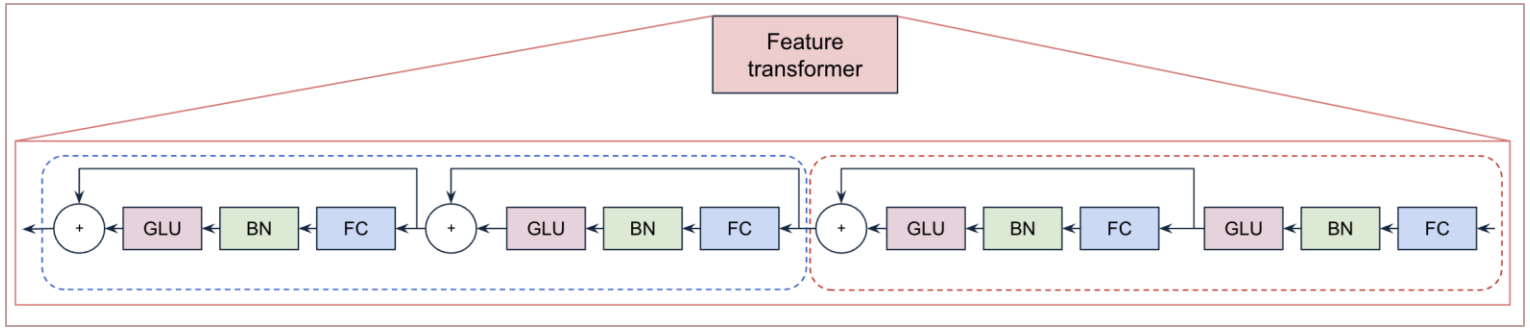


Zeming Lin *et al*. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 379,1123-1130(2023).DOI:10.1126/science.ade2574

TabNet to extract gene expression data

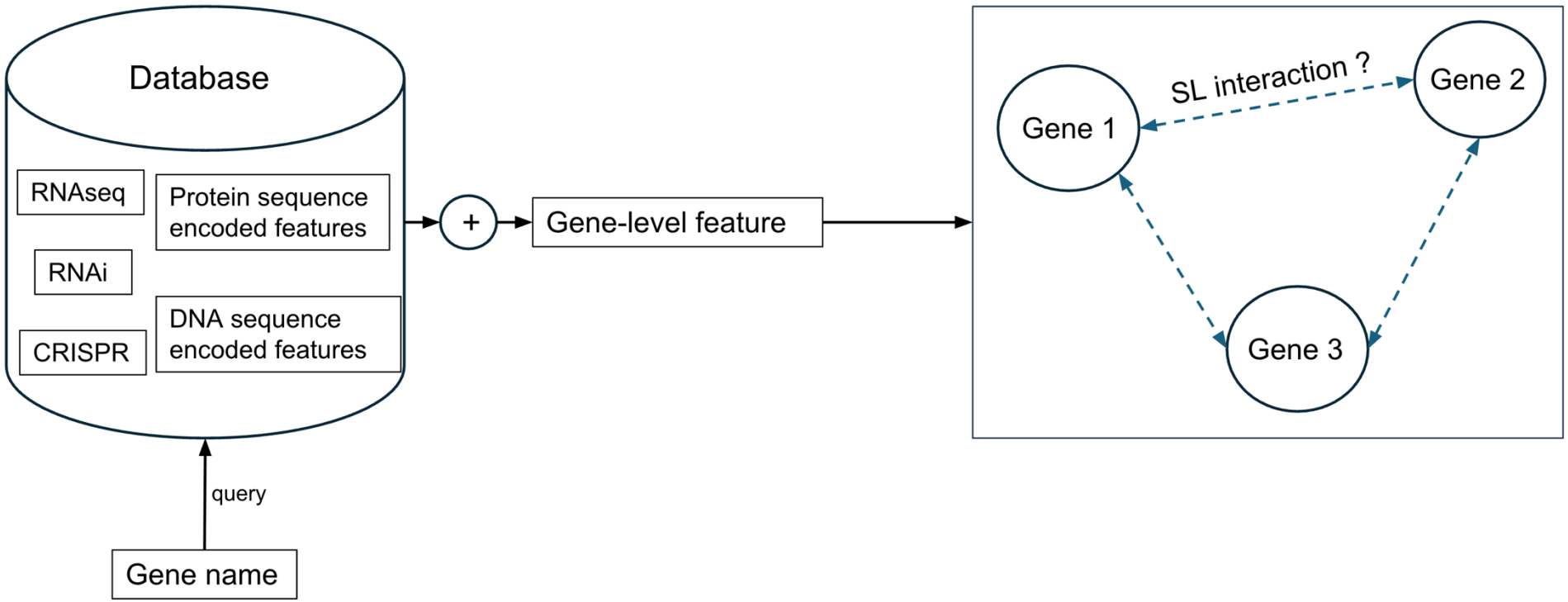






Arik, S. Ö., & Pfister, T. (2021, May). Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 35, No. 8, pp. 6679-6687).

SLGNN to predict possible SL interaction



Yan Zhu, Yuhuan Zhou, Yang Liu, Xuan Wang, Junyi Li, SLGNN: synthetic lethality prediction in human cancers based on factor-aware knowledge graph neural network, *Bioinformatics*, Volume 39, Issue 2, February 2023, btad015, https://doi.org/10.1093/bioinformatics/btad015

Table of contents

$$Specificity = \frac{TN}{TN + FP}$$

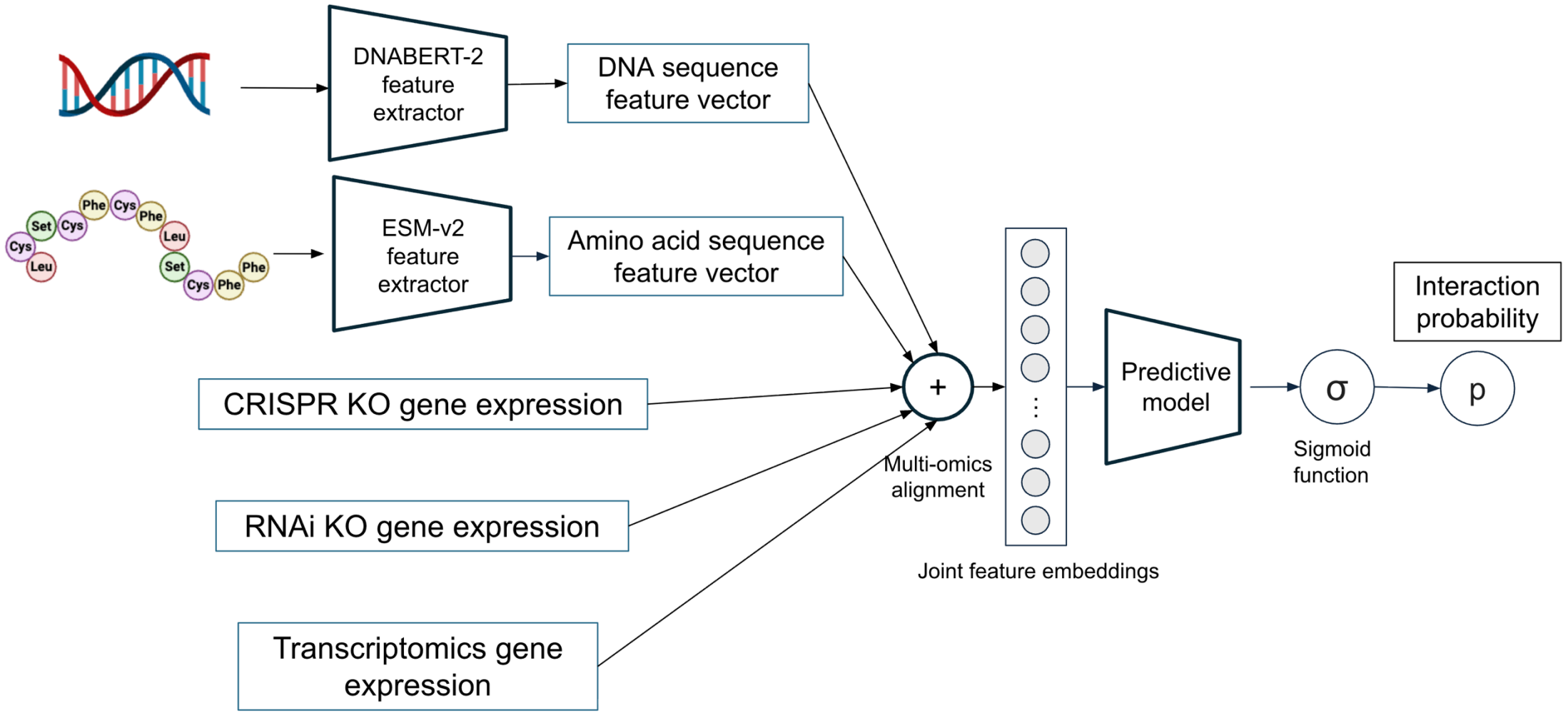
$$Sensitivity = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Benchmark different model configurations for single-omics SL prediction

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	KO Gene	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
SLGNN			45.87	83.04	53.44
TabNet			91.44	**95.41**	92.12
DNABERT-2 (117M) + Logistic regression	Genomics	DNA transcripts	95.62	71.83	90.77
SLGNN			50.60	82.93	57.18
DNABERT-2 (117M) + XGBoost			98.53	**90.84**	96.96
ESM-2 + XGBoost		Amino acid sequences	98.82	90.72	**97.17**
ESM-2 + DNABERT-2 + Logistic Regression		DNA sequence + amino acid sequences	**98.95**	90.11	97.14
ESM-2 + DNABERT-2 + XGBoost		DNA sequence + amino acid sequences	98.50	90.00	96.77

Integrating multi-omics & machine learning to predict synthetic lethality



Benchmark different model configurations for multi-omics SL prediction

Predictive model	Omics	Feature type	SPEC (5615)	SEN (1436)	ACC (7051)
Highest SPEC model (XGB)	Transcriptomics	RNASeq	97.84	86.52	95.53
Highest SPEC model (LR)	KO Gene	RNAi + CRISPR	98.23	87.51	96.05
Highest SPEC model (ESM + XGB)	Genomics	Protein sequences	98.95	90.11	97.14
Logistic Regression	Multi-omics	RNAi + CRISPR + RNAseq	**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		RNAi + CRISPR + RNAseq + DNA transcripts + Amino acid sequences	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 of contents

Conclusions

- A large synthetic lethality interaction in cancer cell lines dataset with 7791 positive SL interaction and 30 960 negative SL interaction.
- **Synthesized and processed 5 different omics-data** for every gene in the SL dataset, including: RNAi, CRISPR, DNA sequences, protein sequences, and transcriptomics.
- Benchmarked and compared different predictive methods

→ Evaluate the advantages of using multi-omics data in SL interaction prediction.

Conclusions

- Experimental results show that using **more advanced models** leads to the **improvement in the predictive performance**.
- Multi-omics data enhances model's performance in both specificity and accuracy, compared to single-omics settings.
- By using 5 different omics data and multi-modal model combination, achieves a specificity of **99.75%** and an accuracy of **97.88%** in predicting SL interaction.

Thank you for your concentration