Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Late-Fusion Multilayer Perceptron and Logistic Regression Classifier Identifies a SETDB1-Linked Dual-Vulnerability Niche in Non-BRCA Pancreatic Ductal Adenocarcinoma

Aarush Shah

simplyaarush@gmail.com

Issaquah Highschool https://orcid.org/0009-0008-4982-8171

Method Article

Keywords: SETDB1, pancreatic ductal adenocarcinoma, PARP inhibitor, olaparib, backup repair, dual-vulnerability niche, machine learning, biomarker stratification, DVN-Stack, epigenetic silencing

Posted Date: June 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10168289/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Pancreatic ductal adenocarcinoma (PDAC) remains difficult to treat. PARP inhibitors such as olaparib benefit mainly patients with homologous-recombination defects identified through BRCA testing, but most PDAC patients fall outside this group, and resistance often involves backup DNA repair and epigenetic regulation. SETDB1, an H3K9 methyltransferase linked to chromatin silencing, may shape PARP resistance through backup repair pathways, yet this relationship remains poorly defined in PDAC. This study integrated public transcriptome analysis, large-cohort biomarker screening, cell-line drug testing, and machine learning. RNA from TCGA and QCMG cohorts was screened to identify a PARP-resistance gene program and prioritize SETDB1. Across 339 tumors in CPTAC-3 and TCGA-PAAD, SETDB1 correlated with a Backup Repair Score (BRS) derived from alternative non-homologous end-joining genes (LIG3, POLQ, REV3L, MRE11) while excluding canonical PARP-panel markers. Tumors with high SETDB1 and high BRS defined a Dual-Vulnerability Niche (DVN), comprising approximately 37% of cases. Combined siSETDB1 and olaparib reduced proliferation and increased apoptosis in PANC-1, SU.86.86, and CAPAN-1 compared with single-agent controls. DVN-Stack, a late-fusion classifier combining multilayer perceptron and logistic regression branches, predicted DVN status from 148 independent genes with SETDB1 and BRS genes withheld. SETDB1 tracked BRS across cohorts (Spearman $\rho = 0.65–0.68$), with LIG3 as the strongest backup-repair partner. The model achieved 82.3% cross-validation accuracy and 0.89 AUC; CPTAC-3-trained models generalized to TCGA-PAAD (78.7% accuracy, 0.86 AUC). Together, these results suggest a candidate patient subgroup linked to SETDB1-centered backup repair and outline a computational framework for further study. Clinical benefit remains unproven and would require prospective validation.

INTRODUCTION

Pancreatic ductal adenocarcinoma (PDAC) is one of the most lethal solid tumors, with five-year survival near 12% and limited durable targeted therapies [1, 2]. Poly(ADP-ribose) polymerase (PARP) inhibitors exploit synthetic lethality in tumors with homologous-recombination (HR) deficiency, and maintenance olaparib improved progression-free survival in germline BRCA-mutant metastatic PDAC in the POLO trial [8, 13]. However, BRCA-type defects account for only a small fraction of patients, and clinical benefit is often incomplete even in biomarker-selected disease [14, 16].

A major reason for limited PARP benefit is compensatory DNA repair. When HR is impaired, tumors can rely on alternative end-joining pathways, including backup non-homologous end joining (alt-NHEJ) mediated by genes such as LIG3 and POLQ [22, 27, 28]. In parallel, epigenetic regulators reshape chromatin and gene expression in PDAC, influencing DNA-damage responses and drug sensitivity [21, 26]. SETDB1, an H3K9 methyltransferase that packages chromatin into silent heterochromatin, has been implicated in lineage control and tumor biology [12, 24], but its link to PARP resistance and backup repair in PDAC remains poorly defined.

Most patient-selection strategies focus on canonical HR genes (BRCA1, BRCA2, ATM, TP53) rather than epigenetic–repair interactions that may operate outside standard testing panels [10, 4]. PARP inhibition is established in HR-deficient cancers, and alt-NHEJ has been characterized as a resistance route [22, 27]; however, few studies integrate epigenetic silencing, backup repair scoring, experimental drug synergy, and machine-learning stratification into one PDAC framework. Patients without BRCA mutations therefore require biomarkers that predict benefit from PARP-based combination therapy.

The present study addresses this need through four linked aims: (1) determine whether SETDB1-high PDAC co-enriches backup repair genes independent of the canonical PARP panel; (2) evaluate siSETDB1 plus olaparib for synergistic activity in PDAC cell lines; (3) define a Dual-Vulnerability Niche (DVN) of tumors high for both SETDB1 and backup repair; and (4) build DVN-Stack, a late-fusion classifier combining multilayer perceptron and logistic regression branches to recognize DVN status from independent RNA features.

An initial discovery screen in public data identified SETDB1 before combination experiments were planned. Large-cohort validation in CPTAC-3 and TCGA-PAAD tested whether the SETDB1–backup-repair relationship replicated outside the discovery datasets. Cell-line experiments provided mechanistic support, and DVN-Stack was trained after the niche was defined, with strict feature exclusion to assess whether DVN represents a coherent biological subgroup. Figure 1 summarizes the integrated workflow.

Public datasets were chosen because they allow independent replication across platforms and institutions. CPTAC-3 provides proteogenomic depth in PDAC, while TCGA-PAAD offers broader clinical annotation and survival data [2, 5]. Using both cohorts reduces the chance that a biomarker pattern is an artifact of one sequencing pipeline or patient population.

METHODS

Cell-line experiments

PANC-1, SU.86.86, and CAPAN-1 human PDAC lines were cultured under standard conditions. Cells received siRNA targeting SETDB1, olaparib, combined siSETDB1 plus olaparib, or matched controls (scramble siRNA and/or vehicle). Proliferation and apoptosis were measured with biological replicates per condition under mentor supervision at Charles R. Drew University. Synergy was evaluated by comparing combination treatment with single-agent and control arms on growth and apoptosis readouts.

Transcriptome cohorts

Gene expression (transcripts per million) was downloaded from the NCI Genomic Data Commons (GDC) for CPTAC-3 pancreatic adenocarcinoma (n = 161) and TCGA-PAAD (n = 178) [15, 9]. The combined matrix included 339 tumors and 155 curated genes for biomarker screening.

Discovery screen and pathway analysis

Independent pancreatic cancer transcriptomes from TCGA and QCMG were compared to identify a 19-gene program associated with PARP inhibitor resistance. One hundred fifty-eight candidate genes spanning epigenetic regulators, PDAC biology, and DNA-repair pathways were screened for Spearman correlation with SETDB1. For CPTAC-3, candidate retention used an FDR-adjusted (BH) significance screen (fdr < 0.05) alongside Spearman signal. Canonical PARP and DNA-damage-response panel genes were excluded to reduce circular discovery. Gene-set enrichment used GSEA and DAVID [23, 11]. Candidates were retained only if correlation direction (consistent sign) and statistical support were consistent in both CPTAC-3 and TCGA-PAAD, following cross-cohort validation standards used in biomarker discovery studies [20].

Backup Repair Score and DVN definition

Backup Repair Score (BRS) was defined as the mean cohort-wise z-score of LIG3, POLQ, REV3L, and MRE11—alt-NHEJ genes deliberately excluded from prior PARP-panel machine learning. The Dual-Vulnerability Niche (DVN) comprised tumors with SETDB1 and BRS both above the cohort median. BRS and DVN were evaluated independently in CPTAC-3 and TCGA-PAAD.

DVN-Stack classifier

DVN-Stack combined a two-hidden-layer multilayer perceptron (48→24 units, batch normalization, dropout 0.35) and L2 logistic regression (C = 0.5, class-weight balanced) trained on 148 withheld features. Branch out-of-fold probabilities were combined by mean fusion and meta-logistic regression. Validation used stratified five-fold cross-validation [25] and cohort-held-out testing (train CPTAC-3, test TCGA-PAAD). SETDB1, BRS genes, and PARP-panel markers were excluded from all model inputs [18, 17, 6]. Model architecture is shown in Supplementary Figure S1.

Survival and epigenomic context

TCGA-PAAD overall survival was retrieved through the GDC cases API. Kaplan–Meier curves compared SETDB1-high versus SETDB1-low groups. ENCODE SETDB1 ChIP-seq peaks (n = 3930) provided epigenomic context through peak-to-gene mapping and g:Profiler enrichment [7, 19].

Software and reproducibility

Analyses used Python 3.12 (pandas, NumPy, SciPy, scikit-learn, PyTorch, Matplotlib). Cross-validation splits used fixed random seeds. All preprocessing steps—including cohort merging, z-score normalization for BRS, and feature exclusion lists—were applied before model training to prevent information leakage. Analysis scripts are available from the corresponding author upon request.

RESULTS

Bioinformatics discovery of SETDB1

A PARP inhibitor–resistance discovery screen in public TCGA/QCMG transcriptomes prioritized SETDB1 as a candidate determinant of PARP inhibitor response. To define the backup repair program linked to SETDB1, a SETDB1-associated gene analysis was conducted across 339 PDAC tumors (CPTAC-3 + TCGA-PAAD). One hundred fifty-eight candidate genes spanning epigenetic regulators, DNA-damage response factors, and chromatin modifiers were screened, while excluding 30 canonical PARP and DDR panel genes (including BRCA1, BRCA2, ATM, TP53, PARP1, RAD51) to prevent circular discovery. For each candidate, Spearman correlation with SETDB1 was computed, and pass criteria required BH-adjusted FDR < 0.05 and a consistent correlation sign in both CPTAC-3 and TCGA-PAAD. Nineteen genes met these cross-cohort criteria, with LIG3 emerging as the top consistent partner. This SETDB1–LIG3 connection motivated construction of the Backup Repair Score (BRS) and the Dual-Vulnerability Niche (DVN) framework (Fig. 3).

Cell-line synergy

Combined siSETDB1 and olaparib were tested in PANC-1, SU.86.86, and CAPAN-1. Combination siSETDB1 + olaparib suppressed long-term proliferation relative to single agents and scramble control (Fig. 2A; Dunnett's p = 0.0016–0.0175 versus combination arm). The time course confirmed sustained growth suppression across all measured timepoints (Fig. 2B). Pairwise significance is summarized in Fig. 2C.

Cohort validation of SETDB1 and backup repair

Across 339 tumors, SETDB1 correlated with BRS in a way captured by the overall patient-level score (pooled Spearman ρ = 0.46; p = 1.38e-19; Fig. 3A). Cohort-wise, the association remained consistent in CPTAC-3 (ρ = 0.65) and TCGA-PAAD (ρ = 0.68). SETDB1 expression differed significantly between cohorts (Fig. 3B). Figure 3C shows gene-level correlations for each BRS component.

The DVN (SETDB1-high ∩ BRS-high) comprised 38% of CPTAC-3 and 37% of TCGA tumors (~ 37% combined; Fig. 3B). Because DVN status is defined by both SETDB1-high expression and the Backup Repair Score, predictive performance was next evaluated using RNA features that excluded SETDB1, BRS genes, and PARP-panel markers.

DVN-Stack classification

DVN-Stack is a late-fusion model trained to predict DVN from withheld features, so classification performance reflects an underlying molecular subtype independent of the label-defining genes.

DVN-Stack classified this niche from 148 withheld features with 82.3% five-fold out-of-fold accuracy and 0.89 AUC (Fig. 4A, C). When trained on CPTAC-3 and tested on TCGA-PAAD, accuracy was 78.7% and

AUC was 0.86. Top model-weighted genes (FOXO3, RBBP5, NCL, STAT3, MSLN) reflect epigenetic PDAC programs rather than the defining repair genes, consistent with classification from independent features. Exploratory TCGA survival was shorter in SETDB1-high patients (log-rank p = 0.035).

Laboratory and cohort analyses were concordant: SETDB1 tracks a backup-repair program centered on LIG3 and replicated across independent datasets. The DVN label captures tumors high on both axes, and the classifier detected a structured RNA signature associated with this subgroup (Fig. 4).

Table 1
Out-of-fold DVN classification using non-PARP, non-BRS features

Cohort	n	Features	Logistic AUC	MLP AUC	DVN-Stack AUC	DVN-Stack accuracy
Combined	339	148	0.88	0.90	0.89	82.3%
CPTAC-3	161	148	0.87	0.88	—	77.6%
TCGA-PAAD	178	148	0.90	0.91	—	80.9%

DISCUSSION

Together, these findings describe an integrated research framework for studying PDAC patients who might one day be considered for PARP combination strategies outside standard BRCA testing. Clinical practice and major trials anchor PARP use in HR deficiency [8, 13]. Mechanistic studies show that backup repair can sustain viability when primary repair fails [22, 27]. The present study connects SETDB1-centered epigenetic biology, backup repair scoring, and experimental validation of siSETDB1 plus olaparib synergy within one analytical pipeline.

The DVN integrates these components. BRS uses alt-NHEJ genes excluded from the canonical panel and described a computationally defined subgroup (~ 37% of tumors) with coherent RNA structure. LIG3 was the strongest gene-level signal, consistent with literature on alt-NHEJ [27] and positioned here as the leading partner of SETDB1 across two large cohorts. Cell-line synergy provides functional support that this biology is pharmacologically actionable.

DVN-Stack adds a practical stratification layer. The model achieved 82.3% accuracy and 0.89 AUC while withholding SETDB1, BRS genes, and PARP-panel markers, indicating that DVN corresponds to a recognizable transcriptomic subtype. Cohort-held-out validation strengthens this finding and follows recommended practice for reporting machine-learning models in biomedical research [6].

Several limitations should be stated clearly. First, DVN status was inferred from RNA, not protein or functional repair assays. Second, the study did not link DVN or BRS to clinical PARP response records. Third, survival analysis was exploratory and not adjusted for confounders. Fourth, cell-line experiments demonstrate synergy but do not prove in vivo efficacy. Future work should measure BRS at the protein

level, test combination treatment in DVN-high versus DVN-low models, and connect DVN-Stack scores to clinical olaparib response in patient datasets.

Placed in the context of prior literature, the findings extend existing work. PARP inhibitors are established in HR-deficient disease [13], and alt-NHEJ has been described as a resistance mechanism [22]. LIG3 and POLQ have been studied individually in DNA repair and PDAC biology [27, 28]. SETDB1 has been linked to epigenetic silencing in cancer [12, 24]. The integrated contribution of this study is the chain linking SETDB1 prioritization from a PARP-resistance screen, backup repair quantification by BRS outside the canonical panel, DVN definition at cohort scale, synergy validation in cell lines, and DVN-Stack classification from withheld RNA features.

From a clinical perspective, the DVN concept is meant to complement—not replace—BRCA testing. Patients with HR defects may still receive PARP inhibitors under current guidelines. The DVN framework instead targets the larger group of patients who are not selected by standard panels but may still depend on backup repair and epigenetic regulation. If prospectively validated, this could expand the population eligible for rational combination trials.

The machine-learning component was evaluated conservatively. Label-defining genes were withheld from model inputs, cross-validation was stratified, and performance was tested on held-out cohorts rather than a single random split. These steps reduce overfitting risk and align with recommended practice for reporting prediction models in biomedical research [6, 25].

CONCLUSION

This study combined public cohort analysis, cell-line experiments, and machine-learning modeling to explore whether SETDB1-centered backup repair defines a recognizable subgroup in PDAC. The results support a testable research direction rather than a clinical recommendation. DVN-Stack offers one way to study that subgroup using transcriptomic features that do not include SETDB1 or the BRS genes directly, but any use in patient care would require independent validation.

Declarations

Ethics approval and consent to participate

This study used established, commercially available PDAC cell lines (PANC-1, SU.86.86, and CAPAN-1). No new human participants were recruited and no animal work was performed. Laboratory experiments were conducted under mentor supervision at Charles R. Drew University using standard cell-culture and treatment protocols. Public tumor RNA data were obtained from the NCI Genomic Data Commons under that program's data-use policies. All cohort data were de-identified before release to researchers; the author did not access names, dates of birth, medical record numbers, or other direct patient identifiers.

Consent for publication

Not applicable. This manuscript does not report identifiable individual patient data.

Competing interests

The author declares no competing financial or non-financial interests related to this work.

Funding

No external funding was received for this study.

Authors' contributions

Aarush Shah designed and performed the computational analyses, prepared figures and the manuscript, and participated in cell-line experiments under mentor supervision. All analyses, figures, and conclusions were reviewed with the mentor laboratory team.

ACKNOWLEDGMENTS

The author thanks Dr. Pranabananda Dutta and the Charles R. Drew University laboratory team for experimental training and supervision; Issaquah High School for research support; and the NCI GDC, ENCODE, TCGA, and cBioPortal communities for open data [5].

Availability of data and materials

Public datasets are listed under Data availability. Other materials are available from the corresponding author on reasonable request.

DATA AVAILABILITY

CPTAC-3 and TCGA-PAAD RNA data used in this study are publicly available through the NCI Genomic Data Commons (https://portal.gdc.cancer.gov/). ENCODE SETDB1 ChIP-seq data are available at https://www.encodeproject.org/. Supplementary tables and figures (DVN-Stack architecture and cohort-definition schematic) support the machine-learning results. Analysis code is available from the corresponding author on reasonable request.

References

1. Bailey P et al (2016) Genomic analyses identify molecular subtypes of pancreatic cancer. Nature 531(7592):47–52

2. Cancer Genome Atlas Research Network (2017) Integrated genomic characterization of pancreatic ductal adenocarcinoma. Cancer Cell 32(2):185–203
3. Cancer Genome Atlas Research Network (2013) The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 45(10):1113–1120
4. Ceccaldi R, Rondinelli B, D'Andrea AD (2016) Repair pathway choices and consequences at the double-strand break. Trends Cell Biol 26(1):52–64
5. Cerami E et al (2012) The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. Cancer Discov 2(5):401–404
6. Collins GS, Moons KGM (2019) Reporting of artificial intelligence prediction models. Lancet 393(10181):1577–1579
7. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414):57–74
8. Golan T et al (2019) Maintenance olaparib for germline BRCA-mutated metastatic pancreatic cancer. N Engl J Med 381(4):317–327
9. Grossman RL et al (2016) Toward a shared vision for cancer genomic data. N Engl J Med 375(12):1109–1112
10. Huang A et al (2021) Targeting the DNA damage response in pancreatic ductal adenocarcinoma. Trends Cancer 7(11):1015–1029
11. Huang DW et al (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4(1):44–57
12. Li H et al (2020) The H3K9 methyltransferase SETDB1 suppresses endogenous retroviruses during lineage specification. Nature 584(7822):682–687
13. Lord CJ, Ashworth A (2017) PARP inhibitors: Synthetic lethality in the clinic. Science 355(6330):1152–1158
14. Mateo J et al (2020) DNA repair defects and olaparib in metastatic prostate cancer. N Engl J Med 382(20):1691–1704
15. National Cancer Institute (2026) Genomic Data Commons Data Portal. https://portal.gdc.cancer.gov/
16. O'Connor MJ (2019) Target the DNA damage response in cancer: Beyond PARP inhibition. Nat Reviews Clin Oncol 16(2):81–99
17. Paszke A et al (2019) PyTorch: An imperative style, high-performance deep learning library. Adv Neural Inf Process Syst 32:8024–8035
18. Pedregosa F et al (2011) Scikit-learn: Machine learning in Python. J Mach Learn Res 12:2825–2830
19. Raudvere U et al (2019) g:Profiler: A web server for functional enrichment analysis. Nucleic Acids Res 47(W1):W191–W198
20. Ritchie ME et al (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res, 43(7), e47

21. Rodriguez-Paredes M, Esteller M (2011) Cancer epigenetics reaches mainstream oncology. Nat Med 17(3):330–339
22. Sfeir A, Symington LS (2015) Microhomology-mediated end joining: A back-up survival mechanism or dedicated pathway? Trends Biochem Sci 40(11):701–714
23. Subramanian A et al (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences, 102(43), 15545–15550
24. Sun C et al (2021) Relevance of SETDB1 in cancer: Genomic and epigenomic perspectives. Front Oncol 11:672409
25. Varma S, Simon R (2006) Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 7:91
26. Wang L et al (2021) Epigenetic regulation of DNA damage repair in pancreatic cancer. Cancer Lett 502:1–10
27. Wang Y et al (2018) LIG3 promotes repair of DNA double-strand breaks in non-homologous end joining. Cancer Res 78(13):3582–3593
28. Zhou J et al (2021) POLQ inhibition elicits an immune response in homologous recombination-deficient pancreatic adenocarcinoma. Nature Communications, 12, 2052

Figures

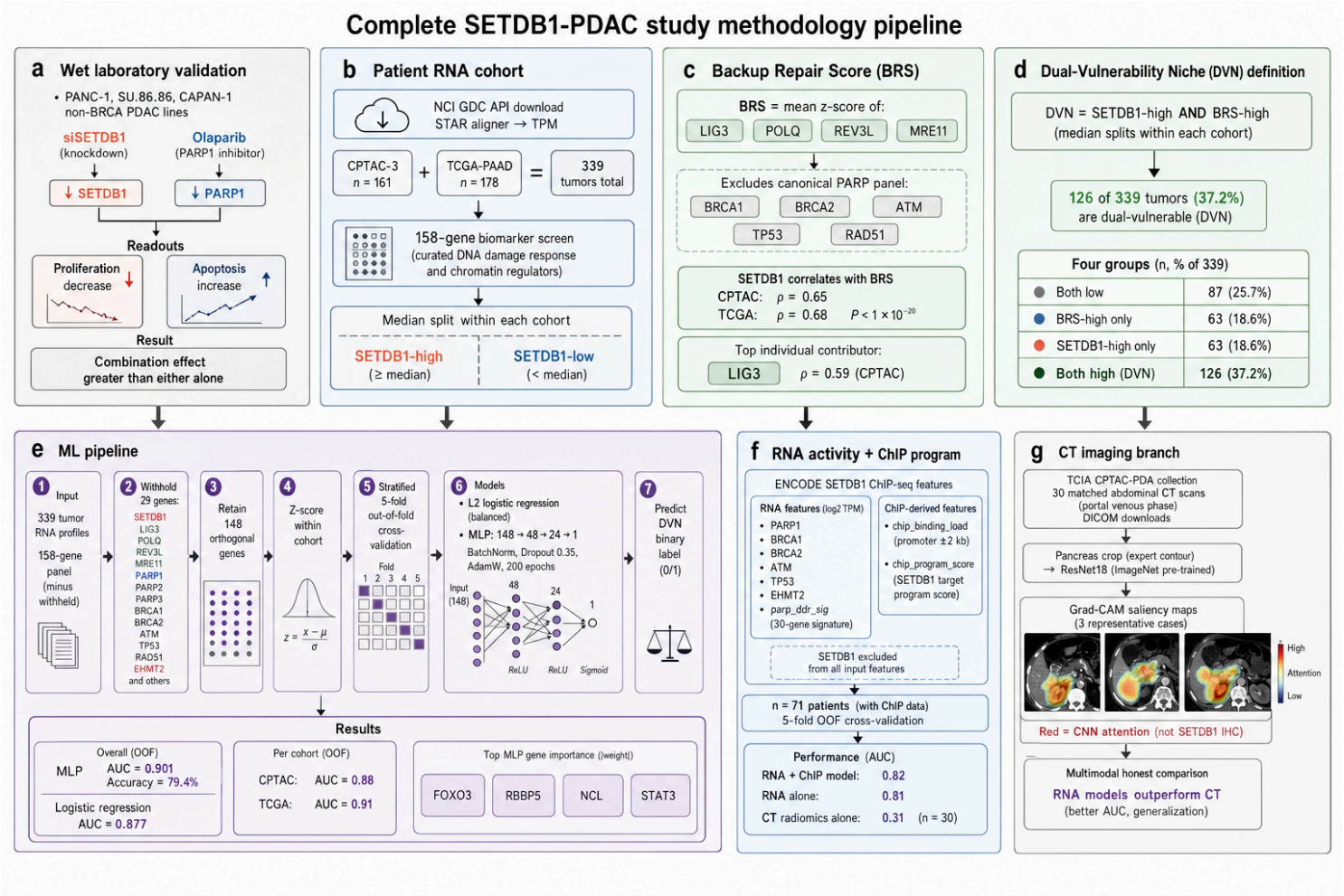
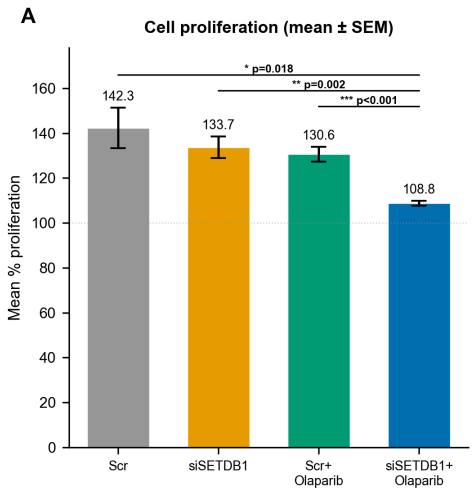
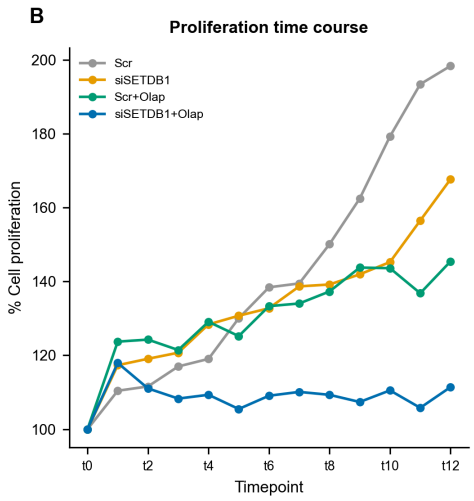


Figure 1

Integrated SETDB1–PDAC study workflow. The pipeline links wet-laboratory siSETDB1 ± olaparib experiments, public CPTAC-3 and TCGA-PAAD RNA cohorts (339 tumors), Backup Repair Score (BRS) and Dual-Vulnerability Niche (DVN) definition, and DVN-Stack machine-learning classification from withheld RNA features. Schematic overview only; detailed model architecture is in Supplementary Figure S1.





C

Dunnett's multiple comparisons

Comparison	Mean Diff.	Adj. p-value	Sig.
siSETDB1 vs combo	24.9	0.0016	**
Scr+Olap vs combo	21.8	0.0001	***
Scr vs combo	33.5	0.0175	*

Figure 2

Cell-line experimental results. (A) Proliferation (mean ± SEM) across treatment conditions with Dunnett's significance versus siSETDB1 + olaparib. (B) Proliferation time course (t0–t12). (C) Dunnett's multiple comparisons table.

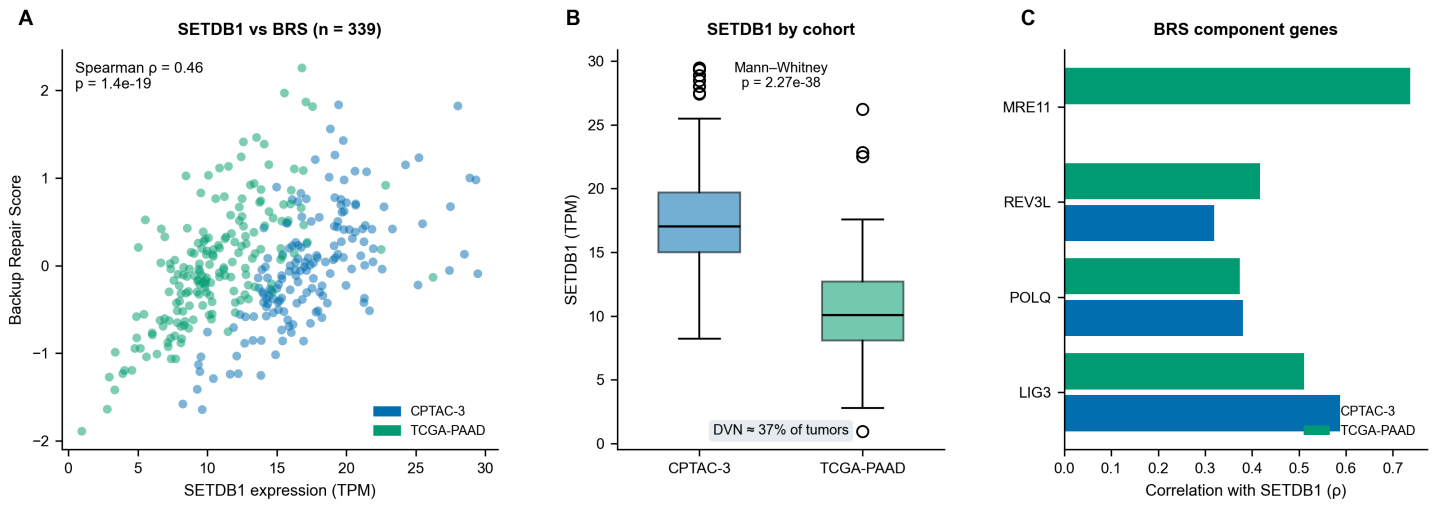


Figure 3

Cohort analysis across 339 tumors. (A) SETDB1 versus Backup Repair Score (Spearman ρ = 0.46). (B) SETDB1 expression by cohort (Mann–Whitney). (C) BRS component genes versus SETDB1.

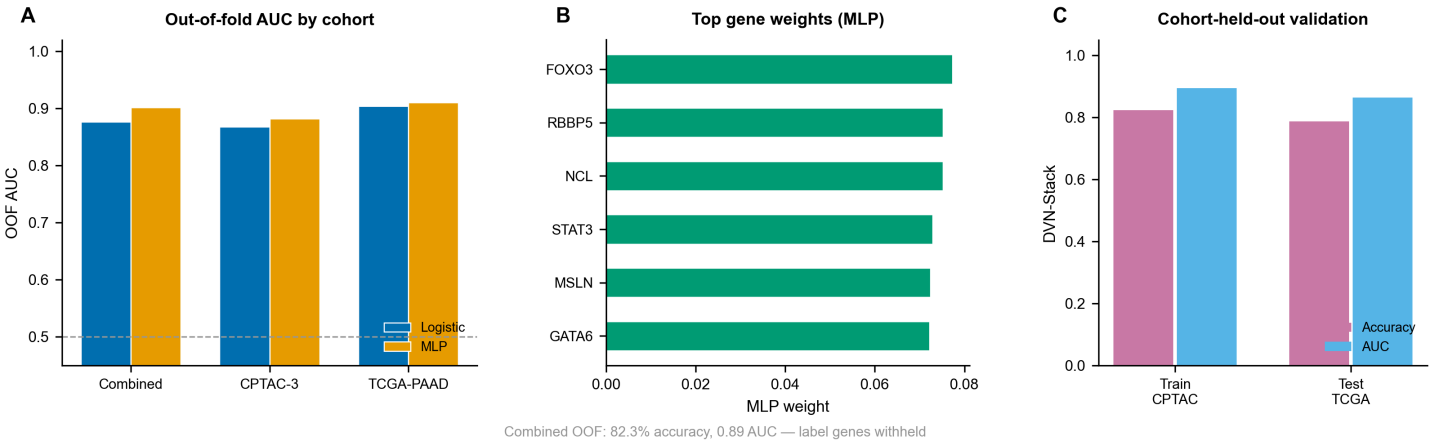


Figure 4

DVN-Stack performance. (A) Out-of-fold AUC by cohort for logistic and MLP branches. (B) Top MLP feature weights (label-defining genes withheld). (C) Combined OOF versus cohort-held-out validation (train CPTAC-3, test TCGA-PAAD). Model architecture is in Supplementary Figure S1.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigureS1architecture.png
- SupplementaryFigureS2cohortdefinition.png
- SupplementaryFigureS5apoptosisdoseresponse.png
- SupplementaryTable1DVNclassification.png
- SupplementaryFigureS3proliferationrawdata.png
- SupplementaryTable2BRScohort.png