

Explainable multi-omics modeling for risk stratification in pancreatic ductal adenocarcinoma

Bingrong Chen¹, Jianuo Chen², Zhen Cao¹, Qiyao Zhang¹, Shangcheng Yan¹, Hongtao Cao¹, Jia Yu², Fang Wang², Ziwen Liu¹

¹Department of General Surgery, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China; ²State Key Laboratory of Medical Molecular Biology, Department of Biochemistry and Molecular Biology, Haihe Laboratory of Cell Ecosystem, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences, School of Basic Medicine, Peking Union Medical College, Beijing, China

Contributions: (I) Conception and design: J Yu, F Wang, Z Liu; (II) Administrative support: F Wang; (III) Provision of study materials or patients: Z Cao, Q Zhang, S Yan, H Cao; (IV) Collection and assembly of data: B Chen, J Chen; (V) Data analysis and interpretation: B Chen, J Chen; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Jia Yu, PhD; Fang Wang, PhD. State Key Laboratory of Medical Molecular Biology, Department of Biochemistry and Molecular Biology, Haihe Laboratory of Cell Ecosystem, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences, School of Basic Medicine, Peking Union Medical College, No. 5 Dongdantsiantao, Dongcheng District, Beijing 100005, China. Email: j-yu@ibms.pumc.edu.cn; wo_wfang@hotmail.com; Ziwen Liu, MD, PhD. Department of General Surgery, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, No. 1 Shuaifuyuan Road, Dongcheng District, Beijing 100730, China. Email: liuziwen@pumch.cn.

Background: Pancreatic ductal adenocarcinoma (PDAC) remains one of the most lethal malignancies due to a lack of reliable tools for individualized risk stratification. A comprehensive understanding of the multi-omics landscape may uncover clinically applicable biomarkers and inform precision prognostic assessment. This study aims to establish a prognostic model directly from the complete omics landscape and extract biomarkers.

Methods: We developed prognostic models using multi-omics data from a PDAC proteogenomic cohort comprising 75 deceased tumor samples. An independent cohort of 63 deceased PDAC cases from The Cancer Genome Atlas (TCGA)-pancreatic adenocarcinoma (PAAD) was used for external validation. Logistic regression models with least absolute shrinkage and selection operator (LASSO) regularization were constructed, and SHapley Additive exPlanations (SHAP) were applied to evaluate feature importance and identify signature genes. Model selection was based on the average area under the receiver operating characteristic curve (AUROC) across cross-validation folds. Functional validation was performed in PANC-1 cells by knockdown (KD) or overexpression (OE) of representative microRNA-, RNA-, and proteomics-derived signature genes, followed by Cell Counting Kit-8 (CCK-8) proliferation and Transwell migration assays.

Results: Systematic evaluation of 120 multi-omics combinations identified a top-performing prognostic model integrating RNA, microRNA, proteomics, and mutation features. This model achieved a mean AUROC of 0.92 ± 0.11 and accuracy of 0.87 ± 0.01 on internal validation, and 0.99 ± 0.00 and 0.98 ± 0.01 on the TCGA test set. The sensitivity, specificity, precision, recall and F1 scores on the TCGA test set were 0.98 ± 0.01 , 0.97 ± 0.02 , 0.98 ± 0.02 , 0.98 ± 0.01 , 0.98 ± 0.01 , respectively. SHAP analysis revealed interpretable and clinically relevant prognostic biomarkers, many of which are implicated in immune signaling, metabolic regulation, and cell cycle control. Importantly, modulation of representative signature genes in PANC-1 cells significantly altered proliferation and migration in directions consistent with model-predicted risk associations.

Conclusions: Our findings demonstrate that explainable multi-omics machine learning frameworks can identify robust prognostic biomarkers and achieve highly accurate survival prediction in PDAC. Functional validation further supports the biological relevance of these signatures, underscoring their translational potential for personalized risk assessment.

Keywords: Pancreatic ductal adenocarcinoma (PDAC); multi-omics; machine learning; biomarker; prognosis

Submitted Sep 02, 2025. Accepted for publication Feb 25, 2026. Published online Mar 27, 2026.

doi: 10.21037/gs-2025-396

View this article at: <https://dx.doi.org/10.21037/gs-2025-396>

Introduction

Background

Pancreatic ductal adenocarcinoma (PDAC) is among the most aggressive and lethal malignancies of the gastrointestinal tract, characterized by late presentation and a dismal prognosis, with a 5-year survival rate of remain 13% (1). PDAC currently has one of the lowest survival rates among all cancers and is projected to account for approximately 51,750 deaths in 2024 alone (2,3). Two major clinical challenges hinder progress in PDAC management: the lack of effective early diagnostic strategies and limited therapeutic efficacy (4,5). Approximately 80–85% of patients are diagnosed at an advanced or metastatic stage, due to non-specific early symptoms and the absence of routine screening (6). Even for those eligible for surgical resection—the only potentially curative intervention—the 5-year survival remains below 20%. Most patients receive cytotoxic chemotherapy as first-line treatment, yet median overall survival rarely exceeds 12 months, and recent therapeutic advances have yielded only marginal improvements (7). Even though National Comprehensive Cancer Network (NCCN) guidelines recommend platinum-based regimens and PARP inhibitors (e.g., FOLFIRINOX) for BRCA1/2-mutant and NTRK inhibitor (larotrectinib) for NTRK

gene fusions patients (8), and survival benefits have been demonstrated (9), marked survival heterogeneity persists among patients with similar staging, mutations, and treatments (10). These limitations underscore an urgent need for improved molecular stratification and the discovery of clinically applicable biomarkers to guide personalized therapeutic intervention in PDAC.

Rationale and knowledge gap

For years, PDAC risk has been assessed primarily based on family history, lifestyle, and clinical risk factors (6). However, such criteria account for only a minority of cases. Some studies have leveraged longitudinal health trajectories to predict PDAC risk (11), yet these approaches often rely on structured medical records that are unavailable for the general population. Genomic studies have identified recurrent mutations in genes such as *KRAS*, *TP53*, *CDKN2A*, *SMAD4*, *BRCA1/2*, and *NTRK* (12). However, with the exception of *NTRK*, which has shown prognostic relevance, these mutations typically do not differ significantly between early- and late-stage tumors (13), and are not strongly associated with overall survival (14), limiting their utility as prognostic biomarkers.

The advent of high-throughput sequencing technologies has enabled transcriptomic subtyping of PDAC, uncovering distinct molecular classes with differential responses to therapy (15). At the proteomic and phosphoproteomic levels, mass spectrometry and immunohistochemistry have identified expression signatures correlated with clinical outcomes (16). Nevertheless, substantial intra-subtype heterogeneity in patient survival remains, suggesting that more complex biological mechanisms are at play. Recent studies have explored integrative multi-omics strategies to capture systems-level insights into PDAC biology (17), offering a promising avenue for uncovering novel prognostic biomarkers and therapeutic targets.

However, mining clinically informative signals from high-dimensional, small-sample multi-omics datasets poses major computational and statistical challenges. Recent advances in machine learning have improved patient stratification and outcome prediction in oncology (18). Interpretable algorithms such as SHapley Additive

Highlight box

Key findings

- SHapley Additive exPlanations-based interpretation revealed clinically relevant biomarkers, advancing the translational utility of multi-omics machine learning in pancreatic ductal adenocarcinoma.

What is known and what is new?

- Single-omics could be used to build up prognosis model but lacked sufficient predictive power.
- An optimal multi-omics combination (RNA, microRNA, proteomics, mutation) enabled robust survival prediction and external validation in The Cancer Genome Atlas cohort.

What is the implication, and what should change now?

- It is necessary to integrate more comprehensive omics information to delineate the prognostic characteristics of pancreatic cancer.

exPlanations (SHAP) enable transparent feature attribution, enhancing biomarker discovery and mechanistic interpretation (19). Notably, most prior studies have focused on single-omics analyses (e.g., transcriptomics or epigenomics), limiting their ability to capture cross-modal regulatory interactions. Integrating multi-omics data significantly increases feature dimensionality, exacerbating the “ $p \gg n$ ” problem and compromising model robustness (20). Traditional solutions include feature preselection or sample size expansion, but these can introduce selection bias or are often infeasible. In contrast, L1-regularized logistic regression [least absolute shrinkage and selection operator (LASSO)] offers a principled statistical framework for enforcing sparsity while maintaining predictive power in high-dimensional settings (21).

Objective

In this study, we employed LASSO-regularized logistic regression to construct multi-omics prognostic models based on six omics types and clinical variables in a cohort of 75 deceased PDAC patients, and validated their performance in an independent The Cancer Genome Atlas (TCGA)-pancreatic adenocarcinoma (PAAD) cohort comprising 63 cases with matched multi-omics profiles. In contrast to previous studies that rely on predefined gene panels or single-omics frameworks, our approach demonstrates the feasibility and utility of directly mining high-dimensional multi-omics data for clinically relevant prognostic biomarkers through interpretable, sparse machine learning models. We present this article in accordance with the TRIPOD reporting checklist (available at <https://gs.amegroups.com/article/view/10.21037/gs-2025-396/rc>).

Methods

Data sources

All multi-omics data used in the training and validation of machine learning models in this study were obtained from the study by “Proteogenomic characterization of pancreatic ductal adenocarcinoma” (22). Specifically, proteomic and phosphoproteomic data were downloaded from the Proteomic Data Commons (PDC, <https://pdc.cancer.gov/pdc/>), while genomic mutation, transcriptomic (mRNA), circRNA, and microRNA data were accessed via the Genomic Data Commons (GDC, <https://gdc.cancer.gov/>). Clinical data were derived from Table S1 of the same publication.

For testing the survival prediction models, independent

validation data were retrieved from TCGA (<https://www.cancer.gov/ccg/research/genome-sequencing/tcga>). Using the Cohort Builder, we selected the TCGA program and then chose the TCGA-PAAD project. We included samples that simultaneously contained the following three experimental strategies: RNA-Seq, miRNA-Seq, and Reverse Phase Protein Array (RPPA). Furthermore, only samples with data types including Gene Expression Quantification, miRNA Expression Quantification, and Protein Expression Quantification were retained. To ensure accurate survival annotations, we filtered for tumor tissues with a vital status of “Dead” and extracted survival time information. A total of 63 such samples were obtained. This study was conducted in accordance with the Declaration of Helsinki and its subsequent amendments.

Clinical feature encoding and data preprocessing

A total of 38 clinical features were selected and encoded. Categorical variables were transformed into one-hot encoded vectors, while ordinal variables were encoded using ordinal encoding. For tumor-node-metastasis (TNM) staging and resection status (R), due to the presence of “unknown” labels, we applied a dual-encoding strategy to ensure deterministic inputs: ordinal encoding was used to preserve staging order, and binary encoding was used to indicate presence of unknowns (e.g., “unknown” as 1, “known” as 0 for N, M, cM, and R). For T stage, a specific binary variable T_invasion was used to denote T3 or T4 stages, indicating tissue or vascular invasion. All missing [not available (NA)] values were encoded as 0.

After encoding, all features were normalized. Specifically:

- (I) For features assumed to follow a normal distribution, Z-score normalization was applied;
- (II) For features with long-tailed distributions, log-transformation [$\log(x+1)$] was used;
- (III) Ordinally encoded variables were scaled via Min-Max normalization;
- (IV) One-hot encoded variables were left unnormalized.

Formulas:

Z-score normalization:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \quad [1]$$

where μ_j and σ_j are the mean and standard deviation of feature j , computed from the training or validation set. If $\sigma_j = 0$, raw values were retained to avoid division by zero.

Log normalization:

$$z_{ij} = \ln(x_{ij} + 1) \quad [2]$$

where x_{ij} is the original feature value. A pseudo-count of 1 was added to handle zeros and avoid undefined logarithmic values.

Min-Max normalization:

$$z_{ij} = \frac{x_{ij} - x_j^{\min}}{x_j^{\max} - x_j^{\min}} \quad [3]$$

where $x_j^{\min}$ and $x_j^{\max}$ are the minimum and maximum values of feature j across all samples.

Machine learning models

Feature construction and multi-omics integration: for each omics modality, gene expression data were represented as a feature vector, in which each element corresponded to the normalized expression level of a specific gene or molecular feature. To achieve multi-omics integration, feature vectors from different omics layers were concatenated to form a unified global feature vector for each sample. This concatenation-based integration strategy preserves modality-specific information while enabling joint modeling across layers. The dimensionality of the resulting global vector served as the input dimension for downstream machine learning models.

Model selection and training strategy: given the high dimensionality of the integrated feature space (approximately 10^4 features), we employed logistic regression with L1 regularization to reduce computational complexity and mitigate overfitting. Regularization facilitates feature selection by shrinking less informative coefficients toward zero, thereby identifying key prognostic biomarkers while maintaining model interpretability. Accordingly, logistic regression was selected as the primary modeling framework.

The survival classification task was formulated as a binary classification problem: predicting whether the survival time is less than 1 year. Separate logistic regression models were trained for each omics modality.

Each logistic regression model consists of a linear layer followed by a sigmoid activation function. Given a feature vector $x = [x_1, x_2, \dots, x_d] \in \mathbb{R}^d$, the model computes:

$$Z = w^T x + b \quad [4]$$

$$\theta(z) = \frac{1}{1 + e^{-z}} \quad [5]$$

$$p = \theta(z) \quad [6]$$

where $w \in \mathbb{R}^d$ is the weight vector, b is the bias term, and $\theta()$ is the sigmoid function mapping to a probability in $(0, 1)$. The output p represents the predicted probability of being a tumor sample or having a survival time <1 year.

Training procedure

Models were trained on an NVIDIA V100 GPU for 1,000 epochs using Binary Cross Entropy (BCE) loss with L1 regularization. The Adam optimizer was used for gradient updates. To ensure generalizability, we employed 10-fold cross-validation. For each fold, we evaluated classification accuracy and area under the receiver operating characteristic curve (AUROC) on both training and validation sets, and the model checkpoint with the highest validation AUROC was saved.

Loss function:

$$\begin{aligned} \text{Loss} &= \frac{1}{N} \sum_{i=1}^N \mathcal{L}(y_i, f(x_i)) + \lambda \sum_{j=1}^d |w_j| \\ &= -\frac{1}{N} \sum_{i=1}^N [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] + \lambda \sum_{j=1}^d |w_j| \end{aligned} \quad [7]$$

where N is the number of samples, y_i is the ground truth label, $p_i = f(x_i)$ is the predicted probability, w_j is the model weight, and λ is the L1 regularization strength.

Feature selection and regularization strategy

To address the $P \gg n$ problem and reduce overfitting, we used L1-regularized logistic regression (LASSO) for feature selection. The procedure was:

- (I) Perform 10-fold cross-validation with LASSO regression to identify the optimal regularization strength λ by minimizing mean squared error (MSE).
- (II) Use the selected λ to fit a sparse logistic regression model and extract features with non-zero weights.
- (III) Repeat the cross-validation until at least 10 folds achieve accuracy >0.70.
- (IV) In each qualifying fold, extract the top 200 and bottom 200 features by weight magnitude. Features appearing in this range in at least 10 out of all folds (allowing two exceptions) were designated as positive and negative “signature” features.

Reported AUROC values represent the mean of models with accuracy >0.70 across repeated 10-fold cross-

validation.

Ablation experiments

Two ablation studies were conducted:

- (I) Removal of top-ranked signature genes identified from the survival prediction model to assess their contribution;
- (II) Removal of signature genes missing in the TCGA dataset to validate the model's robustness on external cohorts.

Both studies were performed on the full PDAC proteogenomic dataset (no train/validation split). For a trained survival model $f\theta: \mathbb{R}^d \rightarrow [0,1]$, and a signature feature set $\mathcal{F}_s \subseteq \{1, \dots, d\}$, we created an ablated input:

$$\begin{aligned} \mathbf{x}^{(s)} &= \mathbf{x} \odot \mathbf{m}^{(s)} \\ \mathbf{m}_i^{(s)} &= \begin{cases} 0 & \mathbf{x}_i \in \mathcal{F}_s \\ 1 & \mathbf{x}_i \notin \mathcal{F}_s \end{cases} \end{aligned} \quad [8]$$

Then computed:

$$\mathbf{p}^{(s)} = \sigma(\mathbf{w}^T \mathbf{x}^{(s)} + \mathbf{b}) \quad [9]$$

All evaluations were performed with fixed model parameters, altering only the input features.

SHAP analysis

For features with non-zero weights in the trained models, SHAP values were computed to quantify feature importance. The analysis was performed on the combined training and validation data using the selected fold model, implemented via the Python shap package. Within the RNA panel, we primarily focused on the contribution of mRNA features.

Model testing on TCGA

To validate generalizability, the trained models were applied to the TCGA test set. We ensured consistency by aligning the test feature vectors with the training data in terms of feature order, name, and dimension. Missing features in TCGA were imputed with zeros. Shared features were filled with normalized values from TCGA.

For the top survival model, we used the 12 models from cross-validation folds, applied to TCGA test samples with a default threshold of $P \geq 0.5$ for binary classification. Evaluation metrics included sensitivity, specificity, precision, recall, F1 score, accuracy, AUROC, and the area under the

precision-recall curve (PRAUC).

microRNA target gene prediction

To investigate the regulatory roles of microRNA signature genes, we used the miRDB (23) (<https://mirdb.org/>) and TargetScanHuman 8.0 (24) (https://www.targetscan.org/vert_80/). Predicted target genes were filtered based on:

- (I) miRDB: target score ≥ 80 (or all predictions if none met this threshold);
- (II) TargetScan: top 100 genes ranked by cumulative weighted context++ score.

Cell line and culture

Panc-1 (RRID: CVCL_0480) human PDAC cells were used in this study. The cell line was obtained from the Laboratory of General Surgery, Peking Union Medical College Hospital (Dr. Yuanyang Wang). Panc-1 cells were maintained in Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% fetal bovine serum (FBS), penicillin (100 U/mL), and streptomycin (100 $\mu\text{g/mL}$). Cells were cultured under standard conditions at 37 °C in a humidified incubator with 5% CO₂. The cell line was routinely tested and confirmed to be free of Mycoplasma contamination.

Plasmid construction

Knockdown (KD) and overexpression (OE) constructs targeting KDM4C, RAB41, and miR-1251-5p were generated for functional studies. For gene silencing, short hairpin RNA (shRNA) sequences specifically targeting each gene were cloned into a KD vector to suppress endogenous gene expression (Table S1). For OE, the coding sequence (CDS) of the target gene was cloned into an expression vector to enable ectopic expression. OE plasmids were introduced into cells to achieve transient gene expression (Table S2). Empty vector plasmids were used as negative controls in OE experiments while scramble shRNA in KD experiments.

Cell transfection

Cells were seeded at a density of 1×10^5 cells per well in 6-well plates and cultured overnight in antibiotic-free medium to allow adherence. KD or OE plasmids (3 μg per well) targeting KDM4C, RAB41, or miR-1251-5p were transfected using 9 μL EZ Trans Cell Transfection

Reagent (Efficient; AC04L091, Life-iLab, Shanghai, China) according to the manufacturer's instructions.

RNA extraction and quantitative real-time polymerase chain reaction (qRT-PCR)

Total RNA was extracted using TRIzol™ reagent (Invitrogen, Carlsbad, USA) according to the manufacturer's instructions. RNA concentration and purity were determined using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, Carlsbad, USA).

For KDM4C and RAB41, total RNA was reverse-transcribed into complementary DNA (cDNA), containing reverse transcriptase, random primers, RNase inhibitor, oligo(dT) primer, dNTP mixture, and reaction buffer, following the manufacturer's protocol. For miR-1251-5p, cDNA synthesis was performed using a stem-loop reverse transcription primer specific for miR-1251-5p after RNA extraction, to ensure selective amplification of mature miRNA. Except for the reverse transcription step, subsequent procedures were identical to those used for KDM4C and RAB41.

qRT-PCR was carried out on a 7500 Fast Real-Time PCR System (Applied Biosystems, Thermo Fisher Scientific) using SYBR Premix Ex Taq II (Takara Bio, Beijing, China). The cycling conditions were 95 °C for 30 s (initial denaturation), followed by 40 cycles of 95 °C for 5 s and 60 °C for 34 s. Relative gene expression levels were calculated using the $2^{-\Delta\Delta C_t}$ method. Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) was used as an internal control for KDM4C and RAB41, and U6 small nuclear RNA was used as the endogenous control for miR-1251-5p. Primer sequences are listed in [Table S3](#).

Western blot (WB) analysis

Western blotting was performed to assess the protein expression of KDM4C and RAB41. miR-1251-5p was not analyzed by WB, as it is a non-coding RNA. Cells were harvested and washed twice with cold phosphate-buffered saline (PBS), then lysed on ice using immunoprecipitation (IP) lysis buffer. Lysates were centrifuged, and the supernatants were collected and mixed with 5× loading buffer. Samples were boiled at 100 °C for 10 min before separation by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) at 120 V. Proteins were transferred onto polyvinylidene difluoride (PVDF) membranes under constant current (300 mA). Membranes were blocked with 5% non-fat milk in

Tris Buffered Saline with Tween 20 (TBST) for 2 h at room temperature, followed by overnight incubation at 4 °C with primary antibodies diluted in blocking buffer. After washing, membranes were incubated with appropriate secondary antibodies for 2 h at room temperature. Protein bands were visualized using an enhanced chemiluminescence system (Tanon 4600, Tanon Science & Technology, Shanghai, China).

Cell proliferation assay [Cell Counting Kit-8 (CCK-8)]

Cell viability was assessed using a CCK-8 (LABLEAD, CK001). Cells were seeded into 96-well plates at a density of 1,000 cells per well and allowed to adhere for 12 h; this time point was designated as day 0. Thereafter, at the same time each day (days 1, 2, and 3), CCK-8 reagent was added according to the manufacturer's instructions, and cells were incubated for 2 h. Absorbance was then measured at 430 nm (OD430) using a microplate reader.

Transwell migration assay

Cell migration was evaluated using Transwell chambers with an 8-μm pore size (24-well format). Prior to the assay, cells were serum-starved for 6 h to minimize the influence of serum-derived factors. Cells were then trypsinized, centrifuged, and washed 1–2 times with PBS. The cell pellet was resuspended in serum-free medium to prepare a single-cell suspension. A total of 5×10^4 cells in 100 μL serum-free medium were seeded into the upper chamber of each Transwell insert. The lower chamber was filled with 675 μL culture medium containing 30% FBS as a chemoattractant. Care was taken to avoid bubble formation beneath the insert; if bubbles were present, inserts were lifted and repositioned to ensure proper contact with the medium. Cells were incubated under standard culture conditions for 48 h. After incubation, the medium was removed, and the inserts were gently washed twice with PBS. Non-migrated cells on the upper surface of the membrane were carefully removed using a moist cotton swab. Migrated cells on the lower surface were fixed with 4% paraformaldehyde for 10 min, followed by staining with 0.1% crystal violet for 15–20 min. Excess stain was gently rinsed off with PBS (1–2 washes) before imaging and quantification.

Statistical analysis

All Student's t-tests conducted in this study were two-tailed and assumed unequal variances between groups (Welch's

t-test). These analyses were performed using the *scipy* module in Python. For non-parametric comparisons, the Mann-Whitney U test was used with default parameters implemented in the same package. Where applicable, statistical significance was defined as $P < 0.05$ unless otherwise specified. Prior to each test, data distributions were visually inspected to assess normality and variance homogeneity. No statistical methods were used to pre-determine sample size, and no samples or data points were excluded from the analyses. All analyses were conducted using Python (version 3.8.17) with relevant libraries including *numpy*, *pandas*, and *scipy*.

Results

Patient demographics

The training and validation cohorts for prognostic model development were derived from a comprehensive PDAC

proteogenomic dataset (22). Specifically, 75 deceased PDAC patients with complete multi-omics profiles—including RNA (RNA-seq), circRNA (RNA-seq), microRNA (miRNA-seq), proteomics (TMT), and phosphoproteomics (TMT)—were analyzed. An independent external test cohort comprising 63 deceased patients from the TCGA-PAAD project was utilized for validation and included RNA, microRNA, and proteomics data. No significant differences in sex, age, or overall survival were observed across patient groups (Mann-Whitney U test, *Table 1*).

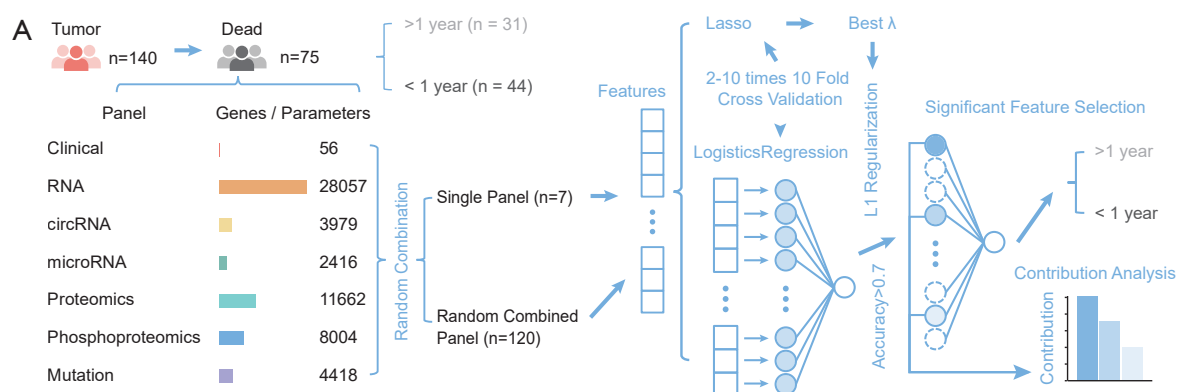
Survival prediction model development

Given the limited sample size, we dichotomized survival into <1 versus >1 year based on a 365-day threshold approximating the cohort median (*Table 1*). Seven single-panel models were trained—including six omics layers and one clinical feature panel—using cross-validated logistic regression (*Figure 1A*).

Table 1 Patient demographics

Variable	PDAC proteogenomic cohort				TCGA cohort			P value
	Censored	Dead			Dead			
		Total	<1 year	>1 year	Total	<1 year	>1 year	
Female	28	38	21	17	34	12	22	0.40
Male	37	37	23	14	29	16	13	0.18
Age, years	65.0 (62.0–71.0)	64.0 (56.5–70.5)	65.0 (56.5–72.0)	63.0 (57.0–69.0)	66.0 (60.5–74.0)	65.0 (59.2–75.2)	66.0 (62.5–73.0)	>0.99
Follow-up, days	628.0 (358.2–743.5)	311.0 (193.5–459.0)	232.5 (115.5–291.0)	599.0 (432.5–752.0)	394.0 (234.5–580.5)	221.5 (138.0–269.8)	511.0 (463.0–646.5)	>0.99

Data are presented as number or median (IQR). IQR, interquartile range; PDAC, pancreatic ductal adenocarcinoma; TCGA, The Cancer Genome Atlas.



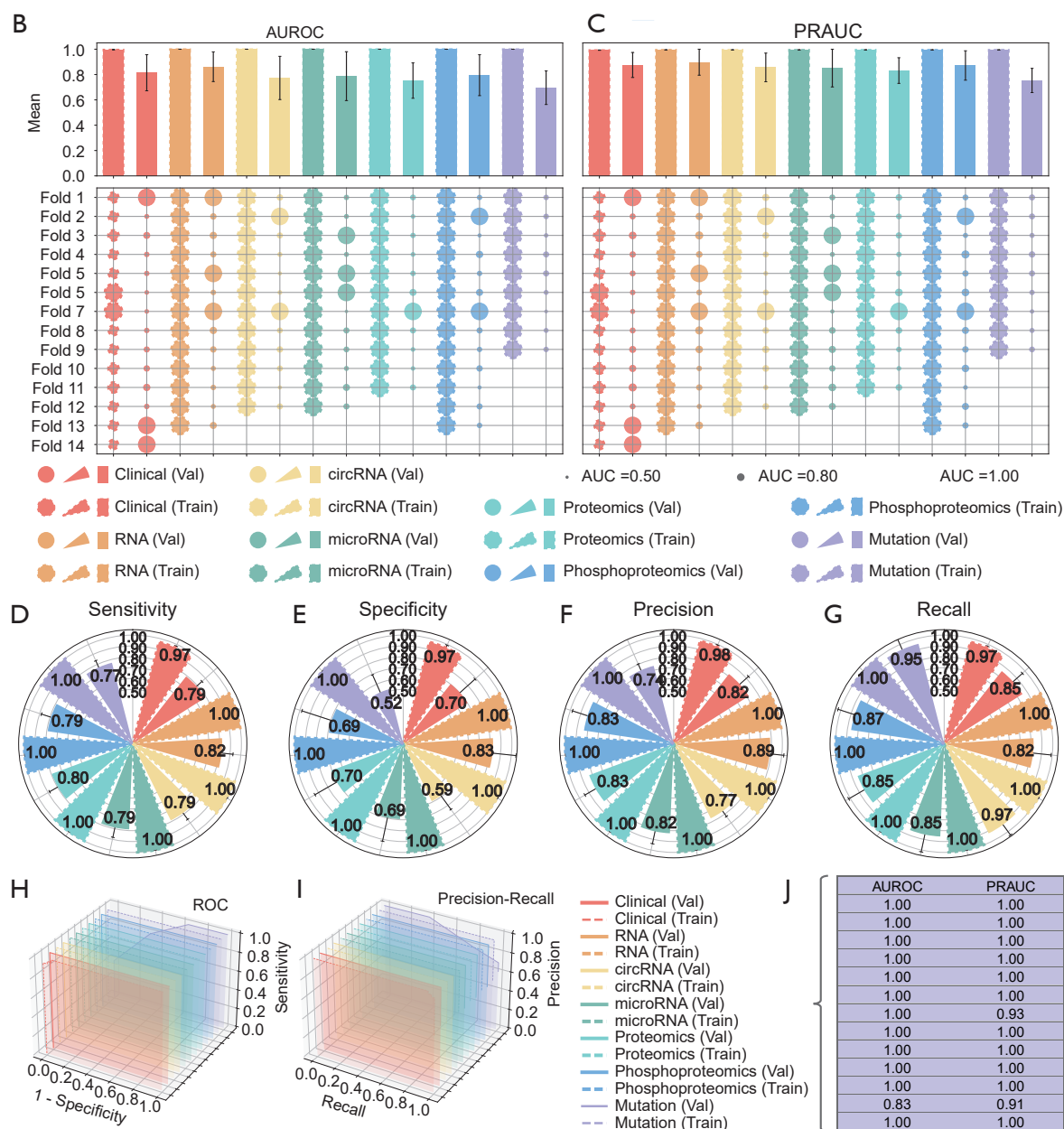


Figure 1 Architecture and training performance of survival prediction models using single omics panels. (A) Workflow of survival prediction model development. (B,C) AUROC and PRAUC values of the selected fold for clinical, RNA, circRNA, microRNA, proteomics, phosphoproteomics, and mutation panels in training and validation sets, including mean values. (D-G) Mean sensitivity, specificity, precision, and recall for training and validation sets across all panels. (H,I) ROC and precision-recall curves for the best-performing generation of each model on training and validation sets. (J) AUROC and PRAUC values of the best-performing generation for both training and validation sets. Model performances are color-coded: clinical (red), RNA (orange), circRNA (yellow), microRNA (green), proteomics (blue), phosphoproteomics (purple), and mutation (gray). Solid borders indicate validation set results; dashed borders indicate training set results. AUROC, average area under the receiver operating characteristic curve; AUC, area under the curve; PRAUC, the area under the precision-recall curve; ROC, receiver operator characteristic curve.

Excluding the mutation panel (due to insufficient model convergence), single-panel models achieved AUROC values of 0.33–1.00 and PRAUC of 0.52–1.00 (*Figure 1B,1C*). In validation, mean sensitivity, specificity, precision, and recall ranged from 0.77–0.82, 0.52–0.83, 0.74–0.89, and 0.82–0.97, respectively (*Figure 1D–1G*). Best-performing folds achieved AUROC and PRAUC of 0.83–1.00 and 0.91–1.00, respectively (*Figure 1H–1J*). While several single-panel models yielded strong performance in select folds, their overall predictive accuracy and consistency remained variable across omics types. These limitations highlighted the potential value of integrating complementary molecular information to improve prognostic robustness.

To explore multi-omics synergy, we constructed combinatorial models from random omics panel sets (*Figure 1A*). After filtering for accuracy >0.70 and repeating 10-fold cross-validation, multi-panel models consistently reached perfect training metrics. In validation, they achieved mean sensitivity, specificity, precision, recall, and F1 score ranging from 0.69–0.98, 0.36–0.92, 0.68–0.95, 0.67–0.98, and 0.74–0.90 (*Figure 2A*). The R^2 distribution of the marginal distributions of AUROC and PRAUC for the best models of 2 to 7 randomly combined panels ranges from 0.77 to 0.87 (*Figure 2B*). Corresponding accuracy, AUROC, and PRAUC were 0.73–1.00, 0.73–1.00, and 0.83–1.00 (*Figure 2C–2E*). The top five model combinations, based on mean AUROC across at least 10 runs, all reached perfect AUROC and PRAUC (1.00) in both training and validation sets (*Figure 2F,2G*). These top-performing combinations included: Top 1—RNA, microRNA, proteomics, and mutation; Top 2—clinical, RNA, proteomics, and mutation; Top 3—clinical, circRNA, and mutation; Top 4—microRNA, circRNA, and mutation; and Top 5—clinical and RNA panels. The best-performing model (Top 1: RNA + microRNA + proteomics + mutation) achieved accuracy of 0.71–1.00 (mean =0.87), AUROC of 0.67–1.00 (mean =0.92), and PRAUC of 0.80–1.00 (mean =0.95).

Feature selection from the optimal prognostic model

We identified 654 non-zero-weight features across 12 cross-validation runs for the Top1 model (*Figure 3A*). Feature ablation confirmed the predictive sufficiency of these features: models trained exclusively on the 654 features—including 381 RNA transcripts, 98 microRNAs, 172 proteomic markers, and 3 mutation features—retained comparable performance, whereas their exclusion resulted in significant performance degradation in more than half of

the models (*Figure 3B–3D*).

SHAP analysis was used to rank feature importance, and the top 20 genes per omics panel were visualized (*Figure 3E–3G*). Gene Ontology (GO) enrichment revealed transcriptional and metabolic processes for RNA features (*Figure 3H*), and SMAD signaling, autophagy, and energy metabolism for proteomics features (*Figure 3I*). Predicted targets of microRNA signatures (via TargetScanHuman 8.0 and miRDB) were enriched for acidic amino acid transport and secretion pathways (*Figure 3J,3K*). Notably, key molecular drivers within these signatures exhibited strong mechanistic links to pancreatic cancer biology: in RNA signatures, RNA processing regulators (HNRNPA2B1, RBMY1B) and metabolic enzymes (NDUFB8, PKLR) were enriched, potentially orchestrating transcriptional reprogramming and metabolic adaptation. Proteomic markers included metastasis-associated junctional proteins (CLDN5, MAGI2) and signaling effectors (EPHA2), aligning with SMAD-driven invasion and autophagic survival. microRNA signatures (miR-6727-3p, miR-5010-3p, miR-2278, miR-6501-5p) converged on secretion regulation and cytoskeletal dynamics through targeting transporters (SLC1A4, SLC7A11) and adhesion regulators (CAPZA1, RhoA, FBXW5), with enriched pathways including acidic amino acid transport, cell migration, and autophagic flux.

Cox regression on top-ranked genes identified CHPT1 (RNA) and miR-6501-5p (microRNA) as significantly associated with poor prognosis ($P<0.001$), while KDM4C (proteomics) showed a trend toward significance ($P=0.057$; *Figure S1*).

External validation using TCGA-PAAD cohort

To test generalizability, we evaluated survival models on the TCGA-PAAD cohort, retaining 25,446 RNA, 600 microRNA, and 172 proteomics features overlapping with our internal dataset. Among them, 326, 18, and 4 genes overlapped with our identified signature features, respectively (*Figure 4A*).

Masking of unavailable features in cross-validation showed minimal impact on internal performance (*Figure 4B*). On the TCGA dataset, the models maintained high prediction accuracy (*Figure 4C,4D*): accuracy, AUROC, and PRAUC were 0.94–1.00, 0.97–1.00, and 0.97–1.00, respectively. The sensitivity, specificity, precision, recall and F1 scores were 0.98–1.00, 0.97–1.00, 0.98–1.00, 0.98–1.00, 0.98–1.00, respectively. While sensitivity, specificity, recall, and AUROC remained consistent with internal results,

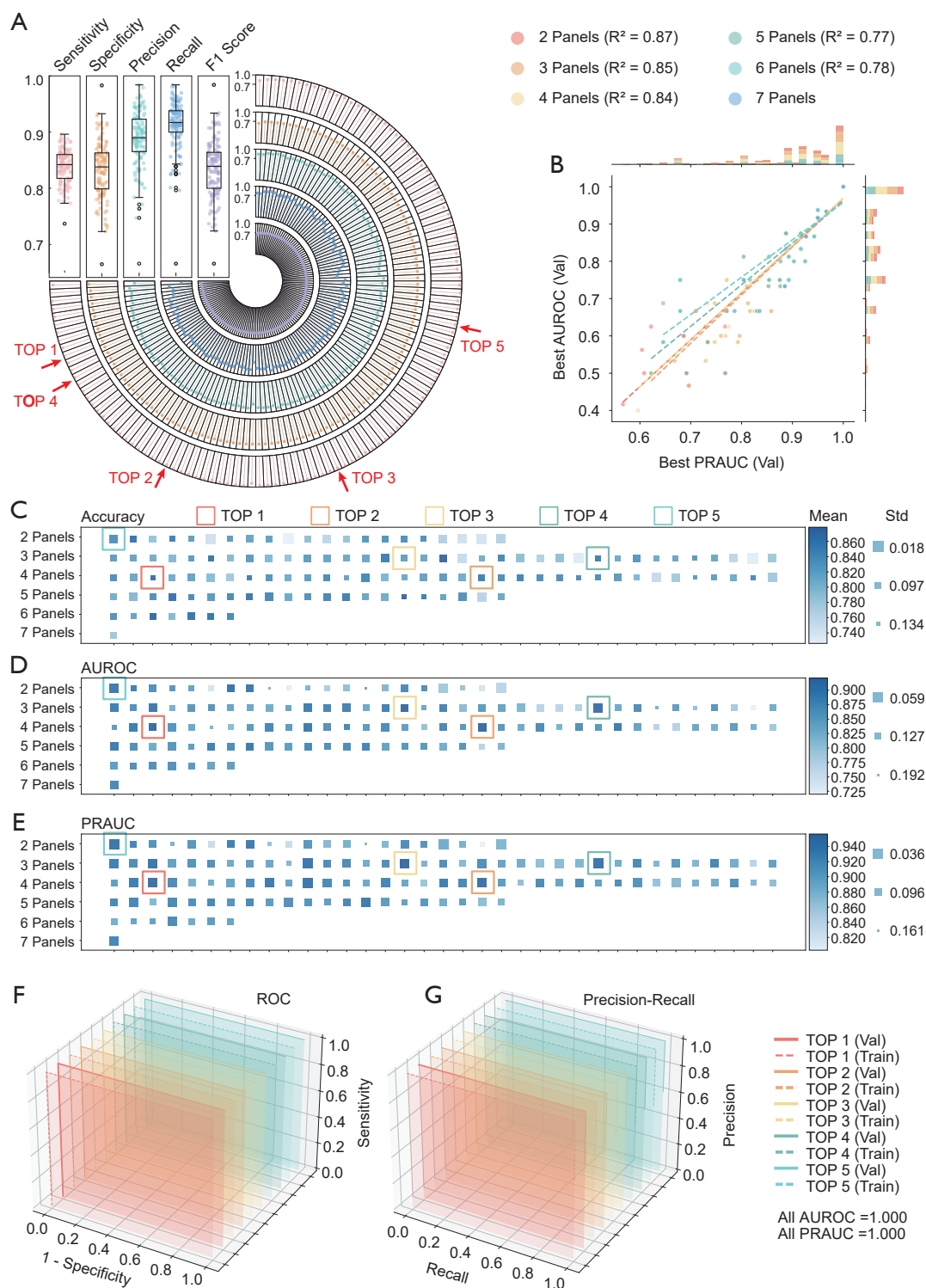


Figure 2 Training performance of random combined panel-based survival prediction models. (A) Average and distribution of sensitivity, specificity, precision, recall, and F1 score (arranged outer-to-inner in donut plots and left-to-right in box plots) for selected folds of 120 random combined panel models. Top 1–5 combinations are indicated by arrows. (B) Marginal distributions of AUROC and PRAUC for best-performing models using 2–7 randomly combined panels. (C–E) Heatmaps of average accuracy, AUROC, and PRAUC for models

using combinations of 2–7 panels. Color intensity indicates mean values; square size indicates standard deviation. Top 1–5 random combined panels are highlighted with red, orange, yellow, green, and cyan borders. (F,G) ROC and precision-recall curves of training and validation sets for the best-performing generation of the Top 1–5 random combined panels. AUROC, average area under the receiver operating characteristic curve; PRAUC, the area under the precision-recall curve; ROC, receiver operator characteristic curve; Std, standard deviation.

precision, accuracy, PRAUC ($P<0.05$), and F1 score ($P<0.01$) were significantly reduced (Figure 4C). All metrics except specificity were significantly different from the original unmasked models ($P<0.05$; Figure S2). The AUROC and PRAUC values of the 12 sets of cross-validation parameters of the TOP 1 prognostic model in the repeated tests of the TCGA-PAAD cohort were distributed from 0.97–1.00 and 0.98–1.00, respectively (Figure 4E–4G).

Modulation of signature genes alters malignant phenotypes in PANC-1 cells

To functionally validate the prognostic signature genes, we selected one top-ranked, previously uncharacterized gene from each of the RNA, proteomics, and microRNA panels (as of September 2025): RAB41, KDM4C, and miR-1251-5p. All three candidates were highly expressed in the poor-prognosis group. KD and OE constructs were generated for each gene in PANC-1 cells. Scramble shRNA served as the negative control for the KD group, whereas an empty expression plasmid (designated as “OE vector”) was used as the control for the OE group (Figure 5A,5B). Cell proliferation and migration were assessed using CCK-8 and Transwell assays, respectively. Silencing KDM4C ($P=0.001$) and miR-1251-5p ($P=0.02$) significantly reduced proliferation rates, while KD of RAB41 showed a similar inhibitory trend ($P=0.056$). Conversely, OE of RAB41 ($P<0.001$), KDM4C ($P<1\times 10^{-4}$), and miR-1251-5p ($P=0.041$) significantly enhanced proliferation (Figure 5C–5E; Table S4). In migration assays, KD of RAB41 ($P=0.003$), KDM4C ($P=0.002$), and miR-1251-5p ($P=0.04$) significantly impaired migratory capacity. In contrast, OE of RAB41 ($P<0.0001$), KDM4C ($P=0.046$), and miR-1251-5p ($P=0.008$) promoted migration, with miR-1251-5p OE inducing a particularly pronounced increase in migratory activity (Figure 5F,5G; Table S5). Collectively, these findings demonstrate that modulation of representative multi-omics signature genes directly influences PDAC malignant phenotypes, in a manner consistent with model-predicted risk associations.

Discussion

Key findings

In this study, we developed multi-omics-based prognostic models for PDAC and demonstrated that integrative approaches consistently outperformed single-omics models, whose AUROCs ranged from 0.70 to 0.86 (19). A compact set of signature genes was identified, and ablation experiments confirmed their essentiality: removing them degraded model performance to near-random, while retaining only signature genes preserved full accuracy. Among modalities, RNA and microRNA features contributed most to prognostic power, underscoring their central roles. For key genes identified from the microRNA, RNA, and proteomics layers, we further performed *in vitro* experiments to evaluate their functional roles. Modulating the expression of these candidate genes significantly altered malignant phenotypes of cancer cells, and these effects were consistent with the trends observed in our predictive models, thereby supporting the biological validity of the computational findings.

Strengths and limitations

Our approach combined diverse omics layers, systematically evaluated model interpretability through SHAP analysis, and validated robustness using ablation studies. These methodological strengths improved confidence in the biological relevance of identified features. However, several limitations should be acknowledged. First, model training was conducted using data derived from a specific cohort. Although external validation was performed using the TCGA dataset, the overall sample size in this study remains relatively modest compared with other large-scale machine learning-based multi-omics investigations. Therefore, our findings require confirmation in larger, prospective clinical cohorts. In particular, the sample sizes for microRNA and phosphoproteomics were modest, limiting statistical power. Residual batch effects could still confound model generalizability. Moreover, higher-order feature interactions

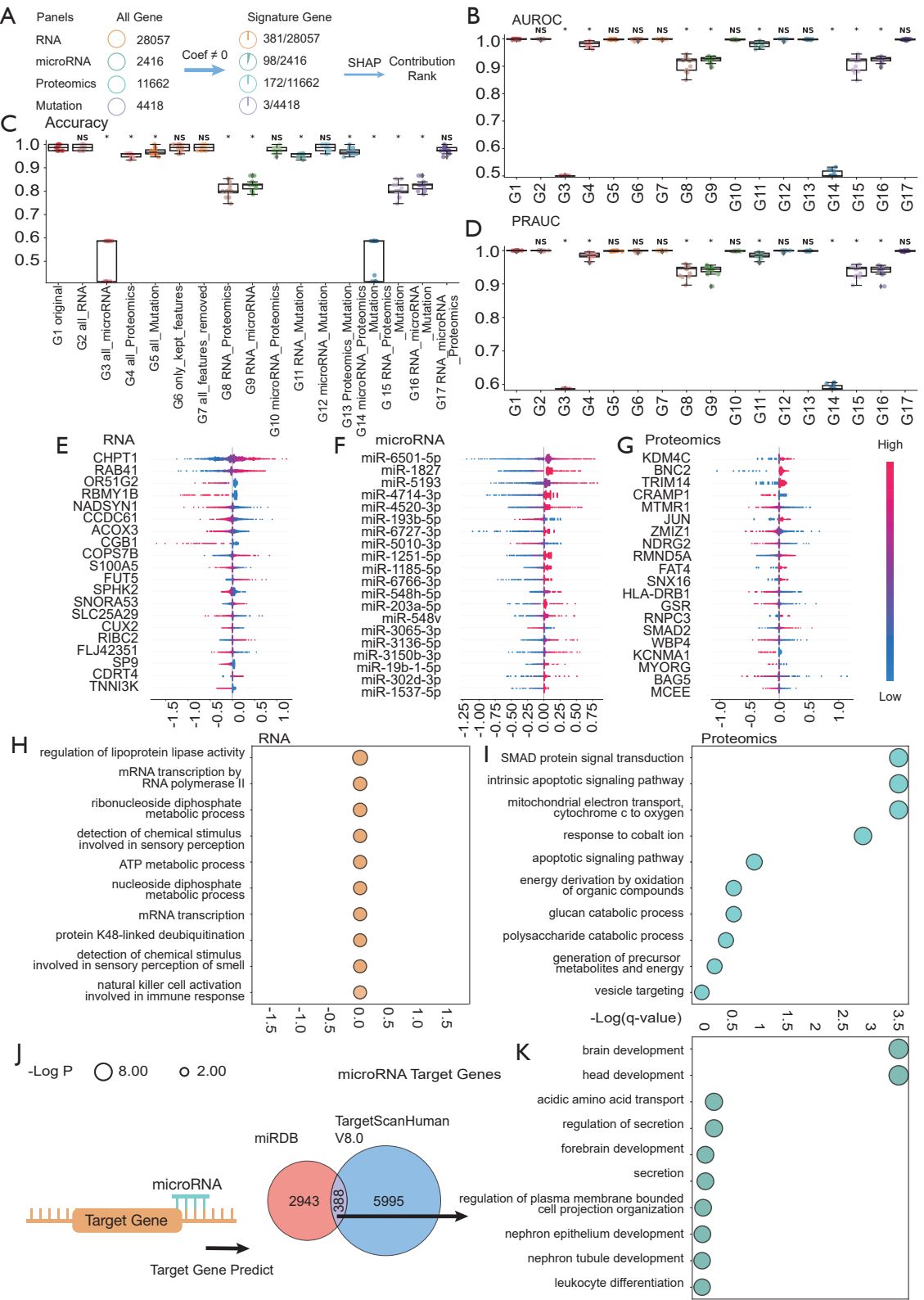


Figure 3 Feature selection and workflow of the Top 1 survival prediction model. (A) Pipeline for extracting signature genes from the Top 1 survival prediction model. (B-D) Impact of masking individual or combined panels of signature genes on model accuracy, AUROC, and

PRAUC. NS, no significant difference; *, significant difference. G1–G17: Group 1 to group 17, which is the group number. The Group names of B–D are the same. In (B,D) diagrams, only the group number should be used to refer to them. (E–G) SHAP value distributions for the top 20 RNA, microRNA, and proteomics features. Red and blue points concentrated on the left or right indicate tendencies to predict >1 or <1 year survival, respectively. (H,I) GO biological process enrichment terms for RNA and proteomics signature genes. (J) Workflow for identifying target genes of microRNA signature genes. (K) GO biological process enrichment terms for microRNA target genes. (I) and (K) share the X-axis scale. AUROC, average area under the receiver operating characteristic curve; GO, Gene Ontology; PRAUC, the area under the precision-recall curve; ROC, receiver operator characteristic curve; SHAP, SHapley Additive exPlanations.

were not fully captured, and validation was retrospective, necessitating future prospective studies. Although *in vitro* experiments were conducted to preliminarily confirm the functional impact of identified biomarkers on malignant phenotypes, *in vivo* validation using animal models and deeper mechanistic investigations are still lacking. Future studies should extend these findings through comprehensive mechanistic exploration and translational research to establish their clinical applicability.

Comparison with similar research

Inspired by a colorectal cancer study employing an ensemble learning framework for prognosis, several PDAC prognostic models have emerged. One such model based on 96 immune-related miRNAs achieved AUCs ranging from 0.647 to 0.789 in independent datasets (25). Another constructed from 31 hypoxia- and lactate metabolism-related genes reported test AUCs from 0.583 to 0.903 (26). These models, driven by interest gene clusters and machine learning ensembles, often focus on local patterns and may underrepresent global omics interdependencies. In our own models, the average AUROC of single-omics prognostic models ranged from 0.70 to 0.86 (19), leaving room for improvement. Indeed, model accuracy is a prerequisite for reliable feature interpretation.

Explanations of findings

Further functional analysis revealed that these signature genes mapped to critical cancer-related pathways. RNA signature genes were enriched in transcriptional regulation and nucleotide metabolism. Several markers correlated with poor prognosis, in especial RAB41 (also known as ISY1-RAB43) has been reported to be highly expressed in oral squamous cell carcinoma, where it promotes proliferation, invasion, and epithelial-mesenchymal transition (EMT) phenotypes (27). However, its role in PDAC remains uncharacterized. Interestingly, intermittent

teriparatide treatment has been shown to downregulate ISY1-RAB43 mRNA expression in bone marrow-derived mesenchymal stem cells (28), suggesting that teriparatide may represent a potential therapeutic candidate (29). Conversely, ACOX3—linked to reduced triglyceride accumulation—was associated with improved survival (30). Six compounds have been identified that stably bind to ACOX3 and enhance its expression (31), although further validation is required to determine whether therapeutic upregulation of ACOX3 can translate into survival benefit. Proteomics-derived signature genes enriched for SMAD signaling, autophagy, and energy metabolism pathways. Prognostically adverse proteins are mainly mediators of resistance, invasion, and immune evasion. Of particular interest, recent evidence indicates that KDM4C promotes PDAC cell proliferation and stemness. Pharmacologic inhibition of KDM4C using the pan-KDM4 inhibitor TACH107 effectively suppresses PDAC growth in preclinical models (32). TACH107 has entered phase I clinical trials and has demonstrated antitumor activity across multiple solid tumors (33), underscoring the translational relevance of our findings.

MicroRNA signature target gene analysis implicated both neural and oncogenic pathways. Poor outcomes were associated with miR-6727-3p (targeting CAPZA1) and miR-5010-3p (MAGEE2) (34,35), whereas protective markers included miR-2278 (RhoA) and miR-6501-5p (FBXW5) (36,37), pointing to mechanisms involving cytoskeletal remodeling, immune regulation, and metastasis. Notably, MUC4—associated with gemcitabine resistance and PDAC aggressiveness—was identified in our proteomic signatures (38), further validating the biological relevance of our findings. MicroRNAs are therapeutically actionable: their expression can be modulated by small molecules, or by delivery of tumor-suppressive miRNAs or inhibitors/antagonists targeting oncogenic miRNAs (39,40). For miR-1827 and miR-1251-5p—both associated with poor prognosis—corresponding inhibitory circRNAs (circ_0000808 and circ_0008362) have been identified, with

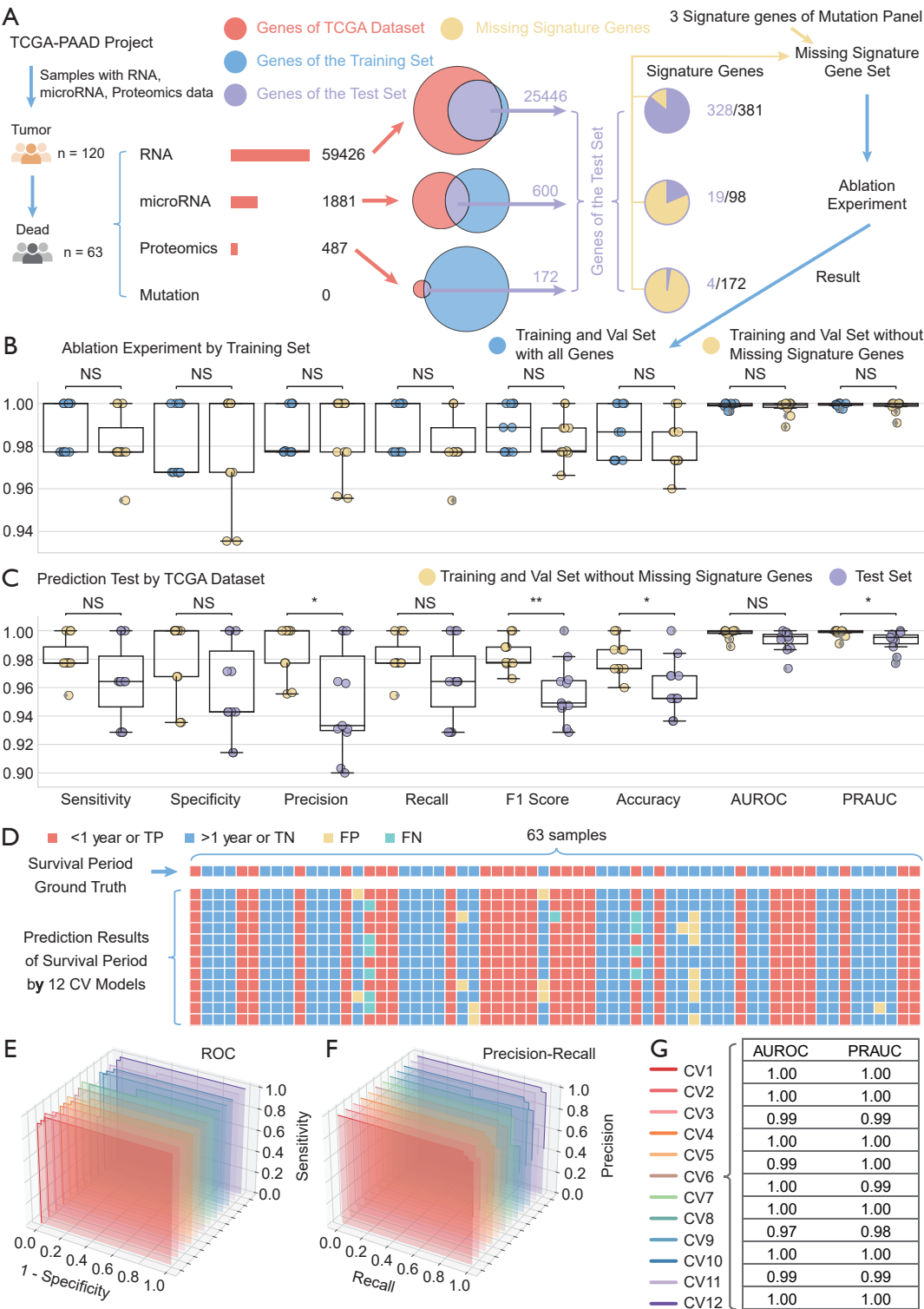


Figure 4 External validation of the Top 1 survival prediction model. (A) Workflow for test cohort selection from the TCGA-PAAD project and handling of missing signature genes. (B) Performance comparison between the masked Top 1 model (with missing signature genes excluded) and the original model on the PDAC proteogenomic cohort, measured by sensitivity, specificity, precision, recall, F1 score, accuracy, AUROC, and PRAUC. (C) Performance comparison on the test cohort (TCGA) with and without masking missing signature

genes. (D) Ground truth survival categories and predictions across the 12 selected folds of the Top 1 model for 63 test samples. Red: <1 year or TP; blue: >1 year or TN; yellow: FP; cyan: FN. (E-G) ROC and precision-recall curves, and AUROC/PRAUC values for predictions on the test set by all 12 cross-validation models. NS, no significant difference; *, $P<0.05$; **, $P<0.01$. CV, cross-validation; AUROC, average area under the receiver operating characteristic curve; FN, false negative; FP, false positive; PAAD, pancreatic adenocarcinoma; PDAC, pancreatic ductal adenocarcinoma; PRAUC, the area under the precision-recall curve; ROC, receiver operator characteristic curve; TCGA, The Cancer Genome Atlas; TN, true negative; TP, true positive.

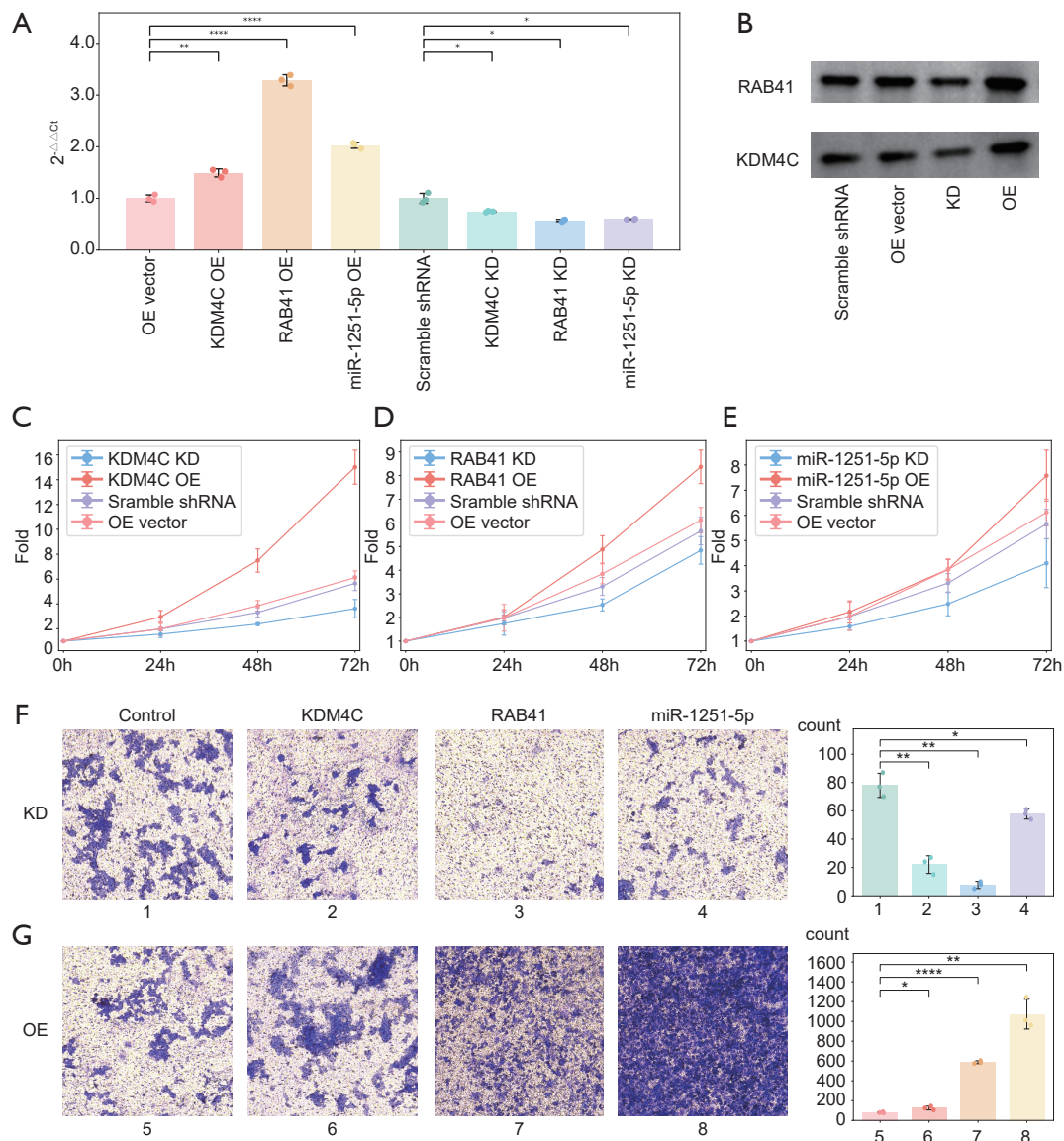


Figure 5 *In vitro* validation of representative signature genes in regulating malignant phenotypes of PDAC cells. (A) qRT-PCR confirmation of knockdown and overexpression efficiency. (B) Western blot validation of KDM4C and RAB41 protein expression. (C-E) Cell proliferation assessed by CCK-8 assay. Blue: KD; red: overexpression; purple: scramble shRNA control; pink: empty vector control. (F,G) Transwell migration assay showing representative staining images and quantitative analysis, while the control of KD group was scramble shRNA and the control of OE group was OE vector. Staining: crystal violet, magnification: 40 $\times$. NS, not significant; *, $P<0.05$; **, $P<0.01$; ****, $P<0.0001$. CCK-8, Cell Counting Kit-8; KD, knockdown; OE, overexpression; PDAC, pancreatic ductal adenocarcinoma; qRT-PCR, quantitative real-time polymerase chain reaction.

circ_0008362 successfully delivered via extracellular vesicles in experimental systems (41,42). Overall, the majority of prognostic features identified in this study align closely with established molecular mechanisms underlying tumor progression and therapeutic resistance.

Despite these promising results, several limitations must be acknowledged. Sample sizes for certain omics layers, particularly microRNA and phosphoproteomics, were modest, potentially limiting power. Residual batch effects may confound model generalizability, despite ablation-based mitigation. While SHAP analysis offered interpretable insights, higher-order interactions remain to be explored. Finally, our models were retrospectively validated; future prospective studies are essential to establish clinical utility.

Implications and actions needed

Our results highlight RNA, microRNA and proteomics features as highly informative biomarkers for prognosis and suggest that multi-omics integration more faithfully reflects the systemic regulatory architecture of tumors. This framework lays the groundwork for precision risk stratification and tailored therapeutic strategies in PDAC. Future research should prioritize prospective validation, deeper modeling of higher-order feature interactions, and expansion of underpowered omics layers. Ultimately, such integrative approaches may improve patient-specific prognostic accuracy and guide more effective clinical decision-making. Looking ahead, it will be essential to further integrate multi-omics discoveries with detailed clinical data. Future machine learning efforts should focus on linking multi-omics landscapes with clinical phenotypes, including patient symptoms, physical examination findings, laboratory parameters, and heterogeneity in treatment response. Such integrative analyses will facilitate the translation of molecular insights into clinical practice, providing stronger evidence for individualized diagnostic and therapeutic strategies in PDAC.

Conclusions

This study demonstrates the utility of integrative machine learning in uncovering clinically relevant biomarkers and improving prognostic prediction in PDAC. By systematically evaluating single-omics and multi-omics models across transcriptomic, proteomic, and regulatory

layers, we highlight the complementary contributions of RNA and microRNA in survival prediction and emphasize the biological interpretability enabled by sparse modeling and SHAP-based feature attribution. Importantly, functional assays in PANC-1 cells confirmed that representative prognostic signature genes directly modulate proliferative and migratory phenotypes in directions consistent with model-predicted risk associations. This concordance between computational prediction and experimental validation strengthens the biological credibility of our identified signatures. Together, our findings suggest that multi-omics integration—when coupled with interpretable algorithms and functional validation—provides a powerful framework for identifying robust molecular signatures that reflect the complex regulatory architecture of PDAC. Future work should prioritize prospective validation in larger cohorts and integration of spatial and temporal molecular data to further enhance translational impact in precision oncology.

Acknowledgments

None.

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at <https://gs.amegroups.com/article/view/10.21037/gs-2025-396/rc>

Data Sharing Statement: Available at <https://gs.amegroups.com/article/view/10.21037/gs-2025-396/dss>

Peer Review File: Available at <https://gs.amegroups.com/article/view/10.21037/gs-2025-396/prf>

Funding: This research was supported by the Noncommunicable Chronic Diseases-National Science and Technology Major Project (No. 2025ZD0552400), National Nature Science Foundation of China (Nos. 82172727 and 81572459), Nature Science Foundation of Beijing (No. 7202164), Chinese Academy of Medical Sciences Innovation Fund for Medical Sciences (CIFMS) (Nos. 2021-I2M-1-002 and 2023-I2M-2-002), and National High Level Hospital Clinical Research Funding (Nos. 2022-PUMCH-D-001 and 2022-PUMCH-B-004).

Conflicts of Interest: All authors have completed the ICMJE

uniform disclosure form (available at <https://gs.amegroups.com/article/view/10.21037/gs-2025-396/coif>). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. This study was conducted in accordance with the Declaration of Helsinki and its subsequent amendments.

Open Access Statement: This is an Open Access article distributed in accordance with the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License (CC BY-NC-ND 4.0), which permits the non-commercial replication and distribution of the article with the strict proviso that no changes or edits are made and the original work is properly cited (including links to both the formal publication through the relevant DOI and the license). See: <https://creativecommons.org/licenses/by-nc-nd/4.0/>.

References

1. Rozenfurt E, Eibl G. Pancreatic cancer: molecular pathogenesis and emerging therapeutic strategies. *Signal Transduct Target Ther* 2026;11:6.
2. Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. *CA Cancer J Clin* 2024;74:12-49.
3. Shi H, Qin C, Zhao Y, et al. Lymphatic metastasis in pancreatic cancer: from bedside to bench and back. *Cancer Metastasis Rev* 2025;44:79.
4. Qi J, Li M, Wang L, et al. National and subnational trends in cancer burden in China, 2005-20: an analysis of national mortality surveillance data. *Lancet Public Health* 2023;8:e943-55.
5. Mastrantoni L, Chiaravalli M, Spring A, et al. Comparison of first-line chemotherapy regimens in unresectable locally advanced or metastatic pancreatic cancer: a systematic review and Bayesian network meta-analysis. *Lancet Oncol* 2024;25:1655-65.
6. Pereira SP, Oldfield L, Ney A, et al. Early detection of pancreatic cancer. *Lancet Gastroenterol Hepatol* 2020;5:698-710.
7. Golan T, Hammel P, Reni M, et al. Maintenance Olaparib for Germline BRCA-Mutated Metastatic Pancreatic Cancer. *N Engl J Med* 2019;381:317-27.
8. Tempero MA, Malafa MP, Al-Hawary M, et al. Pancreatic Adenocarcinoma, Version 2.2021, NCCN Clinical Practice Guidelines in Oncology. *J Natl Compr Canc Netw* 2021;19:439-57.
9. Momtaz P, O'Connor CA, Chou JF, et al. Pancreas cancer and BRCA: A critical subset of patients with improving therapeutic outcomes. *Cancer* 2021;127:4393-402.
10. Conroy T, Castan F, Lopez A, et al. Five-Year Outcomes of FOLFIRINOX vs Gemcitabine as Adjuvant Therapy for Pancreatic Cancer: A Randomized Clinical Trial. *JAMA Oncol* 2022;8:1571-8.
11. Placido D, Yuan B, Hjaltelin JX, et al. A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. *Nat Med* 2023;29:1113-22.
12. Doebele RC, Drilon A, Paz-Ares L, et al. Entrectinib in patients with advanced or metastatic NTRK fusion-positive solid tumours: integrated analysis of three phase 1-2 trials. *Lancet Oncol* 2020;21:271-82.
13. McIntyre CA, Grimont A, Park J, et al. Distinct clinical outcomes and biological features of specific KRAS mutants in human pancreatic cancer. *Cancer Cell* 2024;42:1614-1629.e5.
14. Yadav S, Kasi PM, Bamlet WR, et al. Effect of Germline Mutations in Homologous Recombination Repair Genes on Overall Survival of Patients with Pancreatic Adenocarcinoma. *Clin Cancer Res* 2020;26:6505-12.
15. Collisson EA, Bailey P, Chang DK, et al. Molecular subtypes of pancreatic cancer. *Nat Rev Gastroenterol Hepatol* 2019;16:207-20.
16. Hwang WL, Jagadeesh KA, Guo JA, et al. Single-nucleus and spatial transcriptome profiling of pancreatic cancer identifies multicellular dynamics associated with neoadjuvant treatment. *Nat Genet* 2022;54:1178-91.
17. Tong Y, Sun M, Chen L, et al. Proteogenomic insights into the biology and treatment of pancreatic ductal adenocarcinoma. *J Hematol Oncol* 2022;15:168.
18. Wang L, Liu Z, Liang R, et al. Comprehensive machine-learning survival framework develops a consensus model in large-scale multicenter cohorts for pancreatic cancer. *Elife* 2022;11:e80150.
19. Liu Z, Liu L, Weng S, et al. Machine learning-based integration develops an immune-derived lncRNA signature for improving outcomes in colorectal cancer. *Nat Commun* 2022;13:816.
20. Candes EJ, Tao T. Near-Optimal Signal Recovery From Random Projections: Universal Encoding Strategies? *IEEE Transactions on Information Theory* 2006;52:5406-25.
21. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 1996;58:267-88.

22. Cao L, Huang C, Cui Zhou D, et al. Proteogenomic characterization of pancreatic ductal adenocarcinoma. *Cell* 2021;184:5031-5052.e26.
23. Chen Y, Wang X. miRDB: an online database for prediction of functional microRNA targets. *Nucleic Acids Res* 2020;48:D127-31.
24. McGeary SE, Lin KS, Shi CY, et al. The biochemical basis of microRNA targeting efficacy. *Science* 2019;366:eaav1741.
25. Ge J, Ge J, Tang G, et al. Machine learning-based identification of biomarkers and drugs in immunologically cold and hot pancreatic adenocarcinomas. *J Transl Med* 2024;22:775.
26. Zheng Y, Yang Y, Xiong Q, et al. Establishment and Verification of a Novel Gene Signature Connecting Hypoxia and Lactylation for Predicting Prognosis and Immunotherapy of Pancreatic Ductal Adenocarcinoma Patients by Integrating Multi-Machine Learning and Single-Cell Analysis. *Int J Mol Sci* 2024;25:11143.
27. Wang J, Ouyang S, Zhao S, et al. SP1-Mediated Upregulation of circFAM126A Promotes Proliferation and Epithelial-Mesenchymal Transition of Oral Squamous Cell Carcinoma via Regulation of RAB41. *Front Oncol* 2022;12:715534.
28. Vrščaj LA, Marc J, Ostanek B. Differential expression of lncRNAs and mRNAs in bone marrow-derived mesenchymal stem cells under continuous and intermittent teriparatide treatment. *Biomed Pharmacother* 2025;189:118293.
29. Kido T, Tabatabai ZL, Chen X, et al. Potential dual functional roles of the Y-linked RBMY in hepatocarcinogenesis. *Cancer Sci* 2020;111:2987-99.
30. Yu G, Cheng CJ, Lin SC, et al. Organelle-Derived Acetyl-CoA Promotes Prostate Cancer Cell Survival, Migration, and Metastasis via Activation of Calmodulin Kinase II. *Cancer Res* 2018;78:2490-502.
31. Huang H, Yang M, Fu Z, et al. Integrating multi-omics to reveal the causal effect of gut microbiota on post-traumatic stress disorder and potential drug target analysis. *Eur J Psychotraumatol* 2025;16:2593212.
32. Shaheen MT, Dhebat S, Rajapakshe KI, et al. The Lysine Demethylase KDM4C Is an Oncogenic Driver and Regulates ERK Activity in KRAS-Mutant Pancreatic Ductal Adenocarcinoma. *Cancer Res Commun* 2026;6:245-59.
33. Chandhasin C, Dang V, Perabo F, et al. TACH101, a first-in-class pan-inhibitor of KDM4 histone demethylase. *Anticancer Drugs* 2023;34:1122-31.
34. Huang D, Cao L, Zheng S. CAPZA1 modulates EMT by regulating actin cytoskeleton remodelling in hepatocellular carcinoma. *J Exp Clin Cancer Res* 2017;36:13.
35. Arora M, Kumari S, Singh J, et al. Downregulation of Brain Enriched Type 2 MAGEs Is Associated With Immune Infiltration and Poor Prognosis in Glioma. *Front Oncol* 2020;10:573378.
36. Tuntithavornwat S, Shea DJ, Wong BS, et al. Giant obscurin regulates migration and metastasis via RhoA-dependent cytoskeletal remodeling in pancreatic cancer. *Cancer Lett* 2022;526:155-67.
37. Liang Y, Chen P, Wang S, et al. SCF(FBXW5)-mediated degradation of AQP3 suppresses autophagic cell death through the PDPK1-AKT-MTOR axis in hepatocellular carcinoma cells. *Autophagy* 2024;20:1984-99.
38. Sagar S, Leiphrakpam PD, Thomas D, et al. MUC4 enhances gemcitabine resistance and malignant behaviour in pancreatic cancer cells expressing cancer-associated short O-glycans. *Cancer Lett* 2021;503:91-102.
39. Yang H, Liu Y, Chen L, et al. MiRNA-Based Therapies for Lung Cancer: Opportunities and Challenges? *Biomolecules* 2023;13:877.
40. Vahabi M, Dehni B, Antomás I, et al. Targeting miRNA and using miRNA as potential therapeutic options to bypass resistance in pancreatic ductal adenocarcinoma. *Cancer Metastasis Rev* 2023;42:725-40.
41. Cai Y, Dong Z, Wang J. Circ_0000808 promotes the development of non-small cell lung cancer by regulating glutamine metabolism via the miR-1827/SLC1A5 axis. *World J Surg Oncol* 2022;20:329.
42. Lin X, He SQ, Shan SK, et al. Endothelial cells derived extracellular vesicles promote diabetic arterial calcification via circ_0008362/miR-1251-5p/Runx2 axial. *Cardiovasc Diabetol* 2024;23:369.

Cite this article as: Chen B, Chen J, Cao Z, Zhang Q, Yan S, Cao H, Yu J, Wang F, Liu Z. Explainable multi-omics modeling for risk stratification in pancreatic ductal adenocarcinoma. *Gland Surg* 2026;15(4):91. doi: 10.21037/gs-2025-396