scientific reports

OPEN

Machine learning reveals glycolytic key gene in gastric cancer prognosis

Nan Li[1,4], Yuzhe Zhang[2,4], Qianyue Zhang[1,4], Hao Jin[1], Mengfei Han[1], Junhan Guo[3] & Ye Zhang[2]✉

Glycolysis is recognized as a central metabolic pathway in the neoplastic evolution of gastric cancer, exerting profound effects on the tumor microenvironment and the neoplastic growth trajectory. However, the identification of key glycolytic genes that significantly affect gastric cancer prognosis remains underexplored. In this work, five machine-learning algorithms were used to elucidate the intimate association between the glycolysis-associated gene phosphofructokinase fructose-bisphosphate 3 (PFKFB3) and the prognosis of gastric cancer patients. Validation across multiple independent datasets confirmed the prognostic significance of PFKFB3. Further, we delved into the functional implications of PFKFB3 in modulating immune responses and biological processes within gastric cancer patients, as well as its broader relevance across multiple cancer types. Results underscore the potential of PFKFB3 as a prognostic biomarker and therapeutic target in gastric cancer. Our project can be found at https://github.com/PiPiNam/ML-GCP.

Keywords Machine learning, Gene identification, Prognostic, Gastric cancer, PFKFB3

Gastric cancer (GC) is the fifth most prevalent malignant tumor globally and causes the second leading cancer-related mortality[1–5]. The initial symptoms of GC are often subtle, leading to a late-stage diagnosis for the majority of patients, which consequently results in a poor prognosis for GC patients[6–11]. Despite the significant improvements in diagnostic and therapeutic technologies that have greatly enhanced the quality of life and survival duration for patients, the overall prognosis for GC remains grim, with postoperative recurrence and drug resistance being frequent complications[12,13]. An in-depth investigation into the prognostic factors of gastric cancer facilitates the early identification of high-risk cohorts, enabling early diagnosis and intervention, thereby enhancing patient survival rates and quality of life.

Tumors undergo complex metabolic reprogramming during their evolution, which has increasingly drawn the attention of researchers due to its implications for tumor energy metabolism[14–18]. Among the myriad metabolic activities, glycolysis plays a pivotal role in the growth and dissemination of tumors[19–21]. The intricate processes of glycolysis not only provide the essential energy and biosynthetic precursors required for rapid proliferation of gastric cancer (GC) but also alter the tumor microenvironment to facilitate cancer progression[22]. Aerobic glycolysis, even under both hypoxic and normoxic conditions, is utilized by tumor cells as the primary means of energy provision[23,24]. This implies that, even in the presence of ample oxygen, cancer cells can produce large amounts of lactic acid, supplying the tumor with energy and compounds that support growth[17,20]. Some studies have indicated that genes associated with glycolysis can be effectively used to assess the prognosis of GC patients. Therefore, targeting genes involved in the aerobic glycolysis pathway is vital for illuminating the distinctive metabolic features of GC. Personalized therapeutic strategies derived from this understanding hold the potential to efficiently impede tumor progression, thereby enhancing therapeutic efficacy while minimizing side effects. Despite the paramount importance of such strategies, the fundamental mechanisms driving GC glycolysis remain partially uncharted. Consequently, it becomes imperative to conduct further investigations into potential prognostic biomarkers in GC by leveraging a comprehensive set of glycolytic genes.

The rapid expansion of genomic technologies to characterize healthy and diseased patient populations provides unprecedented solutions to the pathophysiological drivers of cancer and many other diseases. In 2018, The Cancer Genome Atlas (TCGA) completed a 10-year study of 33 tumor types in approximately 11,000

[1]China Academy of Electronics and Information Technology, National Engineering Research Center for Public Safety Risk Perception and Control by Big Data (RPP), Beijing, China. [2]The First Laboratory of Cancer Institute, The First Hospital of China Medical University, Shenyang, China. [3]Center for Reproductive Medicine, Henan Key Laboratory of Reproduction and Genetics, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China. [4]Nan Li, Yuzhe Zhang and Qianyue Zhang contributed equally to this work. ✉email: zhangyecmu@163.com

patients[25,26]. A major goal of the project was to provide publicly available datasets to help improve diagnostic methods, treatment standards, and ultimately prevent cancer[27]. In our study, we utilized TCGA's extensive clinical and RNA-seq datasets to study gene expression profiles relevant to our research goals.

In recent years, the application of artificial intelligence and machine learning methods to study the prognosis of cancer has become a hot topic in bioinformatics, demonstrating immense potential in identifying key genes associated with diseases[28,29]. These methods have emerged as an important tool for screening biomarkers that affect the prognosis of gastric cancer (GC)[30]. However, few studies integrate multiple machine-learning algorithms to select biomarkers influencing the prognosis of GC based on the expression of various genes and their prognostic outcomes[31]. Each of the machine learning models has different assumptions and optimization strategies that reveal the potential relationship between numerous genetic factors and gastric cancer prognosis from various perspectives, which increases the chances of discovering potential biomarkers. For example, a tree-based model provides feature importance and decision paths, while a support vector machine can highlight the decision boundary[32]. In addition, combining multiple models can reduce the risk of overfitting, as different models may have various sensitivities to noise and outliers in the data. If one model is sensitive to a particular type of noise, other models may still be able to predict the outcome accurately. Using multiple models allows us to validate our results against each other. If the biomarkers identified by one model are also considered important in another, this increases the credibility of these markers as true biomarkers[33,34].

This study aims to utilize bioinformatics methods to obtain gene matrices related to GC patients from the TCGA database and apply five machine-learning algorithms to screen for important feature genes that affect the prognosis of GC. By exploring their potential as biomarkers, this research seeks to provide new references for the clinical diagnosis and treatment of gastric cancer.

Materials and methods

This section begins with the overall investigation procedure, followed by a detailed presentation of the experimental methodologies and outcomes for each respective part (Fig. 1).

Dataset and preprocessing

TCGA-STAD (The Cancer Genome Atlas Stomach Adenocarcinoma) is the only gastric cancer dataset available for this study. Clinical data and RNA-seq data used in this study were obtained from The Cancer Genome Atlas

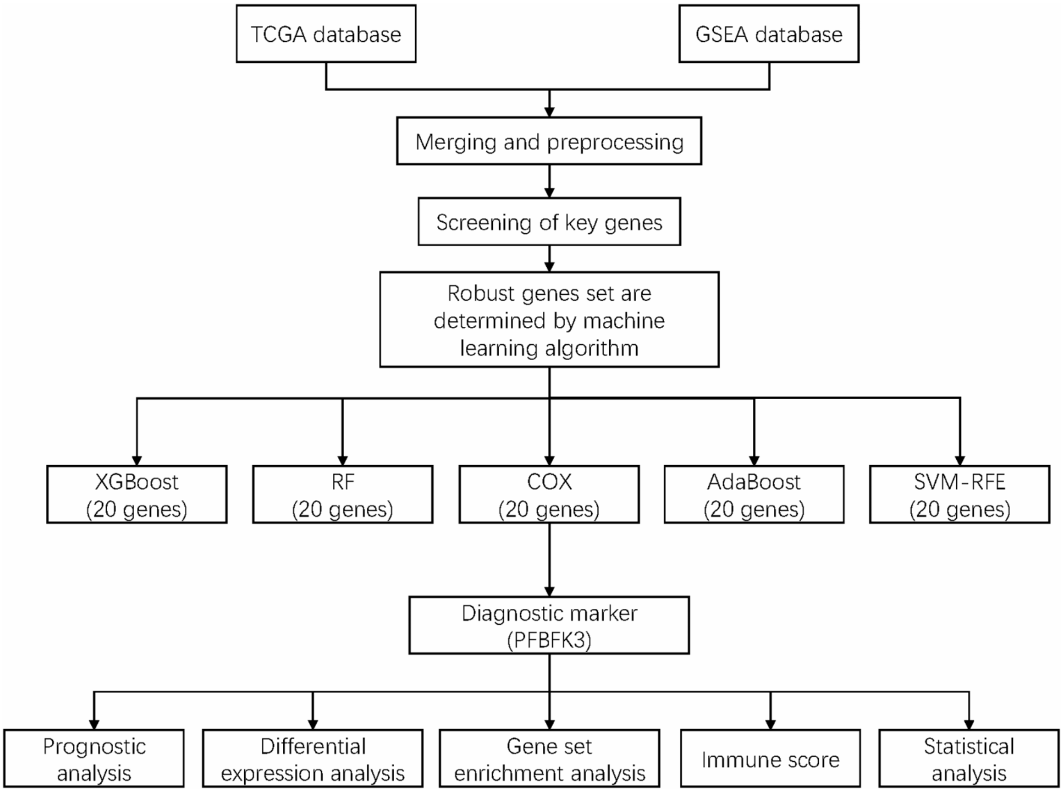

Fig. 1. The flowchart depicting the investigation procedure. We obtained data on gastric cancer patients and glycolysis-related genes from the TCGA and GSEA databases. The genes were then screened using machine learning methods. Dataset Preprocessing, XGBoost, COX proportional hazards model, AdaBoost Adaptive Boosting, RF random forest, SVM-RFE support vector machine-recursive feature elimination. For the screened genes, we performed prognostic analysis, enrichment analysis, immune correlation analysis and pan-cancer analysis.

(TCGA) database. The Glycolysis Gene Set is derived from the Gene Set Enrichment Analysis (GSEA) database, which provides a curated collection of genes associated with specific biological processes and pathways. The Glycolysis Gene Set organized by GSEA includes a subset of genes known to be involved in the glycolysis pathway. For the convenience of readers, a list of abbreviations for various cancers along with their full names is available at the TCGA's official resources page, which can be found at TCGA Study Abbreviations. In this study, we used gene expression data from the TCGA database. RNA-seq is a high-throughput sequencing technology that can comprehensively and accurately detect the expression level of gene transcripts.

GSE28541, GSE34942, GSE15459 and GSE26253 from the GEO database were used to validate the prognostic value of PFKFB3[35–38]. GSE29272 from the GEO database was used to validate the differential expression of PFKFB3 in gastric cancer[39]. IMvigor210 cohort was used to explore the role of PFKFB3 in the immunotherapy cohort[40]. The GSE28541, GSE34942, GSE15459, GSE26253 and GSE29272 datasets are based on microarray technology.

The gastric cancer patient dataset needs to be preprocessed to make it more compatible with the input requirements of various machine learning algorithms. Initially, we conducted data cleaning, which involved the removal of noise and duplicate data, as well as the imputation of missing values. In this step, we first eliminated samples that lacked critical information (such as patient survival data) and features with a high rate of missing data (exceeding 10%). Specifically, for the TCGA-STAD dataset, which documents the expression (continuous variables) of 60,660 genes in 409 gastric cancer patients, we separately calculated the median expression of each gene in the total patient samples and replaced the missing values with it. For the SURVIVAL dataset, which records patient information and survival, we separately counted the multitude values of the discrete variables needed for subsequent analysis and used them to fill in the missing values. Subsequently, the normalization of discrete and continuous features using one-hot coding and z-score methods respectively. By identifying the overlap between TCGA and key glycolysis module genes, we recognized 75 overlapping gene regions. Ultimately, we compiled a dataset comprising 75 genes across 371 samples, along with their respective clinical data. And by comparing with the known gastric cancer-causing genes obtained from the Malacard database, we found that there are 12 overlapping genes among them. After data cleaning and standardization, we obtained the final dataset for diagnostic marker screening.

The experiments conducted in this manuscript utilize gastric cancer-related data from the TCGA database, which includes the transcriptome expression data and clinical data of gastric cancer patients. Since this research complies with the established protocols of TCGA and GSEA, there is no need for ethical review or patient-informed consent.

Design of machine learning algorithms

To ensure that the genes selected are highly interpretable, we have integrated five machine learning algorithms for the screening of significant prognostic biomarkers in gastric cancer: Random Forest (RF), XGBoost, Cox Proportional Hazards Model, AdaBoost, and Support Vector Machine with Recursive Feature Elimination (SVM-RFE).

We primarily utilize these methods for feature selection, specifically the identification of key genes. Specifically, we modeled regressions individually using these algorithms, characterized by the expression of 75 genes and labeled by the OS time of the clinical sample (the time span from diagnosis initiation to the last diagnosis of the sample). For instance, the original Support Vector Machine determines a classification hyperplane based on the support vectors, and the fitted hyperplane can be used to predict a continuous target value in a regression task. When combined with a recursive feature elimination method, the SVM-RFE approach is effectively used for feature selection in our research. Python packages including "sklearn," "future," "lifelines," "xgboost," "scikit-survival," "numpy," and "pandas" were to implement the machine learning techniques in this study[41–43].

Optimal parameter selection in our research was conducted using fivefold cross-validation, and we employed the GridSearchCV method from the "sklearn" package to conduct an optimal hyperparameter search. It systematically explores a grid of predefined hyperparameter values and evaluates every possible combination. For each combination, it performs multiple iterations of model training on a dataset partitioned into training and validation sets, followed by an assessment of model performance through cross-validation. Ultimately, it returns the best model that yields the highest cross-validated performance. Subsequently, those optimized machine models were used to identify 20 key genes that influence the prognosis of gastric cancer (GC) from each model, and then determined the intersection among these screened genes to uncover the most critical genes.

Prognostic analysis

The prognostic data from the UCSC Xena database were leveraged to analyze Overall Survival (OS), Disease-Specific Survival (DSS), Disease-Free Interval (DFI), and Progression-Free Interval (PFI). By employing univariate Cox regression and the Kaplan–Meier model, we evaluated the prognostic influence of screened genes on each malignant tumor-specific prognostic type, and depicted the results in a heatmap. The optimal cutoff grouping is implemented using the surv_cutpoint function in the R package survminer and specifies that the sample size under any grouping scheme is not less than 5% of the total population[44]. The implementation of optimal cutoff grouping relies on the method of selecting the most significant p-value[45,46]. Differences in survival between the two groups were compared by Cox regression Furthermore, we utilized the BEST database[47] to explore the relationship between screened genes and patient prognosis in external datasets and to study the correlation between screened genes expression in cancer patients and their response to immunotherapy.

Differential expression analysis and enrichment analysis

The DESeq R package was employed to perform a differential expression analysis of PFKFB3[48]. Following this, Gene Set Enrichment Analysis (GSEA) was conducted using the clusterProfiler R package based on the outcomes of the differential expression analysis[49]. The genes identified as differentially expressed were then analyzed using Gene Ontology (GO) analysis, which encompassed Biological Process (BP), Cellular Component (CC), and Molecular Function (MF), along with Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis.

Immunization scores

The "estimate" and "ssgsea" algorithms were employed to assess the immune infiltration scores and the infiltration patterns of immune cells in patients with TCGA-STAD[50,51].

Statistical analysis

To assess the correlation between two groups, Pearson's linear correlation coefficient was employed. For survival analysis, the Kaplan–Meier method was used in conjunction with the log-rank test to compare survival distributions between groups. Statistical significance was set at a threshold of $p < 0.05$, with the following notation used to denote levels of significance: *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$. This approach ensures both a rigorous examination of relationships within the data and a clear communication of the strength and importance of observed associations and differences.

Results

Screening and validation of diagnostic markers

In order to select specific gastric cancer-related genes from a large gene pool, five machine learning algorithms were implemented from the Python scikit-learn and XGBoost package: SVM-RFE, AdaBoost, RF, XGBoost and Cox Proportional Hazards Model to discern genes that affect the prognosis of gastric cancer (Fig. 2). To guarantee the validity and robustness of the gene feature selection models, we adopted the widely-used fivefold cross-validation method and grid search algorithm in the machine learning domain to achieve the best model and enhance the algorithms' performance in gene feature selection.

Furthermore, to ensure the comparability of the machine learning models' performance on this task, we calculated the average of the minimum MSE obtained from three runs for each model and used it as the model's MSE. We then computed the relative MSE values for each model, using the minimum MSE as a baseline for comparison, in order to visually the performance differences between the models. Since the Cox proportional hazards model does not typically utilize MSE as an evaluation metric, its results are not presented here. The results in the Table 1 demonstrate that the performance of the four machine learning models is comparable for this task. By utilizing this methodology, we ensured that the feature genes identified were closely correlated with the prognosis and survival of gastric cancer patients.

The filtered dataset consists of only 75 genes; therefore, selecting approximately one-quarter of these genes—specifically, 20 features—is reasonable given the total number of features available. We believe this percentage will encompass the genes that most significantly influence survival time while preserving the model's simplicity and

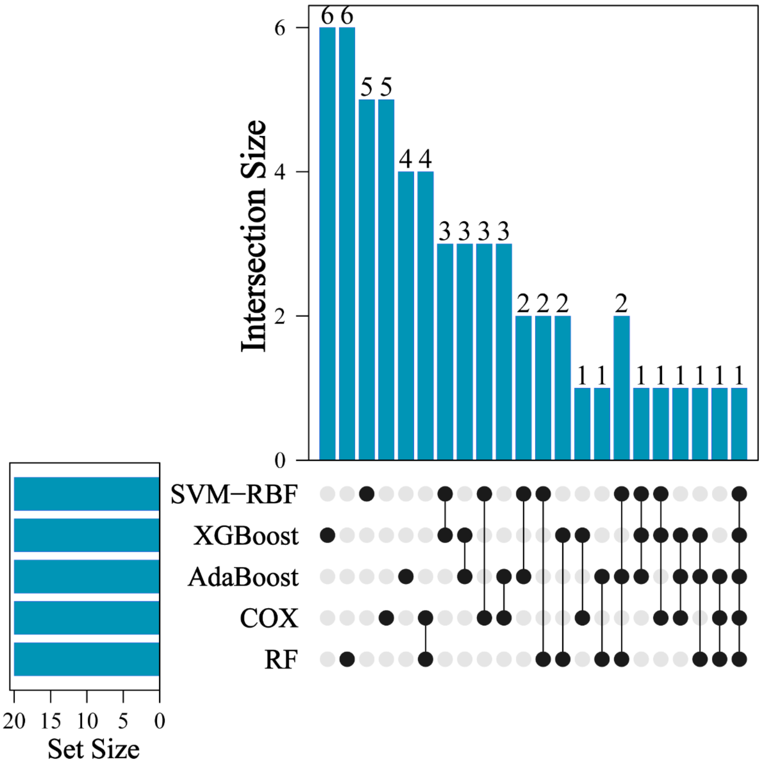

Fig. 2. The Venn diagram that utilizes five machine learning methods to screen key gene.

Model	Train MSE (relative)	Test MSE (relative)
Random forest	1.00	1.13
XGBoost	1.04	1.04
AdaBoost	1.09	1.00
SVM-RBF	1.18	1.16

Table 1. The relative MSE results for each model.

interpretability. Each model independently identified the top 20 most influential feature genes from statistically significant univariate variables. Utilizing a Venn diagram, we analyzed the overlap among the genes selected by the different algorithms. Ultimately, we identified PFKFB3 as a key gene affecting gastric cancer prognosis, consistently selected by all five algorithms, highlighting its critical role as an overlapping gene influencing patient outcomes. To address potential variability from train-test splits, we performed repeated fivefold cross-validation (3 repeats) and observed that PFKFB3 was consistently selected, confirming its robustness.

Prognostic value of PFKFB3 in gastric cancer

Research findings indicate that elevated expression of PFKFB3 in gastric cancer tissues is often associated with poorer Overall Survival (OS) (Fig. 3A). We extended our analysis using the Cox regression model on four publicly accessible extensive datasets of gastric cancer patients (Fig. 3B, C). The outcomes underscored an association between PFKFB3 expression levels and the survival duration of these patients, proposing that PFKFB3 may act as a potential biomarker for individuals with gastric cancer.

Furthermore, we discovered that PFKFB3 also exhibited robust prognostic predictive capabilities within subgroups categorized by a spectrum of clinical features (Fig. 3D–I).

An integrative analysis of the correlation between PFKFB3 and clinical characteristics within the TCGA-STAD cohort revealed elevated PFKFB3 expression in patient groups with diverse clinical and pathological profiles (Fig. 4A–H). Both univariate and multivariate Cox regression analyses established PFKFB3 as an independent prognostic indicator for the Overall Survival (OS) of gastric cancer patients (Fig. 4I, J). Figure 4J displays the results of multivariate Cox regression analysis, adjusting for potential confounding factors. PFKFB3 expression remained an independent prognostic indicator for OS (hazard ratio = 1.423, p = 0.05). In light of these results, we developed a nomogram model (Fig. 4K). This model exhibited superior predictive capabilities, and its precision was substantiated by the calibration curve (Fig. 4L).

Biological functions and activities of PFKFB3

In a clinical trial, we further analyzed the relationship between PFKFB3 and the response to immunotherapy. Within the IMvigor210 cohort 2018 (anti-PD-L1), patients exhibiting a heightened response to immunotherapy demonstrated reduced expression of PFKFB3 (Fig. 5A). Subsequent survival analysis indicated that patients with diminished PFKFB3 expression who received anti-PD-L1 treatment had an extended Overall Survival (OS) (Fig. 5B). Moreover, a Receiver Operating Characteristic (ROC) curve analysis predicated on PFKFB3 expression was performed to evaluate its potential as a predictor of the immunotherapy response, yielding an area under the curve (AUC) of 0.63 within the IMvigor210 cohort 2018 (anti-PD-L1) (Fig. 5C). Employing the ssGSEA and ESTIMATE algorithms, we determined the levels of immune cell infiltration and immune scores among gastric cancer patients, subsequently categorizing them into two groups based on PFKFB3 expression levels (Fig. 5D, F). The findings suggest that patients with elevated PFKFB3 expression levels have markedly increased immune infiltration compared to those with reduced expression. Correlation analyses further revealed a positive correlation between PFKFB3 and both immune cell presence and immune scoring (Fig. 5E, G, H, I).

Immunotherapy and related analysis

In order to delve into the potential mechanisms underlying the influence of PFKFB3 on gastric cancer (GC), the TCGA-STAD dataset was employed to dichotomize GC patients into high and low expression cohorts based on the median expression levels of PFKFB3 (Fig. 6A). Following this stratification, differential expression analysis was executed and subsequently conducted Gene Ontology (GO) and KEGG analyses on the differentially expressed genes (Fig. 6B–D).

Results implicated PFKFB3 in processes and diseases including insulin secretion, growth factor activity, chylomicron metabolism, digestion and absorption of proteins and fats, and diabetes. Furthermore, Gene Set Enrichment Analysis (GSEA) elucidated the biological processes modulated by PFKFB3, indicating involvement in pathways such as glucose metabolism, glycolysis, insulin signaling, and apoptosis. Gene set variant analysis confirmed these findings, emphasizing a link between PFKFB3 and processes such as metastasis, apoptosis, inflammation, and hypoxia (Fig. 6E).

Expression levels of PFKFB3 in pan-cancer and prognostic analysis

To illustrate the association between PFKFB3 expression levels and carcinogenesis as well as tumor progression, we assessed the differential expression of PFKFB3 in a pan-cancer context (Fig. 7A, B). The analysis of the consolidated TCGA_GTEx data revealed upregulation of PFKFB3 across a spectrum of cancers, with particularly significant increases in cholangiocarcinoma (CHOL), head and neck squamous cell carcinoma (HNSC), thyroid carcinoma (THCA), stomach adenocarcinoma (STAD), and esophageal carcinoma (ESCA). Examination of PFKFB3 expression in tumor tissues and their corresponding normal pairs demonstrated a high degree of

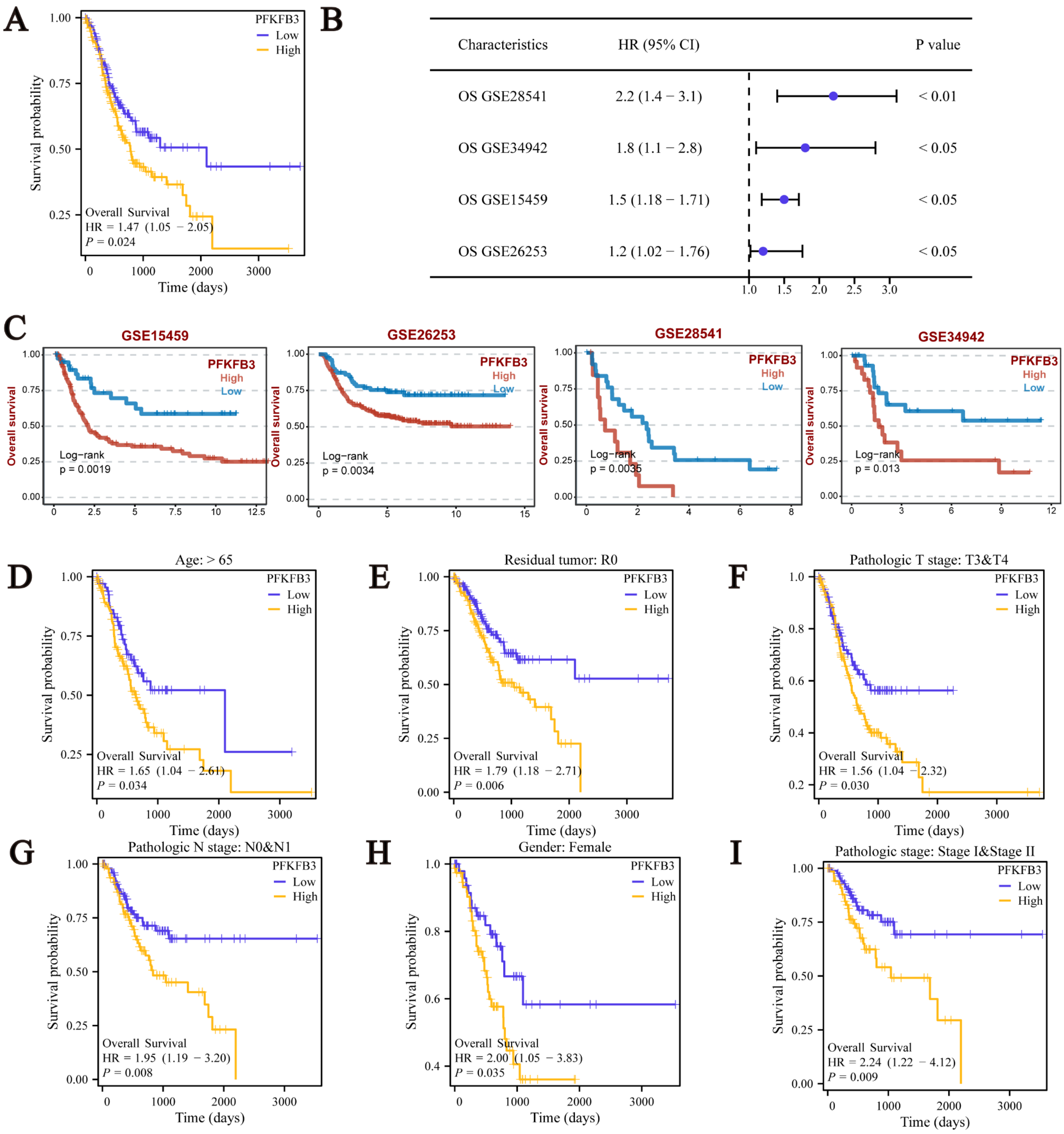

Fig. 3. Prognostic Value of PFKFB3 in Gastric Cancer. (**A**) Kaplan–Meier (KM) curves for high and low expression groups of PFKFB3. (**B**, **C**) The prognostic value of PFKFB3 in four external gastric cancer datasets. (**D**–**I**) The prognostic value of PFKFB3 in multiple subgroups that grouped according to clinical features.

concordance. We also identified a linkage between PFKFB3 expression and pan-cancer prognosis (Fig. 7C). Employing diverse analytical approaches to evaluate the influence of PFKFB3 on overall survival, disease-specific survival, disease-free interval, and progression-free interval, we determined that elevated PFKFB3 expression is associated with an adverse prognosis in adenoid cystic carcinoma (ACC), kidney renal papillary cell carcinoma (KIRP), stomach adenocarcinoma (STAD), and liver hepatocellular carcinoma (LIHC).

Conversely, in kidney renal clear cell carcinoma (KIRC), elevated PFKFB3 expression exerts a protective effect. Forest plots provided a clearer depiction of the specific effects of PFKFB3 expression on the overall survival of various tumor types (Fig. 7D). Kaplan–Meier survival analysis further substantiated the prognostic value of PFKFB3 as a biomarker for ACC, KIRP, KIRC, and LIHC (Fig. 7E–H).

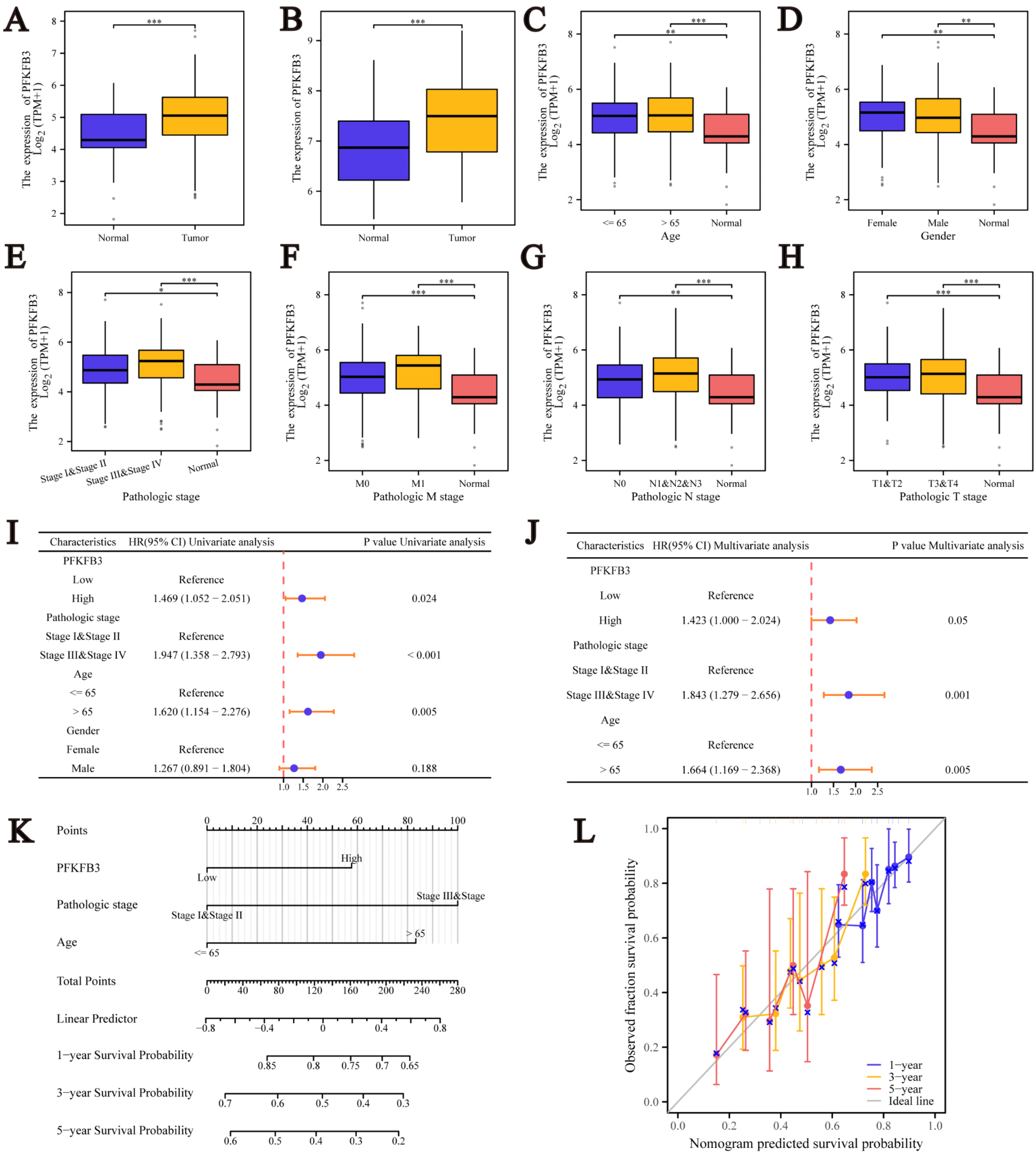

Fig. 4. (**A**) Within the TCGA-STAD dataset, an upregulation of PFKFB3 expression is observed in cancer patients. (**B**) Validation of elevated PFKFB3 expression in tumor tissues is provided by external datasets. (**C–H**) Across patient populations characterized by diverse clinical and pathological features, an increase in PFKFB3 expression is consistently noted. (**I, J**) The results of the univariate and multivariate Cox regression analyses are shown separately using forest plots. Risk ratios and 95% confidence intervals for PFKFB3 expression versus other clinical characteristics are shown. (**K**) A nomogram model constructed based on PFKFB3 expression levels. (**L**) The calibration curve is depicted, illustrating the model's predictive accuracy.

Discussion

The identification and investigation of gastric cancer biomarkers are crucial for the precise prognostication of patients with gastric cancer and for the realization of personalized therapeutic strategies. In this study, we harnessed an ensemble of five machine learning algorithms—Support Vector Machine with Recursive Feature

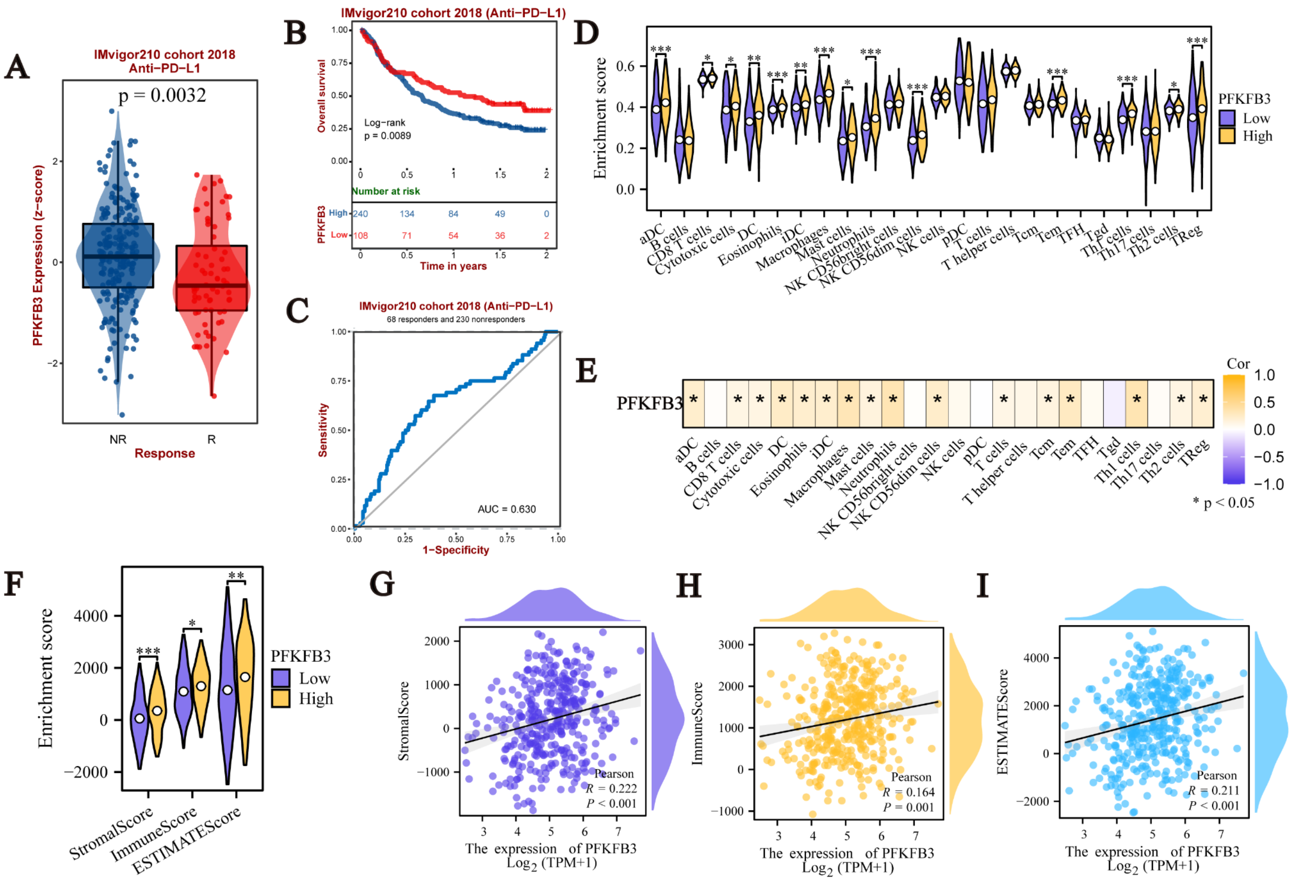

Fig. 5. (**A**) Patients exhibiting high PFKFB3 expression show a diminished response rate to immunotherapy. (**B**) Individuals with reduced PFKFB3 expression demonstrate an extended survival period when treated with this immunotherapy protocol. (**C**) The ROC curve is constructed based on the expression levels of PFKFB3 among patients in the immunotherapy cohort. (**D**, **F**) Differences in immune cell invasion levels and immune scores of gastric cancer patients in the PFKFB3 and high and low groups of gastric cancer patients (based on ssGSEA algorithm and estimate algorithm, respectively) (**E**) A correlation heatmap of PFKFB3 in relation to immune cells. (**G**–**I**) Scatter plots representing the correlation between PFKFB3 expression and three distinct immune scores.

Elimination (SVM-RFE), Random Forest (RF), XGBoost, Cox Proportional Hazards Model, and AdaBoost—to analyze gastric cancer datasets from The Cancer Genome Atlas (TCGA) and a glycolysis gene database. Through this integrative approach, we have successfully identified PFKFB3 as a significant prognostic factor.

PFKFB3 is an enzymatic catalyst for the synthesis of fructose-2,6-bisphosphate (F-2,6-BP), a key molecule in glycolysis that is ubiquitously present across various cellular contexts[52–55]. An increasing body of research indicates elevated expression of PFKFB3 in numerous neoplastic cells, including those of colorectal cancer[56–58]. Despite this, the specific role of PFKFB3 in the processes of tumor invasion and metastasis remains largely elusive, with few reports detailing the underlying mechanisms. Consequently, our differential analysis between groups with disparate levels of PFKFB3 expression sheds light on the enzyme's potential oncogenic mechanisms, paving the way for more profound investigations into its role. Within the IMvigor210 cohort 2018 (anti-PD-L1), there exists a significant correlation between PFKFB3 expression levels and patient responses to immunotherapy, as well as overall prognosis. Further analysis of the immune scoring and immune cell infiltration between the high and low PFKFB3 expression groups revealed a higher degree of immune infiltration in the group with elevated PFKFB3 expression. These findings intimate a regulatory role for PFKFB3 in tumor immune cell infiltration. Pan-cancer analysis of overall survival, disease-specific survival, disease-free interval, and progression-free interval has demonstrated that elevated PFKFB3 expression is indicative of an unfavorable prognosis in liver hepatocellular carcinoma (LIHC), stomach adenocarcinoma (STAD), adenoid cystic carcinoma (ACC), and kidney renal papillary cell carcinoma (KIRP).

This study suggests that PFKFB3 may be a prognostic biomarker for gastric cancer, a finding that has important implications for the diagnosis and treatment of gastric cancer. As a prognostic biomarker, we demonstrated through multiple datasets that PFKFB3 expression is upregulated in gastric cancer and predicts the prognosis of gastric cancer patients. In this study, we also combined the expression level of PFKFB3 with clinicopathologic features, including pathologic grade and age, to construct a more comprehensive prognostic assessment system. However, cancer development and progression usually involve the aberrant expression of multiple genes, so the

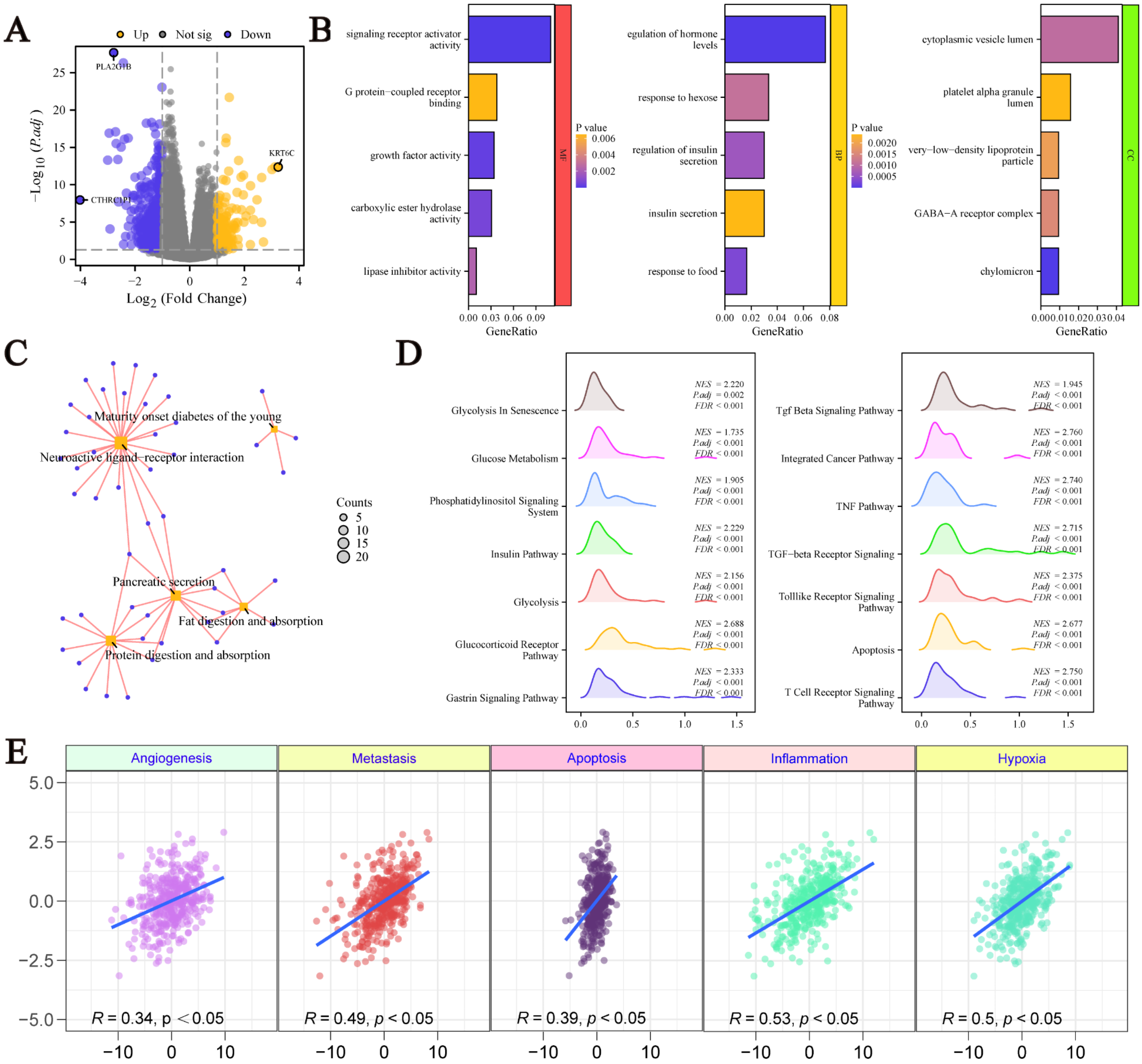

Fig. 6. (**A**) The volcano plot illustrates the differential analysis outcomes between the high and low expression groups of PFKFB3. (**B**) Gene Ontology (GO) analysis of the differentially expressed genes is presented. Counts represent the number of differentially expressed genes in each specific GO term or KEGG pathway. (**C**) KEGG pathway analysis of the differentially expressed genes is depicted. (**D**) Gene Set Enrichment Analysis (GSEA) focused on PFKFB3 is shown. (**E**) Pearson correlation between PFKFB3 and GSVA score.

predictive ability of a single biomarker may be limited[59,60]. In future clinical applications, combining PFKFB3 with other gastric cancer-related biomarkers (e.g., CEA, CA19-9, etc.) may be able to improve the accuracy and reliability of prediction[61,62]. In addition, to ensure the reliability of PFKFB3 as a prognostic biomarker, it needs to be rigorously validated and evaluated, including reproducibility experiments in different patient populations, and combination with other biomarkers[63,64].

Further advancing our investigation, we employed a variety of methods, including differential expression analysis and Gene Set Enrichment Analysis (GSEA), to analyze the selected PFKFB3 and validate its prognostic value across multiple datasets. By investigating the role of PFKFB3 in pan-cancer, we demonstrated that PFKFB3 is a potentially valuable prognostic biomarker for adrenocortical carcinoma, kidney renal papillary cell carcinoma, kidney renal clear cell carcinoma and liver hepatocellular carcinoma. Future studies will require larger sample sizes and more refined experimental designs to further validate and deepen these findings.

In this study, four machine learning methods with comparable performance on the regression task and Cox proportional hazards model were used for gene selection. However, integrating the genes identified by different methods may face comparability challenges. Specifically, differences in algorithmic structures and performance on the regression task can impact the consistency of selected genes. Although the chosen machine learning methods are widely recognized in prior studies, the potential influence of these comparability issues still warrants further exploration.

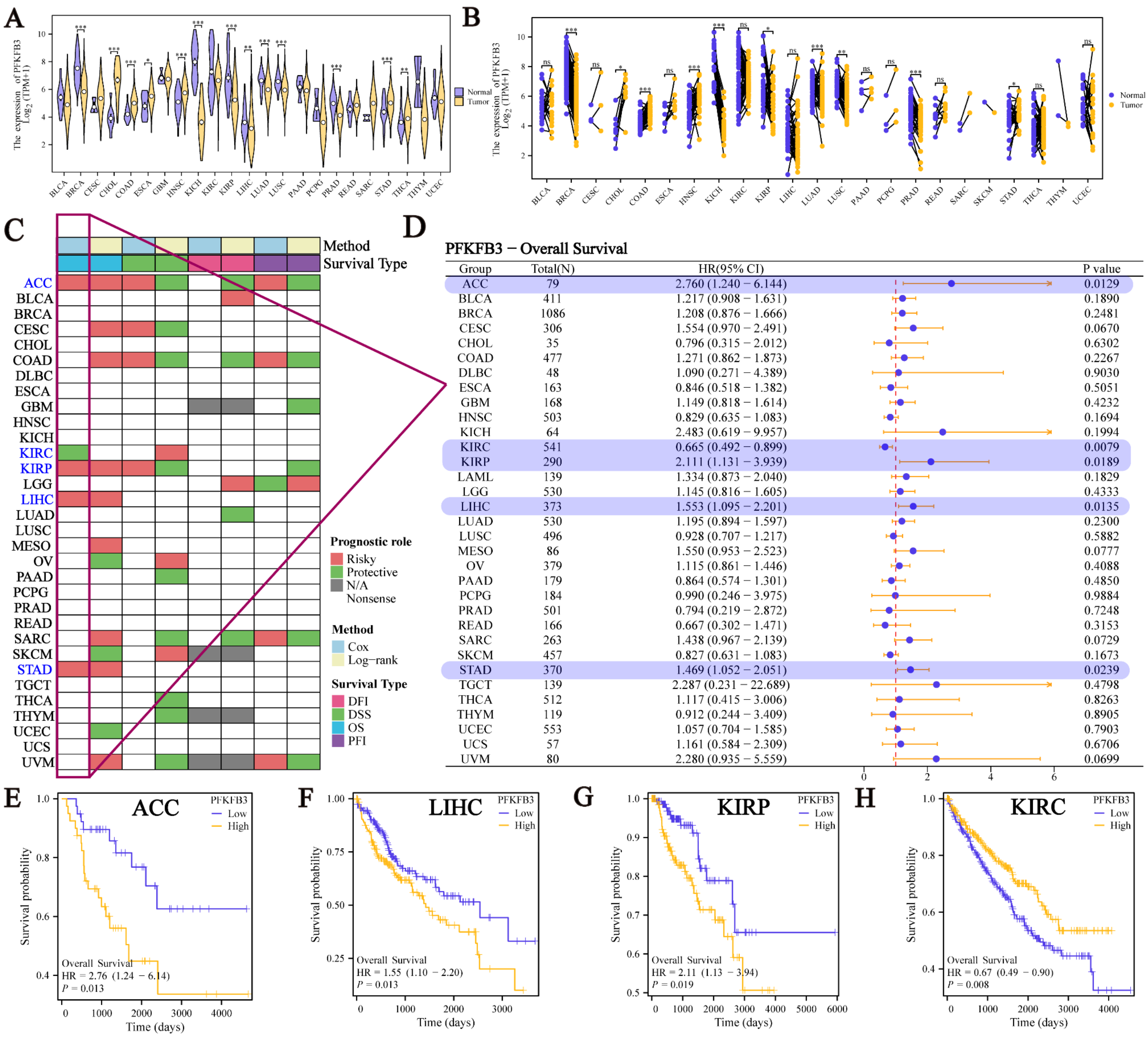

Fig. 7. Prognostic Value of PFKFB3 in Pan-Cancer. (**A**) Analysis of PFKFB3 expression in pan-cancer using paired tumor/normal samples from the TCGA and GTEx databases. (**B**) Expression of PFKFB3 in paired samples from 18 tumors within the TCGA database. (**C**) Correlation between PFKFB3 expression and Overall Survival (OS), Disease-Specific Survival (DSS), Disease-Free Interval (DFI), and Progression-Free Interval (PFI). Only p-values < 0.05are displayed. (**D**) The relationship between PFKFB3 and Overall Survival (OS). (**E**–**H**) Prognostic value of PFKFB3 in ACC, LIHC, KIRP, and KIRC.

Conclusions

In conclusion, the present study employed five machine learning methodologies to pinpoint the glycolysis-associated gene PFKFB3 as a prognostic biomarker for gastric cancer, with validation performed on several external datasets. Furthermore, we explored the function of PFKFB3 in immune and biological processes. Future research could involve extensive clinical cohort studies to further investigate the role of PFKFB3 as a gastric cancer biomarker and to assess its potential as a target for therapeutic intervention.

Data availability

All data generated or analyzed during this study are included in this published article. The data used to support the findings of the present study are available from the corresponding author upon request.

Code availability

Code for this study is publicly available on GitHub: https://github.com/PiPiNam/ML-GCP.

Received: 6 August 2024; Accepted: 7 March 2025

Published online: 13 March 2025

References

1. Wang, C., Zhang, Y., Zhang, Y. & Li, B. A bibliometric analysis of gastric cancer liver metastases: Advances in mechanisms of occurrence and treatment options. *Int. J. Surg.* **110**, 2288–2299 (2024).
2. Thrift, A. P. & El-Serag, H. B. Burden of gastric cancer. *Clin. Gastroenterol. Hepatol.* **18**, 534–542 (2020).
3. Siegel, R. L., Miller, K. D., Fuchs, H. E. & Jemal, A. Cancer Statistics, 2021. *CA Cancer J. Clin.* **71**, 7–33 (2021).
4. López, M. J. et al. Characteristics of gastric cancer around the world. *Crit. Rev. Oncol. Hematol.* **181**, 103841 (2023).
5. Joshi, S. S. & Badgwell, B. D. Current treatment and recent progress in gastric cancer. *CA Cancer J. Clin.* **71**, 264–279 (2021).
6. Oliveira, C., Pinheiro, H., Figueiredo, J., Seruca, R. & Carneiro, F. Familial gastric cancer: Genetic susceptibility, pathology, and implications for management. *Lancet Oncol.* **16**, e60-70 (2015).
7. Catalano, V. et al. Gastric cancer. *Crit. Rev. Oncol. Hematol.* **54**, 209–241 (2005).
8. Hohenberger, P. & Gretschel, S. Gastric cancer. *Lancet* **362**, 305–315 (2003).
9. Smyth, E. C., Nilsson, M., Grabsch, H. I., van Grieken, N. C. & Lordick, F. Gastric cancer. *Lancet* **396**, 635–648 (2020).
10. Venerito, M., Link, A., Rokkas, T. & Malfertheiner, P. Gastric cancer—clinical and epidemiological aspects. *Helicobacter* **21**(Suppl 1), 39–44 (2016).
11. Thrift, A. P., Wenker, T. N. & El-Serag, H. B. Global burden of gastric cancer: Epidemiological trends, risk factors, screening and prevention. *Nat. Rev. Clin. Oncol.* **20**, 338–349 (2023).
12. Puliga, E., Corso, S., Pietrantonio, F. & Giordano, S. Microsatellite instability in gastric cancer: Between lights and shadows. *Cancer Treat. Rev.* **95**, 102175 (2021).
13. Chia, N.-Y. & Tan, P. Molecular classification of gastric cancer. *Ann. Oncol.* **27**, 763–769 (2016).
14. Pietrobon, V. Cancer metabolism. *J. Transl. Med.* **19**, 87 (2021).
15. Zhao, L. et al. Impacts and mechanisms of metabolic reprogramming of tumor microenvironment for immunotherapy in gastric cancer. *Cell Death Dis.* **13**, 378 (2022).
16. Xu, X. et al. Metabolic reprogramming and epigenetic modifications in cancer: From the impacts and mechanisms to the treatment potential. *Exp. Mol. Med.* **55**, 1357–1370 (2023).
17. Tan, Y. et al. Metabolic reprogramming from glycolysis to fatty acid uptake and beta-oxidation in platinum-resistant cancer cells. *Nat. Commun.* **13**, 4554 (2022).
18. Sun, C. et al. Spatially resolved multi-omics highlights cell-specific metabolic remodeling and interactions in gastric cancer. *Nat. Commun.* **14**, 2692 (2023).
19. DeBerardinis, R. J. & Chandel, N. S. We need to talk about the Warburg effect. *Nat. Metab.* **2**, 127–129 (2020).
20. Kadam, W., Wei, B. & Li, F. Metabolomics of gastric cancer. *Adv. Exp. Med. Biol.* **1280**, 291–301 (2021).
21. Liberti, M. V. & Locasale, J. W. The Warburg effect: How does it benefit cancer cells?. *Trends Biochem. Sci.* **41**, 211–218 (2016).
22. Zhang, L. et al. Clk1-regulated aerobic glycolysis is involved in glioma chemoresistance. *J. Neurochem.* **142**, 574–588 (2017).
23. Vallée, A., Lecarpentier, Y., Guillevin, R. & Vallée, J.-N. The influence of circadian rhythms and aerobic glycolysis in autism spectrum disorder. *Transl. Psychiatry* **10**, 400 (2020).
24. Ma, S. et al. NPAS2 promotes aerobic glycolysis and tumor growth in prostate cancer through HIF-1A signaling. *BMC Cancer* **23**, 280 (2023).
25. Shi, X. et al. Building a translational cancer dependency map for the cancer genome atlas. *Nat. Cancer* **5**, 1176–1194 (2024).
26. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The cancer genome atlas (TCGA): An immeasurable source of knowledge. *Contemp. Oncol.* **19**, A68 (2015).
27. Weinstein, J. N. et al. The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
28. Handelman, G. S. et al. eDoctor: Machine learning and the future of medicine. *J. Intern. Med.* **284**, 603–619 (2018).
29. Greener, J. G., Kandathil, S. M., Moffat, L. & Jones, D. T. A guide to machine learning for biologists. *Nat. Rev. Mol. Cell. Biol.* **23**, 40–55 (2022).
30. Do, D. T. & Le, N. Q. K. Using extreme gradient boosting to identify origin of replication in *Saccharomyces cerevisiae* via hybrid features. *Genomics* **112**, 2445–2451 (2020).
31. Uddin, S., Khan, A., Hossain, M. E. & Moni, M. A. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med. Inform. Decis. Mak.* **19**, 281 (2019).
32. Lu, T.-P. et al. Developing a prognostic gene panel of epithelial ovarian cancer patients by a machine learning model. *Cancers* **11**, 270 (2019).
33. Yokoyama, S. et al. Predicted prognosis of patients with pancreatic cancer by machine learning. *Clin. Cancer Res.* **26**, 2411–2421 (2020).
34. Ma, B., Geng, Y., Meng, F., Yan, G. & Song, F. Identification of a sixteen-gene prognostic biomarker for lung adenocarcinoma using a machine learning method. *J. Cancer* **11**, 1288–1298 (2020).
35. Subhash, V. V. et al. Anti-tumor efficacy of Selinexor (KPT-330) in gastric cancer is dependent on nuclear accumulation of p53 tumor suppressor. *Sci. Rep.* **8**, 12248 (2018).
36. Oh, S. C. et al. Clinical and genomic landscape of gastric cancer with a mesenchymal phenotype. *Nat. Commun.* **9**, 1777 (2018).
37. Lee, J. et al. Nanostring-based multigene assay to predict recurrence for gastric cancer patients after surgery. *PLoS One* **9**, e90133 (2014).
38. Ny, C. et al. Regulatory crosstalk between lineage-survival oncogenes KLF5, GATA4 and GATA6 cooperatively promotes gastric cancer development. *Gut* **64** (2015).
39. Wang, G. et al. Comparison of global gene expression of gastric cardia and noncardia cancers from a high-risk population in china. *PloS ONE* **8** (2013).
40. Necchi, A. et al. Atezolizumab in platinum-treated locally advanced or metastatic urothelial carcinoma: Post-progression outcomes from the phase II IMvigor210 study. *Ann. Oncol.* **28** (2017).
41. Sheng, K. L. et al. An integrated approach to biomarker discovery reveals gene signatures highly predictive of cancer progression. *Sci. Rep.* **10**, 21246 (2020).
42. Li, W., Yin, Y., Quan, X. & Zhang, H. Gene expression value prediction based on XGBoost algorithm. *Front. Genet.* **10** (2019).
43. Pölsterl, S. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *J. Mach. Learn. Res.* **21**, 212:8747-212:8752 (2020).
44. Yang, H. et al. PESSA: A web tool for pathway enrichment score-based survival analysis in cancer. *PLOS Comput. Biol.* **20**, e1012024 (2024).
45. Mizuno, H., Kitada, K., Nakai, K. & Sarai, A. PrognoScan: A new database for meta-analysis of the prognostic value of genes. *BMC Med. Genomics* **2**, 18 (2009).
46. Ozhan, A., Tombaz, M. & Konu, O. SmulTCan: A Shiny application for multivariable survival analysis of TCGA data with gene sets. *Comput. Biol. Med.* **137**, 104793 (2021).
47. Liu, Z. et al. BEST: A web application for comprehensive biomarker exploration on large-scale data in solid tumors. *J. Big Data* **10**, 165 (2023).
48. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

49. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
50. Yoshihara, K. et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**, 2612 (2013).
51. Barbie, D. A. et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **462**, 108–112 (2009).
52. Xiao, M., Liu, D., Xu, Y., Mao, W. & Li, W. Role of PFKFB3-driven glycolysis in sepsis. *Ann. Med.* **55**, 1278–1289 (2023).
53. De Bock, K. et al. Role of PFKFB3-driven glycolysis in vessel sprouting. *Cell* **154**, 651–663 (2013).
54. Zeng, H. et al. Suppression of PFKFB3-driven glycolysis restrains endothelial-to-mesenchymal transition and fibrotic response. *Signal Transduct. Target Ther.* **7**, 303 (2022).
55. Thirusangu, P. et al. PFKFB3 regulates cancer stemness through the hippo pathway in small cell lung carcinoma. *Oncogene* **41**, 4003–4017 (2022).
56. Han, J. et al. Interleukin-6 stimulates aerobic glycolysis by regulating PFKFB3 at early stage of colorectal cancer. *Int. J. Oncol.* **48**, 215–224 (2016).
57. Yu, H. et al. Metabolic reprogramming and AMPKα1 pathway activation by caulerpin in colorectal cancer cells. *Int. J. Oncol.* **50**, 161–172 (2017).
58. Lei, L. et al. A potential oncogenic role for PFKFB3 overexpression in gastric cancer progression. *Clin. Transl. Gastroenterol.* **12**, e00377 (2021).
59. Iden, C. R. et al. Circulating tumor DNA predicts recurrence and survival in patients with resectable gastric and gastroesophageal junction cancer. *Gastric Cancer* https://doi.org/10.1007/s10120-024-01556-9 (2024).
60. An, X. et al. Combined influence of physical activity and C-reactive protein to albumin ratio on mortality among older cancer survivors in the United States: A prospective cohort study. *Eur. Rev. Aging Phys. Act.* **21**, 26 (2024).
61. van Hootegem, S. J. M., Pittacolo, M. & Lagarde, S. M. Extended follow-up in patients with gastric cancer-applicable to western patients?. *JAMA Surg.* https://doi.org/10.1001/jamasurg.2024.4279 (2024).
62. Leijonmarck, W., Mattsson, F. & Lagergren, J. Neoadjuvant chemotherapy in relation to long-term mortality in individuals cured of gastric adenocarcinoma. *Gastric Cancer* https://doi.org/10.1007/s10120-024-01558-7 (2024).
63. Zhang, Y. et al. Cell polarity-related gene PTK7, a Potential diagnostic biomarker in pan-cancer. *Curr. Med. Chem.* https://doi.org/10.2174/0109298673313999240816103054 (2024).
64. Kawakami, H. et al. Real-world effectiveness and safety of trastuzumab-deruxtecan in Japanese patients with HER2-positive advanced gastric cancer (EN-DEAVOR study). *Gastric Cancer* https://doi.org/10.1007/s10120-024-01555-w (2024).

Acknowledgements

The author thanks my colleagues and friends for their help, which enabled me to complete this article.

Author contributions

Conceptualization, N.L.; methodology, Y.Z., J.G; algorithms, N.L.; validation, Q.Z.; formal analysis, Y.Z.; investigation, H.J.; resources, M.H.; writing—original draft preparation, N.L.; writing—review and editing, Y.Z. and Q.Z. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by the Natural Science Foundation of China (Grant No. U21B2026) and National Natural Science Foundation of China (No. 82073244).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.