Big Data Mining and Analytics

Research Article

Machine learning prediction of hepatocellular carcinoma informed by 3D-genome selected circRNAs

Shanshan Yin[1,2,3,†], Wenqian Yan[1,2,†], Chongchen Pang[1,2,†], Liangjie Liu[1,2], Mofan Feng[1,2], Zhangli Yuan[1,2], Haoyan Chen[4], Xuanwen Bao[5,6], Weijia Fang[5,6], Xin Jin[7], Weiping Zhu[7], Mika Torhola[8], Klaus Förger[8], Henna Kujanen[8], Lintai Da[3], Hao Li[9], Jacek Jemielity[10], Yuqing Lou[11], Guang He[1,2,12,13], Joanna Kowalska[14], Yi Shi[1,2,12,13,✉]

Cite this article: Yin SS, Yan WQ, Pang CC, et al. *Big Data Min. Anal.* https://doi.org/10.26599/BDMA.2025.9020125

ABSTRACT: Hepatocellular Carcinoma (HCC) remains a major contributor to cancer mortality, with poor prognosis driven by limited options for early detection and intervention. Circular RNAs (circRNAs), a class of non-coding RNAs with tissue-specific expression and regulatory functions, have emerged as potential biomarkers and therapeutic targets. In this work, we investigate the predictive value of circRNAs in HCC using public datasets GSE155949 (49 HCC + 49 normal samples) and GSE121714 (20 HCC + 10 normal samples), and a suite of machine learning approaches, including logistic regression, support vector machines, Light Gradient Boosting Machine (LightGBM), and deep neural networks. To capture circRNA contributions from multiple perspectives, we construct six distinct feature panels, one of which incorporates their spatial distribution within the Three-Dimensional (3D) genome. Our results show that circRNAs selected according to 3D genome loci improves prediction accuracy and enhances generalization by mitigating feature collinearity in small-sample settings. These findings highlight the advantages of integrating 3D genome information into circRNA-based predictive modeling and offer a scalable framework for biomarker discovery and precision medicine in HCC.

KEYWORDS: Hepatocellular Carcinoma (HCC); Three-Dimensional (3D) genome; Circular RNAs (circRNA); biomarker; Machine Learning

1 Introduction

Hepatocellular Carcinoma (HCC) is a highly aggressive malignancy with a steadily increasing global incidence and is now recognized as one of the leading causes of cancer-related deaths worldwide[1, 2]. Despite progress in oncology, conventional diagnostic approaches, such as imaging and systemic therapies, remain insufficient to fully address the biological complexity of HCC[3]. The absence of reliable early biomarkers and limited treatment options for advanced disease stages continue to drive its poor prognosis and high mortality rates[4]. In this context, growing interest has focused on non-coding RNAs, particularly circular RNAs (circRNAs), to better understand their contributions to HCC pathogenesis. CircRNAs, defined by their covalently closed loop structures, represent a novel class of non-coding RNAs with tissue-specific expression patterns and roles in gene regulation, splicing, and cell signaling[5, 6]. Emerging evidence indicates that

† These authors contributed equally to this work.

[1] Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders, Shanghai Jiao Tong University, 1954 Huashan Road, Shanghai 200030, China. [2] Shanghai Key Laboratory of Psychotic Disorders and Brain Science and Technology Research Center, Shanghai Jiao Tong University, 1954 Huashan Road, Shanghai 200030, China. [3] Key Laboratory of Systems Biomedicine (Ministry of Education), Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University, Shanghai 200240, China. [4] State Key Laboratory of Systems Medicine for Cancer, NHC Key Laboratory of Digestive Diseases, Division of Gastroenterology and Hepatology, School of Medicine, Renji Hospital, Shanghai Jiao Tong University, Shanghai Institute of Digestive Disease, Shanghai Cancer Institute, Shanghai 200001, China. [5] Department of Medical Oncology, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou 310003, China. [6] National Key Laboratory of Advanced Drug Delivery and Release Systems, Zhejiang University, Hangzhou 310058, China. [7] Department of Hepatic Surgery, Fudan University Shanghai Cancer Center, Fudan University, Shanghai 200032, China. [8] Atostek Oy, Hermiankatu 3 A, Tampere 33720, Finland. [9] Department of Oncology, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China. [10] Centre of New Technologies, University of Warsaw, Warsaw 00-927, Poland. [11] Department of Respiratory Medicine, Shanghai Chest Hospital, Shanghai Jiao Tong University, Shanghai 200025, China. [12] NHC Key Laboratory of Medical Embryogenesis and Developmental Molecular Biology & Shanghai Key Laboratory of Embryo and Reproduction Engineering, Shanghai 200040, China. [13] Shanghai Institute of Medical Genetics, Shanghai Children's Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200040, China. [14] Division of Biophysics, Institute of Experimental Physics, Faculty of Physics, University of Warsaw, Warsaw 00-927, Poland.

✉ Corresponding author. E-mail: yishi@sjtu.edu.cn

Received: September 26, 2025; Revised: December 3, 2025; Accepted: December 22, 2025

© The Author(s) 2026. This is an open access article under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0, http://creativecommons.org/licenses/by/4.0/).

Just Accepted

circRNAs are not merely splicing byproducts but act as functional regulators in HCC, influencing tumor initiation[7], metastasis[8], and therapeutic resistance[9, 10]. Furthermore, their promise as biomarkers for early detection, prognosis, and therapeutic intervention in HCC is increasingly being investigated[11]. Collectively, this expanding body of work highlights circRNAs as compelling candidates for elucidating the molecular mechanisms of HCC and for advancing innovative diagnostic and treatment strategies[12].

In HCC, the aberrant expression of circRNAs and their multifaceted regulatory roles in gene expression and cellular signaling have attracted increasing attention, as accumulating evidence points to their active involvement in tumorigenesis[13]. One of the most extensively studied mechanisms is their capacity to act as "miRNA sponges", whereby circRNAs sequester specific microRNAs (miRNAs) and prevent their interaction with target messenger RNAs (mRNAs). By modulating intracellular miRNA availability, circRNAs exert downstream effects on post-transcriptional gene regulation and can either enhance or suppress tumor-associated pathways depending on the molecular context[14]. Beyond their role in miRNA regulation, circRNAs also form dynamic interactions with RNA-Binding Proteins (RBPs), thereby modulating diverse aspects of RNA metabolism, including transcription, splicing, stability, and subcellular localization of mRNAs[15]. Through these regulatory functions, circRNAs participate in the fine-tuning of key cellular processes such as proliferation, apoptosis, invasion, and metastasis, all of which are essential hallmarks of HCC progression[16]. Moreover, emerging studies have revealed that circRNAs influence major oncogenic and tumor-suppressive signaling cascades, including the Wnt/β-catenin, p53, and PI3K/Akt pathways, further underscoring their central role in liver cancer biology[17, 18]. Taken together, these findings portray circRNAs not as passive byproducts of splicing but as versatile regulators with the ability to reshape gene regulatory networks in cancer. Their diverse functions highlight their promise as early diagnostic biomarkers, prognostic indicators, and therapeutic targets in HCC, providing new opportunities for precision medicine and the development of innovative treatment strategies[19, 20].

Recent advances in bioinformatics and molecular biology have illuminated the possibility that circRNAs are not isolated regulators but rather integral components of complex intracellular regulatory networks. Increasing evidence suggests that the Three-Dimensional (3D) organization of the genome exerts profound effects on chromatin accessibility, transcriptional dynamics, and overall gene expression programs[21]. Studies on the 3D genome have highlighted its essential role in coordinating interactions among promoters, enhancers, and other regulatory elements, thereby shaping cellular identity and function. Within this spatial framework, circRNAs are emerging as molecules whose activity may be tightly linked to their genomic loci and structural positioning. Their specific localization within the 3D genome architecture could influence not only transcriptional regulation but also long-range chromatin interactions, providing an additional layer of control over gene regulatory networks. This perspective opens new avenues for understanding how circRNAs contribute to HCC pathogenesis, as it integrates both molecular function and nuclear spatial organization[22].

Feature selection in HCC prediction plays a pivotal role in enhancing machine learning model performance by prioritizing the most biologically relevant and informative biomarkers—in this case, circRNAs—while simultaneously reducing noise and mitigating overfitting. Such targeted selection not only improves diagnostic and prognostic accuracy but also yields valuable insights into the molecular mechanisms that drive HCC progression. However, the challenge of limited clinical retrospective datasets, often constrained by the high cost and complexity of generating large-scale omics data such as circRNA profiling, introduces potential pitfalls. In small sample settings, features that are linearly dependent may appear orthogonal, leading to the inadvertent selection of collinear biomarkers that undermine model generalization capacity when exposed to independent datasets. This limitation highlights a critical barrier in developing reliable circRNA-based predictive models for HCC. To address this issue, we explored whether diversifying circRNA selection according to their spatial distribution within the 3D genome could enhance generalization ability by reducing redundancy and capturing biologically complementary signals. Our findings demonstrated that circRNAs positioned in distinct 3D genomic loci significantly improved prediction performance, not only during cross-validation but also when evaluated on external testing sets, suggesting that spatial genome architecture confers added predictive value.

In this study, we utilize two publicly available circRNA expression datasets, GSE155949 and GSE121714, from the GEO database[23, 24]. Multiple machine learning algorithms—Logistic Regression (LR), Support Vector Machine (SVM)[25], Light Gradient Boosting Machine (LightGBM)[26], and Deep Neural Network (DNN)[27]—are applied to construct and evaluate a series of circRNA-based feature panels. Six complementary feature panels are developed to capture circRNA contributions from multiple biological and computational perspectives: (i) all circRNAs, (ii) differentially expressed circRNAs, (iii) pathway-enriched circRNAs, (iv) 3D genome clustering–diversified circRNAs, (v) circRNAs with the highest Spearman correlation values, and (vi) a concise circRNA panel optimized for parsimony. This multi-dimensional design provides a more comprehensive lens to dissect the multilayered regulatory mechanisms of circRNAs in HCC. Among these, the incorporation of 3D genome-derived circRNA features proves particularly significant, demonstrating robust predictive capacity across multiple models and outperforming other feature selection strategies in terms of accuracy and generalizability. By conducting extensive comparative analyses, this work not only underscores the importance of integrating spatial genome information into feature engineering but also provides a methodological framework for overcoming the inherent feature collinearity problem associated with limited training data. Ultimately, our findings contribute to the development of more precise circRNA biomarker panels for early HCC prediction and open avenues for translational applications in biomarker-driven diagnostics and personalized therapeutic strategies.

2 Method

2.1 circRNA dataset curation

After exploring all the HCC circRNA expression datasets available from Gene Expression Omnibus (GEO), as shown in Table S1, which is in the Electronic Supplementary Material (ESM) of the online version of this article. We utilizes two circRNA expression datasets, namely GSE155949 and GSE121714. The GSE155949 dataset consists of 98 samples, evenly divided into 49 HCC and 49 normal cases. The GSE121714 dataset, in contrast, comprises 30 samples, including 20 HCC and 10 normal. These datasets are chosen because of their direct relevance to HCC research and their complementary sample diversity, together providing a robust foundation for predictive modeling and biomarker analysis.

Just Accepted

To build reliable models, we combine the datasets and randomly partition the samples into training and test cohorts. Two-thirds of the data—46 HCC and 39 normal samples, totaling 85 cases—are designated as the training set for model development and hyperparameter optimization. The remaining one-third—23 HCC and 20 normal samples, totaling 43 cases—form the independent test set, reserved strictly for performance evaluation and generalization assessment. Within the training cohort, we implement 100 iterations of 10-fold cross-validation under four established machine learning algorithms, ensuring rigorous and reproducible training. This stratified sampling strategy maintains balanced representation of HCC and normal cases across both sets, thereby minimizing bias and strengthening the reliability of downstream evaluation.

During preprocessing, we apply z-score normalization to all samples within the dataset, to standardize feature distributions. This procedure transforms data such that each feature has a mean of zero and a standard deviation of one, effectively correcting for variability in measurement ranges or platform-specific differences, as shown in in the ESM. By harmonizing the datasets, z-score normalization eliminates inconsistencies, facilitates cross-dataset integration, and provides a stable foundation for model learning. Importantly, this step reduces the risk of overfitting or underfitting and enhances the robustness of predictions across independent validation.

Together, these carefully designed preprocessing and data partitioning strategies establish a reproducible and unbiased analytical pipeline. By ensuring dataset comparability, model stability, and evaluation fairness, our framework provides a solid basis for investigating the molecular determinants of HCC and for identifying circRNA features with predictive and translational potential.

2.2 Feature panel construction

2.2.1 All circRNAs (Panel-All)

The training and testing cohorts each contains expression profiles for 10 592 circRNAs, collectively designated as Panel-All, which serves as the comprehensive baseline for evaluating predictive performance. This panel encompasses all available circRNAs without any prior filtering or feature selection, thereby providing an unbiased reference point for subsequent comparisons. To systematically assess predictive power, we apply four machine learning algorithms—LR, SVM, LightGBM, and DNN—across the full dataset. Figure 1(a) presents the hierarchical clustering analysis of the 10 592 circRNAs across the 85 training samples (two-thirds of the combined GSE155949 and GSE121714 datasets). Clustering is performed using the Spearman correlation coefficient, which reveals multiple distinct circRNA clusters exhibiting strong collinearity. The presence of these highly correlated groups indicates that a substantial portion of circRNAs encode redundant information, thereby contributing little to the discriminatory power needed to distinguish HCC from normal samples. The Principal Component Analysis (PCA) analysis shown in Fig. 1(b) also reveals that the principal components alone cannot discriminate HCC from control, indicating circRNA panel selection is necessary for building accurate HCC prediction models. Complementing this, Figs. 1(c) and 1(d) depict the distribution of correlation coefficients between circRNA expression levels and the outcome labels (HCC vs. normal). The analysis demonstrates that the majority of circRNAs exhibit weak or negligible correlation ($r < 0.5$), while only a small subset achieves moderate correlation values ($0.5 < r < 0.7$). This pattern highlights a key challenge in circRNA-based prediction: Most circRNAs provide limited individual discriminative power, and reliance on the full dataset risks diluting meaningful signal with noise.

Collectively, these results underscore the critical importance of employing advanced feature selection strategies to refine circRNA panels. Without targeted selection, the redundancy and low predictive strength of the majority of circRNAs constrain the performance of models trained on Panel-All, limiting their capacity to achieve robust and generalizable predictions in the context of HCC.

2.2.2 Significant differentially expressed circRNAs (Panel-Sgnf)

To identify significant differentially expressed circRNAs, we conduct a Volcano plot analysis to systematically evaluate changes in expression levels between HCC and normal samples. This approach allows us to visualize and quantify both the magnitude of expression changes and their statistical significance, facilitating the selection of biologically meaningful circRNAs. We apply thresholds of $|\log_2(\text{fold change})| \geqslant 0.15$, which corresponds to at least a slight fold change in expression, and p-value <0.05. This cutoff setting is to avoid prematurely excluding circRNAs that, while showing modest overall changes, may hold discriminatory value in specific subgroups. These criteria enable the identification of both upregulated and downregulated circRNAs, as highlighted in Fig. 2(a), where the Volcano plot visually distinguishes these categories. From this analysis, we curate a panel of 2369 circRNAs, referred to as Panel-Sgnf, encompassing a balanced representation of circRNAs with potential roles in HCC pathogenesis. The inclusion of both upregulated and downregulated circRNAs ensures a comprehensive perspective on the molecular changes associated with HCC. This curated panel serves as a foundational dataset for further downstream analyses, such as pathway enrichment and machine learning predictions, aimed at uncovering the biological mechanisms and predictive potential of circRNAs in HCC. By narrowing the dataset to significant candidates, Panel-Sgnf reduces noise and redundancy, enhancing the reliability and interpretability of subsequent analyses.

2.2.3 circRNAs from enriched pathway (Panel-Pathway)

To further refine circRNA selection and prioritize candidates with stronger biological relevance to HCC, we conduct Gene Ontology (GO) enrichment analysis on the 2369 circRNAs comprising Panel-Sgnf. Enrichment testing is performed across the three major ontology categories—Biological Process (BP), Cellular Component (CC), and Molecular Function (MF)—resulting in the identification of 4917 enriched pathways, as the most enriched pathways shown in Figs. 2(b)–2(d). Applying threshold of adjusted p-value < 0.1, we narrow this set to 299 significantly enriched pathways. From each of these pathways, we select the top five circRNAs ranked by their correlation coefficients with the outcome label (HCC vs. normal). This systematic filtering yields a panel of 1349 circRNAs, designated as Panel-Pathway. By integrating both enrichment-based functional evidence and correlation-driven feature prioritization, this panel is designed to capture circRNAs more directly implicated in HCC biology.

Specifically, the Gene Set Enrichment Analysis (GSEA) for the differentially expressed circRNAs is performed by first mapping the circRNAs to their host genes (GeneSymbol) using the circBase database. CircRNAs without corresponding GeneSymbol entries in circBase are excluded. The resulting list of host genes is then used for enrichment analysis. We employ the Enrichr API via the gseapy Python package to conduct GO enrichment analysis,

Just Accepted

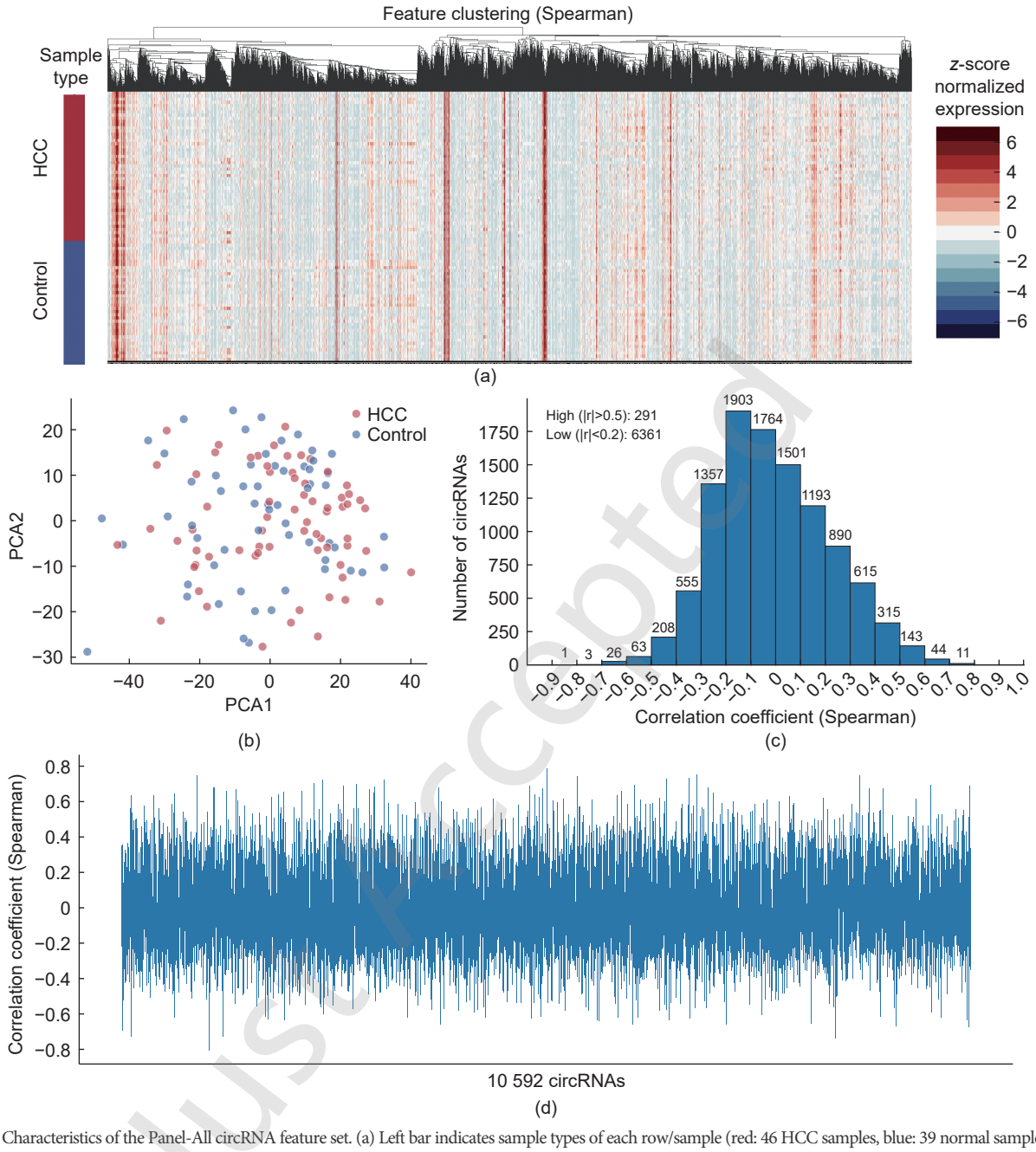


Fig. 1 Characteristics of the Panel-All circRNA feature set. (a) Left bar indicates sample types of each row/sample (red: 46 HCC samples, blue: 39 normal samples). The column-wise hierarchically clustered heatmap indicates the clustering results of the 10 592 circRNAs under Spearman correlation coefficient. (b) PCA analysis of the HCC and control samples. (c) Histogram distribution of correlation coefficients between 10 592 circRNAs and the outcome. (d) Detailed Spearman correlation coefficients between each of the circRNAs and the outcome.

including biological process (GO_BP), cellular component (GO_CC), and molecular function (GO_MF) categories. The analysis is performed using the default statistical model implemented in Enrichr, and pathways with an adjusted p-value <0.05 are considered significant. This approach leverages host gene information to infer potential functional roles of circRNAs.

The construction of Panel-Pathway not only provides a feature set with enhanced functional interpretability but also highlights circRNAs that are likely involved in key tumor-associated processes. This biologically informed panel offers a valuable resource for investigating the molecular mechanisms underlying HCC and contributes to clarifying the diverse roles of circRNAs in liver cancer initiation and progression.

2.2.4 3D-Genome-selected circRNA Panel (Panel-3DG)

We collect and processed high-throughput chromatin conformation capture (Hi-C) data from the IMR90 cell line to construct a 3D genome model covering all 23 chromosomes. The Hi-C data is processed according to the protocol from TopDom[28], and the genome is segmented into bins of 500 kilobases, which serves as the fundamental units of the model[29]. Each circRNA is mapped to the bin containing its genomic starting position, and the corresponding spatial coordinates ($< x, y, z >$) of the bin are assigned to that circRNA as its 3D genome location.

To explore spatial organization, we apply k-means clustering to the circRNAs associated with each enriched pathway, using their 3D coordinates as input features. The quality of clustering is

Under Review Accepted

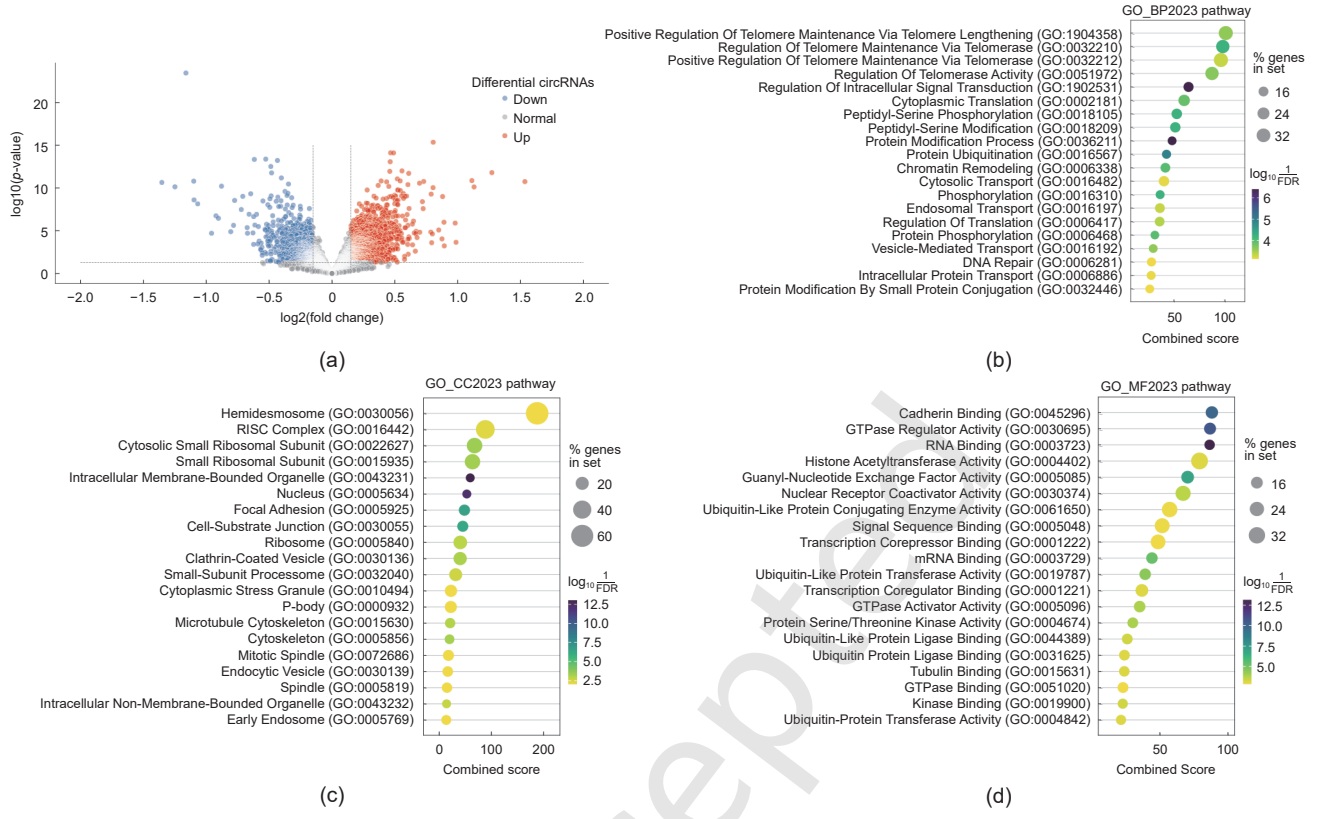


Fig. 2 Differential-expression and functional-enrichment analyses used to construct Panel-Sgnf and Panel-Pathway. (a) Volcano plot for differential expression analysis, with *x*-axis $\log_2$(fold change) and *y*-axis $-\log_{10}(p\text{-value})$. (b) GO enrichment test on BP term set. (c) GO enrichment test on CC term set. (d) GO enrichment test on MF term set.

quantified using Silhouette scores, which are compared against two background distributions of Silhouette scores to evaluate statistical significance. Higher Silhouette values reflect stronger spatial clustering of circRNAs within the 3D genome. Pathways are considered spatially enriched if their Silhouette scores exceed 0.35, while both background distributions are below 0.25 (with background circRNAs drawn from all enriched pathways yielding Silhouette scores <0.3). From each qualifying spatial cluster, we select the top five circRNAs ranked by their correlation coefficients with the HCC versus normal outcome labels. This procedure results in the identification of 191 circRNAs, collectively designated as Panel-3DG, representing features informed by both pathway enrichment and spatial genome organization.

2.2.5 TopN circRNAs with highest correlation coefficients (Panel-TopN)

To evaluate whether the HCC prediction efficacy of the 191 circRNAs from Panel-3DG surpasses that of circRNAs not selected based on 3D genome spatial clustering, we construct a comparison panel. Specifically, we select the top-191 circRNAs from the 2369 differentially expressed circRNAs identified in Panel-Sgnf, ranking them by their Spearman correlation coefficients with the HCC/normal outcome. This new panel, referred to as Panel-TopN, provides a set of features based solely on statistical correlation without incorporating spatial genomic information. The rationale behind constructing Panel-TopN is to isolate the predictive value added by 3D genome-based feature selection and determine if the spatial clustering of circRNAs contributes to improved model performance. By comparing the performance of predictive models trained on Panel-3DG and Panel-TopN, we aim to uncover whether the incorporation of 3D genomic context enhances the biological relevance and predictive accuracy of circRNA-based panels for HCC classification. This analysis is crucial for assessing the added value of integrating genomic structural data into feature selection strategies for machine learning models.

2.2.6 Concise circRNA panel (Panel-Concise)

For practical applications such as the development of diagnostic reagent kits that require rapid result generation, or in scenarios where whole-genome circRNA profiling is not feasible, we design highly compact circRNA panels consisting of only 3–4 features optimized for each machine learning model. These reduced sets are referred to as Panel-Concise. The construction of these panels follows a sequential forward heuristic search strategy. Specifically, all 191 circRNAs from Panel-3DG are first ranked in descending order based on their individual correlation coefficients with the HCC/normal outcome. CircRNAs are then iteratively added one at a time: If the inclusion of a candidate improves predictive performance, as measured by the Area Under the Receiver Operating Characteristic Curve (AUROC), it is retained; otherwise, it is excluded. This stepwise approach ensures that each circRNA in the final panel contributes incremental predictive value. In comparison, we selecte top-4 most differentially expressed circRNA as "positive control" panel named Panel-DETop and perform prediction under the same cross validation setting. As Fig. S2 in the ESM indicates, the prediction performance is way worse than the Panel-Concise which is enabled by 3D-genome-selected circRNAs.

Through this multi-level design process, we ultimately generate six distinct circRNA panels. We hypothesize that Panel-3DG, owing to its functional enrichment and spatial diversification, will achieve superior predictive performance compared to broader or

Just Accepted

less structured panels such as Panel-All, Panel-Sgnf, Panel-Pathway, and Panel-TopN. By contrast, Panel-Concise, with its minimalist architecture, is particularly suited for use cases demanding simplicity, speed, and cost-effectiveness—such as point-of-care diagnostics or analyses in resource-limited clinical settings. Together, these panels provide a flexible framework that balances predictive accuracy with practical feasibility, facilitating translational applications of circRNA-based biomarkers in HCC prediction.

2.3 Prediction models

Four widely adopted machine learning prediction models—LR, SVM, LightGBM, and DNN—are employed for each of the six circRNA panels. Detailed model parameter settings are shown in Table S2 in the ESM. To evaluate the predictive performance of each model-panel combination, we conduct 100 iterations of 10-fold cross-validation and standalone training–test, ensuring stability and robust generalization. The AUROC metric is used to measure the performance of the models.

By training multiple models on each panel and performing a comprehensive comparison, we analyze their predictive power across six feature panels: Panel-All, Panel-Sgnf, Panel-Pathway, Panel-3DG, Panel-TopN, and Panel-Concise. The parameters for each machine learning model are fine-tuned to maximize cross-validation AUROC scores. In addition to the cross-validation process, standalone testing datasets are used to evaluate the performance of the models on unseen data after model construction and parameter optimization. This additional evaluation step aims to verify the generalization ability of the models on previously unseen samples. The AUROC values calculated on the testing datasets provide a reliable measure of the classification performance for binary prediction tasks. This process not only enhances the robustness of the prediction models but also provides strong evidence for their practical application. The workflow is summarized in Fig. 3.

3 Result

3.1 Spatial distribution of circRNAs in 3D genome

Following GO enrichment analysis of the differentially expressed circRNAs, we map each circRNA to its corresponding 3D genomic position using the IMR90-derived 3D genome model. The IMR90 Hi-C data are adopted as a representative one, considering the whole-genome level Hi-C contact maps of six widely adopted Hi-C cell lines, namely HUVEC, IMR90, GM12878, NHEK, KBM7, and K562, are highly correlated to each other, as Fig. S3 in the ESM demonstrates. K-means clustering is subsequently performed on the circRNAs based on their spatial genome coordinates $< x, y, z >$. To assess the quality of clustering, Silhouette scores are calculated for circRNAs within each enriched pathway, as well as for the 3D genome background and across all enriched pathways collectively. The distribution of Silhouette scores, illustrated in Fig. 4, indicates that circRNAs included in Panel-3DG are derived from distinct genomic loci within the 3D nuclear space. This spatial diversification is expected to reduce the likelihood of collinearity, thereby enhancing the robustness of HCC versus normal prediction. For comparison, two types of background genes are used: all genes in the same cell line with available 3D genomic coordinates ("3D genome background") and all circRNAs identified in our study ("cross-pathway background"). The optimal number of clusters for the background genes is determined according to the cluster number at which the circRNAs of the current pathway achieve the highest silhouette score.

In addition, Fig. 5 presents the *k*-means clustering results for circRNAs associated with the top three enriched GO pathways from each of the three categories—GO_BP, GO_CC, and GO_MF. The results show that when circRNAs are selected from different spatial clusters, the resulting panels encompass heterogeneous loci distributed across the 3D genome. Such heterogeneity in spatial origin contributes to minimizing

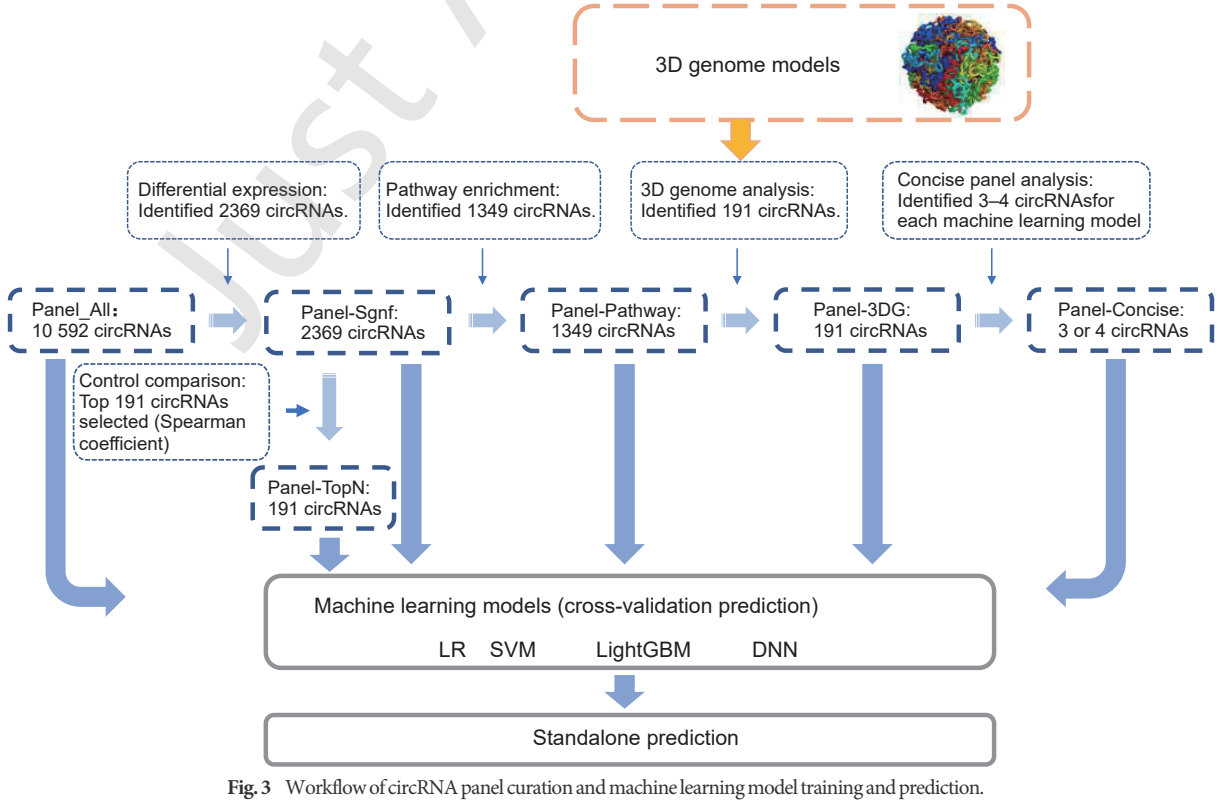


Fig. 3 Workflow of circRNA panel curation and machine learning model training and prediction.

Just Accepted

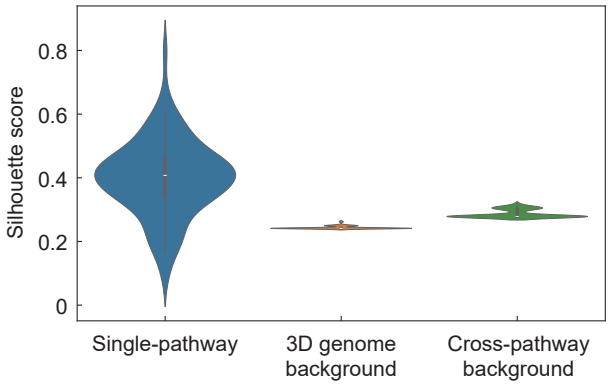


Fig. 4 Silhouette score distribution of circRNAs from each 3D genome cluster of each enriched pathway (left), the Silhouette score distribution of all genes' 3D genome loci (middle), and the Silhouette score distribution of circRNAs from each enriched pathway (right).

redundancy within the panel and supports the development of feature sets with improved generalizability for predictive modeling.

Taken together, these findings provide a strong rationale for incorporating 3D genome information into circRNA feature selection strategies, as spatially diversified panels can mitigate collinearity risks and enhance predictive performance in HCC classification tasks.

3.2 Cross-validation predictions

For two-thirds of the whole dataset, we conduct cross-validation-based training and predictions. For each of the six circRNA panels, we conduct four widely adopted machine learning methods, i.e., LR, SVM, LightGBM, and DNN, under 100 times 10-fold cross-validation. We then calculate the AUROC as demonstrated in Fig. 6.

The cross-validation AUROC results indicate that LR performs consistently well across all panels, with Panel-3DG showing the most significant effect, achieving an AUROC of 0.980. Similarly, SVM demonstrates strong performance across all panels, particularly excelling in Panel-3DG, where it achieves the highest average AU-ROC value of 0.984. While LightGBM's performance is slightly lower than the other models, it improves noticeably in Panel-3DG, reaching an average AUROC of 0.972. DNN performs stably overall but slightly decreases performance for Panel-Concise, while still achieving an average AUROC of 0.973 in Panel-3DG.

These results demonstrate that, over 100 iterations of 10-fold cross-validation, Panel-3DG outperforms all other panels in HCC/normal prediction, underscoring the value of circRNA selection based on 3D genome distribution. Both SVM and LR prove to be reliable model choices, particularly for Panel-3DG. However, the diversity of panels reflects different biological aspects, suggesting that selecting specific panels flexibly can better reveal the functional roles of circRNAs in HCC prediction.

3.3 Standalone dataset predictions

We allocate two-thirds of the dataset for training the machine learning models and use the remaining one-third as an

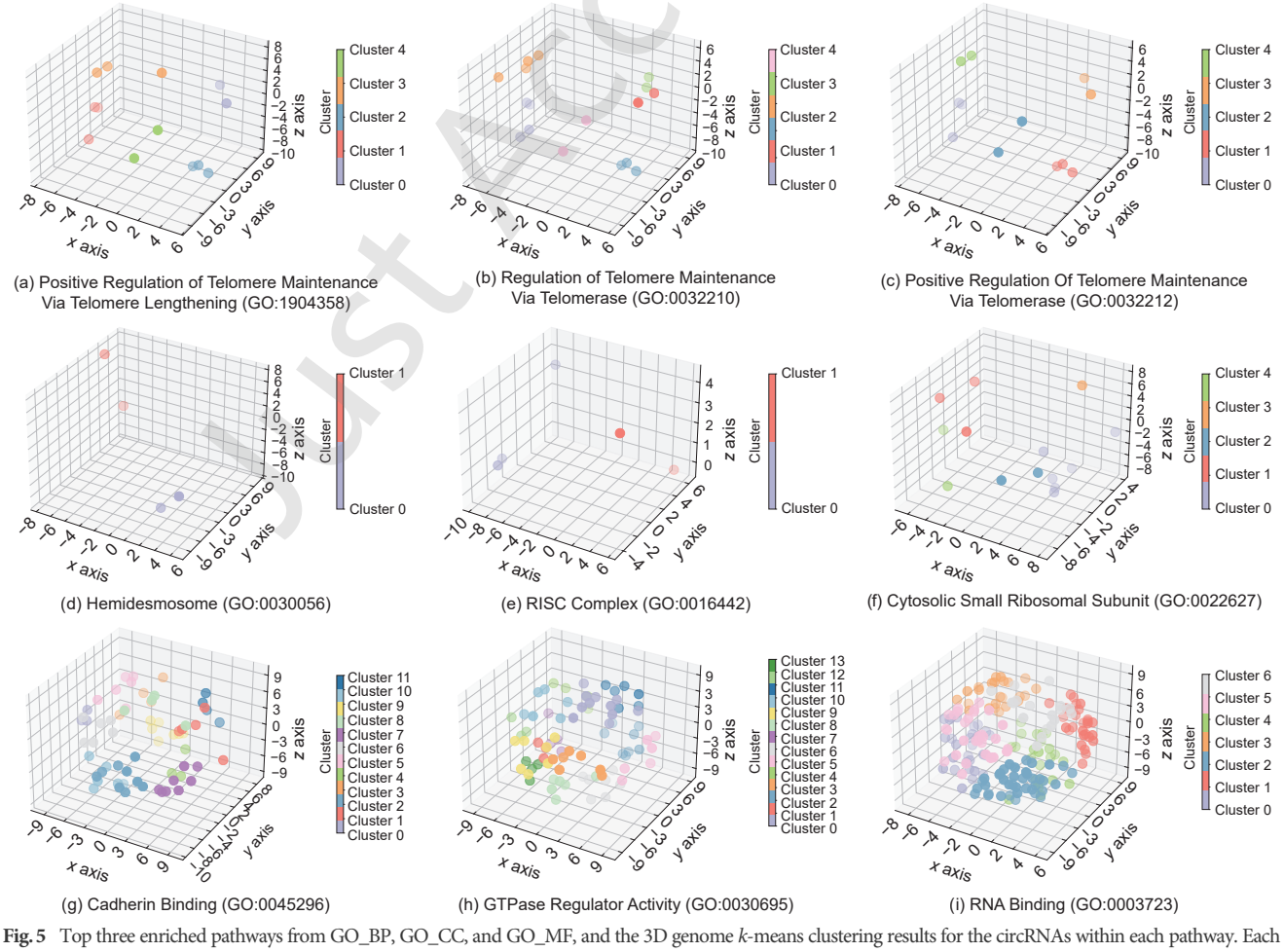


Fig. 5 Top three enriched pathways from GO_BP, GO_CC, and GO_MF, and the 3D genome *k*-means clustering results for the circRNAs within each pathway. Each dot of different color represents a circRNA belong to a specific cluster.

Just Accepted

清华大学出版社 Tsinghua University Press | SciOpen

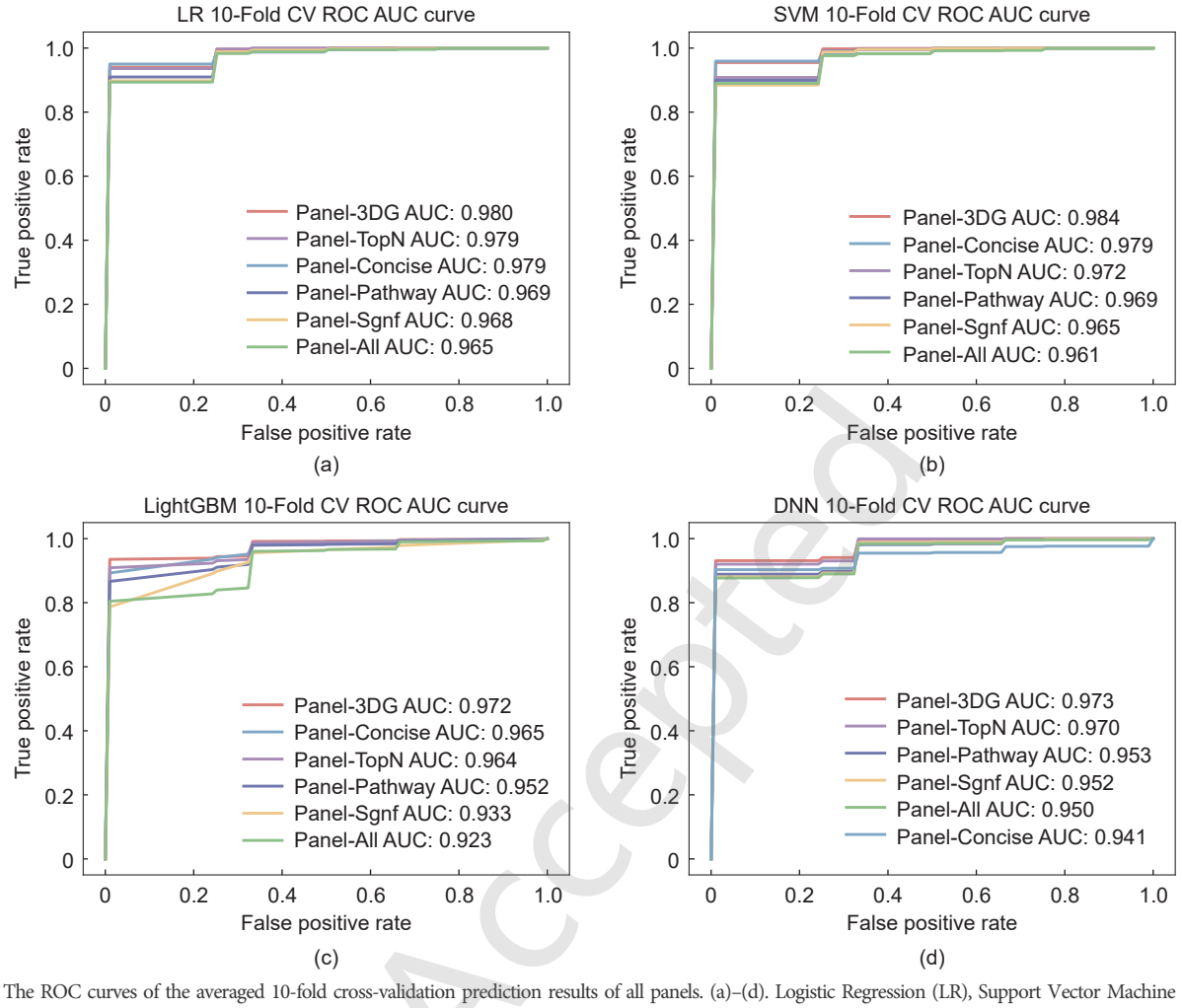


Fig. 6 The ROC curves of the averaged 10-fold cross-validation prediction results of all panels. (a)–(d). Logistic Regression (LR), Support Vector Machine (SVM), LightGBM and Deep Neural Network (DNN) based 10-fold cross-validation prediction AUC curves of the six circRNA panels, respectively.

independent test set to evaluate predictive efficacy. Model performance is assessed using the AUROC and the Area Under the Precision-Recall curve (AUPR), providing complementary perspectives on classification accuracy and reliability. As illustrated in Fig. 7, LR demonstrates consistently strong performance across all feature panels, with its best results observed on Panel-Pathway, Panel-3DG, and Panel-Concise. The combination of high accuracy and stability suggests that LR is a dependable baseline for HCC prediction. SVM also achieves competitive results, performing particularly well on Panel-3DG, which highlights the model's ability to capitalize on circRNA features informed by 3D genomic structure. LightGBM shows slightly weaker overall performance on the test set compared with other methods, yet demonstrates clear gains on Panel-Pathway and Panel-3DG, indicating that tree-based models can effectively leverage pathway enrichment and structural features. DNN performs robustly across all panels, excelling on Panel-Pathway, Panel-3DG, and Panel-Concise, reflecting their capacity to capture complex, multi-dimensional relationships within circRNA data.

Across the six feature panels, Panel-3DG consistently produces the strongest results, underscoring the critical value of incorporating 3D genome-informed circRNA features. By diversifying circRNAs according to their spatial genomic loci, Panel-3DG not only improves predictive accuracy but also enhances model generalization. These findings demonstrate the importance of integrating 3D genome structure into biomarker discovery and suggest that Panel-3DG provides a more precise and biologically meaningful foundation for early HCC prediction. Ultimately, this approach offers theoretical support for advancing personalized treatment strategies and developing circRNA-based diagnostic tools in liver cancer.

To further validate whether 3D-genome-selected circRNA panel has added value comparing to widely adopted feature selection methods in machine learning, we apply LinearSVC with an L1 norm to the 10 592 circRNAs and select 206 circRNAs. The LR, SVM, LightGBM, and DNN models are trained using the same grid-search hyperparameter optimization strategy. The result is demonstrated in Fig. S4 in the ESM, indicating 3D-genome-selected circRNA panel has more generalizable discriminatory power.

4 Discussion

This study underscores the considerable promise of circRNAs as biomarkers for HCC prediction, with a particular emphasis on the integration of diverse feature panels and the pivotal role of 3D genome-guided circRNA selection (Panel-3DG). By explicitly incorporating spatial genomic information, Panel-3DG consistently outperforms other panels, achieving superior predictive accuracy across multiple machine learning models. The strong AUROC values observed—especially for SVM and LR—highlight the effectiveness of selecting circRNAs from distinct 3D genomic loci to alleviate feature collinearity, a

Just Accepted

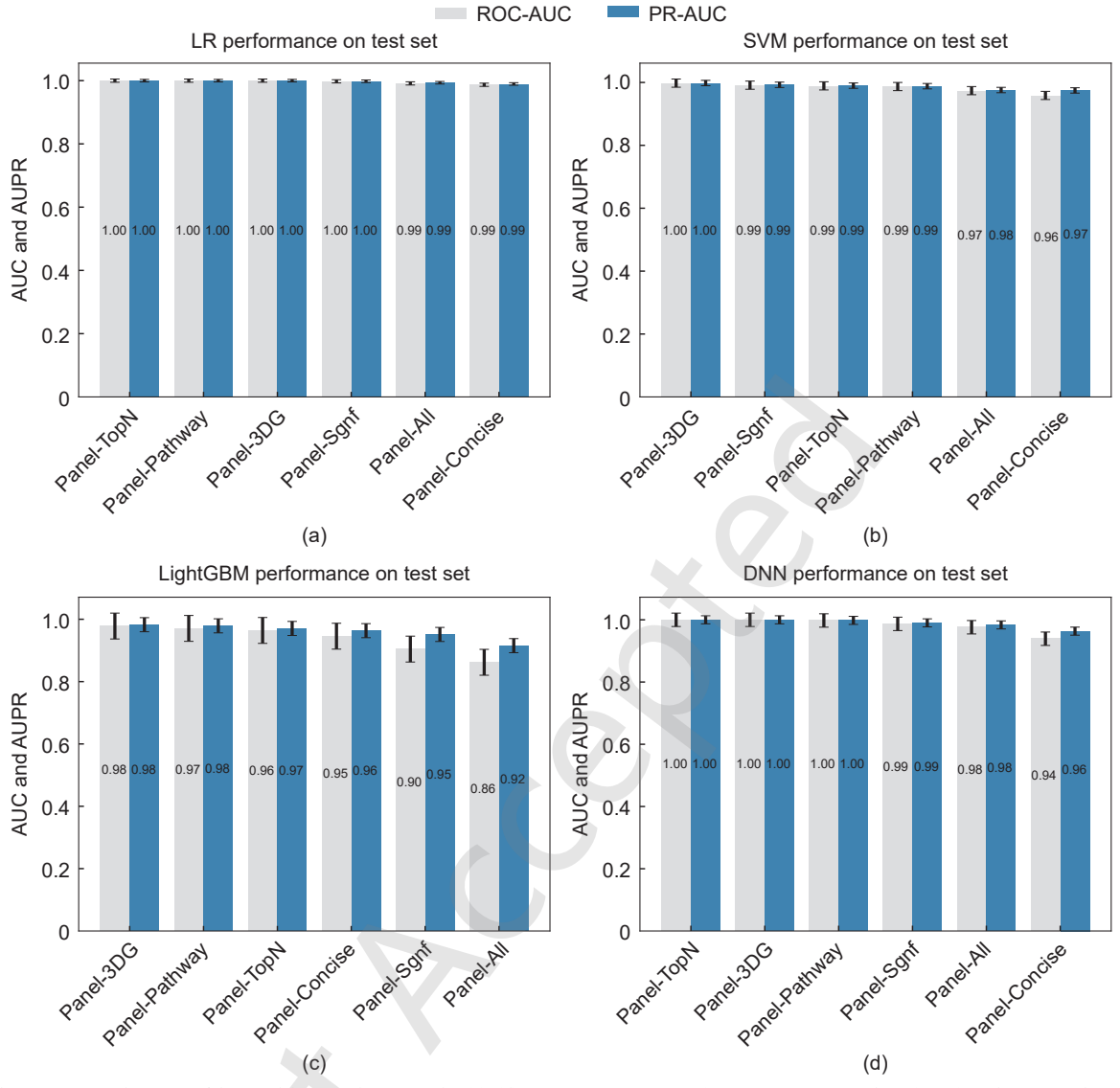


Fig. 7 The AU-ROC and AU-PR of the standalone prediction results. (a)–(d). Logistic Regression (LR), Support Vector Machine (SVM), LightGBM and Deep Neural Network (DNN) based standalone dataset prediction of the six circRNA panels, respectively.

challenge that is especially pronounced in small-scale omics datasets. This diversified feature selection strategy not only improves classification accuracy but also enhances generalization to unseen data, addressing a key limitation of earlier approaches reliant on broader, less structured feature sets.

Beyond predictive performance, Panel-3DG provides biologically meaningful insights by integrating the spatial organization of circRNAs within the 3D genome. This framework aligns with accumulating evidence that nuclear architecture profoundly influences transcriptional regulation and cellular function, suggesting that circRNAs located in different genomic compartments may participate in distinct signaling and regulatory pathways central to HCC pathogenesis. The combination of spatial mapping and pathway enrichment further improves interpretability, revealing how selected circRNAs may contribute to oncogenic processes and tumor-suppressive mechanisms. Notably, while LightGBM and DNNs are slightly less effective than SVM and LR on simpler panels, their predictive capacities increase substantially when applied to 3D genome-informed and pathway-enriched features, demonstrating their strength in capturing complex, multi-layered biological interactions.

In our Panel-3DG and Panel-Concise circRNA sets, several circRNAs have been implicated in the development of HCC, further underscoring the biological significance of our feature selection strategy. Notably, hsa_circ_0089372 has been proposed to function as a ceRNA by sponging miR-346[30], thereby regulating downstream genes involved in cell cycle control and chromosomal organization. Although its direct role in HCC remains to be experimentally validated, the oncogenic role of miR-346 in HCC provides a plausible mechanistic basis for the contribution of hsa_circ_0089372 to tumorigenesis. By contrast, hsa_circ_0066568 (circROBO1) has been experimentally demonstrated to promote HCC progression via the miR-130a-5p/CCNT2 axis[31]. Specifically, circROBO1 downregulates miR-130a-5p, leading to upregulation of CCNT2 and enhanced proliferation and tumor growth, while this effect can be reversed by miR-130a-5p mimics or CCNT2 knockdown. Together, these findings illustrate how circRNAs identified by our 3D genome-guided strategy may not only improve prediction performance but also reveal molecular mechanisms central to HCC biology.

Despite these promising results, several limitations warrant attention in future work. The relatively modest sample size and reliance on only two datasets may constrain generalizability; larger, more heterogeneous cohorts will be essential for rigorous

Just Accepted

validation. But the primary contribution of this work lies in the methodological innovation—introducing a 3D genome-informed perspective for circRNA selection—which offers a novel framework that can guide future validation studies. Moreover, although this study demonstrates the predictive and functional significance of 3D genome-informed circRNAs, the precise molecular mechanisms through which these circRNAs exert their influence in HCC remain poorly understood. The present findings are computational and exploratory, therefore, the need for functional experiments (e.g., knockdown/overexpression of top circRNAs) will be needed to validate the claims of predictive utility. Other functional validation experiments, coupled with integrative multi-omics analyses—including transcriptomics, proteomics, and epigenomics—will also be crucial to elucidating the mechanistic basis of these observations.

In conclusion, this study establishes the utility of 3D genome-guided circRNA panels as a powerful tool for HCC prediction, offering a novel framework for biomarker discovery and translational application. By bridging computational strategies with biological context, our findings advance the development of more accurate and generalizable predictive models, with direct implications for early diagnosis, personalized treatment, and improved clinical outcomes in HCC management. This integrative approach represents a significant step toward harnessing advanced genomic and computational methodologies to address the persistent challenges of liver cancer detection and therapy.

Author Contributions

Shanshan Yin, Wenqian Yan and Chongchen Pang contributed equally to this work. Shanshan Yin, Wenqian Yan, Chongchen Pang and Liangjie Liu performed data curation and preprocessing; Shanshan Yin, Mofan Feng, Zhangli Yuan, and Haoyan Chen carried out methodology and machine learning model development; Xuanwen Bao, Weijia Fang, Xin Jin, and Weiping Zhu handled 3D genome data processing and analysis; Mika Torhola, Klaus Förger, and Henna Kujanen contributed bioinformatics and statistical analysis; Lintai Da, Hao Li, Jacek Jemielity, and Weijia Fang provided experimental validation and clinical input. Shanshan Yin, Wenqian Yan, Chongchen Pang and Liangjie Liu prepared Figures. The original draft was written by Shanshan Yin and Yi Shi, and reviewed and edited by Yi Shi, Guang He, Joanna Kowalska, and Yuqing Lou, Guang He, Joanna Kowalska, and Yi Shi supervised the study, and Yi Shi and Guang He, acquired funding. All authors reviewed and approved the final manuscript.

Acknowledgment

This work was supported by the National Key Research and Development Program (Nos. 2022YFE0125300, 2024YFC2707002, and 2022YFE0125300), the Innovation Program of Shanghai Municipal Education Commission (No. 2023ZKZD16), the National Natural Science Foundation of China (No. 82071262), the Shanghai Municipal Science and Technology Major Project (Nos. 2017SHZDZX01 and 20JC1418600), the Key Technology Breakthrough Program of Ningbo Sci-Tech Innovation YONGJIANG 2035 (No. 2024Z221), the Shanghai Jiao Tong University Medicine-Engineering Fund (Nos. YG2026LC14, YG2025QNA46, YG2023ZD26, YG2022ZD024, YG2022QN111, YG2023LC14, and YG2024QNA59), and the Business Finland for project Machine Learning and 3D Genome based Cancer Early Prediction and Personalized Neoantigen Identification to Atostek Oy.

Electronic Supplementary Material

Supplementary material (circRNA expression datasets, detailed hyperparameter settings of machine learning models, data normalization results, Hi-C data correlation analysis, and additional prediction performance evaluations, including 2 tables and 4 figures) is available in the online version of this article.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

References

[1] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram and A. Jemal, Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, *CA Cancer J. Clin.*, vol. 74, no. 3, pp. 229–263, 2024.

[2] A. Ozakyol, Global epidemiology of hepatocellular carcinoma (HCC epidemiology), *J. Gastrointest. Cancer*, vol. 48, no. 3, pp. 238–240, 2017.

[3] R. Chou, C. Cuevas, R. Fu, B. Devine, N. Wasson, A. Ginsburg, B. Zakher, M. Pappas, E. Graham and S. D. Sullivan, Imaging techniques for the diagnosis of hepatocellular carcinoma: A systematic review and meta-analysis, *Ann. Intern. Med.*, vol. 162, no. 10, pp. 697–711, 2015.

[4] M. R. Toh, E. Y. T. Wong, S. H. Wong, A. W. T. Ng, L. H. Loo, P. K. H. Chow and J. Ngeow, Global epidemiology and genetics of hepatocellular carcinoma, *Gastroenterology*, vol. 164, no. 5, pp. 766–782, 2023.

[5] J. Salzman, C. Gawad, P. L. Wang, N. Lacayo and P. O. Brown, Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types, *PLoS One*, vol. 7, no. 2, p. e30733, 2012.

[6] S. Memczak, M. Jens, A. Elefsinioti, F. Torti, J. Krueger, A. Rybak, L. Maier, S. D. Mackowiak, L. H. Gregersen, M. Munschauer, et al., Circular RNAs are a large class of animal RNAs with regulatory potency, *Nature*, vol. 495, no. 7441, pp. 333–338, 2013.

[7] J. Wei, M. Li, C. Xue, S. Chen, L. Zheng, H. Deng, F. Tang, G. Li, W. Xiong, Z. Zeng, et al., Understanding the roles and regulation patterns of circRNA on its host gene in tumorigenesis and tumor progression, *J. Exp. Clin. Cancer Res.*, vol. 42, no. 1, p. 86, 2023.

[8] J. Li, X. Dong, X. Kong, Y. Wang, Y. Li, Y. Tong, W. Zhao, W. Duan, P. Li, Y. Wang, et al., Circular RNA hsa_circ_0067842 facilitates tumor metastasis and immune escape in breast cancer through HuR/CMTM6/PD-L1 axis, *Biol. Direct*, vol. 18, no. 1, p. 48, 2023.

[9] M. He, Y. Pan, C. You and H. Gao, CircRNAs in cancer therapy tolerance, *Clin. Chim. Acta*, vol. 558, p. 119684, 2024.

[10] T. B. Hansen, T. I. Jensen, B. H. Clausen, J. B. Bramsen, B. Finsen, C. K. Damgaard and J. Kjems, Natural RNA circles function as efficient microRNA sponges, *Nature*, vol. 495, no. 7441, pp. 384–388, 2013.

[11] Y. Tao, Q. Zhang, H. Wang, X. Yang and H. Mu, Alternative splicing and related RNA binding proteins in human health and disease, *Signal Transduct. Target. Ther.*, vol. 9, no. 1, p. 26, 2024.

[12] X. K. Ma, S. N. Zhai and L. Yang, Approaches and challenges in genome-wide circular RNA identification and quantification, *Trends Genet.*, vol. 39, no. 12, pp. 897–907, 2023.

[13] Y. Zhang and Y. Wang, Circular RNAs in hepatocellular carcinoma: Emerging functions to clinical significances, *Front. Oncol.*, vol. 11, p. 667428, 2021.

[14] T. Abaza, M. K. A. El-Aziz, K. A. Daniel, P. Karousi, M. Papatsirou, S. A. Fahmy, N. M. Hamdy, C. K. Kontos and R. A. Youness, Emerging role of circular RNAs in hepatocellular carcinoma immunotherapy, *Int. J. Mol. Sci.*, vol. 24, no. 22, p. 16484, 2023.

Just Accepted

[15] Z. D. Li, Y. L. Li, J. Lu, S. Liang, C. Zhang and L. H. Zeng, Recent research progress of circular RNAs in hepatocellular carcinoma, *Front. Oncol.*, vol. 13, p. 1192386, 2024.
[16] H. M. Bedair, E. A. El-Banna, E. A. Ahmed, M. G. Elhelbawy, A. Abdelfattah, F. A. Khalaf, M. Abdel-Samiee and A. El Sharnoby, Evaluation of circular RNA SMARCA5 as a novel biomarker for hepatocellular carcinoma, *Asian Pac. J. Cancer Prev.*, vol. 25, no. 4, pp. 1411–1417, 2024.
[17] X. Li, X. Yin, H. Bao and C. Liu, Circular RNA ITCH increases sorafenib-sensitivity in hepatocellular carcinoma via sequestering miR-20b-5p and modulating the downstream PTEN-PI3K/Akt pathway, *Mol. Cell. Probes*, vol. 67, p. 101877, 2023.
[18] Z. Liu, F. Yang, Z. Xiao and Y. Liu, Review of novel functions and implications of circular RNAs in hepatocellular carcinoma, *Front. Oncol.*, vol. 13, p. 1093063, 2023.
[19] Y. Ma, T. Wang, X. Zhang, P. Wang and F. Long, The role of circular RNAs in regulating resistance to cancer immunotherapy: Mechanisms and implications, *Cell Death Dis.*, vol. 15, no. 5, p. 312, 2024.
[20] D. Fei, F. Wang, Y. Wang, J. Chen, S. Chen, L. Fan, L. Yang, Q. Ren, S. Duangmano, F. Du, et al., Circular RNA ACVR2A promotes the progression of hepatocellular carcinoma through mir-511-5p targeting PI3K-Akt signaling pathway, *Mol. Cancer*, vol. 23, no. 1, p. 159, 2024.
[21] T. He, J. Huang, H. Liang, B. Zhong, G. Xu and X. Zhu, Circular RNA hsa_circ_0005239 contributes to hepatocellular carcinoma cell migration, invasion, and angiogenesis by targeting the miR-34a-5p/PD-L1 axis, *Cell Biol. Int.*, vol. 47, no. 9, pp. 1519–1534, 2023.
[22] L. Sun, B. Yeerkenbieke, A. Yuemaierabola, F. Liu, D. Yeerxiati, X. Dong and W. Guo, Expression of circular RNA has-circ-0009158 and identification of associated miRNA-mRNA network in hepatocellular carcinoma, *Am. J. Transl. Res.*, vol. 16, no. 2, pp. 415–431, 2024.
[23] J. Han, T. Thurnherr, A. Y. F. Chung, B. K. P. Goh, P. K. H. Chow, C. Y. Chan, P. C. Cheow, S. Y. Lee, T. K. H. Lim, S. S. Chong, et al., Clinicopathological-associated regulatory network of deregulated circRNAs in hepatocellular carcinoma, *Cancers (Basel)*, vol. 13, no. 11, p. 2772, 2021.
[24] P. Gao, Y. Yang, X. Li, Q. Zhao, Y. Liu, C. Dong, Y. Zhang and D. Liu, Circular RNA hsa_circ_0098181 inhibits metastasis in hepatocellular carcinoma by activating the Hippo signaling pathway via interaction with eEF2, *Ann. Hepatol.*, vol. 28, no. 5, p. 101124, 2023.
[25] H. Drucker, D. Wu and V. N. Vapnik, Support vector machines for spam categorization, *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 1048–1054, 1999.
[26] D. Gupta, U. Kose, B. Le Nguyen and S. Bhattacharyya, *Artificial Intelligence for Data-Driven Medical Diagnosis*. Berlin, Germany: De Gruyter, 2021.
[27] Y. LeCun, Y. Bengio and G. Hinton, Deep learning, *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[28] H. Shin, Y. Shi, C. Dai, H. Tjong, K. Gong, F. Alber and X. J. Zhou, TopDom: An efficient and deterministic method for identifying topological domains in genomes, *Nucleic Acids Res.*, vol. 44, no. 7, p. e70, 2016.
[29] Y. Shi, Z. Guo, X. Su, L. Meng, M. Zhang, J. Sun, C. Wu, M. Zheng, X. Shang, X. Zou, et al., DeepAntigen: A novel method for neoantigen prioritization via 3D genome and deep sparse learning, *Bioinformatics*, vol. 36, no. 19, pp. 4894–4901, 2020.
[30] H. He, Z. Shen, Q. Gu and R. Xie, Construction of ceRNA network and identification of two differentially expressed circRNAs in hepatocellular carcinoma by bioinformatic analysis, *Int. J. Clin. Exp. Pathol.*, vol. 13, no. 11, pp. 2727–2737, 2020.
[31] H. Meng, R. Li, Y. Xie, Z. Mo, H. Zhai, G. Zhang, G. Liang, X. Shi and B. Zhou, Nanoparticles mediated circROBO1 silencing to inhibit hepatocellular carcinoma progression by modulating miR-130a-5p/CCNT2 axis, *Int. J. Nanomedicine*, vol. 18, pp. 1677–1693, 2023.



Shanshan Yin received the BEng degree from Dezhou University in 2016 and the MEng degree from University of Science and Technology Beijing, China, in 2018. She is currently a PhD candidate at Shanghai Jiao Tong University, China, with 5 publications. Her research interests include machine learning and computational biology.



Wenqian Yan received Bachelor's degree from Northwest A&F University in 2023. Currently a PhD student at Shanghai Jiao Tong University and has 4 publications. Research interests include cancer genetics and immunotherapy.



Chongchen Pang Received Bachelor's degree from Shanghai Jiao Tong University in 2024. Currently a PhD student at Shanghai Jiao Tong University and has 3 publications. Research interests include bioinformatics and computational biology.



Liangjie Liu Received Bachelor's degree, Master's degree, and PhD degree from Shanghai Jiao Tong University in 2016, 2018 and 2026. Currently a postdoctoral research fellow at Shanghai Jiao Tong University and has 31 publications. Research interests include bioinformatics and computational biology.



Mofan Feng Received Bachelor's degree from Sichuan University in 2020, and PhD degree from Shanghai Jiao Tong University in 2025. Currently a postdoctoral research fellow at Shanghai Jiao Tong University and has 14 publications. Research interests include computational biology.



Zhangli Yuan Received Bachelor's degree from Shanghai Jiao Tong University in 2024. Currently a PhD student at Academy for Advanced Interdisciplinary Studies, Peking University, with research interests in bioinformatics.

Just Accepted



Haoyan Chen Received Bachelor's degree from Southeast University in 2004, and PhD degree from Shanghai Jiao Tong University School of Medicine in 2009. Currently a professor at Shanghai Jiao Tong University School of Medicine. He received the Yangtze River Scholar award in 2026. Research interests include bioinformatics, oncology and microbiome.



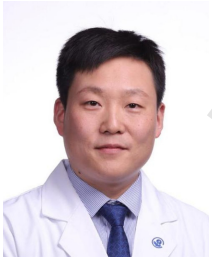
Xuanwen Bao Received his degree from the Technical University of Munich. Currently an attending physician in the Oncology Department of Zhejiang University Hospital, and holds the position of Principal Investigator at Zhejiang University. His research focuses on artificial intelligence, bioinformatics, the tumor immune microenvironment, and spatial biology.



Weijia Fang Dr. Fang is currently a professor and the Director of the Third Division of Medical Oncology at Zhejiang University Hospital and a Chief Physician. His research focuses on CAR-T therapy for solid tumors and the tumor immune microenvironment.



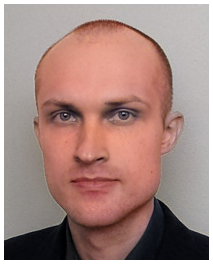
Xin Jin Received the master's degree from Fudan University in 2024. Currently a PhD student at Fudan University, majoring in hepatic oncology. Research interests include data mining and the application of artificial intelligence in medicine.



Weiping Zhu Received the PhD degree from Fudan University in 2017. Currently a surgeon at Fudan University Shanghai Cancer Center, specializing in hepatic oncology. Research interests include the application of artificial intelligence in oncology.



Mika Torhola Mika Torhola is the Chairman of the Board of Atostek Ltd., a Finnish technology company, a position he has held since September 20, 2024. Previously, he served as the company's Managing Director (CEO) for 11 years, during which he gained extensive experience in business development, strategic planning, and operational leadership.



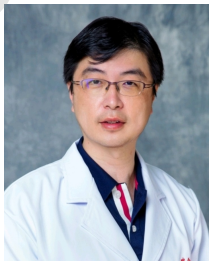
Klaus Förger Received the PhD degree from Aalto University in 2015. Now a Senior Software Designer at Atostek Ltd. Research interests include bioinformatics and computational algorithm.



Henna Kujanen Received Bachelor and Master's degree from Tampere University in 2019 and 2024. Currently project manager and IT professional with special expertise in healthcare software and health data at Atostek Ltd.



Lintai Da Received PhD degree fro Peking University in 2010. Now a tenured associate professor at Shanghai Jiao Tong University. Research interests include computational structural biology and bioinformatics.



Hao Li Received PhD degree from Shanghai Jiao Tong University in 2013. Currently an Associate Chief Physician at Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, specializing in gastrointestinal oncology. Research interests include cancer immunotherapy and targeted therapy.



Jacek Jemielity Received his PhD degree in Chemistry from the University of Warsaw in 2002. Currently a Professor and Director of the Centre of New Technologies at the University of Warsaw, where he leads the Laboratory of Chemical Biology. His research focuses on chemically modified mRNA, particularly the synthesis and application of mRNA cap analogues and modified nucleotides for therapeutic applications.



Yuqing Lou Received his M.D. and PhD degrees and currently serves as Director of the Department of Science and Education and Associate Chief Physician in the Department of Respiratory Medicine at Shanghai Chest Hospital, Shanghai Jiao Tong University School of Medicine. He is also an Associate Research Fellow at the School of Medicine. His research focuses on lung cancer and thoracic oncology, and has published 20 academic papers.

Just Accepted



Guang He Received her PhD degree from China Medical University in 2002. Currently a Research Professor at Shanghai Jiao Tong University and Assistant Director of the Bio-X Research Institute. Her research focuses on the genetic basis of psychiatric disorders, particularly schizophrenia, and the molecular mechanisms underlying the effects of prenatal environmental factors on mental health. She has published more than 120 academic papers and has led several national research projects, including projects under the National Key Research and Development Program of China.



Joanna Kowalska Received her PhD degree from the University of Warsaw in 2010. Currently an Assistant Professor and research group leader at the University of Warsaw. Her research focuses on chemical biology, RNA modifications, mRNA cap structures, and the development of chemically modified RNA molecules for therapeutic applications.



Yi Shi Received his PhD degree from the University of Alberta in 2012 and subsequently completed postdoctoral training at the University of Southern California. He is currently an Associate Professor at Shanghai Jiao Tong University. His research interests include machine learning-driven personalized cancer vaccines, cancer early prediction and subtyping, 3D chromatin architecture, cancer genetics, and cancer immunotherapy. He has published 108 academic papers.

Just Accepted