Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Mime: A flexible machine-learning framework to construct and visualize models for clinical characteristics prediction and feature selection

Hongwei Liu [a,b,1], Wei Zhang [a,b,1], Yihao Zhang [a,b,1], Abraham Ayodeji Adegboro [a,b], Deborah Oluwatosin Fasoranti [a,b], Luohuan Dai [a,b], Zhouyang Pan [a,b], Hongyi Liu [a,b], Yi Xiong [a,b], Wang Li [a,b], Kang Peng [b,c], Siyi Wanggou [a,b,*], Xuejun Li [a,b,*]

[a] *Department of Neurosurgery, Xiangya Hospital, Central South University, Changsha, Hunan 410008, China*
[b] *Hunan International Scientific and Technological Cooperation Base of Brain Tumor Research, Xiangya Hospital, Central South University, Changsha, Hunan 410008, China*
[c] *Department of Radiology, Xiangya Hospital, Central South University, Changsha, Hunan 410008, China*

ARTICLE INFO

Keywords:
Machine learning
Prediction models
Mime
PIEZO1
R package
GitHub

ABSTRACT

The widespread use of high-throughput sequencing technologies has revolutionized the understanding of biology and cancer heterogeneity. Recently, several machine-learning models based on transcriptional data have been developed to accurately predict patients' outcome and clinical response. However, an open-source R package covering state-of-the-art machine-learning algorithms for user-friendly access has yet to be developed. Thus, we proposed a flexible computational framework to construct a machine learning-based integration model with elegant performance (Mime). Mime streamlines the process of developing predictive models with high accuracy, leveraging complex datasets to identify critical genes associated with prognosis. An in silico combined model based on de novo PIEZO1-associated signatures constructed by Mime demonstrated high accuracy in predicting the outcomes of patients compared with other published models. Furthermore, the PIEZO1-associated signatures could also precisely infer immunotherapy response by applying different algorithms in Mime. Finally, SDC1 selected from the PIEZO1-associated signatures demonstrated high potential as a glioma target. Taken together, our package provides a user-friendly solution for constructing machine learning-based integration models and will be greatly expanded to provide valuable insights into current fields. The Mime package is available on GitHub (https://github.com/l-magnificence/Mime).

1. Introduction

The widespread use of high-throughput sequencing technologies have profoundly impacted the understanding of biology and cancer heterogeneity [1]. An increasing number of researchers have identified specific molecular features associated with disease progression, patient outcomes and therapeutic response from sequencing data [2,3]. These selective signatures provide a comprehensive overview of particular biological processes that regulate transcriptional networks of cancer cells [4]. However, due to the diversity and vastness of features, rational computational strategies are urgently needed to identify critical genes for disease.

Machine learning (ML), a branch of computer science that learns from complex datasets to develop a high-accuracy predictive model, has become a popular tool in medical research [5]. Several diagnostic and prognostic models based on machine learning have been developed from transcriptional data for various cancer types [6–10]. Since the performance of machine learning models from large amounts of transcriptional data can vary, it is often recommended to compare the results derived from several methods and select the optimal one for further application [6,11,12]. However, given various formats of input and output and system parameters supported by different in silico approaches, such as CoxBoost [13] and superpc [14], comparative analysis can become extremely complex. Till date, no software has been developed that covers state-of-the-art machine learning algorithms for user-friendly access, which systematically addresses the above problems (Table 1). Thus, integrating sequencing data with different machine learning algorithms will significantly expand and provide valuable

[*] Corresponding authors at: Department of Neurosurgery, Xiangya Hospital, Central South University, Changsha, Hunan 410008, China.
E-mail addresses: siyi.wanggou@gmail.com (S. Wanggou), lxjneuro@csu.edu.cn (X. Li).

[1] Hongwei Liu, Wei Zhang and Yihao Zhang have contributed equally to this work and share first authorship.

https://doi.org/10.1016/j.csbj.2024.06.035
Received 6 March 2024; Received in revised form 27 June 2024; Accepted 27 June 2024
Available online 29 June 2024

insights into current fields.

Here, we developed Mime, an open-source R package with elegant performance that simplified the procedure of constructing machine learning-based integration models from transcriptomic data. Mime mainly provides four implementations for the exploration of prediction models and candidate genes from large-scale features in one-stop: (i) establishment of prognostic models by integrating 10 machine learning algorithms, (ii) construction of binary response models by applying 7 machine learning algorithms, (iii) core feature selections related with prognosis by 8 machine learning methods and (iv) visualization of the performance of each model. We used de novo PIEZO1-associated signatures identified from primary glioblastoma cells and several publicly available cohorts as an example to demonstrate the workflow of Mime in detail and show its overall capabilities.

2. Methods

2.1. Acquisition and pre-processing of sample cohorts with survival information

Nine glioma datasets, comprising mRNA expression profiles and corresponding clinical and genomics feature data were acquired from publicly accessible databases. Notably, transcriptional data from the TCGA-Glioma dataset (n = 702), which encompasses the TCGA low-grade glioma (LGG) and glioblastoma cohorts, were retrieved from XENA (https://xena.ucsc.edu/) in June 2023. TCGA genomic feature profiles were acquired from UCSCXenaShiny [15]. Four external datasets, namely CGGA.325 (n = 325), CGGA.693 (n = 693), CGGA.1018 (n = 1018), and CGGA.array (n = 301) were obtained from the Chinese Glioma Genome Atlas (CGGA) (http://www.cgga.org.cn/). Clinical and transcriptomic annotations of 168 patients with RNA sequencing data for at least two time points were procured from the Glioma Longitudinal Analysis Consortium (GLASS) (https://www.synapse.org/glass). Based on diagnostic time points, we divided the cohorts into two datasets: the

Table 1
Overview of common machine-learning tools for model construction.

Tool	Algorithm	Model abbrev	Platform	Function	Multiple methods combination	Models comparison	Focused visualization	Source
randomForestSRC	Random forests	RSF	R	Survival, Classification and Feature selection	No	No	No	CRAN, GitHub
glmnet	Generalized linear model via penalized maximum likelihood	Enet, Lasso and Ridge	R	Survival and Feature selection	No	No	No	CRAN
survival	Stepwise Cox proportional hazards regression model	StepCox	R	Survival and Feature selection	No	No	No	CRAN, GitHub
CoxBoost	Cox proportional hazards model by componentwise likelihood-based boosting	CoxBoost	R	Survival and Feature selection	No	No	No	CRAN
plsRcox	Partial least squares regression generalized linear model	plsRcox	R	Survival	No	No	No	CRAN, GitHub
superpc	Supervised principal component analysis	superpc	R	Survival	No	No	No	CRAN, GitHub
gbm	Generalized boosted regression model	GBM	R	Survival	No	No	No	CRAN, GitHub
survivalsvm	Survival support vector analysis	survivalsvm	R	Survival	No	No	No	CRAN, GitHub
caret	Classification training using different methods including Naïve Bayes, AdaBoost Classification Trees, Boosted Logistic Regression, k-Nearest Neighbors, Random Forest and Support Vector Machines with Class Weights	nb, svmRadialWeights, rf, kknn, adaboost and LogitBoost	R	Classification	No	No	No	CRAN, GitHub
cancerclass	Nearest-centroid classification	cancerclass	R	Classification	No	No	No	Bioconductor
Boruta	A wrapper around a Random Forest classification algorithm	Boruta	R	Feature selection	No	No	No	CRAN
xgboost	Extreme gradient boosting	Xgboost	R	Feature selection	No	No	No	CRAN, GitHub
e1071	Support vector machine recursive feature elimination	SVM-REF	R	Feature selection	No	No	No	CRAN
Mime	Integrating above algorithms	-	R	Survival, Classification and Feature selection	Yes	Yes	Yes	GitHub

primary glioma cohort (n = 168) and the recurrent glioma cohort (n = 168). We also obtained GSE108474 (n = 314) and GSE16011 (n = 276) from the Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo). We excluded samples without complete survival information from these nine glioma cohorts. The data from the Illumina HiSeq platform was converted into the transcripts per kilobase million (TPM) format. Clean microarray data was obtained by correcting the background, performing quantile normalization and logarithmic transformation. The TCGA-Glioma dataset was employed as the training data to screen for the best predictive model, while the remaining eight glioma cohorts served as independent validation datasets.

2.2. Acquisition and pre-processing of sample cohort with immune checkpoint inhibitor therapy

Eighteen cohorts of pre-treatment samples with immune checkpoint inhibitors (ICI) were collected from published studies and resources, comprising a total of 1059 patients across eight cancer types (296 responders and 763 non-responders). To obtain an integrated dataset, the R package sva (v3.40.0) was used to remove batch effects, and samples lacking response information were excluded. The unified group was arbitrarily segregated into two groups, designated as the training dataset (70 %, n = 730) and the validation dataset (30 %, n = 312). Supplementary Table S1 offers detailed information about the above cohorts.

2.3. Developing the optimal prognostic model with the 117 machine learning combinations

Here, based on the literature, we provided a novel machine-learning framework to develop the optimal model for predicting the prognosis of patients based on the input variables and the provided cohorts [6]. First, we conducted univariate Cox regression analysis to identify prognostic features from the input variables. Genes exhibiting a p-value less than 0.05 were recognized as having prognostic significance. Second, the genes were entered into the machine learning framework. This framework incorporates ten classical machine learning algorithms, i.e., random forest (RSF), elastic network (Enet), stepwise Cox (StepCox), CoxBoost, partial least squares regression for Cox (plsRcox), supervised principal components (superpc), generalized boosted regression models (GBM), survival support vector machine (survivalsvm), Ridge, and least absolute shrinkage and selection operator (Lasso). Possible variable selection filters included Lasso, StepCox, CoxBoost, and RSF, each with distinct core parameters. 117 combinations were integrated into the computational framework with K-fold cross-validation for model construction on the training dataset. The precise parameters of the ten machine learning algorithms can be found in the original code and Supplementary Table S2. The model with the mean of the highest C-index in the validation cohorts was the most valuable predictive model with the best accuracy and lower risk of overfitting.

2.4. Comparing the optimal model with previously published predictive models

To compare the optimal predictive signature developed through the computational framework with other published predictive signatures, we gathered an exhaustive selection of prognostic signatures for glioma, encompassing LGG and glioblastoma. A total of 33, 22, and 40 different predictive signatures were identified for prognosticating the outcomes of glioma, LGG, and glioblastoma patients, respectively. The characteristics that distinguish the signature and their respective coefficients have been furnished in the R package Mime and presented in Supplementary Table S3.

2.5. Constructing models for predicting the binary variable with seven machine learning algorithms

Seven standard machine learning algorithms: Naïve Bayes (nb), AdaBoost Classification Tree (adaboost), cancerclass, Random Forest (rf), Boost Logistic Regression (LogiBoost), Weighted k-Nearest Neighbor Classifier (kknn), and Support Vector Machines with Class Weights (svmRadialWeights), were implemented to create a model for the prediction of a binary variable [16–21]. The detailed parameters of each are shown in the original code and Supplementary Table S4. As cancerclass required no parameters, the entire training dataset was utilized to train the model. To determine the optimal model, fivefold-cross validation was performed, with each resampling repeated 10 times. The model developed by the seven algorithms that produced the best area under the curve (AUC) was regarded as the optimal model for downstream analysis.

2.6. Comparing the developed response model with other predictive models for ICI therapy

To examine the effectiveness of the binary predictive model that we created based on ICI cohorts, we collected 13 previously published ICI response signatures, such as PDL1. Sig, IMS.Sig, and TRS.Sig [22–33]. The algorithms and code script were obtained from the original studies. We developed a visualization function in the R package, Mime, to compare the performance of all 13 signatures and our model. Further details on the 13 ICI signatures are available in Supplementary Table S5.

2.7. Screening out the core variables for prognosis with eight machine learning algorithms

Here, a novel computational framework consisting of eight machine learning algorithms was constructed to screen out the core prognostic variables based on the provided transcriptomic profile and corresponding survival information. The framework implementation consisted of three steps. Firstly, the prognostic factors were identified through univariate Cox regression analysis. Secondly, eight machine learning algorithms related to survival analysis were evaluated, comprising Lasso, Enet, Boruta, CoxBoost, RSF, eXtreme Gradient Boosting (Xgboost), StepCox, and Support Vector Machine Recursive Feature Elimination (SVM-REF), to identify the most crucial features. We executed LASSO 1000 times with different seeds and determined the core features as those that constituted variables that had been selected more than 50 times. The stepwise Cox regression analysis included the direction parameters ’forward’, ’backward’ and ’both’, whereas Enet involved 9 alpha parameters ranging from 0.1 to 0.9. Thus, 18 distinct models for screening fundamental characteristics using diverse parameters exists and can be found in the original code and Supplementary Table S6. Thirdly, we selected the most frequently filtered variables as core features based on their selection frequency.

2.8. Determining PIEZO1-associated signature

RNA-seq data of primary glioblastoma cell lines G508 and G532 treated with scrambled shRNA (Control) and PIEZO1 shRNA were acquired from GSE113261. Its raw sequencing data were mapped to the hg38 reference genome through HISAT2 and StringTie to obtain a raw count matrix. R package DESeq2 (v1.32.0) was used for differential gene expression analysis. Differentially expressed genes (DEGs) with log2 Fold Change > 2 (or < -2) and adjusted P-value < 0.05 were defined as significant. All up-regulated, and down-regulated DEGs among G508 and G532 cell lines between scrambled and PIEZO1 shRNA were intersected respectively to identify PIEZO1-associated signatures with high confidence.

2.9. Enrichment analysis

SDC1-regulated genes between high- and low-expression groups in TCGA, CGGA.325, CGGA.693 and GSE16011 respectively were identified by R package limma (v3.48.0). The criteria for gene filtering were set as log2 Fold Change > 0.5 (or < −0.5) and adjusted P-value < 0.05. Then, the GSEA algorithm in R package clusterProfiler (v4.7.1) was applied based on average values of log2 Fold Change in four datasets to conduct Gene Ontology (GO) enrichment analysis.

2.10. Survival analysis

Patients were divided into high and low survival groups according to the median value of a numeric variable for overall survival (OS) analysis as described by previous studies [34]. Hazard ratios (HRs) with 95 % confidence intervals (CI), log-rank P values and Kaplan-Meier curves were calculated and plotted by R package survival (v3.3–1) and survminer (v0.4.9). The Multivariate Cox proportional hazard model was executed by the R packages ezcox (v1.0.2).

2.11. Statistical analysis

All the data analysis and graph generation were completed in R (v4.1.3), Adobe Photoshop software and BioRender.com. The C-index of different models was determined by concordance from the Cox regression analysis based on the risk score calculated by the specific model. The receiver operating characteristic curve (ROC) used to predict binary categorical variables was conducted by the pROC package (v1.18.0). The time-dependent area under the ROC curve (AUC) for survival variables was implemented via the survivalROC package (v1.0.3). All

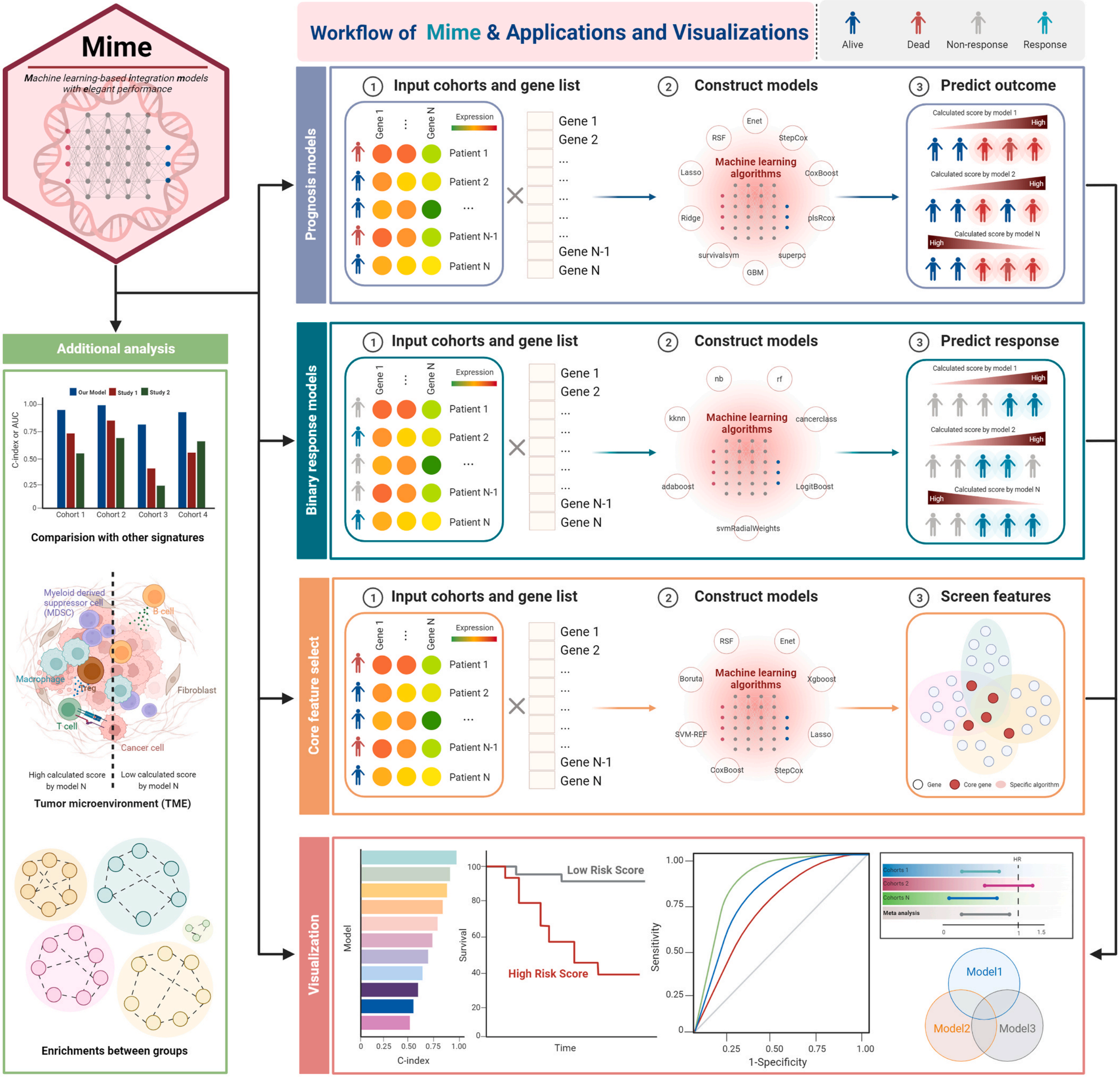


Fig. 1. A schematic diagram of Mime. Mime streamlined the process of developing models for accurately predicting outcomes and therapeutic responses of patients, leveraging complex datasets to identify critical genes associated with prognosis.

related packages used to build Mime are listed in Supplementary Table S7. Correlations between variables were explored using Pearson or Spearman coefficients. Continuous variables normally distributed between the binary groups were compared using a T-test. Categorical variables were compared using the Chi-Squared test. P-value < 0.05 was considered statistically significant, and all statistical tests were two-sided.

3. Results

3.1. Overview of Mime

The schematic diagram of Mime is depicted in Fig. 1, and a comprehensive comparison between Mime and other tools is listed in Table 1. Mime's user-free analysis involves three steps: data input, model construction, and results visualization. The first step for users is to acquire multiple cohorts containing transcriptional sequencing data with information on survival or clinical response to therapy as well as a gene set as inputs to Mime. Then, Mime applies various machine learning algorithms to train models for predicting the outcome or clinical response of patients, which could also screen core features from a large number of genes. Finally, Mime provides several graphs to help users interpret the results. In this process, the overall capacity of each model is comprehensively estimated, and users could select the optimal model for further analysis. Furthermore, Mime could compare the AUC or C-index of the optimal model with other established models derived from previous studies if provided by the user. In addition, the difference in immune cell infiltration and biological enrichments between samples with high-risk and low-risk scores calculated by utilizing the optimal model could also be determined in Mime. A detailed description of all functions and instructions of Mime are provided on GitHub.

3.2. Establishment of de novo prognosis models associated with PIEZO1 by Mime

Since PIEZO1-mediated mechanotransduction is essential for promoting glioma aggression, we used PIEZO1-asscociated signatures identified in primary glioblastoma cells from previous studies [35] as well as public glioma cohorts as an example to illustrate the application of Mime (Fig. 2A). RNA sequencing was performed on primary G508 and G532 cell lines with PIEZO1 knockdown by shRNA (Fig. 2A). In total, there were 89 shared down-regulated genes and 38 shared up-regulated genes among G508 and G532 after PIEZO1 knockdown, which were defined as PIEZO1-associated signatures (PIAS) (Fig. 2B). These signatures and 9 glioma transcriptomic datasets including one training cohort and eight validation cohorts were then used to construct models by integrating 10 machine learning algorithms in Mime. Among 117 models constructed by Mime, StepCox[forward]-Ridge combined model (STRICOM) had the highest C-index mean among the validation cohorts as well as in all other cohorts indicating its outstanding performance (Fig. 2C, Supplementary Fig. S1A). Indeed, the expression level of most genes in STRICOM were decreased in the PIEZO1-knockdown cells (Supplementary Fig. S1B). We further separated glioma patients into high-risk and low-risk groups according to the median risk score calculated by Mime based on STRICOM and determined its survival probability in each cohort. Interestingly, patients with high-risk score had significantly worse outcomes in all cohorts (Fig. 2D). These results demonstrated that Mime made it easy for users to build prognostic models based on the provided gene set and datasets.

3.3. Power evaluation of optimal model

As AUC is another metrics used to evaluate a prognostic model, we performed a time-dependent ROC curve analysis of STRICOM through Mime. Of note, the 1-year and 5-year AUC predicted by STRICOM ranked first with the highest mean of AUC in the validation cohorts, although the 3-year AUC was not ranking first (Fig. 3A). In particular, some validation cohorts such as CGGA, GLASS and GSE16011, also presented high AUC compared with the TCGA training cohort predicted by STRICOM (Fig. 3B, Supplementary Fig. S1C). The low power of the model in GSE108474 may be due to the quality of microarray data. To determine the prognostic effect of STRICOM, we performed a meta-analysis of univariate COX regression via Mime, which showed that the score calculated by STRICOM was the glioma risk factor (Fig. 3C). Having uncovered some known molecular biomarkers for glioma, we further performed multivariate COX regression analysis and found that the score calculated by STRICOM was an independent prognostic factor taking into account gender, age at diagnosis, WHO grade, IDH mutation, 1p/19q codeletion and MGMT promoter methylation in multiple datasets (Fig. 3D). Together, these results revealed that the optimal model constructed by Mime demonstrated high accuracy in predicting the outcomes of patients.

3.4. Comparison of established models based on gene expression

Recently, many prognostic and predictive models based on machine learning have been applied in glioma with the development of next-generation sequencing [36]. To comprehensively compare the performance of STRICOM with other published models in glioma, we retrieved 95 models from previous studies, which were packaged in Mime. However, users can also provide their models of a specific disease to Mime for comparison. In our study, we performed univariate Cox regression for each model across all datasets to compare the relationship between models and prognosis and noticed that STRICOM was significantly associated with worse outcomes across all cohorts compared with other models (Fig. 4A). Furthermore, when comparing the C-index, STRICOM displayed more excellent performance than most models in almost all cohorts (Fig. 4B). Similarly, 1-year, 3-year and 5-year AUC of STRICOM also ranked among the best across almost all cohorts compared with other models (Fig. 4C, Supplementary Fig. S2A-B). Collectively, these results suggested that STRICOM had better extrapolation potential and could be conveniently compared with other models in Mime.

3.5. Depicting the microenvironment and genome landscape shaped by STRICOM

In order to facilitate downstream analysis for users after establishing a prognostic model, Mime integrated the immune infiltration and tumor microenvironment signatures from R packages immunedeconv [37,38] and IOBR [39], allowing users to visualize results quickly. Through tumor microenvironment analysis, we observed that, in both TCGA and CGGA cohorts, the immune infiltration scores were higher in the high-risk group compared to the low-risk group for STRICOM (Fig. 5A-B). Additionally, in the TCGA-Glioma cohort, the expression of many important immune-related genes, such as CXCL10, CD276, and TNFRSF14, was highly correlated with the risk score of the STRICOM model (Fig. 5C). Furthermore, we found a significant correlation between the high-risk group based on STRICOM and genomic features such as higher loss of heterozygosity, CNA alteration fraction, homologous recombination deficiency (HRD) score, non-silent mutations, and aneuploidy score, indicating a higher level of genomic instability in the high-risk group (Fig. 5D-F). These results might explain, to some extent, why there is an apparent stratification in the prognosis of patients based on scores calculated by STRICOM.

3.6. Development of predictive models for therapeutic response by Mime

Several clinical trials of immune checkpoint inhibitor treatment have been ongoing in various cancer types. However, only a relatively small proportion of patients respond to it [40]. Here, we pooled 18 cohorts in which patients received anti-PD(L)− 1 or anti-CTLA4 therapy, with a

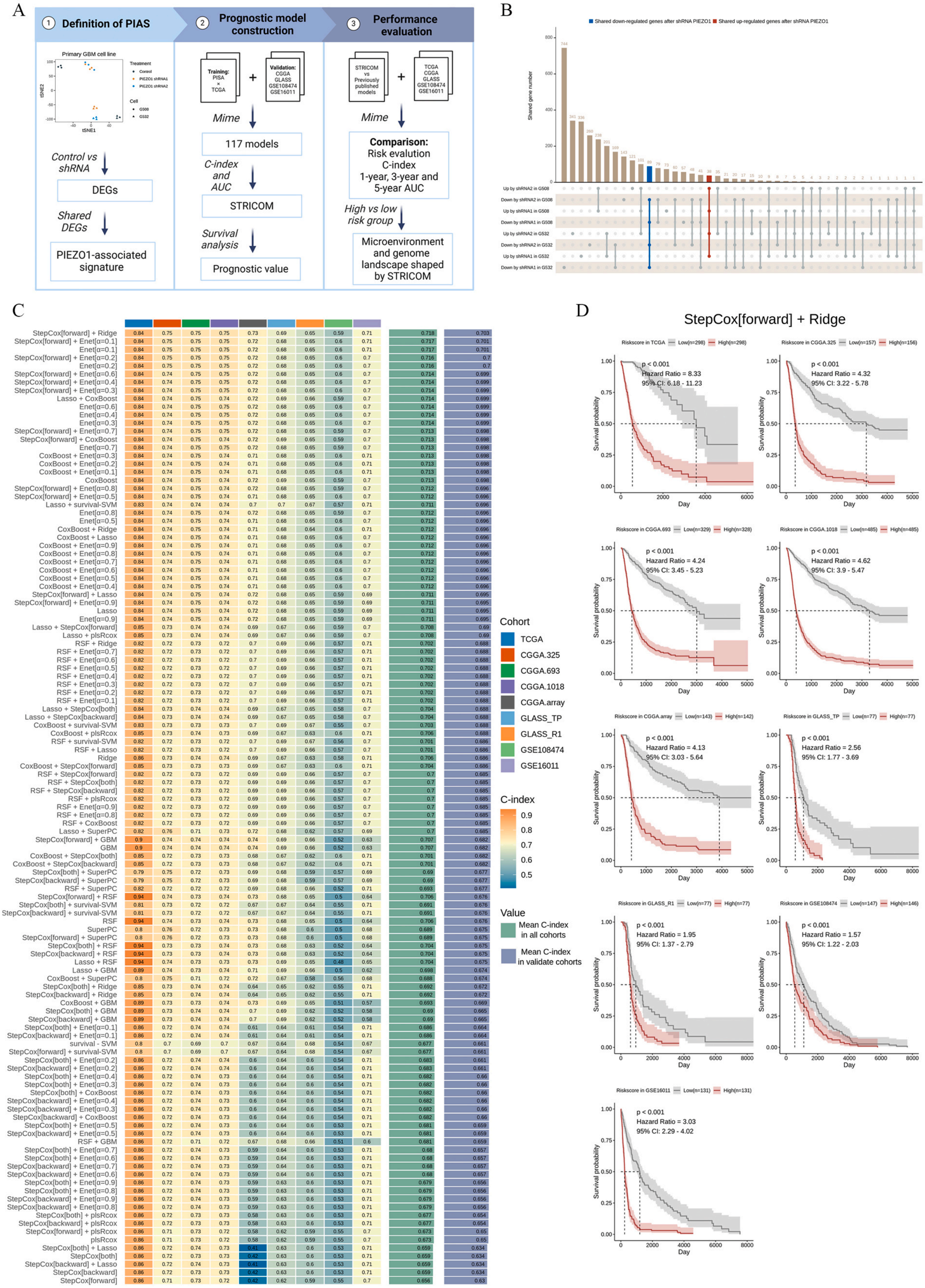


Fig. 2. Construction of prognostic models based on PIEZO1-associated signature. A. A workflow about the construction of prognostic models based on PIEZO1-associated signature. B. DEGs identified between control and PIEZO1 knockdown condition in G508 and G532. Top histogram: number of DEGs intersected in multiple conditions. C. C-index of each model among different cohorts sorted by the average of C-index in validation cohorts. D. The relation between risk score calculated by StepCox[forward]-Ridge combined model and outcome of patients in different cohorts.

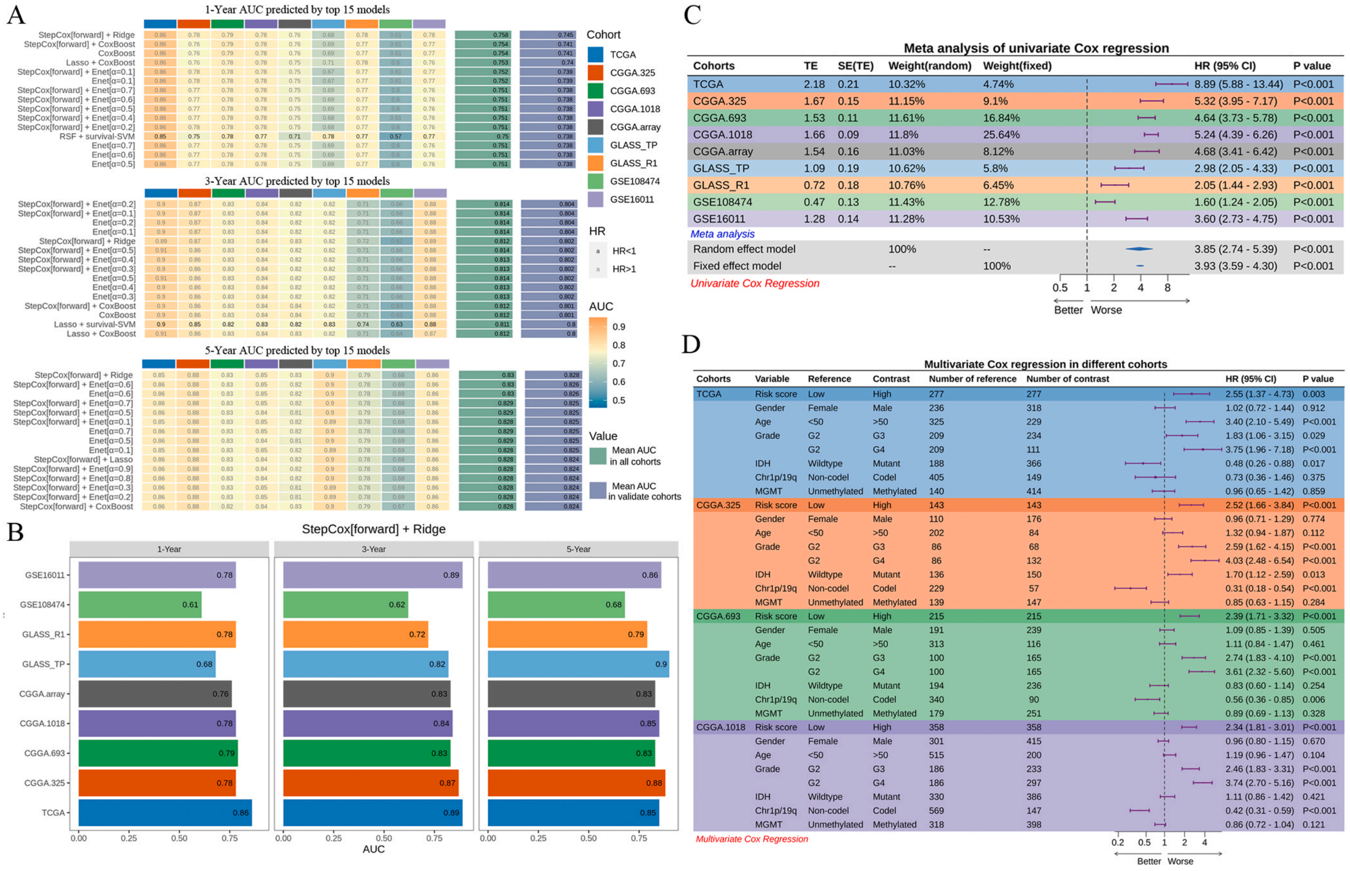


Fig. 3. Performance of prognostic models. A. 1-year, 3-year and 5-year AUC of top 15 models among different cohorts sorted by the average AUC in validation cohorts. The black font numbers indicate that the risk score calculated by this model predicted a better outcome in the corresponding cohort, otherwise indicate a worse outcome prediction. B. 1-year, 3-year and 5-year AUC of StepCox[forward]-Ridge combined model among different cohorts. C. Meta-analysis of univariate cox result of StepCox[forward]-Ridge combined model among different cohorts. D. Multivariate cox result of StepCox[forward]-Ridge combined model in four independent cohorts.

total of 1042 patients divided into a training set (70 %) and a validation set (30 %) similar to the process of previous studies [41]. Next, we provided PIAS to Mime in order to construct predicting models by using seven different machine learning algorithms (Fig. 6A). Notably, the AUC of Adaptive Boosting model (adaboost) achieved 1 in training set and 0.674 in validation set, which were better than other models (Fig. 6B). In addition, the performance of our developed model was more excellent than other previously published ICI-related models when comparing AUC (Fig. 6C). Taken together, Mime covered the main models for users to analyze the potentials of specific signatures comprehensively.

3.7. Identification of critical genes via Mime

To investigate the potential genes for in-depth study, PIAS and TCGA glioma cohort were also provided to Mime for core feature selection by different algorithms and we selected one of the top featured genes for further analysis (Fig. 7A). Most of the top-selected genes (Fig. 7B), such as AQP1, TOP2A and GJB2, are well-known as crucial targets in various diseases [42–44]. Intriguingly, SDC1, a member of the syndecan proteoglycan family also named CD138, had been reported to enhance the radioresistance of glioblastoma by influencing the fusion of autophagosomes with lysosomes [45,46]. Thus, we chose SDC1 as an example to further illustrate its potential role in glioma. Indeed, the expression of SDC1 was reduced when PIEZO1 was knocked down in both G508 and G532 (Fig. 7C). Consistent with our findings, public datasets consisting of TCGA, CGGA, GLASS, GSE108474 and GSE16011, also showed a significant positive correlation between PIEZO1 and SDC1 (Fig. 7D). Besides, high expression of SDC1 was significantly associated with higher grade, malignant histology, IDH wild type, 1p/19q non-co-deletion and mesenchymal subtype in glioma cohorts (Fig. 7E). In order to demonstrate the biological mechanisms involved in SDC1, we separated patients into high-expression and low-expression groups based on the median expression level of SDC1 for the identification of DEGs. Then, DEGs alone in four independent datasets presented in Fig. 7E were intersected to obtain SDC1-regulated genes, which were used to perform GO enrichment analysis (Fig. 7A). In total, there were 636 shared upregulated genes and 295 shared downregulated genes in SDC1 high-expression group. Enrichment network suggested that upregulated genes were associated with the regulation of cell cycle, cytoskeleton organization, extracellular matrix organization, cell-substrate adhesion and cellular response to stimulus (Fig. 7F). In contrast, downregulated genes were associated with neurotransmitter signaling and synaptic structure (Fig. 7F). These changes in biological processes associated with SDC1 are consistent with the known roles of PIEZO1 in regulating glioma aggression [35]. Collectively, Mime-identified genes show high potential as a target in source disease.

3.8. A web application for data visualization of Mime

Mime offers a variety of visualizations for researchers to explore intricate connections between biology and computational models (Fig. 1). These visualizations are also organized into four modules (“Data Upload”, “Prognosis”, “Binary Classification” and “Core Feature Selection”) in a shiny application (https://shiny.pedharmony.ac.cn/mimevis/) where a wet-bench scientist with little computational programming background can feel comfortable to explore their data. The

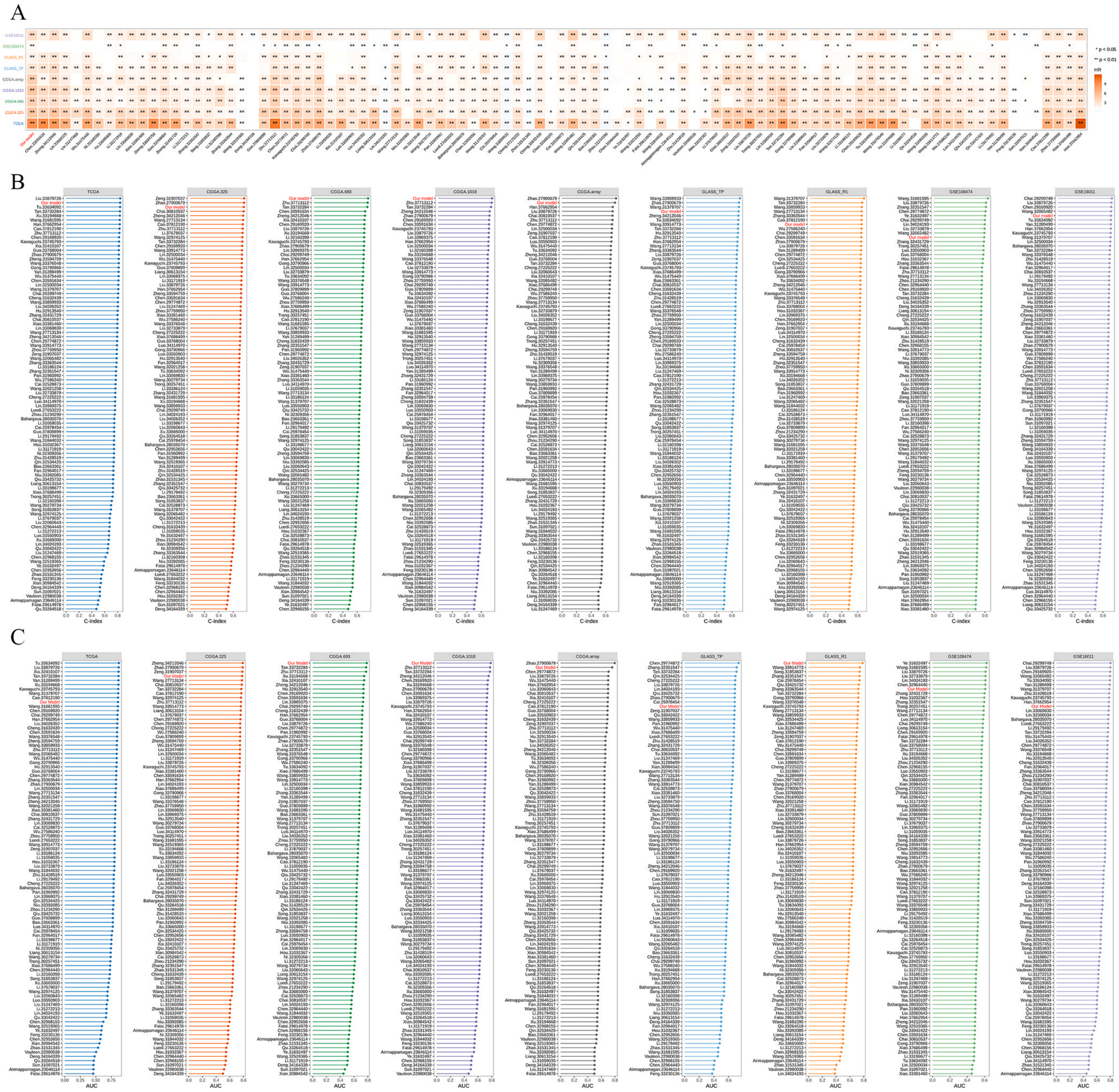


Fig. 4. Comparison with previously established models in glioma. A. HR of StepCox[forward]-Ridge combined model and 95 published models across 9 cohorts. B. C-index of StepCox[forward]-Ridge combined model and 95 published models across 9 cohorts. C. 1-year AUC predicted by StepCox[forward]-Ridge combined model and 95 published models across 9 cohorts.

well-established user guide of shiny application is illustrated in Fig. 8, outlining the fundamental exploration when using this platform. It typically consists of four steps: (i) users need to upload the local outputs of Mime which are saved as RDS file, (ii) choose corresponding type of module for further analysis, (iii) adjust various parameters such as dataset, model name, cut-off value and top number, (iv) click the Submit button and Download button to obtain visualization results. Due to the limited amount of computational power and acceptable file size for the webserver, we still recommend that users install Mime and run it in local.

4. Discussion

As the application of machine learning expands across various domains such as weather prediction and recommendation engines, an increasing number of researchers recognize the potential to apply novel machine learning methods developed in other fields to the medical field [47–49]. This has led to the emergence of a plethora of machine learning-based prognostic models. While researchers have the flexibility to choose a particular class of machine learning algorithm from different perspectives, the comparison and selection of the optimal model remain a complex task due to the diverse requirements in file formats, operating environments, and other factors across different machine learning methods. It is challenging for a single user to comprehensively compare the effectiveness of various algorithms on the same training dataset and validation datasets to obtain an advanced prognostic model.

Recently, more researchers have begun using combinations of various machine learning algorithms for more accurate and stable model

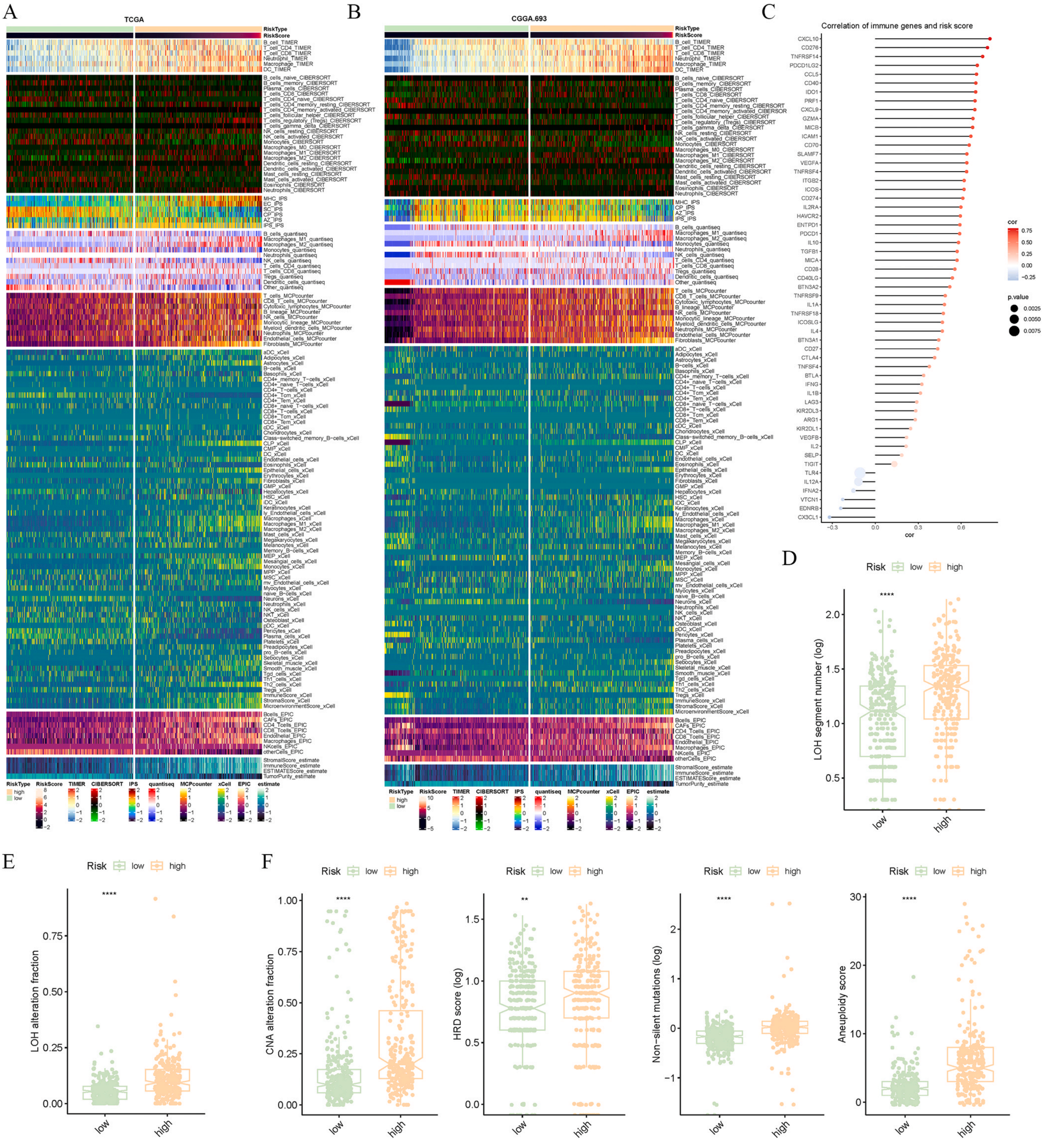


Fig. 5. Correlation between risk score and immune or genome signatures. A. Relationship between risk score calculated by StepCox[forward]-Ridge combined model and microenvironment signatures deconvoluted by different methods in TCGA glioma cohort. Method IPS was from package IOBR, while other methods were from package immunedeconv. B. Same as A but in CGGA.693 cohort. C. Correlation between risk score and various immune genes. D-F. Correlation between risk score and loss of heterozygosity segment number (D), loss of heterozygosity alteration fraction (E), CNA alteration fraction, homologous recombination deficiency score, non-silent mutations, and aneuploidy score (F), respectively. * $P < 0.05$, * * $P < 0.01$, * ** $P < 0.001$, * ** * $P < 0.0001$.

construction [6–8]. To facilitate a more straightforward evaluation of the strengths and weaknesses of different models, we developed the R package Mime to simplify the process of building machine learning ensemble models from transcriptome data. For example, we selected the PIAS in glioma for prognostic model construction. After comparing the predictive performance of 117 different machine learning models, we identified the outstanding model STRICOM. Simultaneously, we reviewed 95 glioma prognostic models published in recent years, and in comparison, STRICOM remained one of the most superior models. This example not only highlights the practicality of Mime but also underscores the tremendous potential of machine learning in prognosis research. Besides, Mime integrates core functions such as response

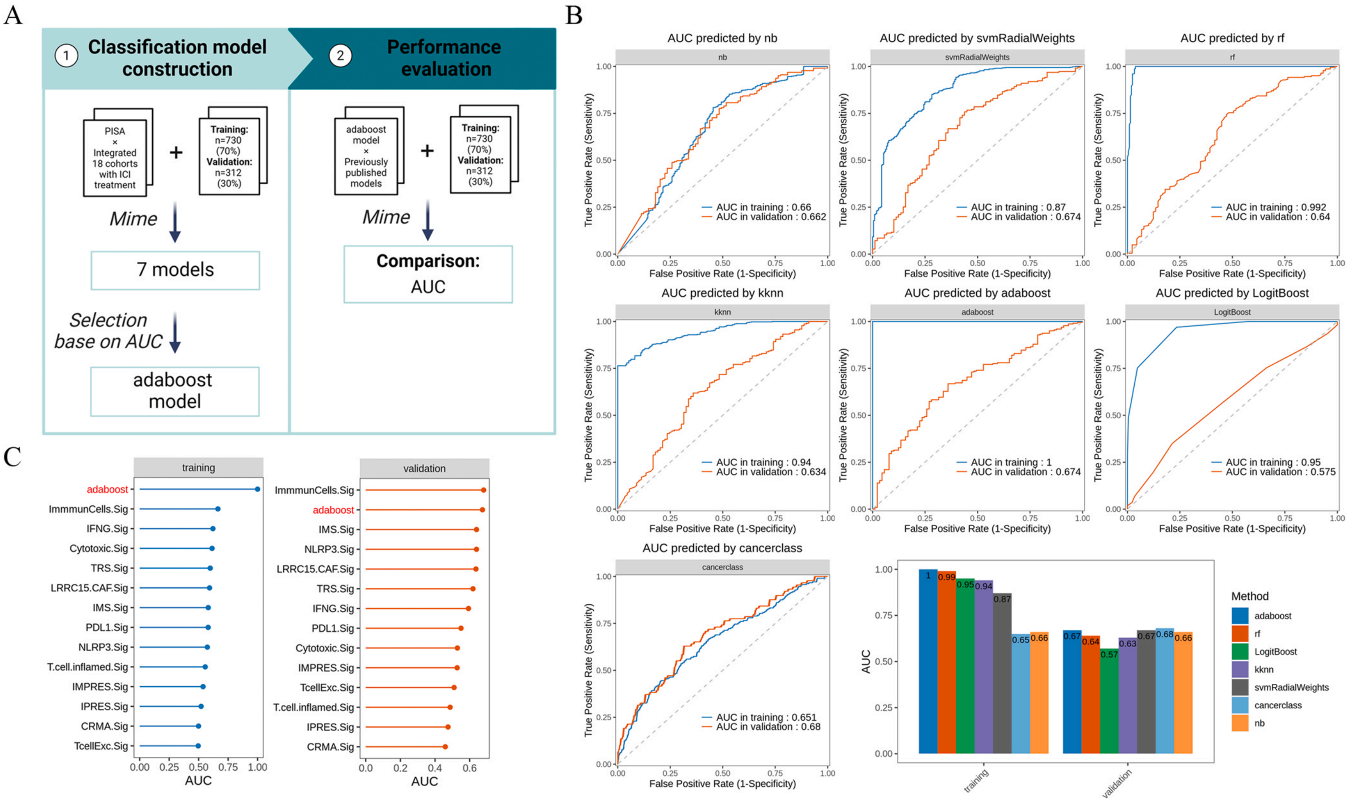


Fig. 6. Construction of predictive models for immunotherapy benefits. A. A workflow about the construction of predictive models for immunotherapy benefits based on PIEZO1-associated signature. B. ROC curves of each model to predict the benefits of immunotherapy in training and validation datasets. Bottom right histogram: The distribution of AUC predicted by 7 machine-learning models in the training and validation dataset. C. AUC predicted by adaboost model and 13 published models across training and validation datasets.

model construction, feature selection, immune infiltration analysis and data visualization, enhancing a deep understanding of models for researchers and revealing the crucial functions these high-value potential genes may play in diseases. Based on these tentative explorations, users can further perform downstream analyses to validate corresponding biological functions and phenotypes for specific features.

In general, Mime is a state-of-the-art tool for comprehensive and convenient machine-learning model construction. It emphasizes building the most commonly used models at once with user- friendly visual representations, eliminating the mastery for different machine-learning algorithms. Besides, Mime-filtered variables from large pools associated with disease progression can be helpful for both researchers and clinical practitioners to uncover novel insights. Mime also has the potential to guide biomarker development and contribute to personalized medicine by bridging the gap between computational biology and cancer research. Although we used transcriptional data as an example to demonstrate the applications of Mime, it can support other input data such as proteomic data, radiomic data, clinical biological indicator data and other numerical matrices with clinical information of patients.

Mime has some limitations. Firstly, Mime integrates a vast number of machine learning algorithms, and due to variations in users' operating environments, Mime's computational speed may be slow in certain extreme cases, with high computational resource requirements on large-scale datasets. Secondly, Mime's performance is highly dependent on the quality and consistency of user input data. Poor data quality or non-compliance with standards may impact the accuracy and reliability of the model. Our future research may include further improving Mime's computational speed and addressing compatibility issues in complex operating scenarios. Thirdly, there are two types of parameters in machine learning algorithms including parameters learned from training data and hyperparameters set before starting the learning process. Hyperparameters often define higher-level concepts about the model, such as model complexity or learning ability. Thus, additional hyperparameter tuning approaches are necessary for Mime to improve its ability in the future. Furthermore, AUC as an evaluation metric may be too optimistic in the case of unbalanced data sets. Therefore, it might be more appropriate to consider other metrics for handling imbalanced datasets, such as error rate, precision and recall. Finally, as new machine learning models are developed in the future, we will incorporate additional models in the next version of Mime to cater to the needs of users. In order to promote the use of Mime and ensure software maintenance, we will constantly pay attention to the problems encountered by users as reported in the Issues section on GitHub and solve these problems rapidly. Additionally, we will test Mime periodically to make sure it works well in different versions of R when some related packages or R platform updates are available.

5. Conclusion

In summary, our study provides a comprehensive and powerful open-source R package, Mime, making it easier for researchers to integrate various machine learning algorithms to better analyze specific signatures. We hope that Mime can enable more researchers to create stable, reliable, and robust predictive models using machine learning, providing deeper insights into the field.

Supplementary data

Supplementary data related to this article are available online.

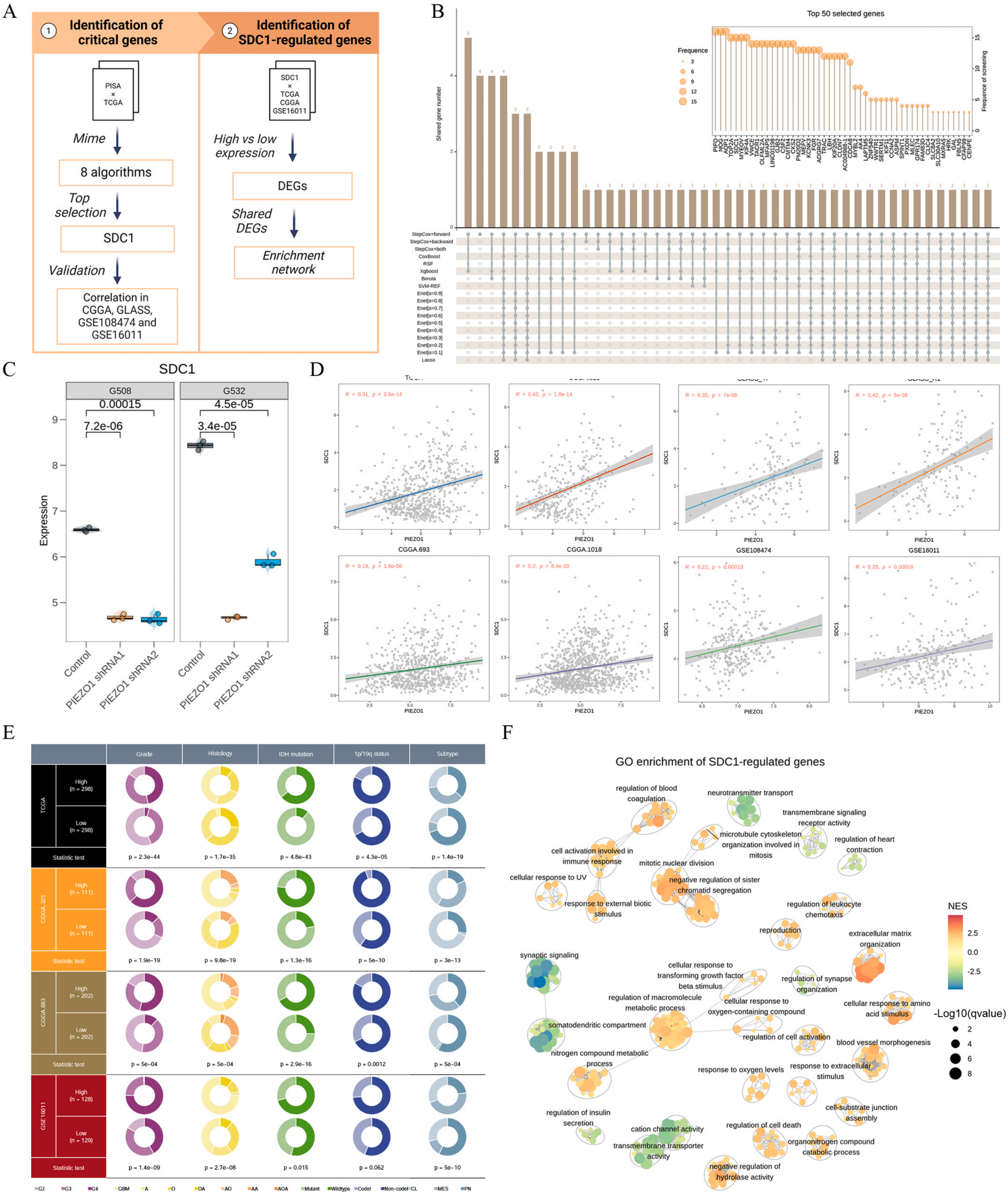
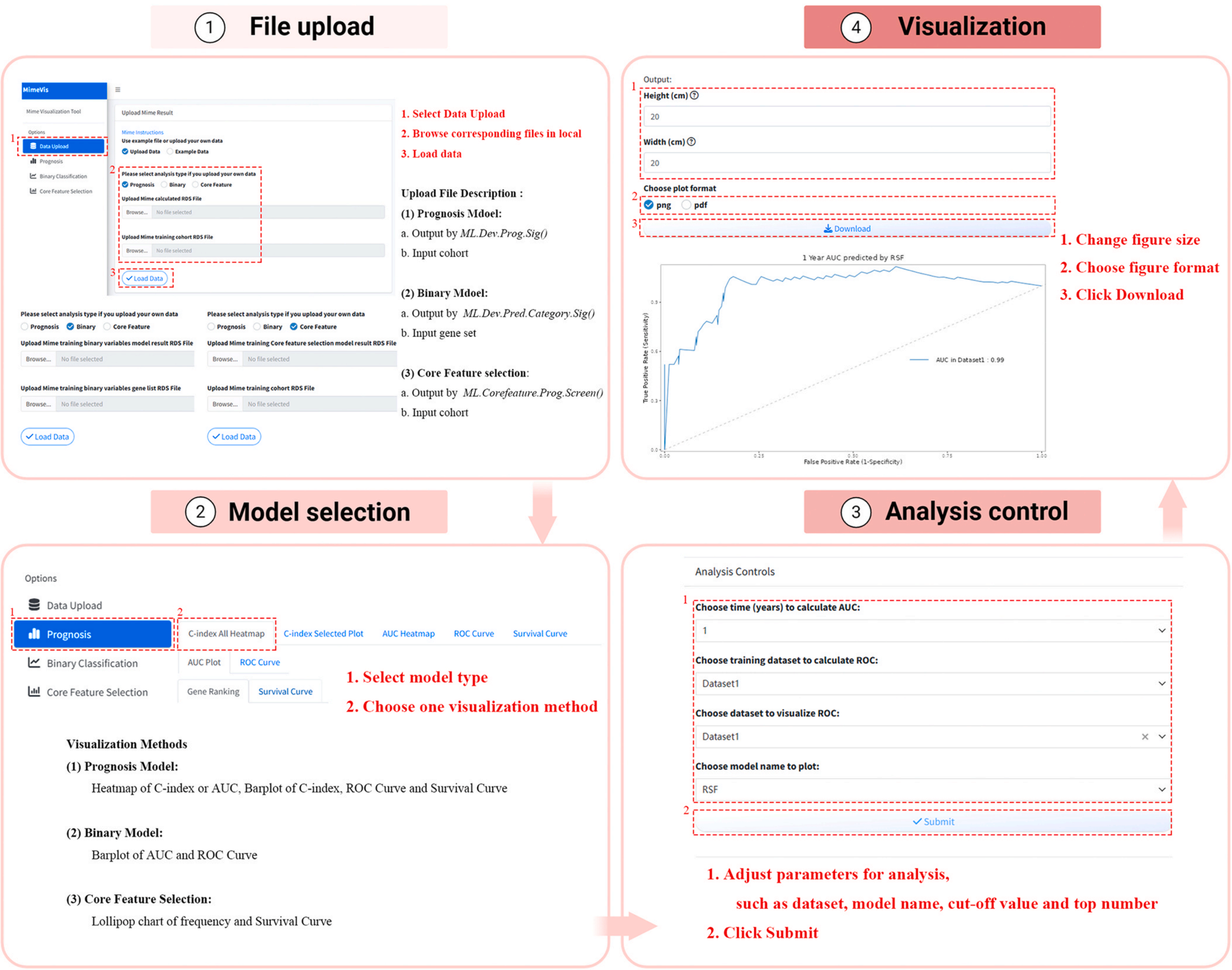


Fig. 7. Characteristic of SDC1 in glioma. A. A workflow of the identification of critical genes based on PIEZO1-associated signature. B. Prognosis-associated genes selected by different machine-learning algorithms. Top histogram: number of genes intersected by multiple models. Top right chart: Frequency of genes selected by different models. C. Expression level of SDC1 between control and PIEZO1 knockdown condition in G508 and G532. Statistic test: t-test. D. Pearson correlation between PIEZO1 and SDC1 in different cohorts. E. The relationship between expression of SDC1 and other clinical features (Grade, Histology, IDH mutation, 1p/19q status and transcriptional subtypes) in TCGA, CGGA.325, CGGA.693 and GSE16011 cohorts. Statistic test: chi-square test. F. GO enrichment of SDC1-regulated genes. Normal enrichment score (NES) > 0 indicated up-regulated processes in SDC1 high-expression group otherwise down-regulated processes in SDC1 high-expression group.



① File upload

1. Select Data Upload
2. Browse corresponding files in local
3. Load data

Upload File Description :

(1) Prognosis Mdoel:
a. Output by *ML.Dev.Prog.Sig()*
b. Input cohort

(2) Binary Mdoel:
a. Output by *ML.Dev.Pred.Category.Sig()*
b. Input gene set

(3) Core Feature selection:
a. Output by *ML.Corefeature.Prog.Screen()*
b. Input cohort

Please select analysis type if you upload your own data: ● Prognosis ○ Binary ○ Core Feature
Upload Mime calculated RDS File: Browse... No file selected
Upload Mime training cohort RDS File: Browse... No file selected

Please select analysis type if you upload your own data: ○ Prognosis ● Binary ○ Core Feature
Upload Mime training binary variables model result RDS File: Browse... No file selected
Upload Mime training binary variables gene list RDS File: Browse... No file selected

Please select analysis type if you upload your own data: ○ Prognosis ○ Binary ● Core Feature
Upload Mime training Core feature selection model result RDS File: Browse... No file selected
Upload Mime training cohort RDS File: Browse... No file selected

② Model selection

1. Select model type
2. Choose one visualization method

Visualization Methods

(1) Prognosis Model:
Heatmap of C-index or AUC, Barplot of C-index, ROC Curve and Survival Curve

(2) Binary Model:
Barplot of AUC and ROC Curve

(3) Core Feature Selection:
Lollipop chart of frequency and Survival Curve

③ Analysis control

Analysis Controls
Choose time (years) to calculate AUC: 1
Choose training dataset to calculate ROC: Dataset1
Choose dataset to visualize ROC: Dataset1
Choose model name to plot: RSF

1. Adjust parameters for analysis,
such as dataset, model name, cut-off value and top number
2. Click Submit

④ Visualization

Output:
Height (cm): 20
Width (cm): 20
Choose plot format: ● png ○ pdf

1. Change figure size
2. Choose figure format
3. Click Download

Fig. 8. A fundamental exploration within shiny application of Mime. Users can interactively obtain visualization results through uploading the local outputs of Mime.

Ethics approval and consent to participate

Not applicable.

Funding

This work is supported by the National Natural Science Foundation of China (Grant No. 82270825).

CRediT authorship contribution statement

Yihao Zhang: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Wang Li:** Validation, Investigation. **Kang Peng:** Validation, Investigation. **Siyi Wanggou:** Writing – review & editing, Investigation, Funding acquisition, Conceptualization. **Hongwei Liu:** Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Xuejun Li:** Writing – review & editing, Investigation, Funding acquisition, Conceptualization. **Wei Zhang:** Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis. **Luohuan Dai:** Validation, Investigation. **Zhouyang Pan:** Validation, Investigation. **Hongyi Liu:** Validation, Investigation. **Yi Xiong:** Validation, Investigation. **Abraham Ayodeji Adegboro:** Writing – original draft, Validation. **Deborah Oluwatosin Fasoranti:** Writing – original draft, Validation.

Declaration of Competing Interest

All authors declared that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Data Availability

Data used to support our work are all publicly available, and the data information can be acquired from the Supplementary Materials. Raw codes for the package are available on GitHub (https://github.com/l-magnificence/Mime). The shiny application of Mime is deployed on our team's server (https://shiny.pedharmony.ac.cn/mimevis/).

Acknowledgments

We thank all participants and investigators involved in the data production including TCGA, CGGA, SRA database, and GEO database.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.06.035.

References

[1] Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies. Mol Cell 2015;58(4):586–97.

[2] Adam G, Rampášek L, Safikhani Z, Smirnov P, Haibe-Kains B, Goldenberg A. Machine learning approaches to drug response prediction: challenges and recent progress. NPJ Precis Oncol 2020;4:19.

[3] Ding L, Bailey MH, Porta-Pardo E, Thorsson V, Colaprico A, Bertrand D, et al. Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. Cell 2018;173(2):305–20. e310.

[4] Hanahan D. Hallmarks of cancer: new dimensions. Cancer Discov 2022;12(1):31–46.

[5] Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2015;13:8–17.

[6] Liu Z, Liu L, Weng S, Guo C, Dang Q, Xu H, et al. Machine learning-based integration develops an immune-derived lncRNA signature for improving outcomes in colorectal cancer. Nat Commun 2022;13(1):816.

[7] Sundar R, Barr Kumarakulasinghe N, Huak Chan Y, Yoshida K, Yoshikawa T, Miyagi Y, et al. Machine-learning model derived gene signature predictive of paclitaxel survival benefit in gastric cancer: results from the randomised phase III SAMIT trial. Gut 2022;71(4):676–85.

[8] Zhang Z, Chen L, Chen H, Zhao J, Li K, Sun J, et al. Pan-cancer landscape of T-cell exhaustion heterogeneity within the tumor microenvironment revealed a progressive roadmap of hierarchical dysfunction associated with prognosis and therapeutic efficacy. EBioMedicine 2022;83:104207.

[9] Zhang W, Dang R, Liu H, Dai L, Liu H, Adegboro AA, et al. Machine learning-based investigation of regulated cell death for predicting prognosis and immunotherapy response in glioma patients. Sci Rep 2024;14(1):4173.

[10] Zhang W, Zhu Y, Liu H, Zhang Y, Liu H, Adegboro AA, et al. Pan-cancer evaluation of regulated cell death to predict overall survival and immune checkpoint inhibitor response. NPJ Precis Oncol 2024;8(1):77.

[11] Cai Z, Poulos RC, Liu J, Zhong Q. Machine learning for multi-omics data integration in cancer. iScience 2022;25(2):103798.

[12] Hindocha S, Charlton TG, Linton-Reid K, Hunter B, Chan C, Ahmed M, et al. A comparison of machine learning methods for predicting recurrence and death after curative-intent radiotherapy for non-small cell lung cancer: development and validation of multivariable clinical prediction models. EBioMedicine 2022;77:103911.

[13] Binder H, Allignol A, Schumacher M, Beyersmann J. Boosting for high-dimensional time-to-event data with competing risks. Bioinforma (Oxf, Engl) 2009;25(7):890–6.

[14] Bair E, Tibshirani R. Semi-supervised methods to predict patient survival from gene expression data. PLoS Biol 2004;2(4). E108.

[15] Wang S, Xiong Y, Zhao L, Gu K, Li Y, Zhao F, et al. UCSCXenaShiny: an R/CRAN package for interactive analysis of UCSC Xena data. Bioinforma (Oxf, Engl) 2022;38(2):527–9.

[16] Vougas K, Sakellaropoulos T, Kotsinas A, Foukas GP, Ntargaras A, Koinis F, et al. Machine learning and data mining frameworks for predicting drug response in cancer: an overview and a novel in silico screening process based on association rule mining. Pharm Ther 2019;203:107395.

[17] Huang X, Hu M, Sun T, Li J, Zhou Y, Yan Y, et al. Multi-kingdom gut microbiota analyses define bacterial-fungal interplay and microbial markers of pan-cancer immunotherapy across cohorts. Cell Host Microbe 2023;31(11):1930–43. e1934.

[18] Zhang J, Ma G, Peng S, Hou J, Xu R, Luo L, et al. Risk factors and predictive models for peripherally inserted central catheter unplanned extubation in patients with cancer: prospective, machine learning study. J Med Internet Res 2023;25:e49016.

[19] Zhou Y, Smith J, Keerthi D, Li C, Sun Y, Mothi SS, et al. Longitudinal clinical data improves survival prediction after hematopoietic cell transplantation using machine learning. Blood Adv 2023.

[20] Duerr R, Dimartino D, Marier C, Zappile P, Wang G, Francois F, et al. Selective adaptation of SARS-CoV-2 Omicron under booster vaccine pressure: a multicentre observational study. EBioMedicine 2023;97:104843.

[21] Granata V, Fusco R, De Muzio F, Brunese MC, Setola SV, Ottaiano A, et al. Radiomics and machine learning analysis by computed tomography and magnetic resonance imaging in colorectal liver metastases prognostic assessment. Radio Med 2023;128(11):1310–32.

[22] Auslander N, Zhang G, Lee JS, Frederick DT, Miao B, Moll T, et al. Robust prediction of response to immune checkpoint blockade therapy in metastatic melanoma. Nat Med 2018;24(10):1545–9.

[23] Ayers M, Lunceford J, Nebozhyn M, Murphy E, Loboda A, Kaufman DR, et al. IFN-gamma-related mRNA profile predicts clinical response to PD-1 blockade. J Clin Invest 2017;127(8):2930–40.

[24] Cui C, Xu C, Yang W, Chi Z, Sheng X, Si L, et al. Ratio of the interferon-gamma signature to the immunosuppression signature predicts anti-PD-1 therapy response in melanoma. NPJ Genom Med 2021;6(1):7.

[25] Dominguez CX, Muller S, Keerthivasan S, Koeppen H, Hung J, Gierke S, et al. **Single-cell RNA sequencing reveals stromal evolution into LRRC15(+) myofibroblasts as a determinant of patient response to cancer immunotherapy**. Cancer Discov 2020;10(2):232–53.

[26] Jerby-Arnon L, Shah P, Cuoco MS, Rodman C, Su MJ, Melms JC, et al. A cancer cell program promotes T cell exclusion and resistance to checkpoint blockade. Cell 2018;175(4):984–97. e924.

[27] Ju M, Bi J, Wei Q, Jiang L, Guan Q, Zhang M, et al. Pan-cancer analysis of NLRP3 inflammasome with potential implications in prognosis and immunotherapy in human cancer. Brief Bioinform 2021;22(4).

[28] Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. Cell 2015;160(1-2):48–61.

[29] Shukla SA, Bachireddy P, Schilling B, Galonska C, Zhan Q, Bango C, et al. Cancer-germline antigen expression discriminates clinical outcome to CTLA-4 blockade. Cell 2018;173(3):624–33. e628.

[30] Thompson JC, Hwang WT, Davis C, Deshpande C, Jeffries S, Rajpurohit Y, et al. Gene signatures of tumor inflammation and epithelial-to-mesenchymal transition (EMT) predict responses to immune checkpoint blockade in lung cancer with high accuracy. Lung Cancer 2020;139:1–8.

[31] Topalian SL, Hodi FS, Brahmer JR, Gettinger SN, Smith DC, McDermott DF, et al. Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. N Engl J Med 2012;366(26):2443–54.

[32] Xiong D, Wang Y, You M. A gene expression signature of TREM2(hi) macrophages and gammadelta T cells predicts immunotherapy response. Nat Commun 2020;11(1):5084.

[33] Yan M, Hu J, Ping Y, Xu L, Liao G, Jiang Z, et al. Single-cell transcriptomic analysis reveals a tumor-reactive T cell signature associated with clinical outcome and immunotherapy response in melanoma. Front Immunol 2021;12:758288.

[34] Bagaev A, Kotlov N, Nomie K, Svekolkin V, Gafurov A, Isaeva O, et al. Conserved pan-cancer microenvironment subtypes predict response to immunotherapy. Cancer Cell 2021;39(6):845–65. e847.

[35] Chen X, Wanggou S, Bodalia A, Zhu M, Dong W, Fan JJ, et al. A feedforward mechanism mediated by mechanosensitive ion channel PIEZO1 and tissue mechanics promotes glioma aggression. Neuron 2018;100(4):799–815. e797.

[36] Zhang N, Zhang H, Wu W, Zhou R, Li S, Wang Z, et al. Machine learning-based identification of tumor-infiltrating immune cell-associated lncRNAs for improving outcomes and immunotherapy responses in patients with low-grade glioma. Theranostics 2022;12(13):5931–48.

[37] Sturm G, Finotello F, List M. Immunedeconv: an R package for unified access to computational methods for estimating immune cell fractions from bulk RNA-sequencing data. Methods Mol Biol 2020;2120:223–32.

[38] Sturm G, Finotello F, Petitprez F, Zhang JD, Baumbach J, Fridman WH, et al. Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. Bioinformatics 2019;35(14):i436–45.

[39] Zeng D, Ye Z, Shen R, Yu G, Wu J, Xiong Y, et al. IOBR: multi-omics immuno-oncology biological research to decode tumor microenvironment and signatures. Front Immunol 2021;12:687975.

[40] Sharma P, Siddiqui BA, Anandhan S, Yadav SS, Subudhi SK, Gao J, et al. The next decade of immune checkpoint therapy. Cancer Discov 2021;11(4):838–57.

[41] Zhang Z, Wang ZX, Chen YX, Wu HX, Yin L, Zhao Q, et al. Integrated analysis of single-cell and bulk RNA sequencing data reveals a pan-cancer stemness signature predicting immunotherapy response. Genome Med 2022;14(1):45.

[42] Guo Z, Zhang H, Liu X, Zhao Y, Chen Y, Jin J, et al. Water channel protein AQP1 in cytoplasm is a critical factor in breast cancer local invasion. J Exp Clin Cancer Res: CR 2023;42(1):49.

[43] Uusküla-Reimand L, Wilson MD. Untangling the roles of TOP2A and TOP2B in transcription and cancer. Sci Adv 2022;8(44):eadd4920.

[44] Kersbergen CJ, Babola TA, Kanold PO, Bergles DE. Preservation of developmental spontaneous activity enables early auditory system maturation in deaf mice. PLoS Biol 2023;21(6):e3002160.

[45] Zeng L, Zheng W, Liu X, Zhou Y, Jin X, Xiao Y, et al. SDC1-TGM2-FLOT1-BHMT complex determines radiosensitivity of glioblastoma by influencing the fusion of autophagosomes with lysosomes. Theranostics 2023;13(11):3725–43.

[46] Zheng W, Chen Q, Liu H, Zeng L, Zhou Y, Liu X, et al. SDC1-dependent TGM2 determines radiosensitivity in glioblastoma by coordinating EPG5-mediated fusion of autophagosomes with lysosomes. Autophagy 2023;19(3):839–57.

[47] Hashizume T, Ying BW. Challenges in developing cell culture media using machine learning. Biotechnol Adv 2023:108293.

[48] Roisman LC, Kian W, Anoze A, Fuchs V, Spector M, Steiner R, et al. Radiological artificial intelligence - predicting personalized immunotherapy outcomes in lung cancer. NPJ Precis Oncol 2023;7(1):125.

[49] Kim HJ, Gong EJ, Bang CS. Application of machine learning based on structured medical data in gastroenterology. Biomim (Basel, Switz) 2023;8(7).