SurProGenes: Survival Risk-Ordered Representation of Cancer Patients and Genes for the Identification of Prognostic Genes

Junetae Kim [*1] **Kyoungsuk Park** [*1] **Hanseok Jeong** [2] **Youngwook Kim** [3] **Jeongseon Kim** [1] **Sun-Young Kim** [1]

Abstract

Identifying prognostic genes associated with patient survival is an important goal in cancer genomics, as this information could inform treatment approaches and improve patient outcomes. However, the identification of prognostic genes is complicated by the high dimensionality of genetic data, which makes their identification computationally intensive. Furthermore, most cancer genomics studies lack appropriate low-risk groups against which to compare. To address these issues, we present a framework that identifies candidate prognostic genes by integrating representation learning and statistical analysis approaches. Specifically, we propose a collaborative filtering-derived mechanism to represent patients in order of their survival risk, facilitating their dichotomization. We also propose a mechanism that allows embedded gene vectors to be polarized on the extremities of, or centered on, both reference axes to facilitate recommendations. Restricting our analysis to a few representative genes within each cluster allowed for the efficient identification of prognostic genes. Finally, we demonstrate the potential of this proposed framework for identifying prognostic genes.

1. Introduction

Cancer remains a leading cause of death globally, and identifying prognostic biomarkers is an essential goal in the field of cancer genomics (Raman et al., 2019; Tran et al., 2022). The recent application of genomic, transcriptomic, and proteomic technologies to the field has resulted in the development of various genetic biomarkers associated with increased mortality (Campbell et al., 2020; Docking et al., 2021; Sundar et al., 2022). Accordingly, a more comprehensive understanding of the risks associated with individual genes may inform treatment decisions and impact patient outcomes.

The identification of biomarkers is facilitated when there is a clear dichotomy between high- and low-risk groups (Raman et al., 2019). In practice, however, this dichotomy can be difficult to achieve, because most cancer genomics studies are performed exclusively on cancer populations and lack a low-risk comparison group (i.e., normal subjects) (Suntsova et al., 2019).

One approach to help rectify this limitation is to dichotomize cancer patients into short- and long-term survivors based on their median gene expression (Raman et al., 2019; Alves et al., 2021), or the median hazard ratio (HR) of the Cox proportional-hazards (CPH) model (Cai et al., 2019; Docking et al., 2021). However, this approach is limited as the groups are divided equally under the strong assumption that the dichotomy exists at these median values (DeCoster et al., 2011; Altman & Royston, 2006). Other researchers have attempted to cluster genes according to their molecular patterns and then explore differences in the survival risks of these gene clusters (Beer et al., 2002; Witkiewicz et al., 2015). This approach is powerful, as it can group numerous gene features to facilitate interpretation (Campbell et al., 2020); however, its drawbacks are that it stratifies patients based exclusively on genetic patterns while ignoring differences in survival risk in the underlying groups, and explores the associations between gene clusters rather than those between individual genes and mortality.

Here, we propose a new framework[1] that addresses these issues. We dichotomize patients based on their survival risk while recommending individual prognostic genes expressed at statistically different levels between the resulting groups.

[*]Equal contribution [1]Department of Cancer AI & Digital Health, Graduate School of Cancer Science and Policy, National Cancer Center, Goyang, Republic of Korea [2]Department of Electrical and Computer Engineering, University of Seoul, Seoul, Republic of Korea [3]Department of Cancer Biomedical Science, Graduate School of Cancer Science and Policy, National Cancer Center, Goyang, Republic of Korea. Correspondence to: Junetae Kim <lyjune0070@gmail.com>, Kyoungsuk Park <bluemk00@gmail.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

[1]The code is available at `https://github.com/JunetaeKim/SurProGenes`.

The overall framework, from data preprocessing to the recommendation of prognostic genes, is depicted in Figure 4, and the highlights of our work are as follows:

Risk-ordered representation: We propose a novel mechanism to represent patient entities in order of risk in a vector space, based on the principle of collaborative filtering (CF) (Sarwar et al., 2001). This mechanism allows both the survival risk, inferred from the time-to-event data (Cox, 1972), and multiple gene expression patterns to be addressed during the embedding process.

Patient dichotomization: We developed a method to divide the patients into low- and high-risk groups based on risk differences, using the distances between the patient vectors.

Gene recommendation: We developed a mechanism to represent the genes, correlate them with survival risk, and cluster them. Moreover, we propose a mechanism for recommending prognostic genes that differ significantly between the dichotomous risk groups.

Evaluation: We evaluated the recommendations regarding patient dichotomization and the statistical differences in gene variables between the dichotomous groups. One challenge with the statistical evaluation of genomic data is that hypothesis testing for every individual gene is computationally heavy. Thus, to reduce computational demands and facilitate our evaluation, we focused on a subset of candidate genes near the centroids of their respective gene clusters.

Pan-cancer analysis: Pan-cancer analysis involves identifying common mutational patterns between different cancer types. It has been receiving increased attention as it provides a deeper understanding of cancer (Campbell et al., 2020). Following this trend, we simultaneously identified prognostic genes across multi-cancer cohorts.

The major contributions of this study are as follows:

- We developed a mechanism to represent and dichotomize patient entities in order of survival risk using time-to-event data. This mechanism can be extended to other CF models dealing with survival risk.
- We developed a mechanism to induce feature vectors (i.e., items) to be either polarized on the extremities of the two reference axes or centered on both of them. This mechanism facilitates the recommendation of features that differ significantly between the two groups (i.e., users).
- We developed a framework that incorporates representation learning into the task of defining test and comparison groups for statistical hypothesis testing.

2. Related works

Gene expression: Gene expression, the process by which the proteins that make up an organism are formed by genes, can be quantified by protein levels (Crick, 1970). Expression levels are influenced by various factors such as environmental factors, gene mutations, and molecules (Hallgrimsson & Hall, 2005). Normal and cancer cells show differences in expression; thus, studies have been conducted to identify genes related to cancer development and survival risk based on these differences (Raman et al., 2019; Alves et al., 2021; Cai et al., 2019; Docking et al., 2021). The expression differences between groups of interest are assessed using various statistical models (Robinson & Smyth, 2008).

Survival analysis: Survival analysis is a method of modeling the time remaining until death (Cox, 1972). The CPH model has been commonly employed for survival analysis in the field of cancer-genomics (Raman et al., 2019; Cox, 1972). The risk of $subject_i$ at time t can be modeled as $\lambda(t|X_i) = \lambda_0(t)\exp(X_i B)$, where X, B, and $\lambda_0(t)$ represent the covariates, coefficients, and baseline hazard, respectively. To estimate B and $\lambda_0(t)$, the likelihood function, which divides the risk of the subjects by the cumulative risk of the non-event group, is maximized. Moreover, time-to-event data, which consist of time lengths between observations and the start of follow-up (TL), and an indicator of whether the event occurred are required (Cox, 1972).

Recently, several studies have attempted to develop machine learning-based survival models using gene expression data. These studies involved performance comparisons between machine learning and conventional models (Ching et al., 2018), training methods with data from multiple cancer cohorts (Kuruc et al., 2022), processing methods for high-dimensional genetic data (Qiu et al., 2020), and feature discovery (Sundar et al., 2022). However, no attempt has been made to represent patient entities in order of their survival risk.

Collaborative filtering with side-information: As a representative recommendation algorithm, CF learns an association between two heterogeneous entities, such as a relationship between users and items (Koren et al., 2009). In recent years, various attempts have been made to introduce side-information into CF models to improve their recommendation performance. Representative examples of such side-information include genres in movie recommendation (Singh & Gordon, 2008), social relations in political-party prediction (Nickel et al., 2011), and topics in article recommendation (Wang & Blei, 2011). Among the various reported methods, those that are similar to this work use side-information for regularization. Specifically, they regularize model training so that the process of embedding the entities depends on the side-information (Dong et al., 2017). Despite best practices in these areas, no attempt has been

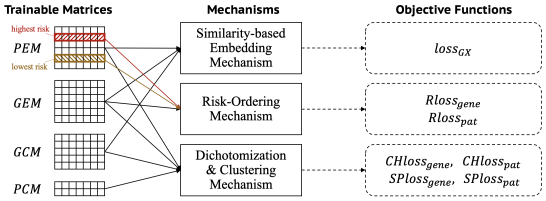


Figure 1. Outline of the proposed model.

made to incorporate survival risk, a key indicator in the medical field, into the patient embedding process.

Survival risk clustering: The methods for survival risk clustering can be divided into two broad streams. The first stream is characterized by obtaining survival risk strata based on given gene clusters without clustering the survival risks themselves (Beer et al., 2002; Witkiewicz et al., 2015). Recently, various machine learning algorithms have been developed for clustering genes while projecting them into a low-dimensional space (Karim et al., 2020).

The second stream, which is similar to our work, is characterized by clustering the survival risks themselves. Previous methods in this stream model the cumulative hazard function (CHF) as a mixture of the distribution components, $\sum_{k=1}^{K} P(z=k|x)\lambda(t|X)$ (Nagpal et al., 2021; Jeanselme et al., 2022); the individual hazard risks are assigned to certain clusters according to the mixture weight, $P(Z|x)$. Our study differs from these previous studies in two respects. In terms of purpose, previous studies focused on algorithm development for CHF clustering itself, while we established a methodological strategy for identifying statistically different features among the dichotomous groups discovered by the model. In terms of mechanism, our algorithm incorporates survival risk into the representations of the patients and genes, which has not been attempted previously, allowing the embedded vectors to be ordered by survival risk.

3. Methodology

The proposed model comprises three novel mechanisms (illustrated in Figure 1). We used the term 'dichotomization' for patient clustering because the patients were divided into two groups: high- and low-risk. The following notations and definitions were used in this study.

Let $v, w \in \mathbb{R}^{1\times E}$ be row vectors and $M \in \mathbb{R}^{D\times E}$ be a matrix. Denoting $\|v\|_2$ as the *L2-norm* of v, we defined $\hat{v} = \frac{v}{\|v\|_2}$, with $\hat{M}$ denoting a matrix whose rows have been L2-normalized. The *cosine similarity* of two vectors is defined as $S_c(v,w) = \hat{v} \times \hat{w}^T$. Denoting $\mathbb{I} = \{x \in \mathbb{R} | 0 \leq x \leq 1\}$, we defined the *arccosine angle distance* ($ArcD$) in two ways as follows: First, the $ArcD$ between *two vectors* is defined as

$$D_\theta(v,w) = \frac{1}{\pi}\cos^{-1}(S_c(v,w)) \in \mathbb{I} \tag{1}$$

and second, the $ArcD$ between *a vector and a matrix* is defined as $D_\theta(v,M) = \min_{1\leq k\leq D} D_\theta(v, M_k)$, where M_k is the k-th row of M.

In this study, P and G represent the number of patients and genes, respectively. The hyperparameters C_P and C_G represent the number of patient and gene clusters, respectively, and E is the dimension size of the embedding vector space. Herein, we set $C_P = 2$, $C_G = 5$, and $E = 50$. We denote $\texttt{PEM} \in \mathbb{R}^{P\times E}$ and $\texttt{GEM} \in \mathbb{R}^{G\times E}$ as the embedded matrices of the patients and genes, respectively. Moreover, $\texttt{PCM} \in \mathbb{R}^{C_P\times E}$ and $\texttt{GCM} \in \mathbb{R}^{C_G\times E}$ represent matrices for the centroids of the patient and gene clusters, respectively. These matrices are trainable weights, denoted as $W = \{\texttt{PEM}, \texttt{GEM}, \texttt{PCM}, \texttt{GCM}\}$. W is learned by minimizing $Total\ Loss$, the sum of the objective functions (Figure 1), as expressed in the following equation:

$$\widehat{W} = \underset{W}{\operatorname{argmin}}\, Total\ Loss. \tag{2}$$

3.1. Similarity-based embedding mechanism (SEM)

Like the mechanisms in other CF models, SEM represents the patient and gene entities in a new vector space (Figure 2(a)). These entities are projected to minimize the difference between the cosine similarity-based predictand and actual gene expression. We also introduced the term $\tilde{\omega}_g = \max_{1\leq k\leq C_G} S_c(\texttt{GEM}_g, \texttt{GCM}_k)$ to multiply the predictand by $\tilde{\omega}_g$. This allows W to be trained such that the maximum similarity between a gene entity and its cluster centroid is proportional to the predictand. Hence, highly expressed genes should gather around the centroid and be representative of their respective cluster. As the true gene expression was normalized between 0 and 1, the predictand was also scaled. Thus, the predictand, the *weighted gene expression*, $\widehat{\texttt{WGX}} \in \mathbb{I}^{P\times G}$, is defined as

$$\widehat{\texttt{WGX}}_{p,g} = 0.5 + 0.5 \times S_c(\texttt{PEM}_p, \texttt{GEM}_g) \times \tilde{\omega}_g. \tag{3}$$

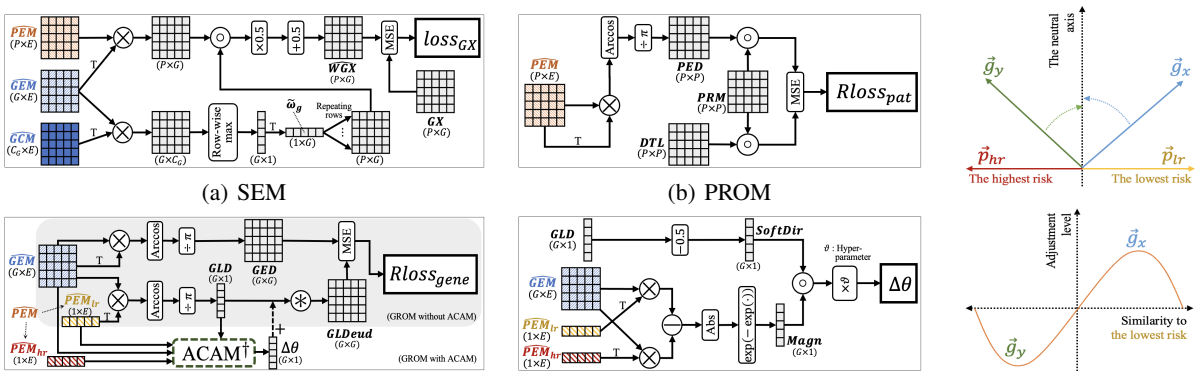
Accordingly, the proposed objective function, $loss_{GX}$, is

$$loss_{GX} = \frac{1}{PG}\sum_{p=1}^{P}\sum_{g=1}^{G}(\texttt{GX}_{p,g} - \widehat{\texttt{WGX}}_{p,g})^2, \tag{4}$$

where $\texttt{GX}_{p,g}$ is the normalized *true gene expression* for the g-th gene of the p-th patient.

3.2. Risk ordering mechanism (ROM)

ROM quantifies survival risk and sorts the patient and gene entities in order of their risk during the embedding process. Inspired by the CPH model estimation (Cox, 1972), ROM was designed for time-to-event data.



(a) SEM (b) PROM (c) GROM (d) ACAM (e) Intuition of ACAM

Figure 2. The proposed model. T, $\otimes$, and $\odot$ denote transposition, matrix multiplication, and element-wise multiplication, respectively. In (c), the operation $\circledast$ with single input $v \in \mathbb{R}^G$ returns the matrix $M \in \mathbb{R}^{G\times G}$ defined as $M_{i,j} = \|v_i - v_j\|_2$. The main models with ACAM are marked with †. In (e), the upper figure describes the soft-direction according to cosine similarity to either $\vec{p}_{hr}$ or $\vec{p}_{lr}$, and the lower figure represents the arccosine angle adjustment level according to similarity to the lowest risk.

3.2.1. PATIENT-ORIENTED ROM (PROM)

Figure 2(b) describes the structure of PROM. The matrix $\texttt{PED}$ is defined as $\texttt{PED}_{i,j} = D_\theta(\texttt{PEM}_i, \texttt{PEM}_j)$. The exogenous matrix $\texttt{DTL}$ is defined as the normalized values (between 0 and 1) of the TL differences among all patients.

$\texttt{DTL}$ only contains information about the time length differences among the patients without including death or survival events. Hence, $\texttt{PRM}$ was introduced, restricting the model training to $\texttt{DTL}$ values that showed comparable risks during the ROM process.

For example, let us suppose there are three patients, patient A, dead patient B, and surviving patient C. Given $TL_B < TL_A < TL_C$, herein, we may consider that patient B has an *obviously higher risk* while patient C has an *obviously lower risk* than patient A. Based on this intuition, the matrix $\texttt{PRM}$ is defined as

$$\texttt{PRM}_{i,j} = \begin{cases} 1, & \text{if } (\text{TL}_i > \text{TL}_j)\ \&\ (\text{event}_j = \text{death}), \\ 1, & \text{if } (\text{TL}_i < \text{TL}_j)\ \&\ (\text{event}_j = \text{censored}), \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

Ultimately, the proposed objective function for PROM is

$$Rloss_{pat} = \frac{1}{P^2}\sum_{i=1}^{P}\sum_{j=1}^{P}\texttt{PRM}_{i,j}(\texttt{DTL}_{i,j} - \texttt{PED}_{i,j})^2. \quad (6)$$

3.2.2. GENE-ORIENTED ROM (GROM)

As with the patient vectors, the gene vectors were embedded to correlate with the survival risk. Assuming all the patients were correctly ordered according to their risks during PROM, all the genes can be ordered efficiently, based on a specific patient vector as a reference. The vector $\texttt{PEM}_{lr}$ of the *lowest risk patient*, the survivor with the longest TL, was chosen as the reference, owing to its obvious and understandable baseline nature.

Technically, GROM (Figure 2(c)) orders the gene entities by equalizing (1) $\texttt{GED}$, which comprises the $ArcD$s among $\texttt{GEM}_i$, and (2) $\texttt{GLDeud}$, which comprises the differences among the $ArcD$s of $\texttt{GEM}_i$ and the reference vector $\texttt{PEM}_{lr}$. Specifically, the matrices $\texttt{GED}$ and $\texttt{GLDeud}$ are defined as $\texttt{GED}_{i,j} = D_\theta(\texttt{GEM}_i, \texttt{GEM}_j)$ and $\texttt{GLDeud}_{i,j} = \|\texttt{GLD}_i - \texttt{GLD}_j\|_2$, respectively. Here, $\texttt{GLD}$ is a vector of the $ArcD$s between $\texttt{GEM}_i$ and $\texttt{PEM}_{lr}$, defined as $\texttt{GLD}_i = D_\theta(\texttt{GEM}_i, \texttt{PEM}_{lr})$. Accordingly, the proposed objective function for GROM is

$$Rloss_{gene} = \frac{1}{G^2}\sum_{i=1}^{G}\sum_{j=1}^{G}(\texttt{GLDeud}_{i,j} - \texttt{GED}_{i,j})^2. \quad (7)$$

3.2.3. ARCCOSINE ANGLE ADJUSTMENT MECHANISM (ACAM)

To facilitate prognostic gene recommendations, the gene entities should be clearly distinguished between the genes with low and high expression values. Thus, we propose an additional mechanism, ACAM.

Let us suppose there are two gene vectors $\vec{g}_x$ and $\vec{g}_y$, with high but not very high cosine similarity to the lowest risk patient vector $\vec{p}_{lr}$ and to the highest risk patient vector $\vec{p}_{hr}$, respectively (Figure 2(e)). ACAM is the process of adjusting $ArcD$ between $\vec{g}_x$ and $\vec{p}_{lr}$ or $ArcD$ between $\vec{g}_y$ and $\vec{p}_{hr}$ according to their cosine similarities, so that $\vec{g}_x$ or $\vec{g}_y$ rotates toward the neutral axis during the embedding process (top of Figure 2(e)). The closer the similarity is to 1 or 0, the less the angle is adjusted, and the closer the similarity is to

0.5, the more the angle is adjusted (bottom of Figure 2(e)).

Technically, as shown in Figure 2(d), the adjustment term $\Delta\theta \in \mathbb{R}^{1\times G}$ is defined as follows:

$$\Delta\theta_g = SoftDir_g \times Magn_g \times \vartheta, \tag{8}$$

where ϑ is a hyperparameter for the weight. The *soft-direction*, $SoftDir_g$ is the rotation direction of $\texttt{GEM}_g$, defined as $SoftDir_g = 0.5 - \texttt{GLD}_g$. If $\texttt{GLD}_g$ is less than 0.5, $\texttt{GEM}_g$ is close to $\texttt{PEM}_{lr}$, otherwise it is close to $\texttt{PEM}_{hr}$, the vector of the dead patient with the shortest TL. Hence, given $\texttt{PEM}_{lr}$ as the reference, a negative $\texttt{GLD}_g$ rotates the angle, pulling $\texttt{GEM}_g$ towards the neutral axis, and vice versa. Additionally, the *rotation magnitude*, $Magn_g$ is defined as

$$Magn_g = \exp\left(-\exp(\Delta LH_g)\right), \text{where} \tag{9}$$

$$\Delta LH_g = |S_c(\texttt{GEM}_g, \texttt{PEM}_{lr}) - S_c(\texttt{GEM}_g, \texttt{PEM}_{hr})|. \tag{10}$$

Thus, ACAM can change the adjustment level according to $SoftDir$ and $Magn$. Further details are presented in Appendix B.

When ACAM is applied, GLD and GLDeud are replaced by $\overline{\texttt{GLD}}$ and $\overline{\texttt{GLDeud}}$, defined as $\overline{\texttt{GLD}} = \texttt{GLD} + \Delta\theta$ and $\overline{\texttt{GLDeud}}_{i,j} = \|\overline{\texttt{GLD}}_i - \overline{\texttt{GLD}}_j\|_2$, respectively. Thus, the proposed objective function for GROM with ACAM is

$$\overline{Rloss}_{gene} = \frac{1}{G^2}\sum_{i=1}^{G}\sum_{j=1}^{G}(\overline{\texttt{GLD}}\texttt{eud}_{i,j} - \texttt{GED}_{i,j})^2. \tag{11}$$

3.3. Dichotomization and clustering mechanism

In this work, the patients were dichotomized to define low-risk and high-risk groups, and the genes were clustered to recommend candidates close to the cluster centroids. For a generalizable expression, we used the term EM for PEM and GEM, and CM for PCM and GCM. The principle of this clustering is to train the centroids, CMs. Then, $\texttt{EM}_i$, with $\hat{k}$ defined as $\hat{k} = \text{argmax}_k\, S_c(\texttt{EM}_i, \texttt{CM}_k)$, is assigned to $\texttt{CM}_{k=\hat{k}}$.

In general, higher intra-cluster cohesion and higher inter-cluster separation indicate better clustering performance (Zhou & Gao, 2014). Thus, we proposed two losses for clustering: $CHloss$ to reduce the within-cluster $ArcD$ and $SPloss$ to widen the inter-cluster $ArcD$. The detailed functions for the gene and patient clustering are presented in Appendix C.

$$CHloss = \frac{1}{N}\sum_{i=1}^{N} D_\theta(\texttt{EM}_i, \texttt{CM}) \tag{12}$$

$$SPloss = \frac{1}{M}\sum_{j=1}^{M}\left(1 - \min_{k\neq j} D_\theta(\texttt{CM}_j, \texttt{CM}_k)\right) \tag{13}$$

4. Experiments and evaluation

4.1. Data sources and preprocessing

RNA sequencing gene expression data for eight tumor cohorts from The Cancer Genome Atlas were analyzed in this study. The method of downloading the data via GitHub repositories was introduced in a previous study (Kuruc et al., 2022).

The eight tumor cohorts were grouped into four tumor types, as per previous research (Kuruc et al., 2022). Specifically, these were: (1) the GLIOMA group, which included 153 cases (102 deaths) of glioblastoma multiforme and 473 cases (82 deaths) of low-grade brain glioma; (2) the KIPAN group, which included 64 cases (8 deaths) of kidney chromophobe, 516 cases (159 deaths) of kidney renal clear cell carcinoma, and 252 cases (29 deaths) of kidney renal papillary cell carcinoma; (3) the BRCA group, which included 1,012 cases (101 deaths) of invasive breast carcinoma; and (4)the COLO group, which included 339 cases (54 deaths) of colon adenocarcinoma and 133 cases (10 deaths) of rectum adenocarcinoma.

Our analyses were restricted to genes having variance values in the first quartile (Kuruc et al., 2022). Expression values, which were the predictands of the models, were normalized to between 0 and 1 after log-transformation (Robinson & Smyth, 2008). Normalization was performed separately for each group, as we expected the characteristics of the four tumor types to differ. Then, all 2,942 patients and 17,526 gene entities were assigned integer identifiers indicating the embedded vectors to be trained after sorting the patients by the event (descending) and TL (ascending). Also, DTL and PRM were generated based on the time-to-event data.

4.2. Performance comparisons using various settings

4.2.1. Ablation study settings

Five ablation models, developed according to five mechanisms, were evaluated. In addition, the performance differences between three RCFR_AC models developed with different hyperparameter ϑ values for ACAM were investigated. Details of the model settings are presented in Table 1.

4.2.2. Evaluation metrics

All models were evaluated based on eight metrics. For model-level assessment, metrics initially measured at the tumor level were summarized at the model level, as shown below.

Significant differences in survival risk: Statistical tests were performed for the four tumor types to test the null hypothesis that there were no differences in survival risk between the dichotomous groups. For these tests, CPH

Table 1. Model settings. Real numbers in the ACAM column represent hyperparameter value ϑ. Main models are indicated by a superscript †. * and ** indicate $\texttt{WGX}^{*}_{p,g} = 0.5 + 0.5 \times S_c(\texttt{PEM}_p, \texttt{GEM}_g)$ and $\texttt{WGX}^{**}_{p,g} = (\texttt{PEM}_p \times \texttt{GEM}_g^T) \times \tilde{\omega}_g$, respectively.

	Model	SEM	PROM	GROM	GCM	ACAM
Ablation	CFR	$Eq.3$	-	-	$Eq.14$ & 16	-
	$\text{RCFR}_{\text{NoGROM}}$	$Eq.3$	$Eq.6$	-	$Eq.14$ & 16	-
	RCFR	$Eq.3$	$Eq.6$	$Eq.7$	$Eq.14$ & 16	-
	$\text{RCFR\_AC}_{\text{NoCL}}$	*	$Eq.6$	$Eq.11$	-	3.0
	$\text{RCFR\_AC}_{\text{NoSIM}}$	**	$Eq.6$	$Eq.11$	$Eq.14$ & 16	3.0
Main	$\text{RCFR\_AC}^{\dagger 1}$	$Eq.3$	$Eq.6$	$Eq.11$	$Eq.14$ & 16	1.0
	$\text{RCFR\_AC}^{\dagger 2}$	$Eq.3$	$Eq.6$	$Eq.11$	$Eq.14$ & 16	2.0
	$\text{RCFR\_AC}^{\dagger 3}$	$Eq.3$	$Eq.6$	$Eq.11$	$Eq.14$ & 16	3.0

regression analyses were conducted using a binary variable indicating the group to which each patient belonged (Cox, 1972). The maximum, Sig^{Surv}, of the p-values for the four tumor types was used for the summarized metric.

Degree of size balance between patient groups: The ratio of the minimum group size to the maximum group size was employed to evaluate the degree of size balance between the dichotomous groups. The ratio, Bal^{Dicho}, was aggregated at the model level by selecting the minimum value of the ratios for the four tumor types.

Effect size of the differences in survival risk: The effect size of the differences in survival risk between the dichotomous groups was examined for the four tumor types. The effect size was measured based on an absolute coefficient value of the binary variable in the CPH regression (Cox, 1972). Among the effect sizes for the four tumor types, the minimum, ES^{Surv}_{min}, and mean, ES^{Surv}_{mean}, were employed.

Significant differences in gene expression: Statistical tests were conducted with the null hypothesis of no difference in the expression of the K selected genes between the dichotomous groups. The set of K genes comprised those with embedded vectors closest to $\texttt{GCM}$. Here, K was determined by the formula: $K = \sum_{i=1}^{5} \min(Kn, \texttt{GCE}n_i)$, where Kn is a hyperparameter indicating the maximum number of genes for selection, and $\texttt{GCE}n_i$ is the number of genes in $\texttt{GCM}_i$. Accordingly, potential K genes that show differences between the patient groups should be recommended as candidates. For testing, a Welch's t-test was performed for each candidate by tumor type, and each p-value, adjusted by the Benjamini-Hochberg method (Benjamini & Hochberg, 1995), was obtained. Next, the ratio of significant genes ($p < 0.05$) for each tumor type was calculated. Finally, the minimum, $SigR^{Gene}_{min}$, and mean, $SigR^{Gene}_{mean}$, of the ratios for the four tumor types were employed.

Effect size of differences in gene expression: The effect size of the differences in gene expression between the dichotomous groups was evaluated. The effect size for each candidate gene was initially obtained using absolute Cohen's d (Cohen, 2013). Then, the effect sizes for all candidate genes were summarized at the tumor level. For the model-level evaluation, the minimum, ES^{Gene}_{min}, and mean, ES^{Gene}_{mean}, of the effect sizes for the four tumor types were employed.

Algorithm 1 Model evaluation and selection

INPUT: $M = (m_{t,i,j}) \in \mathbb{R}^{T \times I \times J}$, MTR
$\Omega = \{\}$
A. for $\hat{t} \in \{1, \ldots, T\}$
  $\mu_{i,j} \leftarrow \{m_{t,i,j} \mid t = \hat{t}\}$
  B. for $\hat{j} \in MTR$
    $\widetilde{\mu}_{i,j=\hat{j}} \leftarrow \frac{\mu_{i,j=\hat{j}} - \min_i \mu_{i,j=\hat{j}}}{\max_i \mu_{i,j=\hat{j}} - \min_i \mu_{i,j=\hat{j}}}$
  end
  C. $\widehat{\mu}_i \leftarrow \frac{1}{J}\left\{(1 - \tilde{\mu}_{i,j=Sig^{Surv}}) + \sum_{j \neq Sig^{Surv}}^{MTR} \widetilde{\mu}_{i,j}\right\}$
  D. $\Omega \leftarrow \Omega \cup \{\mu_{i^*,j} \mid i^* = \arg\max_i \widehat{\mu}_i, j \in MTR\}$
end
OUTPUT: Ω

4.2.3. Evaluation approaches

A major consideration for the evaluation was the potential non-reproducibility of the gradient-based training (Beam et al., 2020). Thus, each model type was independently trained and evaluated five times for 100 epochs. Another challenge was the ambiguity of when to stop training the model, as the objective function may not be directly related to the identification performance. Thus, all performance throughout the 100 epochs in the five training sessions was evaluated, and the five best learning results for each model type were selected by Algorithm 1.

Algorithm 1 functions as follows. First, it takes M, the eight metric values over 100 epochs in the five training sessions for each model type, and MTR, the metric names such that $|MTR| = J$. Herein, M is a tensor in which the t, i, and j axes represent the training sessions, epochs, and metrics, respectively. To summarize, the algorithm averages the eight normalized metrics to generate a summarized metric and then returns the eight metric values with the largest summarized metric for each training session.

Although all eight metrics are important, Sig^{Surv} was prioritized over the others for the following reason. Specifically, to test statistical differences in variables of interest between groups, the groups to be compared with each other must be hypothesized a priori. In this study, the groups were hypothesized to be high-risk and low-risk patients. Thus, the seven metrics are only valid when preceded by significant differences in the survival risk between the dichotomous patients. Accordingly, the identification performance was evaluated based only on the metric values with Sig^{Surv} less

Table 2. Performance in eight metrics of each type of model. Neg and Pos refer to each task identifying the genes associated with survival positively and negatively, respectively. The metric was obtained with Kn of 100. Models written in gray with $Sig^{Surv} > 0.05$ were excluded from the evaluation.

Sign		Model	Sig^{Surv}	Bal^{Dicho}	ES^{Surv}_{min}	ES^{Surv}_{mean}	$SigR^{Gene}_{min}$	$SigR^{Gene}_{mean}$	ES^{Gene}_{min}	ES^{Gene}_{mean}
Neg	Ablation	~~CFR~~	0.590	0.77%	0.095	0.426	0.20%	15.67%	0.125	0.414
		~~$\text{RCFR}_{\text{NoGROM}}$~~	0.995	0.48%	1.201	8.306	0.00%	4.69%	0.109	0.678
		~~RCFR~~	0.146	48.60%	0.64	1.415	52.60%	67.50%	0.217	0.376
		~~$\text{RCFR\_AC}_{\text{NoCL}}$~~	0.186	6.59%	0.514	1.252	19.80%	59.15%	0.205	0.491
		~~$\text{RCFR\_AC}_{\text{NoSIM}}$~~	0.398	4.54%	3.513	6.739	0.12%	4.79%	0.165	0.31
	Main	$\text{RCFR\_AC}^{\dagger 1}$	0.033	**44.57%**	**0.82**	**1.583**	33.00%	57.38%	0.17	0.326
		$\text{RCFR\_AC}^{\dagger 2}$	**0.014**	30.35%	0.676	1.264	33.92%	64.44%	0.192	0.303
		$\text{RCFR\_AC}^{\dagger 3}$	0.020	23.85%	0.625	1.127	**68.78%**	**80.90%**	**0.241**	**0.427**
Pos	Ablation	~~CFR~~	0.563	1.77%	0.101	0.353	0.00%	20.00%	0.136	0.338
		~~$\text{RCFR}_{\text{NoGROM}}$~~	0.995	0.48%	1.201	8.306	0.00%	1.17%	0.109	0.678
		~~RCFR~~	0.360	1.79%	0.507	1.054	3.16%	48.22%	0.589	0.793
		~~$\text{RCFR\_AC}_{\text{NoCL}}$~~	0.181	16.42%	2.171	2.728	12.80%	33.20%	0.162	0.357
		$\text{RCFR\_AC}_{\text{NoSIM}}$	**0.024**	4.09%	**2.047**	**2.377**	16.64%	67.45%	0.125	**0.501**
	Main	~~$\text{RCFR\_AC}^{\dagger 1}$~~	0.242	7.03%	0.745	1.84	21.52%	53.67%	0.257	0.434
		$\text{RCFR\_AC}^{\dagger 2}$	0.055	14.43%	0.795	1.677	25.24%	53.69%	0.121	0.435
		$\text{RCFR\_AC}^{\dagger 3}$	0.035	**30.46%**	0.556	1.34	**48.89%**	**76.86%**	**0.268**	0.406

than or equal to 0.05.

We independently evaluated the performance in identifying each positive and negative gene for a prognosis for the following reasons. All $S_c(\texttt{GEM}_i, \texttt{GEM}_k)$, $S_c(\texttt{PEM}_i, \texttt{PEM}_k)$, and $S_c(\texttt{GEM}_i, \texttt{PEM}_k)$ within the same group are likely to be highly correlated owing to the SEM and ROMs, which may dichotomize patients so that the expression levels of the majority of the genes in the group are similar. In contrast, W is trained to have low similarity between the groups, so the expression levels between the groups are different. Thus, depending on the learning result of W, which may vary for each epoch, the overall expression distributions within a group (i.e., high risk) can be either higher (i.e., negative) or lower (i.e., positive) than that in the counter group (i.e., low risk). Hence, the positive (Pos) and negative (Neg) gene identification performance was measured for the learning result of each epoch.

4.2.4. Evaluation results

The best results for each model are shown in Table 2. Models with Sig^{Surv} greater than 0.05 were excluded from the evaluation, leaving only one type of model, $\text{RCFR\_AC}^{\dagger i}$. Among them, $\text{RCFR\_AC}^{\dagger 3}$ showed the best performance in eight out of 16 metrics. $\text{RCFR\_AC}_{\text{NoSIM}}$ with the same conditions as $\text{RCFR\_AC}^{\dagger 3}$ except for SEM showed the best performance in three metrics in Pos. However, it is questionable that $\text{RCFR\_AC}_{\text{NoSIM}}$ showing Sig^{Surv} of 0.398 in Neg receives a good evaluation. These results suggest that the mechanisms proposed in this study work well. Also, given that the performance depends on the ϑ in ACAM, tuning ϑ may foster better performance. Further evaluation is discussed in Appendix E.

4.3. Performance comparisons by candidate size

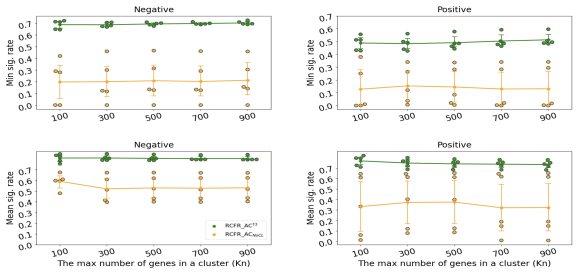
$SigR^{Gene}_{min}$ and $SigR^{Gene}_{mean}$ according to Kn from 100 to 900 for $\text{RCFR\_AC}^{\dagger 3}$ and $\text{RCFR\_AC}_{\text{NoCL}}$ were assessed (Figure 3(a)). For $\text{RCFR\_AC}_{\text{NoCL}}$, where all conditions other than the application of GCM are equivalent to $\text{RCFR\_AC}^{\dagger 3}$, the Kn candidate genes were randomly chosen. $\text{RCFR\_AC}^{\dagger 3}$ significantly outperformed $\text{RCFR\_AC}_{\text{NoCL}}$ over the entire K range, suggesting that it is better to evaluate identification performance using candidate genes close to the cluster centroids, rather than using randomly recommended genes. Moreover, the consistent results across K in $\text{RCFR\_AC}^{\dagger 3}$ may indicate the potential of evaluating candidate genes instead of all the genes.

4.4. Genomic domain-based evaluation

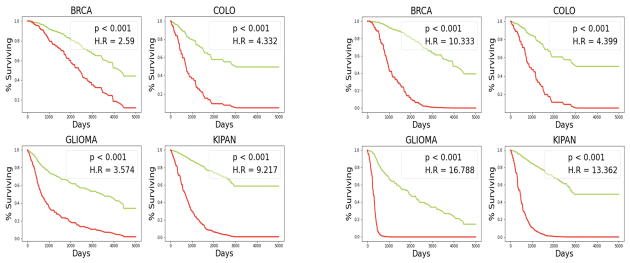
Genomic domain-based evaluations were performed based on $\text{RCFR\_AC}^{\dagger 3}$, which generally showed good performance across all metrics. Each best result for Pos and Neg was selected from those with Sig^{Surv} less than 0.01 and Bal^{Dicho} greater than 0.4. The reason for considering Bal^{Dicho} is that a size imbalance between the dichotomous groups may increase false negatives by increasing $Type$ II errors (Hsieh et al., 2003). The selection was performed through steps B, C, and D in Algorithm 1.

CPH analysis with the dichotomous groups as an independent variable revealed significant differences in survival risk between the groups for all tumor types (Figure 3(b)). Specifically, all results had p-values less than 0.001 and an HR greater than 2.59.

The identified genes were further evaluated in terms of ES



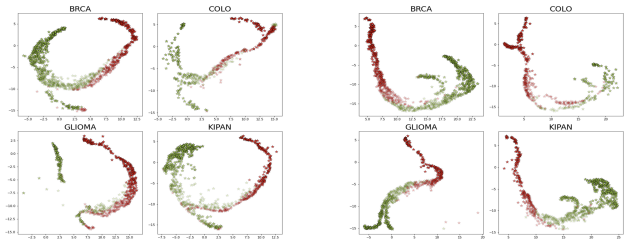
(a)



(b)

Sign	Tumor	Gene	ES	Ref.
Neg	BRCA	DDX5	0.458	(Mazurek et al., 2012)
	COLO	IGF1	0.971	(Sekharam et al., 2003)
	GLIOMA	EIF4G1	0.699	(Jaiswal et al., 2019)
	KIPAN	LASP1	0.627	(Viney et al., 2007)
Pos	BRCA	MIR320A	0.410	(Wang et al., 2015)
	COLO	PEBP1	0.607	(Al-Mulla et al., 2006)
	GLIOMA	HNRNPU	0.458	(Sapir et al., 2022)
	KIPAN	TIMP3	0.739	(Wang et al., 2014)

(c)



(d)

Figure 3. (a) Performance of $SigR^{Gene}_{min}$ and $SigR^{Gene}_{mean}$ according to Kn by Neg(left) and Pos(right). (b) Survival curves according to cancer type by Neg(left) and Pos(right). HR stands for the hazard ratio. (c) The identified genes with the largest absolute effect sizes (ES). (d) Two-dimensional representation of patient-embedded vectors ordered by survival risk according to Neg(left) and Pos(right). The darker the red, the closer $\texttt{PEM}_i$ is to $\texttt{PEM}_{hr}$, and the darker the green, the closer $\texttt{PEM}_i$ is to $\texttt{PEM}_{lr}$.

(i.e., Cohen's d) and consistency with previous findings. Figure 3(c) presents one of the top five genes with the highest ESs within the homogeneous tumor type, all of which have been previously reported as prognostic genes. A discussion of the overall distributions of the ESs is covered in Appendix E. All eight results were found to be consistent with existing knowledge. For instance, IGF1 was found to increase invasion and induce resistance to apoptosis in colon cancer cells, while LASP1, which is highly expressed in kidney cancer patients, has been suggested to be a negative prognostic marker. In addition, we confirmed the positive correlation between the expression of PEBP1 and disease-free survival in patients with primary colorectal cancer. Although we found no studies that clearly demonstrate a relationship between HNRNPU and brain-related cancer prognosis, the loss of HNRNPU was shown to induce rapid cell death in both postmitotic neurons and neural progenitors, causing brain disorders. These results are based on the genes recommended with a Kn of 100, so increasing the Kn size may allow more diverse prognostic genes to be analyzed in detail. Thus, further studies are needed to enhance and validate the performance of the proposed recommendation algorithm.

4.5. Patient representation results

To evaluate PROM, we visualized the embedded vectors of the best training result of RCFR_AC$^{\dagger 3}$ after reducing them to two dimensions using UMAP (Figure 3(d)) (McInnes et al., 2018). Overall, patients tended to be well ordered by their survival risk and grouped with similar risks, but some patients differed greatly from the majority. While more research is needed to explore the underlying cause, we speculate that it was caused by insufficient hyperparameter tuning in either of our proposed models or in UMAP.

5. Conclusions and Future Work

Herein, we proposed a framework to identify prognostic genes by dichotomizing cancer patients based on their survival risk and considering their genetic characteristics. To summarize the key results, RCFR_AC generally achieved the best performance, overcoming a limitation of CFR in which the results do not extend to statistical interpretation. Recently, most attempts to elucidate data patterns using representation learning have lacked statistical interpretability. However, statistical interpretation remains the most important practice in many scientific fields. Accordingly, our framework contributes to overcoming this limitation by bridging the gap between representation learning and statistical testing associated with survival risk in cancer patients.

Our work has great potential to be extended beyond cancer genomics. Notably, the mechanism for ordering the embedded vectors by survival risk can be applied in other fields dealing with time-to-event data while ACAM can be applied to feature recommendation tasks that show significant differences between groups.

Despite these contributions, we recognize the preliminary nature of our proposed algorithm. One primary concern that needs improvement is addressing genes with similar overall expression levels in both groups. Such genes were not adequately considered in our study, owing to the reliance of our algorithm on cosine similarity. We initially anticipated that neglecting such cases would not impact the identification of prognostic genes, as they would likely be non-significant and not recommended. However, we later realized that this issue could adversely affect the SEM process, reducing the performance of the embedding that should ideally reveal differences in expression between the groups. Consequently, we separately evaluated the performance of our algorithm in identifying each positive and negative gene for prognosis, indicating that further refinement is required before implementing our algorithm in practice.

Moreover, to improve evaluation efficiency, future work should focus on developing an objective function directly related to identification performance, while optimizing the training time. In addition, more in-depth analyses of our findings may yield more insights. Specifically, investigations into the relationships between commonly expressed genes recommended by our model would be beneficial as a primary focus of pan-cancer analysis (Campbell et al., 2020).

Acknowledgment

This research was financially supported by grants from the National Research Foundation of Korea (NRF) funded by the Korean Government (Grant No. 2021R1A4A1032861), and from the National Cancer Center of Korea (Grant No. 2310800-1).

References

Al-Mulla, F., Hagan, S., Behbehani, A. I., Bitar, M. S., George, S. S., Going, J. J., García, J. J. C., Scott, L., Fyfe, N., Murray, G. I., and Kolch, W. Raf kinase inhibitor protein expression in a survival analysis of colorectal cancer patients. *Journal of Clinical Oncology*, 24(36): 5672–5679, dec 2006. doi: 10.1200/jco.2006.07.5499.

Altman, D. G. and Royston, P. The cost of dichotomising continuous variables. *BMJ*, 332(7549):1080, May 2006.

Alves, A. d. F., Moura, A. C. d., Andreolla, H. F., Veiga, A. B. G. d., Fiegenbaum, M., Giovenardi, M., and Almeida, S. Gene expression evaluation of antioxidant enzymes in patients with hepatocellular carcinoma: RT-qPCR and bioinformatic analyses. *Genet. Mol. Biol.*, 44(2): e20190373, April 2021.

Beam, A. L., Manrai, A. K., and Ghassemi, M. Challenges to the reproducibility of machine learning models in health care. *JAMA*, 323(4):305–306, January 2020.

Beer, D. G., Kardia, S. L., Huang, C.-C., Giordano, T. J., Levin, A. M., Misek, D. E., Lin, L., Chen, G., Gharib, T. G., Thomas, D. G., Lizyness, M. L., Kuick, R., Hayasaka, S., Taylor, J. M., Iannettoni, M. D., Orringer, M. B., and Hanash, S. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine*, 8(8):816–824, jul 2002. doi: 10.1038/nm733.

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.*, 57(1):289–300, January 1995.

Cai, L., Lin, S., Girard, L., Zhou, Y., Yang, L., Ci, B., Zhou, Q., Luo, D., Yao, B., Tang, H., Allen, J., Huffman, K., Gazdar, A., Heymach, J., Wistuba, I., Xiao, G., Minna, J., and Xie, Y. LCE: an open web portal to explore gene expression and clinical associations in lung cancer. *Oncogene*, 38(14):2551–2564, April 2019.

Campbell, P. J., Getz, G., and ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature*, 578(7793):82–93, February 2020.

Ching, T., Zhu, X., and Garmire, L. X. Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS Comput. Biol.*, 14 (4):e1006076, April 2018.

Cohen, J. *Statistical power analysis for the behavioral sciences*. Routledge, London, England, 2 edition, May 2013.

Cox, D. R. Regression models and life-tables. *J. R. Stat. Soc.*, 34(2):187–202, January 1972.

Crick, F. Central dogma of molecular biology. *Nature*, 227 (5258):561–563, August 1970.

DeCoster, J., Gallucci, M., and Iselin, A.-M. R. Best practices for using median splits, artificial categorization, and their continuous alternatives. *J. Exp. Psychopathol.*, 2(2): 197–209, May 2011.

Docking, T. R., Parker, J. D. K., Jädersten, M., Duns, G., Chang, L., Jiang, J., Pilsworth, J. A., Swanson, L. A., Chan, S. K., Chiu, R., Nip, K. M., Mar, S., Mo, A., Wang, X., Martinez-Høyer, S., Stubbins, R. J., Mungall, K. L., Mungall, A. J., Moore, R. A., Jones, S. J. M., Birol, İ., Marra, M. A., Hogge, D., and Karsan, A. A clinical transcriptome approach to patient stratification and therapy selection in acute myeloid leukemia. *Nat. Commun.*, 12(1):2474, April 2021.

Dong, X., Yu, L., Wu, Z., Sun, Y., Yuan, L., and Zhang, F. A hybrid collaborative filtering model with deep structure for recommender systems. *Proc. Conf. AAAI Artif. Intell.*, 31(1), February 2017.

Hallgrimsson, B. and Hall, B. K. (eds.). *Variation*. Academic Press, San Diego, CA, June 2005.

Hsieh, F. Y., Lavori, P. W., Cohen, H. J., and Feussner, J. R. An overview of variance inflation factors for sample-size calculation. *Evaluation & the Health Professions*, 26(3): 239–257, sep 2003. doi: 10.1177/0163278703255230.

Hu, H., Li, Z., Li, X., Yu, M., and Pan, X. ScCAEs: deep clustering of single-cell RNA-seq via convolutional autoencoder embedding and soft k-means. *Briefings in Bioinformatics*, 23(1), sep 2021. doi: 10.1093/bib/bbab321.

Hua, B. and Springer, M. Widespread cumulative influence of small effect size mutations on yeast quantitative traits. *Cell Syst.*, 7(6):590–600.e6, December 2018.

Jaiswal, P. K., Koul, S., Palanisamy, N., and Koul, H. K. Eukaryotic translation initiation factor 4 gamma 1 (EIF4g1): a target for cancer therapeutic intervention? *Cancer Cell International*, 19(1), aug 2019. doi: 10.1186/s12935-019-0947-2.

Jeanselme, V., Tom, B., and Barrett, J. Neural survival clustering: Non-parametric mixture of neural networks for survival clustering. In Flores, G., Chen, G. H., Pollard, T., Ho, J. C., and Naumann, T. (eds.), *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 92–102. PMLR, 07–08 Apr 2022. URL https://proceedings.mlr.press/v174/jeanselme22a.html.

Karim, M. R., Beyan, O., Zappa, A., Costa, I. G., Rebholz-Schuhmann, D., Cochez, M., and Decker, S. Deep learning-based clustering approaches for bioinformatics. *Briefings in Bioinformatics*, 22(1):393–415, feb 2020. doi: 10.1093/bib/bbz170.

Koren, Y., Bell, R., and Volinsky, C. Matrix factorization techniques for recommender systems. *Computer (Long Beach Calif.)*, 42(8):30–37, August 2009.

Kuruc, F., Binder, H., and Hess, M. Stratified neural networks in a time-to-event setting. *Brief. Bioinform.*, 23(1), January 2022.

Lin, M., Lucas, H. C., and Shmueli, G. Too big to fail: Large samples and the p-value problem. *Information Systems Research*, 24(4):906–917, dec 2013. doi: 10.1287/isre.2013.0480.

Mazurek, A., Luo, W., Krasnitz, A., Hicks, J., Powers, R. S., and Stillman, B. DDX5 regulates DNA replication and is required for cell proliferation in a subset of breast cancer cells. *Cancer Discovery*, 2(9):812–825, sep 2012. doi: 10.1158/2159-8290.cd-12-0116.

McInnes, L., Healy, J., and Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. 2018.

Morrow, E. H. and Ingleby, F. C. Detecting differential gene expression in blastocysts following pronuclear transfer. *BMC Res. Notes*, 10(1):97, February 2017.

Nagpal, C., Yadlowsky, S., Rostamzadeh, N., and Heller, K. Deep cox mixtures for survival regression. In Jung, K., Yeung, S., Sendak, M., Sjoding, M., and Ranganath, R. (eds.), *Proceedings of the 6th Machine Learning for Healthcare Conference*, volume 149 of *Proceedings of Machine Learning Research*, pp. 674–708. PMLR, 06–07 Aug 2021. URL https://proceedings.mlr.press/v149/nagpal21a.html.

Nickel, M., Tresp, V., and Kriegel, H.-P. A three-way model for collective learning on multi-relational data. In *In Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011.

Northcott, P. A., Buchhalter, I., Morrissy, A. S., Hovestadt, V., Weischenfeldt, J., Ehrenberger, T., Gröbner, S., Segura-Wang, M., Zichner, T., Rudneva, V. A., Warnatz, H.-J., Sidiropoulos, N., Phillips, A. H., Schumacher, S., Kleinheinz, K., Waszak, S. M., Erkek, S., Jones, D. T. W., Worst, B. C., Kool, M., Zapatka, M., Jäger, N., Chavez, L., Hutter, B., Bieg, M., Paramasivam, N., Heinold, M., Gu, Z., Ishaque, N., Jäger-Schmidt, C., Imbusch, C. D., Jugold, A., Hübschmann, D., Risch, T., Amstislavskiy, V., Gonzalez, F. G. R., Weber, U. D., Wolf, S., Robinson, G. W., Zhou, X., Wu, G., Finkelstein, D., Liu, Y., Cavalli, F. M. G., Luu, B., Ramaswamy, V., Wu, X., Koster, J., Ryzhova, M., Cho, Y.-J., Pomeroy, S. L., Herold-Mende, C., Schuhmann, M., Ebinger, M., Liau, L. M., Mora, J., McLendon, R. E., Jabado, N., Kumabe, T., Chuah, E., Ma, Y., Moore, R. A., Mungall, A. J., Mungall, K. L., Thiessen, N., Tse, K., Wong, T., Jones, S. J. M., Witt, O., Milde, T., Von Deimling, A., Capper, D., Korshunov, A., Yaspo, M.-L., Kriwacki, R., Gajjar, A., Zhang, J., Beroukhim, R., Fraenkel, E., Korbel, J. O., Brors, B., Schlesner, M., Eils, R., Marra, M. A., Pfister, S. M., Taylor, M. D., and Lichter, P. The whole-genome landscape of medulloblastoma subtypes. *Nature*, 547(7663):311–317, July 2017.

Qiu, Y. L., Zheng, H., Devos, A., Selby, H., and Gevaert, O. A meta-learning approach for genomic survival analysis. *Nat. Commun.*, 11(1):6350, December 2020.

Raman, P., Zimmerman, S., Rathi, K. S., de Torrenté, L., Sarmady, M., Wu, C., Leipzig, J., Taylor, D. M., Tozeren, A., and Mar, J. C. A comparison of survival analysis methods for cancer gene expression RNA-Sequencing data. *Cancer Genet.*, 235-236:1–12, June 2019.

Robinson, M. D. and Smyth, G. K. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, 9(2):321–332, April 2008.

Sapir, T., Kshirsagar, A., Gorelik, A., Olender, T., Porat, Z., Scheffer, I. E., Goldstein, D. B., Devinsky, O., and Reiner, O. Heterogeneous nuclear ribonucleoprotein u (HNRNPU) safeguards the developing mouse cortex. *Nature Communications*, 13(1), jul 2022. doi: 10.1038/s41467-022-31752-z.

Sarwar, B., Karypis, G., Konstan, J., and Reidl, J. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the tenth international conference on World Wide Web - WWW '01*, New York, New York, USA, 2001. ACM Press.

Sekharam, M., Zhao, H., Sun, M., Fang, Q., Zhang, Q., Yuan, Z., Dan, H. C., Boulware, D., Cheng, J. Q., and Coppola, D. Insulin-like growth factor 1 receptor enhances invasion and induces resistance to apoptosis of colon cancer cells through the akt/bcl-x(l) pathway. *Cancer research*, 63:7708–7716, November 2003. ISSN 0008-5472.

Singh, A. P. and Gordon, G. J. Relational learning via collective matrix factorization. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*. ACM Press, 2008. doi: 10.1145/1401890.1401969.

Sundar, R., Barr Kumarakulasinghe, N., Huak Chan, Y., Yoshida, K., Yoshikawa, T., Miyagi, Y., Rino, Y., Masuda, M., Guan, J., Sakamoto, J., Tanaka, S., Tan, A. L.-K., Hoppe, M. M., Jeyasekharan, A. D., Ng, C. C. Y., De Simone, M., Grabsch, H. I., Lee, J., Oshima, T., Tsuburaya, A., and Tan, P. Machine-learning model derived gene signature predictive of paclitaxel survival benefit in gastric cancer: results from the randomised phase III SAMIT trial. *Gut*, 71(4):676–685, April 2022.

Suntsova, M., Gaifullin, N., Allina, D., Reshetun, A., Li, X., Mendeleeva, L., Surin, V., Sergeeva, A., Spirin, P., Prassolov, V., Morgan, A., Garazha, A., Sorokin, M., and Buzdin, A. Atlas of RNA sequencing profiles for normal human tissues. *Sci. Data*, 6(1):36, April 2019.

Tran, K. B., Lang, J. J., Compton, K., and GBD 2019 Cancer Risk Factors Collaborators. The global burden of cancer attributable to risk factors, 2010-19: a systematic analysis for the global burden of disease study 2019. *Lancet*, 400 (10352):563–591, August 2022.

Viney, R. L., Morrison, A. A., van den Heuvel, L. P., Ni, L., Mathieson, P. W., Saleem, M. A., and Ladomery, M. R. A proteomic investigation of glomerular podocytes from a denys-drash syndrome patient with a mutation in the wilms tumour suppressor geneWT1. *PROTEOMICS*, 7 (5):804–815, mar 2007. doi: 10.1002/pmic.200600666.

Wang, B., Yang, Z., Wang, H., Cao, Z., Zhao, Y., Gong, C., Ma, L., Wang, X., Hu, X., and Chen, S. Microrna-320a inhibits proliferation and invasion of breast cancer cells by targeting rab11a. *American journal of cancer research*, 5:2719–2729, 2015. ISSN 2156-6976.

Wang, C. and Blei, D. M. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11*. ACM Press, 2011. doi: 10.1145/2020408.2020480.

Wang, Z., Famulski, K., Lee, J., Das, S. K., Wang, X., Halloran, P., Oudit, G. Y., and Kassiri, Z. TIMP2 and TIMP3 have divergent roles in early renal tubulointerstitial injury. *Kidney International*, 85(1):82–93, jan 2014. doi: 10.1038/ki.2013.225.

Witkiewicz, A. K., McMillan, E. A., Balaji, U., Baek, G., Lin, W.-C., Mansour, J., Mollaee, M., Wagner, K.-U., Koduru, P., Yopp, A., Choti, M. A., Yeo, C. J., McCue, P., White, M. A., and Knudsen, E. S. Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets. *Nature Communications*, 6(1), apr 2015. doi: 10.1038/ncomms7744.

Zhou, H. B. and Gao, J. T. Automatic method for determining cluster number based on silhouette coefficient. *Advanced Materials Research*, 951:227–230, may 2014. doi: 10.4028/www.scientific.net/amr.951.227.

Appendix

A. Research framework

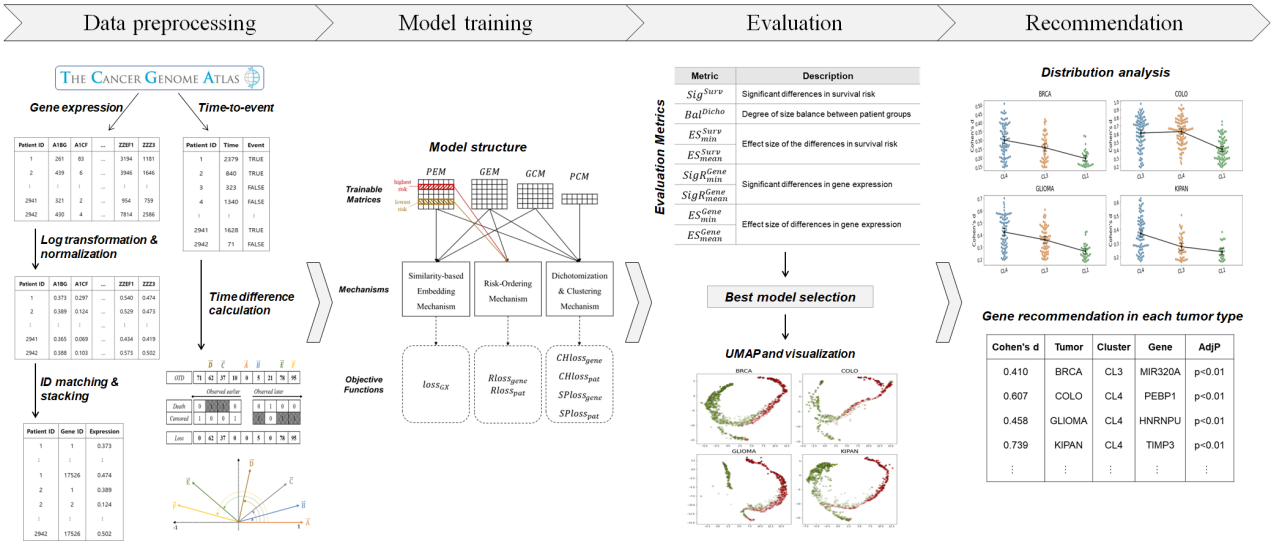


Figure 4. Research framework.

Figure 4 illustrates the framework of the process from data preprocessing to gene recommendation. The details of this process are thoroughly described across sections 3 and 4 in the main manuscript.

B. Further explanation of the ACAM adjustment level

Recall equation (10):

$$\Delta LH_g = |S_c(\texttt{GEM}_g, \texttt{PEM}_{lr}) - S_c(\texttt{GEM}_g, \texttt{PEM}_{hr})|.$$

With the $SoftDir$ and $Magn$ introduced in 3.2.3, ACAM can change the adjustment level according to ΔLH. A larger ΔLH_g value indicates that $\texttt{GEM}_g$ is more polarized to either $\texttt{PEM}_{lr}$ or $\texttt{PEM}_{hr}$. The larger the ΔLH_g value, the less the angle should be adjusted; thus, $Magn_g$ is the function of ΔLH_g transformed in reverse order based on a negative exponential function. However, even when given $\max(\Delta LH) = 2$ as the cosine similarity ranges from -1 to 1, the angle still rotates because $\exp(-\max(\Delta LH)) \approx 0.135$. Thus, the exponential function is composed as shown in equation (9),

$$Magn_g = \exp\left(-\exp(\Delta LH_g)\right),$$

because $\exp(-\exp(\max(\Delta LH))) \approx 0.001$.

The closer $\texttt{GEM}_g$ is to $\texttt{PEM}_{lr}$ or $\texttt{PEM}_{hr}$, the higher the positive or negative intensity of $SoftDir_g$, respectively, but the lower the intensity of $Magn_g$. Thus, this mechanism allows $\Delta\theta$ to have one periodic property, as shown in the bottom of Figure 2(e).

C. Objective functions for dichotomization and clustering mechanism

In 3.3, $CHloss$ and $SPloss$ have been described without any distinction between gene and patient clustering. Here, equations (12) and (13) for gene and patient clustering, respectively, are rewritten as follows:

$$CHloss_{gene} = \frac{1}{G}\sum_{g=1}^{G} \max\{D_\theta(\texttt{GEM}_g, \texttt{GCM}) - \xi_1, 0\}, \tag{14}$$

$$CHloss_{pat} = \frac{1}{P}\sum_{p=1}^{P} \max\{D_\theta(\texttt{PEM}_p, \texttt{PCM}) - \xi_2, 0\}, \tag{15}$$

$$SPloss_{gene} = \frac{1}{C_G}\sum_{i=1}^{C_G}\max\{\xi_3 - (D_\theta^s(\texttt{GCM}))_i, 0\}, \tag{16}$$

$$SPloss_{pat} = \frac{1}{C_P}\sum_{i=1}^{C_P}\max\{\xi_4 - (D_\theta^s(\texttt{PCM}))_i, 0\}, \tag{17}$$

where

$$(D_\theta^s(\texttt{GCM}))_i = \min_{k\neq i} D_\theta(\texttt{GCM}_i, \texttt{GCM}_k),$$
$$(D_\theta^s(\texttt{PCM}))_i = \min_{k\neq i} D_\theta(\texttt{PCM}_i, \texttt{PCM}_k),$$

and the ξ_is are positive hyperparameters.

D. Constraints of balance between dichotomous groups

Extremely rare members in a cluster may lead to biased statistical results (Hu et al., 2021; Hsieh et al., 2003). To balance the sizes of the dichotomous groups, we constrained the mean cosine similarities between each row $\texttt{PCM}_i$ and all $\texttt{PEM}_p$ vectors to be similar. Thus, the following additional objective function was introduced:

$$Bloss = \frac{1}{C_P^2 - C_P}\sum_{i=1}^{C_P}\sum_{j=1, j\neq i}^{C_P}\max(\overline{DS}_{i,j} - \xi_5, 0), \tag{18}$$

where

$$\overline{DS}_{i,j} = (\overline{\texttt{PCS}}_i - \overline{\texttt{PCS}}_j)^2,$$
$$\overline{\texttt{PCS}}_i = \frac{1}{P}\sum_{p=1}^{P} S_c(\texttt{PEM}_p, \texttt{PCM}_i),$$

and ξ_5 is a positive hyperparameter.

E. Further evaluation

As our recommendation framework is first being proposed in this study, evaluating the identification performance for prognostic genes from multidimensional perspectives is worthwhile. Thus, we conducted three additional assessments: further performance evaluation using the eight metrics, an investigation of performance changes over epochs, and an exploration of the distribution of effect sizes. Performance in identifying genes that are positively (Pos) and negatively (Neg) associated with survival risk was assessed independently.

E.1. Performance using the eight metrics

We further investigated the performance of each of the eight metrics in terms of patient dichotomization and the statistical significance of the prognostic genes. The eight metrics, Sig^{Surv}, Bal^{Dicho}, ES_{min}^{Surv}, ES_{mean}^{Surv}, $SigR_{min}^{Gene}$, $SigR_{mean}^{Gene}$, ES_{min}^{Gene}, and ES_{mean}^{Gene} are described in 4.2.2 in the manuscript. Additionally, to account for the variance of the results, we examined all the performance metrics of the five best independently trained models.

E.1.1. Performance in dichotomization

For the evaluation of patient dichotomization by their survival risk, Sig^{Surv}, Bal^{Dicho}, ES_{min}^{Surv}, and ES_{mean}^{Surv} were investigated. Summarizing the key findings, most models except the RCFR_AC models, showed high values in Sig^{Surv} (Figure 5(A)), indicating that they failed to dichotomize patients by survival risk. The performance of these models on the remaining metrics may not be valid because the groups for hypothesis testing are incorrectly defined.

By focusing on the performance of a model type, the $\text{RCFR}_{\text{NoGROM}}$ models showed high values for both Sig^{Surv} (i.e., low performance) and ES_{mean}^{Surv} (i.e., high performance) metrics (Figure 5(A) and (D)), unlike other models. These results may be associated with low performance in Bal^{Dicho} (Figure 5(B)), where Bal^{Dicho} values close to 0 indicate that there are rare members in a group. Accordingly, although some outliers contribute to large ES_{mean}^{Surv} values, their small sample size may

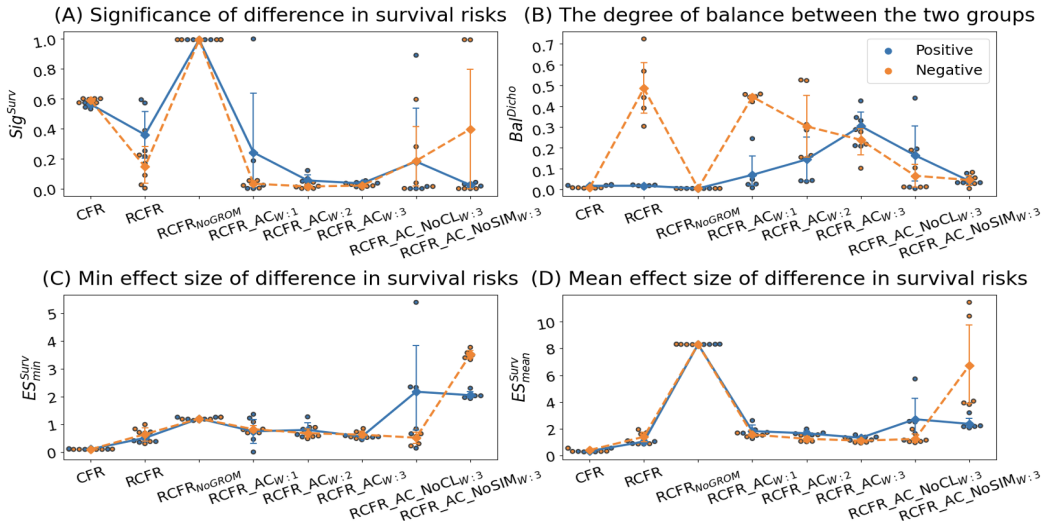


Figure 5. Performance of patient dichotomization. (A), (B), (C), and (D) represent Sig^{Surv}, Bal^{Dicho}, ES^{Surv}_{min}, and ES^{Surv}_{mean}, respectively. The main manuscript describes the model settings in Table 1.

lead to large variance, resulting in a statistically insignificant difference in survival risk between the dichotomous groups (Lin et al., 2013).

Only the RCFR_AC models showed significant performance in the priority metric, Sig^{Surv} and comparable performance in the other metrics, suggesting that the SEM, ROMs, CMs, and ACAM proposed in this study work well. However, because some results from the five best models within each model type were somewhat inconsistent, further studies are needed to evaluate performance based on more training results. In addition, research on developing a method that can minimize the inconsistency of learning results owing to the initialization of the weight matrices is needed.

E.1.2. Performance in gene significance

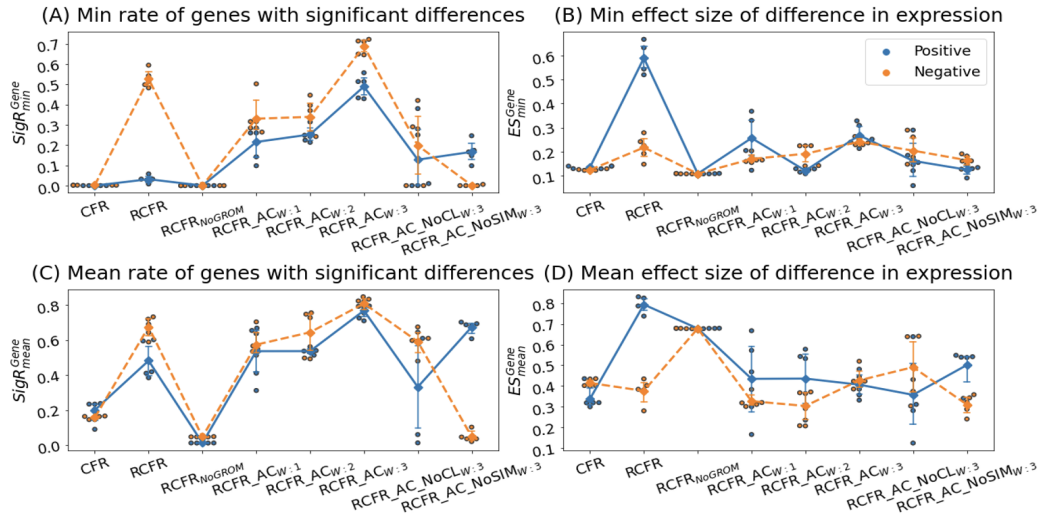


Figure 6. Performance in the significance of prognostic genes. (A), (B), (C), and (D) represent $SigR^{Gene}_{min}$, ES^{Gene}_{min}, $SigR^{Gene}_{mean}$, and ES^{Gene}_{mean}, respectively. The main manuscript describes the model settings in Table 1.

For assessment of the identification of prognostic genes with statistical differences in expression between the dichotomous groups, $SigR^{Gene}_{min}$, $SigR^{Gene}_{mean}$, ES^{Gene}_{min}, and ES^{Gene}_{mean} were explored (Figure 6). As with the evaluation of patient dichotomization, there were inconsistent results among the five best models within the same model type.

In terms of the statistical significance, $SigR^{Gene}_{min}$ and $SigR^{Gene}_{mean}$, the RCFR_AC models consistently showed high performance (Figure 6(A) and (C)). Although the RCFR models showed high performance in Neg, given high values (i.e., low performance) in Sig^{Surv} (Figure 5(A)), we cannot conclude that the results show the significant differences in the gene expression between the groups divided by survival risk.

In terms of the mean effect size, ES^{Gene}_{mean}, the RCFR$_{\text{NoGROM}}$ models showed excellent performance in both Neg and Pos (Figure 6(D)). However, the low performance in $SigR^{Gene}_{min}$ and $SigR^{Gene}_{mean}$ devalues the results (Figure 6(A) and (C)). Also, the low performance in Sig^{Surv} (Figure 5(A)) suggests the results may not represent the effect size of the differences in the gene expression between the groups divided by survival risk.

Furthermore, even though the RCFR models performed excellently on ES^{Gene}_{min} in Pos (Figure 6(B)), the performance of the significant differences in the gene expression between the dichotomous patient groups was very poor (Figure 6(A)). Additionally, like the RCFR$_{\text{NoGROM}}$ models, the RCFR models could not dichotomize patients by survival risk (Figure 5(A)).

For the RCFR_AC models, the performance of $SigR^{Gene}_{min}$ and $SigR^{Gene}_{mean}$ tends to depend on the hyperparameter, ϑ, used for ACAM (Figure 6(A) and (C)). While this may demonstrate that ACAM functions well, it motivates further research to investigate optimal hyperparameter tuning.

E.2. Performance comparisons over epochs

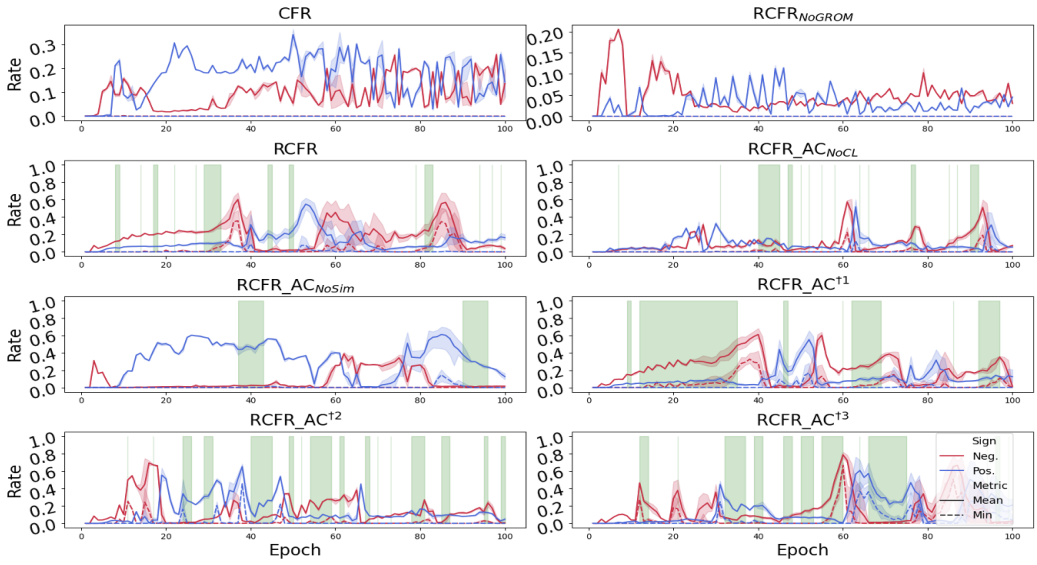


Figure 7. Performance of $SigR^{Gene}_{min}$ and $SigR^{Gene}_{mean}$ by Neg or Pos over all epochs. The red and blue lines represent the performance of Neg and Pos, respectively. The solid and dotted lines represent the mean and min metric performance, respectively. Green-shaded intervals indicate significant differences in survival risk between the dichotomous groups ($Sig^{Surv} < 0.05$).

The performance of $SigR^{Gene}_{min}$ and $SigR^{Gene}_{mean}$ by Neg or Pos was evaluated according to each epoch (Figure 7). Over 100 epochs, the CFR and RCFR$_{\text{NoGROM}}$ models did not result in significant differences in survival risk between dichotomous groups ($Sig^{Surv} > 0.05$). Furthermore, there was no result showing high performance in both Pos and Neg simultaneously, when Sig^{Surv} was less than 0.05. The possible causes related to this phenomenon are discussed in 4.2.3 of the main text. Additionally, given that the gene expression distributions of groups of interest are often similar to each other (Morrow & Ingleby, 2017; Hua & Springer, 2018), small differences in the distributions may lead to these results.

E.3. Distribution of effect sizes

We further investigated the effect size distributions of the gene expression differences. We prepared swarm-plots to visualize the distributional characteristics of the effect sizes based on the absolute Cohen's d values for significant genes (adjusted-$p < 0.05$) (Cohen, 2013). The effect sizes were obtained from the best RCFR_AC$^{\dagger 3}$ model with a Kn of 100. (For Kn, see 4.2.2 Significant differences in gene expression.) Each best result for Pos and Neg was selected from those with Sig^{Surv} less than 0.01 and Bal^{Dicho} greater than 0.4. Although we set the number of gene clusters to five (i.e., $C_G = 5$) for the

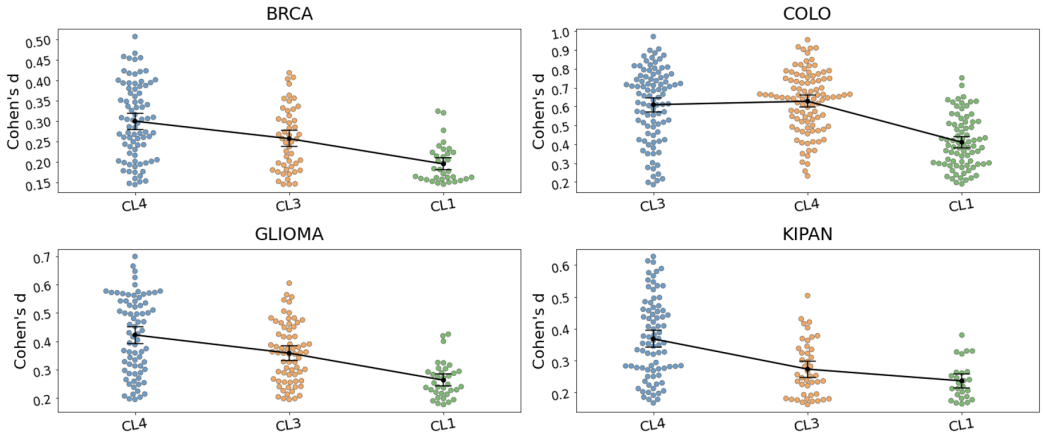


Figure 8. Effect size distributions by Neg. BRCA, COLO, GLIOMA, and KIPAN represent breast-, colorectal-, brain-, and kidney-related cancers, respectively.

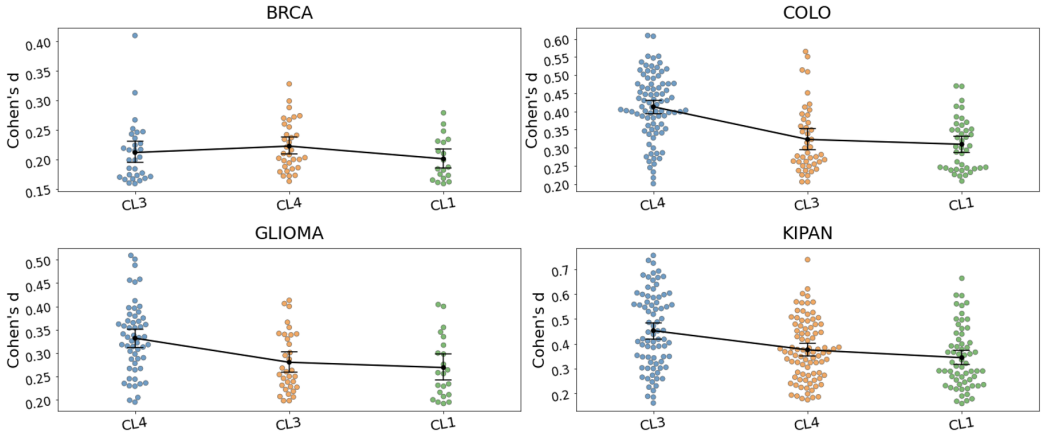


Figure 9. Effect size distributions by Pos. BRCA, COLO, GLIOMA, and KIPAN represent breast-, colorectal-, brain-, and kidney-related cancers, respectively.

RCFR_AC$^{\dagger 3}$ in this work, only three clusters identified significant prognostic genes in both Neg and Pos (Figure 8 and Figure 9).

For Neg, the effect sizes in COLO were higher overall than those in the other cancer groups (Figure 8). Contrastingly, the overall effect sizes were the lowest in the BRCA group. Unlike the performance for Neg, for Pos, the effect sizes were higher overall in KIPAN than in the other groups (Figure 9). Additionally, like the results for Neg, the effect sizes in BRCA were the lowest compared to the other groups.

Finally, it is worth noting that the overall effect sizes varied across clusters. This may indicate the ability of RCFR_AC$^{\dagger 3}$ to group the prognostic genes with differences in effect sizes between clusters. As this ability relates to molecular subtyping, a major focus of cancer genomics (Northcott et al., 2017), research that investigates these clustering results in depth in terms of cancer genomics and further improves clustering performance should be encouraged.