Briefings in Bioinformatics, 2026, 27, bbag290
https://doi.org/10.1093/bib/bbag290
Published: 4 June 2026
Problem Solving Protocol

OXFORD

CSIE: cancer subtyping via inference and ensemble

Dao Tran[1], Yen Thi-Hai Pham[2], Hung N. Luu[2], Juli Petereit[3], Manuel A. Andrade-Rodriguez[4], Phi Bya[1], Tin Nguyen[1,5,*]

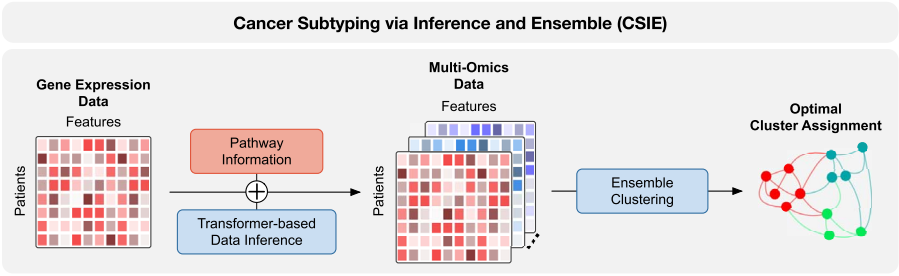
[1]Department of Industrial and Systems Engineering, Wayne State University, 4815 4th St, Detroit, MI 48201, United States
[2]Dr. Mary and Ron Neal Cancer Center, Houston Methodist Research Institute, 6670 Bertner Ave, Houston, TX 77030, United States
[3]Nevada Bioinformatics Center, University of Nevada, Reno, 1664 N Virginia St, Reno, NV 89557, United States
[4]Department of Agriculture, Veterinary and Rangeland Sciences, University of Nevada, Reno, 1664 N Virginia St, Reno, NV 89557, United States
[5]Karmanos Cancer Institute, Wayne State University School of Medicine, 4100 John R St, Detroit, MI 48201, United States

*Corresponding author. Department of Industrial and Systems Engineering, Wayne State University, 4815 4th St, Detroit, MI 48201, United States. E-mail: tin@wayne.edu

Abstract

While multi-omics integration is the gold standard for precision oncology, its clinical utility is severely hampered by the incomplete data problem, where cost and technical barriers often leave researchers with only single-omics profiles. Our manuscript introduces CSIE (cancer subtyping via inference and ensemble), a framework that bridges this gap by using a novel transformer-based inference module which incorporates systems-level knowledge to accurately infer missing omics layers from gene expression data. Furthermore, CSIE employs an ensemble clustering module that simultaneously integrates multi-omics data via different similarity metrics and clustering algorithms to capture molecular patterns of cancer subtypes. The robustness of CSIE is validated through extensive benchmarking against 12 state-of-the-art methods across 66 cancer datasets with over 15 000 patients and 22 diverse data modalities/platforms. Our results demonstrate that CSIE significantly outperforms existing tools, particularly in scenarios with incomplete data. This work shifts the paradigm from requiring exhaustive data collection to leveraging biological intelligence for data completion, offering a scalable solution for high-resolution cancer subtyping in real-world clinical settings. All source code of CSIE and scripts for regenerating results reported in this article are available at https://github.com/tinnlab/CSIE.

Graphical Abstract



Keywords cancer subtyping, multi-omics, data inference, transformer, pathway information, consensus clustering

Introduction

Cancer encompasses a broad spectrum of diseases, ranging from highly malignant, metastatic, and aggressive cancers to benign lesions with a low risk of progression or death. To capture the dynamic nature of cancer development, researchers employ various genome-wide profiling techniques across multiple biological levels, including genomics (DNA), transcriptomics (RNA), epigenomics (gene regulation), proteomics (proteins), and metabolomics (metabolites). Analyzing multi-omics data provides a comprehensive view of cancer evolution, molecular subtypes, and potential risks. These insights are critical for effective personalized treatment and prognosis [1–4].

Received: January 12, 2026. **Revised:** April 26, 2026. **Accepted:** May 11, 2026

Multi-omics integration has led to significant discoveries in cancer research. For example, integrating transcriptomic, epigenomic, and proteomic data from leukemic blood cells has identified cancer-specific processes in blood differentiation and crucial markers for leukemia subtypes [5]. Similarly, multi-omics studies have resulted in the discovery of novel subgroups and new therapeutic targets across various cancer types, including breast cancer [6], liver cancer [7], lung cancer [8], and brain cancer [9], among others [10–13].

Many integrative methods have been developed for cancer subtyping, which can be grouped into three main categories: consensus-based models, shared representation methods, and similarity-based approaches. Consensus-based models identify clusters within each omics dataset and combine the individual clustering results into an optimal final grouping. Early approaches like CC [14], RDCCE [15], BCC [16], MDI [17], and ECC [18] focus on combining cluster assignments, while recent methods such as KLIC [19], CancerSubtypes [20], MOVICS [21], ClustOmics [22], Subtype-WESLR [23], and DSCC [24] integrate multiple clustering algorithms. Shared representation methods first generate a low-dimensional common representation across all data types and then apply clustering to discover subtypes; examples include scMNMF [25], Subtype-GAN [26], SMRT [27], SCFA [28], intNMF [29], LRACluster [30], iClusterBayes [31], iClusterPlus [32], iCluster [33, 34], MRGCN [35], DLSF [36], and DSIR [37]. Finally, similarity-based approaches construct a similarity matrix (e.g. a patient-to-patient network) for each individual data type, mathematically combine them into a single overall similarity matrix, and then partition this matrix using a clustering algorithm. Prominent examples are PartIES [38], SNF [39], NEMO [40], PINS [41, 42], CIMLR [43], ANF [44], hMKL [45], MDICC [46], and iMKL [47].

Despite the variety of available methods, current subtyping approaches face significant challenges. First, many state-of-the-art methods rely on specific omics types or particular data platforms, and often require completely matched samples across all data types. However, these ideal conditions are rarely met in practice, where typically only gene expression data are available. As a result, existing subtyping methods usually exhibit limited performance on incomplete datasets. Second, most of the available subtyping approaches only employ Euclidean distance metrics to measure the differences among samples while ignoring alternative metrics. Directional distance formulations, such as angular and cosine-based metrics, which have gained increasing traction in deep learning research [48–50], have the potential to enhance the discriminative power of subtyping methods. Third, existing tools often fail to consider the vital role of pathway information, which is fundamental for understanding the convergence of molecular variations into similar biological functions. The incorporation of systems-level knowledge has increased the performance of models developed for cancer recurrence and cancer survival prediction [51, 52], and can lead to meaningful subgroup discoveries in disease subtyping.

To address these challenges, we introduce a new subtyping approach named Cancer Subtyping via Inference and Ensemble (CSIE). CSIE adopts a transformer-based data inference module that utilizes pathway information to capture meaningful cross-omics relationships, and generates important molecular types (miRNA, DNA methylation) from gene expression datasets when these data are missing. The method also employs an ensemble clustering module that leverages different distance metrics in constructing patient affinity matrices and applies a multi-stage fusion technique to effectively integrate these networks across all available omics types. Additionally, CSIE combines different clustering algorithms in its ensemble clustering module to derive the optimal subgroups for each dataset. To demonstrate the advantages of the proposed method, we compare CSIE against 12 current state-of-the-art approaches: consensus clustering (CC) [14], CIMLR [43], SNF [39], LRACluster [30], intNMF [29], ANF [44], NEMO [40], MRGCN [35], hMKL [45], MDICC [46], DLSF [36], and DSIR [37]. Our benchmarking involves an extensive analysis of 66 cancer datasets with over 15 000 patients obtained from the Genomic Data Commons, or formerly known as The Cancer Genome Atlas (GDC/TCGA), cBioPortal, NCBI GEO, and published articles.

Materials and methods

Data processing

We use all available omics types for the downloaded datasets, including: mRNA, miRNA expression, DNA methylation, copy number variations (CNVs), somatic mutations, protein, and metabolite levels. Among these data types, only CNVs encompasses gene-level features, while others comprise features at different levels. Each omics type may include multiple data formats, which are treated as separate data matrices in our analysis. For instance, mRNA includes raw counts, Transcripts Per Million (TPM), Fragments Per Kilobase of transcript per Million fragments mapped (FPKM), and upper quartile FPKM (FPKMuq); miRNA expression contains Reads Per Million (miRNArpm) and Isoform (miRNAiso). We follow the data processing pipeline introduced in our previous work [24]. We start with gene-level aggregation for mRNA, miRNA, DNA methylation, and protein quantification, filtering out the genes not associated with Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. We then perform log2 transformation as well as replace missing values with zeros for mRNA, miRNA, DNA methylation, CNVs, protein levels, and metabolomics data (see Supplementary Fig. S2 for alternative imputation techniques). For somatic mutations, we count the number of each single-base substitution type (among C>A, C>G, C>T, T>A, T>C, T>G) for each sample.

Overall pipeline

Figure 1 illustrates the CSIE framework for cancer subtyping, which comprises two primary modules: data inference (Fig. 1A) and ensemble clustering (Fig. 1B). The data inference module is designed to reconstruct missing omics types, while the ensemble clustering module integrates multi-omics data to partition patients into distinct subtypes.

In the data inference module, miRNA (miRNArpm and miRNAiso) and DNA methylation data, if they are missing, are generated from gene expression profiles (Fig. 1A). The inference models for these three data types share a common architecture: an encoder to transform the source data (gene expression) into pathway embedding matrices, and a decoder to reconstruct the target data from these embeddings. We optimize the parameters for each model using gene expression, miRNA, and methylation data aggregated from all GDC/TCGA datasets.

The ensemble clustering module processes the multi-omics data by combining various data types, distance metrics, and clustering algorithms (Fig. 1B). For each distance metric, the method first constructs the patient similarity networks across all omics types before combining them into a single consensus affinity matrix. Next, the method

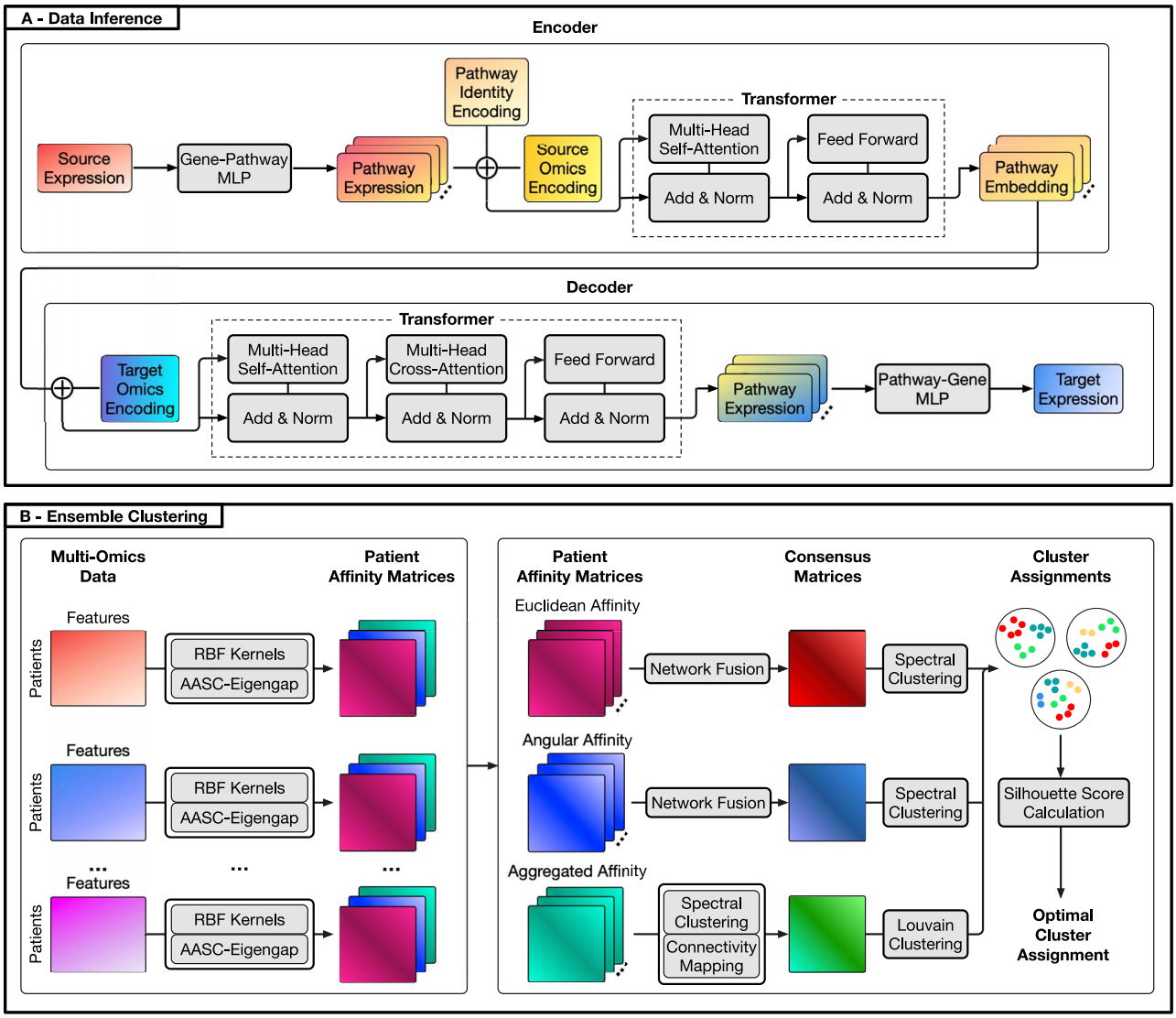


Figure 1 The overall analysis pipeline of CSIE encompassing two main modules: (A) The data inference module uses a gene-pathway MLP encoder and transformer block to convert gene expression data into pathway embeddings, which a decoder transforms into target omics data (miRNArpm, miRNAiso, or DNA methylation); (B) The ensemble clustering module generates and combines Euclidean, angular, and aggregated affinity matrices per omics input, applies spectral clustering and the Louvain algorithm, and selects the final cancer subtyping result using the highest Silhouette score.

employs spectral clustering and community detection methods to partition these consensus matrices, ultimately selecting the clustering solution with the highest Silhouette score.

Inference of miRNA and DNA methylation

The goal of the data inference module is to generate important data types from gene expression data if such data types are missing from the input. In this article, we use the same model architecture to generate miRNA (miRNArpm and miRNAiso) and DNA methylation. For each of the three data types, we train the parameters separately using all data samples from GDC/TCGA, which have gene expression, miRNArpm, miRNAiso, and DNA methylation data. We merge all GDC/TCGA samples into a large dataset and then use it to fine-tune parameters for each of the three data types (miRNArpm, miRNAiso, and DNA methylation). The overall architecture of the module consists of an encoder and a decoder. The encoder transforms the source data (gene expression) into pathway embedding matrices, whereas the decoder generates the target omics data (miRNArpm, miRNAiso, and DNA methylation) from the pathway embeddings.

Encoder

The encoder transforms the input gene expression vector of each patient into a pathway embedding matrix (pathways by latent features), in which rows represent KEGG pathways (353 pathways) and columns represent latent features (256 features). The encoder consists of one *gene-pathway multi-layer perceptron (MLP)*, two encoding matrices *(pathway identity encoding, source omics encoding)*, and one *transformer block*.

The *gene-pathway MLP* encompasses one gene-pathway layer initialized with KEGG pathway-gene membership and two fully connected layers—each with layer normalization, Rectified Linear Unit (ReLU) activation, and dropout regularization. The *gene-pathway MLP*

aggregates gene-level expression into pathway activities, transforming the expression vector of each patient into a pathway expression matrix. These expression matrices are then augmented with the *pathway identity encoding* and *source omics encoding* before being passed to the *transformer block*. The two encodings are learnable matrices randomly initialized with the same dimensions as pathway expression matrices.

The *pathway identity encoding* is comprised of unique encoding vectors, in which each vector represents a pathway. The addition of this encoding allows the model to distinguish among pathways and prioritize important pathways during training. The *source omics encoding* consists of a single encoding vector, replicated across all pathways, which informs the model of the molecular type being used as the source omics. The incorporation of this encoding enables learning omics-specific transformation patterns during cross-omics translation.

The *transformer block* consists of two components: a self-attention mechanism [53], and a two-layer feed-forward network, each with ReLU activation and dropout. A residual connection and layer normalization are employed around each of these components. The *transformer block* transforms each augmented pathway expression matrix into a pathway embedding matrix with the same dimensions. During this process, each expression matrix is treated as an embedding of a sequence of 353 tokens (pathways). The self-attention mechanism computes context-rich representation for each pathway by allowing every pathway to attend to all others in the sequence, which helps capture inter-pathway relationships. The feed-forward network enables learning more complex, non-linear transformations of each pathway's representation via ReLU activation, thus further refining pathway embeddings.

Decoder

The decoder employs a reverse architecture, compared to the encoder, to reconstruct the target data type from the pathway embedding matrices. Specifically, it encompasses one encoding matrix (*target omics encoding*), one *transformer block*, and one *pathway-gene MLP*. Similar to the *source omics encoding*, the *target omics encoding* is a learnable matrix randomly initialized, which informs the model of the target omics type and guides the model to generate omics-specific output patterns. The decoder combines the pathway embedding matrices (from the encoder) with the *target omics encoding* before passing them to the *transformer block*.

The *transformer block* of the decoder comprises a self-attention mechanism, a cross-attention mechanism [54], and a two-layer feed-forward network, each with ReLU activation and dropout. Residual connections and layer normalization are still applied around each of the three components. The *transformer block* reconstructs pathway expression matrices from the pathway embedding matrices. The addition of the cross-attention mechanism helps the model learn which cross-omics pathway relationships are biologically meaningful for accurate omics-translation. Unlike the self-attention mechanism which queries and attends to the same data, the cross-attention mechanism queries the decoder's self-attention output while attending to the encoder's output. As a result, this mechanism enables the model to compute attention weights that determine the relevance of each source pathway to each target pathway.

The *pathway-gene MLP* employs two layers with layer normalization, ReLU activation and dropout regularization. This block first aggregates pathway-level expression via mean pooling, producing a single pathway expression vector for each patient, then projects each vector into a full expression vector of the target omics type.

Training

To prepare for the training of each model, we merge all GDC/TCGA datasets into one merged dataset and select only samples that have both the source omics (gene expression in mRNAtpm format) and target omics (miRNArpm, miRNAiso, or DNA methylation). We also intersect genes between the two modalities (source and target omics), keeping only genes present in both. The merged dataset is then partitioned into training (80%) and validation (20%) sets using stratified sampling which preserves the distribution of cancer types across splits. Finally, z-score transformation is performed for each gene, in which the mean and standard deviation are computed from the training set and applied to both training and validation data.

Each model is trained using a bidirectional learning approach that simultaneously optimizes two translation directions: forward and reverse. For the forward direction, we use gene expression as the source omics and leverage other omics types (miRNArpm, miRNAiso, or DNA methylation) as the target omics. For the reverse direction, we use other molecular data as the source and gene expression as the target. In each training iteration, both directional translations are computed and their training losses are jointly minimized, enabling the model to learn bidirectional cross-omics relationships. We employ the *AdamW* optimizer [55] with an initial learning rate of 10^{-5} and weight decay of 10^{-6} for regularization. To prevent overfitting, we incorporate a dropout regularization rate of 0.1, gradient clipping (maximum norm of 1.0), and early stopping with a patience of 20 epochs based on the validation loss. The best-performing model, determined by the lowest validation loss, is saved along with the data normalization parameters (mean and standard deviation) to enable consistent preprocessing during inference.

Loss functions

We employ a composite loss function that includes three components: *reconstruction loss* ($\mathcal{L}_{\text{re}}$), *cycle consistency loss* ($\mathcal{L}_{\text{cy}}$), and *pathway consistency loss* ($\mathcal{L}_{\text{pa}}$). The total loss is defined as follows:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{re}} + \lambda_2 \mathcal{L}_{\text{cy}} + \lambda_3 \mathcal{L}_{\text{pa}}$$

where $\lambda_1 = 1.0, \lambda_2 = 0.1, \lambda_3 = 0.01$ by default. These weights represent their relative importance in the total loss.

The *reconstruction loss* directly measures accuracy of the omics-translation task in both translation directions. Let us denote n as the number of patients, x_i as the source expression for patient i, y_i as the target expression, and $\hat{y}_i = f(x_i)$ as the source-to-target translation, f^{-1} as the reverse translation, and MSE as mean squared error. The *reconstruction loss* is calculated as:

$$\mathcal{L}_{\text{re}} = \frac{1}{n} \sum_{i=1}^{n} \left[\text{MSE}\left(f(x_i), y_i\right) + \text{MSE}\left(f^{-1}(y_i), x_i\right) \right]$$

The *cycle consistency loss* enforces invertibility by ensuring that sequential translations through both modalities return to the original input. It is computed as follows:

$$\mathcal{L}_{\text{cy}} = \frac{1}{n} \sum_{i=1}^{n} \left[\text{MSE}\left(f^{-1}(\hat{y}_i), x_i\right) + \text{MSE}\left(f(f^{-1}(y_i)), y_i\right) \right]$$

The *pathway consistency loss*, computed for each KEGG pathway, minimizes the discrepancy between the average expression of pathway genes in the original and reconstructed data. It ensures that functionally related gene-level features maintain coherent expression patterns across different omics types. The loss is calculated as follows:

$$\mathcal{L}_{\text{pa}} = \frac{1}{n|P||G_p|^2} \sum_{i=1}^{n} \sum_{p \in P} \sum_{g \in G_p} [(x_{ig} - \hat{y}_{ig})^2 + (y_{ig} - f^{-1}(y_{ig}))^2]$$

where P is the set of pathways, G_p is the set of genes in pathway p, x_{ig} and y_{ig} are the source and target values for gene g, $\hat{y}_{ig}$ represents the estimated target value for gene g, and $f^{-1}(y_{ig})$ represents the estimated source value for gene g.

Data inference

We perform inference for each of three data types separately. Given an input gene expression matrix, we remove genes in the input data that are not present during training and replace missing expression values with zeros before performing a z-score transformation. Next, we perform cumulative distribution function (CDF) matching to align the distribution of the input data with that of the training data [56]. CDF matching works by ranking the expression values for each gene in the new data, then mapping these ranks to corresponding quantiles in the distribution of the training data and inferring new expression values accordingly.

We use the trained models and preprocessed data to generate data matrices of the target omics types, which are in z-score scale. We then apply inverse z-score transformation, leveraging training statistics of each target modality, to convert the generated data matrices to their normal scale. Finally, we also perform data trimming on the scale-reversed generated data. Specifically, DNA methylation values are clipped to $[0, 1]$ while miRNArpm and miRNAiso values are forced to be non-negative.

Ensemble clustering

The ensemble clustering module measures patient similarities by applying different affinities to each data type, then merges these matrices to generate consensus patient networks. Next, spectral clustering and community detection algorithms partition these networks into distinct clusters. Finally, the optimal assignment is selected based on the Silhouette score.

Constructing patient affinity matrices

Given a set of L molecular data matrices, let X_l denote the data matrix for omics l. For each X_l, we generate three distinct affinity matrices: (i) Euclidean affinity, (ii) angular affinity, and (iii) aggregated affinity.

The Euclidean and angular affinity matrices are computed using the radial basis function kernel:

$$M_l(i,j) = \exp\left(-\frac{d(x_{li}, x_{lj})^2}{2\sigma^2}\right)$$

where $M_l(i,j)$ represents the affinity matrix between samples i and j for the l omics type, x_{li} and x_{lj} are their respective profiles, $d(x_{li}, x_{lj})$ denotes the distance between them, and σ is a normalizing parameter that controls how quickly the similarity decays with distance. For the Euclidean affinity matrix, $d(x_{li}, x_{lj})$ is the Euclidean distance, and σ is defined as the 60^{th} percentile of all calculated pairwise Euclidean distances. For the angular affinity matrix, $d(x_{li}, x_{lj})$ is the angle in radians between the two vectors, scaled by π, while σ is set to 0.5.

Previous studies have emphasized the impact of local structures in affinity matrices on the performance of similarity-based clustering algorithms [57, 58]. However, focusing on local connections exclusively can lead to an undesirable inflation of the number of clusters when analyzing small datasets. Therefore, we tailor our approach based on the sample size of each dataset, selecting between full and local affinity matrices accordingly. We employ full affinity matrices (following the formula above) for datasets with fewer than 100 samples and leverage local affinity matrices for datasets with 100 samples or more. Adopting a K-nearest neighbor (KNN) approach [40], we generate the local affinity matrices as follows:

$$LM_l(i,j) = \frac{M_l(i,j) \cdot I(j \in \eta_{li})}{\sum_{r \in \eta_{li}} M_l(i,r)} + \frac{M_l(i,j) \cdot I(i \in \eta_{lj})}{\sum_{r \in \eta_{lj}} M_l(r,j)}$$

where $LM_l(i,j)$ represents the local affinity matrix, η_{li} denotes the set of the 10 nearest neighbors of sample i, and $I(j \in \eta_{li})$ is an indicator function equal to 1 if $j \in \eta_{li}$, and 0 otherwise.

Next, we integrate the Euclidean and angular affinity matrices (in either full or local forms) into an aggregated affinity matrix using the AASC-Eigengap algorithm previously developed [24]. The AASC-Eigengap algorithm iteratively searches for the optimal linear combination of the two component affinity matrices to yield an aggregated affinity matrix with the largest eigengap. In practice, this algorithm may introduce noise when applied to large datasets, especially for those with at least 500 samples. Therefore, for such datasets, we transform the aggregated affinity matrix into its local affinity matrix, employing a neighborhood size of 20 for each sample instead of 10.

Network partitioning

We integrate the affinity matrices (Euclidean, angular, or aggregated) across all data types into their corresponding consensus matrices. The consensus Euclidean and angular affinity matrices are calculated as follows:

$$CM(i,j) = \frac{1}{|OT(i,j)|} \sum_{l \in OT(i,j)} A_l(i,j), A \in \{M, LM\}$$

where $CM(i,j)$ represents the consensus affinity matrix (generated from either Euclidean or angular affinity matrices), $OT(i,j)$ is the set of omics types available for both samples i and j, and $|OT(i,j)|$ is its cardinality.

We also generate a consensus connectivity matrix from the aggregated affinity matrices. For each matrix, we select the number of clusters via the eigengap method and apply spectral clustering to obtain a temporary partitioning of patients. We then transform each partitioning into a binary connectivity matrix, in which 1 indicates that two samples belong to the same cluster, and 0 indicates otherwise. Subsequently, we apply the equation above, replacing the affinity matrices with the connectivity matrices from all omics types, to derive the consensus connectivity matrix.

Given the three consensus matrices, we apply spectral clustering to the first two (the consensus Euclidean and angular affinity matrices) and use the Louvain community detection algorithm to partition the third one (the aggregated affinity matrix). Since Louvain is stochastic, we perform multiple runs and select the solution with the highest modularity score [59]. After clustering, we obtain three different partitionings for the same set of patients. We evaluate each partitioning using the Silhouette score [60] based on a distance matrix derived by

fusing the aggregated affinity matrices of all omics types, and subtracting the resulting matrix from 1. Finally, we output the partitioning with the highest Silhouette score.

Results

We perform a comprehensive analysis of 66 cancer datasets (46 multi-omics and 20 single-omics datasets) covering 28 tissues and over 15 000 samples (Supplementary Table S1). First, we compare the performance of CSIE against 12 state-of-the-art methods using Cox *P*-value and empirical *P*-value. Second, we analyze the usefulness of the identified subtypes as an additional covariate for risk prediction. Finally, we provide an in-depth pathway analysis of the discovered subtypes for the pan-gastrointestinal (pan-GI) cancer.

The 46 multi-omics datasets comprise 33 datasets from GDC/TCGA, nine from cBioPortal, and four from published articles. For TCGA data, we download all 12 matrices measuring mRNA expression, miRNA expression, DNA methylation, CNV, and somatic mutations from GDCs (https://gdc.cancer.gov/), along with clinical variables and survival information. We also download their protein data from Proteomic Data Commons (https://pdc.cancer.gov/pdc/) and metabolite data from Benedetti et al. [61]. Similarly, we download all available data matrices for the nine datasets from cBioPortal (https://www.cbioportal.org/). The four additional datasets are from published articles: P23918603 [62], P24316975 [63], P30244973 [64], and P33577785 [65]. For these datasets, we use preprocessed metabolite levels and gene expression data from Benedetti et al. [61]. Finally, the 20 single-omics datasets containing only gene expression are all downloaded from NCBI GEO (https://www.ncbi.nlm.nih.gov/geo/).

Subtyping of 66 cancer datasets

We benchmark the performance of CSIE against twelve state-of-the-art approaches, including Consensus Clustering (CC) [14], Cancer Integration via Multikernel Learning (CIMLR) [43], Similarity Network Fusion (SNF) [39], Low-Rank Approximation based multi-omics data Clustering (LRACluster) [30], Integrative Non-negative Matrix Factorization (intNMF) [29], Affinity Network Fusion (ANF) [44], NEighborhood based Multi-Omics clustering (NEMO) [40], Multi-Reconstruction Graph Convolutional Network (MRGCN) [35], Hierarchical Multi-Kernel Learning (hMKL) [45], Multi-omics Data Integration for Clustering to identify Cancer subtypes (MDICC) [46], Deep Latent Space Fusion (DLSF) [36] and Deep Structure Integrative Representation (DSIR) [37]. For CSIE, we perform data processing and inference as described in section Materials and methods. Note that we do not perform imputation on any multi-omics datasets. We perform data inference strictly on single-omics datasets (mRNA datasets) to impute miRNA and methylation data from mRNA data. For other methods, we select the top 8000 highly variable features from each data matrix and apply the processing procedure suggested by each method. The feature size of 8000 is chosen to align with CSIE's post-processing data dimensionality, ensuring a fair assessment (see Supplementary Fig. S3 for the impact of gene filtering). We execute each method using their default parameters. All methods can automatically determine the number of subtypes.

We use three pre-defined metrics to quantify the performance of the competing methods. First, we utilize the Cox proportional hazards (PH) model [66] to assess the statistical significance in survival differences across the identified subtypes (where smaller *P*-values indicate better performance). Second, we also assess the statistical significance of these survival differences using an empirical *P*-value method that mitigates potential bias associated with the number of subtypes (Supplementary Fig. S4). Third, we calculate the Concordance Index (C-Index) [67] to evaluate whether incorporating the clustering information improves the discrimination power of survival prediction (where higher C-Indices indicate better performance). We also perform the Schoenfeld residual test [68] of the PH assumption (Supplementary Fig. S5). With the exception of DSIR, the remaining 12 methods satisfy the assumption across the vast majority of the 66 datasets.

Table 1 presents the Cox *P*-values, which are computed using the *survival* package (https://cran.r-project.org/web/packages/survival/). NA entries in the table indicate analysis failures. Only CSIE, NEMO, and MRGCN are able to perform subtyping for all datasets. CSIE, NEMO, and MRGCN are capable of handling missing data across data types, which allows them to analyze all available samples for each dataset. In contrast, the remaining methods require completely matched samples among data types, restricting their analysis to sample intersection. These methods crash when analyzing TCGA-Breast Invasive Carcinoma (BRCA), TCGA-Colon Adenocarcinoma (COAD), TCGA-Ovarian Serous Cystadenocarcinoma (OV), difg-glass, luad-oncosg-2020, rectal-msk-2022, sclc-ucologne-2015, and GSE13041 datasets. Overall, CSIE outperforms all competitors in identifying subtypes with significantly different survival profiles in most datasets (48 out of 66). The next best methods are NEMO, hMKL, SNF and ANF, with significant Cox *P*-values in 31, 21, 20, and 20 datasets, respectively. All remaining methods yield significant *P*-values in less than 20 datasets.

Figure 2 shows the −log10 *P*-values of both the Cox and the empirical *P*-values of all 13 methods across 66 cancer datasets. The distribution of the −log10 Cox *P*-values of CSIE has a median value of 1.76, which is substantially higher than those of all other methods. The second best method, NEMO, has a median of 1.13, a value that falls below the commonly used 5% significance threshold (−log10 of 0.05 is ∼1.3). Similarly, the −log10 empirical *P*-value of CSIE has a median of 1.54. This considerably exceeds that of the second best method, NEMO, with a median of 0.88. Overall, CSIE achieves the highest number of statistically significant results: 48 datasets based on the Cox *P*-value and 42 datasets based on the empirical *P*-value. Detailed empirical *P*-values for each method and each dataset can be found in Supplementary Table S2.

To gain further insights, we separately assess the methods on multi-omics and single-omics datasets. As shown in Fig. 3 for the 46 multi-omics datasets, the −log10 *P*-values of CSIE exhibits a median value of 1.93, which is substantially higher than all competitors. Similarly, the −log10 empirical *P*-value median for CSIE is 1.74, considerably exceeding those of comparison methods. CSIE also yields the highest number of datasets with significant *P*-values: 34 by Cox *P*-value and 32 by empirical *P*-value. Additionally, CSIE maintains strong performance across the 20 single-omics datasets, as shown in Fig. 4. We include the results for all methods in both scenarios: (i) using the combination of the original and imputed data, and (ii) using only the original mRNA data. When incorporating the imputed data, the number of datasets with significant Cox *P*-values increases for 10 out of 13 methods (CSIE, CC, SNF, LRACluster, intNMF, ANF, hMKL, MDICC, DLSF, and DSIR). Similarly, the number of datasets with significant empirical *P*-values also increases for 8 out of 13 methods (CSIE, CC, SNF, LRACluster, intNMF, ANF, DLSF, and DSIR). CSIE with the imputed data outperforms existing

Table 1 Cox P-values and number of subtypes, formatted as **P-value (#Subtypes)**, identified by all methods for 66 cancer datasets. Significant P-values are defined as less than 0.05.

Dataset	CSIE	CC	CIMLR	SNF	LRA- Cluster	intNMF	ANF	NEMO	MRGCN	hMKL	MDICC	DLSF	DSIR
TCGA-ACC	5E-7(3)	2E-2(3)	6E-2(6)	2E-4(3)	7E-1(2)	8E-4(4)	1E-3(3)	**3E-8(5)**	6E-1(2)	5E-3(8)	1E-1(3)	7E-3(4)	5E-4(5)
TCGA-BLCA	**1E-4(10)**	1E-1(3)	3E-1(8)	1E-1(2)	3E-2(2)	5E-2(2)	8E-1(2)	7E-4(7)	6E-1(2)	3E-1(9)	2E-1(2)	5E-1(7)	2E-1(10)
TCGA-BRCA	**4E-2(10)**	NA(NA)	NA(NA)	NA(NA)	NA(NA)	NA(NA)	NA(NA)	5E-1(2)	5E-1(2)	NA(NA)	NA(NA)	NA(NA)	NA(NA)
TCGA-CESC	**8E-3(10)**	6E-1(3)	4E-2(7)	1E-1(2)	3E-1(2)	5E-1(4)	9E-2(2)	1E-1(10)	9E-1(2)	5E-2(10)	6E-1(2)	1E-2(5)	7E-2(8)
TCGA-CHOL	8E-1(2)	7E-1(3)	1E+0(2)	3E-1(2)	3E-1(2)	4E-1(3)	**2E-1(2)**	5E-1(7)	2E-1(2)	NA(NA)	1E+0(2)	NA(NA)	2E-1(2)
TCGA-COAD	**2E-2(10)**	NA(NA)	NA(NA)	NA(NA)	NA(NA)	NA(NA)	NA(NA)	8E-1(3)	1E-1(2)	NA(NA)	NA(NA)	NA(NA)	NA(NA)
TCGA-DLBC	8E-1(2)	6E-1(3)	4E-1(2)	9E-1(2)	6E-1(4)	**2E-1(2)**	4E-1(2)	6E-1(3)	3E-1(2)	NA(NA)	8E-1(2)	NA(NA)	NA(NA)
TCGA-ESCA	5E-1(2)	2E-3(5)	3E-1(2)	4E-3(2)	**4E-4(3)**	2E-2(2)	4E-3(2)	4E-1(2)	1E-1(2)	1E-2(3)	5E-3(2)	3E-1(3)	9E-1(9)
TCGA-GBM	**7E-3(9)**	2E-2(3)	6E-1(2)	6E-1(2)	3E-1(2)	1E-1(2)	4E-2(3)	2E-1(8)	6E-2(2)	9E-1(3)	2E-2(4)	NA(NA)	3E-2(3)
TCGA-HNSC	6E-3(6)	5E-1(3)	5E-1(8)	3E-1(2)	1E+0(2)	1E-1(3)	2E-1(2)	**2E-3(6)**	1E-1(2)	4E-1(4)	4E-1(2)	2E-1(4)	7E-1(9)
TCGA-KICH	8E-5(2)	**4E-6(3)**	6E-1(5)	1E-1(2)	3E-1(4)	2E-5(4)	4E-3(2)	2E-5(8)	2E-1(2)	4E-2(5)	2E-2(2)	9E-3(6)	4E-1(3)
TCGA-KIRC	3E-9(4)	3E-1(3)	2E-1(6)	6E-1(2)	5E-1(2)	5E-1(3)	7E-1(2)	**2E-10(5)**	8E-10(2)	1E-1(7)	4E-1(3)	2E-1(3)	8E-2(10)
TCGA-KIRP	**2E-16(6)**	4E-1(3)	8E-3(10)	2E-2(2)	2E-2(2)	4E-2(4)	3E-2(2)	1E-4(2)	3E-1(2)	1E-4(9)	6E-1(5)	8E-2(6)	2E-2(9)
TCGA-LAML	4E-4(7)	1E-1(3)	4E-1(8)	3E-2(3)	2E-2(2)	2E-1(5)	3E-2(3)	**2E-4(6)**	3E-1(2)	5E-1(10)	8E-1(2)	1E-1(6)	3E-2(4)
TCGA-LGG	**2E-30(10)**	6E-8(3)	3E-3(4)	1E-15(2)	4E-3(2)	6E-3(2)	1E-10(2)	7E-30(4)	7E-8(2)	3E-2(2)	7E-1(3)	2E-7(7)	9E-1(8)
TCGA-LIHC	2E-3(9)	4E-1(3)	1E-1(2)	8E-1(2)	4E-2(2)	1E-1(2)	7E-1(2)	5E-1(2)	**3E-4(2)**	8E-1(2)	6E-1(3)	1E-3(5)	3E-2(8)
TCGA-LUAD	**5E-5(10)**	5E-1(3)	9E-2(3)	4E-1(2)	2E-1(2)	7E-2(5)	9E-2(2)	2E-1(2)	5E-2(2)	5E-1(3)	3E-1(2)	9E-1(4)	8E-1(6)
TCGA-LUSC	**8E-3(10)**	5E-2(3)	2E-1(3)	9E-1(2)	2E-1(3)	3E-2(2)	2E-1(2)	3E-2(2)	1E-1(2)	3E-1(2)	3E-2(2)	2E-1(5)	1E-1(8)
TCGA-MESO	**7E-4(3)**	2E-1(3)	3E-1(6)	3E-2(2)	5E-1(2)	1E-2(3)	3E-2(3)	1E-1(3)	2E-3(2)	2E-3(8)	2E-1(2)	9E-3(3)	3E-1(2)
TCGA-OV	4E-2(8)	NA(NA)	NA(NA)	NA(NA)	NA(NA)	NA(NA)	NA(NA)	**1E-3(4)**	4E-1(2)	NA(NA)	NA(NA)	NA(NA)	NA(NA)
TCGA-PAAD	1E-2(10)	1E-2(3)	2E-1(3)	1E-2(3)	2E-2(3)	4E-1(3)	1E-2(3)	5E-3(2)	3E-1(2)	9E-2(6)	5E-1(2)	2E-1(5)	**8E-5(7)**
TCGA-PCPG	9E-1(10)	6E-1(3)	2E-1(2)	1E-1(3)	4E-1(2)	6E-1(2)	**1E-1(2)**	7E-1(2)	2E-1(2)	2E-1(9)	5E-1(2)	2E-1(4)	3E-1(5)
TCGA-PRAD	**3E-2(10)**	3E-1(3)	5E-1(6)	7E-1(2)	9E-1(2)	7E-1(2)	9E-1(2)	4E-1(2)	2E-1(2)	3E-1(8)	8E-1(2)	7E-1(5)	6E-2(9)
TCGA-READ	6E-1(2)	5E-1(3)	6E-1(7)	3E-1(2)	3E-1(2)	7E-1(3)	7E-1(2)	7E-1(3)	7E-1(2)	**3E-6(9)**	5E-1(3)	2E-1(6)	4E-1(3)
TCGA-SARC	3E-4(9)	1E-3(4)	4E-1(6)	8E-1(2)	8E-1(2)	2E-1(2)	2E-2(3)	3E-1(3)	6E-1(2)	**4E-9(8)**	3E-1(2)	6E-2(6)	4E-2(7)
TCGA-SKCM	9E-2(2)	1E+0(3)	**8E-2(8)**	5E-1(5)	5E-1(3)	3E-1(2)	4E-1(5)	2E-1(10)	3E-1(2)	5E-1(10)	5E-1(3)	7E-1(6)	5E-1(3)
TCGA-STAD	**2E-2(3)**	2E-1(3)	1E-1(3)	3E-1(3)	4E-1(3)	1E-1(5)	4E-1(3)	4E-2(3)	8E-1(2)	3E-1(6)	9E-2(2)	4E-1(7)	8E-2(6)
TCGA-TGCT	1E-1(9)	3E-1(3)	6E-1(7)	1E-1(2)	1E-1(2)	**1E-1(4)**	1E-1(2)	7E-1(3)	5E-1(2)	2E-1(5)	4E-1(3)	6E-1(5)	7E-1(7)
TCGA-THCA	1E-2(10)	1E+0(3)	**1E-3(10)**	6E-3(3)	4E-1(2)	7E-1(3)	2E-2(3)	4E-2(4)	3E-1(2)	2E-2(5)	6E-1(2)	4E-1(4)	5E-1(10)
TCGA-THYM	4E-2(7)	7E-1(3)	8E-1(6)	4E-1(3)	4E-1(3)	6E-1(3)	3E-1(3)	**1E-3(5)**	1E-2(2)	6E-3(9)	4E-1(2)	5E-1(4)	8E-1(6)
TCGA-UCEC	1E-4(10)	3E-1(3)	2E-2(7)	1E-3(2)	3E-2(2)	6E-3(2)	5E-3(2)	**1E-5(4)**	1E-3(2)	3E-2(10)	3E-1(2)	5E-4(4)	1E+0(8)
TCGA-UCS	4E-1(4)	4E-1(3)	6E-1(4)	5E-1(3)	2E-1(2)	5E-1(3)	4E-1(3)	3E-2(10)	8E-1(2)	3E-1(6)	**3E-2(2)**	6E-1(4)	4E-1(3)
TCGA-UVM	**5E-5(2)**	5E-1(3)	1E-1(5)	7E-3(2)	1E-1(2)	2E-2(2)	7E-3(2)	2E-2(9)	9E-1(2)	3E-1(8)	1E-2(3)	3E-1(7)	1E-1(4)
coad_silu_2022	4E-1(7)	4E-1(3)	8E-1(10)	4E-2(2)	7E-1(2)	9E-1(3)	5E-2(2)	1E-1(2)	1E-1(2)	**6E-3(9)**	1E-1(2)	4E-2(5)	6E-1(8)
difg_glass_2019	2E-2(10)	3E-2(3)	1E-1(4)	5E-3(2)	2E-1(2)	3E-2(3)	5E-3(2)	**4E-4(8)**	3E-3(2)	1E-1(3)	6E-1(6)	NA(NA)	8E-3(4)
difg_glass	7E-16(8)	NA(NA)	NA(NA)	NA(NA)	NA(NA)	NA(NA)	NA(NA)	**2E-23(3)**	3E-7(2)	NA(NA)	NA(NA)	NA(NA)	NA(NA)
luad_cptac_gdc	**6E-5(10)**	2E-1(3)	6E-3(4)	1E-2(2)	2E-3(2)	3E-4(2)	1E-1(4)	4E-2(4)	7E-4(2)	5E-1(2)	3E-1(2)	2E-2(3)	6E-2(10)
luad_oncosg_2020	1E-2(10)	NA(NA)	NA(NA)	NA(NA)	NA(NA)	NA(NA)	NA(NA)	**6E-3(3)**	3E-2(2)	NA(NA)	NA(NA)	NA(NA)	NA(NA)
ohnca_cptac_gdc	7E-1(2)	5E-1(3)	5E-1(6)	**4E-1(2)**	6E-1(4)	5E-1(5)	6E-1(2)	9E-1(6)	6E-1(2)	1E+0(10)	7E-1(2)	9E-1(5)	4E-1(10)
rcc_cptac_gdc	**4E-5(8)**	1E-3(3)	9E-2(5)	7E-4(3)	8E-2(2)	3E-1(2)	2E-3(3)	2E-3(3)	6E-3(2)	6E-3(9)	1E+0(2)	3E-1(5)	3E-2(5)
rectal_msk_2022	6E-1(8)	NA(NA)	NA(NA)	NA(NA)	NA(NA)	NA(NA)	NA(NA)	**3E-1(8)**	7E-1(2)	NA(NA)	NA(NA)	NA(NA)	NA(NA)
sclc_ucologne_2015	**4E-3(10)**	NA(NA)	NA(NA)	NA(NA)	NA(NA)	NA(NA)	NA(NA)	3E-1(8)	5E-3(2)	NA(NA)	NA(NA)	NA(NA)	NA(NA)

(continued)

Table 1 Continued.

Dataset	CSIE	CC	CIMLR	SNF	LRA- Cluster	intNMF	ANF	NEMO	MRGCN	hMKL	MDICC	DLSF	DSIR
P23918603	4E-2(2)	1E-1(3)	4E-2(4)	2E-1(2)	6E-2(2)	3E-2(2)	2E-1(2)	4E-1(6)	5E-1(2)	5E-2(5)	1E-1(4)	**2E-2(6)**	1E-1(5)
P24316975	**3E-2(2)**	5E-2(3)	1E-1(2)	3E-1(2)	9E-2(2)	5E-2(2)	3E-1(2)	2E-1(2)	1E-1(2)	2E-1(2)	8E-1(2)	5E-1(3)	4E-1(10)
P30244973	3E-2(10)	4E-1(3)	**2E-4(10)**	6E-1(3)	4E-1(3)	2E-2(4)	1E-1(3)	9E-1(6)	6E-1(2)	7E-1(3)	6E-1(3)	1E-1(5)	5E-1(7)
P33577785	9E-2(2)	4E-1(3)	**7E-3(4)**	8E-1(2)	4E-1(2)	7E-1(2)	8E-1(2)	8E-1(4)	6E-1(2)	2E-2(7)	1E+0(2)	6E-1(3)	8E-1(9)
GSE103479	7E-1(2)	1E+0(3)	9E-1(5)	1E+0(3)	9E-1(2)	2E-1(2)	8E-1(3)	9E-1(4)	8E-1(3)	9E-1(5)	7E-1(2)	**2E-1(3)**	NA(NA)
GSE13041	**9E-2(9)**	NA(NA)	NA(NA)	NA(NA)	NA(NA)	NA(NA)	NA(NA)	5E-1(4)	3E-1(2)	NA(NA)	NA(NA)	NA(NA)	NA(NA)
GSE1456	**3E-4(7)**	1E-3(3)	8E-2(3)	9E-2(3)	6E-3(2)	8E-2(3)	2E-1(3)	2E-2(6)	6E-1(2)	1E-1(10)	2E-2(2)	8E-2(6)	9E-2(4)
GSE150615_1	**3E-16(8)**	2E-3(4)	4E-3(2)	4E-3(2)	4E-3(2)	6E-2(3)	4E-3(2)	2E-2(5)	3E-1(2)	3E-2(6)	8E-1(2)	1E-2(4)	NA(NA)
GSE150615_2	4E-3(7)	2E-1(3)	3E-4(6)	1E-1(4)	1E-1(4)	1E-1(3)	9E-2(4)	2E-1(6)	3E-3(2)	**1E-5(8)**	6E-1(3)	1E-1(7)	NA(NA)
GSE17536	1E-2(6)	9E-2(3)	5E-3(7)	6E-1(4)	3E-3(2)	5E-2(2)	1E-1(4)	1E-1(7)	3E-1(2)	**5E-4(10)**	8E-1(2)	2E-3(3)	NA(NA)
GSE17537	5E-3(2)	1E-1(3)	2E-1(7)	3E-1(3)	1E-1(6)	6E-2(2)	2E-1(3)	**2E-4(9)**	1E-1(2)	5E-1(3)	4E-1(2)	9E-2(4)	NA(NA)
GSE20685	2E-4(6)	6E-1(3)	4E-1(3)	4E-1(2)	8E-1(2)	4E-1(2)	5E-1(2)	2E-2(5)	4E-2(2)	**9E-5(7)**	5E-2(2)	1E-3(6)	NA(NA)
GSE21501	4E-2(4)	**3E-4(5)**	2E-1(10)	1E-1(3)	3E-1(2)	3E-1(2)	9E-2(3)	2E-3(10)	2E-1(2)	6E-2(3)	2E-1(2)	5E-1(4)	NA(NA)
GSE42669	2E-1(2)	2E-1(3)	3E-1(9)	5E-2(3)	1E-1(2)	3E-1(3)	5E-2(3)	**1E-3(6)**	7E-2(2)	3E-1(6)	6E-1(2)	3E-1(4)	NA(NA)
GSE4412	3E-2(3)	7E-3(3)	3E-2(9)	2E-1(3)	5E-2(2)	2E-1(2)	2E-2(2)	**1E-3(9)**	5E-2(2)	2E-2(9)	7E-2(2)	2E-1(3)	3E-2(8)
GSE57495	8E-1(3)	4E-1(3)	5E-1(6)	2E-1(4)	5E-1(3)	2E-1(2)	2E-1(4)	**2E-4(9)**	5E-2(2)	1E-1(7)	8E-1(3)	5E-1(3)	NA(NA)
GSE61335	3E-2(5)	4E-2(3)	1E-1(7)	1E-2(3)	8E-3(2)	NA(NA)	1E-2(3)	7E-2(8)	**3E-3(2)**	2E-1(9)	2E-2(2)	4E-2(3)	7E-2(7)
GSE62452	4E-2(2)	5E-1(3)	**2E-2(5)**	5E-2(2)	3E-1(2)	3E-1(2)	1E-1(2)	8E-2(7)	1E-1(2)	1E-1(9)	4E-1(3)	4E-1(3)	NA(NA)
GSE71729	3E-2(8)	1E-1(3)	**7E-3(10)**	3E-1(3)	8E-1(2)	9E-1(2)	2E-1(3)	1E-1(10)	9E-1(2)	2E-1(5)	8E-1(2)	1E+0(3)	NA(NA)
GSE72951	4E-1(5)	6E-1(3)	8E-1(4)	**4E-2(4)**	5E-1(2)	3E-1(2)	2E-1(4)	9E-1(4)	1E+0(2)	8E-1(7)	1E-1(2)	6E-1(4)	NA(NA)
GSE74187	4E-2(4)	2E-3(3)	4E-2(8)	**2E-3(2)**	2E-1(3)	9E-3(2)	**2E-3(2)**	7E-2(9)	6E-3(2)	9E-2(10)	2E-1(2)	2E-1(3)	NA(NA)
GSE78229	3E-2(2)	3E-1(3)	1E-1(8)	**2E-2(2)**	2E-1(2)	3E-1(3)	2E-1(2)	4E-1(7)	4E-1(2)	6E-1(7)	2E-1(2)	3E-1(3)	NA(NA)
GSE85916	4E-1(2)	5E-1(3)	1E-2(8)	1E+0(2)	4E-2(2)	3E-2(2)	4E-1(2)	8E-3(6)	7E-1(2)	**2E-3(7)**	2E-1(2)	9E-2(4)	NA(NA)
GSE87211	**6E-3(8)**	5E-1(3)	3E-1(3)	3E-1(2)	7E-2(2)	6E-1(2)	4E-1(2)	1E-1(3)	7E-1(2)	5E-1(5)	5E-1(2)	4E-1(3)	NA(NA)
Sig. datasets	**48**	**16**	**17**	**20**	**14**	**16**	**20**	**31**	**19**	**21**	**8**	**14**	**10**
Ave. clus. no.	**6.14**	**3.1**	**5.64**	**2.5**	**2.36**	**2.7**	**2.55**	**5.27**	**2.02**	**6.55**	**2.4**	**4.5**	**6.56**

Cells with bold text have the most significant *P*-value in each row. The last two rows report the number of datasets with significant *P*-values and the average number of subtypes.

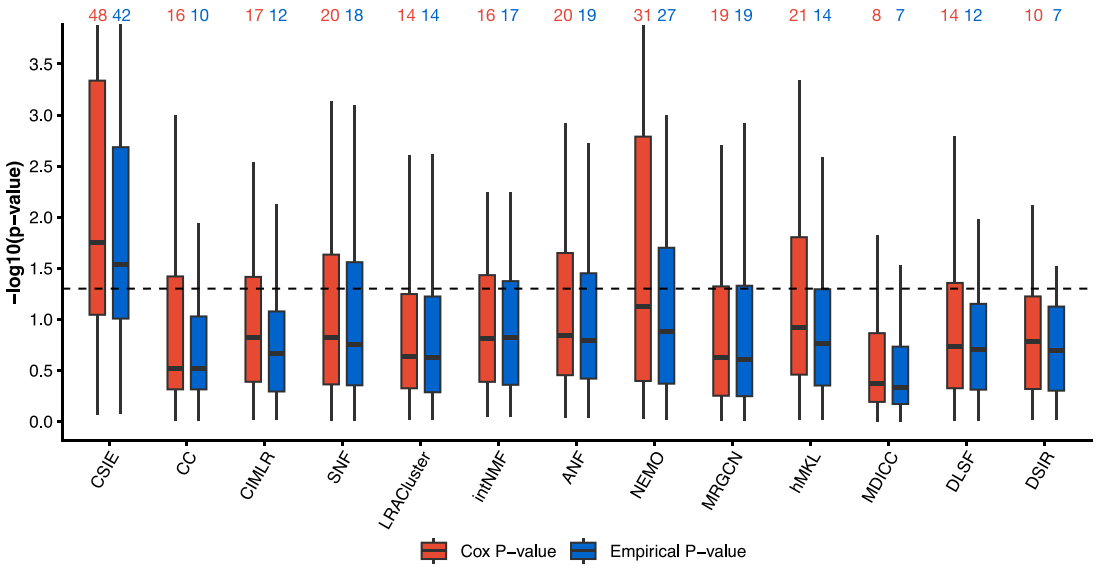


Figure 2 Distribution of Cox *P*-values and empirical *P*-values (in minus log10) for subtypes identified by 13 methods across all 66 cancer datasets, where the number on top of each box represents the total number of datasets with significant *P*-values, the dashed horizontal line denotes the statistical significance threshold corresponding to 5%, and CSIE has the highest median values of 1.76 and 1.54 for Cox and empirical *P*-values, respectively.

methods, achieving the highest number of significant datasets and the highest median −log10 Cox and empirical *P*-values. This robust performance validates the dual contribution of the CSIE framework: the efficacy of the data inference module and the superiority of the subtyping algorithm.

Among the 13 subtyping methods under evaluation, only CSIE, NEMO, and MRGCN can handle missing data. The remaining 10 methods lack this functionality, which restricts the above analysis to sample intersection for them. To evaluate their performance on complete sample sets, we conduct an additional analysis where missing values (NAs) are replaced with zeros for these 10 methods. This enables subtyping on the union of samples within each dataset. Supplementary Fig. S1 shows that six methods (CC, CIMLR, SNF, intNMF, MDICC, and DLSF) have more datasets with significant *P*-values when using the sample union. However, these methods remain inferior to NEMO and are substantially outperformed by CSIE.

To perform a more comprehensive analysis, we also calculate the FDR-corrected *P*-values [69] (Supplementary Table S6) and hazard ratios (Supplementary Table S7) for each method across 66 datasets. CSIE remains the top method, achieving significant FDR-corrected *P*-values in 34 (Cox) and 32 (empirical) datasets, compared to 22 and 16 for NEMO, the second-best method. Similarly, CSIE has the highest number of datasets with high hazard ratios (HR > 2). Figure 5 shows both statistics (FDR-corrected *P*-values and hazard ratios) in log scale. CSIE outperforms all other methods by a large margin, having the highest number of datasets that are both significant and have high hazard ratios (log2 HR > 1). Specifically, CSIE achieves 33 such datasets compared to 20 for NEMO, the second-best method. The performance gap between the two top methods using hazard ratios and corrected Cox *P*-values (an increase from 20 to 33, or 65%) is even higher than the gap when using nominal Cox *P*-values (an increase from 31 to 48, or 54%). We also conduct the same analysis using median HRs (Supplementary Fig. S6). The conclusion remains consistent but the gap in performance is even larger where CSIE outperforms NEMO by 80% (15 to 27).

Next, we evaluate the usefulness of CSIE-derived subtypes for risk prediction. Specifically, we examine the performance of a risk prediction method under two different scenarios: (i) using only clinical variables (age, height, gender, etc.) as predictors, and (ii) incorporating the subtype assignment of each subtyping method as an additional covariate. For each dataset, we perform 5-fold cross-validation and measure the C-Index on the validation sets. At each iteration, we perform subtyping within the training set and then we select the top 2000 features for each omics type based on the ANOVA's F-statistic [70]. Given a testing sample, we assign the clustering label using KNN and cosine distance. A blockForest model [71] was used to predict risk scores using clinical variables and the assigned subtype labels. As shown in Fig. 6, CSIE achieves the highest mean C-Index value of 0.663, compared to 0.65 obtained by NEMO and MRGCN, the second-best methods. Detailed C-Index for each method and dataset can be found in Supplementary Table S3.

In all of the above analyses, we use the overall survival as the ground truth to benchmark the performance of the 13 subtyping methods. To understand whether CSIE can achieve equally good results for tumor metastasis and recurrence, we conduct additional analyses using 37 cancer datasets containing recurrence information and 27 datasets containing metastasis information (Supplementary Table S1). For the recurrence analysis, we define disease-free survival (DFS) as the time from study entry until recurrence. Similarly, metastasis-free survival (MFS) is defined as the time from study entry until the observation of metastasis. We evaluate the subtyping methods using Cox *P*-value and empirical *P*-value calculated using DFS and MFS. In all analyses, CSIE outperforms all other 12 methods, yielding more significant Cox and empirical *P*-values (Supplementary Section 3).

We analyze the 37 datasets with recurrence information (33 multi-omics and 4 single-omics), covering 25 tissues and over 11 000 samples. Supplementary Fig. S7 shows the Cox and empirical *P*-values (−log10 scale) for all 13 methods. The −log10 Cox *P*-values for CSIE exhibit a median of 1.35, substantially higher than that of the

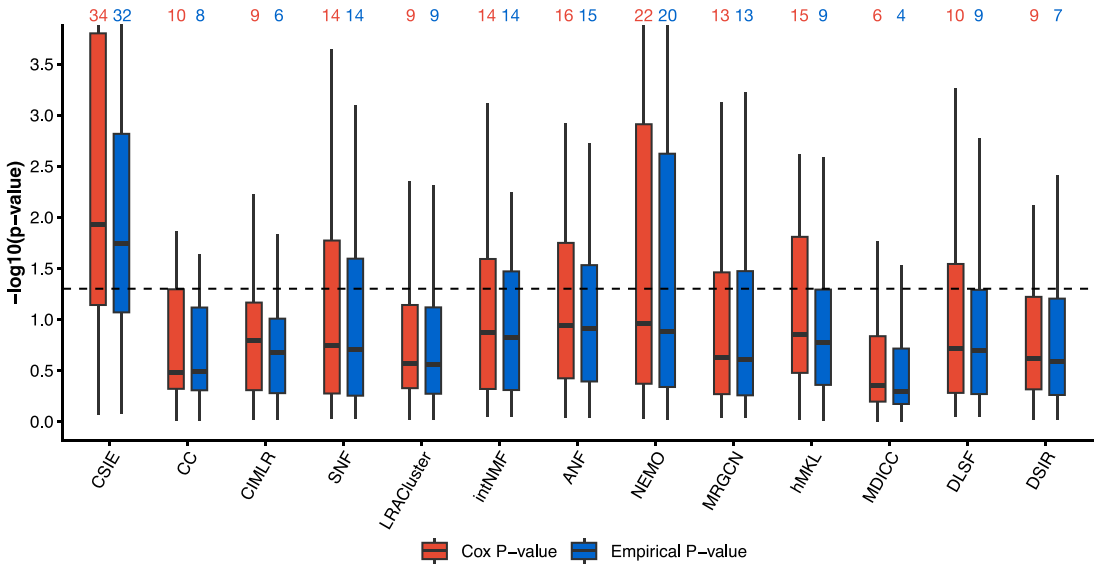


Figure 3 Distribution of Cox P-values and empirical P-values (in minus log10) for subtypes identified by each method across 46 multi-omics cancer datasets, where the number on top of each box represents the total number of datasets with significant P-values, the dashed horizontal line denotes the statistical significance threshold corresponding to 5%, and CSIE has the highest median values of 1.93 and 1.74 (for Cox and empirical P-values, respectively).

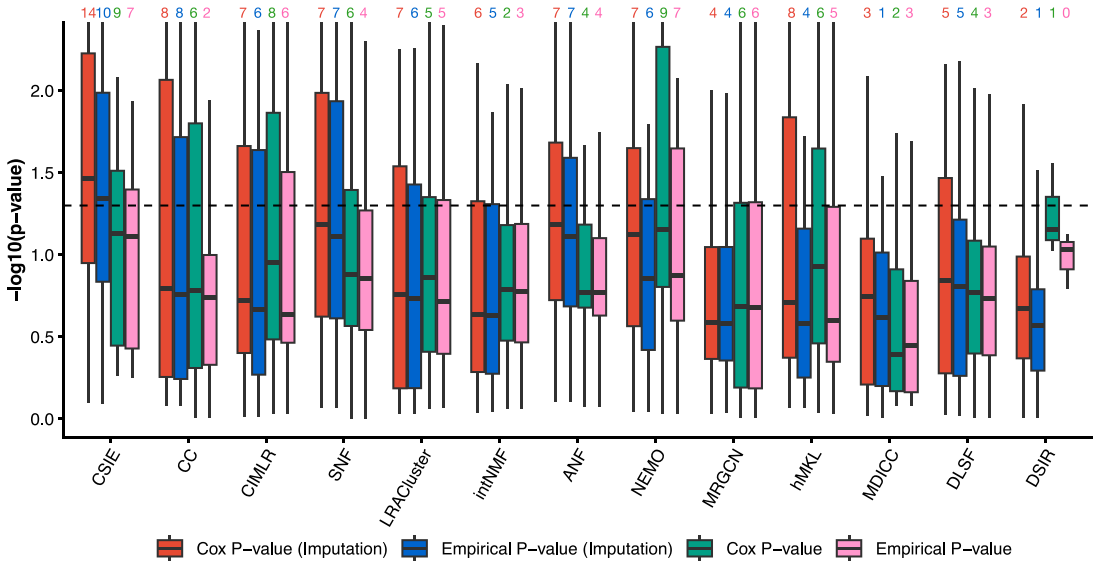


Figure 4 Cox and empirical P-values (−log10 scale) for subtypes identified by 13 methods across 20 single-omics datasets with and without data inference, where incorporating the imputed miRNA and DNA methylation data increases the number of datasets with significant Cox P-values and empirical P-values for most of the methods, and CSIE with the imputed data achieves the highest number of significant datasets and the highest median −log10 Cox and empirical P-values.

second-best method, NEMO, which has a median of 0.69. Similarly, the median −log10 empirical P-value for CSIE is 1.25, exceeding the result of the second-best method, intNMF, which has a median of 0.64. CSIE also achieves the highest number of datasets with significant Cox and empirical P-values (19 and 15, respectively). CC has the second-highest number for Cox P-values (9), while NEMO and MRGCN have the second-highest for empirical P-values (7). The P-values for each dataset and method are provided in Supplementary Table S8.

Similarly, we analyze the 27 datasets with metastasis information (26 multi-omics and one single-omics), covering 21 tissues and over 8000 samples. Supplementary Fig. S8 shows the Cox and empirical P-values (−log10 scale) for all 13 methods across these datasets. The median −log10 Cox and empirical P-values for CSIE are 1.60 and 1.55, respectively, substantially higher than those of any comparison method. The second-best method, NEMO, obtains median −log10 Cox and empirical P-values of 1.06 and 0.87, respectively. Additionally, CSIE achieves the highest number of statistically significant results, with 15 datasets reaching significance for both metrics. NEMO has the second-highest number for both Cox (11) and empirical P-values (12). The detailed results can be found in Supplementary Table S9.

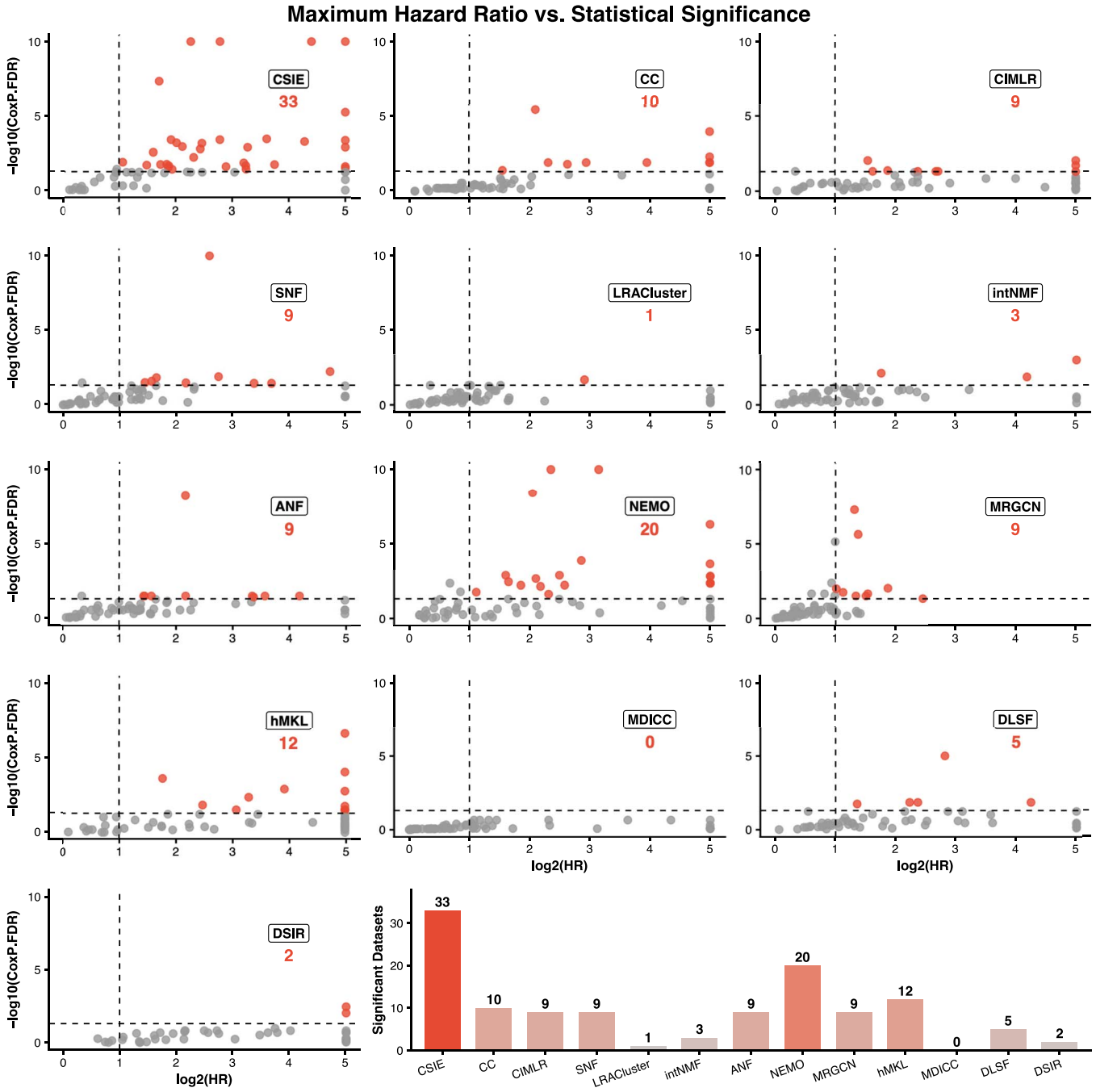


Figure 5 Performance assessment of 13 subtyping methods across 66 datasets using FDR-corrected Cox P-values and maximum hazard ratios, where the x-axis represents the hazard ratio (log2 scale) between the most aggressive and most benign subtypes, the y-axis represents the FDR-corrected Cox P-values ($-$log10 scale), and CSIE has the highest number of datasets (33) that satisfy both statistical significance (FDR-corrected P-value $< .05$) and high hazard ratio (log2 HR > 1).

Case study of gastrointestinal cancer (pan-GI)

We perform an in-depth analysis of the pan-GI cancer dataset, which is compiled by merging four TCGA cohorts: COAD, Liver Hepatocellular Carcinoma (LIHC), Pancreatic Adenocarcinoma (PAAD), and Rectum Adenocarcinoma (READ). To uncover the molecular mechanisms underlying patient survival, we perform molecular subtyping using multi-omics data and CSIE, without incorporating any clinical variables or original cancer labels. This analysis identifies seven subtypes (Fig. 7). Subtype 1 has 293 patients: 206 COAD (70.3%), 83 READ (28.3%), and 4 PAAD (1.4%) patients. Subtype 2 encompasses 221 patients: 152 (68.8%) from COAD, and 69 (31.2%) from READ. Subtype 3 includes 99 patients: 88 (88.9%) from COAD, and 11 (11.1%) from READ. Subtype 4 encompasses 230 patients: 229 (99.6%) from LIHC, and 1 (0.4%) from READ. Subtype 5 contains 139 patients: 138 (99.3%) from LIHC, and 1 (0.7%) from PAAD. Subtype 6 has 63 patients: 61 (96.8%) from PAAD, and 2 (3.2%) from LIHC. Finally, subtype 7 has 122 samples: 117 PAAD (95.9%), 3 READ (2.5%), and 2 COAD (1.6%).

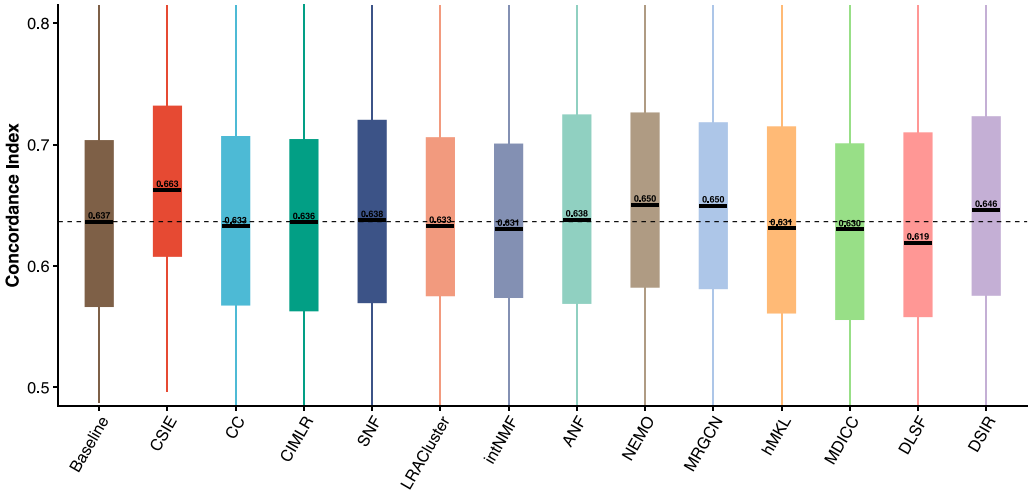


Figure 6 Evaluation of prediction models incorporating subtype labels identified by the 13 subtyping methods across all cancer datasets with clinical data, where we adopt a nested cross-validation strategy in which subtype labels are learned exclusively from the training set and then assigned to samples in the validation set using a KNN and cosine distance, followed by training a blockForest model on either clinical variables alone (baseline) or clinical variables combined with subtype labels from each method, and performance evaluation on the validation set using the C-Index.

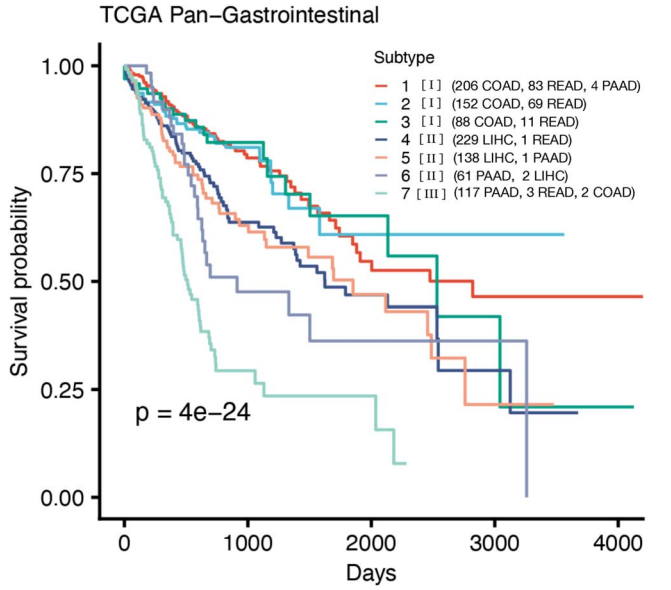


Figure 7 Kaplan–Meier survival analysis of subtypes identified by CSIE for the pan-GI dataset (TCGA-GI), where the horizontal axis represents the days passed after entry into the study and the vertical axis represents the estimated survival probability.

Essentially, CSIE effectively separates pan-GI patients based on their molecular patterns. The seven subtypes, each with a mixture of different cancer types (COAD, LIHC, PAAD, and READ), exhibit profoundly different survival profiles with a highly significant Cox *P*-value of 4×10^{-24}. Interestingly, subtype 7, which shows the lowest survival, consists predominantly of PAAD, the deadliest malignancy within the gastrointestinal tract, with the five-year survival rate of only 13%, compared to 64% of colorectal cancer and 22% of liver cancer [72].

To investigate the mechanisms underlying the survival outcomes of pan-GI patients, we compare the lowest-survival subtype (subtype 7) against all other subtypes using our previously developed consensus pathway analysis (CPA) [73–75]. Briefly, CPA performs differential analysis to identify the differentially expressed (DE) genes and then performs consensus analysis of four pathway analysis methods, ORA [76], GSA [77], FGSEA [78], and PADOG [79]. CPA outputs the list of DE genes, and the list of KEGG pathways that are identified as significantly impacted across all four pathway methods.

Our analysis identifies 2535 genes that have adjusted *P*-values smaller than 1% and absolute log2 fold-changes larger than 1 (Supplementary Table S4). Among these, 89% (2266 genes) are up-regulated. Seven of the top 10 DE genes, INS, GCG, PRSS1, CPA1, PRSS2, PNLIP, and CPB1, are well-known for their encoding pancreatic exocrine enzymes and endocrine hormones, which are frequently affected during pancreatic cancer development [80, 81]. Previous studies have uncovered somatic mutations of PRSS1, CPA1, and CPB1 in patients with pancreatic cancer and chronic pancreatitis as well as the contribution of INS and GCG dysregulation to metabolic derangements in pancreatic cancer [82–84]. The remaining three genes, CLDN18, CTSE, and MUC5AC, are associated with the development of aggressive colorectal cancer subtypes. For instance, CLDN18 functions as a tumor promoter in colorectal cancer, and its overexpression is an indicator of lymph node metastasis [85, 86]. Similarly, overexpression of MUC5AC and CTSE has been linked to chemotherapy resistance in colorectal cancer patients [87, 88].

Figure 8 presents the pathway analysis results, with the horizontal axis displaying the enrichment scores and the vertical axis representing the −log10 adjusted *P*-values. Supplementary Table S5 provides the complete statistics for each pathway. Our analysis identifies 19 pathways with adjusted *P*-values smaller than 5%. Notably, all of these pathways are found to be upregulated. Among these, there are four pathways that have adjusted *P*-values less than 1%: the *Ras signaling pathway*, *Rap1 signaling pathway*, *MAPK signaling pathway* and *Leukocyte transendothelial migration* with adjusted *P*-values of 2.5×10^{-3}, 4.5×10^{-3}, 4.6×10^{-3} and 9.9×10^{-3}, respectively.

The *Ras signaling pathway* and *MAPK signaling pathway* govern fundamental cellular processes such as cell proliferation, differentiation, survival, and metabolic homeostasis through tightly controlled activation-inactivation cycles [89]. However, these two pathways are frequently hijacked by oncogenic alterations in pancreatic and colorectal cancers, with KRAS mutations being the most prominent

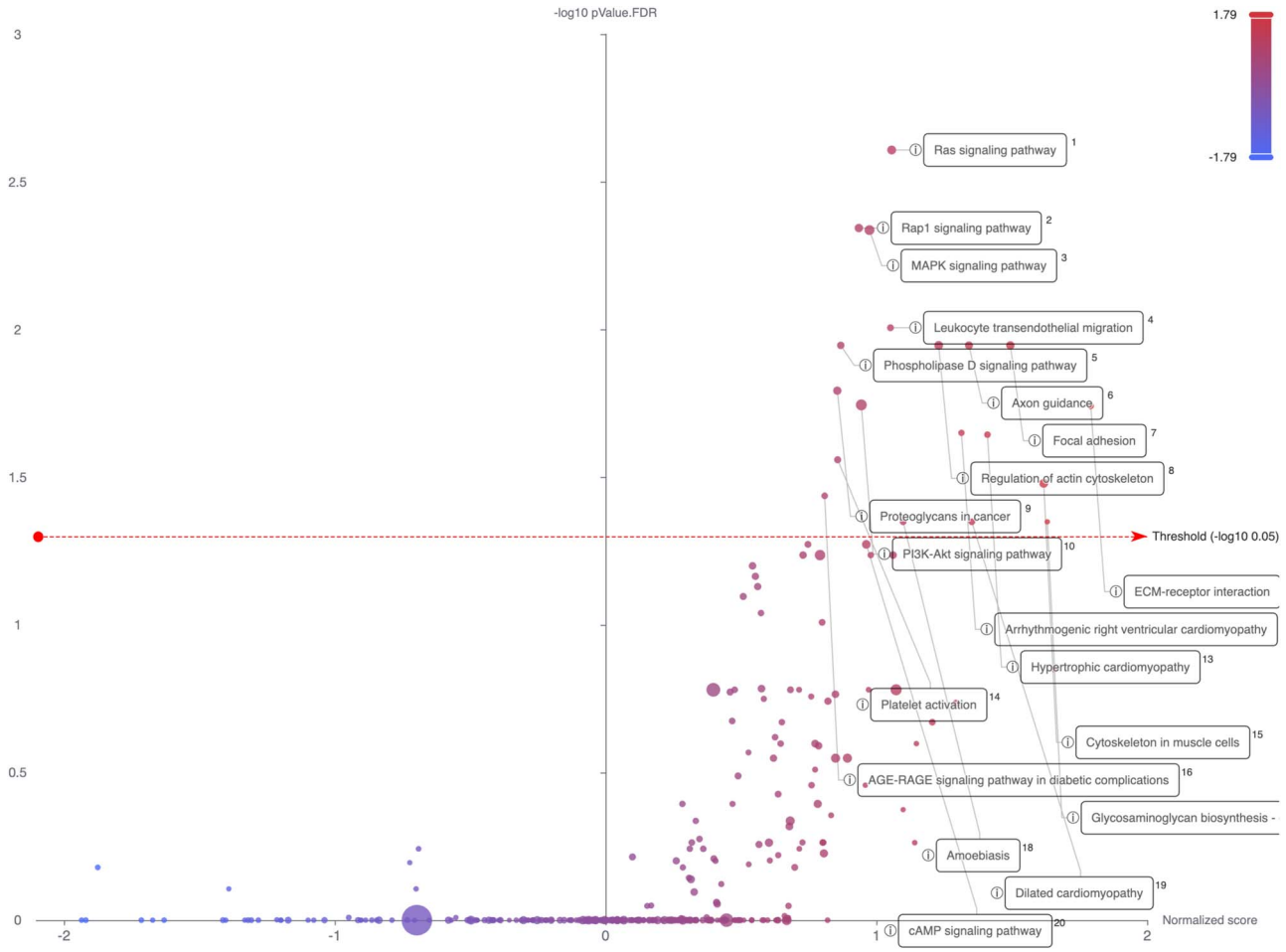


Figure 8 Pathway volcano plot for the CPA of the TCGA-GI gene expression dataset, where the x-axis shows the normalized enrichment score and the y-axis shows the minus log10 of FDR-adjusted *P*-value (pFDR), each point on the figure represents a pathway or gene set, the size of a point is proportional to the number of genes in the corresponding gene set and the color of each point is determined by the normalized enrichment score.

[90, 91]. These mutations constitutively activate the KRAS protein in its GTP-bound state, leading to persistent downstream signaling that promotes uncontrolled proliferation, resistance to apoptosis, angiogenesis, and metabolic reprogramming necessary for cancer cell survival [92, 93]. The concurrent enrichment of both the *Ras* and *MAPK signaling pathways* strongly suggests a coordinated transcriptional dysregulation that further compromises the survival of patients in the studied group.

The *Rap1 signaling pathway* is an important cascade that plays critical roles in cell adhesion, junction formation, and integrin-mediated cell-matrix interactions [94]. Its aberrant signaling can promote Epithelial-Mesenchymal Transition, enhance cell migration and invasion, and facilitate metastatic dissemination, which collectively results in cancer progression and poorer prognosis [95]. Meanwhile, the *leukocyte transendothelial migration pathway* governs the extravasation of immune cells from the bloodstream into tissues and tumors, involving complex interactions between leukocyte integrins, endothelial adhesion molecules, and chemokine gradients [96]. Dysregulation of this pathway may reflect alterations in immune cell infiltration patterns which contribute to immune evasion or indicate an immunosuppressive tumor microenvironment [97]. Alternatively, it may indicate increased inflammatory signaling or the aberrant expression of adhesion molecules caused by tumor cells themselves, facilitating their own transendothelial migration during metastatic spread [98].

Figure 9 demonstrates the strongly connected network of significant pathways, in which a connection between two pathways is established if they share at least 10 genes. This highly interconnected and upregulated network illustrates a prognostic signaling landscape for pan-GI cancer, characterized by aggressive disease progression and reduced patient survival. The underlying mechanism involves tightly coordinated modules driving invasion, chemoresistance, and the formation of a hostile tumor microenvironment.

The cluster of pathways of *ECM-receptor interaction*, *Focal adhesion*, and *Regulation of actin cytoskeleton* naturally forms an integrated mechanosensory network that enables cells to detect, adhere to, and remodel their extracellular environment while coordinating their internal cytoskeletal dynamics for shape changes and migration [99]. High enrichment of this cluster in GI cancer signals a high-risk phenotype, in which cancer cells exploit the strong interconnections among the three pathways to extensively and continually reshape both their external matrix and internal structures to facilitate migration, invasion, and metastatic dissemination [100]. Additionally, these migration-invasion pathways are directly linked to two core signaling hubs that are important for cell proliferation and survival: *PI3K-Akt* and *MAPK signaling pathways* [101]. This structural-to-survival axis

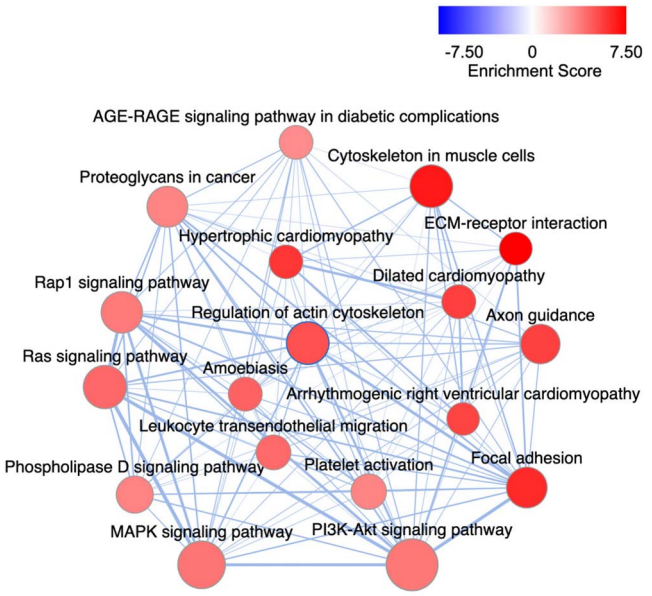


Figure 9 Pathway network of the significant pathways in the TCGA-GI dataset based on the consensus analysis, where each node represents a biological pathway, with node size proportional to the number of genes it contains, edge width indicates the number of shared genes for each pair of two pathways, and the color intensity within each node represents the enrichment score from the consensus result (as shown in the color bar).

coupling means that the functional output of the migration machinery is immediately translated and amplified by the two signalings. In GI cancer, the coordinated hyperactivation of these interconnected modules, often resulted from mutations in upstream regulators such as RAS or PIK3CA, confers a powerful survival advantage for cancer cells [102]. This mechanism directly drives chemoresistance and prevents apoptosis that would normally be triggered by standard chemotherapies [103, 104].

Furthermore, Fig. 9 reveals critical modules that shape the local tumor environment and contribute to overall poor prognosis. The presence of *AGE-RAGE signaling* connects the patient's underlying metabolic status (such as diabetes or obesity) directly to tumor aggression, establishing a chronic inflammatory state [105]. Concurrently, the cluster encompassing the cardiomyopathy nodes (*Hypertrophic*, *Dilated*, etc.) represents a shared genetic signature for severe tissue remodeling and desmoplasia common to many aggressive solid tumors [106, 107].

These findings suggest potential therapeutic treatments focusing on the migration-invasion cluster and the signaling hubs. The former can be targeted through several mechanisms: integrin inhibitors that disrupt ECM-cell adhesion [108], FAK inhibitors that block focal adhesion signaling [109], and Rho/ROCK pathway inhibitors that interfere with cytoskeletal remodeling and cell motility [110]. The latter is already addressed by current FDA-approved targeted therapies for GI cancers. Key agents include inhibitors for KIT/PDGFRA in gastrointestinal stromal tumors such as Imatinib, Trastuzumab for HER2 in gastric cancers [111], and Ramucirumab for VEGFR2 in gastric and colorectal cancers [112, 113]. Beyond these established targets, CLDN18 has emerged as an important opportunity in gastrointestinal malignancies. Zolbetuximab, a monoclonal antibody targeting CLDN18.2, has recently demonstrated significant survival benefits in gastroesophageal junction adenocarcinomas [114].

Conclusion

We present CSIE, a robust framework for cancer subtyping that integrates cross-modal data inference with ensemble clustering. CSIE represents a significant departure from conventional methods as it is the first framework to explicitly generate miRNA and DNA methylation profiles from gene expression data to enhance subtype discovery. Unlike existing Transformer-based models [115–117], CSIE bridges the gap between deep learning and biology by incorporating systems-level information via a gene-pathway MLP and a pathway consistency loss function. Furthermore, CSIE's unique bidirectional training strategy optimizes translation directions within a unified framework, thereby enhancing model robustness. Finally, CSIE features an ensemble clustering module that employs a rigorous and interpretable integration pipeline for diverse omics types.

We benchmark CSIE against 12 state-of-the-art methods (CC, CIMLR, SNF, LRACluster, intNMF, ANF, NEMO, MRGCN, hMKL, MDICC, DLSF, and DSIR) across 66 cancer datasets using three evaluation metrics (Cox *P*-values, empirical *P*-values, and C-Index). CSIE consistently outperforms existing subtyping approaches, achieving more significant *P*-values and higher C-Indices. Through a systematic ablation study, we demonstrate the utility of both the cross-omics inference module and the subtyping algorithm. When incorporating the imputed data, the number of datasets with significant Cox *P*-values increases for the vast majority of methods. Furthermore, CSIE with the imputed data outperforms all comparison methods. These results validate the dual contribution of the CSIE framework: the efficacy of the data inference module and the superiority of the subtyping algorithm. Additionally, a detailed pathway analysis of the pan-GI cancer data demonstrates CSIE's ability to recover known oncogenic processes and uncover potential therapeutic targets in aggressive disease subtypes.

Key Points

- Cancer subtyping via inference and ensemble (CSIE) adopts a transformer-based data inference module to impute missing omics types.
- The method leverages ensemble clustering to identify meaningful disease subtypes.
- The paper presents performance results comparing CSIE with 12 state-of-the-art methods on 66 cancer datasets.
- The paper presents a case study of pan-gastrointestinal cancer.

Author contributions

D.T. developed the method and performed the analysis. P.B., Y.P., and H.L. helped with analysis and case study. J.P. and M.A. helped with analysis results and interpretation. D.T. and T.N. wrote the manuscript. T.N. supervised all aspects of the work. All authors reviewed the manuscript.

Supplementary material

Supplementary material is available at *Briefings in Bioinformatics* online.

Conflict of interest

No competing interest is declared.

Funding

This work is partially supported by the National Science Foundation (NSF: # 2343019 and # 2203236), the National Cancer Institute (NCI: # 1U01CA274573-01A1), the National Institute of General Medical Sciences (NIGMS: # 1R44GM152152-01), the National Institute of Food and Agriculture (NIFA: # 2023-67022-40041), and National Aeronautics and Space Administration (NASA: # 80NSSC22M0255, subaward 23-42). Any opinions, findings, and conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any of the funding agencies.

Data availability

The detailed information of datasets analyzed in this study is provided in Supplementary Table S1. 33 datasets from GDC (https://gdc.cancer.gov/) and PDC (https://pdc.cancer.gov/pdc/) can be accessed using these project IDs: TCGA-ACC, TCGA-BLCA, TCGA-BRCA, TCGA-CESC, TCGA-CHOL, TCGA-COAD, TCGA-DLBC, TCGA-ESCA, TCGA-GBM, TCGA-HNSC, TCGA-KICH, TCGA-KIRC, TCGA-KIRP, TCGA-LAML, TCGA-LGG, TCGA-LIHC, TCGA-LUAD, TCGA-LUSC, TCGA-MESO, TCGA-OV, TCGA-PAAD, TCGA-PCPG, TCGA-PRAD, TCGA-READ, TCGA-SARC, TCGA-SKCM, TCGA-STAD, TCGA-TGCT, TCGA-THCA, TCGA-THYM, TCGA-UCEC, TCGA-UCS, TCGA-UVM.

Nine datasets from cBioPortal (https://www.cbioportal.org/) are public with these IDs: sclc_ucologne_2015, rectal_msk_2022, coad_silu_2022, difg_glass_2019, luad_cptac_gdc, luad_oncosg_2020, ohnca_cptac_gdc, rcc_cptac_gdc, difg_glass.

20 datasets from NCBI GEO (https://www.ncbi.nlm.nih.gov/geo/) can be accessed using these accession IDs: GSE103479, GSE13041, GSE1456, GSE150615, GSE17536, GSE17537, GSE20685, GSE21501, GSE42669, GSE4412, GSE57495, GSE61335, GSE62452, GSE71729, GSE72951, GSE74187, GSE78229, GSE85916, GSE87211. Note that we split GSE150615 into two datasets: GSE150615_1 and GSE150615_2 in our analysis.

Four additional datasets are from published articles: P23918603 (PMID: 23918603), P24316975 (PMID: 24316975), P30244973 (PMID: 30244973), and P33577785 (PMID: 33577785).

Processed data and download instructions, as well as source code of CSIE and scripts for regenerating results reported in this article are available at https://github.com/tinnlab/CSIE.

References

1. Karaman ED, Işik Z. Multi-omics data analysis identifies prognostic biomarkers across cancers. *Med Sci* 2023;**11**:1–24. https://doi.org/10.3390/medsci11030044
2. Kumar-Sinha C, Chinnaiyan AM. Precision oncology in the age of integrative genomics. *Nat Biotechnol* 2018;**36**:46–60. https://doi.org/10.1038/nbt.4017
3. Senft D, Leiserson MDM, Ruppin E *et al.* Precision oncology: the road ahead. *Trends Mol Med* 2017;**23**:874–98. https://doi.org/10.1016/j.molmed.2017.08.003
4. Tran D, Nguyen H, Pham V-D *et al.* A comprehensive review of cancer survival prediction using multi-omics integration and clinical variables. *Brief Bioinform* 2025;**26**:1–17. https://doi.org/10.1093/bib/bbaf150
5. Granja JM, Klemm S, McGinnis LM *et al.* Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nat Biotechnol* 2019;**37**:1458–65. https://doi.org/10.1038/s41587-019-0332-7
6. Burstein MD, Tsimelzon A, Poage GM *et al.* Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer. *Clin Cancer Res* 2015;**21**:1688–98. https://doi.org/10.1158/1078-0432.CCR-14-0432
7. Chaudhary K, Poirion OB, Liangqun L *et al.* Deep learning–based multi-omics integration robustly predicts survival in liver cancer. *Clin Cancer Res* 2018;**24**:1248–59. https://doi.org/10.1158/1078-0432.CCR-17-0853
8. Wang C, Li J, Chen J *et al.* Multi-omics analyses reveal biological and clinical insights in recurrent stage I non-small cell lung cancer. *Nat Commun* 2025;**16**:1–19. https://doi.org/10.1038/s41467-024-55068-2
9. Migliozzi S, Young Taek O, Hasanain M *et al.* Integrative multi-omics networks identify PKCδ and DNA-PK as master kinases of glioblastoma subtypes and guide targeted cancer therapy. *Nat Cancer* 2023;**4**:181–202. https://doi.org/10.1038/s43018-022-00510-x
10. Lindskrog SV, Prip F, Lamy P *et al.* An integrated multi-omics analysis identifies prognostic molecular subtypes of non-muscle-invasive bladder cancer. *Nat Commun* 2021;**12**:1–18. https://doi.org/10.1038/s41467-021-22465-w
11. Mulong D, Dongying G, Xin J *et al.* Integrated multi-omics approach to distinct molecular characterization and classification of early-onset colorectal cancer. *Cell Rep Med* 2023;**4**:100974. https://doi.org/10.1016/j.xcrm.2023.100974
12. Zhao Z, Ding Y, Tran LJ *et al.* Innovative breakthroughs facilitated by single-cell multi-omics: Manipulating natural killer cell functionality correlates with a novel subcategory of melanoma cells. *Front Immunol* 2023;**14**:1–24. https://doi.org/10.3389/fimmu.2023.1196892
13. Charoentong P, Finotello F, Angelova M *et al.* Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. *Cell Rep* 2017;**18**:248–62. https://doi.org/10.1016/j.celrep.2016.12.019
14. Monti S, Tamayo P, Mesirov J *et al.* Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Mach Learn* 2003;**52**:91–118. https://doi.org/10.1023/A:1023949509487
15. Zhiwen Y, Chen H, You J *et al.* Adaptive fuzzy consensus clustering framework for clustering analysis of cancer data. *IEEE/ACM Trans Comput Biol Bioinform* 2014;**12**:887–901. https://doi.org/10.1109/TCBB.2014.2359433
16. Lock EF, Dunson DB. Bayesian consensus clustering. *Bioinformatics* 2013;**29**:2610–6. https://doi.org/10.1093/bioinformatics/btt425
17. Kirk P, Griffin JE, Savage RS *et al.* Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics* 2012;**28**:3290–7. https://doi.org/10.1093/bioinformatics/bts595
18. Liu H, Zhao R, Fang H *et al.* Entropy-based consensus clustering for patient stratification. *Bioinformatics* 2017;**33**:2691–8. https://doi.org/10.1093/bioinformatics/btx167
19. Cabassi A, Kirk PDW. Multiple kernel learning for integrative consensus clustering of omic datasets. *Bioinformatics* 2020;**36**:4789–96. https://doi.org/10.1093/bioinformatics/btaa593
20. Taosheng X, Le TD, Liu L *et al.* CancerSubtypes: an R/Bioconductor package for molecular cancer subtype identification, validation and visualization. *Bioinformatics* 2017;**33**:3131–3.
21. Xiaofan L, Meng J, Zhou Y *et al.* MOVICS: an R package for multi-omics integration and visualization in cancer subtyping. *Bioinformatics* 2020;**36**:5539–41.

22. Brière G, Darbo É, Thébault P *et al.* Consensus clustering applied to multi-omics disease subtyping. *BMC Bioinformatics* 2021;**22**:361. https://doi.org/10.1186/s12859-021-04279-1
23. Song W, Wang W, Dai D-Q. Subtype-WESLR: identifying cancer subtype with weighted ensemble sparse latent representation of multi-view data. *Brief Bioinform* 2022;**23**:1–12.
24. Tran D, Pham V-D, Nguyen H *et al.* DSCC: disease subtyping using spectral clustering and community detection from consensus networks. *Brief Bioinform* 2025;**26**:1–12.
25. Qiu Y, Dong Guo P, Zhao, and Quan Zou. scMNMF: a novel method for single-cell multi-omics clustering based on matrix factorization. *Brief Bioinform* 2024;**25**:1–9.
26. Yang H, Chen R, Li D *et al.* Subtype-GAN: a deep learning approach for integrative cancer subtyping of multi-omics data. *Bioinformatics* 2021;**37**:2231–7. https://doi.org/10.1093/bioinformatics/btab109
27. Nguyen H, Tran D, Tran B *et al.* SMRT: randomized data transformation for cancer subtyping and big data analysis. *Front Oncol* 2021;**11**:725133. https://doi.org/10.3389/fonc.2021.725133
28. Tran D, Nguyen H, Le U *et al.* A novel method for cancer subtyping and risk prediction using consensus factor analysis. *Front Oncol* 2020;**10**:1052. https://doi.org/10.3389/fonc.2020.01052
29. Chalise P, Fridley BL. Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm. *PLoS One* 2017;**12**:e0176278. https://doi.org/10.1371/journal.pone.0176278
30. Dingming W, Wang D, Zhang MQ *et al.* Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: Application to cancer molecular classification. *BMC Genomics* 2015;**16**:1022.
31. Mo Q, Shen R, Guo C *et al.* A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics* 2018;**19**:71–86. https://doi.org/10.1093/biostatistics/kxx017
32. Mo Q, Wang S, Seshan VE *et al.* Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc Natl Acad Sci* 2013;**110**:4245–50.
33. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* 2009;**25**:2906–12. https://doi.org/10.1093/bioinformatics/btp543
34. Shen R, Mo Q, Schultz N *et al.* Integrative subtype discovery in glioblastoma using iCluster. *PLoS One* 2012;**7**:e35236. https://doi.org/10.1371/journal.pone.0035236
35. Yang B, Yang Y, Wang M *et al.* MRGCN: Cancer subtyping with multi-reconstruction graph convolutional network using full and partial multi-omics dataset. *Bioinformatics* 2023;**39**:1–8.
36. Zhang C, Chen Y, Zeng T *et al.* Deep latent space fusion for adaptive representation of heterogeneous multi-omics data. *Brief Bioinform* 2022;**23**:1–15. https://doi.org/10.1093/bib/bbab600
37. Yang B, Yang Y, Xueping S. Deep structure integrative representation of multi-omics data for cancer subtyping. *Bioinformatics* 2022;**38**:3337–42. https://doi.org/10.1093/bioinformatics/btac345
38. Miao Y, Huang X, Wang S. PartIES: a disease subtyping framework with partition-level integration using diffusion-enhanced similarities from multi-omics data. *Brief Bioinform* 2025;**26**:1–11.
39. Wang B, Mezlini AM, Demir F *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods* 2014;**11**:333–7. https://doi.org/10.1038/nmeth.2810
40. Rappoport N, Shamir R. NEMO: cancer subtyping by integration of partial multi-omic data. *Bioinformatics* 2019;**35**:3348–56. https://doi.org/10.1093/bioinformatics/btz058
41. Nguyen T, Tagett R, Diaz D *et al.* A novel approach for data integration and disease subtyping. *Genome Res* 2017;**27**:2025–39. https://doi.org/10.1101/gr.215129.116
42. Nguyen H, Shrestha S, Draghici S *et al.* PINSPlus: a tool for tumor subtype discovery in integrated genomic data. *Bioinformatics* 2019;**35**:2843–6. https://doi.org/10.1093/bioinformatics/bty1049
43. Ramazzotti D, Lal A, Wang B *et al.* Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival. *Nat Commun* 2018;**9**:4453. https://doi.org/10.1038/s41467-018-06921-8
44. Ma T, Zhang A. Integrate multi-omic data using affinity network fusion (ANF) for cancer patient clustering. In: Hu X (ed.), *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. New York: IEEE, 2017, 398–403.
45. Wei Y, Li L, Zhao X *et al.* Cancer subtyping with heterogeneous multi-omics data via hierarchical multi-kernel learning. *Brief Bioinform* 2023;**24**:1–13. https://doi.org/10.1093/bib/bbac488
46. Yang Y, Tian S, Yushan Qiu P *et al.* MDICC: Novel method for multi-omics data integration and cancer subtype identification. *Brief Bioinform* 2022;**23**:1–13.
47. Cao H, Wang T, Zhaoyang X *et al.* Multi-omics data integration for enhanced cancer subtyping via interactive multi-kernel learning. *Brief Bioinform* 2025;**26**:1–13. https://doi.org/10.1093/bib/bbaf687
48. Liu W, Wen Y, Yu Z *et al.* Sphereface: Deep hypersphere embedding for face recognition. In: Liu Y (ed.), *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. New York: IEEE, 2017, 212–20.
49. Wang H, Wang Y, Zheng Z *et al.* Cosface: Large margin cosine loss for deep face recognition. In: Forsyth D (ed.), *Proceedings of the IEEE Conference on Computer vision and Pattern Recognition*. New York: IEEE, 2018, 5265–74.
50. Deng J, Guo J, Xue N *et al.* Arcface: Additive angular margin loss for deep face recognition. In: Gupta A (ed.), *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New York: IEEE, 2019, 4690–9.
51. Lan W, Liao H, Chen Q *et al.* DeepKEGG: A multi-omics data integration framework with biological insights for cancer recurrence prediction and biomarker discovery. *Brief Bioinform* 2024;**25**:1–16.
52. Hao J, Masum M, Oh JH *et al.* Gene-and pathway-based deep neural network for multi-omics data integration to predict cancer survival outcomes. In: Cai Z (ed.), *Bioinformatics Research and Applications: 15th International Symposium, ISBRA 2019, Barcelona, Spain, June 3–6, 2019, Proceedings 15*. Switzerland: Springer, 2019, 113–24.
53. Vaswani A, Shazeer N, Parmar N *et al.* Attention is all you need. In: Guyon I, vonLuxburg U, Bengio S *et al.* (eds.), *Advances in Neural Information Processing Systems*, Vol. **30**. New York: Curran Associates, Inc., 2017, 5998–6008.
54. Gheini M, Ren X, May J. Cross-attention is all you need: adapting pretrained transformers for machine translation. In: Moens M, Huang X, Specia L *et al.* (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Pennsylvania: Association for Computational Linguistics, 2021, 1754–65.
55. Loshchilov I, Hutter F. Decoupled weight decay regularization. In: *7th International Conference on Learning Representations*, 2019.
56. Reichle RH, Koster RD. Bias reduction in short records of satellite soil moisture. *Geophys Res Lett* 2004;**31**:1–4. https://doi.org/10.1029/2004GL020938
57. Xia D-y, Fei W, Zhang X-q *et al.* Local and global approaches of affinity propagation clustering for large scale data. *J Zhejiang Univ-Sci A* 2008;**9**:1373–81. https://doi.org/10.1631/jzus.A0720058

58. Zhu X, Loy CC, Gong S. Constructing robust affinity graphs for spectral clustering. In: Kacprzyk J (ed.), *Proceedings of the IEEE conference on computer vision and pattern recognition*. New York: IEEE, 2014, 1450–7.
59. Newman MEJ. Modularity and community structure in networks. *Proc Natl Acad Sci* 2006;**103**:8577–82. https://doi.org/10.1073/pnas.0601602103
60. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* 1987;**20**:53–65. https://doi.org/10.1016/0377-0427(87)90125-7
61. Benedetti E, Liu EM, Tang C *et al.* A multimodal atlas of tumour metabolism reveals the architecture of gene–metabolite covariation. *Nat Metab* 2023;**5**:1029–44.
62. Zhang G, He P, Tan H *et al.* Integration of metabolomics and transcriptomics revealed a fatty acid network exerting growth inhibitory effects in human pancreatic cancer. *Clin Cancer Res* 2013;**19**:4983–93. https://doi.org/10.1158/1078-0432.CCR-13-0209
63. Terunuma A, Putluri N, Mishra P *et al.* MYC-driven accumulation of 2-hydroxyglutarate is associated with breast cancer prognosis. *J Clin Invest* 2014;**124**:398–412. https://doi.org/10.1172/JCI71180
64. Gentric G, Kieffer Y, Mieulet V *et al.* PML-regulated mitochondrial metabolism enhances chemosensitivity in human ovarian cancers. *Cell Metab* 2019;**29**:156–173.e10. https://doi.org/10.1016/j.cmet.2018.09.002
65. Wang L-B, Karpova A, Gritsenko MA *et al.* Proteogenomic and metabolomic characterization of human glioblastoma. *Cancer Cell* 2021;**39**:509–528.e20. https://doi.org/10.1016/j.ccell.2021.01.006
66. Bewick V, Cheek L, Ball J. Statistics review 12: Survival analysis. *Crit Care* 2004;**8**:389–94. https://doi.org/10.1186/cc2955
67. Antolini L, Boracchi P, Biganzoli E. A time-dependent discrimination index for survival data. *Stat Med* 2005;**24**:3927–44. https://doi.org/10.1002/sim.2427
68. Schoenfeld D. Partial residuals for the proportional hazards regression model. *Biometrika* 1982;**69**:239–41. https://doi.org/10.1093/biomet/69.1.239
69. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol* 1995;**57**:289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x
70. Fisher RA. Statistical methods for research workers. In: Kotz S, Johnson NL (eds.), *Breakthroughs in Statistics: Methodology and Distribution*, Volume 18, pp. 66–70. New York, NY: Springer, 1970.
71. Hornung R, Wright MN. Block forests: random forests for blocks of clinical and omics covariate data. *BMC Bioinformatics* 2019;**20**:1–17.
72. Siegel RL, Kratzer TB, Giaquinto AN *et al.* Cancer statistics, 2025. *CA Cancer J Clin* 2025;**75**:10–45. https://doi.org/10.3322/caac.21871
73. Nguyen H, Tran D, Galazka JM *et al.* CPA: a web-based platform for consensus pathway analysis and interactive visualization. *Nucleic Acids Res* 2021;**49**:W114–24. https://doi.org/10.1093/nar/gkab421
74. Nguyen H, Nguyen H, Maghsoudi Z *et al.* RCPA: an open-source r package for data processing, differential analysis, consensus pathway analysis, and visualization. *Curr Protoc* 2024;**4**:e1036. https://doi.org/10.1002/cpz1.1036
75. Nguyen H, Pham V-D, Nguyen H *et al.* CCPA: cloud-based, self-learning modules for consensus pathway analysis using GO, KEGG and Reactome. *Brief Bioinform* 2024;**25**:bbae222.
76. Tavazoie S, Hughes JD, Campbell MJ *et al.* Systematic determination of genetic network architecture. *Nat Genet* 1999;**22**:281–5. https://doi.org/10.1038/10343
77. Efron B, Tibshirani R. On testing the significance of sets of genes. *Ann Appl Stat* 2007;**1**:107–29. https://doi.org/10.1214/07-AOAS101
78. Korotkevich G, Sukhov V, Budin N *et al.* Fast gene set enrichment analysis. *bioRxiv* 2016;060012.
79. Tarca AL, Drăghici S, Bhatti G *et al.* Down-weighting overlapping genes improves gene set analysis. *BMC Bioinformatics* 2012;**13**:136. https://doi.org/10.1186/1471-2105-13-136
80. Campbell JE, Drucker DJ. Pharmacology, physiology, and mechanisms of incretin hormone action. *Cell Metab* 2013;**17**:819–37. https://doi.org/10.1016/j.cmet.2013.04.008
81. Porterfield M, Zhao P, Han H *et al.* Discrimination between adenocarcinoma and normal pancreatic ductal fluid by proteomic and Glycomic analysis. *J Proteome Res* 2014;**13**:395–407. https://doi.org/10.1021/pr400422g
82. Liu Q, Guo L, Zhang S *et al.* PRSS1 mutation: a possible pathomechanism of pancreatic carcinogenesis and pancreatic cancer. *Mol Med* 2019;**25**:1–11.
83. Tamura K, Yu J, Hata T *et al.* Mutations in the pancreatic secretory enzymes CPA1 and CPB1 are associated with pancreatic cancer. *Proc Natl Acad Sci* 2018;**115**:4767–72.
84. Gilliland TM, Villafane-Ferriol N, Shah KP *et al.* Nutritional and metabolic derangements in pancreatic cancer and pancreatic resection. *Nutrients* 2017;**9**:1–15. https://doi.org/10.3390/nu9030243
85. Chen J, Zhiyuan X, Can H *et al.* Targeting CLDN18.2 in cancers of the gastrointestinal tract: new drugs and new indications. *FrontOncol* 2023;**13**:1–17.
86. Iwaya M, Hayashi H, Nakajima T *et al.* Colitis-associated colorectal adenocarcinomas frequently express claudin 18 isoform 2: implications for claudin 18.2 monoclonal antibody therapy. *Histopathology* 2021;**79**:227–37. https://doi.org/10.1111/his.14358
87. Pothuraju R, Rachagani S, Krishn SR *et al.* Molecular implications of MUC5AC-CD44 axis in colorectal cancer progression and chemoresistance. *Mol Cancer* 2020;**19**:1–14. https://doi.org/10.1186/s12943-020-01156-y
88. Chou C-L, Chen T-J, Tian Y-F *et al.* CTSE overexpression is an adverse prognostic factor for survival among rectal cancer patients receiving CCRT. *Life* 2021;**11**:1–14. https://doi.org/10.3390/life11070646
89. Bahar ME, Kim HJ, Kim DR. Targeting the RAS/RAF/MAPK pathway for cancer therapy: from mechanism to clinical studies. *Signal transduction and targeted. Therapy* 2023;**8**:1–38.
90. Luchini C, Paolino G, Mattiolo P *et al.* KRAS wild-type pancreatic ductal adenocarcinoma: molecular pathology and therapeutic opportunities. *J Exp Clin Cancer Res* 2020;**39**:1–10.
91. Zhu G, Pei L, Xia H *et al.* Role of oncogenic KRAS in the prognosis, diagnosis and treatment of colorectal cancer. *Mol Cancer* 2021;**20**:1–17. https://doi.org/10.1186/s12943-021-01441-4
92. Parikh K, Banna G, Liu SV *et al.* Drugging KRAS: current perspectives and state-of-art review. *J Hematol Oncol* 2022;**15**:1–22.
93. Shi Y, Zheng H, Wang T *et al.* Targeting KRAS: from metabolic regulation to cancer treatment. *Mol Cancer* 2025;**24**:1–21.
94. Kooistra MRH, Dubé N, Bos JL. Rap1: a key regulator in cell-cell junction formation. *J Cell Sci* 2007;**120**:17–22. https://doi.org/10.1242/jcs.03306
95. Zhang Y-L, Wang R-C, Cheng K *et al.* Roles of Rap1 signaling in tumor cell migration and invasion. *Cancer Biol Med* 2017;**14**:90–9. https://doi.org/10.20892/j.issn.2095-3941.2016.0086
96. Muller WA. Mechanisms of leukocyte transendothelial migration. *Annu Rev Pathol Mech Dis* 2011;**6**:323–44. https://doi.org/10.1146/annurev-pathol-011110-130224
97. Fang J, Yue L, Zheng J *et al.* Exploring the crosstalk between endothelial cells, immune cells, and immune checkpoints in the

tumor microenvironment: New insights and therapeutic implications. *Cell Death Dis* 2023;**14**:1–15.
98. Zhang D, Bi J, Liang Q *et al.* VCAM1 promotes tumor cell invasion and metastasis by inducing EMT and Transendothelial migration in colorectal cancer. *Front Oncol* 2020;**10**:1–12. https://doi.org/10.3389/fonc.2020.01066
99. Bachir AI, Horwitz AR, James Nelson W *et al.* Actin-based adhesion modules mediate cell interactions with the extracellular matrix and Neighboring cells. *Cold Spring Harb Perspect Biol* 2017;**9**:1–19. https://doi.org/10.1101/cshperspect.a023234
100. Winkler J, Abisoye-Ogunniyan A, Metcalf KJ *et al.* Concepts of extracellular matrix remodelling in tumour progression and metastasis. *Nat Commun* 2020;**11**:1–19. https://doi.org/10.1038/s41467-020-18794-x
101. Morgos D-T, Stefani C, Miricescu D *et al.* Targeting PI3K/AKT/mTOR and MAPK signaling pathways in gastric cancer. *Int J Mol Sci* 2024;**25**:1–26. https://doi.org/10.3390/ijms25031848
102. Rascio F, Spadaccino F, Rocchetti MT *et al.* The pathogenic role of PI3K/AKT pathway in cancer onset and drug resistance. An updated review. *Cancers* 2021;**13**:1–15. https://doi.org/10.3390/cancers13163949
103. Wang Q, Shi Y-l, Zhou K *et al.* PIK3CA mutations confer resistance to first-line chemotherapy in colorectal cancer. *Cell Death Dis* 2018;**9**:1–11. https://doi.org/10.1038/s41419-018-0776-6
104. Tang YM, Fan Y. Combined KRAS and TP53 mutation in patients with colorectal cancer enhance chemoresistance to promote postoperative recurrence and metastasis. *BMC Cancer* 2024;**24**:1–8. https://doi.org/10.1186/s12885-024-12776-8
105. Palanissami G, Paul SFD. AGEs and RAGE: Metabolic and molecular signatures of the glycation-inflammation axis in malignant or metastatic cancers. *Explor Target Anti-tumor Ther* 2023; **4**:812–49.
106. Zhao J, Lv T, Quan J *et al.* Identification of target genes in cardiomyopathy with fibrosis and cardiac remodeling. *J Biomed Sci* 2018;**25**:1–10. https://doi.org/10.1186/s12929-018-0459-8
107. Henke E, Nandigama R, Ergün S. Extracellular matrix in the tumor microenvironment and its impact on cancer therapy. *Front Mol Biosci* 2020;**6**:1–24. https://doi.org/10.3389/fmolb.2019.00160
108. Desgrosellier JS, Cheresh DA. Integrins in cancer: Biological implications and therapeutic opportunities. *Nat Rev Cancer* 2010;**10**:9–22. https://doi.org/10.1038/nrc2748
109. Bergonzini C, Kroese K, Zweemer AJM *et al.* Targeting integrins for cancer therapy - disappointments and opportunities. *Front Cell Dev Biol* 2022;**10**:1–13. https://doi.org/10.3389/fcell.2022.863850
110. Ning Y, Zheng M, Zhang Y *et al.* RhoA-ROCK2 signaling possesses complex pathophysiological functions in cancer progression and shows promising therapeutic potential. *Cancer Cell Int* 2024;**24**:1–15. https://doi.org/10.1186/s12935-024-03519-7
111. Kelly CM, Sainz LG, Chi P. The management of metastatic GIST: current standard and investigational therapeutics. *J Hematol Oncol* 2021;**14**:1–12.
112. Casak SJ, Fashoyin-Aje I, Lemery SJ *et al.* FDA approval summary: Ramucirumab for gastric cancer. *Clin Cancer Res* 2015;**21**:3372–6. https://doi.org/10.1158/1078-0432.CCR-15-0600
113. Clarke JM, Hurwitz HI. Targeted inhibition of VEGF receptor 2: an update on ramucirumab. *Expert Opin Biol Ther* 2013;**13**:1187–96. https://doi.org/10.1517/14712598.2013.810717
114. Shah MA, Shitara K, Ajani JA *et al.* Zolbetuximab plus CAPOX in CLDN18.2-positive gastric or gastroesophageal junction adenocarcinoma: the randomized, phase 3 GLOW trial. *Nat Med* 2023;**29**:2133–41. https://doi.org/10.1038/s41591-023-02465-7
115. Chen Y, Fan X, Shi C *et al.* A joint analysis of single cell transcriptomics and proteomics using transformer. *NPJ Syst Biol Appl* 2025;**11**:1. https://doi.org/10.1038/s41540-024-00484-9
116. Jing X, Huang D-S, Zhang X. scmFormer integrates large-scale single-cell proteomics and transcriptomics data by multi-task transformer. *Adv Sci* 2024;**11**:2307835. https://doi.org/10.1002/advs.202307835
117. Boyko M, Beliaeva A, Kornilov D *et al.* impuTMAE: Multimodal Transformer with Masked Pre-training for Missing Modalities Imputation in Cancer Survival Prediction. *arXiv preprint arXiv:2508.09195*, 2025.

Briefings in Bioinformatics, 2026, **27**, bbag290
https://doi.org/10.1093/bib/bbag290
Problem Solving Protocol