Pathformer: a biological pathway informed Transformer integrating multi-omics data for disease diagnosis and prognosis

Xiaofan Liu[1,2], Yuhuan Tao[1,2], Zilin Cai[1], Pengfei Bao[1,2], Hongli Ma[1,2], Kexing Li[1], Mengtao Li[3,] *, Yunping Zhu[4,] *, Zhi John Lu[1,2,] *

[1]MOE Key Laboratory of Bioinformatics, Center for Synthetic and Systems Biology, School of Life Sciences, Tsinghua University, Beijing 100084, China.

[2]Institute for Precision Medicine, Tsinghua University, Beijing 100084, China.

[3]Department of Rheumatology and Clinical Immunology, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences, Peking Union Medical College, National Clinical Research Center for Dermatologic and Immunologic Diseases (NCRC-DID), MST State Key Laboratory of Complex Severe and Rare Diseases, MOE Key Laboratory of Rheumatology and Clinical Immunology, Beijing 100730, China

[4]State Key Laboratory of Proteomics, Beijing Proteome Research Center, National Center for Protein Sciences (Beijing), Beijing Institute of Lifeomics, 38 Life Science Park, Changping District, Beijing 102206, China.

*To whom correspondence should be addressed: Zhi John Lu, Tel: +86 10 62789217, E-mail: zhilu@tsinghua.edu.cn; Mengtao Li, E-mail: mengtao.li@cstar.org.cn; Yunping Zhu, Tel.: +86 10 61777058, E-mail: zhuyunping@ncpsb.org.cn.

Abstract

Multi-omics data provide a comprehensive view of gene regulation at multiple levels, which is helpful in achieving accurate diagnosis of complex diseases like cancer. To integrate various multi-omics data of tissue and liquid biopsies for disease diagnosis and prognosis, we developed a biological pathway informed Transformer, Pathformer. It embeds multi-omics input with a compacted multi-modal vector and a pathway-based sparse neural network. Pathformer also leverages criss-cross attention mechanism to capture the crosstalk between different pathways and modalities. We first benchmarked Pathformer with 18 comparable methods on multiple cancer datasets, where Pathformer outperformed all the other methods, with an average improvement of 6.3%-14.7% in F1 score for cancer survival prediction and 5.1%-12% for cancer stage prediction. Subsequently, for cancer prognosis prediction based on tissue multi-omics data, we used a case study to demonstrate the biological interpretability of Pathformer by identifying key pathways and their biological crosstalk. Then, for cancer early diagnosis based on liquid biopsy data, we used plasma and platelet datasets to demonstrate Pathformer's potential of clinical applications in cancer screen. Moreover, we revealed deregulation of interesting pathways (e.g., scavenger receptor pathway) and their crosstalk in cancer patients' blood, providing new candidate targets for cancer microenvironment study.

Keywords: Multi-omics integration; Liquid biopsy; Cancer early diagnosis; Cancer prognosis; Pathway crosstalk

Introduction

Comparing to single type of data, multi-omics data provide a more comprehensive view of gene regulation[1]. Therefore, integrating multi-omics data from tissue and liquid biopsies would be helpful in addressing challenges in disease diagnosis and prognosis, such as deregulated network between different types of molecules and data noise caused by patients' heterogeneity[2]. To integrate multi-omics data of cancer, several supervised methods have been developed, such as mixOmics[3], liNN[4], eiNN[5], liCNN[6], eiCNN[7], MOGONet[8] and MOGAT[9]. Later, the performance and interpretability of multi-omics data integration were further improved using deep learning models informed by biological pathways. For instance, a pathway-associated sparse deep neural network (PASNet) was utilized to predict the prognosis of glioblastoma multiforme (GBM) patients[10]. Recently, P-NET, a sparse neural network integrating multiple molecular features based on a multilevel view of biological pathways, was introduced to predict subtype and survival of prostate cancer patients[11]. In addition, PathCNN based on a convolutional neural network (CNN) was developed to predict the prognosis of GBM patients using principal component analysis (PCA) to define image-like multi-omics pathways[12].

These pathway-informed deep learning methods did not consider the crosstalk between omics and between pathways, although the crosstalk holds biological significance as well as pathway itself[13-15]. Meanwhile, the criss-cross attention mechanism of the Transformer would be very useful to capture the crosstalk information[16]. However, incorporating multi-omics data and their crosstalk information in a Transformer is very challenging: when processing multi-omics data, the multi-modal features are usually multiplied by the number of genes (tens of thousands), producing an extremely long input that is not acceptable by a common Transformer model (usually less than 512). Meanwhile, certain embedding methods for biological data, such as discretization and linear transformation, were introduced in the previous Transformer models[17-19], while biological information was largely lost during these kinds of embedding.

In order to integrate multi-omics data by embedding biological pathway crosstalk without information loss, we introduce a Transformer model, Pathformer, with three key steps to address the above problems. First, it ingeniously concatenates different modalities into a novel compacted multi-modal vector for each gene, which not only saves valuable information but also shortens the input. Second, Pathformer utilizes a sparse neural network based on prior pathway knowledge to transform gene embeddings into pathway embeddings. Third, Pathformer naturally incorporates pathway crosstalk network into a Transformer model with bias to enhance the exchange of information between different pathways and between different modalities (e.g., omics) as well.

Here, we first benchmarked Pathformer and 18 other integration methods in various classification tasks, using

multiple cancer tissue datasets from TCGA. Then, we used Pathformer to integrate various multi-omics data from tissue and liquid biopsies. Through case studies on survival prediction of breast cancer and noninvasive diagnosis of pan-cancer, we revealed interesting pathways, genes, and regulatory mechanisms related to cancer in human tissue and plasma, demonstrating the prediction accuracy and biological interpretability of Pathformer in various clinical applications.

Materials and methods

Overview of Pathformer

Pathformer is mainly designed to integrate various multi-omics data from tissue and liquid biopsies, which can be used for different classification tasks in disease diagnosis and prognosis, such as cancer early detection, cancer staging and survival prediction (**Fig. 1a**). It has 6 modules: 1) biological pathway and crosstalk network calculation module, 2) multi-omics data input module, 3) biological multi-modal embedding module (key module), 4) Transformer module with pathway crosstalk network bias, 5) classification module and 6) biological interpretability module.

Pathformer combines prior biological pathway information (module 1, **Fig. 1a**) with multi-modal data (module 2, **Fig. 1b**) for disease diagnosis and prognosis. It introduces a new embedding method to incorporate biological multi-modal data at both gene level and pathway level: it initiates the process by uniformly transforming different modalities to the gene level through a series of statistical indicators, then concatenates these modalities into compacted multi-modal vectors to define gene embedding, and employs a sparse neural network based on the gene-to-pathway mapping to transform gene embedding into pathway embedding (module 3, **Fig. 1c**). Pathformer then enhances the fusion of information between various modalities and pathways by combining pathway crosstalk networks with Transformer encoder (module 4, **Fig. 1a, Supplementary Fig. 1**). Finally, a fully connected layer serves as the classifier for different downstream classification tasks (module 5). In addition, Pathformer uses a biological interpretable module with attention weights and SHapley Additive exPlanations[20] values to identify important genes, pathways, modalities, and their crosstalk or regulation (module 6).

These 6 modules are described in detail below.

Module 1: curation of biological pathways and calculation of initial crosstalk network

We curated 2,289 biological pathways from four public databases including Kyoto Encyclopedia of Genes and Genomes database (KEGG)[21], Pathway Interaction database (PID)[22], Reactome database (Reactome)[23], and

BioCarta Pathways database (BioCarta)[24]. Then, we filtered these pathways by three criteria: the gene number, the overlap ratio with other pathways (the proportion of genes in the pathway that are also present in other pathways), and the number of pathway subsets (the number of pathways included in the pathway). Following the principle of moderate size and minimal overlap with other pathway information, we selected 1,497 pathways with gene number between 15 and 100, or gene number greater than 15 and overlap ratio less than 1, or gene number greater than 15 and the number of pathway subsets less than 5. Next, we used *BinoX*[25] to calculate the crosstalk relationship of pathways and build a pathway crosstalk network with adjacency matrix $\boldsymbol{P} \in \mathbb{R}^{N_p \times N_p}$, N_p=1,497 (more details in **Supplementary Note 1**).

Modules 2 and 3: multi-omics data input and multi-modal embedding

Biological multi-modal data preprocessing and embedding method are two key modules of Pathformer (**Fig. 1b, c**). In module 2 (**Fig. 1b**), to capture more comprehensive regulatory information, we expand biological multi-omics data into multi-modal data, including not only data from different omics sources but also variant features of the same omics, such as RNA splicing, RNA editing, RNA alternative promoter, and so on. To obtain multi-modal data, we use standardized bioinformatics pipeline to calculate different omics or variant features of the same omics from raw sequence reads (more details in **Supplementary Note 5**). These multi-modal data have different dimensions, including nucleotide level, fragment level, and gene level. For example, Pathformer's input for cancer tissue datasets from Cancer Genome Atlas (TCGA)[26] includes gene-level RNA expression, fragment-level DNA methylation, and both fragment-level and gene-level DNA CNV. Modalities and their dimensions level for different datasets are described in **Supplementary Table 1**.

In module 3 (**Fig. 1c**), we proposed a new biological multi-modal embedding method of Pathformer, which consists of gene embedding $\boldsymbol{E}_G$ and pathway embedding $\boldsymbol{E}_P$. We represented biological multi-modal input matrix of a sample as $\boldsymbol{M}$, described as follows:

$$\boldsymbol{M} = [\boldsymbol{M}_1, \boldsymbol{M}_2, \boldsymbol{M}_3, \cdots, \boldsymbol{M}_m], \boldsymbol{M}_i \in \mathbb{R}^{N_{M_i}}, i = 1, 2, 3, \cdots, m$$

, where m is the number of modalities, and N_{M_i} is the length of input for modality i, likes the number of genes for RNA expression, the number of editing sites for RNA editing, and the number of CpG islands for DNA methylation.

Next, we first use a series of statistical indicators to convert different modalities into gene level modal features, and then concatenate these modal features into a compressed multi-modal vector as gene embedding $\boldsymbol{E}_G$, which are calculated as follows:

$$\boldsymbol{E}_G = \boldsymbol{F}_E(\boldsymbol{M}) = [\boldsymbol{V}_1, \boldsymbol{V}_2, \cdots, \boldsymbol{V}_i, \cdots, \boldsymbol{V}_m] = [\boldsymbol{G}_1, \boldsymbol{G}_2, \cdots, \boldsymbol{G}_l, \cdots, \boldsymbol{G}_{N_g}] \in \mathbb{R}^{N_g \times D_g}$$

$$\boldsymbol{V}_i = \boldsymbol{F}_{E_i}(\boldsymbol{M}_i) = [f_*(\boldsymbol{M}_i), \cdots, f_*(\boldsymbol{M}_i)] = [\boldsymbol{V}_i^1, \cdots, \boldsymbol{V}_i^{e_i}] \in \mathbb{R}^{N_g \times e_i}, f_* \in [f_1, \cdots, f_8]$$

, where $\boldsymbol{G}_l = [\boldsymbol{V}_1^{(l)}, \cdots, \boldsymbol{V}_i^{(l)}, \cdots, \boldsymbol{V}_m^{(l)}] \in \mathbb{R}^{D_g}$ is gene embedding of the *l*th gene and is a compacted multi-modal vector; $\boldsymbol{V}_i$ is the modal features matrix of modality *i*; $\boldsymbol{V}_i^j$ is the *j*th dimension of modal features matrix for modality *i*; N_g is the number of genes, $D_g = e_1 + e_2 + \cdots + e_m$ is the dimension of gene embedding; e_i is the dimension of modal features matrix for modality *i*; $\boldsymbol{F}_E$ is the conversion function, which uses statistical indicators to uniformly convert different modalities into gene level; $\boldsymbol{F}_{E_i}$ is the conversion function of modality *i*, and each modality's function is constructed from distinct statistical indicator functions f_* (more details in **Supplementary Table 1**). These statistical indicator functions include gene level score (f_1), count (f_2), minimum (f_3), maximum (f_4), mean (f_5), entropy (f_6), weighted mean in whole gene (f_7) and weighted mean in window (f_8), formulas of which are in **Supplementary Note 2**.

Subsequently, we used the known gene-pathway mapping relationship to develop a sparse neural network based on prior pathway knowledge (PSNN) to transform gene embedding $\boldsymbol{E}_G$ into pathway embedding $\boldsymbol{E}_P$, as described below:

$$\boldsymbol{E_P} = \boldsymbol{W}_{\boldsymbol{sparse}}^{\boldsymbol{T}} \boldsymbol{E_G} + \boldsymbol{B}, \boldsymbol{E_P} \in \mathbb{R}^{N_p \times D_p}$$

, where N_p is the number of pathways, $D_p = D_g$ is the dimension of pathway embedding, $\boldsymbol{W}_{sparse} \in \mathbb{R}^{N_g \times N_p}$ is a learnable sparse weight matrix, and $\boldsymbol{B}$ is a bias term. $\boldsymbol{W}_{sparse}$ is constructed based on the known relationship between pathways and genes. When the given gene and the pathway are irrelevant, the corresponding element of $\boldsymbol{W}_{sparse}$ will always be 0. Otherwise, it needs to be learned through training. Therefore, pathway embedding is a dynamic embedding method. The PSNN can not only restore the mapping relationship between genes and pathways, but also capture the different roles of different genes in pathways, and can preserve the complementarity of different modalities. Additionally, this biological multi-modal embedding step does not require additional gene selection, thereby avoiding bias and overfitting problems resulting from artificial feature selection.

Module 4: Transformer module with pathway crosstalk network bias

We developed the Transformer module based on criss-cross attention (CC-attention) with bias for data fusion of pathways, modalities, and their crosstalk (**Supplementary Fig. 1**). This module has 3 blocks, each containing multi-head column-wise self-attention (col-attention), multi-head row-wise self-attention (row-attention), layer normalization, GELU activation, residual connection, and network update. Particularly, col-attention is used to enhance the exchange of information between pathways, with the pathway crosstalk network matrix serving as the

bias for col-attention to guide the flow of information. Row-attention is employed to facilitate information exchange between different modalities, and the updated pathway embedding matrix is used to update the pathway crosstalk network matrix by calculating the correlation between pathways.

Multi-head column-wise self-attention contains 8 heads, each head is a mapping of $\boldsymbol{Q_1}$,$\boldsymbol{K_1}$,$\boldsymbol{V_1}$,$\boldsymbol{P}$, which are query vector, key vector, and value vector of pathway embedding $\boldsymbol{E}_P$ and pathway crosstalk network matrix $\boldsymbol{P}$, respectively. First, we represented the *h*th column-wise self-attention by $\boldsymbol{A}_{col}^{(h)}$, calculated as follows:

$$\boldsymbol{A}_1^{(h)} = (\boldsymbol{Q}_1\boldsymbol{K}_1^T)/\sqrt{d}, \boldsymbol{Q}_1 = \boldsymbol{E}_P\boldsymbol{W}_{Q_1}^{(h)}, \boldsymbol{K}_1 = \boldsymbol{E}_P\boldsymbol{W}_{K_1}^{(h)}, \boldsymbol{V}_1 = \boldsymbol{E}_P\boldsymbol{W}_{V_1}^{(h)}$$

$$\boldsymbol{A}_{col}^{(h)} = \mathrm{dropout}_{0.2}(\mathrm{softmax}(\boldsymbol{A}_1^{(h)} + \boldsymbol{P})) \cdot \boldsymbol{V}_1^{(h)}$$

, where $h = 1,2,\cdots,H$ is the *h*th head; *H* is the number of heads; $\boldsymbol{W}_{Q_1}^{(h)} \in \mathbb{R}^{D_p\times d}$, $\boldsymbol{W}_{K_1}^{(h)} \in \mathbb{R}^{D_p\times d}$, $\boldsymbol{W}_{V_1}^{(h)} \in \mathbb{R}^{D_p\times d}$ are the weight matrices as parameters; *d* is the attention dimension; $\mathrm{dropout}_{0.2}$ is a dropout neural network layer with a probability of 0.2; and softmax is the normalized exponential function.

Next, we merged multi-head column-wise self-attention and performed a series of operations as follows:

$$\boldsymbol{g}_1^{(h)} = sigmoid(\boldsymbol{E}_P\boldsymbol{W}_{g_1}^{(h)})$$

$$\boldsymbol{U}_1' = \boldsymbol{U}_1 + \boldsymbol{E}_P, \boldsymbol{U}_1 = \sum_{h=1}^{H}(\boldsymbol{g}_1^{(h)} \circ \boldsymbol{A}_{col}^{(h)}) \cdot \boldsymbol{W}_{U_1}^{(h)}$$

$$\boldsymbol{O}_1 = \mathrm{dropout}_{0.2}(\mathrm{GELU}(\mathrm{LayerNorm}(\boldsymbol{U}_1') \cdot \boldsymbol{W}_{O_{11}})) \cdot \boldsymbol{W}_{O_{12}} + \boldsymbol{U}_1'$$

, where $h = 1,2,\cdots,H$ is the *h*th head; *H* is the number of heads; $\circ$ is the matrix dot product; $\boldsymbol{W}_{g_1}^{(h)} \in \mathbb{R}^{D_p\times d}$, $\boldsymbol{W}_{U_1}^{(h)} \in \mathbb{R}^{d\times D_p}$, $\boldsymbol{W}_{O_{11}} \in \mathbb{R}^{D_p\times o}$, $\boldsymbol{W}_{O_{12}} \in \mathbb{R}^{o\times D_p}$ are the weight matrices as parameters; *o* is a constant; LayerNorm is the layer normalization function; GELU is the distortion of RELU activation function; and $\mathrm{dropout}_{0.2}$is a dropout neural network layer with a probability of 0.2.

Multi-head row-wise self-attention enables information exchange between different modalities. It is a regular dot-product attention. It also contains 8 heads, and the *h*th row-wise self-attention, i.e., $\boldsymbol{A}_{row}^{(h)}$, is calculated as follows:

$$\boldsymbol{A}_2^{(h)} = (\boldsymbol{Q}_2\boldsymbol{K}_2^T)/\sqrt{d}, \boldsymbol{Q}_2 = \boldsymbol{E}_P^T\boldsymbol{W}_{Q_2}^{(h)}, \boldsymbol{K}_2 = \boldsymbol{E}_P^T\boldsymbol{W}_{K_2}^{(h)}, \boldsymbol{V}_2 = \boldsymbol{E}_P^T\boldsymbol{W}_{V_2}^{(h)}$$

$$\boldsymbol{A}_{row}^{(h)} = \mathrm{dropout}_{0.2}(\mathrm{softmax}(\boldsymbol{A}_2^{(h)})) \cdot \boldsymbol{V}_2^{(h)}$$

, where $h = 1,2,\cdots$, h is the *h*th head; *H* is the number of heads; $\boldsymbol{W}_{Q_2}^{(h)} \in \mathbb{R}^{N_p\times d}$, $\boldsymbol{W}_{K_2}^{(h)} \in \mathbb{R}^{N_p\times d}$, $\boldsymbol{W}_{V_2}^{(h)} \in \mathbb{R}^{N_p\times d}$ are the weight matrices as parameters; *d* is the attention dimension; $\mathrm{dropout}_{0.2}$ is a dropout neural network layer with a probability of 0.2; and softmax is the normalized exponential function.

Subsequently, we merged multi-head row-wise self-attention and performed a series of operations. The formulas are as follows:

$$\boldsymbol{g}_2^{(h)} = sigmoid(\boldsymbol{E}_P^T \boldsymbol{W}_{g_2}^{(h)})$$

$$\boldsymbol{U}_2' = \boldsymbol{\beta} * \boldsymbol{U}_2 + \boldsymbol{E}_P^T, \boldsymbol{U}_2 = \sum_{h=1}^{H} (\boldsymbol{g}_2^{(h)} \circ \boldsymbol{A}_{row}^{(h)}) \cdot \boldsymbol{W}_{U_2}^{(h)}$$

$$\boldsymbol{O}_2 = dropout_{0.2}(\mathrm{GELU}(\mathrm{LayerNorm}(\boldsymbol{U}_2') \cdot \boldsymbol{W}_{O_{21}})) \cdot \boldsymbol{W}_{O_{22}} + \boldsymbol{U}_2'$$

, where $h = 1,2,\cdots,$ h is the *h*th head; *H* is the number of heads; $\circ$ is the matrix dot product; $\boldsymbol{W}_{g_2}^{(h)} \in \mathbb{R}^{N_p \times d}$, $\boldsymbol{W}_{U_2}^{(h)} \in \mathbb{R}^{d \times N_p}$, $\boldsymbol{W}_{O_{21}} \in \mathbb{R}^{N_p \times o}$, $\boldsymbol{W}_{O_{22}} \in \mathbb{R}^{o \times N_p}$ are the weight matrices as parameters; o is a constant; $\boldsymbol{\beta}$ is a constant coefficient for row-attention; LayerNorm is the layer normalization function; GELU is the distortion of RELU activation function; and $dropout_{0.2}$ is a dropout neural network layer with a probability of 0.2. $\boldsymbol{O}_2$ is pathway embedding input of the next Transformer block. In other words, when $\boldsymbol{E}_P$ is $\boldsymbol{E}_P^{(0)}$, $\boldsymbol{O}_2$ is $\boldsymbol{E}_P^{(1)}$, superscripts with parenthesis represent data at different block.

Then, we used the updated pathway embedding $\boldsymbol{O}_2$ to update the pathway crosstalk network. We exploited the correlation between embedding vectors of two pathways to update the corresponding element of the pathway crosstalk network matrix. The formula is as follows:

$$\boldsymbol{P}' = (\boldsymbol{P} \cdot \boldsymbol{P}^T)/N_p$$

, where $\boldsymbol{P}'$ is the updated pathway crosstalk network matrix of next Transformer block. In other words, when $\boldsymbol{P}'$ is $\boldsymbol{P}^{(1)}$, $\boldsymbol{P}$ is $\boldsymbol{P}^{(0)}$, superscripts with parenthesis represent data at different block.

Module 5: classification module

Given the classification tasks in disease diagnosis and prognosis, we used the fully connected neural network as the classification module to transform pathway embedding encoded by the Transformer module into the probability for each label. Three fully connected neural networks each have 300, 200, and 100 neurons, with dropout probability $dropout_c$, which is hyperparameter. More details of the classification module are described in **Supplementary Note 3**.

Model 6: biological interpretability module

The biological interpretable module enables us to calculate the contribution of each modality, identify important pathways and their key genes, and uncover the most critical pathway crosstalk subnetworks.

To calculate the contribution of each omics and each modality, we first integrated all matrices of row-attention maps into one matrix by element-wise average. Then, we averaged this average row-attention matrix along with columns as the attention weights of modalities. More details are described in **Supplementary Note 4**.

To identify important pathways and their key genes, we used SHapley Additive exPlanations[20] (SHAP value) to calculate the contribution of each feature. It is an additive explanation model inspired by coalitional game theory, which regards all features as "contributors". SHAP value is the value assigned to each feature, which explains the relationship between modalities, pathways, genes and classification, implemented by “SHAP” package of *Python* v3.6.9. Then, pathways with the top 15 SHAP values in the classification task are considered as important pathways. For each pathway, genes with top 5 SHAP values are considered as the key genes of the pathway. The core modality on which one gene depends indicates that the SHAP value of that gene ranks higher on this modality than on the others. More details are described in **Supplementary Note 4**.

Particularly, the pathway crosstalk network matrix is used to guide the direction of information flow, and updated according to updated pathway embedding in each Transformer block. Therefore, the updated pathway crosstalk network contains not only the prior information in the initial network (module 1) but also the multi-modal data information derived from the Transformer module (module 4), which represents the specific regulatory mechanism in each classification task. We defined the sub-network score through SHAP value of each pathway in sub-network, so as to find foremost sub-network for prediction, that is, hub module of the updated pathway crosstalk network. The calculation of the sub-network score can be divided into four steps: average pathway crosstalk network matrix calculation, network pruning, sub-network boundary determination, and score calculation. More details of sub-network score calculations are described in **Supplementary Note 4**.

Experimental settings

Data collection and preprocessing

We assayed both tissue biopsy and liquid biopsy data in this study. First, for benchmark testing on cancer diagnosis and prognosis, we collected multiple datasets of different cancer types from TCGA (tissue data) to evaluate the classification performance, including 10 datasets for early- and late-stage classification, and 10 datasets for low- and high-risk survival classification (**Supplementary Fig. 2**). In addition, we also collected and processed two types of body fluid datasets: the plasma dataset (373 samples assayed by total cell-free RNA-seq[27,28]) and the platelet dataset (918 samples assayed by blood platelet RNA-seq[29,30]). Through our biological information pipeline, 3 and 7 biological modalities were derived from the TCGA (tissue biopsy) datasets and the liquid biopsy datasets, respectively. More details are described in **Supplementary Notes 5 and 7**.

Model training and test

We implemented Pathformer's network architecture using the “PyTorch” package in *Python* v3.6.9, and our codes

can be found in the GitHub repository (https://github.com/lulab/Pathformer). For model training and test, we used 5-fold cross-validation, and repeated it twice by shuffling. Before evaluating the performance on the test set, we optimized hyperparameters (e.g., learning rate, dropout probability of classification and constant coefficient for row-attention) and epoch numbers inside the training set only. More details of model training and test are described in **Supplementary Note 6**.

Evaluation criteria

When evaluating the classification performance, we used at least three evaluation indicators, area under the receiver operating characteristic curve (AUC), weighted-averaged F1 score (F1score_weighted), and macro-averaged F1 score (F1score_macro). Notably, we prioritized F1score_macro as the main evaluation criterion in this paper. This choice stems from the imbalance of sub-classes in our data, where F1score_macro stands out as a fairer and more robust indicator compared to other metrics such as AUC.

Results

Benchmark of Pathformer and 18 multi-omics data integration methods using TCGA data

We conducted a meticulous benchmark of Pathformer and other 18 multi-omics data integration methods for various classification tasks in cancer diagnosis and prognosis (**Fig. 2**). These methods can be categorized into 3 types. Type I includes early and late integration methods based on conventional classifiers, such as support vector machine (SVM), logistic regression (LR), random forest (RF), and extreme gradient boosting (XGBoost). Type II includes partial least squares-discriminant analysis (PLSDA) and sparse partial least squares-discriminant analysis (sPLSDA) of mixOmics[3]. Type III consists of deep learning-based integration methods, i.e., eiNN[5], liNN[4], eiCNN[7], liCNN[6], MOGONet[8], MOGAT[9], P-NET[11] and PathCNN[12]. Among these, eiNN and eiCNN are early integration methods based on NN and CNN; liNN and liCNN are late integration methods based on fully connected neural network (NN) and convolutional neural network (CNN); MOGONet and MOGAT are multi-modal integration methods based on graph neural network; P-NET and PathCNN are representative multi-modal integration methods that combines pathway information.

To evaluate the performance, we tested the methods on multiple TCGA datasets for two tasks, cancer survival prediction and cancer staging. DNA methylation, DNA copy number variation (CNV), and RNA expression were used as input. The results of optimal hyperparameter combination for each dataset are listed in **Supplementary Table 2**. Considering the numbers of sub-classes in the TCGA data are imbalanced, we utilized the macro-averaged F1 score as the primary evaluation metric for hyperparameter optimization and performance evaluation (**Fig. 2 and**

261 **Tables 1-2**). We also list the other evaluation indicators (e.g., AUC) in **Supplementary Table 3**.

262 In general, Pathformer significantly performed better than the other 18 integration methods in terms of
263 F1score_macro score (**Fig. 2a**) and cross-validation variances (**Fig. 2b**). In cancer low- and high-risk survival
264 classification tasks, comparing to the other eight deep learning methods (Type III), Pathformer's F1score_marco
265 showed average improvements between 6.3% and 14.6%. When comparing to eiXGBoost, which performed best in
266 the conventiaonl machine learning methods (Types I and II), Pathformer's F1score_marco showed an average
267 improvement of 8.3%. In early- and late-stage classification tasks, comparing to the deep learning methods (Type
268 III), Pathformer's F1score_marco showed average improvements between 5.1% and 12%. Compared to eiXGBoost,
269 Pathformer's F1score_marco showed an average improvement of 6.2%. Moreover, Pathformer also showed good
270 generalization ability in terms of stability (**Fig. 2b**).

271 The detailed performance comparisons of Pathformer and other integration methods for different cancer types
272 are shown in **Tables 1-2** and **Supplementary Table 3**. In survival classifications, Pathformer achieved the highest
273 F1score_macro and F1score_weighted in all the 10 datasets, and the highest AUC in 7 of 10 datasets. In stage
274 classifications, Pathformer achieved the highest F1score_macro in 9 of 10 datasets, the highest F1score_weighted
275 in 8 of 10 datasets, and the highest AUC in 6 of 10 datasets.

276 Ablation analysis of Pathformer

277 We used ablation analysis to evaluate the contributions of different input modalities and calculation modules in the
278 Pathformer model, based on 9 datasets for cancer survival prediction (**Fig. 3**) and 9 datasets for cancer staging
279 (**Supplementary Fig. 5**). The pan-cancer dataset of the 10 datasets was not used here. First, we evaluated different
280 input modalities, including RNA expression, DNA methylation, DNA CNV, and a combination thereof. By
281 comparing the performances of seven models on cancer survival risk classification, we discovered that the model
282 with all three modalities as input achieved the best performance, followed by RNA expression and DNA CNV
283 combined model and DNA methylation-only model (**Fig. 3a**). Furthermore, we observed that the performances of
284 models with single modality can vary greatly between datasets. For example, DNA methylation-only model
285 performed better than RNA expression-only and DNA CNV-only in the LUSC, LIHC, and LGG datasets, but the
286 opposite performances were observed in the LUAD and BLCA datasets. Ablation analysis of different modalities
287 on cancer stage classification showed similar results (**Supplementary Fig. 5a**). These findings underscore the
288 distinct behaviors of different modalities in different cancer types, highlighting the necessity of multi-modal data
289 integration in various cancer stage and survival risk classification tasks.

290 Next, we evaluated the essentialities of different calculation modules in Pathformer. We developed 4 models,
291 namely CC-attention, Transformer, PSNN, and NN, which successively remove one to multiple modules of
292 Pathformer. The "CC-attention" model is Pathformer without pathway crosstalk network bias. The "Transformer"
293 model is Pathformer without pathway crosstalk network bias and row-attention, using only normal attention
294 mechanism and pathway embeddings. The "PSNN" model directly uses classification module with pathway
295 embedding as input. The "NN" model directly uses classification module with gene embedding as input. As shown
296 in **Fig. 3b** and **Supplementary Fig. 5b**, the complete Pathformer achieved the best classification performance, while
297 the performance of CC-Attention, Transformer, PSNN, and NN decreased successively. This indicated that pathway
298 crosstalk network, attention mechanism, and pathway embedding are all integral components of Pathformer. In
299 particular, CC-attention exhibited significantly poorer classification performance compared to Pathformer,
300 providing strong evidence for the necessity of incorporating pathway crosstalk in Pathformer.

301 Biological significance revealed by Pathformer for breast cancer prognosis using tissue data

302 To further understand the decision-making process of Pathformer and validate the reliability of its biological
303 interpretability, we showed a case study on breast cancer survival risk classification. We demonstrated that
304 Pathformer can use attention weights and SHAP values to reveal modalities, pathways, and genes related to breast
305 cancer prognosis, which align with biological knowledges (**Fig. 4**).

306 First, at the omics and modality level, we visualized the contributions of different modalities for breast cancer
307 survival risk classification by the attention weights (**Fig. 4a**). The contribution of transcriptomic data was greatest
308 in breast cancer prognostic prediction, which is consistent with the results of ablation analysis (**Fig. 3a**) and findings
309 from other literatures[31,32]. Combining with the results of stage classification using breast cancer tissue multi-modal
310 data (**Supplementary Figs. 6a**), we observed that the contributions of the three modalities in both tasks were similar,
311 but the contributions of different features for DNA methylation in breast cancer prognosis classification were
312 comparable. In other words, the contributions of various features in the same modality were also different for
313 different classification tasks. These findings further validated the necessity of multi-modal integration and biological
314 multi-modal embedding process.

315 Next, at the pathway and gene levels, we identified key pathways with top 15 SHAP values and key genes with
316 top 5 SHAP values for each pathway in breast cancer survival risk classification (**Fig. 4b**). Then, we presented a
317 hub module of the updated pathway crosstalk network (**Fig. 4c**). Here, *complex I biogenesis* pathway, which
318 participates in cancer cell proliferation and metastasis[33], was identified as the most critical pathway in the

classification and a key node in the hub module of the updated pathway crosstalk network. Five mitochondrial genes (MT-ND4, MT-ND1, MT-ND3, MT-ND6, and MT-ND2) were identified as key genes of the *complex I biogenesis* pathway in breast cancer prognostic classification by Pathformer, which is consistent with the literatures[34]. In addition, compared to stage classification, *Anaplastic Lymphoma Kinase (ALK) signaling* pathway is specific to breast cancer prognosis classification. ALK promotes cell proliferation by inhibiting apoptosis[35]. Overexpression, deletion, and copy number variations of ALK are associated with different breast cancer subtypes[36-38], especially triple negative breast cancer with poor prognosis[39]. Furthermore, through the hub module of the updated pathway crosstalk network (**Fig. 4c**), we discovered a series of regulatory relationships related to immune response, centered around *FCERI-mediated MAPK activation* pathway[38], which are crucial to prognosis classification. These pathways impact the tumor immune response by influencing the energy supply[40], activation[41], and signal transduction[42] of immune cells.

Subsequently, to facilitate a more intuitive understanding of the impact of active pathways identified by Pathformer on breast cancer survival risk classification, we depicted survival curves comparing patients with high and low scores in active pathways (**Fig. 4d** and **Supplementary Fig. 7**). The pathway score for each sample was obtained by averaging across different dimensions of pathway embedding. Log-rank tests indicated that most active pathways identified by Pathformer, like *complex I biogenesis* pathway, significantly influenced patient survival.

Performance of Pathformer for the noninvasive diagnosis of cancer using liquid biopsy data

In clinical practice, cancer diagnosis involves not only using tissue data for cancer staging but also using liquid biopsy data (i.e., plasma) for noninvasive early detection and screen. The later has even greater clinical significance because early detection substantially increases five-year survival rate of cancer patients. For instance, colon cancer's 5-year survivals were reported as 93.2% for stage I, and only 8.1% for stage IV[43]. Therefore, we applied Pathformer to liquid biopsy data, aiming to classify cancer patients from healthy controls. We curated two types of cell-free RNA sequencing (cfRNA-seq) data derived from plasma and platelet, respectively. We then calculated seven RNA-level modalities as Pathformer's multi-modal input, including RNA expression, RNA splicing, RNA editing, RNA alternative promoter (RNA alt. promoter), RNA allele-specific expression (RNA ASE), RNA single nucleotide variations (RNA SNV), and chimeric RNA (**Supplementary Note 5**). Because these seven modalities of RNA may have information redundancy, we selected the best modality combination based on 2 times 5-fold cross validations (**Supplementary Note 8**). The results showed that the plasma data with 7 modalities and the platelet data with 3 modalities obtain the best performances (AUCs > 0.9) (**Table 3** and **Supplementary Fig. 9**). Because cancer screen

348 usually requires high specificity, we particularly report sensitivities on 99% specificity in **Table 3**, all of which are
349 above 45%. It is worth noting that the sensitivity is still above 45% on 99% specificity in the plasma data even for
350 the early-stage cancer patients, showing its potential for early cancer diagnosis. In addition, we showed that
351 Pathformer's performance was superior to the other representative integration methods using the liquid biopsy data
352 (**Supplementary Table 4**).

353 Biological significance revealed by Pathformer in cancer patient's blood

354 Based on the above analysis, we also gained new insight into the deregulated alterations in plasma through
355 Pathformer's biological interpretability module. We revealed three interesting pathways deregulated in cancer
356 patients' blood. The first one is *binding and uptake of ligands* (e.g., oxidized low-density lipoprotein, oxLDL) *by*
357 *scavenger receptors* pathway, which was identified as the most active pathway in the plasma data (**Fig. 5a**). It is
358 well established that scavenger receptors play a crucial role in cancer prognosis and carcinogenesis by promoting
359 the degradation of harmful substances and accelerating the immune response[44]. Scavenger receptors are also closely
360 related to the transport process of vesicles, such as stabilin-1[45]. Meanwhile, HBB, HBA1, HBA2, FTH1, HSP90AA1
361 were identified as key genes in this pathway. HBB, as a known biomarker, has been demonstrated to be
362 downregulated in gastric cancer blood transcriptomics[46]. HSP90AA1 is also a potential biomarker for various
363 cancers[47], especially in the blood[48].

364 The other two interesting pathways, *DAP12 signaling* pathway and *DAP12 interactions* pathway, were
365 revealed in both plasma and platelet data (**Fig. 5a and Supplementary Fig. 10a**). DAP12 triggers natural killer cell
366 immune responses against certain tumor cells[49], which is regulated by platelet[50]. Among the top 5 key genes of
367 DAP12 related pathway, B2M has been reported as a serum protein encoding gene and a widely recognized tumor
368 biomarker[51]; HLA-E and HLA-B were reported as cancer biomarkers in tissue and plasma[52,53].

369 Furthermore, Pathformer provides insight into the interplay between various biological processes and their
370 impact on cancer progression by updating pathway crosstalk network (**Fig. 5b**). In the plasma data, the link between
371 *binding and uptake of ligands by scavenger receptors* pathway and *iron uptake and transport* pathway was a novel
372 addition to the updated network. These two pathways lack the crosstalk relationship in the original network before
373 training, but there are literature reports that SCARA5 (member of scavenger receptor class A) can function as a
374 ferritin receptor[54]. The crosstalk between two pathways was amplified by Pathformer in plasma dataset, probably
375 because they were important for classification and shared the same key gene, FTH1 (gene encoding the heavy
376 subunit of ferritin), one of two intersecting genes between the two pathways. In summary, Pathformer's updated

pathway crosstalk network visualizes the information flow between pathways related to cancer classification task in the liquid biopsy data, providing novel insight into the crosstalk of biological pathways in cancer patients' plasma.

Conclusion and Discussion

Pathformer successfully applied a Transformer model to the integrate multi-modal data for cancer diagnosis and prognosis. Particularly, it introduced a novel biological embedding method based on the compacted multi-modal vectors (**Fig. 1b**). Moreover, it utilized the criss-cross attention mechanism of Transformer to capture crosstalk between biological pathways and regulation between modalities (i.e., different omics).

Clinical Applications of Pathformer. Pathformer can be applied to various classification tasks in disease diagnosis and prognosis, such as early detection, cancer staging, and survival prediction. Its prediction performance and biological interpretability were demonstrated through substantial benchmark and case studies focusing on cancer prognosis and noninvasive diagnosis. Moreover, this framework is adaptable for the diagnosis and prognosis of other complex diseases, like autoimmune disease, neurodegenerative diseases, etc.

Potential targets revealed in cancer patients' blood. Particularly, we revealed the scavenger receptor related pathway and DAP12 related pathway in cancer patients' blood, which are associated with extracellular vesicle transport[45] and immune response[49], respectively. We even found a new cancer-related pathway crosstalk in blood, which is between *binding and uptake of ligands by scavenger receptors* pathway and *iron uptake and transport* pathway. These results provide candidate targets for the mechanism study of cancer microenvironment and immune system, and even new targets for cancer treatment.

Limitations of Pathformer and future directions. Pathformer used genes involved in pathways from four public databases, all of which consist of protein-coding genes. However, a substantial body of literature has reported that noncoding RNAs are also crucial in cancer prognosis and diagnosis[55]. Therefore, incorporating noncoding RNAs and their related functional pathways into Pathformer would be promising for future work. In addition, we used the multi-modal features derived from cfRNA-seq only in the application of liquid biopsy, because the published cell-free multi-omics datasets[28] are usually too small to be train-and-tested. Another flaw of Pathformer is the computing memory issue. Pathway embedding of Pathformer has prevented memory overflow of Transformer module caused by long inputs. However, when adding more pathways or gene sets (e.g., transcription factors), Pathformer still faces the issue of memory overflow. In the future work, we may introduce linear attention to further improve computational speed.

Declarations

Data availability

All datasets used in this study are publicly available for academic research usages. The details of usage are also fully illustrated in Methods and Supplementary Notes.

Code availability

Source code for data preprocessing and model training is freely available at Github (https://github.com/lulab/Pathformer) with detailed instructions. Source code for comparing the other methods is also included.

Consent for publication

All authors have approved the manuscript and agree with the publication.

Competing interests

The authors declare that they have no competing interests.

Funding and Acknowledgements

This work is supported by National Natural Science Foundation of China (32170671, 8237061277), Tsinghua University Guoqiang Institute Grant (2021GQG1020), Tsinghua University Initiative Scientific Research Program of Precision Medicine (2022ZLA003), Bioinformatics Platform of National Center for Protein Sciences (Beijing) (2021-NCPSB-005). This study was also supported by Bayer Micro-funding, Bio-Computing Platform of Tsinghua University Branch of China National Center for Protein Sciences.

References

1 Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome biology* **18**, 1-15 (2017).

2 Tarazona, S., Arzalluz-Luque, A. & Conesa, A. Undisclosed, unmet and neglected challenges in multi-omics studies. *Nature Computational Science* **1**, 395-402 (2021).

3 Rohart, F., Gautier, B., Singh, A. & Lê Cao, K.-A. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS computational biology* **13**, e1005752 (2017).

4 Kuru, H. I., Tastan, O. & Cicek, A. E. MatchMaker: a deep learning framework for drug synergy prediction. *IEEE/ACM transactions on computational biology and bioinformatics* **19**, 2334-2344 (2021).

5 Preuer, K. *et al.* DeepSynergy: predicting anti-cancer drug synergy with Deep Learning. *Bioinformatics* **34**, 1538-1546 (2018).

6 Islam, M. M. *et al.* An integrative deep learning framework for classifying molecular subtypes of breast cancer. *Computational and structural biotechnology journal* **18**, 2185-2199 (2020).

7 Fu, Y. *et al.* A gene prioritization method based on a swine multi-omics knowledgebase and a deep learning model. *Communications Biology* **3**, 502 (2020).

8 Wang, T. *et al.* MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nat Commun* **12**, 3445 (2021). https://doi.org:10.1038/s41467-021-23774-w

9 Xing, X. *et al.* in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 556-561 (IEEE).

10 Hao, J., Kim, Y., Kim, T. K. & Kang, M. PASNet: pathway-associated sparse deep neural network for prognosis prediction from high-throughput data. *BMC Bioinformatics* **19**, 510 (2018). https://doi.org:10.1186/s12859-018-2500-z

11 Elmarakeby, H. A. *et al.* Biologically informed deep neural network for prostate cancer discovery. *Nature* **598**, 348-352 (2021).

12 Oh, J. H. *et al.* PathCNN: interpretable convolutional neural networks for survival prediction and pathway analysis applied to glioblastoma. *Bioinformatics* **37**, i443-i450 (2021). https://doi.org:10.1093/bioinformatics/btab285

13 Li, Y., Agarwal, P. & Rajagopalan, D. A global pathway crosstalk network. *Bioinformatics* **24**, 1442-1447 (2008).

14 Banoth, B. *et al.* Stimulus-selective crosstalk via the NF-κB signaling system reinforces innate immune response to alleviate gut infection. *Elife* **4**, e05648 (2015).

15 Jiang, J. & Hui, C.-c. Hedgehog signaling in development and cancer. *Developmental cell* **15**, 801-812 (2008).

16 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).

17 Cui, H. *et al.* scGPT: Towards building a foundation model for Single-Cell multi-omics using generative AI. *bioRxiv*, 2023.2004. 2030.538439 (2023).

18 Osseni, M. A., Tossou, P., Laviolette, F. & Corbeil, J. MOT: a Multi-Omics Transformer for multiclass classification tumour types predictions. *BioRxiv*, 2022.2011. 2014.516459 (2022).

19 Theodoris, C. V. *et al.* Transfer learning enables predictions in network biology. *Nature*, 1-9 (2023).

20 Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **30** (2017).

21 Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28**, 27-30 (2000).

22 Schaefer, C. F. *et al.* PID: the pathway interaction database. *Nucleic acids research* **37**, D674-D679 (2009).

23 Croft, D. *et al.* Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research* **39**, D691-D697 (2010).
24 Nishimura, D. BioCarta. *Biotech Software & Internet Report: The Computer Software Journal for Scient* **2**, 117-120 (2001).
25 Ogris, C., Guala, D., Helleday, T. & Sonnhammer, E. L. A novel method for crosstalk analysis of biological networks: improving accuracy of pathway annotation. *Nucleic acids research* **45**, e8-e8 (2017).
26 Cancer Genome Atlas Research Network, J. The cancer genome atlas pan-cancer analysis project. *Nat. Genet* **45**, 1113-1120 (2013).
27 Chen, S. *et al.* Cancer type classification using plasma cell-free RNAs derived from human and microbes. *eLife* **11**, e75181 (2022).
28 Tao, Y. *et al.* Cell-free multi-omics analysis reveals potential biomarkers in gastrointestinal cancer patients' blood. *Cell Reports Medicine* **4** (2023).
29 Best, M. G. *et al.* RNA-Seq of tumor-educated platelets enables blood-based pan-cancer, multiclass, and molecular pathway cancer diagnostics. *Cancer cell* **28**, 666-676 (2015).
30 Best, M. G. *et al.* Swarm intelligence-enhanced detection of non-small-cell lung cancer using tumor-educated platelets. *Cancer cell* **32**, 238-252. e239 (2017).
31 Huang, Z. *et al.* SALMON: survival analysis learning with multi-omics neural networks on breast cancer. *Frontiers in genetics* **10**, 166 (2019).
32 Tong, L., Wu, H. & Wang, M. D. Integrating multi-omics data by learning modality invariant representations for improved prediction of overall survival of cancer. *Methods* **189**, 74-85 (2021).
33 Urra, F. A., Muñoz, F., Lovy, A. & Cárdenas, C. The mitochondrial complex (I) ty of cancer. *Frontiers in oncology* **7**, 118 (2017).
34 Kopinski, P. K., Singh, L. N., Zhang, S., Lott, M. T. & Wallace, D. C. Mitochondrial DNA variation and cancer. *Nature Reviews Cancer* **21**, 431-445 (2021).
35 Motegi, A., Fujimoto, J., Kotani, M., Sakuraba, H. & Yamamoto, T. ALK receptor tyrosine kinase promotes cell growth and neurite outgrowth. *Journal of cell science* **117**, 3319-3329 (2004).
36 Robertson, F. M. *et al.* Presence of anaplastic lymphoma kinase in inflammatory breast cancer. *Springerplus* **2**, 1-12 (2013).
37 Perez-Pinera, P., Chang, Y., Astudillo, A., Mortimer, J. & Deuel, T. Anaplastic lymphoma kinase is expressed in different subtypes of human breast cancer. *Biochemical and biophysical research communications* **358**, 399-403 (2007).
38 Siraj, A. K. *et al.* ALK alteration is a frequent event in aggressive breast cancers. *Breast Cancer Research* **17**, 1-12 (2015).
39 Sachdev, J. C. *et al.* (American Society of Clinical Oncology, 2013).
40 Mills, E. L., Kelly, B. & O'Neill, L. A. Mitochondria are the powerhouses of immunity. *Nature immunology* **18**, 488-498 (2017).
41 Feske, S. Calcium signalling in lymphocyte activation and disease. *Nature Reviews Immunology* **7**, 690-702 (2007).
42 Bhat, M. Y. *et al.* Comprehensive network map of interferon gamma signaling. *Journal of cell communication and signaling* **12**, 745-751 (2018).
43 O'Connell, J. B., Maggard, M. A. & Ko, C. Y. Colon cancer survival rates with the new American Joint Committee on Cancer sixth edition staging. *Journal of the National Cancer Institute* **96**, 1420-1425 (2004).
44 Ryu, S., Howland, A., Song, B., Youn, C. & Song, P. I. Scavenger receptor class A to E involved in various cancers. *Chonnam medical journal* **56**, 1-5 (2020).
45 Kzhyshkowska, J., Gratchev, A. & Goerdt, S. Stabilin‐1, a homeostatic scavenger receptor with multiple functions. *Journal of cellular and molecular medicine* **10**, 635-649 (2006).
46 Lee, I.-S. *et al.* A blood-based transcriptomic signature for noninvasive diagnosis of gastric cancer. *British Journal of Cancer* **125**, 846-853 (2021).

47 Zuehlke, A. D., Beebe, K., Neckers, L. & Prince, T. Regulation and function of the human HSP90AA1 gene. *Gene* **570**, 8-16 (2015).

48 Zhang, P. j. *et al.* Genes expression profiling of peripheral blood cells of patients with hepatocellular carcinoma. *Cell biology international* **36**, 803-809 (2012).

49 Campbell, K. S. & Colonna, M. DAP12: a key accessory protein for relaying signals by natural killer cell receptors. *The international journal of biochemistry & cell biology* **31**, 631-636 (1999).

50 Placke, T., Kopp, H.-G. & Salih, H. R. Modulation of natural killer cell anti-tumor reactivity by platelets. *Journal of innate immunity* **3**, 374-382 (2011).

51 Cooper, E. & Plesner, T. Beta‐2‐microglobulin review: Its relevance in clinical oncology. *Medical and Pediatric Oncology* **8**, 323-334 (1980).

52 Zeestraten, E. *et al.* Combined analysis of HLA class I, HLA-E and HLA-G predicts prognosis in colon cancer patients. *British journal of cancer* **110**, 459-468 (2014).

53 Liu, L. *et al.* A three-platelet mRNA set: MAX, MTURN and HLA-B as biomarker for lung cancer. *Journal of Cancer Research and Clinical Oncology* **145**, 2713-2723 (2019).

54 Yu, B. *et al.* Interactions of ferritin with scavenger receptor class A members. *Journal of Biological Chemistry* **295**, 15727-15741 (2020).

55 Qi, P., Zhou, X.-y. & Du, X. Circulating long non-coding RNAs in cancer: current status and future perspectives. *Molecular cancer* **15**, 1-11 (2016).

Figure Legends

Figure 1. Overview of Pathformer

Schematic of Pathformer (**a**), which integrates multi-oimcs data of tissue and liquid biopsies for disease diagnosis and prognosis. Pathformer has 6 modules: 1) biological pathway and crosstalk network calculation module, 2) multi-omics data input module (**b**), 3) biological multi-modal embedding module (**c**), 4) Transformer module with pathway crosstalk network bias, 5) classification module and 6) biological interpretability module. F_E, conversion function in the gene embedding; G, gene; P, pathway; W, weight of pathway-based sparse neural network.

Figure 2. Performance comparison among multiple multi-omics data integration methods

Average macro-averaged F1 score (**a**) and its standard deviation (**b**) are shown for each method on the TCGA datasets (all cancer types) for cancer low- and high-risk survival classification and early- and late-stage classification, respectively. Error bars are from 5-fold cross-validation repeated twice (10 values) of all datasets.

Figure 3. Ablation analysis of Pathformer for different input modalities and different calculation modules

a. Different types of input modalities (omics data types) were used as input for TCGA cancer low- and high-risk survival classification. **B.** Ablation analysis of different calculation modules in Pathformer. Error bars are from 2 times 5-fold cross-validation across 9 datasets, representing 95% confidence intervals. CC-attention, Pathformer without pathway crosstalk network bias; Transformer, only based on normal attention and pathway embedding; PSNN, only classification module with pathway embedding; NN, only classification module with gene embedding.

Figure 4. Biological interpretation of the breast cancer survival data using Pathformer

a. Contributions of different modalities for breast cancer (BRCA) survival risk classification calculated by attention weights (averaging attention maps of row-attention). **b**. Important pathways and their key genes with top SHapley Additive exPlanations (SHAP) values for BRCA survival risk classification. Among the key genes, different colors represent different pillar modalities of the genes. **c**. A hub module of the updated pathway crosstalk network for BRCA survival risk classification. Color depth and size of node represents the degree of node. Line thickness represents the weight of edge. All links are predicted by Pathformer, where known links are reported by the initial crosstalk network and new links are new predictions. **d**. Keplan-Meier curves of the most active pathway selected identified by Pathformer. P-value calculated through Log-Rank test.

576 **Figure 5. Biological interpretation of the cancer patients' plasma data using Pathformer**

577 **a**. Important pathways and their key genes revealed by Pathformer in the plasma cell free RNA-seq data when
578 classifying cancer patients from healthy controls. The pathways and their key genes were selected with top SHAP
579 values. Among the key genes, different colors represent different pillar modalities (e.g., RNA expression, RNA
580 editing, etc) of the genes. **b**. Hub modules of pathway crosstalk network are shown for plasma cell free RNA-seq
581 data. Color depth and size of node represent the degree of node. Line thickness represents the weight of edge. All
582 links are predicted by Pathformer, where known links are reported by the initial crosstalk network and new links
583 are new predictions.

584

Figure 1

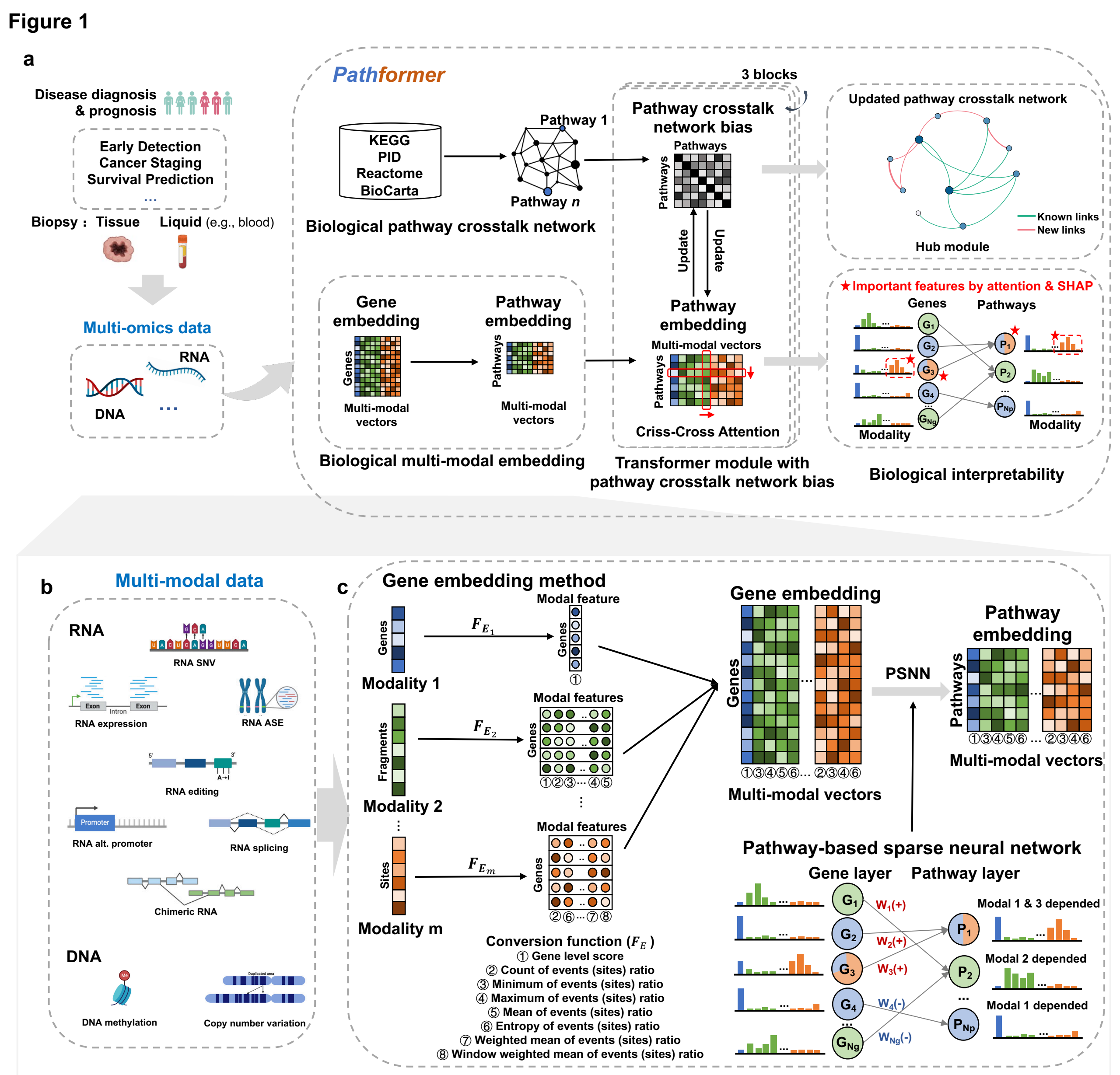

a
Disease diagnosis & prognosis
Early Detection
Cancer Staging
Survival Prediction
...
Biopsy : Tissue
Liquid (e.g., blood)
Multi-omics data
RNA
DNA
...
Pathformer
KEGG
PID
Reactome
BioCarta
Pathway 1
Pathway n
Biological pathway crosstalk network
Gene embedding
Pathway embedding
Genes
Pathways
Multi-modal vectors
Multi-modal vectors
Biological multi-modal embedding
3 blocks
Pathway crosstalk network bias
Pathways
Pathways
Update
Update
Pathway embedding
Multi-modal vectors
Pathways
Criss-Cross Attention
Transformer module with pathway crosstalk network bias
Updated pathway crosstalk network
Known links
New links
Hub module
★Important features by attention & SHAP
Genes
Pathways
Modality
Modality
Biological interpretability
b
Multi-modal data
RNA
RNA SNV
Exon
Intron
Exon
RNA expression
RNA ASE
RNA editing
Promoter
RNA alt. promoter
RNA splicing
Chimeric RNA
DNA
DNA methylation
Copy number variation
c
Gene embedding method
Genes
Modality 1
Modal feature
Fragments
Modality 2
Modal features
Sites
Modality m
Modal features
Conversion function (F_E)
① Gene level score
② Count of events (sites) ratio
③ Minimum of events (sites) ratio
④ Maximum of events (sites) ratio
⑤ Mean of events (sites) ratio
⑥ Entropy of events (sites) ratio
⑦ Weighted mean of events (sites) ratio
⑧ Window weighted mean of events (sites) ratio
Gene embedding
Genes
Multi-modal vectors
PSNN
Pathway embedding
Pathways
Multi-modal vectors
Pathway-based sparse neural network
Gene layer
Pathway layer
Modal 1 & 3 depended
Modal 2 depended
Modal 1 depended

Figure 2

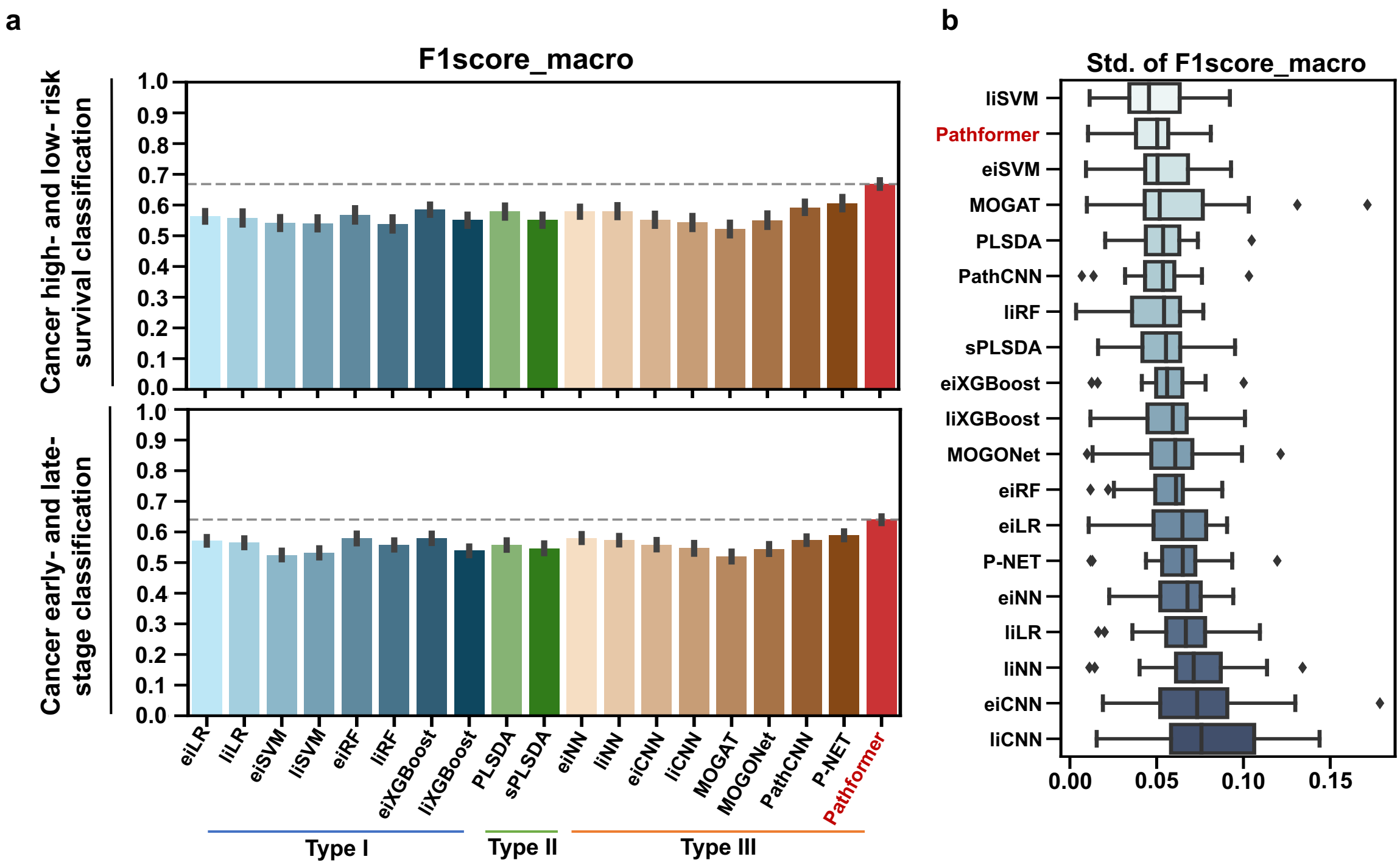

a
F1score_macro
Cancer high- and low- risk survival classification
Cancer early- and late-stage classification
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
eiLR
liLR
eiSVM
liSVM
eiRF
liRF
eiXGBoost
liXGBoost
PLSDA
sPLSDA
eiNN
liNN
eiCNN
liCNN
MOGAT
MOGONet
PathCNN
P-NET
Pathformer
Type I
Type II
Type III
b
Std. of F1score_macro
liSVM
Pathformer
eiSVM
MOGAT
PLSDA
PathCNN
liRF
sPLSDA
eiXGBoost
liXGBoost
MOGONet
eiRF
eiLR
P-NET
eiNN
liLR
liNN
eiCNN
liCNN
0.00
0.05
0.10
0.15

Table 1. Performance comparison among multiple multi-omics data integration methods for cancer survival prediction

	BRCA[1]	KIRC	LUAD	LUSC	HNSC	BLCA	LIHC	SKCM	LGG	Pan-cancer[2]
eiLR	0.571	0.584	0.530	0.520	0.529	0.479	0.497	0.557	0.689	0.674
eiRF	0.494	0.636	0.443	0.477	0.497	0.475	0.579	0.609	0.759	0.709
eiSVM	0.467	0.656	0.487	0.471	0.466	0.459	0.436	0.630	0.640	0.705
eiXGBoost	0.564	0.653	0.530	0.519	0.476	0.549	0.550	0.594	0.723	0.694
liLR	0.569	0.619	0.479	0.481	0.512	0.456	0.487	0.554	0.718	0.696
liRF	0.473	0.622	0.453	0.452	0.478	0.438	0.428	0.609	0.724	0.712
liSVM	0.468	0.614	0.438	0.508	0.470	0.475	0.439	0.561	0.714	0.710
liXGBoost	0.523	0.585	0.481	0.491	0.483	0.504	0.500	0.532	0.715	0.697
PLSDA	0.561	0.660	0.444	0.496	0.527	0.506	0.514	0.635	0.778	0.669
sPLSDA	0.549	0.617	0.450	0.495	0.521	0.465	0.442	0.593	0.726	0.658
eiNN	0.573	0.640	0.503	0.483	0.537	0.497	0.593	0.583	0.695	0.687
eiCNN	0.510	0.548	0.462	0.529	0.512	0.474	0.578	0.582	0.631	0.688
liNN	0.576	0.631	0.521	0.480	0.523	0.518	0.551	0.590	0.702	0.713
liCNN	0.536	0.596	0.504	0.501	0.487	0.460	0.465	0.489	0.706	0.697
MOGONet	0.510	0.632	0.455	0.497	0.469	0.470	0.476	0.599	0.706	0.676
MOGAT	0.466	0.637	0.438	0.434	0.462	0.443	0.448	0.605	0.595	0.689
PathCNN	0.558	0.643	0.522	0.560	0.525	0.486	0.598	0.551	0.766	0.712
P-NET	0.640	0.661	0.551	0.509	0.559	0.470	0.578	0.641	0.715	0.733
Pathformer	**0.673**	**0.688**	**0.633**	**0.615**	**0.606**	**0.601**	**0.651**	**0.694**	**0.787**	**0.735**

[1] abbreviation of cancer type according to the TCGA terms; [2]pan-cancer dataset contains 33 cancer types of TCGA terms.
10 TCGA datasets are tested. Average macro-averaged F1 scores are listed for the two unbalanced classes, high- and low-risk survival cancer patients. Each value is the mean of 5-fold cross-validation repeated twice (10 values).

Table 2. Performance comparison among multiple multi-omics data integration methods for cancer staging.

	BRCA[1]	KIRC	LUAD	LUSC	STAD	BLCA	LIHC	SKCM	THCA	Pan-cancer[2]
eiLR	0.510	0.647	0.506	0.501	0.541	0.612	0.567	0.579	0.613	0.631
eiRF	0.488	0.723	0.503	0.509	0.540	0.647	0.528	0.571	0.660	0.622
eiSVM	0.451	0.661	0.460	0.467	0.528	0.555	0.464	0.551	0.550	0.557
eiXGBoost	0.518	0.686	0.526	0.498	0.548	0.639	0.540	0.535	0.664	0.629
liLR	0.522	0.654	0.473	0.506	0.581	0.624	0.496	0.548	0.626	0.620
liRF	0.454	0.704	0.497	0.469	0.543	0.611	0.507	0.570	0.622	0.604
liSVM	0.436	0.675	0.460	0.470	0.552	0.564	0.438	0.557	0.601	0.567
liXGBoost	0.475	0.670	0.471	0.483	0.493	0.542	0.526	0.516	0.601	0.608
PLSDA	0.456	0.710	0.462	0.487	0.565	0.609	0.521	0.562	0.627	0.572
sPLSDA	0.450	0.717	0.503	0.465	0.540	0.568	0.552	0.514	0.596	0.559
eiNN	0.545	0.669	0.549	0.515	0.555	0.636	0.569	0.614	0.528	0.617
eiCNN	0.522	0.647	0.563	0.501	0.505	0.634	0.584	0.553	0.469	0.606
liNN	0.535	0.681	0.531	0.528	0.518	0.616	0.565	0.582	0.547	0.630
liCNN	0.518	0.643	0.559	0.473	0.482	0.565	0.586	0.557	0.468	0.618
MOGONet	0.466	0.646	0.461	0.476	0.574	0.553	0.474	0.526	0.644	0.624
MOGAT	0.463	0.637	0.493	0.459	0.562	0.554	0.461	0.528	0.574	0.473
PathCNN	0.525	0.694	0.554	0.534	0.521	0.566	0.563	0.540	0.595	0.643
P-NET	0.522	0.624	0.543	0.526	0.537	0.660	0.612	0.571	0.632	**0.664**
Pathformer	**0.573**	**0.726**	**0.629**	**0.564**	**0.596**	**0.706**	**0.616**	**0.646**	**0.690**	0.657

[1] abbreviation of cancer type according to the TCGA terms; [2]pan-cancer dataset contains 21 cancer types of TCGA terms.
10 TCGA datasets are tested. Average macro-averaged F1 scores are listed for the two unbalanced classes, early- (stage I and II) and late-stage (stage III and IV) cancer patients. Each value is the mean of 5-fold cross-validation repeated twice (10 values).

Figure 3

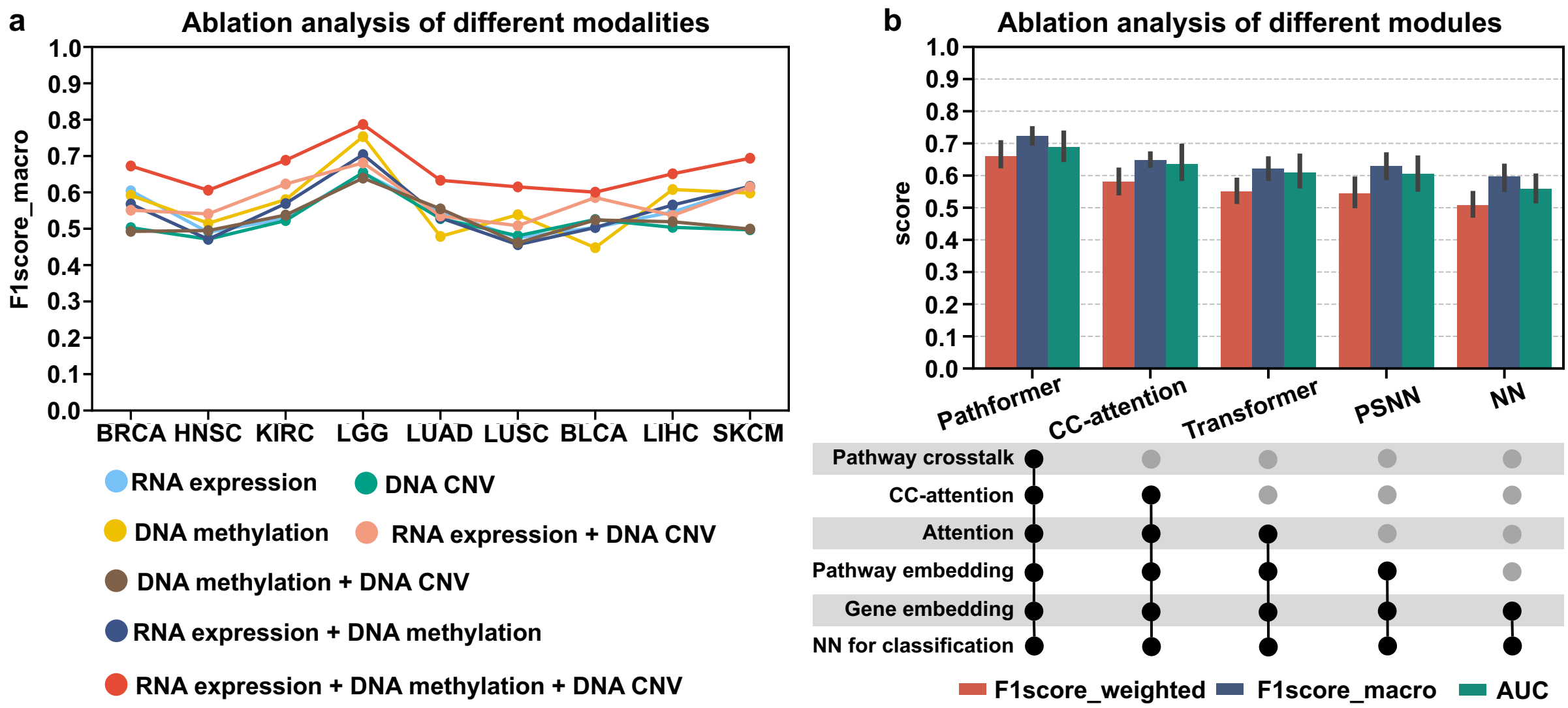

a
Ablation analysis of different modalities
F1score_macro
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
BRCA HNSC KIRC LGG LUAD LUSC BLCA LIHC SKCM
RNA expression
DNA CNV
DNA methylation
RNA expression + DNA CNV
DNA methylation + DNA CNV
RNA expression + DNA methylation
RNA expression + DNA methylation + DNA CNV
b
Ablation analysis of different modules
score
Pathformer
CC-attention
Transformer
PSNN
NN
Pathway crosstalk
CC-attention
Attention
Pathway embedding
Gene embedding
NN for classification
F1score_weighted
F1score_macro
AUC

Figure 4

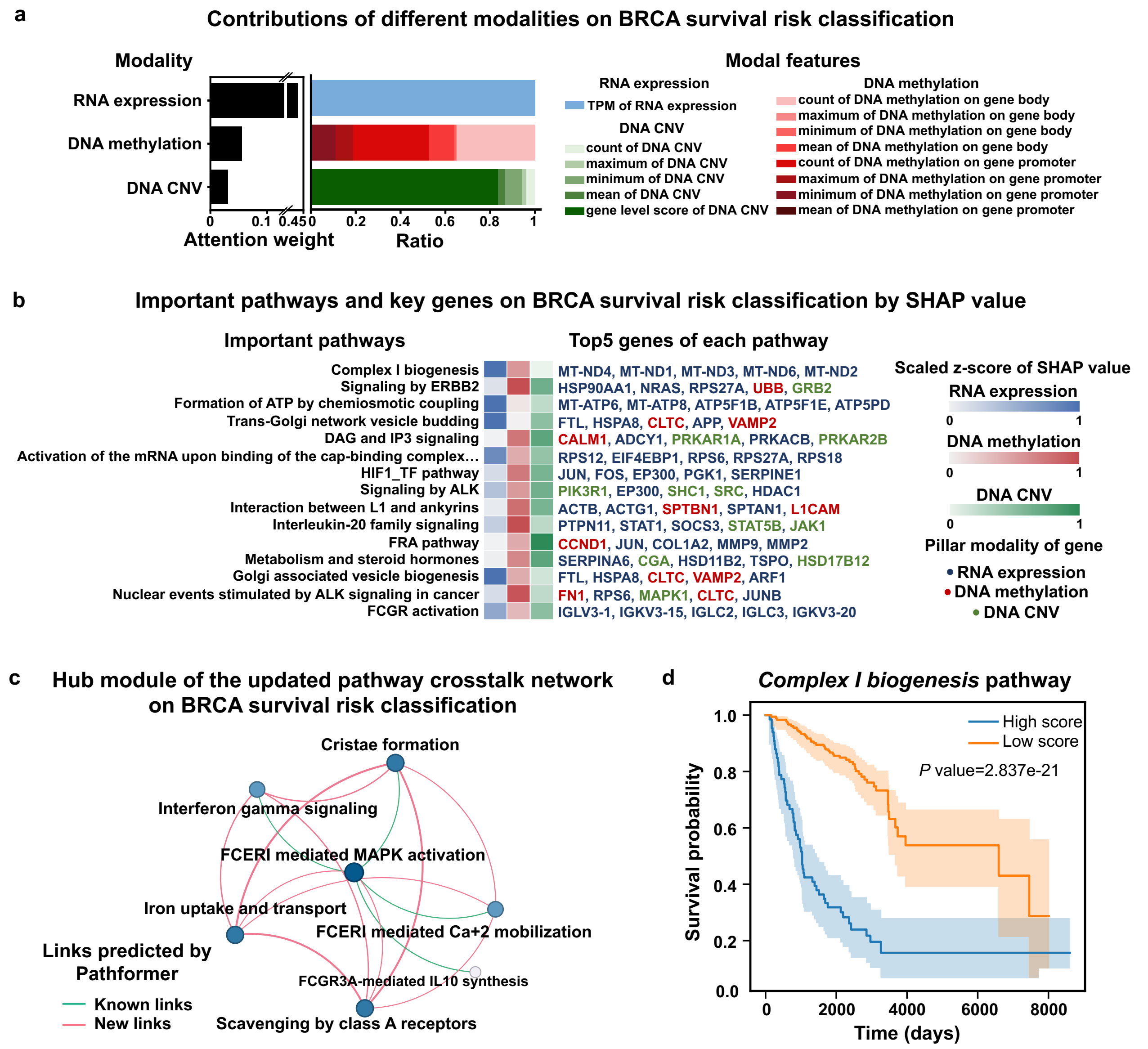
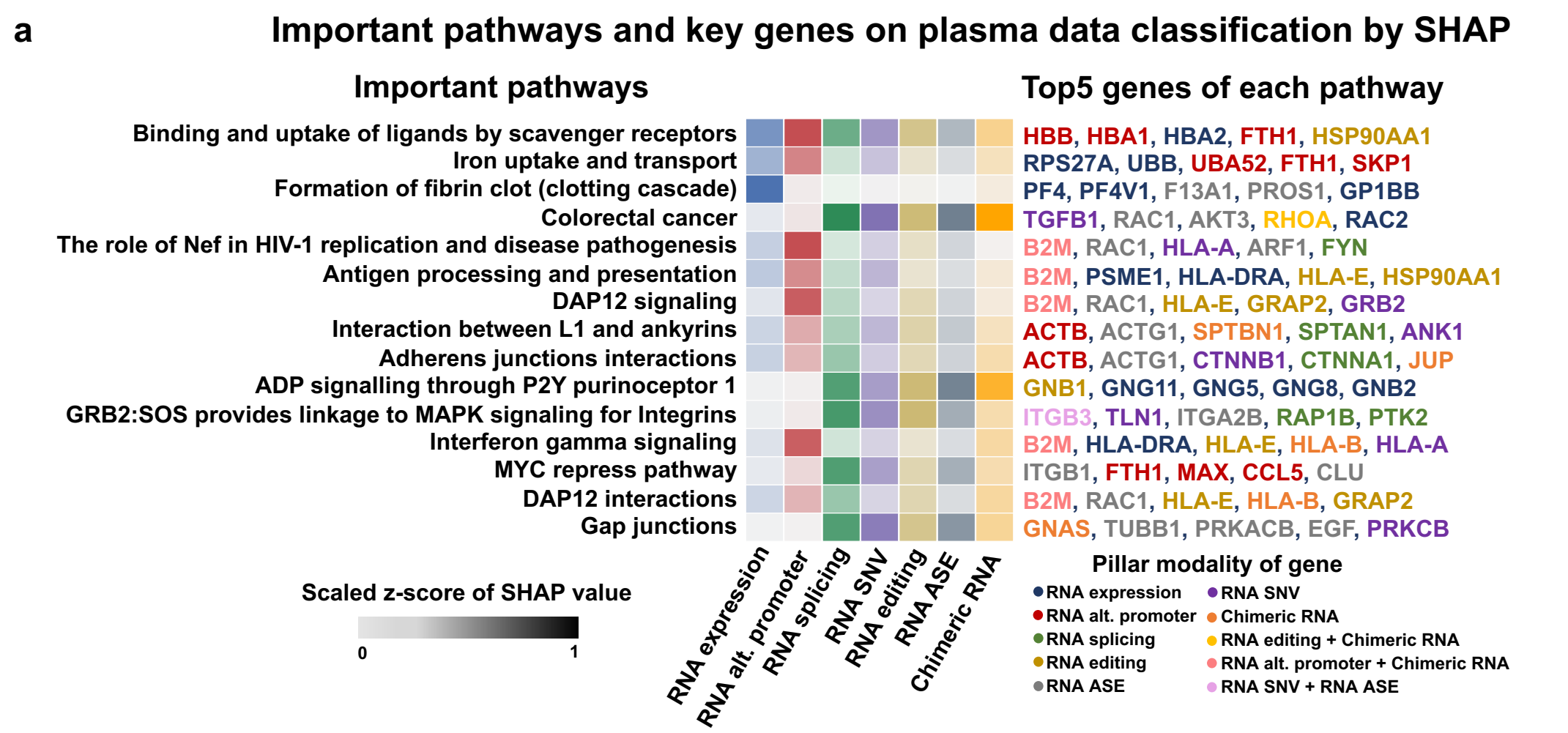

a
Contributions of different modalities on BRCA survival risk classification
Modality
RNA expression
DNA methylation
DNA CNV
0
0.1
0.45
Attention weight
0
0.2
0.4
0.6
0.8
1
Ratio
Modal features
RNA expression
TPM of RNA expression
DNA CNV
count of DNA CNV
maximum of DNA CNV
minimum of DNA CNV
mean of DNA CNV
gene level score of DNA CNV
DNA methylation
count of DNA methylation on gene body
maximum of DNA methylation on gene body
minimum of DNA methylation on gene body
mean of DNA methylation on gene body
count of DNA methylation on gene promoter
maximum of DNA methylation on gene promoter
minimum of DNA methylation on gene promoter
mean of DNA methylation on gene promoter
b
Important pathways and key genes on BRCA survival risk classification by SHAP value
Important pathways
Top5 genes of each pathway
Complex I biogenesis
MT-ND4, MT-ND1, MT-ND3, MT-ND6, MT-ND2
Signaling by ERBB2
HSP90AA1, NRAS, RPS27A, UBB, GRB2
Formation of ATP by chemiosmotic coupling
MT-ATP6, MT-ATP8, ATP5F1B, ATP5F1E, ATP5PD
Trans-Golgi network vesicle budding
FTL, HSPA8, CLTC, APP, VAMP2
DAG and IP3 signaling
CALM1, ADCY1, PRKAR1A, PRKACB, PRKAR2B
Activation of the mRNA upon binding of the cap-binding complex…
RPS12, EIF4EBP1, RPS6, RPS27A, RPS18
HIF1_TF pathway
JUN, FOS, EP300, PGK1, SERPINE1
Signaling by ALK
PIK3R1, EP300, SHC1, SRC, HDAC1
Interaction between L1 and ankyrins
ACTB, ACTG1, SPTBN1, SPTAN1, L1CAM
Interleukin-20 family signaling
PTPN11, STAT1, SOCS3, STAT5B, JAK1
FRA pathway
CCND1, JUN, COL1A2, MMP9, MMP2
Metabolism and steroid hormones
SERPINA6, CGA, HSD11B2, TSPO, HSD17B12
Golgi associated vesicle biogenesis
FTL, HSPA8, CLTC, VAMP2, ARF1
Nuclear events stimulated by ALK signaling in cancer
FN1, RPS6, MAPK1, CLTC, JUNB
FCGR activation
IGLV3-1, IGKV3-15, IGLC2, IGLC3, IGKV3-20
Scaled z-score of SHAP value
RNA expression
0
1
DNA methylation
0
1
DNA CNV
0
1
Pillar modality of gene
RNA expression
DNA methylation
DNA CNV
c
Hub module of the updated pathway crosstalk network on BRCA survival risk classification
Cristae formation
Interferon gamma signaling
FCERI mediated MAPK activation
Iron uptake and transport
FCERI mediated Ca+2 mobilization
FCGR3A-mediated IL10 synthesis
Scavenging by class A receptors
Links predicted by Pathformer
Known links
New links
d
Complex I biogenesis pathway
High score
Low score
P value=2.837e-21
Survival probability
1.0
0.8
0.6
0.4
0.2
0.0
0
2000
4000
6000
8000
Time (days)

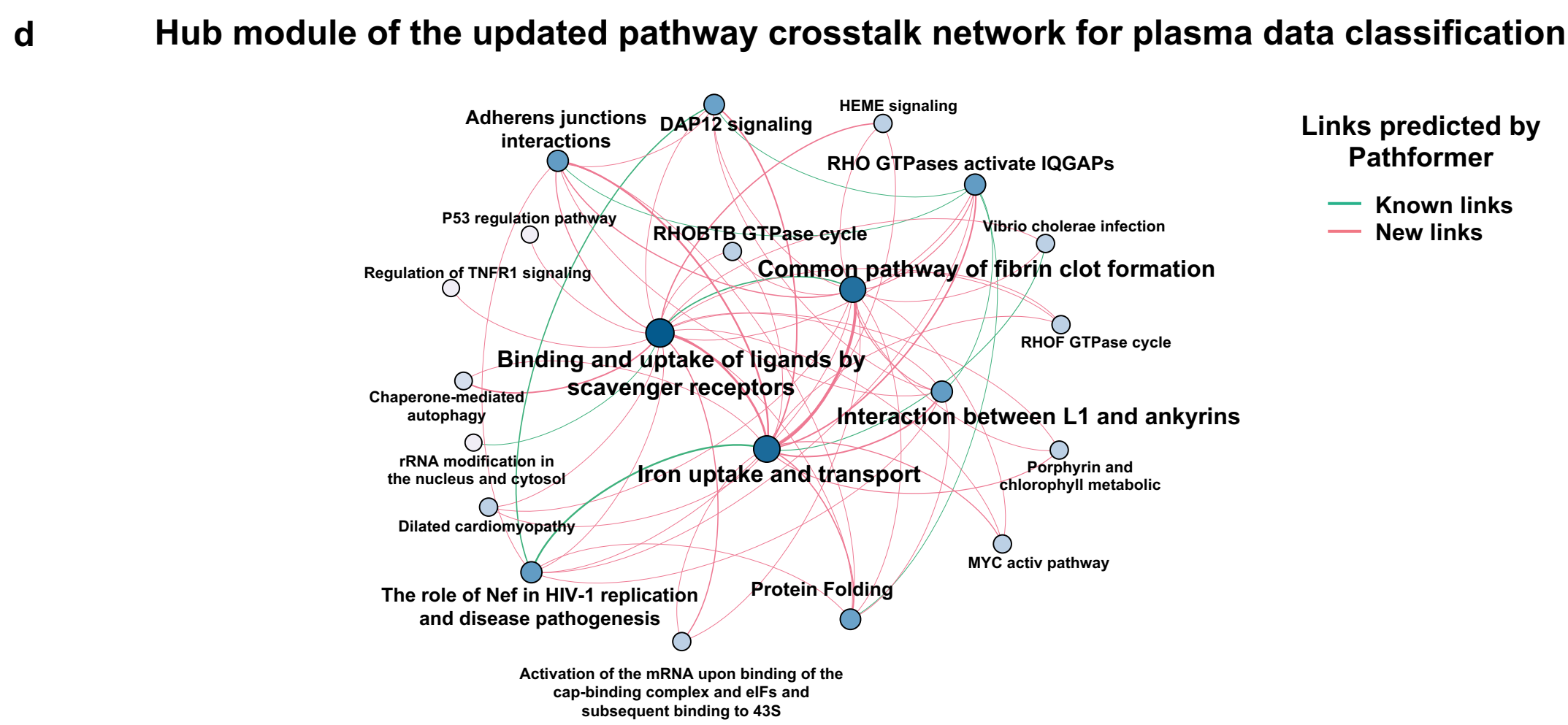
Table 3. Cancer detection performance of Pathformer based on the cell-free RNA liquid biopsy data.

Dataset	macro-averaged F1 score	weighted-averaged F1 score	AUC	Sensitivity (99% specificity)
Plasma[1]	0.843	0.877	0.914	0.488
Plasma (early stage[2])	0.853	0.869	0.916	0.479
Platelet[3]	0.889	0.903	0.938	0.481

[1]all cancer stages from I to IV; [2]cancer stage I and stage II; [3]Stage information not available.
All types of cancer patients are used as positives; the heathy controls are used as negatives. Each value is the mean of 5-fold cross-validation repeated twice (10 values).

Figure 5


a
Important pathways and key genes on plasma data classification by SHAP
Important pathways
Top5 genes of each pathway
Binding and uptake of ligands by scavenger receptors
HBB, HBA1, HBA2, FTH1, HSP90AA1
Iron uptake and transport
RPS27A, UBB, UBA52, FTH1, SKP1
Formation of fibrin clot (clotting cascade)
PF4, PF4V1, F13A1, PROS1, GP1BB
Colorectal cancer
TGFB1, RAC1, AKT3, RHOA, RAC2
The role of Nef in HIV-1 replication and disease pathogenesis
B2M, RAC1, HLA-A, ARF1, FYN
Antigen processing and presentation
B2M, PSME1, HLA-DRA, HLA-E, HSP90AA1
DAP12 signaling
B2M, RAC1, HLA-E, GRAP2, GRB2
Interaction between L1 and ankyrins
ACTB, ACTG1, SPTBN1, SPTAN1, ANK1
Adherens junctions interactions
ACTB, ACTG1, CTNNB1, CTNNA1, JUP
ADP signalling through P2Y purinoceptor 1
GNB1, GNG11, GNG5, GNG8, GNB2
GRB2:SOS provides linkage to MAPK signaling for Integrins
ITGB3, TLN1, ITGA2B, RAP1B, PTK2
Interferon gamma signaling
B2M, HLA-DRA, HLA-E, HLA-B, HLA-A
MYC repress pathway
ITGB1, FTH1, MAX, CCL5, CLU
DAP12 interactions
B2M, RAC1, HLA-E, HLA-B, GRAP2
Gap junctions
GNAS, TUBB1, PRKACB, EGF, PRKCB
Scaled z-score of SHAP value
0
1
RNA expression
RNA alt. promoter
RNA splicing
RNA SNV
RNA editing
RNA ASE
Chimeric RNA
Pillar modality of gene
RNA expression
RNA alt. promoter
RNA splicing
RNA editing
RNA ASE
RNA SNV
Chimeric RNA
RNA editing + Chimeric RNA
RNA alt. promoter + Chimeric RNA
RNA SNV + RNA ASE


d
Hub module of the updated pathway crosstalk network for plasma data classification
Adherens junctions interactions
DAP12 signaling
HEME signaling
RHO GTPases activate IQGAPs
P53 regulation pathway
RHOBTB GTPase cycle
Vibrio cholerae infection
Regulation of TNFR1 signaling
Common pathway of fibrin clot formation
RHOF GTPase cycle
Binding and uptake of ligands by scavenger receptors
Chaperone-mediated autophagy
Interaction between L1 and ankyrins
rRNA modification in the nucleus and cytosol
Iron uptake and transport
Porphyrin and chlorophyll metabolic
Dilated cardiomyopathy
MYC activ pathway
The role of Nef in HIV-1 replication and disease pathogenesis
Protein Folding
Activation of the mRNA upon binding of the cap-binding complex and eIFs and subsequent binding to 43S
Links predicted by Pathformer
Known links
New links