Pathformer: biological pathway crosstalk informed Transformer model based on a multi-modal embedding method

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, 4, *], Yunping Zhu[5, *], 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, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100730, China.

[4]National Clinical Research Center for Dermatologic and Immunologic Diseases (Ministry of Science & Technology), MOE Key Laboratory of Rheumatology and Clinical Immunology, State Key Laboratory of Complex Severe and Rare Diseases, Peking Union Medical College Hospital, Beijing 100730, China.

[5]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

Transformer model is hard to be applied to the integration of biological multi-modal data, mainly due to the challenge of incorporating lots of multi-modal features with efficient memory cost. To address this challenge, we have developed Pathformer, a biological pathway crosstalk informed Transformer model using a novel embedding method based on compacted multi-modal vectors and pathway-based sparse neural network. Pathformer also leverages criss-cross attention mechanism to capture the crosstalk between different pathways and modalities, and utilizes biological interpretability module to reveal key pathways, genes, and regulatory mechanisms. When benchmarking on 20 TCGA datasets, Pathformer substantially outperformed 18 published integration methods, with an average improvement of 6.3%-14.7% in F1 scores for cancer prognosis predictions and 5.1%-12% for stage predictions. Subsequently, we used case studies of prognosis prediction and early cancer diagnosis to demonstrate biological interpretation and clinical utility. For instance, we revealed interesting cancer-related deregulation pathways in blood (e.g., scavenger receptor pathway) from the liquid biopsy data for noninvasive diagnosis.

Keywords: Multi-modal integration; Transformer; Pathway crosstalk; Multi-modal embedding; Cancer diagnosis.

Introduction

The rapid progress in high-throughput technologies has made it possible to curate multi-modal data for disease studies using genome-wide platforms. Compared to single type of data, integrating multi-modal data offers a more comprehensive view of gene regulation in diseases[1] and addresses certain key challenges in disease research, such as heterogeneity and complex molecular interactions[2]. In addition to the previous methods that mainly focus on unsupervised problems, many supervised multi-modal integration methods have recently been developed for biomedical classification problems, such as mixOmics[3], liNN[4], eiNN[5], liCNN[6], eiCNN[7], MOGONet[8] and MOGAT[9]. These methods did not incorporate prior biological knowledge. Therefore, several deep learning studies attempted to include the biological pathway information in the model. For instance, 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 published for the classification of prostate cancer patients[11]. Another method, PathCNN, was developed to predict the prognosis of GBM patient by using principal component analysis (PCA) algorithms to define multi-modal pathway images and a convolutional neural network[12].

All the above methods still did not consider the crosstalk between pathways. The phenomenon where stimulating receptors in one pathway leads to downstream regulatory effects on target genes in other pathways is referred to as pathway crosstalk, which has even more biological significance than pathway itself[13,14]. The criss-cross attention mechanism of Transformer would be very useful to capture the crosstalk information[15]. However, the current embedding methods for biological data in Transformer models[16-18], such as discretization and linear transformation, usually missed the biological information. Meanwhile, multi-modal features were typically multiplied by the number of genes, resulting in long input sequences, which could strain available memory resources or necessitate a feature selection step[18,19].

Here, we propose Pathformer, which combines pathway crosstalk networks and the Transformer encoder with bias for the classification and interpretation of multi-modal data. Pathformer has three essential steps. First, it ingeniously concatenates different modalities into a compacted multi-modal vector for each gene, ensuring that each dimension of the gene embedding retains biological significance. Second, Pathformer aptly utilizes a sparse neural network based on prior pathway knowledge to transform gene embeddings into pathway embeddings, which not only captures valuable information but also addresses memory explosion issue. Third, Pathformer naturally incorporates pathway crosstalk networks into a Transformer model with bias to enhance the exchange of information between different pathways and modalities.

We further compared Pathformer with 18 other integration models on 20 benchmark datasets of the Cancer Genome Atlas (TCGA)[20], and demonstrated its superior performance and biological interpretability on various classification tasks. In addition, we conducted case studies on cancer prognosis prediction and non-invasive diagnosis, revealing interesting pathways, genes, and regulatory mechanisms related to cancer in human tissue and plasma, demonstrating the high accuracy and interpretability of Pathformer in clinical applications.

Materials and methods

Overview of Pathformer

Pathformer is a universal and interpretable multi-modal integration model that can be applied in different biomedical classification tasks, including cancer prognosis predictions, stage predictions, and non-invasive diagnosis (**Fig. 1a**). Specifically, Pathformer model consists of the following six modules: biological pathway and crosstalk network calculation module, multi-modal data input module, biological multi-modal embedding module, Transformer module with pathway crosstalk network bias, classification module and biological interpretability module. Pathformer introduces a novel and universal embedding method for incorporating biological multi-modal data at gene level and pathway level. It begins by uniformly transforming different modalities to the gene level through a series of statistical indicators. Subsequently, it concatenates these modalities into compacted multi-modal vectors to define gene embedding. Finally, it employs a sparse neural network based on the gene-to-pathway mapping to transform gene embedding into pathway embedding. It then enhances the fusion of information between various modalities and pathways by combining pathway crosstalk networks with Transformer encoder. Finally, a fully connected layer serves as the classifier for different downstream classification tasks. In addition, Pathformer uses biological interpretable module with attention weights and SHapley Additive exPlanations[21] values to identify important genes, pathways, modalities, and regulatory processes for providing explanations of the biological mechanisms behind the model from multiple perspectives.

Biological pathways and crosstalk network calculation

We collected 2,289 pathways of four public databases including Kyoto Encyclopedia of Genes and Genomes database (KEGG)[22], Pathway Interaction database (PID)[23], Reactome database (Reactome)[24], and BioCarta Pathways database (BioCarta)[25]. 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*[26] 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**).

Biological multi-modal data input and embedding method

Multi-modal biological data preprocessing and embedding method are crucial components of Pathformer (**Fig. 1b**). 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 TCGA datasets 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**.

We proposed a novel 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 lth 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 jth 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.

Transformer module with pathway crosstalk network bias

Transformer module with pathway crosstalk network bias is the key module of Pathformer model (**Supplementary Fig. 1**). Inspired by the Evoformer model used in AlphaFold2[15] for processing multiple sequences, we developed the Transformer module based on criss-cross attention (CC-attention) with bias for data fusion of pathways and modalities. The Transformer 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.

Classification module

In order to solve the classification tasks, 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**.

Biological interpretability module

In addition, to comprehend Pathformer's decision-making process, we introduce a biological interpretable module. This 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 within the classification.

For calculating contribution of 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**.

For identifying important pathways and their key genes, we used SHapley Additive exPlanations[21] (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**.

In Pathformer, 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 prior information but also multi-modal data information, 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**.

Results and Discussion

Experimental settings

Data collection and preprocessing

For benchmark testing, we collected 20 datasets across different cancer types from TCGA to evaluate classification performance of Pathformer and existing comparison methods, which consists of 10 datasets for early- and late-stage classification, and 10 datasets for low- and high-risk survival classification. Besides, to further verify the effect of Pathformer in non-invasive cancer diagnosis, we also collected three types of body fluid datasets: the plasma dataset (comprising 373 samples assayed by total cell-free RNA-seq[27,28]), the extracellular vesicle (EV) dataset (comprising 477 samples from two studies assayed by exosomal RNA-seq[29,30]), and the platelet dataset (comprising 918 sample from two studies assayed by tumor-educated blood platelet RNA-seq[31,32]). Through our biological information pipeline, totally 3 and 7 biological modalities are obtained for TCGA dataset and liquid biopsy dataset, respectively. More details are described in **Supplementary Fig. 2** and **Supplementary Note 5**.

Model training and test

In this study, 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 divided the labeled dataset into the discovery set (75%) and the validation set (25%) hierarchically. We implemented model training, hyperparameter optimization and model early stopping on the discovery set and tested

230 on the validation set (**Supplementary Fig. 2**).

231 When training the model, we applied cross-entropy loss with class-imbalance weight as the label prediction
232 loss, the ADAM optimizer to train Pathformer, and the cosine annealing learning rate method to optimized learning
233 rate. For hyperparameter optimization, we used grid search with 5-fold cross-validation in the discovery set with
234 the macro-averaged F1 score as the selection criterion. For early stopping, we divided the discovery set into the
235 training set (75%) and the test set (25%) hierarchically, and used the macro-averaged F1 score of the test set as the
236 criterion for stopping training. When testing the model, we used the best model trained with optimal hyperparametric
237 combination in the validation set. More details of model training and test are described in **Supplementary Note 6**.

238 ***Performance evaluation and comparison***

239 We compared the classification performance of Pathformer with four existing types of supervised multi-modal
240 integration methods. The first type consists of eight deep learning-based integration methods, including eiNN[5],
241 liNN[4], eiCNN[7], liCNN[6], MOGONet[8], MOGAT[9], P-NET[11] and PathCNN[12]. eiNN and eiCNN are early integration
242 methods based on NN and CNN. liNN and liCNN are late integration methods based on fully connected neural
243 network (NN) and convolutional neural network (CNN). MOGONet and MOGAT are multi-modal integration
244 methods based on graph neural network. P-NET and PathCNN are representative multi-modal integration methods
245 that combines pathway information. The second type includes partial least squares-discriminant analysis (PLSDA)
246 and sparse partial least squares-discriminant analysis (sPLSDA) of mixOmics[3]. The third type includes early
247 integration methods based on base classifiers, i.e., support vector machine (SVM), logistic regression (LR), random
248 forest (RF), and extreme gradient boosting (XGBoost). The fourth type includes late integration methods based on
249 SVM, LR, RF, and XGBoost. During comparison methods, the multi-modal data were preprocessed with the
250 statistical indicators and features were prefiltered with ANOVA as input (see **Supplementary Note 7**).

251 In order to fairly evaluate the Pathformer model and other integration methods, we use area under the receiver
252 operating characteristic curve (AUC), average F1 score weighted by support (F1score_weighted), and macro
253 covered F1 score (F1score_macro) as evaluation indicators. Especially, we use F1score_macro as the main
254 evaluation criterion because it is a more robust measurement than the other two scores for the imbalanced classes.

255 Pathformer outperforms 18 existing multi-modal integration methods in various classification 256 tasks using 20 benchmark datasets

257 To evaluate the performance of Pathformer, we tested model on prognosis and stage predictions and as benchmark
258 studies. For these tasks, DNA methylation, DNA copy nunumber variation (CNV), and RNA expression were used

as input. For 20 benchmark studies, we performed 2 times 5-fold cross-validation that divided the data into a discovery set (75%) and a validation set (25%) for each test, optimized hyperparameters and on the discovery set, and tested Pathformer with optimal parameters on the validation set. The results of optimal hyperparameter combination for each dataset are listed in **Supplementary Fig. 3** and **Supplementary Table 2**. Considering data imbalance, we utilize the macro-covered F1 score as the primary evaluation metric for hyperparameter optimization and classification performance.

Figure 2 shows the comprehensive comparison results across 20 benchmark datasets for Pathformer and 18 other existing multi-modal integration methods. In general, Pathformer significantly performed better than 18 other integration methods (**Fig. 2a**). In cancer low- and high-risk survival classification tasks, comparing to the other eight deep learning methods (i.e., eiNN, liNN, eiCNN, liCNN, MOGONet, MOGAT, PathCNN and P-NET), Pathformer's F1score_marco showed average improvements between 6.3% and 14.6%. When comparing to eiXGBoost, which performed best in the traditional machine learning methods (i.e., eiLR, eiSVM, eiRF, liLR, liSVM, liRF, liXGBoost, PLSDA and sPLSDA), Pathformer's F1score_marco showed an average improvement of 8.3%. In early- and late-stage classification tasks, comparing to the other eight deep learning methods, Pathformer's F1score_marco showed average improvements between 5.1% and 12%. Compared to eiXGBoost, Pathformer's F1score_marco showed an average improvement of 6.2%.

Then, **Table 1** and **Table 2** respectively illustrate the comparison of F1score_macro for Pathformer and 18 other integration methods in prognosis and stage prediction tasks. Other evaluation indicators are shown in **Supplementary Table 3**. In survival risk classifications, Pathformer achieved the best F1score_macro and F1score_weighted in all the 10 datasets, and the best AUC in 7 of 10 datasets. Furthermore, compared to the top-performing control method, Pathformer demonstrated an 8.2% improvement in F1 score_macro in the LUAD dataset, while exceeding 5% improvement in LUSC, BLCA, LIHC, and SKCM datasets. In stage classifications, Pathformer achieved the best F1score_macro in 9 of 10 datasets, the best F1score_weighted in 8 of 10 datasets, and the best AUC in 6 of 10 datasets. Furthermore, compared to the top-performing control method, Pathformer demonstrated an 6.6% improvement in F1 score_macro in the LUAD dataset, while exceeding 3% improvement in LUSC, BLCA, and SKCM datasets. These results highlight Pathformer's exceptional learning ability. Moreover, in terms of stability, Pathformer also showed significantly better generalization ability than other deep learning algorithms, as indicated by the cross-validation variances (**Fig. 2b**).

Ablation analysis shows that Pathformer benefits from multi-modal integration, pathway crosstalk network, attention mechanism, and embedding method

We used ablation analysis to evaluate the essentialities of each type of data and each module of model in the multi-model data integration of Pathformer, based on nine datasets of cancer survival risk classification (**Fig. 3**) and nine datasets of cancer stage classification (**Supplementary Fig. 4**). First, we evaluated the essentialities of seven different data inputs, including RNA expression, DNA methylation, DNA CNV, and a combination thereof. By comparing the performances of seven models on cancer survival risk classification, we discovered that the model with all three modalities as input achieved the best performance, followed by RNA expression and DNA CNV combined model and DNA methylation-only model (**Fig. 3a**). Furthermore, we observed that the performances of models with single modality can vary greatly between datasets. For example, DNA methylation-only model performed better than RNA expression-only and DNA CNV-only in the LUSC, LIHC, and LGG datasets, but the opposite performances were observed in the LUAD and BLCA datasets. Ablation analysis of different modalities on cancer stage classification showed similar results (**Supplementary Fig. 4a**). These findings suggest that different modalities have disparate behaviors in different cancer types, and emphasized the necessity of multi-modal data integration in various cancer classification tasks.

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

Application of Pathformer to cancer prognosis prediction

Pathformer is designed to incorporate prior biological pathway knowledge to uncover the relevant mechanisms behind cancer classification tasks. To further understand the decision-making process of Pathformer and validate

316 the reliability of its biological interpretability, we showed a case study on breast cancer prognosis prediction. We
317 demonstrated that Pathformer can use attention weights and SHAP values to reveal modalities, pathways, and genes
318 related to breast cancer prognosis, which align with molecular features discussed in biomedical literature (**Fig. 4**).
319 First, at the modality level, we visualized the contributions of different modalities for breast cancer prognostic
320 prediction by the attention weights (**Fig. 4a**). The contribution of transcriptomic data was greatest in breast cancer
321 prognostic prediction, which is consistent with the results of ablation analysis (**Fig. 3a**) and findings from other
322 literatures[33,34]. Combining with the results of stage classification using breast cancer tissue multi-modal data
323 (**Supplementary Figs. 5a**), we observed that the contributions of the three modalities in both tasks were similar,
324 but the contributions of different features for DNA methylation in breast cancer prognosis classification were
325 comparable. In other words, the contributions of various features in the same modality were also different for
326 different classification tasks. These findings further validated the necessity of multi-modal integration and biological
327 multi-modal embedding process.
328 Next, at the pathway and gene level, we identified key pathways with top 15 SHAP value and key genes with
329 top 5 SHAP value of each pathway in breast cancer prognostic classification (**Fig. 4b**). Then, we presented a hub
330 module of the updated pathway crosstalk network (**Fig. 4c**). Here, *complex I biogenesis* pathway which participates
331 in cancer cell proliferation and metastasis[35] was identified as the most critical pathway in the classification and a
332 key node in the hub module of the updated pathway crosstalk network. Five mitochondrial genes (MT-ND4, MT-
333 ND1, MT-ND3, MT-ND6, and MT-ND2) were identified as key genes of the *complex I biogenesis* pathway in breast
334 cancer prognostic classification by Pathformer, which is consistent with the literatures[36]. In addition, compared to
335 stage classification, *Anaplastic Lymphoma Kinase (ALK) signaling* pathway is specific to breast cancer prognosis
336 classification. ALK promotes cell proliferation by inhibiting apoptosis[37]. Overexpression, deletion, and copy
337 number variations of ALK are associated with different breast cancer subtypes[38-40], especially triple negative breast
338 cancer with poor prognosis[41]. Furthermore, through the hub module of the updated pathway crosstalk network (**Fig.
339 4c**), we discovered a series of regulatory relationships related to immune response, centered around *FCERI-
340 mediated MAPK activation* pathway[40], which are crucial to prognosis classification. These pathways impact the
341 tumor immune response by influencing the energy supply[42], activation[43], and signal transduction[44] of immune cells.
342 Subsequently, to facilitate a more intuitive understanding of the impact of active pathways identified by
343 Pathformer on breast cancer prognostic classification, we depicted survival curves between patients with high and
344 low scores in active pathways (**Fig. 4d** and **Supplementary Fig. 6**). The pathway score for each sample was
345 obtained by averaging across different dimensions of pathway embedding. Log-rank tests indicated that most active

346 pathways identified by Pathformer, like *complex I biogenesis* pathway, significantly influenced patient survival.

347 Application of Pathformer to the noninvasive diagnosis of cancer

348 Pathformer aims to construct a versatile integration model capable of accommodating various types and quantities
349 of multi-modal inputs. To further validate the versatility of Pathformer, we conducted a case study focusing on
350 blood multi-modal data integration to classify cancer patients from healthy controls for cancer noninvasive diagnosis.
351 We first curated three types of cfRNA-seq datasets derived from three different blood components: plasma,
352 extracellular vesicle (EV), and platelet. We then calculated seven RNA-level modalities as Pathformer's input,
353 including RNA expression, RNA splicing, RNA editing, RNA alternative promoter (RNA alt. promoter), RNA
354 allele-specific expression (RNA ASE), RNA single nucleotide variations (RNA SNV), and chimeric RNA
355 (**Supplementary Note 5**). Because these seven modalities of RNA may have information redundancy, we selected
356 the best modality combination based on 5-fold cross validations (**Supplementary Note 8**). The results showed that
357 the plasma data with 7 modalities, the EV data with 3 modalities, and the platelet data with 3 modalities obtain the
358 best performances (AUCs > 0.9) (**Table 3** and **Supplementary Fig. 8**). Because cancer screen usually requires high
359 specificity, we particularly report sensitivities on 99% specificity in **Table 3**, all of which are above 40%. It is worth
360 noting that the sensitivity is still above 40% on 99% specificity in the plasma data even for the early-stage cancer
361 patients, showing its potential for early cancer diagnosis.

362 Next, we can gain insight into the deregulated alterations in body fluid (i.e., plasma) for cancer patients vs.
363 healthy controls by Pathformer's biological interpretability module (**Fig. 5**). We examined there interesting
364 pathways. The first is *binding and uptake of ligands* (e.g., oxidized low-density lipoprotein, oxLDL) *by scavenger*
365 *receptors* pathway was identified as the most active pathway in both plasma and EV datasets (**Fig. 5a and**
366 **Supplementary Fig. 9a**). It is well established that scavenger receptors play a crucial role in cancer prognosis and
367 carcinogenesis by promoting the degradation of harmful substances and accelerating the immune response[45].
368 Scavenger receptors are also closely related to the transport process of vesicles, such as stabilin-1[46]. Meanwhile,
369 HBB, HBA1, HBA2, FTH1, HSP90AA1 were identified as key genes in this pathway. HBB, as a known biomarker,
370 has been demonstrated to be downregulated in gastric cancer blood transcriptomics[47]. HSP90AA1 is also a potential
371 biomarker for various cancers[48], especially in the blood[49]. The other interesting pathways are *DAP12 signaling*
372 pathway and *DAP12 interactions* pathway revealed in both plasma and platelet datasets (**Fig. 5a and**
373 **Supplementary Fig. 10a**). DAP12 triggers natural killer cell immune responses against certain tumor cells[50], which
374 is regulated by platelet[51]. Among the top 5 key genes of DAP12 related pathway, B2M has been reported as a serum

375 protein encoding gene and a widely recognized tumor biomarker[52]; HLA-E and HLA-B were reported as cancer
376 biomarkers in tissue and plasma[53,54].

377 In addition, Pathformer provides insight into the interplay between various biological processes and their
378 impact on cancer progression by updating pathway crosstalk network (**Fig. 5b**). In the plasma data, the link between
379 *binding and uptake of ligands by scavenger receptors* pathway and *iron uptake and transport* pathway was a novel
380 addition to the updated network. These two pathways lack the crosstalk relationship in the original network before
381 training, but there are literature reports that SCARA5 (member of scavenger receptor class A) can function as a
382 ferritin receptor[55]. The crosstalk between two pathways was amplified by Pathformer in plasma dataset, probably
383 because they were important for classification and shared the same key gene, FTH1 (gene encoding the heavy
384 subunit of ferritin), one of two intersecting genes between the two pathways. In summary, Pathformer's updated
385 pathway crosstalk network visualizes the information flow between pathways related to cancer classification task
386 in the liquid biopsy data, providing novel insight into the crosstalk of biological pathways in cancer patients' plasma.

387 Discussion

388 Pathformer utilizes a novel biological multi-modal embedding method (**Fig. 1b**) based on multi-modal vector and
389 pathway-based sparse neural network, providing the foundation for applying Transformer model on biological
390 multi-modal data integration. Particularly, we showed that the criss-cross attention mechanism (**Supplementary**
391 **Fig. 1**) contributed to classification tasks by capturing crosstalk between biological pathways and potential
392 regulation between modalities (i.e., multi-omics).

393 ***Applications of Pathformer.*** Pathformer can be widely applied to various biomedical classification tasks
394 such as cancer staging, prognosis, and diagnosis. Compared to other integration methods, Pathformer has
395 demonstrated significant improvements, with an average improvement in F1score_macro over 6.3% in cancer
396 survival data and over 5.1% in cancer staging data. Additionally, we use two case studies on breast cancer prognosis
397 and pan-cancer noninvasive diagnosis, demonstrating the prediction performance and biological interpretability of
398 the Pathformer model. For example, the model highlighted the *complex I biogenesis* pathway, which affects the
399 prognosis of breast cancer patients by participating in tumor cell proliferation and metastasis[35]. The Pathformer
400 model for the liquid biopsy data (i.e., cfRNA-seq) revealed the scavenger receptor related pathway and DAP12
401 related pathway, which are associated with extracellular vesicle transport[46] and immune response[50], respectively.

402 ***Limitations of Pathformer and future directions.*** Pathformer used genes involved in pathways from four
403 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[56]. Therefore, incorporating noncoding RNAs and their related functional pathways into Pathformer would be promising for future work. 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), National Key Research Program of China (2021YFA1301603), 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, Beijing Advanced Innovation Center for Structural Biology, Bio-Computing Platform of Tsinghua University Branch of China National Center for Protein Sciences.

Funding for open access charge: Tsinghua University Guoqiang Institute Grant (2021GQG1020).

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

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

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

17 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).

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

19 Shen, H. *et al.* Generative pretraining from large-scale transcriptomes for single-cell deciphering. *Iscience* **26** (2023).

20 Cancer Genome Atlas Research Network, J. The cancer genome atlas pan-cancer analysis project. *Nat. Genet* **45**, 1113-1120 (2013).

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

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

23 Schaefer, C. F. *et al.* PID: the pathway interaction database. *Nucleic acids research* **37**, D674-D679 (2009).
24 Croft, D. *et al.* Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research* **39**, D691-D697 (2010).
25 Nishimura, D. BioCarta. *Biotech Software & Internet Report: The Computer Software Journal for Scient* **2**, 117-120 (2001).
26 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).
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 tumor status-informative signatures in gastrointestinal cancer patients' plasma. *bioRxiv*, 2023.2001. 2031.526431 (2023).
29 Li, S. *et al.* exoRBase: a database of circRNA, lncRNA and mRNA in human blood exosomes. *Nucleic acids research* **46**, D106-D112 (2018).
30 Yu, S. *et al.* Plasma extracellular vesicle long RNA profiling identifies a diagnostic signature for the detection of pancreatic ductal adenocarcinoma. *Gut* **69**, 540-550 (2020).
31 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).
32 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).
33 Huang, Z. *et al.* SALMON: survival analysis learning with multi-omics neural networks on breast cancer. *Frontiers in genetics* **10**, 166 (2019).
34 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).
35 Urra, F. A., Muñoz, F., Lovy, A. & Cárdenas, C. The mitochondrial complex (I) ty of cancer. *Frontiers in oncology* **7**, 118 (2017).
36 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).
37 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).
38 Robertson, F. M. *et al.* Presence of anaplastic lymphoma kinase in inflammatory breast cancer. *Springerplus* **2**, 1-12 (2013).
39 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).
40 Siraj, A. K. *et al.* ALK alteration is a frequent event in aggressive breast cancers. *Breast Cancer Research* **17**, 1-12 (2015).
41 Sachdev, J. C. *et al.* (American Society of Clinical Oncology, 2013).
42 Mills, E. L., Kelly, B. & O'Neill, L. A. Mitochondria are the powerhouses of immunity. *Nature immunology* **18**, 488-498 (2017).
43 Feske, S. Calcium signalling in lymphocyte activation and disease. *Nature Reviews Immunology* **7**, 690-702 (2007).
44 Bhat, M. Y. *et al.* Comprehensive network map of interferon gamma signaling. *Journal of cell communication and signaling* **12**, 745-751 (2018).
45 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).
46 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).

47 Lee, I.-S. *et al.* A blood-based transcriptomic signature for noninvasive diagnosis of gastric cancer. *British Journal of Cancer* **125**, 846-853 (2021).
48 Zuehlke, A. D., Beebe, K., Neckers, L. & Prince, T. Regulation and function of the human HSP90AA1 gene. *Gene* **570**, 8-16 (2015).
49 Zhang, P. j. *et al.* Genes expression profiling of peripheral blood cells of patients with hepatocellular carcinoma. *Cell biology international* **36**, 803-809 (2012).
50 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).
51 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).
52 Cooper, E. & Plesner, T. Beta‐2‐microglobulin review: Its relevance in clinical oncology. *Medical and Pediatric Oncology* **8**, 323-334 (1980).
53 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).
54 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).
55 Yu, B. *et al.* Interactions of ferritin with scavenger receptor class A members. *Journal of Biological Chemistry* **295**, 15727-15741 (2020).
56 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.

a. Schematic of the Pathformer framework and its applications. Pathformer integrates biological multi-modal data based on pathway crosstalk information and Transformer model, which can be applied to different biomedical classification tasks with high accuracy and better biological interpretability. **b.** Schematic of biological multi-modal embedding. Pathformer supports any type and quantity of multi-modal data as input, including single modal data as well. Biological multi-modal consists of gene embedding and pathway embedding. Gene embedding is derived by converting different modalities into gene level modal features and concatenating these into compacted multi-modal vectors. Pathway embedding is generated from gene embedding through a sparse neural network based on gene-to-pathway mappings. F_E, conversion function in the gene embedding, which uses statistical indicators to convert different modalities into gene level; circles, neurons in the neural network; arrows, represent the direction of information flow; G, gene; P, pathway; W, weight of pathway-based sparse neural network.

Figure 2. Performance comparison between Pathformer and 18 other multi-modal integration methods

a. Comprehensive results of Pathformer and other method on cancer low- and high-risk survival classification and cancer early- and late-stage classification. Bar charts show average macro covered F1 score of different methods on all datasets of different classification tasks. Error bars are from 2 times 5-fold cross-validation across all datasets of different classification tasks. **b.** Standard deviation of macro covered F1 score for various multi-modal integration methods. Each data point in each box represents the standard deviation of 2 times 5-fold cross-validation for the given method on a test dataset of different classification tasks. eiLR, eiSVM, eiRF, eiXGBoost, eiNN and eiCNN refer to the early integration methods, while liLR, liSVM, liRF, liXGBoost, liNN and liCNN refers to the late integration methods.

Figure 3. Ablation analysis of Pathformer for the classification of low- and high-risk cancer patients.

a. Different types of data (modalities) were used as input for TCGA cancer low- and high-risk survival classification. **b.** Ablation analysis of different 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.

585

586 **Figure 4. Breast cancer prognostic related modalities, pathways and genes revealed by Pathformer.**

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

595

596 **Figure 5. Interpretation of cancer patients' plasma data using Pathformer.**

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

604

Figure 1. Overview of Pathformer

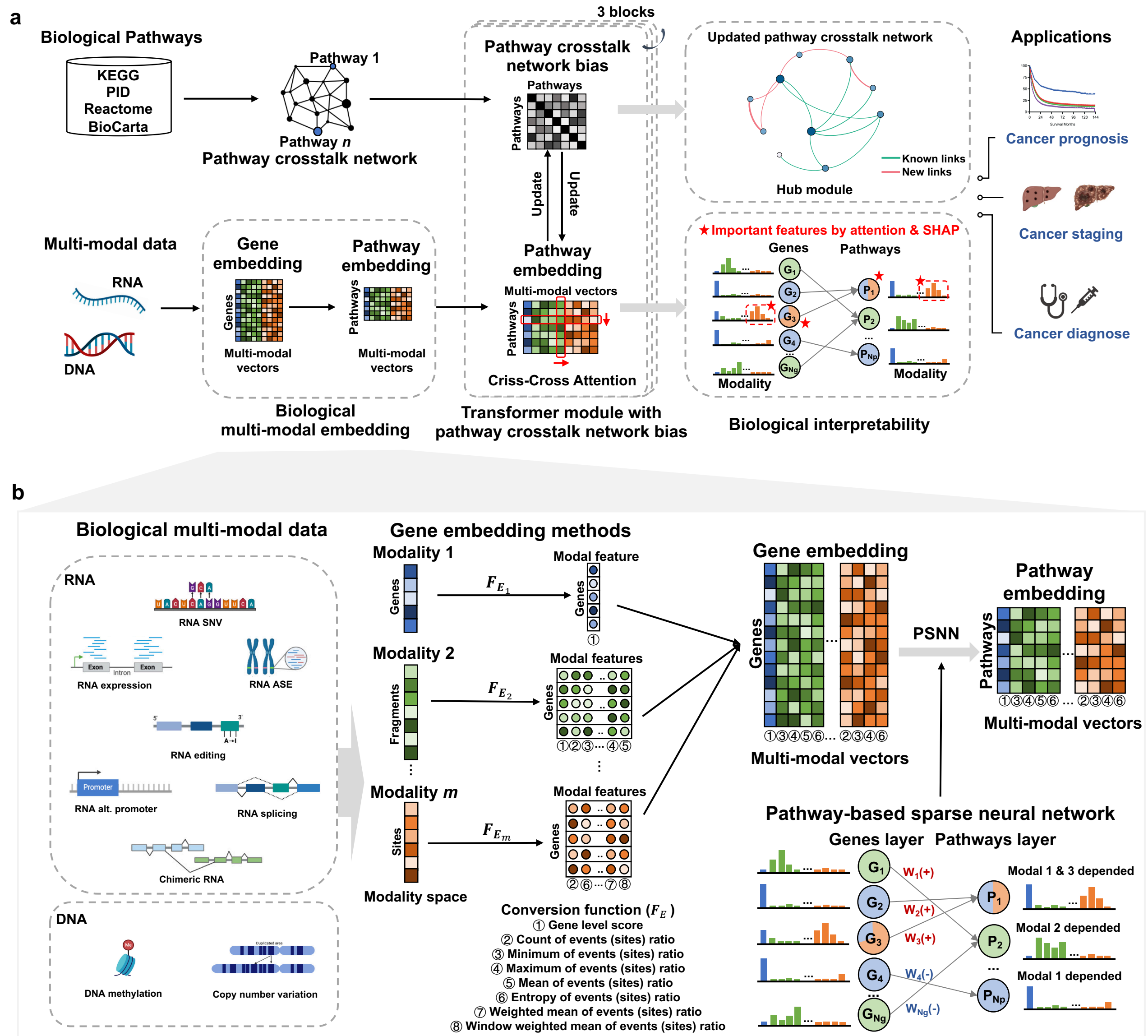

Figure 2. Performance comparison between Pathformer and 18 other multi-modal integration methods

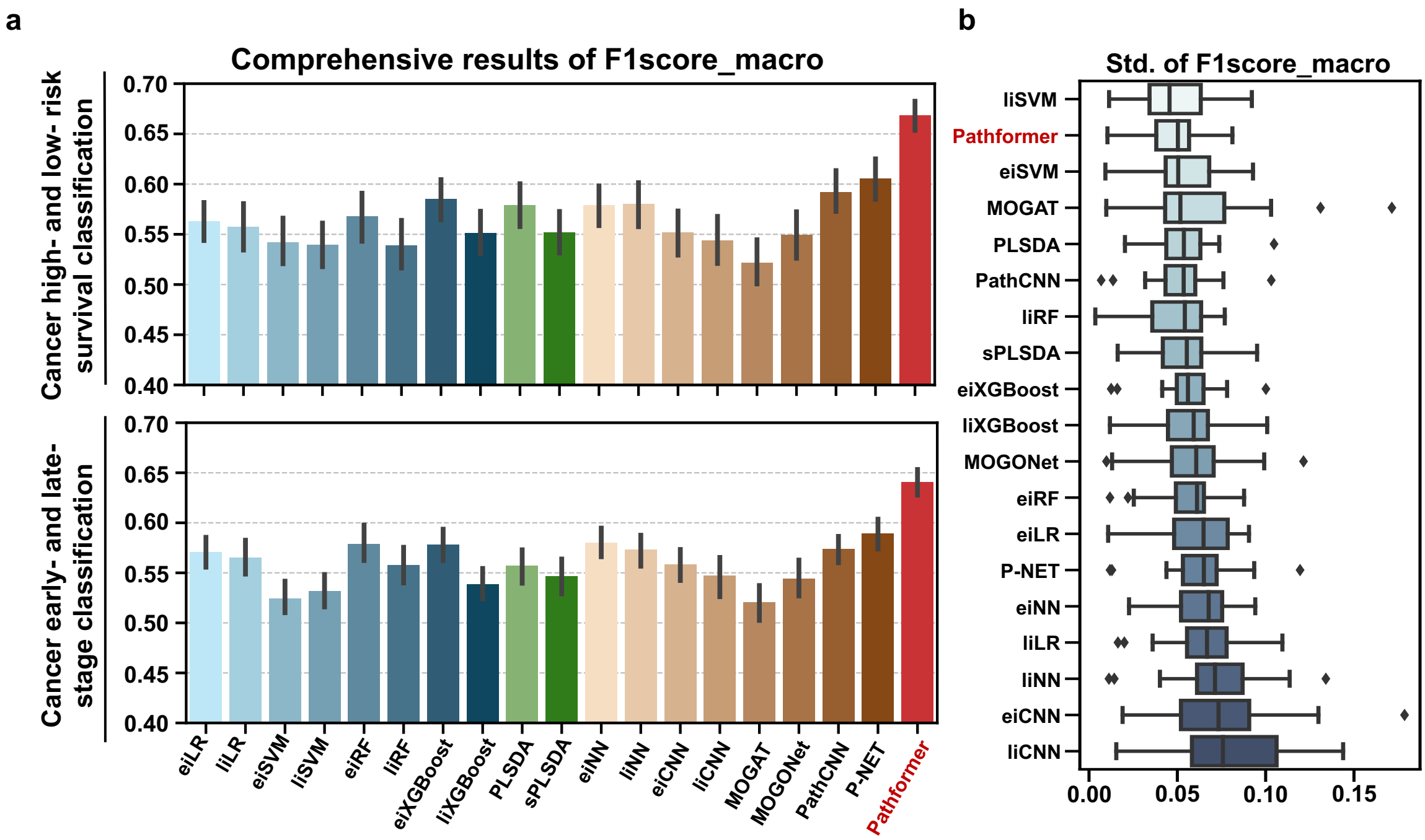

Table 1. Comparison results of Pathformer and 18 other integration models on 10 TCGA datasets of cancer low- and high-risk survival classification in terms of macro covered F1 score.

	BRCA	KIRC	LUAD	LUSC	HNSC	BLCA	LIHC	SKCM	LGG	Pancancer
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**

Table 2. Comparison results of Pathformer and 18 other integration models on 10 TCGA datasets of cancer early- and late-stage classification in terms of macro covered F1 score.

	BRCA	KIRC	LUAD	LUSC	STAD	BLCA	LIHC	SKCM	THCA	Pancancer
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

Figure 3. Ablation analysis of Pathformer for the classification of high- and low- risk cancer patients

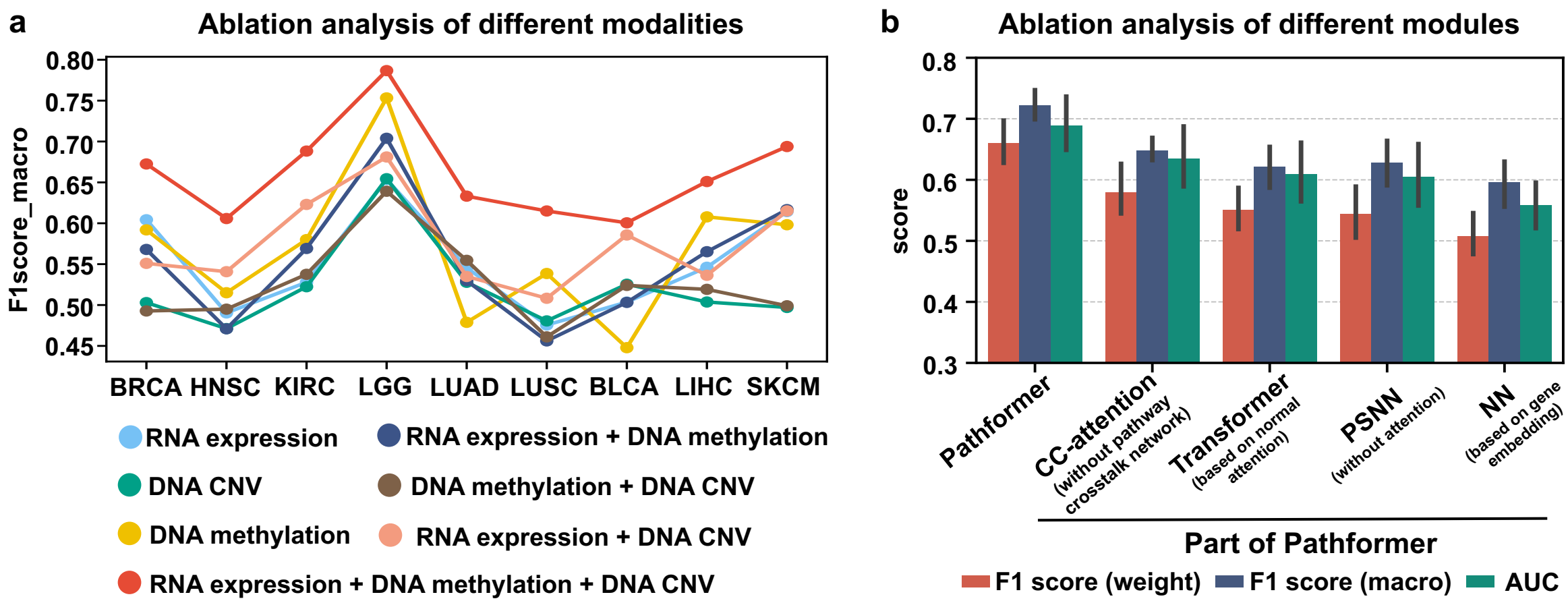

Figure 4. Breast cancer prognostic related modalities, pathways and genes revealed by Pathformer

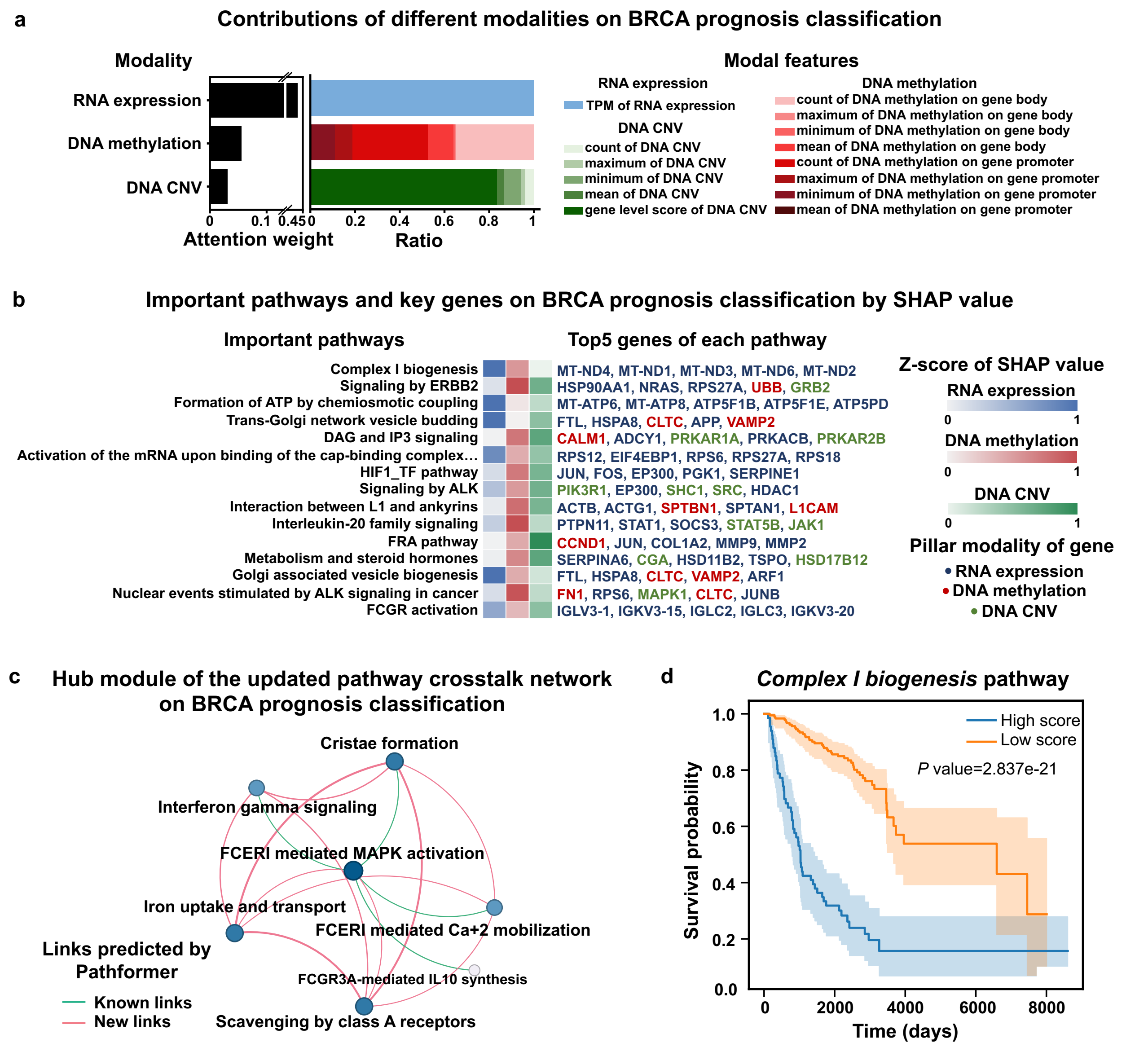

Table 3. Classification performance of three liquid biopsy datasets by Pathformer. Each value is the mean of 5-fold cross-validation.

Dataset	macro covered F1 score	average F1 score weighted by support	AUC	Sensitivity (99% specificity)
Plasma	0.857	0.886	0.914	0.400
Early stage of Plasma	0.849	0.850	0.921	0.425
EV	0.874	0.892	0.940	0.472
Platelet	0.896	0.910	0.943	0.464
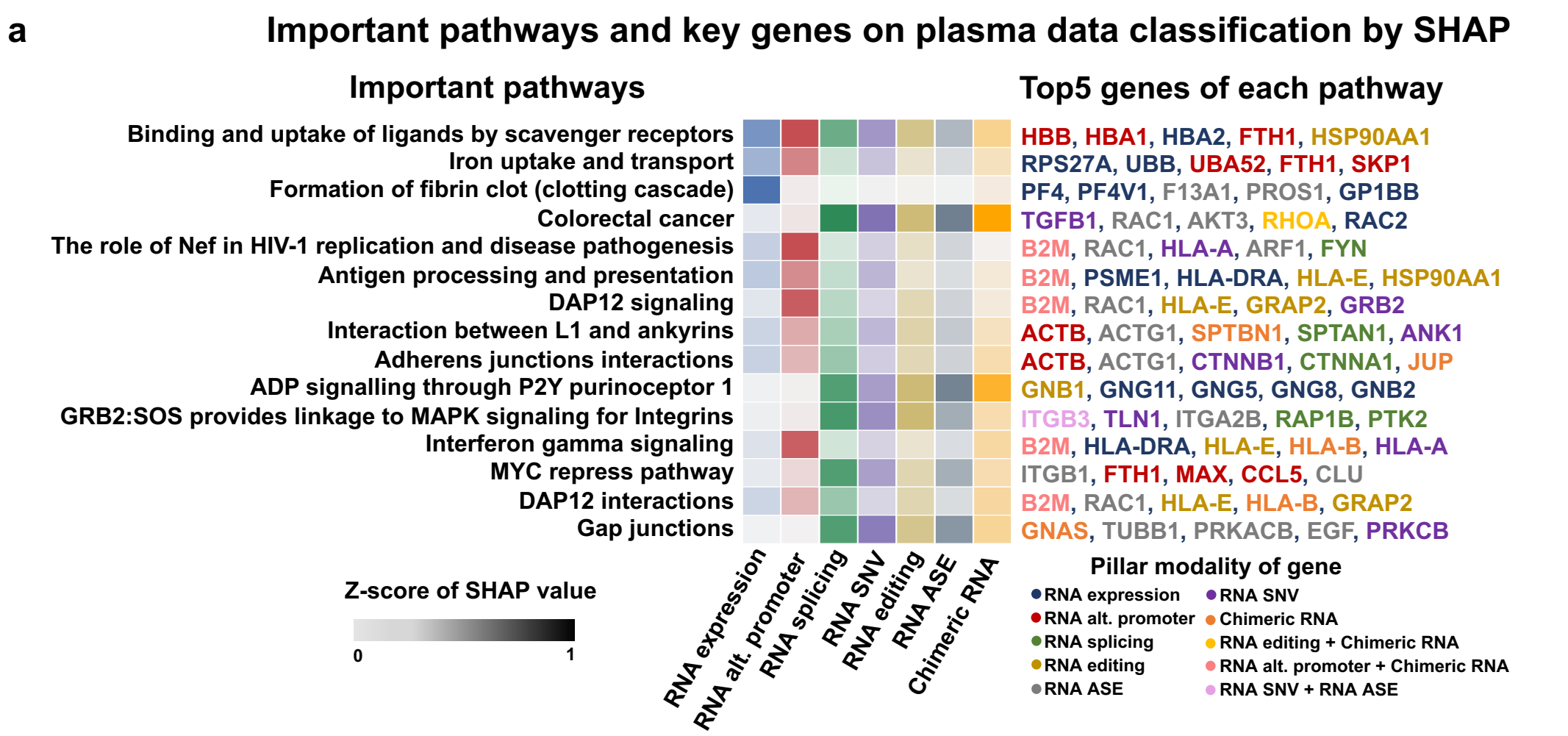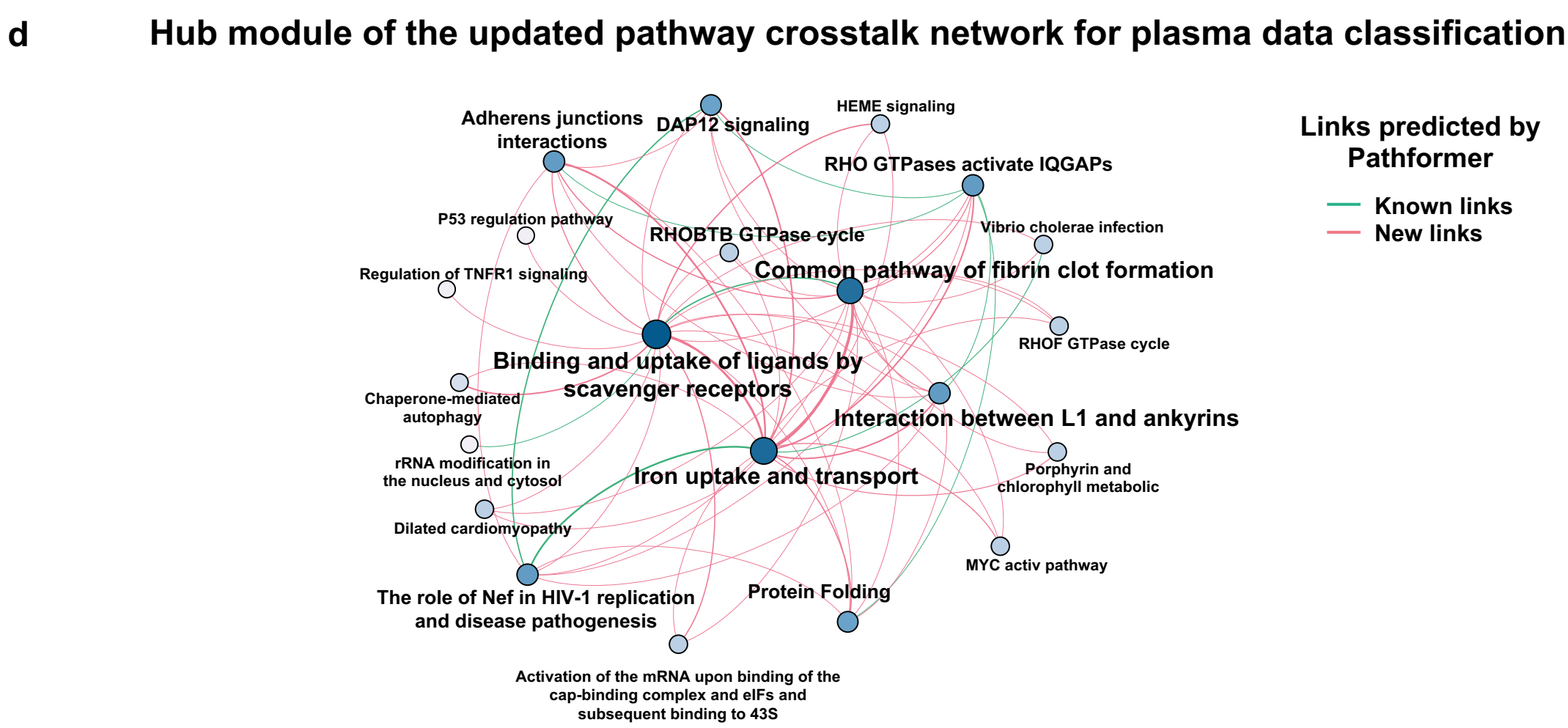
Figure 5. Interpretation of cancer patients' plasma data using Pathformer