Pathformer: biological pathway informed Transformer model integrating multi-modal data

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

Multi-modal biological data can provide comprehensive views of gene regulation and cell development. However, conventional integration methods ignore prior biological knowledge and lack interpretability. Therefore, we developed Pathformer, a biological pathway informed deep learning model based on Transformer with bias to integrate multi-modal data. Pathformer leverages criss-cross attention mechanism to capture crosstalk between different biological pathways and modalities, and utilizes SHapley Additive Explanation method to reveal key pathways, genes, and regulatory mechanisms. When benchmarking on 20 TCGA datasets, Pathformer substantially outperformed other published multi-modal data integration methods in multiple classification tasks. Subsequently, we conducted case studies on breast cancer survival prediction and cancer noninvasive diagnosis, demonstrating Pathformer‘s utilities in biological and clinical applications. Particularly, Pathformer’s biological interpretability helped us reveal interesting cancer-related deregulation pathways in blood (e.g., scavenger receptor pathway and DAP12 signaling transmission pathway) from the liquid biopsy data, which are related to extracellular vesicle transport and immune response.

Keywords: Multi-modal integration; Pathway crosstalk network; Transformer; Interpretability; 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. These platforms can analyze different molecular alterations in the same samples, such as DNA variances (e.g., mutation, methylation, and copy number variance) and RNA alterations (e.g., expression, alternative promoter, splicing, and editing). Integrating these multi-modal data offers a more comprehensive view of gene regulation in diseases than analyzing single type of data[1]. For instance, multi-modal data integration is helpful in addressing certain key challenges of cancer diagnosis and prognosis, such as heterogeneity of intra- and inter-cancer, and complex molecular interactions[2]. Therefore, there is a pressing need for advanced computational methods that uncover interactions of multi-modal data in cancer.

Current algorithms for integrating multi-modal data can be broadly categorized into three groups: early integration models that merge multi-modal data into a single matrix[3,4], late integration models that process each modality separately and then combine their outputs through averaging or maximum voting[5,6], and intermediate integration models that dynamically merge multi-modal data[7,8]. Recently, instead of previous methods that mainly focus on unsupervised problems, several supervised algorithms have been proposed for different biological classification problem. For example, mixOmics uses latent component analysis to find common features among multi-modal data[9]. Wang et al. proposed multi-omics graph convolutional networks (MOGONet), a late integration model that uses graph convolutional networks for modal-specific learning and view correlation discovery network for multi-modal integration[10]. Moon et al. proposed two modal data integration and interpretation algorithm (MOMA) that utilizes attention mechanisms to extract important modules[11]. These methods rely on computational inference to capture relationship between modalities, but ignore the immensely informative prior biological knowledge such as regulatory networks.

To improve the interpretability, several studies have attempted to incorporate prior biological knowledge into deep learning models for multi-modal data integration. For instance, Ma et al. proposed a visible neural network that combines with biological pathways to model the impact of gene interactions on yeast cell growth[12]. Meanwhile, pathway-associated sparse deep neural network (PASNet) was utilized to accurately predict the prognosis of glioblastoma multiforme (GBM) patients[13]. Recently, a sparse neural network integrating multiple molecular features based on a multilevel view of biological pathways, P-net, was published for the classification of prostate cancer patients[14]. Another method, PathCNN, was developed to predict survival of GBM patient by using principal component analysis (PCA) algorithms to define multi-modal pathway images and a convolutional neural network[15]. However, these algorithms rarely considered the synergy and nonlinear relationships between pathways. Given the

complexity of biological systems, understanding the pathway crosstalk is crucial for comprehending more complex biological procession[16], which can help deep learning models better capture multi-modal interactions.

Inspired by these prior works, we propose Pathformer, which combines pathway crosstalk networks and the Transformer encoder with bias for the interpretation and classification of multi-modal data. Recently, Transformer has demonstrated its capability in handling multi-modal tasks in computational fields[17]. At the same time, Transformer has begun to be utilized for processing biological single-modality data[18,19]. However, Transformer hasn't been applied to the biological multi-modal data for lack of reliable biological embedding methods and solutions to the memory explosion posed by the vast amount of gene inputs. These challenges are addressed by Pathformer. First, Pathformer uses multiple statistical indicators of multi-modal data as gene embedding, which comprehensively describes different perspectives of gene information. Second, Pathformer 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 incorporates pathway crosstalk networks into the Transformer model with bias to enhance the exchange of information between different modalities and pathways.

As far as we are aware, Pathformer is the first biological multi-modal integration model that combines prior pathways knowledge and Transformer encoder model. We evaluated Pathformer on 20 benchmark datasets of the Cancer Genome Atlas (TCGA)[20] and demonstrated its superior performance and biological interpretability on various classification tasks, compared to other integration models. In addition, we conducted case studies on cancer staging, survival analysis, and non-invasive diagnosis, utilizing biological interpretability module to reveal 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 stage classification, cancer survival analysis, and cancer diagnosis (**Fig. 1a**). Pathformer utilizes biological pathway network and a Transformer encoder to allow better information fusion and better classification for downstream tasks. 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 uses biological multi-modal data and biological pathway information as input, and define biological multi-modal embedding (gene embedding and 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 to identity important features 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)[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* 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

Multi-modal biological data preprocessing and embedding 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 biological multi-modal embedding method of Pathformer. We represented multi-modal input matrix of a sample as $\boldsymbol{M}$, described as follows:

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

, where m is the number of modalities, and N_g is the number of genes.

Next, we converted the multi-modal input matrix $\boldsymbol{M}$ into gene encoding $\boldsymbol{E}_G$ and pathway encoding $\boldsymbol{E}_P$. We used a series of statistical indicators in different modalities as gene embedding, which are calculated as follows:

$$\boldsymbol{E}_G = \boldsymbol{F}_E(\boldsymbol{M}) = \left[f_{\boldsymbol{E}_1}(\boldsymbol{G}_1),\ f_{\boldsymbol{E}_2}(\boldsymbol{G}_2), \cdots, f_{\boldsymbol{E}_m}(\boldsymbol{G}_m)\right] = \left[\boldsymbol{G}_1^1, \cdots, \boldsymbol{G}_1^{e_1}, \boldsymbol{G}_2^1, \cdots, \boldsymbol{G}_2^{e_2}, \cdots, \boldsymbol{G}_m^1, \cdots, \boldsymbol{G}_m^{e_m}\right] \in \mathbb{R}^{N_g \times D_g}$$

, where e_i is the length of gene embedding for modality i, $\boldsymbol{G}_i^j$ is the jth dimension of gene embedding for modality i, $D_g = e_1 + e_2 + \cdots + e_m$ is the length of gene embedding for all modalities, and $\boldsymbol{F}_E$ is a series of gene embedding functions. $\boldsymbol{F}_E$ uses a series of statistical indicators to uniformly convert the data of different modalities into gene level, and the embedding functions corresponding to different modalities are different (more details in **Supplementary Table 1**). These statistical indicators 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 length 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[25] 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 multi-modal 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 multi-modal embedding and pathway crosstalk network matrix, respectively. First, we represented the *h*th column-wise self-attention by $\boldsymbol{A}_{col}^{(h)}$, calculated as follows:

$$\boldsymbol{A}_1^{(h)} = (\boldsymbol{Q}_1\boldsymbol{K}_1^T)/\sqrt{d}$$

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

, where $h = 1,2,\cdots,H$ is the *h*th head; *H* is the number of heads; $\boldsymbol{Q}_1 = \boldsymbol{E}_P\boldsymbol{W}_{Q_1}^{(h)}$, $\boldsymbol{K}_1 = \boldsymbol{E}_P\boldsymbol{W}_{K_1}^{(h)}$, $\boldsymbol{V}_1 = \boldsymbol{E}_P\boldsymbol{W}_{V_1}^{(h)}$ are linear transformations of the input $\boldsymbol{E}_P$; $\boldsymbol{W}_{Q_1}^{(h)} \in \mathbb{R}^{D_p\times d}$, $\boldsymbol{W}_{K_1}^{(h)} \in \mathbb{R}^{D_p\times d}$, $\boldsymbol{W}_{V_1}^{(h)} \in \mathbb{R}^{D_p\times d}$ 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:

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

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

$$\boldsymbol{U}_1' = \boldsymbol{U}_1 + \boldsymbol{E}_P$$

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

, 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:

$$\boldsymbol{A}_2^{(h)} = (\boldsymbol{Q}_2\boldsymbol{K}_2^T)/\sqrt{d}$$

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

, where $h = 1,2,\cdots,$ h is the *h*th head; *H* is the number of heads; $\boldsymbol{Q}_2 = \boldsymbol{E}_P^T \boldsymbol{W}_{Q_2}^{(h)}$, $\boldsymbol{K}_2 = \boldsymbol{E}_P^T \boldsymbol{W}_{K_2}^{(h)}$, $\boldsymbol{V}_2 = \boldsymbol{E}_P^T \boldsymbol{W}_{V_2}^{(h)}$ are linear transformations of the input $\boldsymbol{E}_P^T$; $\boldsymbol{W}_{Q_2}^{(h)} \in \mathbb{R}^{N_p \times d}$, $\boldsymbol{W}_{K_2}^{(h)} \in \mathbb{R}^{N_p \times d}$, $\boldsymbol{W}_{V_2}^{(h)} \in \mathbb{R}^{N_p \times d}$ 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:

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

$$\boldsymbol{U}_2 = \sum_{h=1}^{H} (\boldsymbol{g}_2^{(h)} \circ \boldsymbol{A}_{row}^{(h)}) \cdot \boldsymbol{W}_{U_2}^{(h)}$$

$$\boldsymbol{U}_2' = \boldsymbol{\beta} * \boldsymbol{U}_2 + \boldsymbol{E}_P^T$$

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

, 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 used averaging attention maps in row-attention to represent the contributions of different modalities, and SHapley Additive exPlanations[26] (SHAP value) to decipher the important pathways and their key genes. Finally, the hub module of the updated pathway crosstalk network represents the most critical regulatory mechanism in classification.

Contribution of each modality

In Pathformer, row-attention is used to facilitate information interaction between different modalities, that is, row-attention map can represent the importance of each modality. According to the trained model, we obtained row-attention maps of 8 heads in 3 blocks for each sample. For the 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, i.e., the contribution of modalities. The calculation is as follows:

$$\boldsymbol{A}_{aver} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{BL}\sum_{b=1}^{BL}\frac{1}{H}\sum_{h=1}^{H} softmax([[\boldsymbol{A}_2^{(h)}]^{(b)}]^{(n)})$$

$$attention\ weight_i = \frac{1}{D_p}\sum_{j=1}^{D_p} a_{ij},\ a_{ij} \text{ is the } i\text{th row and the } j\text{th columns of } \boldsymbol{A}_{aver}$$

, where N is the number of samples, BL is the number of blocks, H is the number of heads, softmax is a normalized exponential function, and $attention\ weight_i$ is the attention weight of dimension i of pathway embedding.

Important pathways and their key genes

SHapley Additive exPlanations[26] (SHAP) 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 pathways, genes and classification, implemented by "SHAP" package of *Python* v3.6.9.

Specifically, we calculated SHAP values of the gene embedding and the pathway embedding encoded by Transformer module corresponding to each sample and each category, denoted as $\boldsymbol{S}_{gn}^{(j)} \in \mathbb{R}^{D_p}$ and $\boldsymbol{S}_{pn}^{(j)} \in \mathbb{R}^{D_p}$ respectively. The SHAP values of genes and pathways are calculated as follows:

$$\text{SHAP}_g = \sum_{j=1}^{d_{out}}\sum_{e=1}^{D_p}\frac{1}{N}\sum_{n=1}^{N}\left|s_{gne}^{(j)}\right|, s_{gie}^{(j)} \in \boldsymbol{S}_{gi}^{(j)}$$

$$\mathrm{SHAP}_p = \sum_{j=1}^{d_{out}} \sum_{e=1}^{D_p} \frac{1}{N} \sum_{n=1}^{N} \left| s_{pne}^{(j)} \right|, s_{pie}^{(j)} \in \boldsymbol{S}_{pi}^{(j)}$$

, where $g = 1,2,\cdots,N_g$ is the gth gene, $g = 1,2,\cdots,N_p$ is the pth pathway, $n = 1,2,\cdots,N$ is the nth sample, $e = 1,2,\cdots,D_p$ is dimension e of pathway embedding, and $j = 1,2,\cdots,d_{out}$ is the jth category of sample.

In addition, the z-score of SHAP values of different modalities for each pathway and gene can demonstrate modal complementarity at the gene level and the pathway level, described as follows:

$$\mathrm{SHAP}_{gi} = \sum_{j=1}^{d_{out}} \sum_{e=e_1+\cdots+e_{i-1}}^{e_i} \frac{1}{N} \sum_{n=1}^{N} \left| s_{gne}^{(j)} \right|, s_{gie}^{(j)} \in \boldsymbol{S}_{gi}^{(j)}$$

$$\mathrm{SHAP}_{pi} = \sum_{j=1}^{d_{out}} \sum_{e=e_1+\cdots+e_{i-1}}^{e_i} \frac{1}{N} \sum_{n=1}^{N} \left| s_{pne}^{(j)} \right|, s_{pie}^{(j)} \in \boldsymbol{S}_{pi}^{(j)}$$

, where $i = 1,\cdots,m$ is the ith modality, e_i is the length of gene embedding and pathway embedding for modality i.

Finally, 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.

Hub module of the updated pathway crosstalk network

In Pathformer, pathway crosstalk network matrix is used to guide the direction of information flow, and updated according to encoded 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 cancer early- and late-stage classification, and 10 datasets for cancer low- and high-risk survival classification. Besides, to further verify the effect of Pathformer in 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 on the validation set (**Supplementary Fig. 2**).

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

Performance evaluation

We compared the classification performance of Pathformer with several existing multi-modal integration methods, including early integration methods based on base classifiers, i.e., nearest neighbor algorithm (KNN), support vector machine (SVM), logistic regression (LR), random forest (RF), and extreme gradient boosting (XGBoost); late integration methods based on KNN, SVM, LR, RF, and XGBoost; partial least squares-discriminant analysis (PLSDA) and sparse partial least squares-discriminant analysis (sPLSDA) of mixOmics[9]; two deep learning-based integration methods, MOGONet[10] and PathCNN[15]. MOGONet is a multi-modal integration method based on graph convolutional neural network. PathCNN is a representative multi-modal integration method that combines pathway

information. During comparison methods, the multi-modal data were preprocessed with the statistical indicators and features were prefiltered with ANOVA as input (see **Supplementary Note 7**).

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

Pathformer outperforms existing multi-modal integration methods in various classification tasks using benchmark datasets

To evaluate the performance of Pathformer, we tested model on cancer early- and late-stage classification and low- and high-risk survival classification as benchmark studies. For these tasks, DNA methylation, DNA copy nunumber variation (CNV), and RNA expression were used as input. For 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**.

Figure 2 shows the comprehensive comparison results of Pathformer and other integration methods on all benchmark datasets. In general, Pathformer significantly performed better than five representative integration methods (**Fig. 2a**). In cancer low- and high- risk survival classification tasks, comparing to XGBoost, which performed best in the traditional machine learning methods (i.e., KNN, RF, SVM, KNN (late), RF (late), SVM (late), XGBoost (late), PLSDA and sPLSDA), Pathformer's F1score_marco showed an average improvement of 8.3%. When comparing to the other deep learning methods (i.e., MOGONet and PathCNN), Pathformer's F1score_marco showed average improvements between 8% and 12%. In early- and late- stage classification tasks, compared to XGBoost, Pathformer's F1score_marco showed an average improvement of 6.2%. Compared to the other deep learning methods, Pathformer's F1score_marco showed average improvements between 7% and 10%. Then, **Table 1** and **Table 2** respectively illustrate the comparison of F1score_macro for all methods in survival risk classification and stage classification tasks. Other evaluation indicators are shown in **Supplementary Table 3**. In both survival risk classification and stage classification, Pathformer achieved the best F1score_macro and F1score_weighted in all the 10 datasets, and the best AUC in 8 of 10 datasets. These results highlight Pathformer's exceptional learning ability. Moreover, in terms of stability, Pathformer also showed significantly better generalization ability than the

310 other deep learning algorithms, as indicated by the cross-validation variances (**Fig. 2b**).

311 Ablation analysis shows that Pathformer benefits from multi-modal integration, attention 312 mechanism and pathway crosstalk network

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

326 Next, we also evaluated the essentialities of different modules in Pathformer. We developed 4 models, namely
327 CC-attention, Transformer, PSNN, and NN, which successively remove one to multiple modules of Pathformer.
328 CC-attention is a model without pathway crosstalk network bias. Transformer is a model without either pathway
329 crosstalk network bias or row-attention. PSNN is a model that directly uses classification module with pathway
330 embedding as input. NN is a model that directly uses classification module with gene embedding as input. As shown
331 in **Fig. 3b** and **Supplementary Fig. 4b**, the complete Pathformer model achieved the best classification performance,
332 while the performance of CC-Attention, Transformer, PSNN, and NN decreased successively. Transformer had a
333 significantly lower classification performance compared to CC-Attention, but no significant improvement compared
334 to PSNN. This indicates that the criss-cross attention mechanism (**Supplementary Fig. 1**) plays a key role in
335 Pathformer, with respect to information fusion and crosstalk between different biological pathways and between
336 different modalities.

Application of Pathformer to the prognostic prediction of cancer

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 the reliability of its biological interpretability, we show a case study on breast cancer prognosis prediction. We demonstrated that Pathformer can use attention weights and SHAP values to reveal modalities, pathways, and genes related to breast cancer prognosis, which align with molecular features discussed in biomedical literature (**Fig. 4**).

First, at the modality level, we visualized the contributions of different modalities for breast cancer prognostic prediction by the attention weights (**Fig. 4a**). The contribution of transcriptomic data was greatest in breast cancer prognostic prediction, which is consistent with the results of ablation analysis (**Fig. 3a**) and findings from other literatures[33,34]. Combining with the results of stage classification using breast cancer tissue multi-modal data (**Supplementary Figs. 5a**), we observed that the contributions of the three modalities in both tasks were similar, but the contributions of different statistical indicators for DNA methylation in breast cancer prognosis classification were comparable. In other words, the contributions of various statistical indicators in the same modality were also different for different classification tasks. These findings further validated the necessity of multi-modal integration and biological multi-modal embedding.

Next, at the pathway and gene level, we identified key pathways with top 15 SHAP value and key genes with top 5 SHAP value of each pathway in breast cancer prognostic classification (**Fig. 4b**). Then, we presented a hub module of the updated pathway crosstalk network (**Fig. 4c**). Here, *complex I biogenesis* pathway which participates in cancer cell proliferation and metastasis[35] 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[36]. 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[37]. Overexpression, deletion, and copy number variations of ALK are associated with different breast cancer subtypes[38-40], especially triple negative breast cancer with poor prognosis[41]. 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[40], which are crucial to prognosis classification. These pathways impact the tumor immune response by influencing the energy supply[42], activation[43], and signal transduction[44] of immune cells.

Subsequently, to facilitate a more intuitive understanding of the impact of active pathways identified by

367 Pathformer on breast cancer prognostic classification, we depicted survival curves between patients with high and
368 low scores in active pathways (**Fig. 4d** and **Supplementary Fig. 6**). The pathway score for each sample was
369 obtained by averaging across different dimensions of pathway embedding. Log-rank tests indicated that most active
370 pathways identified by Pathformer, like *complex I biogenesis* pathway, significantly influenced patient survival.

371 Application of Pathformer to the noninvasive diagnosis of cancer

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

386 Next, we can gain insight into the deregulated alterations in body fluid (i.e., plasma) for cancer patients vs.
387 healthy controls by Pathformer's biological interpretability module (**Fig. 5**). We examined there interesting pathways.
388 The first is *binding and uptake of ligands* (e.g., oxidized low-density lipoprotein, oxLDL) *by scavenger receptors*
389 pathway was identified as the most active pathway in both plasma and EV datasets (**Fig. 5a and Supplementary**
390 **Fig. 9a**). It is well established that scavenger receptors play a crucial role in cancer prognosis and carcinogenesis
391 by promoting the degradation of harmful substances and accelerating the immune response[45]. Scavenger receptors
392 are also closely related to the transport process of vesicles, such as stabilin-1[46]. Meanwhile, HBB, HBA1, HBA2,
393 FTH1, HSP90AA1 were identified as key genes in this pathway. HBB, as a known biomarker, has been
394 demonstrated to be downregulated in gastric cancer blood transcriptomics[47]. HSP90AA1 is also a potential
395 biomarker for various cancers[48], especially in the blood[49]. The other interesting pathways are *DAP12 signaling*

396 pathway and *DAP12 interactions* pathway revealed in both plasma and platelet datasets (**Fig. 5a and**
397 **Supplementary Fig. 10a**). DAP12 triggers natural killer cell immune responses against certain tumor cells[50], which
398 is regulated by platelet[51]. Among the top 5 key genes of DAP12 related pathway, B2M was reported as a serum
399 protein encoding gene and a widely recognized tumor biomarker[52]; HLA-E and HLA-B were reported as cancer
400 biomarkers in tissue and plasma[53,54].

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

411 Discussion

412 Pathformer utilizes a biological multi-modal embedding (**Fig. 1b**) based on pathway-based sparse neural network,
413 providing a demonstration of applying Transformer model on biological multi-modal data integration. Particularly,
414 we showed that the criss-cross attention mechanism (**Supplementary Fig. 1**) contributed to classification tasks by
415 capturing crosstalk between biological pathways and potential regulation between modalities (i.e., multi-omics).

416 ***Applications of Pathformer.*** Pathformer can be widely applied to various biomedical classification tasks
417 such as cancer staging, prognosis, and diagnosis. Compared to the other integration models, Pathformer has
418 demonstrated significant improvements, with an average increase in F1score_macro over 8% in cancer survival data
419 and over 6% in cancer staging data. Additionally, we use two case studies on breast cancer prognosis and pan-cancer
420 noninvasive diagnosis, demonstrating the prediction performance and biological interpretability of the Pathformer
421 model. For example, the model highlighted the *complex I biogenesis* pathway, which affects the prognosis of breast
422 cancer patients by participating in tumor cell proliferation and metastasis[35]. The Pathformer model for the liquid
423 biopsy data (i.e., cfRNA-seq) revealed the scavenger receptor related pathway and DAP12 related pathway, which
424 are associated with extracellular vesicle transport[46] and immune response[50], respectively.

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[56]. Therefore, incorporating noncoding RNAs and their related functional pathways into Pathformer would be a potential 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 Shen, R., Olshen, A. B. & Ladanyi, M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **25**, 2906-2912 (2009).

4 Lando, M. *et al.* Gene dosage, expression, and ontology analysis identifies driver genes in the carcinogenesis and chemoradioresistance of cervical cancer. *PLoS genetics* **5**, e1000719 (2009).

5 Cabassi, A. & Kirk, P. D. Multiple kernel learning for integrative consensus clustering of omic datasets. *Bioinformatics* **36**, 4789-4796 (2020).

6 Wang, T. *et al.* MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nature Communications* **12**, 1-13 (2021).

7 Wang, B. *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nature methods* **11**, 333-337 (2014).

8 Lock, E. F., Hoadley, K. A., Marron, J. S. & Nobel, A. B. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The annals of applied statistics* **7**, 523 (2013).

9 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).

10 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

11 Moon, S. & Lee, H. MOMA: A Multi-Task Attention Learning Algorithm for Multi-Omics Data Interpretation and Classification. *Bioinformatics* (2022). https://doi.org:10.1093/bioinformatics/btac080

12 Ma, J. *et al.* Using deep learning to model the hierarchical structure and function of a cell. *Nature methods* **15**, 290-298 (2018).

13 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

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

15 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

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

17 Hu, R. & Singh, A. in *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 1439-1449.

18 De Waele, G., Clauwaert, J., Menschaert, G. & Waegeman, W. CpG Transformer for imputation of single-cell methylomes. *Bioinformatics* (2021). https://doi.org:10.1093/bioinformatics/btab746

19 Yang, F. *et al.* scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nature Machine Intelligence* **4**, 852-866 (2022).

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

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 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).
26 Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **30** (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, which can be applied to different biological classification problem and can identity key genes, pathways, modalities, and regulatory processes for better biological interpretability. **b.** Schematic of biological multi-modal embedding. Biological multi-modal data has different levels, which can be different omics or different variant features of the same omics. Pathformer supports any type and quantity of multi-modal data as input, including single modal data as well. F_E, statistical indicators in the gene embedding; 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. The weights of the pathway-based sparse neural network represent the importance of different genes in different pathways.

Figure 2. Performance comparison between Pathformer and 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 classification performance 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. XGBoost refers to the early integration methods, while XGBoost (late) 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, Pathformer without either pathway crosstalk network bias or row-attention; PSNN, Pathformer without Transformer module; NN, classification module only.

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

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

622

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

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

631

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

	BRCA	KIRC	LUAD	LUSC	HNSC	BLCA	LIHC	SKCM	LGG	Pancancer
KNN	0.498	0.624	0.508	0.503	0.464	0.453	0.496	0.599	0.634	0.681
LR	0.571	0.584	0.530	0.520	0.529	0.479	0.497	0.557	0.689	0.674
RF	0.494	0.636	0.443	0.477	0.497	0.475	0.579	0.609	0.759	0.709
SVM	0.467	0.656	0.487	0.471	0.466	0.459	0.436	0.630	0.640	0.705
XGBoost	0.564	0.653	0.530	0.519	0.476	0.549	0.550	0.594	0.723	0.694
KNN (late)	0.422	0.610	0.426	0.403	0.440	0.439	0.441	0.556	0.726	0.701
LR (late)	0.569	0.619	0.479	0.481	0.512	0.456	0.487	0.554	0.718	0.696
RF (late)	0.473	0.622	0.453	0.452	0.478	0.438	0.428	0.609	0.724	0.712
SVM (late)	0.468	0.614	0.438	0.508	0.470	0.475	0.439	0.561	0.714	0.710
XGBoost (late)	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
MOGONet	0.510	0.632	0.455	0.497	0.469	0.470	0.476	0.599	0.706	0.676
PathCNN	0.558	0.643	0.522	0.560	0.525	0.486	0.598	0.551	0.766	0.712
Pathformer	**0.673**	**0.688**	**0.633**	**0.615**	**0.606**	**0.601**	**0.651**	**0.694**	**0.787**	**0.735**

634
635

Table 2. Comparison results of Pathformer and other 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
KNN	0.473	0.646	0.487	0.457	0.516	0.619	0.504	0.556	0.537	0.567
LR	0.510	0.647	0.506	0.501	0.541	0.612	0.567	0.579	0.613	0.631
RF	0.488	0.723	0.503	0.509	0.540	0.647	0.528	0.571	0.660	0.622
SVM	0.451	0.661	0.460	0.467	0.528	0.555	0.464	0.551	0.550	0.557
XGBoost	0.518	0.686	0.526	0.498	0.548	0.639	0.540	0.535	0.664	0.629
KNN (late)	0.427	0.631	0.442	0.455	0.521	0.533	0.437	0.525	0.547	0.594
LR (late)	0.522	0.654	0.473	0.506	0.581	0.624	0.496	0.548	0.626	0.620
RF (late)	0.454	0.704	0.497	0.469	0.543	0.611	0.507	0.570	0.622	0.604
SVM (late)	0.436	0.675	0.460	0.470	0.552	0.564	0.438	0.557	0.601	0.567
XGBoost (late)	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
MOGONet	0.466	0.646	0.461	0.476	0.574	0.553	0.474	0.526	0.644	0.624
PathCNN	0.525	0.694	0.554	0.534	0.521	0.566	0.563	0.540	0.595	0.643
Pathformer	**0.573**	**0.726**	**0.629**	**0.564**	**0.596**	**0.706**	**0.616**	**0.646**	**0.690**	**0.657**

640 **Table 3. Classification performance of three liquid biopsy datasets by Pathformer. Each value is the mean of**
641 **5-fold cross-validation.**

Dataset	AUC	average F1 score weighted by support	macro covered F1 score	Sensitivity (99% specificity)
Plasma	0.914	0.886	0.857	0.400
Early stage of Plasma	0.921	0.850	0.849	0.425
EV	0.940	0.892	0.874	0.472
Platelet	0.943	0.910	0.896	0.464

642

Figure 1. Overview of Pathformer

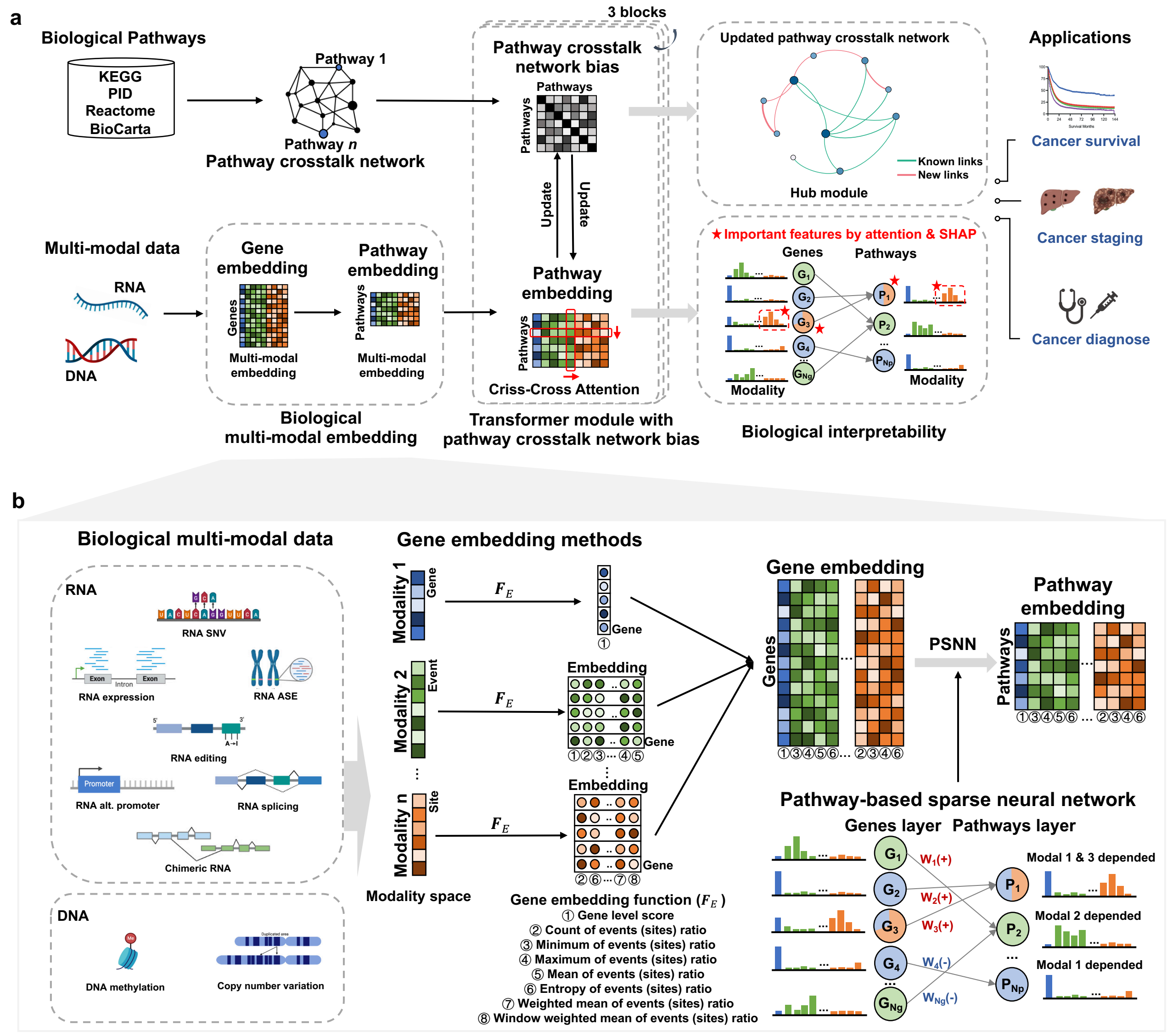

Figure 2. Performance comparison between Pathformer and other multi-modal integration methods

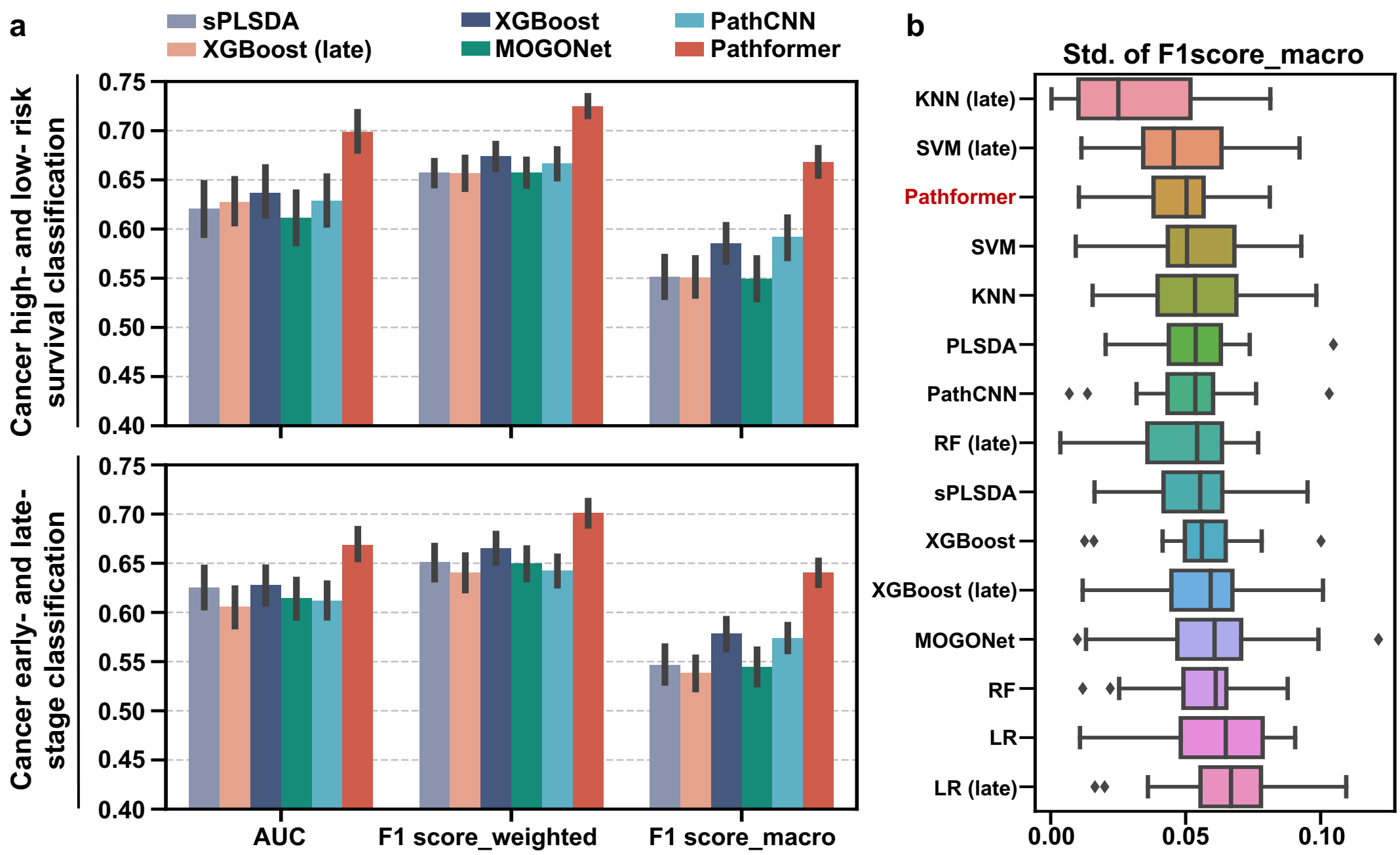

Figure 3. Ablation analysis of Pathformer for the classification of high- and low- risk cancer patients

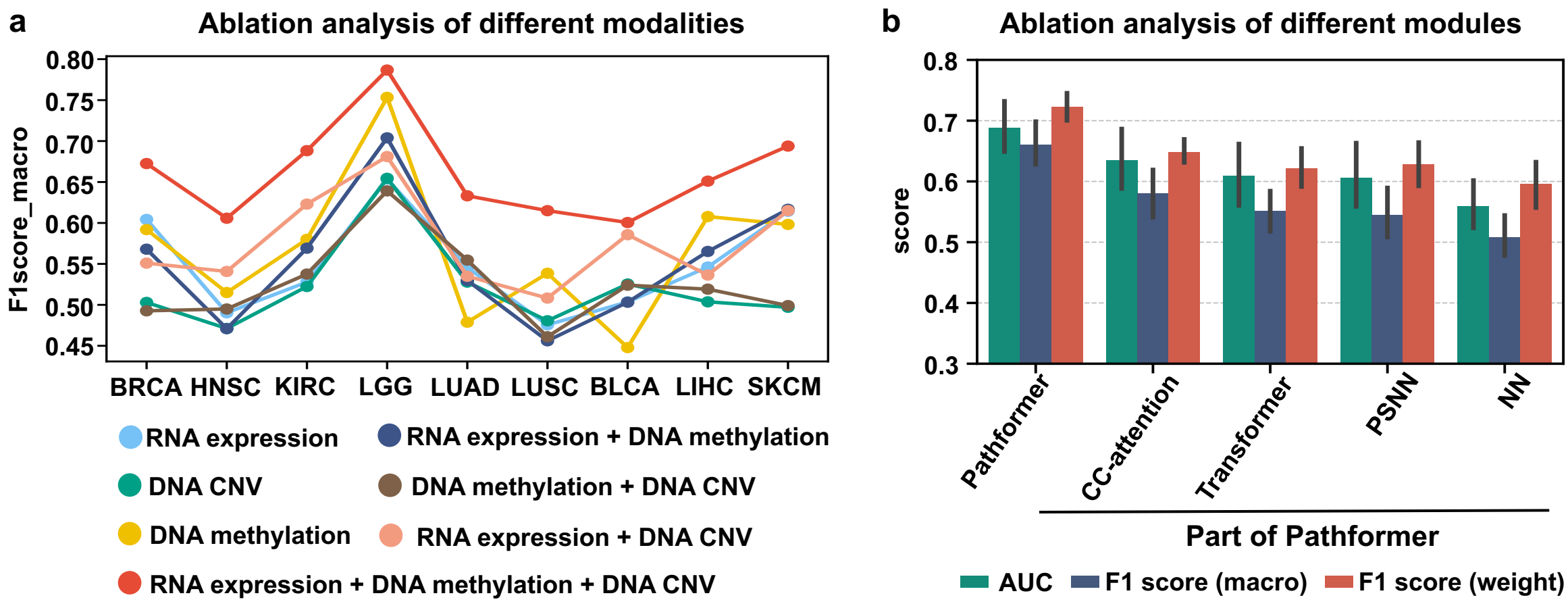

Figure 4. Breast cancer prognostic related modalities, pathways and genes revealed by Pathformer

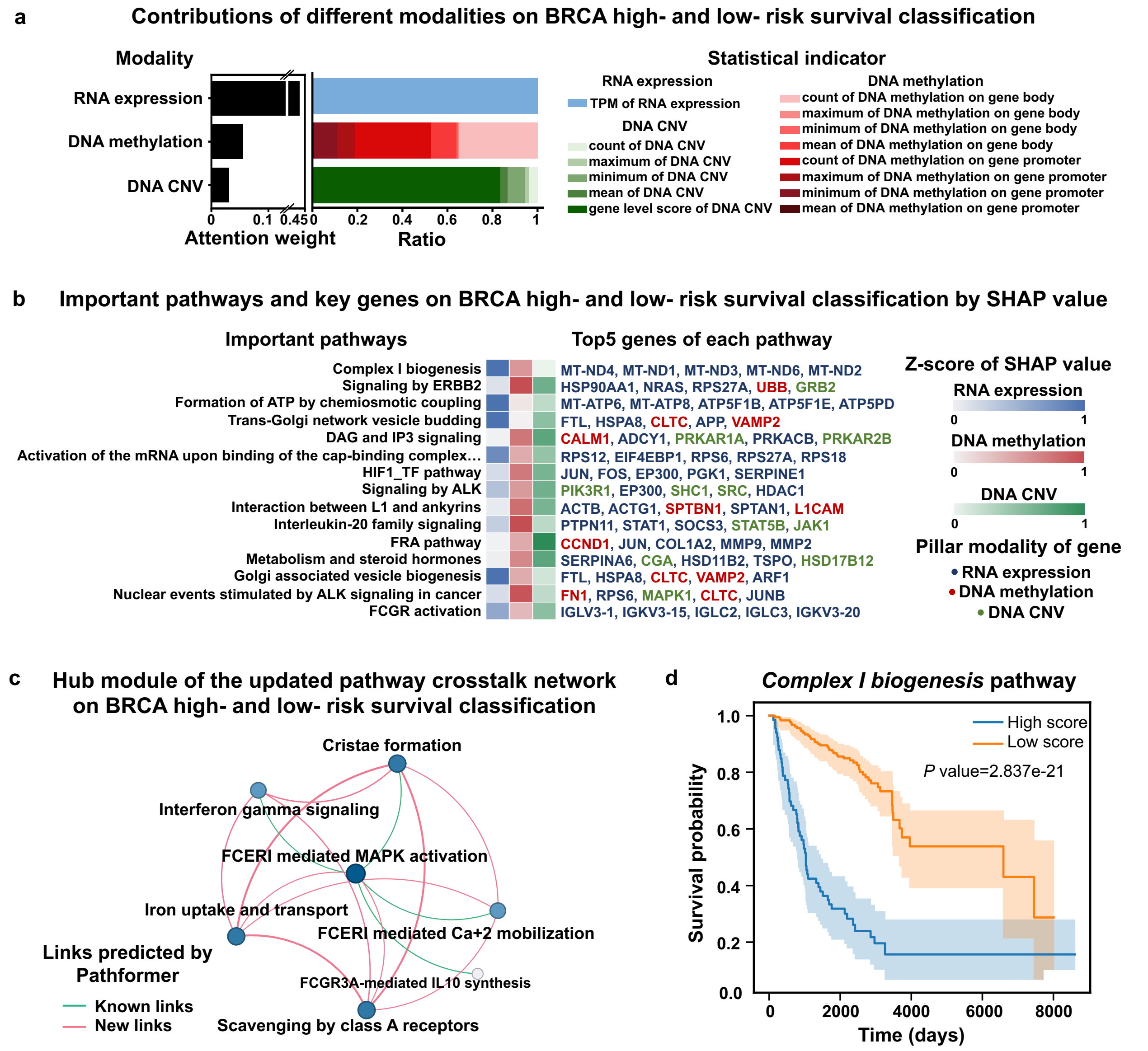
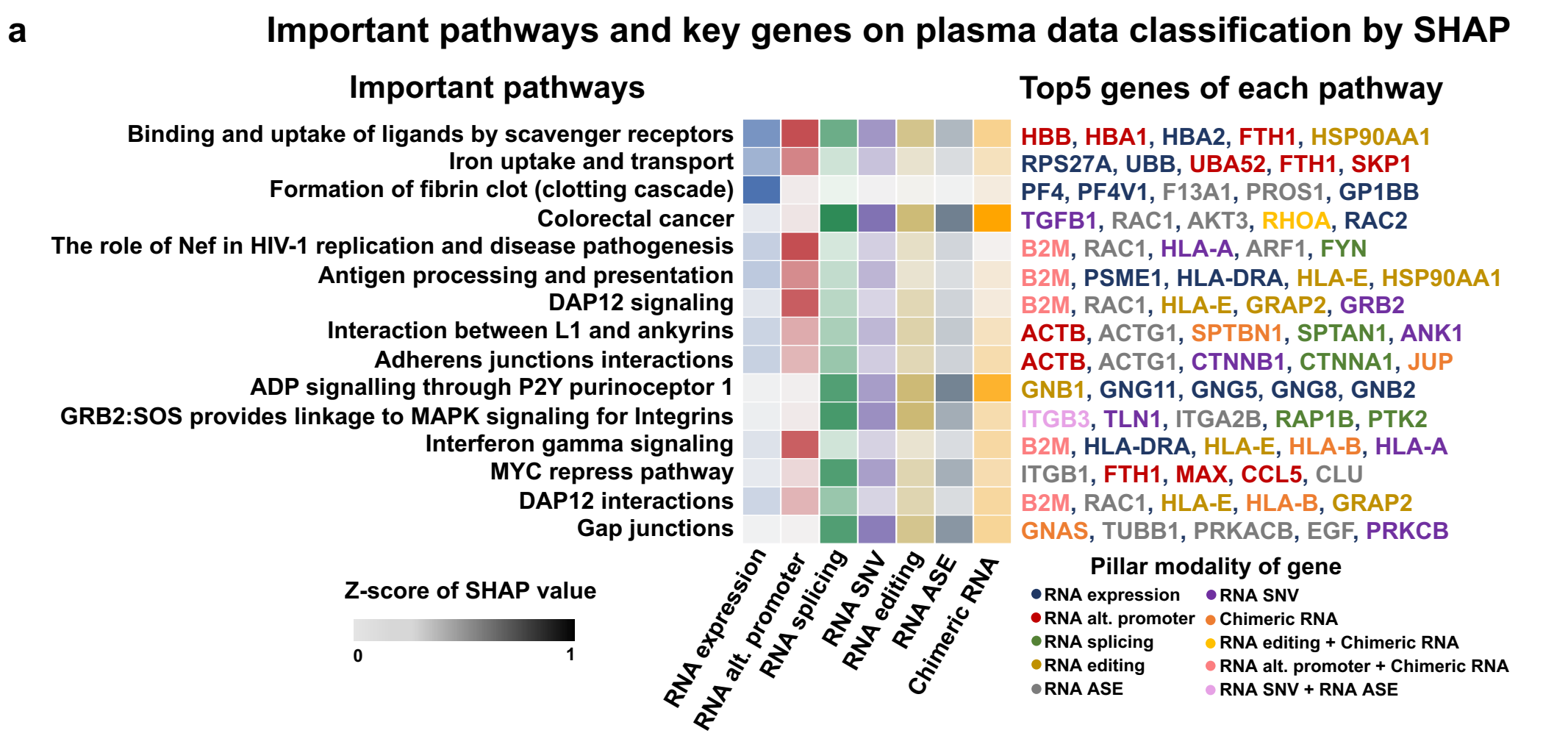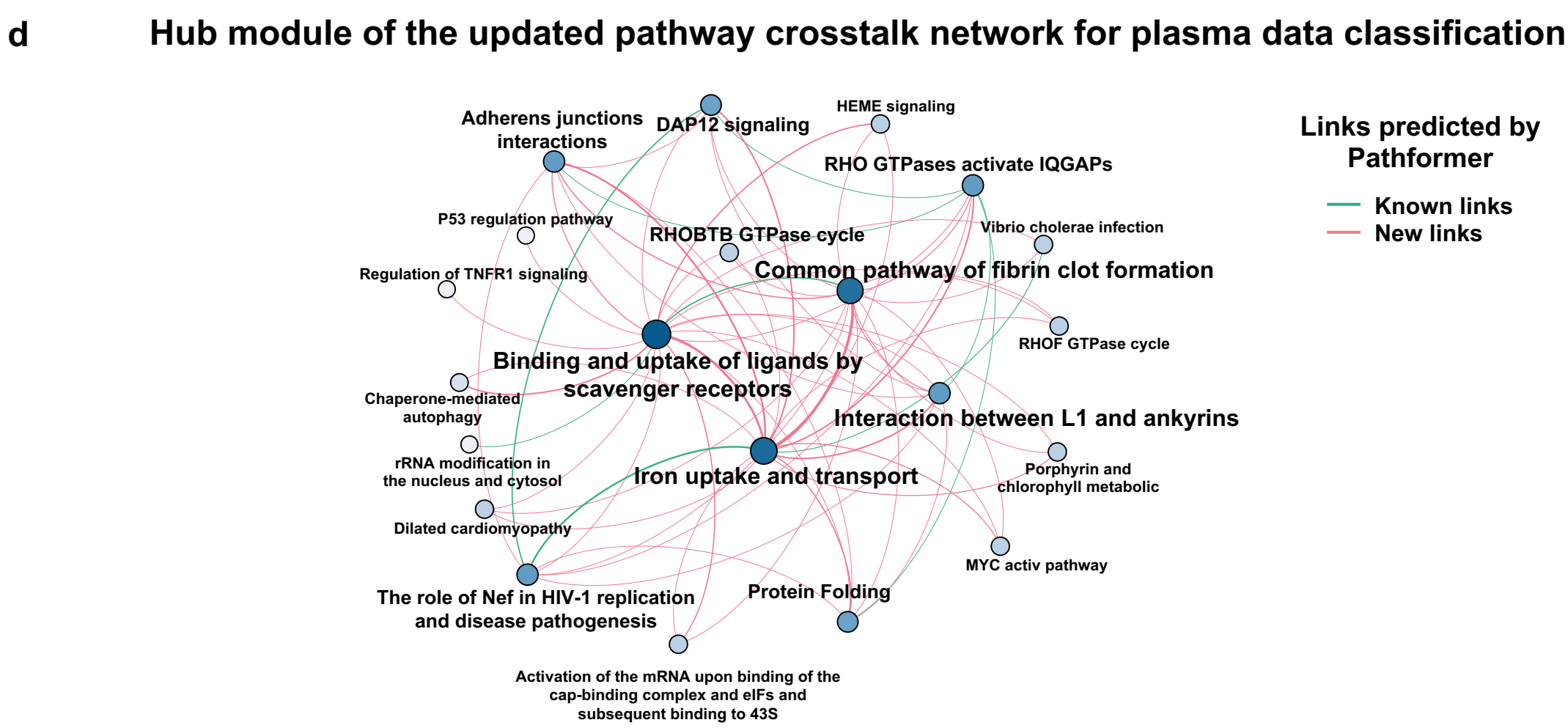
Figure 5. Interpretation of cancer patients' plasma data using Pathformer