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

Xiaofan Liu[1,2], Yuhuan Tao[1,2], Zilin Cai[1], Pengfei Bao[1,2], Hongli Ma[1,2], Kexing Li[1], Yunping Zhu[3,] *, 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]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; Yunping Zhu, Tel.: +86 10 61777058, E-mail: zhuyunping@ncpsb.org.cn.

Abstract

Multi-modal biological data integration can provide comprehensive views of gene regulation and cell development. However, conventional integration methods rarely utilize prior biological knowledge and lack interpretability. To address these challenges, 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 between different modalities (i.e., multi-omics). It also utilizes SHapley Additive Explanation method to reveal key pathways, genes, and regulatory mechanisms. Through benchmark studies on 28 TCGA datasets, we demonstrated the superior performance and interpretability of Pathformer on various cancer classification tasks, compared to other integration models. Furthermore, we applied Pathformer to liquid biopsy multi-modal data integration with high accuracy in cancer diagnosis. Meanwhile, Pathformer revealed interesting molecularly altered pathways in cancer patients' body fluid, such as ligand binding of scavenger receptors, iron transport, and DAP12 signaling transmission, which are related to extracellular vesicle transport, platelet, and immune response.

Keywords: Multi-modal integration; Transformer; Pathway crosstalk network; Cancer diagnosis; Liquid biopsy.

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 (e.g., cancer) 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 classifying diseases. 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 diseases[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 in cancer. Recently, Transformer has demonstrated its capability in handling multi-modal tasks in computational fields[17]. It 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 28 benchmark datasets of the Cancer Genome Atlas (TCGA)[18] and demonstrated its superior performance and biological interpretability on various cancer classification tasks, compared to other integration models. Pathformer was applied to liquid biopsy data, which not only showed high accuracy for noninvasive cancer diagnosis but revealed interesting molecularly altered pathways in human plasma.

Results

The Pathformer model

Pathformer utilizes biological pathway network and a Transformer encoder to allow better information fusion. It has six modules: biological pathway input, pathway crosstalk network calculation, multi-modal data input, biological multi-modal embedding, Transformer module with pathway crosstalk network, and classification module (**Fig. 1a,** see **Methods** for details). 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.

We curated all pathways from four public databases, then selected 1,497 pathways based on the criterion of gene number, overlap ratio with other pathways, and the number of pathway subsets. Next, we used *BinoX*[19], a classic tool for crosstalk analysis, to calculate the crosstalk relationships among the 1,497 pathways. Based on these relationships, we created a pathway crosstalk network as Pathformer's input (see **Methods** and **Supplementary Notes**).

Multi-modal biological data preprocessing and embedding are crucial components of Pathformer (**Fig. 1b**). We preprocessed the raw sequence reads of DNA-seq and RNA-seq into multi-modal data, including DNA methylation, DNA copy number, and different RNA alterations (see **Methods** and **Supplementary Notes**). These multi-modal data are on different levels, such as nucleotide level, fragment level, and gene level, which significantly influence data integration. To address this, we used multiple statistical indicators as gene embeddings to retain the gene diversity across different modalities (see **Fig. 1b** and **Methods**). 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 into pathway embedding. The PSNN has two layers representing genes and pathways, respectively. These two layers are not fully connected, but rather share a connection pruned based on the pathway and gene inclusion relationships. If there is no correlation between a given gene and a given pathway, the connection weight between two neurons is set to be 0; otherwise, it is learned through training (see **Methods**). Therefore, pathway embedding is a dynamic embedding method. The PSNN can not only restore the mapping relationship between genes and pathways, but also identify important genes in different pathways through trained weights, and can transfer the complementarity of modalities at the gene level to the pathway level. 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 is the key module of Pathformer model (**Fig. 1c**). Inspired by the Evoformer model used in AlphaFold2[20] 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. Particularly, multi-head column-wise self-attention (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. Multi-head row-wise self-attention (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. More details of the

Transformer module are described in **Methods**.

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

To evaluate the performance of Pathformer, we tested model on various cancer classification tasks as benchmark studies: cancer early- and late- stage classification (10 TCGA cancer datasets), low- and high- survival risk classification (10 TCGA cancer datasets), and cancer subtype classification (8 TCGA cancer datasets) (see **Supplementary Fig. 1** and **Supplementary Notes**). For these tasks, DNA methylation, DNA CNV, and RNA expression were used as input. For model training and test, 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 (see **Supplementary Fig. 1** and **Methods**). We first optimized hyperparameters using 5-fold cross-validation on the discovery set, with macro-averaged F1 score as the criterion for grid search. The results of optimal hyperparameter combination for each dataset are listed in **Supplementary Fig. 2** and **Supplementary Table 1**. Then, we trained Pathformer using the discovery set with early stopping and tested it on the validation set.

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 Notes**).

Pathformer consistently outperformed the other integration methods in most classification tasks, evaluated by macro-averaged F1 score (F1score_macro) (**Fig. 2**), as well as area under the receiver operating characteristic curve (AUC) and average F1 score weighted by support (F1score_weighted) (**Supplementary Fig. 3** and **Supplementary Table 2**). We showed F1score_macro in the main figure because it is a more robust measurement than the other two scores for the imbalanced classes. In the cancer stage classification and survival classification tasks, Pathformer achieved the best F1score_macro and F1score_weighted in all the 10 datasets, and the best AUC in 8 of 10 datasets. In cancer subtype classification of TCGA, Pathformer achieved the best F1score_macro in 7 of 8 datasets, the best

F1score_weighted in 6 of 8 datasets, and the best AUC in 6 of 8 datasets. Notably, Pathformer substantially outperformed the other methods in the challenging classification tasks like cancer early- and late- stage classification and low- and high- survival risk classification, showing average increases of 11% and 15% in F1score_marco compared with XGBoost, respectively. This highlights Pathformer's exceptional learning ability. Moreover, in terms of stability, Pathformer also showed significantly better generalization ability than the other deep learning algorithms, as indicated by the cross-validation variances (**Supplementary Fig. 4**).

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

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 early- and late- stage classification. First, we evaluated the essentialities of seven different data inputs, including RNA expression, DNA methylation, DNA CNV, and a combination thereof (**Fig. 3a**). By comparing the classification performances of seven models, we discovered that the model with all three modalities as input achieved the best performance, followed by RNA expression-only and DNA methylation-only model. 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 KIRC dataset, but the opposite performances were observed in the LUAD dataset. 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 is a model without pathway crosstalk network bias. Transformer is a model without either pathway crosstalk network bias or row-attention. PSNN is a model that directly uses classification module with pathway embedding as input. NN is a model that directly uses classification module with gene embedding as input. As shown in **Fig. 3b**, the complete Pathformer model achieved the best classification performance, while the performance of CC-Attention, Transformer, PSNN, and NN decreased successively. Transformer had a significantly lower classification performance compared to CC-Attention, but no significant improvement compared to PSNN. This indicates that the criss-cross attention mechanism (**Fig. 1c**) plays a key role in Pathformer, with respect to information fusion and crosstalk between different biological pathways and between different modalities (i.e., multi-omics).

Biological interpretability of the Pathformer model

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[21] (SHAP value) to decipher the important pathways and their key genes (see **Methods**). SHAP value is a post hoc model interpretation method that assigns an importance value to each feature to explain the relationship between features and classification[21]. 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. Finally, the hub module of the updated pathway crosstalk network represents the most critical regulatory mechanism in classification, and is screened by sub-network scores based on SHAP values of pathways. Links of the updated network indicate crosstalk relationships that affect classification tasks (see **Methods**).

Here, we demonstrated the interpretability of Pathformer using the breast cancer subtype classification task as an example (**Fig. 4**). First, at the modality level, we visualized the contributions of different modalities for breast cancer subtype classification by the attention weights (**Fig. 4a**). The contribution of transcriptomic data was greater than 50% in breast cancer subtype classification, which is consistent with the fact that PAM50 is defined based on transcriptomic data[22]. Combining with the results of other classification tasks for breast cancer (**Supplementary Figs. 5a, 6a**), we observed that transcriptome always played a crucial role in various classification tasks; DNA CNV had certain contribution in subtype classification; and DNA methylation contributed substantially in early- and late-stage classification. In addition, the contributions of various statistical indicators in the same modality were also different for different classification tasks. For example, mean of DNA CNV played an important role in subtype classification, while minimum of DNA CNV had greater contribution in stage classification and survival classification. These findings further validated the necessity of multi-modal integration and biological multi-modal embedding.

Next, at the pathway and gene level, we identified the pathways with top 15 SHAP value and the genes with top 5 SHAP value of each pathway as key genes in breast cancer subtype classification (**Fig. 4b**). Then, we presented a hub module of the updated pathway crosstalk network (**Fig. 4c**). Here, *complex I biogenesis* pathway was identified as the most critical pathway in breast cancer subtype classification and a key node in the hub module of the updated pathway crosstalk network. This pathway comprises 57 genes, including mitochondrial genes and protein-coding genes. Complex I participates in the biosynthesis and redox control during cancer cell proliferation and metastasis[23]. Five mitochondrial genes (MT-ND3, MT-ND1, MT-ND4, MT-ND2, and MT-ND6) were identified as key genes of the *complex I biogenesis* pathway in breast cancer subtype classification by Pathformer. These mitochondrial genes have been reported to exhibit distinct patterns in different breast cancer subtypes[24]. In addition, in the hub module of the updated pathway crosstalk network, *complex I biogenesis* pathway was closely related to *TP53-regulated metabolic genes* pathway and *signaling by ERBB4* pathway, and has been identified as the most critical regulatory mechanism for breast cancer subtype classification. According to literatures, TP53 mutation spectrum[25] and ERBB4[26] are biomarkers for breast cancer subtypes.

Moreover, many other important pathways identified by Pathformer for breast cancer subtype classification have also already been reported previously (**Fig. 4b**). For example, the expression of *nucleotide excision repair* pathway is reduced in TNBC, which may affect survival after platinum chemotherapy of patients[27]. RFC4 is the key gene of this pathway, and DNA CNV of RFC4 was reported to play a crucial role in determining individual breast cancer subtypes[28], which is consistent with the prediction of the gene's pillar module by Pathformer. Key genes of *transcription of E2F targets under negative control by p107 and p130 in complex with HDAC1* pathway were identified as E2F1, HDAC1, RBBP4, CCNA2, and CDK1 by Pathformer. Most E2F family genes expressions are significantly up-regulated in TNBC, and are predictive biomarkers of neoadjuvant therapies in patients with ER-positive/HER2-negative tumors[29]. In addition to the transcriptome level, DNV CNV of E2F1 is also a susceptibility factor for breast cancer[30], again consistent with the prediction of the gene's pillar module by Pathformer. HDAC1 is significantly lower in HER2-positive and TNBC compared to luminal A and luminal B[31].

Similarly, we also analyzed important pathways and hub modules of the updated pathway crosstalk network in breast cancer early- and late-stage classification and high- and low-risk survival classification (**Supplementary Figs. 5,6**). We found that *complex I biogenesis* pathway always played a crucial role in different classification tasks of breast cancer, due to its connection between various cancer-related pathways. Particularly, in breast cancer early- and late-stage classification, *iron uptake and transport* pathway had the greatest impact. Supportively, the transport and storage of iron in cells are known to play a key role in carcinogenesis, cell proliferation, and the development of breast cancer[32]. Furthermore, we found that some pathways were more important in early- and late-stage classification than in subtype classification and survival classification, such as *collagen biosynthesis and modifying enzymes* pathway, *Eph/ephrin signaling* pathway, *FRA* pathway, and *G1* pathway. *Roles of LAT2/NTAL/LAB in calcium mobilization* pathway was more important in survival classification than in the other classification tasks, which was consistent with calcium signaling pathway's function in breast cancer cells' proliferation, invasion, apoptosis, and multidrug resistance, and with breast cancer survival[33].

Application of Pathformer to liquid biopsy data for non-invasive cancer diagnosis

Liquid biopsy is a non-invasive detection way with important clinical applications in both cancer diagnosis and status monitoring, which provides comprehensive information on transcriptome dynamics[34]. RNA alterations reflect the complementarity between different levels of information and help to overcome missed detection results of single data to further improve the accuracy of cancer diagnosis. Therefore, we used Pathformer to integrate multi-modal data of liquid biopsies for classifying cancer patients from healthy controls. We applied Pathformer to three cell-free RNA-seq datasets derived from three different blood components: plasma, extracellular vesicle (EV), and platelet datasets (see **Methods**).

We calculated seven RNA-level modalities from RNA-seq data as Pathformer's input, including RNA expression, RNA splicing, RNA editing, RNA alternative promoter (RNA alt. promoter), RNA allele-specific expression (RNA ASE), RNA single nucleotide variations (RNA SNV), and chimeric RNA. From results of 5-fold cross-validation in **Supplementary Fig. 7**, we found that the model with all modalities as input had the best comprehensive performance on three datasets, followed by RNA expression-only model and RNA alt. promoter-only model, and some models with other modalities exhibited great fluctuations on different datasets. In order to effectively integrate information without redundancy, we performed further feature selection based on different modality combinations evaluated by Pathformer. First, we calculated the contributions of each modality and its corresponding statistical indicators (**Fig. 5a**). Similar to results of cross-validation, RNA expression was the core modality across all datasets. Next, we performed 5-fold cross-validation find an optimal modality combination for each dataset (**Fig. 5b, Supplementary Table 3**). We found that plasma dataset with 7 modalities, EV dataset with 3 modalities, and platelet dataset with 3 modalities obtained the best performance. The AUCs were higher than 0.9 for all three datasets. In conclusion, Pathformer effectively integrated multi-modal data from human plasma, and accurately classified cancer patients from healthy controls.

Pathformer reveals deregulated pathways and genes in cancer patients' plasma

Because the Pathformer model has biological interpretability, we used Pathformer to predict cancer related pathways and genes in the above liquid biopsy data (**Fig. 6**). Then, we can gain insight into the deregulated alterations in body fluid (i.e., plasma) for cancer patients vs. healthy controls.

First, in comparison to cancer tissue data (**Fig. 4, Supplementary Fig. 6**), we found that vesicle transport and coagulation related pathways occupied an important position in datasets of various blood components, which is consistent with the characteristics of body fluids (**Fig. 6a-c**). Furthermore, we also observed that active pathways and key genes of plasma dataset were more similar to those in platelet dataset, which is consistent with a recent report showing platelet is a major origin in the plasma cell-free transcriptome[35].

Next, we examined there interesting pathways: one was found in EV data and the others were revealed from platelet data. In both EV and plasma datasets, we found that *binding and uptake of ligands* (e.g., oxidized low-density lipoprotein, oxLDL) *by scavenger receptors* pathway was identified as the most active pathway (**Fig. 6a, b**). It is well established that scavenger receptors play a crucial role in cancer prognosis and carcinogenesis by promoting the degradation of harmful substances and accelerating the immune response through endocytosis, phagocytosis, and adhesion[36]. Scavenger receptors are also closely related to the transport process of vesicles. For example, stabilin-1, a homeostatic receptor, has the potential to impact macrophage secretion by linking extracellular signals and intracellular vesicular processes[37]. Meanwhile, HBB, HBA1, HBA2, FTH1, HSP90AA1 were identified as key genes in this pathway. HBB has been reported as a biomarker in thyroid cancer[38], breast cancer[39], and gastric cancer[40]. It has also been demonstrated that HBB is significantly downregulated in gastric cancer blood transcriptomics[40]. HSP90AA1 has also been demonstrated to be a potential biomarker for various cancers[41], especially in the blood[42].

The other interesting pathways are *DAP12 signaling* pathway and *DAP12 interactions* pathway revealed in both platelet and plasma datasets (**Fig. 6a, c**). DAP12 triggers natural killer cell immune responses against certain tumor cells[43], which is regulated by platelet[44]. Among the top 5 key genes of DAP12 related pathway in both platelet and plasma datasets, B2M was reported as a serum protein encoding gene and a widely recognized tumor biomarker[45]; HLA-E and HLA-B were reported as cancer biomarkers in tissue and plasma[46,47].

In addition, Pathformer provides insight into the interplay between various biological processes and their impact on cancer progression by updating pathway crosstalk network (**Fig. 6d-e**). In the plasma data, the link between *binding and uptake of ligands by scavenger receptors* pathway and *iron uptake and transport* pathway was a novel addition to the updated network (**Fig. 6d**). In other words, this crosstalk relationship was newly predicted by Pathformer. The crosstalk between two pathways was amplified by Pathformer in plasma dataset, probably because they were important for classification and shared the same key gene, FTH1, one of two intersecting genes between the two pathways. However, in platelet dataset, this crosstalk between two pathways was not shown, when the scavenger receptors pathway was not important enough (**Fig. 6e**). In summary, Pathformer's updated pathway crosstalk network visualizes the information flow between pathways related to cancer classification task in the liquid biopsy data, providing novel insight into the cross-talk of biological pathways in cancer patients' plasma.

Discussion

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

Applications of Pathformer. Pathformer will be usefully in many clinical applications like cancer subtyping, staging, prognosis, and diagnosis. For instance, we have demonstrated excellent performance of Pathformer on noninvasive diagnosis of cancer based on multi-modal data of liquid biopsy. The accuracies (AUC scores) of cancer classification in plasma, EV, and platelet datasets were all higher than 90%. Furthermore, the interpretability of the Pathformer model can help researchers gain insights into the complex regulation processes involved in cancer. For instance, Pathformer has identified active pathways consistent with the characteristics of body fluid data, such as binding and uptake of ligands by scavenger receptors, and the DAP12 related pathway, which have been reported to be closely related to extracellular vesicle transport, platelet, and immune response during the development and progression of cancer.

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[48]. 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.

Methods

Data collection and preprocessing

We collected 28 datasets across different cancer types from TCGA to evaluate classification performance of Pathformer and existing comparison methods, which consists of 8 datasets for cancer subtype classification, 10 datasets for cancer early- and late- stage classification, and 10 datasets for cancer low- and high- survival risk 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[49]), the extracellular vesicle (EV) dataset (comprising 477 samples from two studies assayed by exosomal RNA-seq[50,51]), and the platelet dataset (comprising 918 sample from two studies assayed by tumor-educated blood platelet RNA-seq[52,53]). Through our biological information pipeline, totally 4 and 7 biological modalities are obtained for TCGA dataset and liquid biopsy dataset, respectively. More details of data collection and preprocessing are described in **Supplementary Fig. 1** and **Supplementary Notes**.

The Pathformer model

As shown in **Fig. 1**, Pathformer consists of the following six modules: biological pathway input, pathway crosstalk network calculation, multi-modal data input, biological multi-modal embedding, Transformer module with pathway crosstalk network bias, and classification module.

Biological pathways and crosstalk network

We collected 2,289 pathways of four public databases including Kyoto Encyclopedia of Genes and Genomes database (KEGG)[54], Pathway Interaction database (PID)[55], Reactome database (Reactome)[56], and BioCarta Pathways database (BioCarta)[57]. Then, we filtered these pathways by three criteria: 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 1,497 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 Notes**).

Biological multi-modal data input and embedding

Pathformer supports any number of modalities as input which may 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. Pathformer's input for liquid biopsy datasets includes gene-level RNA expression; fragment-level RNA alternative promoter, RNA splicing, and chimeric RNA; and nucleotide-level RNA editing, RNA ASE, and RNA SNV. We represented multi-modal input matrix of a sample as $\boldsymbol{M}$, and converted matrix $\boldsymbol{M}$ into gene encoding $\boldsymbol{E}_G$ and pathway encoding $\boldsymbol{E}_P$. First, we used a series of statistical indicators in different modalities as gene embedding. These statistical indicators include gene level score, count, entropy, minimum, maximum, mean, weighted mean in whole gene, and weighted mean in window. Gene embedding is calculated as follows:

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

, where $\boldsymbol{G}_i$ is modality i, D_g is length of gene embedding for all modalities, $\boldsymbol{F}_E$ 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 the gene level, and the embedding functions corresponding to different modalities are different (more details in **Supplementary Notes**). Then, we used the known biological pathways to construct a sparse neural network for converting the gene embedding $\boldsymbol{E}_G$ into the pathway embedding $\boldsymbol{E}_P$, as described below:

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

, 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.

Transformer module with pathway crosstalk network bias

We employed the Transformer module based on criss-cross attention with pathway crosstalk network bias, which has 3 blocks. Each block of Transformer module contains the following processes: 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. 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 hth 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 hth 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 hth 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 without pathway crosstalk network bias. The hth 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 hth 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 hth 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 Notes**.

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. 1**).

When training the model, we used a normal model learning strategy. We applied cross-entropy loss with classimbalance 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. We used the macro-averaged F1 score as the selection criterion to find the optimal combination of maximum of learning rate∈[1e-4, 1e-5], dropout probability of classification (c)∈[0.3, 0.5], and constant coefficient for row-attention ($\boldsymbol{\beta}$)∈[0.1,1]. 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 Notes**.

Model interpretability

To better understand Pathformer's decisions, we increased the interpretability of Pathformer by calculating contributions of different modalities, important pathways and their key genes, and hub module of the updated pathway crosstalk network.

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[21] (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:

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

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

, 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, we calculated SHAP values of pathways and genes in different modalities, 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 Notes**.

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 (81972798, 32170671), Tsinghua University Spring Breeze Fund (2021Z99CFY022), 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. We also thank Hongli Ma and Kexing Li for helping us edit the text of manuscript.

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, doi:10.1038/s41467-021-23774-w (2021).

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

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, doi:10.1186/s12859-018-2500-z (2018).

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, doi:10.1093/bioinformatics/btab285 (2021).

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

19 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).

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

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

22 Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology* **27**, 1160 (2009).

23 Urra, F. A., Muñoz, F., Lovy, A. & Cárdenas, C. The mitochondrial complex (I) ty of cancer. *Frontiers in oncology* **7**, 118 (2017).

24 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).

25 Silwal-Pandit, L. *et al.* TP53 Mutation Spectrum in Breast Cancer Is Subtype Specific and Has Distinct Prognostic RelevanceTP53 in Breast Cancer. *Clinical Cancer Research* **20**, 3569-3580 (2014).

26 Sundvall, M. *et al.* Role of ErbB4 in breast cancer. *Journal of mammary gland biology and neoplasia* **13**, 259-268 (2008).

27 Ribeiro, E. *et al.* Triple negative breast cancers have a reduced expression of DNA repair genes. *PLoS One* **8**, e66243 (2013).

28 Srihari, S. *et al.* Understanding the functional impact of copy number alterations in breast cancer using a network modeling approach. *Molecular BioSystems* **12**, 963-972 (2016).

29 Oshi, M. *et al.* The E2F pathway score as a predictive biomarker of response to neoadjuvant therapy in ER+/HER2− breast cancer. *Cells* **9**, 1643 (2020).

30 Rocca, M. S. *et al.* E2F1 copy number variations in germline and breast cancer: a retrospective study of 222 Italian women. *Molecular Medicine* **27**, 1-7 (2021).

31 Guo, Q. *et al.* Expression of HDAC1 and RBBP4 correlate with clinicopathologic characteristics and prognosis in breast cancer. *International journal of clinical and experimental pathology* **13**, 563 (2020).

32 Marques, O., da Silva, B. M., Porto, G. & Lopes, C. Iron homeostasis in breast cancer. *Cancer letters* **347**, 1-14 (2014).

33 So, C. L., Saunus, J. M., Roberts-Thomson, S. J. & Monteith, G. R. in *Seminars in cell & developmental biology.* 74-83 (Elsevier).

34 Schwarzenbach, H., Hoon, D. S. & Pantel, K. Cell-free nucleic acids as biomarkers in cancer patients. *Nature Reviews Cancer* **11**, 426-437 (2011).

35 Vorperian, S. K. *et al.* Cell types of origin of the cell-free transcriptome. *Nature biotechnology* **40**, 855-861 (2022).

36 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).

37 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).

38 Onda, M. *et al.* Decreased expression of haemoglobin beta (HBB) gene in anaplastic thyroid cancer and recovory of its expression inhibits cell growth. *British Journal of Cancer* **92**, 2216-2224 (2005).

39 Ponzetti, M. *et al.* Non-conventional role of haemoglobin beta in breast malignancy. *British journal of cancer* **117**, 994-1006 (2017).

40 Lee, I.-S. *et al.* A blood-based transcriptomic signature for noninvasive diagnosis of gastric cancer. *British Journal of Cancer* **125**, 846-853 (2021).

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

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

43 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).

44 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).

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

46 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).

47 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).

48 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).

49 Chen, S. *et al.* Cancer type classification using plasma cell-free RNAs derived from human and microbes. *eLife* **11**, e75181 (2022).

50 Li, S. *et al.* exoRBase: a database of circRNA, lncRNA and mRNA in human blood exosomes. *Nucleic acids research* **46**, D106-D112 (2018).

51 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).

52 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).

53 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).

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

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

56 Croft, D. *et al.* Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research* **39**, D691-D697 (2010).

57 Nishimura, D. BioCarta. *Biotech Software & Internet Report: The Computer Software Journal for Scient* **2**, 117-120 (2001).

Figure Legends

Figure 1. Overview of the Pathformer model.

a. Model architecture of Pathformer. F_E, statistical indicators in the gene embedding. **b.** Calculation of biological multi-modal 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. **c**. A block of Transformer module with pathway crosstalk network bias (3 blocks used in **a**). The pathway embedding matrix is used as input and the pathway crosstalk network matrix is used as bias. N_p, number of pathways; D_p, dimensionality of pathway embedding; h, number of attention heads; d, attention dimension; $\boldsymbol{V}_1$, $\boldsymbol{K}_1$, $\boldsymbol{Q}_1$, $\boldsymbol{A}_1$: vale, key, query and attention map of col-attention; $\boldsymbol{V}_2$, $\boldsymbol{K}_2$, $\boldsymbol{Q}_2$, $\boldsymbol{A}_2$: vale, key, query and attention map of row-attention; +, element-wise addition; ×, matrix multiplication; ∘, matrix dot product; β, constant coefficient for row-attention.

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

Bar charts show the macro-averaged F1 score of different multi-modal integration methods in different classification tasks of TCGA datasets. Error bars are from 2 times 5-fold cross-validation, representing 95% confidence intervals. XGBoost refers to the early integration methods based on gradient boosted tree, while XGBoost (late) refers to the late integration methods based on gradient boosted tree.

Figure 3. Ablation analysis of Pathformer for the classification of early- and late-stage cancer patients.

a. Different types of data (modalities) were used as input for TCGA cancer early- and late-stage classification. **b.** Ablation analysis of different modules in Pathformer. Error bars are from 2 times 5-fold cross-validation across 8 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.

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

a. Contributions of different modalities for breast cancer (BRCA) subtype classification calculated by attention weights (averaging attention maps of row-attention). **b.** Important pathways and their key genes with top SHapley Additive exPlanations (SHAP) values. Among the key genes, different colors represent different pillar modalities of the genes. **c**. A hub module of pathway crosstalk network for BRCA subtype classification. Color depth and size of node represents the degree of node. Line thickness represents the weight of edge. All links are predicted by Pathformer, where known links are reported by the initial crosstalk network and new links are new predictions.

Figure 5. Pathformer integrates multi-modal liquid biopsy data for non-invasive cancer diagnosis.

a. Contributions of different input features and their statistical indicators when classifying cancer patients from healthy controls using three liquid biopsy datasets. All mean represents the sum of mean, weighted mean and window weighted mean. Each type of RNA splicing is the sum of all statistical indicators in this type. **b**. Classification performance of different input combinations. Each value is the mean of 5-fold cross-validation.

Figure 6. Interpretation of the liquid biopsy data using Pathformer.

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

Figure 1. Overview of the Pathformer model.

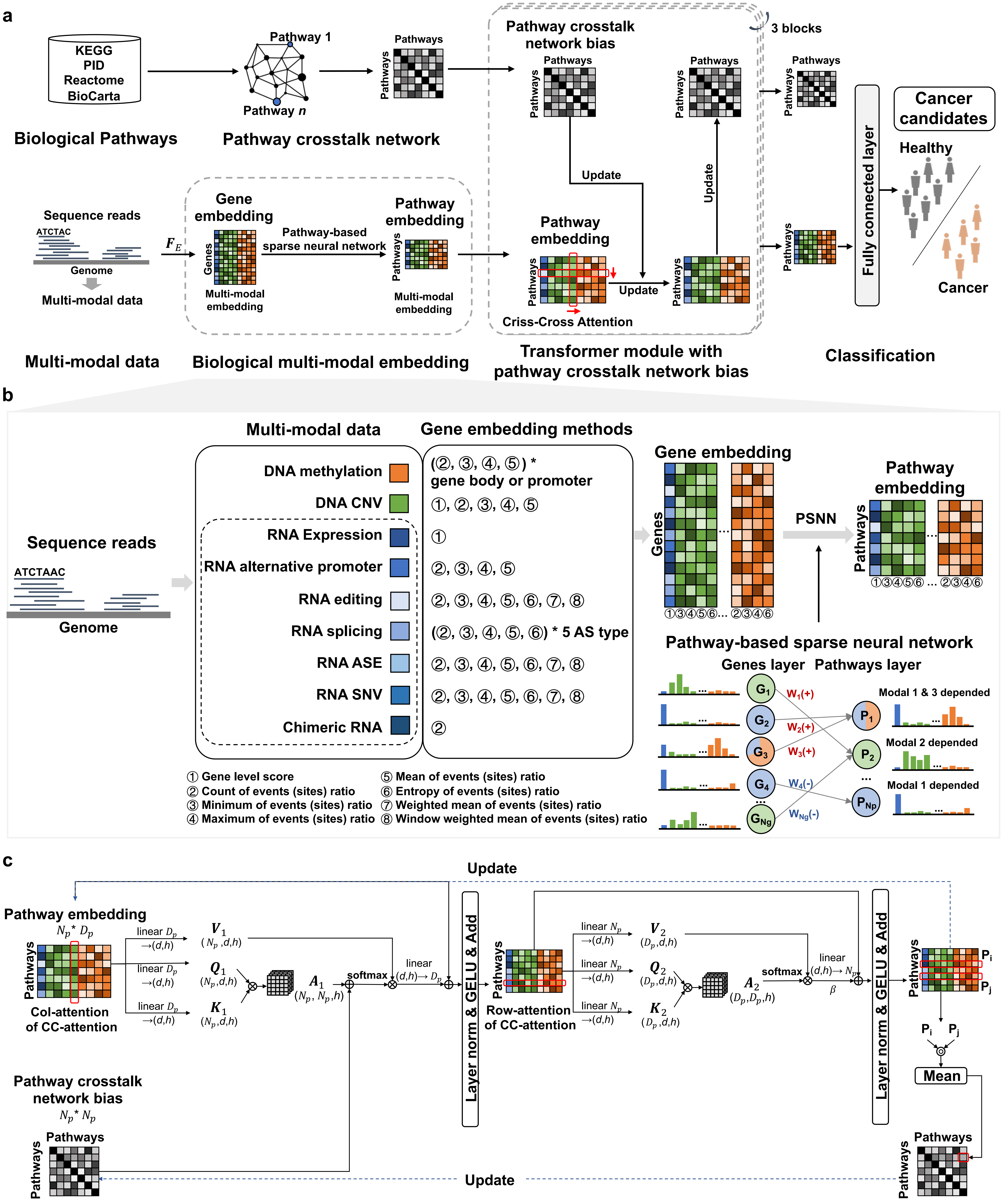


Figure 2. Performance comparison between Pathformer and other multi-modal integration methods

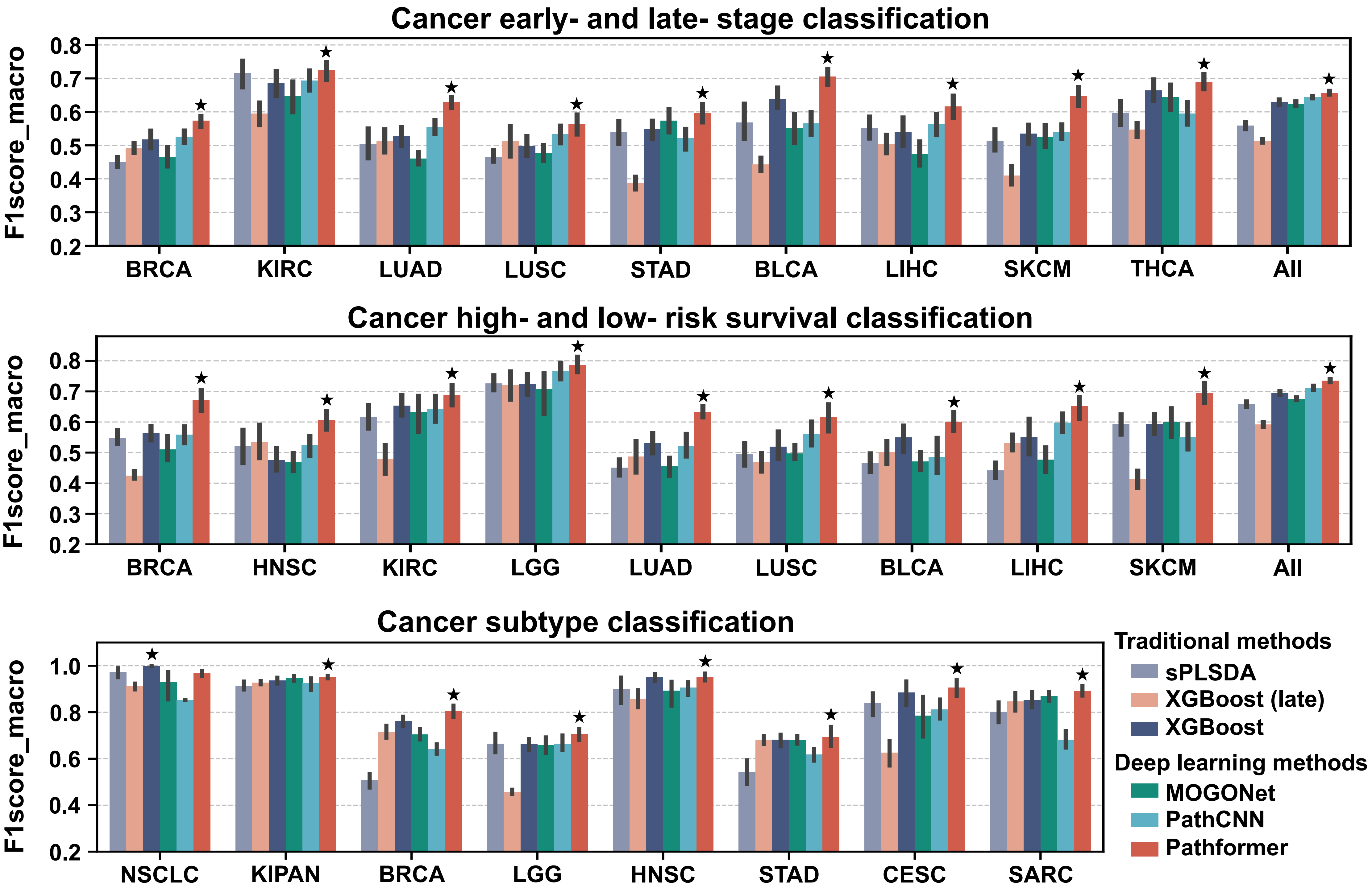


Figure 3. Ablation analysis of Pathformer for the classification of early- and late-stage cancer patients.

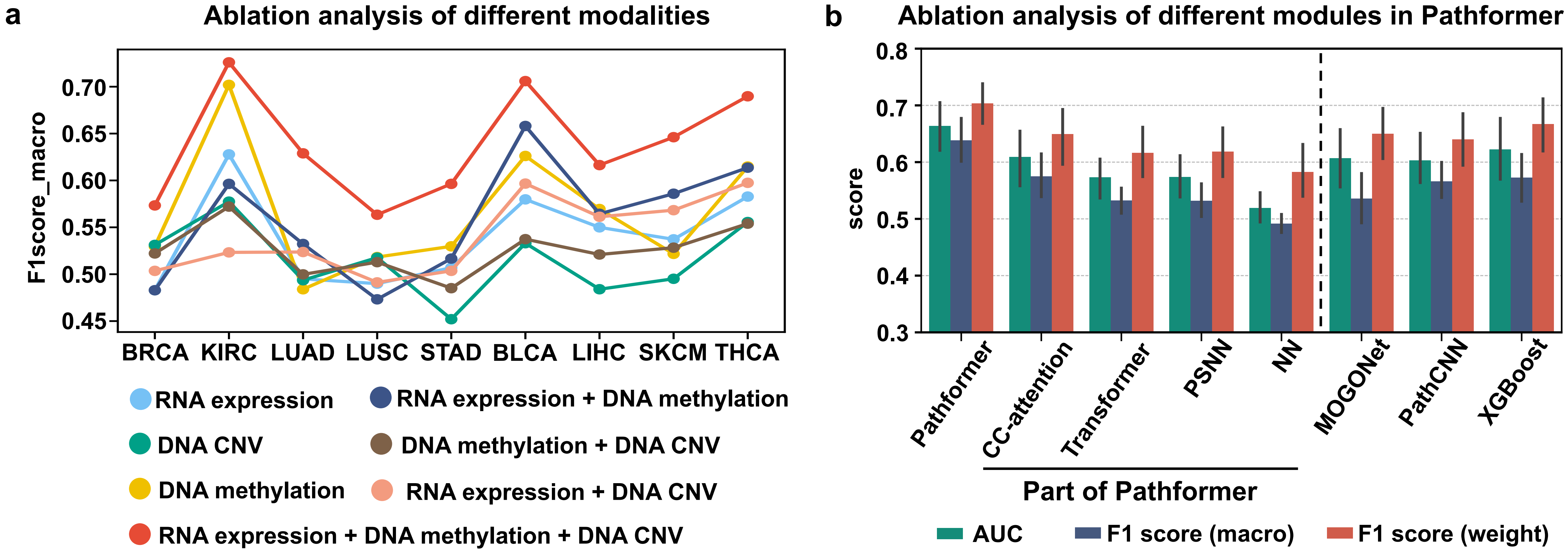


Figure 4. Breast cancer subtype related modalities, pathways and genes revealed by Pathformer

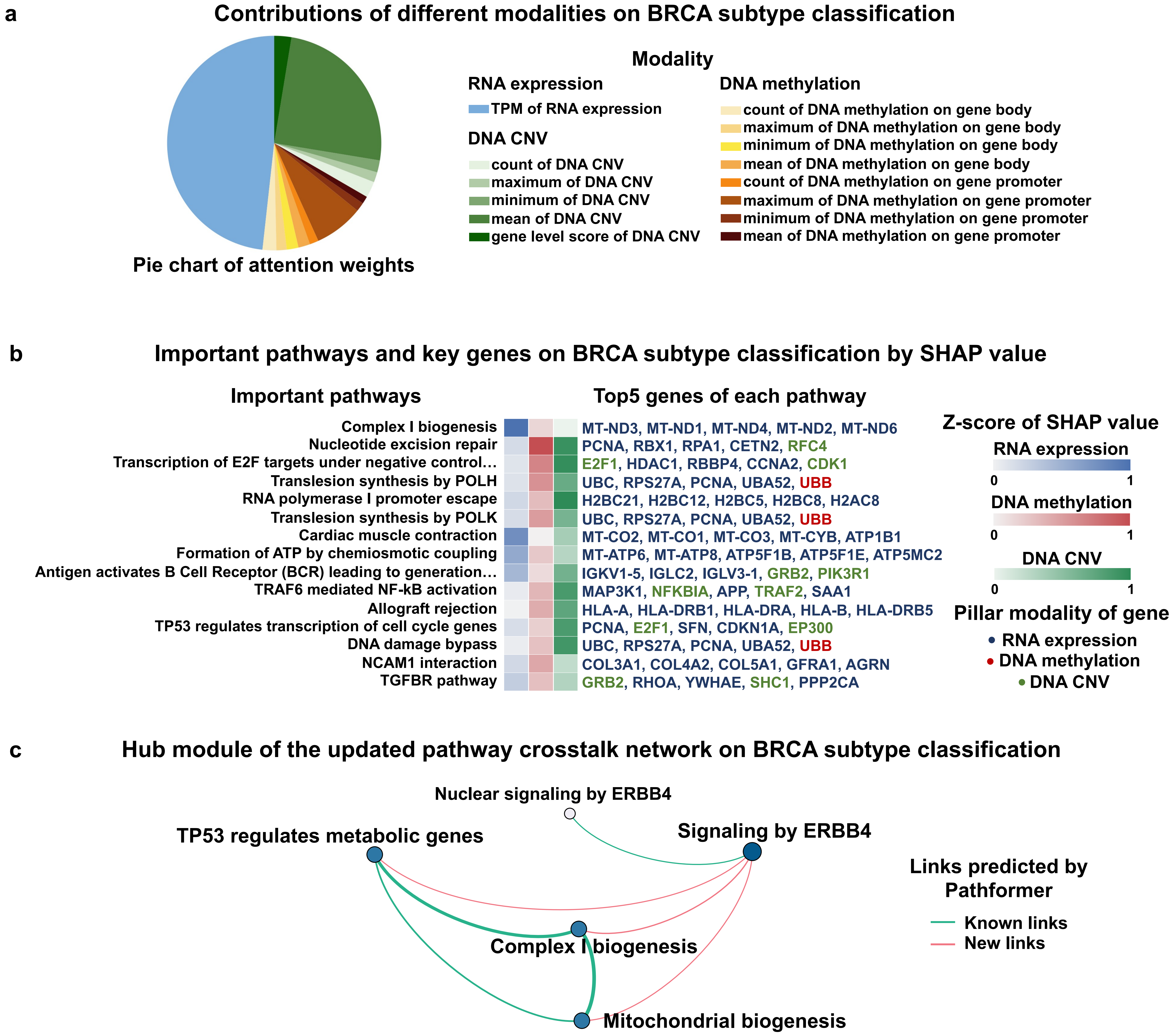


Figure 5. Pathformer integrates multi-modal liquid biopsy data for non-invasive cancer diagnosis

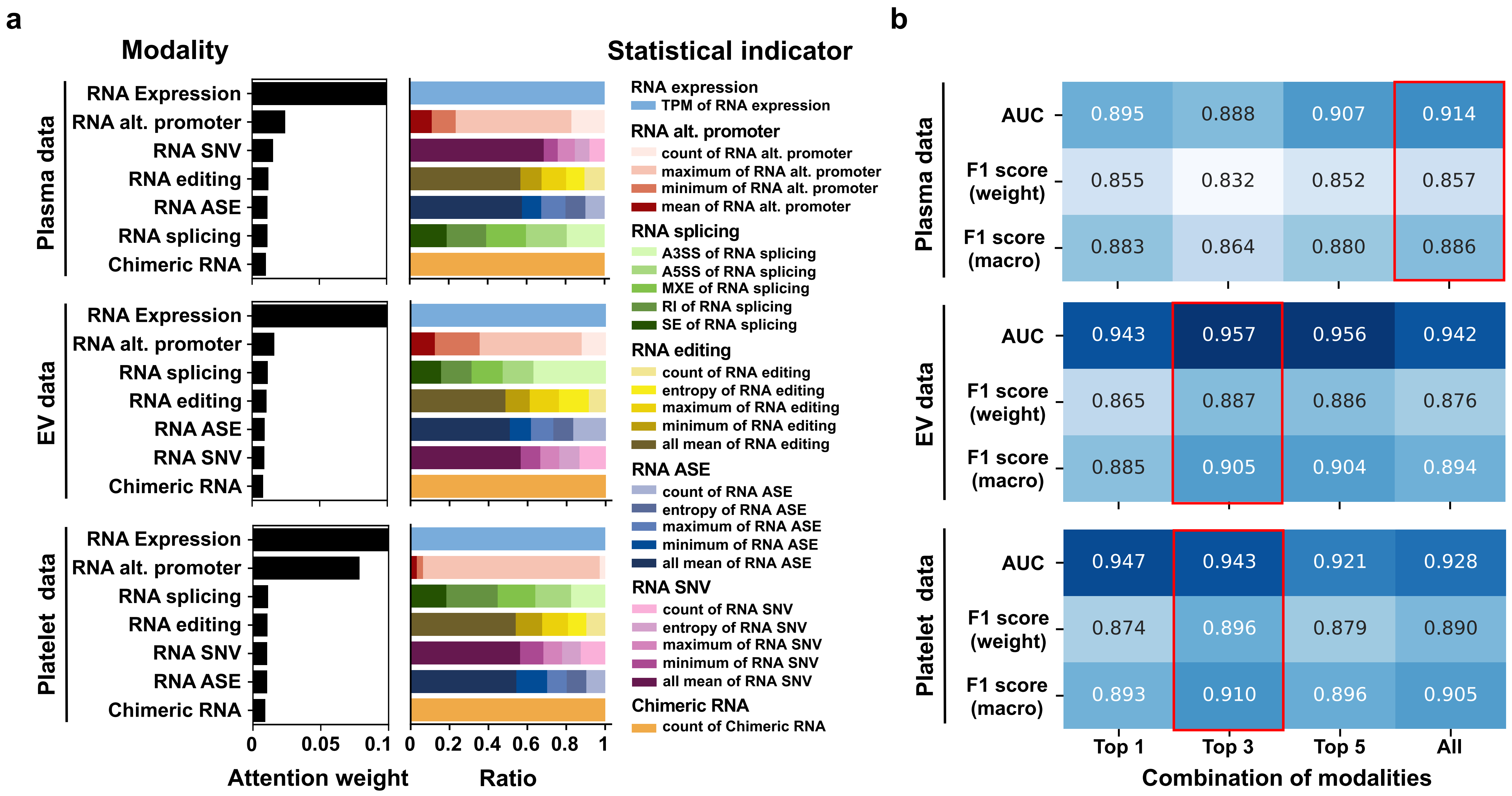
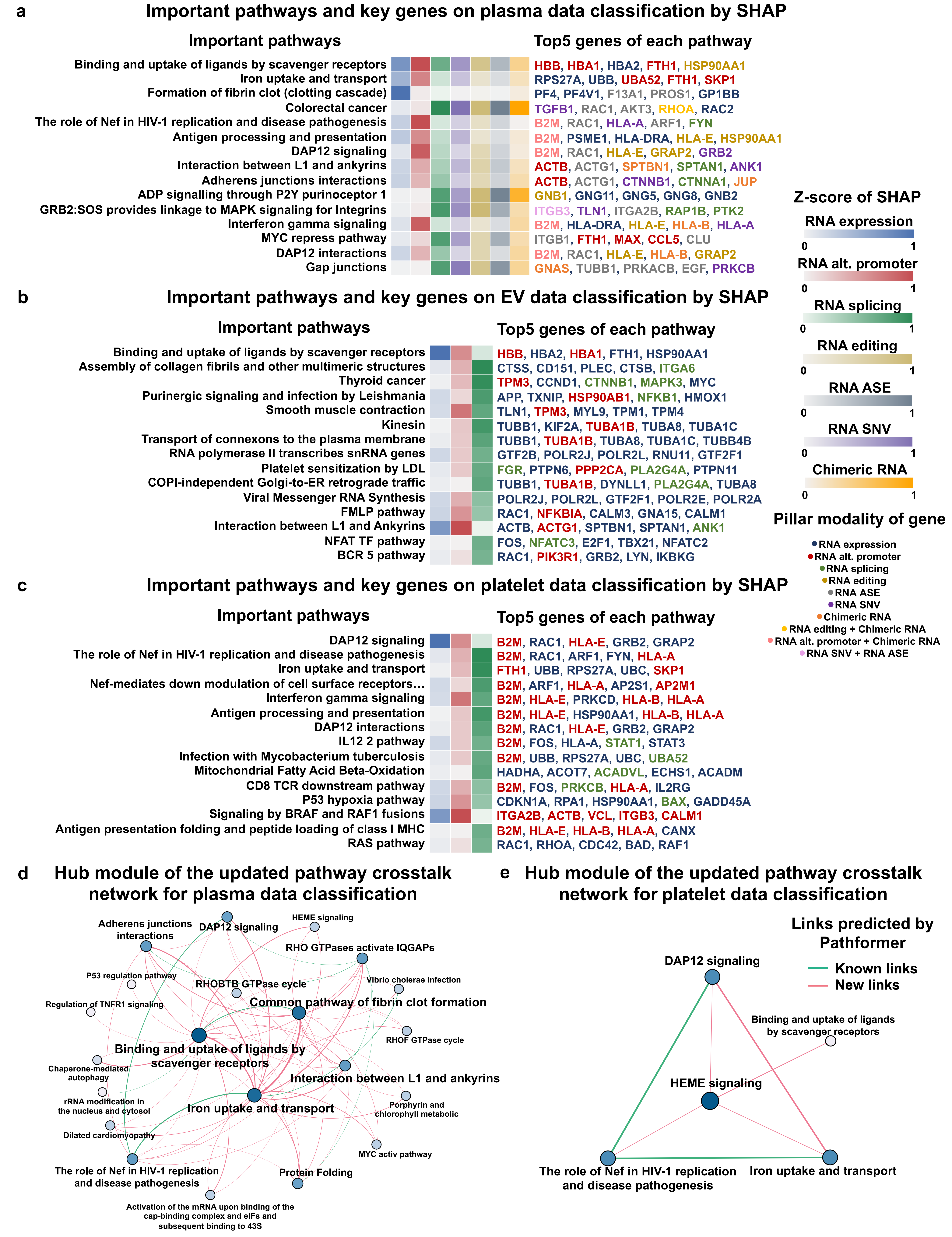


Figure 6. Interpretation of the liquid biopsy data using Pathformer


a
Important pathways and key genes on plasma data classification by SHAP
Important pathways
Top5 genes of each pathway
Binding and uptake of ligands by scavenger receptors
Iron uptake and transport
Formation of fibrin clot (clotting cascade)
Colorectal cancer
The role of Nef in HIV-1 replication and disease pathogenesis
Antigen processing and presentation
DAP12 signaling
Interaction between L1 and ankyrins
Adherens junctions interactions
ADP signalling through P2Y purinoceptor 1
GRB2:SOS provides linkage to MAPK signaling for Integrins
Interferon gamma signaling
MYC repress pathway
DAP12 interactions
Gap junctions
HBB, HBA1, HBA2, FTH1, HSP90AA1
RPS27A, UBB, UBA52, FTH1, SKP1
PF4, PF4V1, F13A1, PROS1, GP1BB
TGFB1, RAC1, AKT3, RHOA, RAC2
B2M, RAC1, HLA-A, ARF1, FYN
B2M, PSME1, HLA-DRA, HLA-E, HSP90AA1
B2M, RAC1, HLA-E, GRAP2, GRB2
ACTB, ACTG1, SPTBN1, SPTAN1, ANK1
ACTB, ACTG1, CTNNB1, CTNNA1, JUP
GNB1, GNG11, GNG5, GNG8, GNB2
ITGB3, TLN1, ITGA2B, RAP1B, PTK2
B2M, HLA-DRA, HLA-E, HLA-B, HLA-A
ITGB1, FTH1, MAX, CCL5, CLU
B2M, RAC1, HLA-E, HLA-B, GRAP2
GNAS, TUBB1, PRKACB, EGF, PRKCB
Z-score of SHAP
RNA expression
RNA alt. promoter
RNA splicing
RNA editing
RNA ASE
RNA SNV
Chimeric RNA
Pillar modality of gene
● RNA expression
● RNA alt. promoter
● RNA splicing
● RNA editing
● RNA ASE
● RNA SNV
● Chimeric RNA
● RNA editing + Chimeric RNA
● RNA alt. promoter + Chimeric RNA
● RNA SNV + RNA ASE
b
Important pathways and key genes on EV data classification by SHAP
Important pathways
Top5 genes of each pathway
Binding and uptake of ligands by scavenger receptors
Assembly of collagen fibrils and other multimeric structures
Thyroid cancer
Purinergic signaling and infection by Leishmania
Smooth muscle contraction
Kinesin
Transport of connexons to the plasma membrane
RNA polymerase II transcribes snRNA genes
Platelet sensitization by LDL
COPI-independent Golgi-to-ER retrograde traffic
Viral Messenger RNA Synthesis
FMLP pathway
Interaction between L1 and Ankyrins
NFAT TF pathway
BCR 5 pathway
HBB, HBA2, HBA1, FTH1, HSP90AA1
CTSS, CD151, PLEC, CTSB, ITGA6
TPM3, CCND1, CTNNB1, MAPK3, MYC
APP, TXNIP, HSP90AB1, NFKB1, HMOX1
TLN1, TPM3, MYL9, TPM1, TPM4
TUBB1, KIF2A, TUBA1B, TUBA8, TUBA1C
TUBB1, TUBA1B, TUBA8, TUBA1C, TUBB4B
GTF2B, POLR2J, POLR2L, RNU11, GTF2F1
FGR, PTPN6, PPP2CA, PLA2G4A, PTPN11
TUBB1, TUBA1B, DYNLL1, PLA2G4A, TUBA8
POLR2J, POLR2L, GTF2F1, POLR2E, POLR2A
RAC1, NFKBIA, CALM3, GNA15, CALM1
ACTB, ACTG1, SPTBN1, SPTAN1, ANK1
FOS, NFATC3, E2F1, TBX21, NFATC2
RAC1, PIK3R1, GRB2, LYN, IKBKG
c
Important pathways and key genes on platelet data classification by SHAP
Important pathways
Top5 genes of each pathway
DAP12 signaling
The role of Nef in HIV-1 replication and disease pathogenesis
Iron uptake and transport
Nef-mediates down modulation of cell surface receptors…
Interferon gamma signaling
Antigen processing and presentation
DAP12 interactions
IL12 2 pathway
Infection with Mycobacterium tuberculosis
Mitochondrial Fatty Acid Beta-Oxidation
CD8 TCR downstream pathway
P53 hypoxia pathway
Signaling by BRAF and RAF1 fusions
Antigen presentation folding and peptide loading of class I MHC
RAS pathway
B2M, RAC1, HLA-E, GRB2, GRAP2
B2M, RAC1, ARF1, FYN, HLA-A
FTH1, UBB, RPS27A, UBC, SKP1
B2M, ARF1, HLA-A, AP2S1, AP2M1
B2M, HLA-E, PRKCD, HLA-B, HLA-A
B2M, HLA-E, HSP90AA1, HLA-B, HLA-A
B2M, RAC1, HLA-E, GRB2, GRAP2
B2M, FOS, HLA-A, STAT1, STAT3
B2M, UBB, RPS27A, UBC, UBA52
HADHA, ACOT7, ACADVL, ECHS1, ACADM
B2M, FOS, PRKCB, HLA-A, IL2RG
CDKN1A, RPA1, HSP90AA1, BAX, GADD45A
ITGA2B, ACTB, VCL, ITGB3, CALM1
B2M, HLA-E, HLA-B, HLA-A, CANX
RAC1, RHOA, CDC42, BAD, RAF1
d
Hub module of the updated pathway crosstalk network for plasma data classification
Adherens junctions interactions
DAP12 signaling
HEME signaling
RHO GTPases activate IQGAPs
P53 regulation pathway
RHOBTB GTPase cycle
Vibrio cholerae infection
Regulation of TNFR1 signaling
Common pathway of fibrin clot formation
RHOF GTPase cycle
Binding and uptake of ligands by scavenger receptors
Chaperone-mediated autophagy
Interaction between L1 and ankyrins
rRNA modification in the nucleus and cytosol
Iron uptake and transport
Porphyrin and chlorophyll metabolic
Dilated cardiomyopathy
MYC activ pathway
The role of Nef in HIV-1 replication and disease pathogenesis
Protein Folding
Activation of the mRNA upon binding of the cap-binding complex and eIFs and subsequent binding to 43S
e
Hub module of the updated pathway crosstalk network for platelet data classification
Links predicted by Pathformer
Known links
New links
DAP12 signaling
Binding and uptake of ligands by scavenger receptors
HEME signaling
The role of Nef in HIV-1 replication and disease pathogenesis
Iron uptake and transport