Computational and Structural Biotechnology Journal 23 (2024) 1715–1724

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research Article

Multi-layered knowledge graph neural network reveals pathway-level agreement of three breast cancer multi-gene assays

Sangseon Lee [a,1], Joonhyeong Park [a,1], Yinhua Piao [b], Dohoon Lee [c,d], Danyeong Lee [e], Sun Kim [b,e,f,g,*]

[a] Institute of Computer Technology, South Korea
[b] Department of Computer Science and Engineering, South Korea
[c] Bioinformatics Institute, South Korea
[d] BK21 FOUR Intelligence Computing, South Korea
[e] Interdisciplinary Program in Bioinformatics, South Korea
[f] Interdisciplinary Program in Artificial Intelligence, Seoul National University, Gwanak-ro 1, Gwanak-gu, Seoul, 08826, South Korea
[g] AIGENDRUG Co., Ltd., Gwanak-ro 1, Gwanak-gu, Seoul, 08826, South Korea

ARTICLE INFO

Keywords:
Knowledge graph
Multi-gene assay
Breast cancer recurrence
Graph neural network
Regulatory landscape

ABSTRACT

Multi-gene assays have been widely used to predict the recurrence risk for hormone receptor (HR)-positive breast cancer patients. However, these assays lack explanatory power regarding the underlying mechanisms of the recurrence risk. To address this limitation, we proposed a novel multi-layered knowledge graph neural network for the multi-gene assays. Our model elucidated the regulatory pathways of assay genes and utilized an attention-based graph neural network to predict recurrence risk while interpreting transcriptional subpathways relevant to risk prediction. Evaluation on three multi-gene assays—Oncotype DX, Prosigna, and EndoPredict—using SCAN-B dataset demonstrated the efficacy of our method. Through interpretation of attention weights, we found that all three assays are mainly regulated by signaling pathways driving cancer proliferation especially RTK-ERK-ETS-mediated cell proliferation for breast cancer recurrence. In addition, our analysis highlighted that the important regulatory subpathways remain consistent across different knowledgebases used for constructing the multi-level knowledge graph. Furthermore, through attention analysis, we demonstrated the biological significance and clinical relevance of these subpathways in predicting patient outcomes. The source code is available at http://biohealth.snu.ac.kr/software/ExplainableMLKGNN.

1. Introduction

Breast cancer has been extensively studied at the molecular level, leading to the widespread use of multi-gene assays as prognostic and predictive biomarkers. In particular, several multi-gene assays, such as Oncotype DX, Prosigna (also known as PAM50) and EndoPredict, are widely used for early-stage, estrogen receptor (ER)-positive, node-negative breast cancer patients. Oncotype DX [1] analyzed 250 genes through methods like literature mining and pathway analysis, identifying 16 target genes associated with recurrence in breast cancer across three clinical studies. Prosigna [2] utilized hierarchical clustering analysis and statistical tests to identify 50 genes as biomarkers that were widely used for defining molecular subtypes of breast cancers (Luminal A, Luminal B, Her2-enriched, Basal). EndoPredict [3] used an outcome-driven strategy to select eight genes strongly associated with 10-year outcomes, integrating clinical factors like tumor size and nodal status to calculate the recurrence score. Further information on these three multi-gene assays is available in the Supplementary Material.

While these multi-gene assays are now widely used for predicting breast cancer recurrence globally, a significant drawback remains in their limited explanatory power. These assays rely on outcome-driven approaches to predict the risk of recurrence. Also, since quantitative reverse transcription PCR (qRT-PCR) and microarray-based gene expression analysis have been the main measurement for the multi-gene assays, the number of selected genes is limited to a few genes for cost effectiveness and flexibility [4,5]. Although the assay genes are re-

* Corresponding author at: Department of Computer Science and Engineering, South Korea.
E-mail address: sunkim.bioinfo@snu.ac.kr (S. Kim).
[1] Equal contribution.

https://doi.org/10.1016/j.csbj.2024.04.038
Received 27 January 2024; Received in revised form 14 April 2024; Accepted 15 April 2024
Available online 22 April 2024

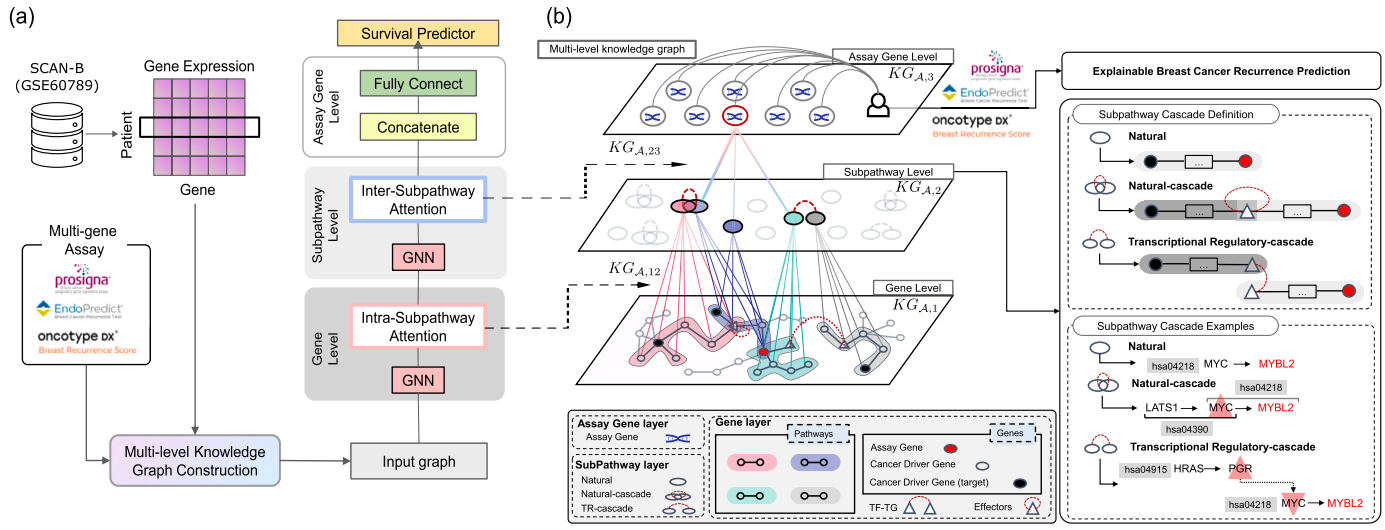

Fig. 1. Overview of the proposed model. (a) Given a gene expression profile vector, the graph neural network predicts the recurrence risk of a given patient by considering the regulatory landscape of assay genes hierarchically. On the multi-layered knowledge graph, our model utilizes two attention-based propagation concepts: *intra-subpathway level* and *inter-subpathway level* propagations. (b) The *multi-layered knowledge graph* consists of three layers: gene-level, subpathway-level, and assay gene-level. Information of gene-level is propagated to assay gene-level through the regulatory subpathways. To capture the complex transcriptional regulatory mechanisms of assay genes, we propose a *subpathway cascade* that identifies all the subpathways that start from cancer driver genes to assay genes while considering both the gene-level and subpathway-level interactions. Examples of the subpathways generated by the method are shown on the right panel.

lated to key mechanisms in cancer cells such as proliferation, apoptosis, ER/PR/HER2 action, invasion [6,7], these marker genes are not informative enough to explain underlying biological mechanisms.

The main question is why different assays, i.e., different gene sets, can predict recurrence or survival. Our research question is "Can we understand the underlying biological mechanisms of these assays for breast cancer recurrence?" If successful, we may be able to improve both the predictive and explanatory power of the assays. Enhancing the explanatory power of the assays, we used transcriptome profiles at the entire cell level to elucidate the underlying biological mechanisms. To use the whole genome level transcriptome data effectively, we need a strategy to address an issue that multiple biological mechanisms and signaling pathways are interdependent due to the multi-step nature of breast cancer recurrence [8–10].

In this study, we constructed a multi-layered knowledge graph with a novel approach, *subpathway cascade*, from curated biological pathway databases. Leveraging the multi-layered knowledge graph, we proposed an interpretable multi-layered graph neural network model featuring attention mechanisms. This model aims to identify the significance of subpathways and predict the risk of recurrence in breast cancer patients. By investigating the three well-known assays (Oncotype DX, Prosigna, EndoPredict) on the SCAN-B dataset [11], we identified shared underlying biological mechanisms including RTK-ERK-ETS-mediated cell proliferation, thyroid hormone-mediated regulation of HIF1A, and JAK-STAT-mediated inflammatory response. With a substantial increase of up to 11% in the c-index of recurrence risk prediction, our model effectively captured essential biological subpathways associated with breast cancers, as demonstrated by robustness to different knowledgebases, survival analysis, and simulation studies. Our contributions are as follows:

- To improve explanatory power of multi-gene assays, we proposed an explainable deep learning model for multi-gene assays of breast cancer recurrence.
- Multi-layered knowledge graphs identified key transcriptional regulatory pathways for three assays: Oncotype DX, Prosigna, and EndoPredict.
- Attention-based graph neural networks improved both the performance of recurrence prediction and the interpretability of regulatory mechanisms at the individual patient level.

2. Materials and methods

In this section, we describe the process of constructing the multi-layered knowledge graph by identifying regulatory subpathways of assay genes, and implementing the explainable graph neural network (Fig. 1).

2.1. Subpathway cascade via transcriptional regulation

To investigate the regulatory subpathways of assay genes (AGs) considering the multi-step nature of recurrence [10], we have developed a novel method named *subpathway cascade*. This approach constructs regulatory subpathways by initiating from cancer driver genes (CDGs) and progressing through the transcriptional regulation by transcription factors (TFs). CDGs are known to be play an important role in development, progression, and even recurrence of cancer [12]. TFs can act as connectors between signaling pathways and other pathways containing target genes (TGs) [13].

2.1.1. Generation schema of the subpathway cascade

The *subpathway cascade* consists of three regulatory layers: (1) *Natural*, (2) *Natural-cascade*, (3) *Transcriptional Regulatory-cascade*. Case (1) represents a single subpathway while cases (2) and (3) assemble two distinct subpathways incorporating a single regulatory mechanism. In *Natural-cascade*, two subpathways share common genes. In *Transcriptional Regulatory-cascade*, TF-TG relationship connects two subpathways. Generating all regulatory subpathways from CDG to AG while considering these cascade cases is inefficient and computationally expensive due to redundant subpathways. To address this challenge, we performed a two-step generating process (Supplementary Figure S1): generating cascading backbones $\mathcal{B}_{\mathcal{A}}$ in Step 1 and subpathway graphs $\mathcal{S}_{\mathcal{A}}$ in Step 2.

Step 1: Generating cascading backbones $\mathcal{B}_{\mathcal{A}}$ The cascading backbones $\mathcal{B}_{\mathcal{A}}$ contain only essential information about subpathways in the form of two sequences (b_g, b_p): b_g is a sequence of hub genes consisting of CDG, cascading mediator genes, and AG in the corresponding subpathway. b_p is a sequence of regulations, such as KEGG pathway identifiers or TF-TG symbol τ between two adjacent hub genes. These

backbones (b_g, b_p) for a pair of i-th CDG and j-th AG, denoted (d_i, a_j), are defined as follows:

Case 1 (Natural): (d_i, a_j) are connected in the same pathway p_x
$b_g = (d_i, a_j)$, $b_p = (p_x)$, where $d_i, a_j \in p_x$

Case 2 (Natural-cascade): two pathways (p_x, p_y) share a gene e
$b_g = (d_i, e, a_j)$, $b_p = (p_x, p_y)$, where $d_i \in p_x, a_j \in p_y$, $e \in p_x$, $e \in p_y$

Case 3 (Transcriptional Regulatory-cascade): TF-TG relationship $\tau = (\alpha, \beta)$ cascades two pathways (p_x, p_y)
$b_g = (d_i, \alpha, \beta, a_j)$, $b_p = (p_x, \tau, p_y)$, where $d_i \in p_x, a_j \in p_y$, $\alpha \in p_x$, $\beta \in p_y$

From the two sequences (b_g, b_p), all the assay-specific cascading backbones $\mathcal{B}_{\mathcal{A}}$ are defined below.

$$\mathcal{B}_{\mathcal{A}} = \bigcup_{(d_i, a_j) \in (\mathcal{D}, \mathcal{A})} \mathcal{B}_{ij} \qquad (1)$$
$$\mathcal{B}_{ij} = \{(b_g, b_p) : \text{cascading backbones of } (d_i, a_j)\}$$

where $\mathcal{D}$ denotes a set of CDGs and $\mathcal{A}$ denotes a set of assay genes.

Step 2: Assay-specific subpathway graphs $\mathcal{S}_{\mathcal{A}}$ Given assay-specific cascading backbones $\mathcal{B}_{\mathcal{A}}$, all regulatory subpathway graphs $\mathcal{S}_{\mathcal{A}}$ are constructed as follows:

$$\mathcal{S}_{\mathcal{A}} = \bigcup_{\mathcal{B}_{ij} \in \mathcal{B}_{\mathcal{A}}} \mathcal{S}_{ij} \text{ where } \mathcal{S}_{ij} = \{s_b : b = (b_g, b_p) \in \mathcal{B}_{ij}\} \qquad (2)$$

$$s_b = \begin{cases} SP(d_i, a_j, p_x) & b \text{ is Case 1,} \\ SP(d_i, e, p_x) \parallel SP(e, a_j, p_y) & b \text{ is Case 2,} \\ SP(d_i, \alpha, p_x) \parallel (\alpha, \beta) \parallel SP(\beta, a_j, p_y) & b \text{ is Case 3} \end{cases} \qquad (3)$$

where s_b is a subpathway graph of the cascading backbone b. $\parallel$ is a concatenation of subpathway sequences. $SP(v_1, v_2, G)$ is a subgraph of the shortest path nodes between (v_1, v_2) in G.

2.1.2. Construction of the assay-specific multi-layered knowledge graph

To construct the regulatory landscape of assays, an assay-specific multi-layered knowledge graph $KG_{\mathcal{A}}$ is constructed from $\mathcal{S}_{\mathcal{A}}$. The knowledge graph consists of three levels: gene, subpathway, and assay in a hierarchy. Based on the types of interacting entities, the graph can be divided into two parts: intra-level and inter-level knowledge graphs.

The gene-level graph $KG_{\mathcal{A},1}$ represents the regulatory mechanism at the gene level. It is constructed from $\mathcal{S}_{\mathcal{A}}$ as a set of nodes V_g representing genes and edges E_g representing regulatory interactions between genes belonging to the subpathways.

The gene-to-subpathway-level graph $KG_{\mathcal{A},12}$ represents the inclusion relationship between genes and subpathways. It is constructed as a set of nodes $V_g \cup V_s$ representing genes and subpathways and edges $E_{g \to s}$ representing the inclusion of genes in subpathways. The edges are constructed through pathway knowledge database.

The subpathway-level graph $KG_{\mathcal{A},2}$ represents the regulatory mechanism at the subpathway level. It is constructed as a set of nodes V_s representing subpathways and edges E_s representing regulatory interactions, named cascades, between subpathways. The edges are constructed using results from subpathway cascades that are either natural cascades or transcriptional regulatory cascades.

The subpathway-to-assay-level graph $KG_{\mathcal{A},23}$ represents the regulatory target relationship from subpathway to assay gene. It is constructed as a set of nodes $V_s \cup V_a$ representing subpathways and assay genes and edges $E_{s \to a}$ representing the regulatory target relationship from subpathway to assay gene.

The assay gene-level graph $KG_{\mathcal{A},3}$ represents the assay genes that are the endpoints of $KG_{\mathcal{A}}$ and are utilized as inputs for recurrence risk prediction.

The intra-subpathway-level knowledge graph is defined as a union of $KG_{\mathcal{A},1}$ and $KG_{\mathcal{A},12}$ that include gene-level interactions and the inclusion relationship between genes and subpathways. The inter-subpathway-level knowledge graph is defined as a union of $KG_{\mathcal{A},2}$, $KG_{\mathcal{A},23}$, and $KG_{\mathcal{A},3}$ that include subpathway-level interactions and the regulatory target relationships from subpathways to assay genes.

2.2. Explainable hierarchical graph neural network for breast cancer recurrence assay

The multi-layered knowledge graph represents the hierarchical regulatory landscape of assay genes. With this hierarchical structure, an *attention-based hierarchical graph neural network* predicts the recurrence risk of patients by considering comprehensive regulatory mechanisms via intra- and inter-subpathway level attention mechanisms.

2.2.1. Hierarchical propagations to generate patient representations

Given a gene expression profile of a patient $\mathbf{u} \in \mathbb{R}^n$ where n is the number of genes in the knowledge graph $KG_{\mathcal{A}}$, the attention-based graph neural network h_θ generates a patient representation $\mathbf{r}_\theta \in \mathbb{R}^d$ through the two propagation schema: Intra- and Inter-subpathway level propagation.

Intra-subpathway level propagation: Formally, given the patient vector $\mathbf{u}$ with the gene-level graph $KG_{\mathcal{A},1}$, we compute the subpathway representation $\mathbf{H}_s$ by using GNNs [14,15] and attention mechanisms [16] as follows:

$$\mathbf{H}_g = \mathrm{GNN}_\theta(\mathbf{u}, KG_{\mathcal{A},1})$$
$$\mathbf{H}_s = \mathrm{Att}(\mathbf{H}_g, KG_{\mathcal{A},12}) \qquad (4)$$

where $\mathbf{H}_g$ is the representation of genes in $KG_{\mathcal{A},1}$ and contains information of gene interactions via GNN with model parameters θ. $\mathrm{Att}(\mathbf{H}, G) = \omega_G(W^T \tanh(W^{(1)}\mathbf{H} + b^{(1)}) + b)\mathbf{H}^T$ where $\omega_G()$ is a softmax function that computes attention weights of genes belonging to a single subpathway. $W^{(1)} \in \mathbb{R}^{d_h \times d_e}$ and $W \in \mathbb{R}^{d_h}$ are the learnable parameters for calculating attention weights where d_e is the embedding size of genes in $\mathbf{H}_g$ and d_h is the latent embedding size for calculation of the attention scores. $b^{(1)}$ and b are the biases.

Inter-subpathway level propagation: While the intra-subpathway level propagation aggregates the information of genes belonging to each subpathway, this does not account for information exchange due to cascading between subpathways. To leverage the subpathway cascade, we performed an inter-subpathway level propagation as follows:

$$\mathbf{H}_{sc} = \mathrm{GNN}_\theta(\mathbf{H}_s, KG_{\mathcal{A},2})$$
$$\mathbf{H}_{\mathcal{A}} = \mathrm{Att}(\mathbf{H}_{sc}, KG_{\mathcal{A},23}) \qquad (5)$$
$$\mathbf{r} = \mathrm{CONCAT}(\mathbf{H}_{\mathcal{A}})$$

where $\mathbf{H}_{sc}$ is the representation vector of subpathway cascade. $\mathbf{H}_{\mathcal{A}}$ is the representation vector of assay genes that is obtained from the attention mechanism via considering importance of the subpathways. CONCAT is a concatenate operation for embeddings of assay genes.

2.2.2. Recurrence risk prediction through survival predictor

Using the patient representation vector $\mathbf{r}_\theta$ generated by h_θ, we apply a downstream MLP survival predictor, denoted as g_ϕ, to measure the patient's risk of breast cancer recurrence in an end-to-end manner. Thus, the hierarchical graph neural network and the survival predictor are co-trained using the average negative log partial likelihood as the training objective $l(\theta, \phi)$ that is a common choice for survival prediction tasks.

$$l(\theta, \phi) = -\frac{1}{N_E} \sum_{i:\, E_i = 1} \left[g_\phi(r_{\theta,i}) - \log \sum_{j \in \mathbf{R}(T_i)} \exp(g_\phi(r_{\theta,j})) \right] \qquad (6)$$

where N_E is the number of recurrences. E_i is a recurrence indicator for i-th patient and $\mathbf{R}(t) = \{j : T_j \geq t\}$ is the set of patients who are at recurrence risk at time t.

2.3. Knowledge databases for transcriptional regulation

In order to generate and evaluate our hierarchical graph neural network for predicting breast cancer recurrence risks, we utilized several databases. From the **KEGG PATHWAY Database** [17], we selected 210 pathways, excluding disease or metabolic pathways, to identify potential subpathways relevant to recurrence development. **TFLink** is a database of transcription factor-target gene regulatory relationships that act as cascading mediators between different subpathways [18]. We collected 839 TFs and 11,797 TGs. **COSMIC Cancer Census** is a database of cancer driver genes that act as the initiating source of regulatory subpathways [12]. Among the cancer hallmark genes, 114 breast cancer related genes were finally selected as starting points to construct the multi-level knowledge graph.

2.4. Experiment settings

Breast cancer recurrence assays are designed to help determine a treatment strategy, such as treatment of additional chemotherapy for high-risk groups, by predicting the future 5-year or 10-year recurrence risk when only endocrine therapy is treated to patients. Therefore, we conducted recurrence risk prediction experiments using SCAN-B dataset (GSE60789) [11], which was further filtered to include early-stage breast cancer samples with survival information and belonging to either the None or Endo (endocrine therapy) treatment groups. The filtered experimental dataset of SCAN-B comprises 3,661 samples. Among them, 327 patients experienced recurrence. Within this group, 228 patients had recurrence within 5 years while 96 experienced recurrence between 5 and 10 years. Additionally, 788 patients were right-censored within 5 years without recurrence and 2,350 were right-censored between 5 and 10 years without experiencing recurrence.

In order to include cancer-related gene information, we constructed one-hot encoding vector for each gene from the 50 hallmark gene set of MSigDB [19,20]. Then, the input features were constructed by concatenating the expression value and one-hot encoding vector for each gene. The maximum training epochs were limited to 200. The performance evaluation experiment was conducted as 10-fold cross validation with random split of train:validation:test = 8:1:1 while holding the hyperparameter settings of the baseline and our model identically. Hyperparameter tuning was performed using Optuna [21], a python library for hyperparameter optimization, and detailed tuning parameters are provided in Table S3.

As baseline models for recurrence risk prediction, we utilized three ML-based methods, one deep-learning (DL) method, and two pathway-based (PB) methods. The list of additional baseline models is outlined below.

- ML Fast Survival SVM (FS_SVM) [22]: Efficient linear survival support vector machine.
- ML GB_Tree [23,24]: Gradient-boosted Cox proportional hazard loss with regression trees as base learner.
- ML Coxnet [25]: Cox's proportional hazard's model with elastic net penalty.
- DL SALMON_E: A modified version of SALMON [26] utilizing only gene expression data.
- PB ssGSEA [27]: A single sample gene set enrichment analysis.
- PB SAS [28,29]: Subsystem activation score for measuring pathway activity using protein-protein interaction network.

Since our multi-level knowledge graph contains fewer than one thousand genes, ML and DL models utilized the top 1,000 genes with large variance as input features. PB models computed input features using the same pathways for our multi-level knowledge graph construction and utilized the three ML models as predictors.

Table 1
C-index of breast cancer risk prediction on SCAN-B dataset with three multi-gene assays. The best results are highlighted in bold.

Assay	Assay Genes (Baseline)	Intra-Subpathway	Inter-Subpathway
Oncotype DX	0.714 ± 0.044	0.606 ± 0.043	$\mathbf{0.741 \pm 0.057}$
Prosigna	0.691 ± 0.058	0.616 ± 0.026	$\mathbf{0.764 \pm 0.043}$
EndoPredict	0.708 ± 0.076	0.635 ± 0.058	$\mathbf{0.742 \pm 0.046}$

3. Results and discussion

3.1. Multi-layered knowledge graph identifies shared regulatory mechanisms of breast cancer recurrence

In this section, utilizing the subpathway cascade, the regulatory landscapes of assays are illustrated in Fig. 2 and Supplementary Figure S2 and S3 for Oncotype DX, EndoPredict, and Prosigna, respectively. In Fig. 2, our subpathway cascade method revealed important biological mechanisms: cellular senescence, apoptosis, and focal adhesion. This is consistent with established knowledge that disrupted signaling pathways affect breast cancer processes. Notably, these diverse regulations were not captured without cascading cases (Supplementary Figure S4-6).

By investigating the regulatory landscape of the three assays, we identified shared underlying biological mechanisms such as RTK-ERK-ETS-mediated cell proliferation, thyroid hormone-mediated regulation of HIF1A, and JAK-STAT-mediated inflammatory response, despite minimal gene overlap between assays. For example, Oncotype DX as a reference, we observed 38.10% gene overlap with Prosigna while the regulatory subpathways exhibited a large overlap of 98.75% (Supplementary Table S1). The most common subpathways were signaling pathways (Supplementary Table S2), acting as important mechanisms in breast cancer recurrence [30,9].

3.2. Multi-layered knowledge graph helps recurrence risk prediction

We evaluated how well the multi-layered knowledge graph predicts breast cancer recurrence. We compared predictive outcomes from three assays: using MLP with assay gene profiles alone (Assay Genes), GNN with only intra-subpathway graph (Intra-Subpathway), and GNN with the multi-layered graph including inter-subpathway connections (Inter-Subpathway). Notably, as shown in Table 1, Inter-Subpathway significantly outperformed the baseline, with up to an 11% increase in c-index, highlighting the importance of inter-subpathway interactions in comprehending breast cancer recurrence mechanisms. In addition, Inter-Subpathway model shows good performance of 5-year recurrence prediction on all the three assays (Supplementary Figure S7). Interestingly, the 'Intra-Subpathway' model exhibits the worst predictive performances across all assays. Without endpoints (i.e., assay genes) to summarize information from learned subpathways, this model faces challenges in integrating data from multiple subpathways, leading to overfitting.

To further explore the predictive performance of a multi-layered knowledge graph, we conducted an additional comparison involving three ML-based methods (FS_SVM, GB_Tree, and Coxnet), one deep-learning (DL) method (SALMON_E), and two pathway-based (PB) methods (ssGSEA and SAS). As shown in Supplementary Table S4, our proposed method leveraging the multi-layered knowledge graph demonstrates better predictive performance compared to most ML, DL, and PB models. While DL models like SALMON_E and ssGSEA + Coxnet demonstrate strong recurrence risk prediction among baseline models, they do not consider assay genes and suffer from limited interpretability. In

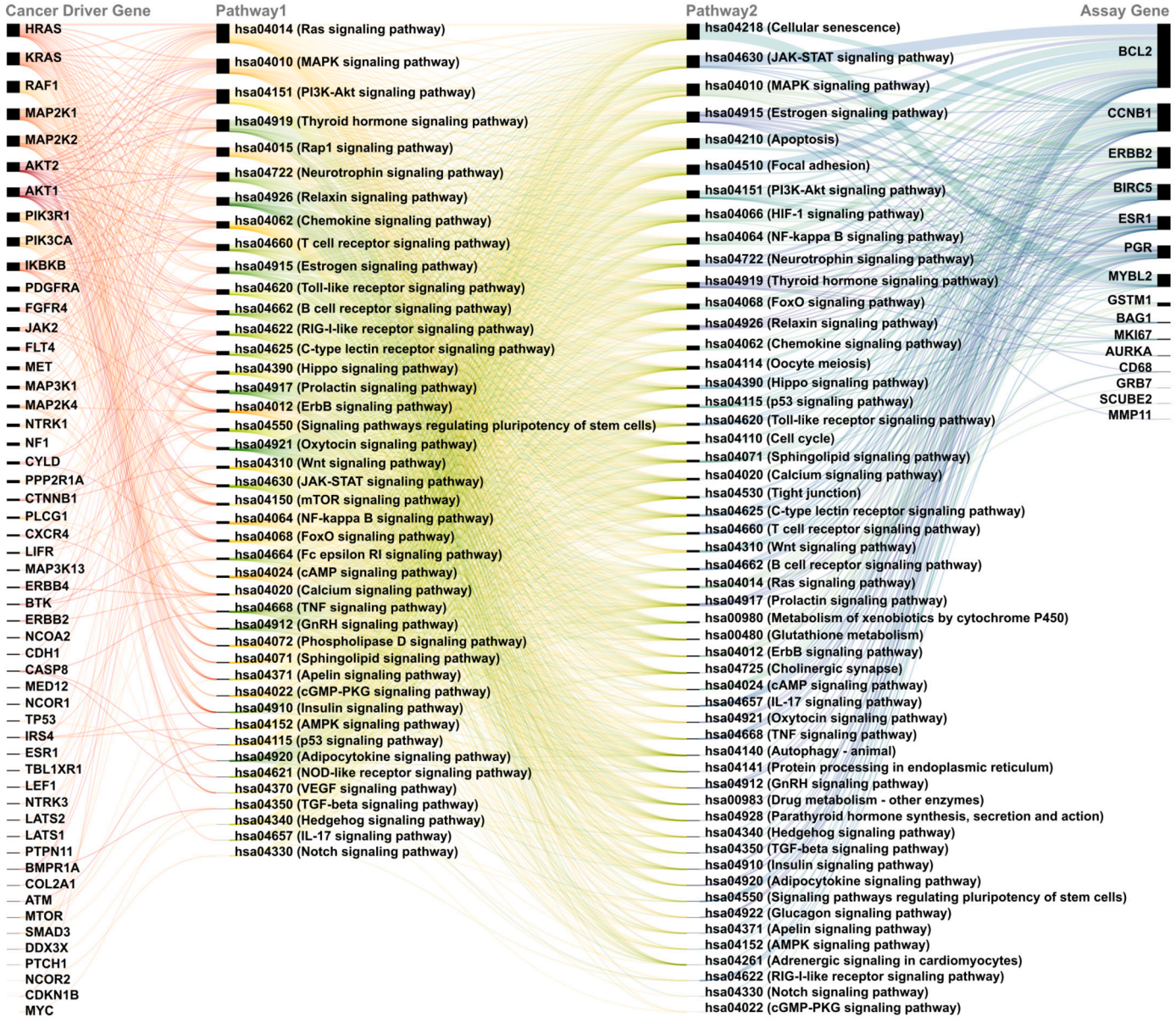

Fig. 2. The regulatory landscape of Oncotype DX via the subpathway cascade. Each combination of lines reaching the Assay Gene (AG) from the Cancer Driver Gene (CDG) represents a single regulatory subpathway. The thickness of the node is calculated based on the number of subpathways passing through the node representing the relative importance of the node to construct the entire landscape in its own column. The line color was determined according to the alphabetical order of each column.

contrast, our proposed method not only achieves the best prediction accuracy but also offers biological explanation in terms of the regulatory mechanisms of assay genes.

3.3. Multi-layered knowledge graph is robust with respect to the different choices of knowledgebase and cancer driver gene sets

To demonstrate the robustness of our proposed approach with respect to the knowledgebases, we collected pathways from Reactome [31] and WikiPathways [32]. To ensure a fair comparison, we excluded disease or metabolome-related pathways from both Reactome and WikiPathways employing the same filtering criteria utilized for the KEGG database.

Table 2 shows that Oncotype DX and EndoPredict maintain consistent performance across different knowledgebases. Although Prosigna performs similarly with KEGG and WikiPathways, its performance slightly decreased with Reactome. Nonetheless, the proposed approach with Prosigna consistently demonstrated superior predictive performance compared to other baseline models.

Table 2
C-index of breast cancer risk prediction on SCAN-B dataset with different pathway databases. The best results are highlighted in bold.

Assay	KEGG	WikiPathways	Reactome
Oncotype DX	0.741 ± 0.057	**0.743 ± 0.049**	0.740 ± 0.040
Prosigna	**0.764 ± 0.043**	**0.764 ± 0.038**	0.756 ± 0.040
EndoPredict	0.742 ± 0.046	**0.743 ± 0.047**	0.742 ± 0.036

In addition, top-highlighted pathways captured by attention analysis highlight biological mechanisms consistently regardless of knowledgebases used for the analysis. Supplementary Figure S8 show the top 20 subpathways for Oncotype DX from KEGG, WikiPathways, and Reactome. In all cases of knowledgebases used for constructing the hierarchical graph model, the 20 subpathways included the Ras and Estrogen

Table 3
C-index of breast cancer risk prediction on SCAN-B dataset with different cancer driver genes. The best results are highlighted in bold. CCHG: Cancer Census Hallmark Gene. BC-CCHG: Breast Cancer realted CCHG.

Assay	BC-CCHG	CCHG
Oncotype DX	**0.741 ± 0.057**	**0.741 ± 0.042**
Prosigna	**0.764 ± 0.043**	0.759 ± 0.047
EndoPredict	**0.742 ± 0.046**	0.736 ± 0.044

Table 4
C-index of breast cancer risk prediction on SCAN-B dataset with 'Merged Assays'. 'Merged Assays' is an union of the three multi-gene assays. The best results are highlighted in bold.

	C-index		C-index
Oncotype DX	0.741 ± 0.057		
Prosigna	0.764 ± 0.043	Merged Assays	**0.771 ± 0.045**
EndoPredict	0.742 ± 0.046		

signaling pathways were commonly identified as well as the PI3K-AKT and MAPK pathways. Similar trends were observed with Prosigna and EndoPredict as well (Supplementary Figure S9-S10). This implies that the proposed multi-level knowledge graph can robustly identify the regulatory mechanisms of assay genes, which could serve as a crucial foundation for exploring the biological mechanisms of breast cancer recurrence.

We further conducted an experiment with a different set of cancer driver genes that is the starting point of our deep learning analysis. The proposed method utilized only breast cancer-related cancer census hallmark genes (BC-CCHG). Instead, we trained an additional model using the entire set of Cancer Census Hallmark Genes (CCHG) as cancer driver genes.

Table 3 demonstrates that predictive performances are either comparable or slightly worse. Concerning cancer driver genes, out of a total of 349 CCHG, 114 genes are associated with BC-CCHG. While augmenting the number of cancer driver genes might enhance connectivity with assay genes, it could also introduce unnecessary complexity as the multi-level knowledge graph expands. The use of CCHG resulted in an average increase of 8.2% in the number of nodes in the gene-level network and 36.6% in the subpathway-level network. Thus, it seems that the model's performance slightly decreases when utilizing CCHG.

Therefore, when constructing a multi-level knowledge graph, it is crucial to curate regulatory subpathways that can connect cancer driver genes and assay genes to build an efficient multi-level knowledge graph. One possible approach to address this is to utilize three multi-gene assays together (called as 'Merged Assays') when constructing the knowledge graph. As shown in Table 4, the 'Merged Assays' achieved a c-index of 0.771, indicating a slight enhancement in recurrence risk prediction performance. The improved prediction performance when all genes in the three multi-gene assays is probably because the assay genes are better linked to regulatory subpathways that play crucial roles in breast cancer recurrence such as the PI3K-AKT signaling, JAK-STAT signaling, or cytokine-cytokine receptor interaction [33–35].

3.4. Attention mechanism captures regulatory subpathways of breast cancer recurrence

To demonstrate the explainability of our model, we analyzed attention weights and identified significant subpathways of three assays: Oncotype DX (Fig. 3), EndoPredict (Supplementary Figure S11), and Prosigna (Supplementary Figure S12).

Fig. 3(a) illustrates the top 20 regulatory subpathways of Oncotype DX using a Sankey diagram, and their assay-level regulations are summarized in 3(b). Among these top 20 subpathways, some exhibit notable attention weights (Fig. 3(c)), especially subpathways from Ras (hsa04014), estrogen (hsa04915), and PI3K-Akt (hsa04151) signaling pathways. These signaling pathways are known to play an important role in regulation of breast cancer cells [36,37,33]. To further investigate the biological relevance of the identified subpathways in terms of breast cancer recurrence, we analyzed the two highlighted subpathways: hsa04014:*RTK→ETS* and hsa04915:*RAS→ER+TF*, which are commonly identified Oncotype DX, Prosigna, and EndoPredict.

Ras signaling subpathway (hsa04014:RTK→ETS): Starting with *RTK* family, this subpathway propagates transcriptional signaling via the *ETS1* transcription factor, finally regulating some Oncotype DX assay genes including *CCNB1*. Fig. 3(d) illustrates strong correlations between cancer driver genes *MET* and *PDGFRA* within the *RTK* family, and their association with *ETS1* and *CCNB1*. Both *MET* and *PDGFRA*, which are strongly correlated with *ETS1*, have proliferative signaling hallmarks [12,38]. Interestingly, *CCNB1*, a proliferation activity marker in the Oncotype DX assay by its overexpression [39], is negatively correlated with *ETS1*, a well-known inhibitor of cell growth and proliferation in breast cancer cells [40].

Estrogen signaling subpathway (hsa04915:RAS→ER+TF): Starting from Ras family, cancer driver genes (*HRAS*, *KRAS*), this subpathway regulates assay genes including *BCL2*, an indicative of increased estrogen signaling in Oncotype DX [41]. Specifically, *BCL2* is regulated by transcription factors including *ESR1*, another assay gene, and *SP1*. Similar with Ras signaling pathway, our correlation analysis shows that cancer driver genes, transcription factors and the assay gene are strongly correlated sequentially (Fig. 3(e)). With apoptotic behaviors of *BCL2* and the Ras family [42,43], we conjecture that the subpathway regulates both estrogen receptor activity and cell apoptosis in breast cancer cells.

3.5. Attention mechanisms can be interpreted with predicted risk

We further examined the association between assay genes and recurrence risk using the assay gene-level layer of the multi-layered knowledge graph. Utilizing the Layer Conductance method [44], we measured neuron attributions in the assay gene-level layer related to risk prediction. High neuron attribution indicates that neuron activation positively impacts higher recurrence risk. Conversely, low or negative attribution suggests that neuron activation hinders risk prediction, resulting in predictions of lower recurrence risk.

Fig. 4(a) shows neuron attributions of assay genes in Oncotype DX, measured on test data. We divided the test data into High/Low groups based on median recurrence risk, and average of neuron attributions were summarized for each group. Among the 15 genes of Oncotype DX, neuronal attributions of 12 genes except for PGR, MYBL2, and CD68 show marked differences between the High and Low groups. On the other hand, in Fig. 4(b), only 4 genes out of 15 genes show significant differences in the expression level fold change. This indicates that assay genes may not be determined only by using differentially expressed genes (DEGs) based on the fold change in gene expression. Thus, it is necessary to consider gene-gene interaction to determine assay genes for recurrence risk. Our model can be used to identify such genes for recurrence risk by looking at activation levels of the neuron for risk prediction. The neuron attribution is calculated by considering the activation value of the neuron based on the activation values of other genes. Thus, the neuron attribution can consider gene interactions. For this reason, neuron attribution shows more clinical relevance than the simple fold change metric (Fig. 4 (c)-(d)). Therefore, the graph neural network utilizing the multi-layered knowledge graph better captured the impact of assay genes on recurrence risk.

To further analyze regulatory biological functions of assay genes, we selected *BCL2* due to substantial differences in neuron attributions. We performed pathway enrichment analysis using genes in the top 10 subpathways linked to BCL2 for each group, respectively (Fig. 4(e)

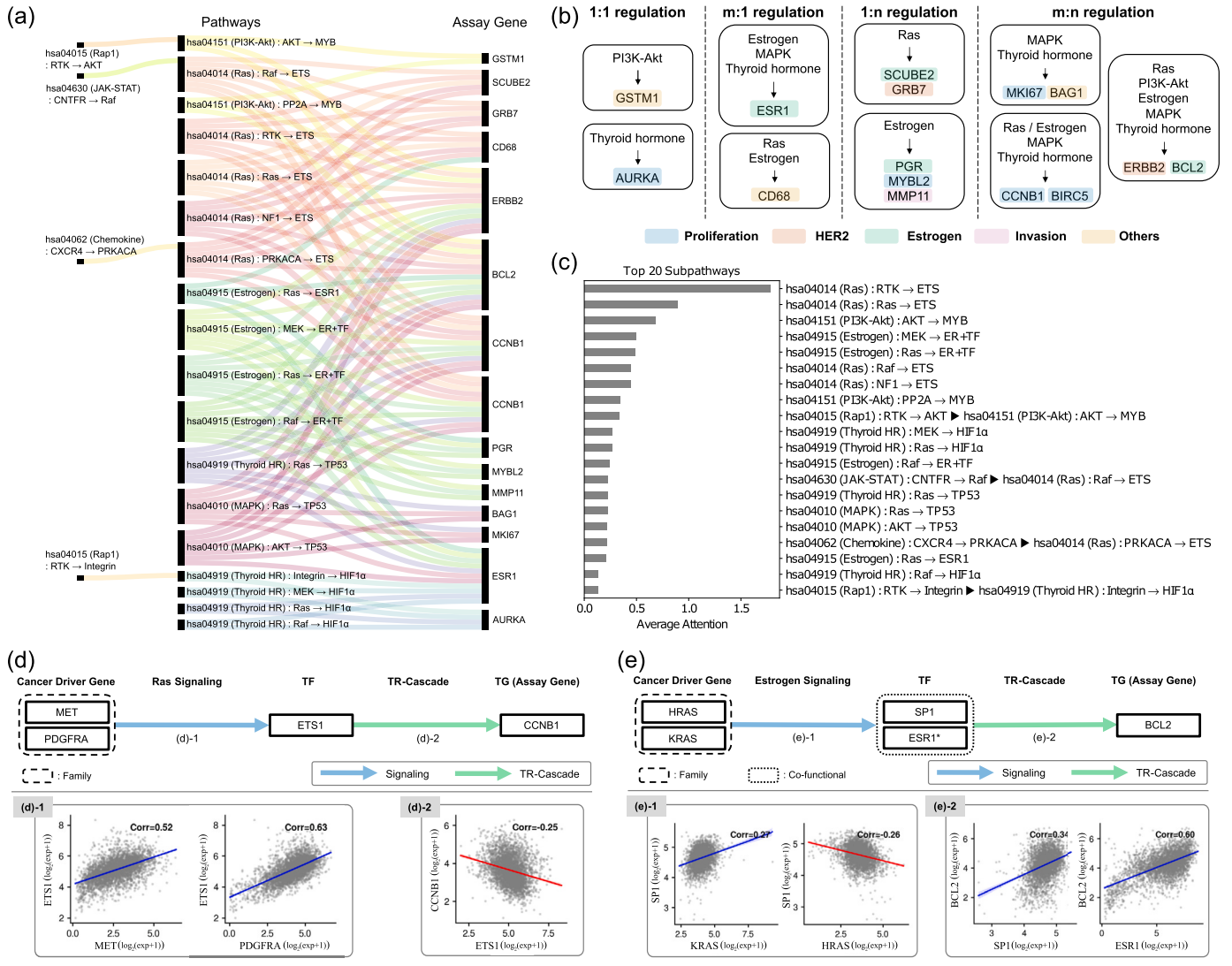

Fig. 3. Subpathways obtained by attention mechanisms in Oncotype DX. (a) The Sankey diagram of top-20 subpathways. Various known cancer-related signaling pathways such as Ras, Estrogen, and PI3K-Akt were identified. (b) Assay gene-level regulation cases of the top-20 subpathways. 5 categories of Oncotype DX markers are provided together. (c) A bar plot representing the top-20 subpathways in order of average attentions. (d), (e) Two case studies of Ras and estrogen signaling subpathways, hsa04014:*RTK→ETS* and hsa04915:*Ras→ER+TF*, respectively. Cancer driver genes, transcription factors, and target assay genes were strongly correlated sequentially.

and (f)). The enriched pathways in the two groups differ significantly. In Fig. 4(e), pathways related to the development and progression of breast cancer cells (e.g., Ras and ErbB signaling) were enriched in the High group. On the other hand, in the Low group, more general-disease related pathways were enriched (Fig. 4(f)). Also, in the Low group, breast cancer-related pathways identified in the High group are also rarely enriched. This result suggests that distinct regulatory subpathways of *BCL2* contributed to recurrence risk. Therefore, our model facilitates the interpretation of associations between assay genes and recurrence risks by suggesting how assay genes are regulated by specific biological mechanisms.

3.6. Survival and simulation studies suggest the clinical potentials of top-3 regulatory subpathways

We investigated the clinical impacts of top-3 regulatory subpathways in Fig. 3(c): hsa04014(Ras):*RTK→ETS*, hsa04151(PI3K-Akt):*AKT→MYB*, and hsa04915(Estrogen):Ras→*ER+TF*. First, we divided test samples into high and low-risk groups based on the expression levels of genes associated with these subpathways. In Fig. 5(a), the subpathways are up-regulated in the high-risk group. Interestingly, overexpression of *CCNB1*, *MYBL2*, and *ESR1*, promoting cell proliferation, can lead to higher recurrence risks [39,45]. To further examine the prognostic potential of the top-3 subpathways, we conducted survival analyses and a simulation study assuming a targeted therapy situation. Details of experimental setup are described in Supplementary Material.

To evaluate the prognostic values of the top-3 subpathways, we performed survival analyses. Since breast cancer recurrence assays are used to predict 5-year or 10-year recurrence risks, we limited the observation period up to 10 years and selected 342 test samples accordingly. We used Kaplan-Meier estimation and log rank test to compare the risks between two groups with three survival outcomes: Overall Survival, Recurrence-Free Intervals, Distant Recurrence-Free Intervals. As shown in Fig. 5(b), these top-3 subpathways were significantly related to all three survival outcomes. Corresponding results for Prosigna and EndoPredict are provided in Supplementary Figure S13.

To further investigate the clinical potential of the top-3 subpathways, we conducted a simulation study to observe changes in predicted risks. As shown in Fig. 5(c), we replaced the gene expression values of

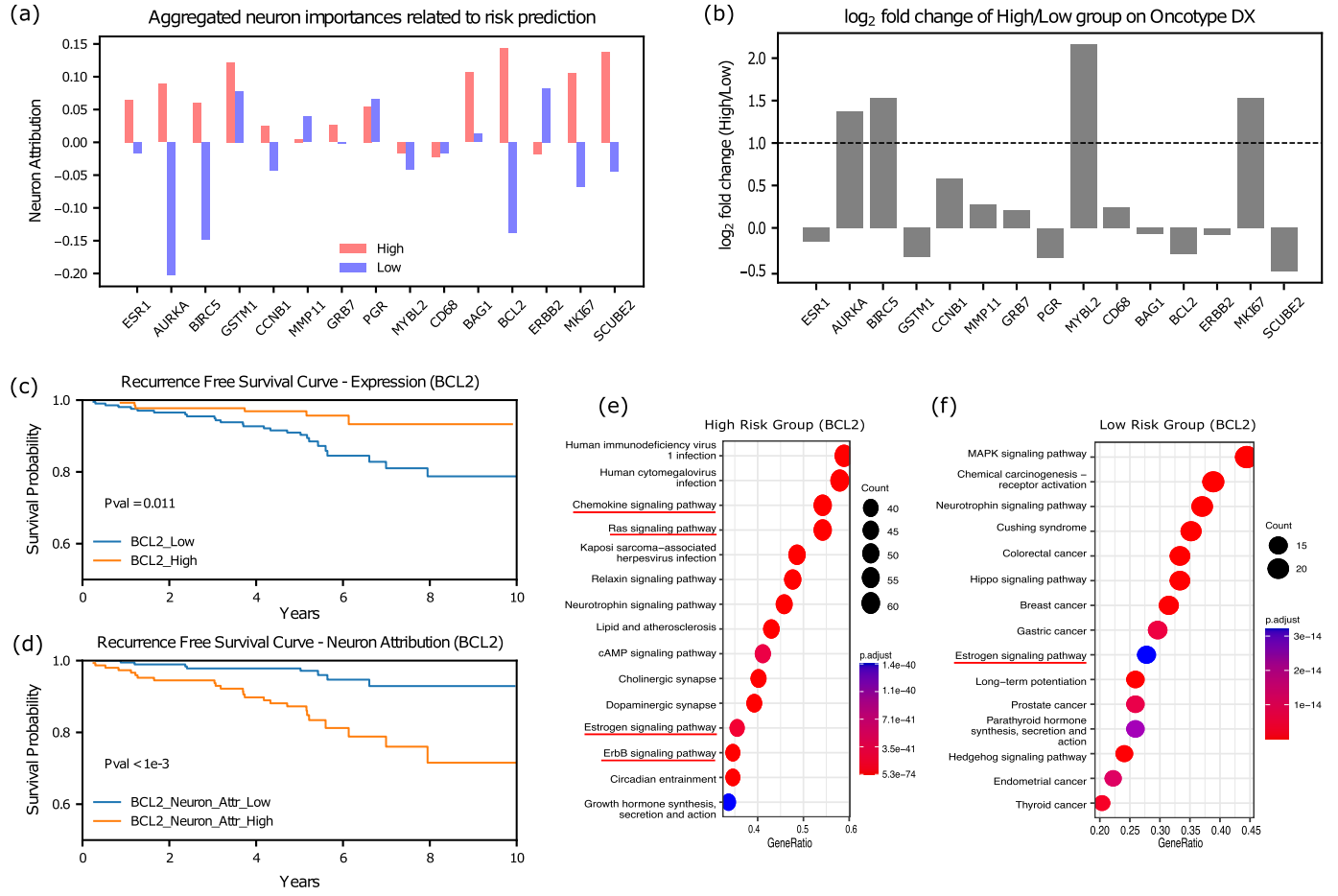

Fig. 4. Impact of assay genes in Oncotype DX for recurrence risk prediction. (a) Relation of neuron activation in assay gene-level with recurrence risk. Attribution of neuron is measured by Layer Conductance. High/Low group is divided by median risk values of test data. As the activation value of the neuron increases, the recurrence risk increases whereas negative neuron attribution means that as the activation value of the neuron increases, the recurrence risk decreases. (b) Log2 fold change (log2FC) of genes in Oncotype DX. The values are calculated as log2FC(High/Low) and averaged on the test data. The black dashed line denotes log2FC is 1. (c) Recurrence free survival outcomes of breast cancer patients with high or low expression levels of BCL2. The high/low groups are determined by the mean gene expression values of BCL2 in the test data. (d) Recurrence free survival outcomes of breast cancer patients with high or low neuron attribution levels of BLC2. The high/low groups are determined by the mean neuron attribution levels of BCL2 in the test data. (e-f) Pathway enrichment results of top 10 highlighted subpathways of *BCL2* for (e) High and (f) Low group, respectively. Pathways associated with breast cancer are highlighted with red underlines.

top-3 subpathways in high-risk samples with the mean values of low-risk group and measured the difference in predicted recurrence risks. Fig. 5(d) shows that up to 45.7% of high-risk samples turned into the low-risk group. Notably, the more subpathways were replaced, the more the rate of changes increased.

4. Conclusion

In this study, we presented a novel multi-layered knowledge graph neural network model that improves both the predictive and explanatory power of three widely used recurrence assays for breast cancer. *Subpathway cascade* identified potential subpathways that are likely to regulate the transcriptomic states of assay genes. Then, the multi-layered knowledge graph was constructed to represent the regulatory landscape of assay genes. By leveraging the multi-layered knowledge graph, our hierarchical graph neural network not only improved risk prediction performance, but also provided explainability through intra- and inter-subpathway attentions. Our key finding is that the three assays, despite using different sets of genes, share common underlying biological mechanisms, such as RTK-ERK-ETS-mediated cell proliferation. Our results would suggest that targeting these key subpathways may have therapeutic potential.

CRediT authorship contribution statement

Sangseon Lee: Conceptualization, Investigation, Writing – original draft, Writing – review & editing. **Joonhyeong Park:** Conceptualization, Investigation, Writing – original draft, Writing – review & editing. **Yinhua Piao:** Investigation, Writing – review & editing. **Dohoon Lee:** Investigation, Writing – review & editing. **Danyeong Lee:** Investigation, Writing – review & editing. **Sun Kim:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2021R1A6A3A01086898), by the National Research Foundation of Korea (NRF) grant funded by the Ko-

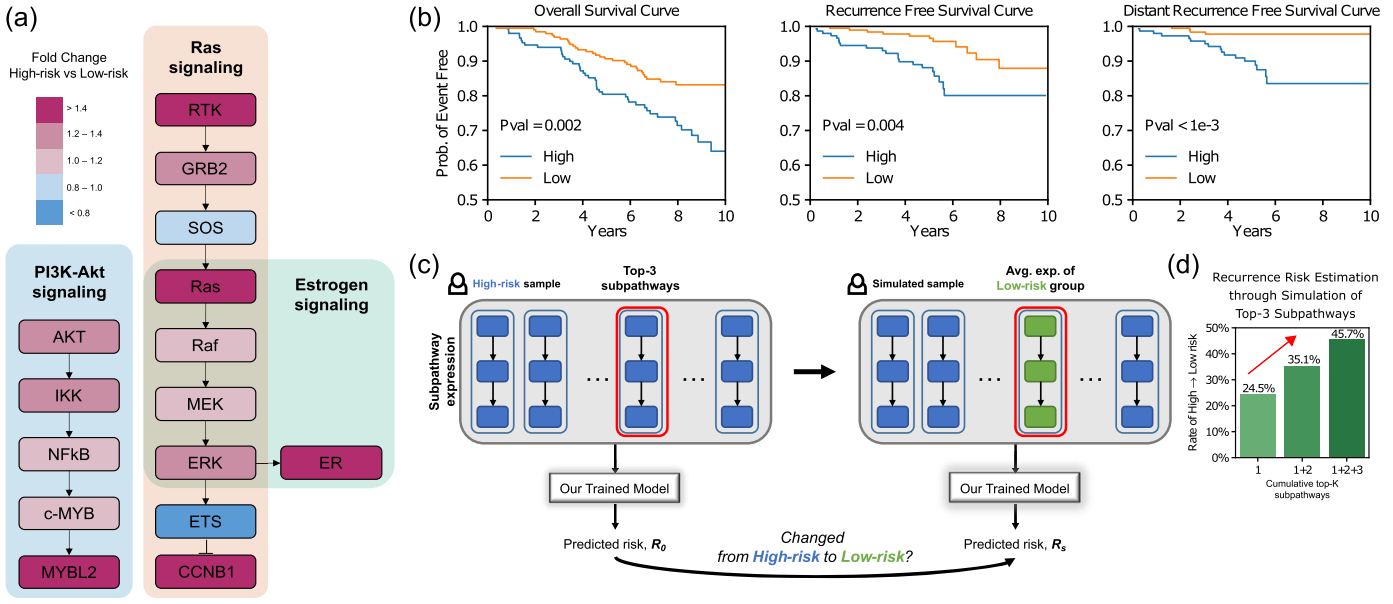

Fig. 5. Clinical potential of top-3 subpathways in Oncotype DX. (a) A subnetwork of top-3 subpathways consists of Ras, PI3K-Akt, and Estrogen signaling subpathways. High-risk (149 test samples) and low-risk (193 test samples) groups were classified from the expression levels of genes belonging to the subnetwork. Fold change values are also indicated. (b) Three 10-year survival curves for high and low risk groups. Significant log rank p-values were observed in all. (c) A schema of simulation study for recurrence risk estimation. (d) Result of simulation study for the high-risk group. Changes in predicted risks were observed by replacing the expression level of genes within the top-3 subpathways as the average of the low-risk group.

rea government (MSIT) (No. NRF-2023R1A2C2006953), by Research Program for Agricultural Science & Technology Development through National Institute of Agricultural Sciences, Rural Development Administration (RS-2023-00231699), by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [NO. 2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University)].

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csbj.2024.04.038.

References

[1] Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 2004;351:2817–26.
[2] Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol 2009;27:1160.
[3] Filipits M, Rudas M, Jakesz R, Dubsky P, Fitzal F, Singer CF, et al. A new molecular predictor of distant recurrence in ER-positive, HER2-negative breast cancer adds independent information to conventional clinical risk factors. Clin Cancer Res 2011;17:6012–20.
[4] Byron SA, Keuren-Jensen V, Kendall R, Engelthaler DM, Carpten JD, Craig DW. Translating rna sequencing into clinical diagnostics: opportunities and challenges. Nat Rev Genet 2016;17:257–71.
[5] Lee H-B, Lee SB, Kim M, Kwon S, Jo J, Kim J, et al. Development and validation of a next-generation sequencing–based multigene assay to predict the prognosis of estrogen receptor–positive, her2-negative breast cancer. Clin Cancer Res 2020;26:6513–22.
[6] Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. J Clin Invest 2011;121:2750–67.
[7] Varga Z, Lebeau A, Bu H, Hartmann A, Penault-Llorca F, Guerini-Rocco E, et al. An international reproducibility study validating quantitative determination of erbb2, esr1, pgr, and mki67 mrna in breast cancer using mammatyper®. Breast Cancer Res 2017;19:1–13.
[8] Blanco MA, Kang Y. Signaling pathways in breast cancer metastasis-novel insights from functional genomics. Breast Cancer Res 2011;13:1–9.
[9] Fares J, Fares MY, Khachfe HH, Salhab HA, Fares Y. Molecular principles of metastasis: a hallmark of cancer revisited. Signal Transduct Targeted Ther 2020;5:1–17.
[10] Gui P, Bivona TG. Evolution of metastasis: new tools and insights. Trends Cancer 2022;8:98–109.
[11] Saal LH, Vallon-Christersson J, Häkkinen J, Hegardt C, Grabau D, Winter C, et al. The Sweden cancerome analysis network-breast (scan-b) initiative: a large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine. Gen Med 2015;7:1–12.
[12] Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The cosmic cancer gene census: describing genetic dysfunction across all human cancers. Nat Rev Cancer 2018;18:696–705.
[13] Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, et al. The human transcription factors. Cell 2018;172:650–65.
[14] Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. In: International conference on learning representations; 2017. Available from:. https://openreview.net/forum?id=SJU4ayYgl.
[15] Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph attention networks. In: International conference on learning representations; 2018.
[16] Shen T, Zhou T, Long G, Jiang J, Pan S, Zhang C. Disan: directional self-attention network for rnn/cnn-free language understanding. In: Proceedings of the AAAI conference on artificial intelligence, vol. 32. 2018.
[17] Kanehisa M, Goto S. Kegg: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28:27–30.
[18] Liska O, Bohár B, Hidas A, Korcsmáros T, Papp B, Fazekas D, et al. Tflink: an integrated gateway to access transcription factor–target gene interactions for multiple species. Database 2022:2022.
[19] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci 2005;102:15545–50.
[20] Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (msigdb) 3.0. Bioinformatics 2011;27:1739–40.
[21] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: a next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining; 2019. p. 2623–31.
[22] Pölsterl S, Navab N, Katouzian A. Fast training of support vector machines for survival analysis. In: Machine learning and knowledge discovery in databases: European conference, ECML PKDD 2015. Springer; 2015. p. 243–59.
[23] Friedman JH. Stochastic gradient boosting. Comput Stat Data Anal 2002;38:367–78.
[24] Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests; 2008.
[25] Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. J Stat Softw 2011;39:1.
[26] Huang Z, Zhan X, Xiang S, Johnson TS, Helm B, Yu CY, et al. Salmon: survival analysis learning with multi-omics neural networks on breast cancer. Front Genet 2019;10:166.
[27] Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, et al. Systematic rna interference reveals that oncogenic kras-driven cancers require tbk1. Nature 2009;462:108–12.

[28] Lim S, Park Y, Hur B, Kim M, Han W, Kim S. Protein interaction network (pin)-based breast cancer subsystem identification and activation measurement for prognostic modeling. Methods 2016;110:81–9.
[29] Lim S, Lee S, Jung I, Rhee S, Kim S. Comprehensive and critical evaluation of individualized pathway activity measurement tools on pan-cancer data. Brief Bioinform 2020;21:36–46.
[30] Rankin EB, Nam J-M, Giaccia AJ. Hypoxia: signaling the metastatic cascade. Trends Cancer 2016;2:295–304.
[31] Milacic M, Beavers D, Conley P, Gong C, Gillespie M, Griss J, et al. The reactome pathway knowledgebase 2024. Nucleic Acids Res 2024;52:D672–8.
[32] Agrawal A, Balcı H, Hanspers K, Coort SL, Martens M, Slenter DN, et al. Wikipathways 2024: next generation pathway database. Nucleic Acids Res 2024;52:D679–89.
[33] Miricescu D, Totan A, Stanescu-Spinu I-I, Badoiu SC, Stefani C, Greabu M. Pi3k/akt/mtor signaling pathway in breast cancer: from molecular landscape to clinical aspects. Int J Mol Sci 2020;22:173.
[34] Manore SG, Doheny DL, Wong GL, Lo H-W. Il-6/jak/stat3 signaling in breast cancer metastasis: biology and treatment. Front Oncol 2022;12:866014.
[35] Martínez-Pérez C, Leung J, Kay C, Meehan J, Gray M, Dixon JM, et al. The signal transducer il6st (gp130) as a predictive and prognostic biomarker in breast cancer. J Person Med 2021;11:618.
[36] Eckert LB, Repasky GA, Ulkü AS, McFall A, Zhou H, Sartor CI, et al. Involvement of ras activation in human breast cancer cell signaling, invasion, and anoikis. Cancer Res 2004;64:4585–92.
[37] Hah N, Danko CG, Core L, Waterfall JJ, Siepel A, Lis JT, et al. A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. Cell 2011;145:622–34.
[38] Qu W, Chen X, Wang J, Lv J, Yan D. Microrna-1 inhibits ovarian cancer cell proliferation and migration through c-met pathway. Clin Chim Acta 2017;473:237–44.
[39] Fang Y, Yu H, Liang X, Xu J, Cai X. Chk1-induced ccnb1 overexpression promotes cell proliferation and tumor growth in human colorectal cancer. Cancer Biol Ther 2014;15:1268–79.
[40] Kim G-C, Lee C-G, Verma R, Rudra D, Kim T, Kang K, et al. Ets1 suppresses tumorigenesis of human breast cancer via trans-activation of canonical tumor suppressor genes. Front Oncol 2020;10:642.
[41] Kawiak A, Kostecka A. Regulation of bcl-2 family proteins in estrogen receptor-positive breast cancer and their implications in endocrine therapy. Cancers 2022;14:279.
[42] Sun J, Shen Q, Lu H, Jiang Z, Xu W, Feng L, et al. Oncogenic ras suppresses ing4-tdg-fas axis to promote apoptosis resistance. Oncotarget 2015;6:41997.
[43] Deniaud E, Baguet J, Mathieu A-L, Marvel J, Leverrier Y, et al. Overexpression of sp1 transcription factor induces apoptosis. Oncogene 2006;25:7096–105.
[44] Dhamdhere K, Sundararajan M, Yan Q. How important is a neuron. In: International conference on learning representations; 2018.
[45] Furth PA, Wang W, Kang K, Rooney BL, Keegan G, Muralidaran V, et al. Overexpression of estrogen receptor α in mammary glands of aging mice is associated with a proliferative risk signature and generation of estrogen receptor α–positive mammary adenocarcinomas. Am J Pathol 2023;193:103–20.