Supplementary materials: "Identifying cellular cancer mechanisms through pathway-driven data integration"

Sam F. L. Windels, Noël Malod-Dognin, and Nataša Pržulj

Contents

1 Data

1.1 Statistics of case and control PPI networks

To create four tissue-specific case and control networks PPI networks, we overlay a generic PPI network sourced from BioGRID with case and control RNA-SEQ gene expression data sourced from the Human Protein Atlas for four types of cancer (see Section 2.4.1 of the main paper). For each of our case and control networks, we present the number of nodes, the number of edges and the network density in Supplementary Table **??**. For each pair of case and control networks, we present the overlap in terms of nodes and edges, measured using the Jaccard Index, in Supplementary Table **??**.

	Nodes	**Edges**	**Density**
Lung Case	11,635	274,202	0.41%
Lung Control	13,590	311,754	0.34%
Colon Case	11,298	267,833	0.42%
Colon Control	13,480	316,247	0.35%
Prostate Case	11,651	275470	0.41%
Prostate Control	13,654	312,201	0.33%
Ovary Case	12,027	286,596	0.40%
Ovary Control	12,626	288,960	0.36%

Supplementary Table 1: **Details case and control PPI networks.** The number of nodes, the number of edges and the density (columns 1-3, respectively), for each of our case and control PPI networks (rows).

	Node overlap	**Edge overlap**
Lung	0.78	0.80
Colon	0.78	0.80
Prostate	0.79	0.79
Ovary	0.81	0.80

Supplementary Table 2: **Overlap between case and control networks.** For each case and control network (rows), the overlap in terms of nodes and edges measured using the Jaccard Index (columns).

1.2 The number of pathways considered per molecular network

We collect pathway data from Reactome (see Section 2.4.2 in the main paper). For each of our four pairs of tissue-specific case and control networks, we consider only those pathways that form a subnetwork of at least ten and up to a hundred nodes when induced on either network (as mentioned in Section 2.6.2 of the main paper). We present the number of pathways considered per tissue type in Supplementary Table **??**. Additionally, for a each of our four tissues, we take the union of the case and control PPI network and measure how many of the nodes (genes) are pathway annotated (using only the selected pathways).

	No. of pathways	**Coverage of nodes in union of case and control PPI**
Lung	1,025	5,266 / 14,118 (37.30%)
Colon	1,027	5,269 / 13,908 (37.88%)
Prostate	1,024	5,227 / 14,096 (37.09%)
Ovary	1,002	5,100 / 13,627 (37.43%)

Supplementary Table 3: **Pathway statistics.** For each of our four tissues considered (rows), we indicate the number of pathways considered (column 1) and how many genes they annotate in the union of the case and control network for the corresponding tissue.

1.3 Statistics of annotation data

We collect cancer annotation data at the pathway gene level from Reactome and COSMIC, respectively (for details, see Section 2.4.3). Here, we present the number of cancer pathways, cancer drivers and prognostic genes per tissue in Supplementary Table **??**.

	Cancer pathways	**Driver genes**	**Prognostic genes**
Lung	61	652	614
Colon	61	647	575
Prostate	61	647	152
Ovary	61	632	475

Supplementary Table 4: **Cancer annotation data statistics.** For each tissue, the number of cancer pathways, driver genes and prognostic genes considered.

2 Methods

To learn the functional organisation of the cell, we introduce GNMTF in Section 2.2.1. To improve how well our model captures the functional organisation of pathways in the cell, we propose PNMTF, which includes the integration of Reactome pathway data, in Section 2.2.2. Below, we detail how we solve both models using multiplicative update rules, and how we initialise both solvers using Singular Value Decomposition (SVD).

2.1 Solving GNMTF

2.1.1 GNMTF Multiplicative update rules

We present our algorithm to solve GNMTF below in Algorithm **??**.

Algorithm 1 Multiplicative update rules GNMTF

Initialise U, S, V using SVD (see Supplementary Section **??**)

for all $t = 0, 1, \ldots, t_{limit} - 1$, or until $S^{t+1} = S^t$, $U^{t+1} = U^t$ and $V^{t+1} = V^t$ **do**

$$S_p^{t+1} = S_p^t \odot \sqrt{\frac{((U_p^t)^T A_{pG_i})^+ + (U_p^t S_p^t)^-}{((U_p^t)^T A_{pG_i})^- + (U_p^t S_p^t)^+}};$$

$$U_p^{t+1} = U_p^t \odot \sqrt{\frac{(A_{pG_i} V^t (S_p^{t+1})^T)^+ + U_p^t (S_p^{t+1} (S_p^{t+1})^T)^-}{(A_{pG_i} V^t (S_p^{t+1})^T)^- + U_p^t (S_p^{t+1} (S_p^{t+1})^T)^+}};$$

$$V^{t+1} = V^t \odot \sqrt{\frac{(A_{pG_i}^T U_p^{t+1} S_p^{t+1})^+ + (V^t (S_p^t + 1)^T (U_p^{t+1})^T U^{t+1} S^{t+1})^-}{(A_{pG_i}^T U^{t+1} S^{t+1})^- + (V^t (S^{t+1})^T (U^{t+1})^T U^{t+1} S^{t+1})^+}};$$

Return the last computed (U, S, V);

2.1.2 GNMTF initialisation

We recall the objective function of our GNMTF model:

$$\min_{U,S,V \geq 0} \sum_{i=0}^{r} \left\| \widetilde{A}_{G_i} - USV^T \right\|_F^2, \text{ s.t.: } V^T V = I. \tag{1}$$

To initialise U, S, and V, we apply a truncated SVD on graphlet adjacency matrix $\widetilde{A}_{G_i}$, for the d largest singular values:

$$W \Sigma Z^T = SVD(\widetilde{A}_{G_i}), \tag{2}$$

where W and Z^T are $n \times d$ dimensional matrices of which the columns are respectively the d left and right singular vectors of $\widetilde{A}_{G_i}$, and Σ is a $d \times d$ diagonal matrix with the d largest singular values of $\widetilde{A}_{G_i}$ on the diagonal.

We initialise the columns of U based on the columns of W. Specifically, we initialise the j^{th} column vector of U, $U[j]$, based on the j^{th} column vector of W, $W[j]$. To do so, we split $W[j]$ into two non-negative vectors: $W[j]^+$, which is a copy of $W[j]$ with all negative values set to 0, and $W[j]^-$, which is a copy of $W[j]$ where all positive values are set to 0 and all negative entries are set to their absolute value. Either $W[j]^+$ or $W[j]^-$ is assigned to $U[j]$, depending on which one has the highest euclidean norm. Analogously, we initialise V based on Z. S is simply initialized by matrix Σ.

2.2 Solving PNMTF

2.2.1 PNMTF multiplicative update rules

We present our algorithm to solve PNMTF below in Algorithm **??**.

Algorithm 2 Multiplicative update rules PNMTF

Initialise U, S, V using SVD (see Supplementary Section **??**)

for all $t = 0, 1, \ldots, t_{limit} - 1$, or until $S^{t+1} = S^t$, $U^{t+1} = U^t$ and $V^{t+1} = V^t$ **do**

$$S_p^{t+1} = S_p^t \odot \sqrt{\frac{((U_p^t)^T H_{pG_i})^+ + (U_p^t S_p^t)^-}{((U_p^t)^T H_{pG_i})^- + (U_p^t S_p^t)^+}};$$

$$U_p^{t+1} = U_p^t \odot \sqrt{\frac{(H_{pG_i} V^t (S_p^{t+1})^T)^+ + U_p^t (S_p^{t+1} (S_p^{t+1})^T)^-}{(H_{pG_i} V^t (S_p^{t+1})^T)^- + U_p^t (S_p^{t+1} (S_p^{t+1})^T)^+}};$$

$$V^{t+1} = V^t \odot \sqrt{\frac{(H_{pG_i}^T U_p^{t+1} S_p^{t+1})^+ + (V^t (S_p^t + 1)^T (U_p^{t+1})^T U^{t+1} S^{t+1})^-}{(H_{pG_i}^T U^{t+1} S^{t+1})^- + (V^t (S^{t+1})^T (U^{t+1})^T U^{t+1} S^{t+1})^+}};$$

Return the last computed (U, S, V);

2.2.2 PNMTF initialisation

We recall the objective function of our PNMTF model:

$$\min_{\substack{U_p \geq 0, \\ S_p \geq 0, \\ V \geq 0}} \sum_{p=0}^{r} \left\| H_{pG_i} - U_p S_p V^T \right\|_F^2, \text{ s.t.: } V^T V = I. \tag{3}$$

The initialisation of PNMTF is analogous to that of GNMTF. To initialize U_p for a given pathway p, we apply a truncated SVD on graphlet adjacency matrix H_{pG_i}, for the d_p largest singular values:

$$W\Sigma Z = SVD(H_{pG_i}). \tag{4}$$

We initialise the columns of U_p based on the columns of W. Specifically, we initialise the j^{th} column vector of U_p, $U_p[j]$, based on the j^{th} column vector of W, $W[j]$. To do so, we split $W[j]$ into two non-negative vectors: $W[j]^+$, which is a copy of $W[j]$ with all negative values set to 0, and $W[j]^-$, which is a copy of $W[j]$ where all positive values are set to 0 and all negative entries are set to their absolute value. Either $W[j]^+$ or $W[j]^-$ is assigned to $U_p[j]$, depending on which one has the highest euclidean norm. S_p is simply initialized by matrix Σ.

To initialize V, we apply a truncated SVD on normalized graphlet adjacency matrix A_{G_i}, computed for the d components corresponding to the d largest singular values:

$$W\Sigma Z = SVD(\widetilde{A}_{G_i}). \tag{5}$$

we initialize V based on Z, initialising the column vectors of V using the same procure we applied to initialize V in GNMTF.

2.3 Ancestor-pathway enrichment analysis

We asses if a groups of pathways embedded nearby in space of pathways are functionally similar according to Reactome in Section 3.1. To do so, we first annotate each pathway with its first level ancestor annotations found higher up in the Reactome ontology, one hop away from the root node. That is, to annotate a given pathway with its ancestor(s), we first find that pathway in the Reactome directed acyclic graph (DAG) encoding the relationships between pathway annotations, from most generic to most specific. From there, we trace the Reactome Ontology DAG upwards (against the direction of the 'is_a' relationships) until we reach the annotations that are one hop away from one of the 23 roots of the Reactome DAG, and use those 92 pathway annotations as ancestor annotations. Then, to assess if a set of pathways is statistically significantly enriched by pathways sharing ancestor annotations, we apply the hyper-geometric test. That is, we consider a set of pathways as a 'sampling without replacement' experiment, in which each time we find a given ancestor, we count that as a 'success'.

The probability of observing the same or higher enrichment (i.e. successes) of the given annotation by chance is equal to:

$$p = 1 - \sum_{i=0}^{X-1} \binom{K}{i} \binom{M-K}{N-i} / \binom{M}{N}, \tag{6}$$

where N is the number of ancestor annotated pathways in the pathway-set, X is the number of pathways annotated with the given ancestor, M is the number of ancestor pathways and K is the number of pathways annotated with the given ancestor over all pathways in the pathway-set. An ancestor annotation is considered to be statistically significantly enriched if its enrichment p-value is lower than or equal to 5% after application of the Benjamini and Hochberg correction for multiple hypothesis testing.

2.4 Measuring prediction accuracy

We apply our NMTF-scores to predict cancer implicated pathways and genes in Sections 3.2 and 3.3, respectively. As wet-lab validation is expensive, we are predominantly interested in the top-scoring entities that are highly likely to be cancer implicated for both types of predictions. So, for both types of predictions, we consider the set of top-scoring entities as a prediction and use the Matthew Correlation Coefficient (MCC) to measure the prediction accuracy:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}, \quad (7)$$

where TP is he number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives. For pathway focused predictions, we use the set of known 'cancer pathways' in Reactome as the set of true positives (see Section 2.6.2). For gene focused predictions, we use the set of known cancer drivers in COSMIC as the set of true positives (see Section 2.6.3). The MMC ranges from -1 to 1, where 1 indicates a perfect prediction, 0 indicates random performance and -1 indicates an inverse prediction.

2.5 Network diffusion

Network diffusion refers to a family of related techniques, which propagate node information along the edges of the network. Here we focus on the diffusion kernel, which we apply to predict cancer related genes in Section 3.3 of the paper.

The diffusion kernel is often called the 'heat kernel', as it can be viewed as describing the flow of heat originating from the nodes across the edges of a graph with time. In network biology nodes typically represent genes and 'heat' on a node represents experimental measurements. Here, our network is the generic PPI network collected from BioGRID, on which we diffuse the differentially expressed genes for a given cancer type (see Section 2.4.1) based on graphlet adjacency A_{G_1}; as explained next.

Network diffusion is based on the *Laplacian matrix*, a network representation related to the adjacency matrix defined as $\mathcal{L} = D - A$, where A is the adjacency matrix and D is the diagonal matrix such that D_{uu} is equal to the degree of node u. We extended heat diffusion to graphlet based heat diffusion by replacing A with the symmetrically normalized graphlet adjacency matrix $\widetilde{A}_{G_k}$ in (**?**).

Concretely, for the set of n nodes in the generic PPI, we encode the differentially expressed genes in a given cancer in vector $P_0 \in \mathbb{R}^n$, where $P_0(u) = 1$ if the gene u is differentially expressed and 0 otherwise. Information is diffused as follows: $P = HP_0$, where H is a diffusion kernel. The diffusion kernel, H^k_α, is defined as the matrix exponential of the Laplacian matrix (**?**):

$$H_\alpha = e^{-\alpha\mathcal{L}}, \tag{8}$$

where the parameter $\alpha \in \mathbb{R}$ controls the level of diffusion.

3 Results

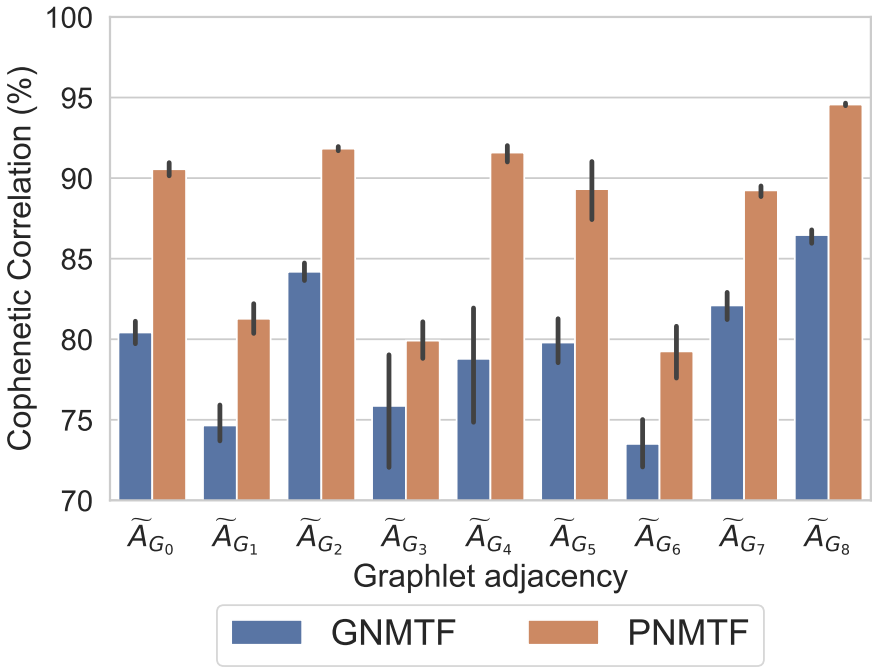
3.1 PNMTF best captures functional organisation of pathways in the cell

In Section 3.1 of the main manuscript we validate that our PNMTF model best captures the functional organisation of pathways in the cell as described by Reactome, compared to a standard NMTF model that does not integrate pathway annotation data (i.e. GNMTF). We consider this the case if when embedding all pathways in the space: (1) the pathway embeddings form dense and well-separable clusters, and (2) pathways in the same cluster are functionally similar according to Reactome. In the main paper, we show that this holds for the control PPI network based on graphlet adjacency A_{G_1}. Here, we present the results for all of our control networks across all graphlet adjacencies.

3.1.1 Intrinsic quality of pathway-clusters in space: cophenetic correlation

For a given network and graphlet adjacency, to measure if the pathway embeddings form dense and well-separable clusters in the embedding space (i.e. have high intrinsic clustering quality), we first embed all pathways in shared space, apply agglomerative hierarchical clustering and measure the intrinsic quality of the hierarchical clustering using the cophenetic correlation (i.e. the Pearson correlation between the cosine distance between two pathways and the height in the linkage tree where their corresponding branches meet). Results averaged over the four control networks across all graphlet adjacencies are presented in Supplementary

Figure **??**. We observe that, with the exception of when based on graphlet adjacency $\widetilde{A}_{G_5}$, PNMTF outperforms GNMTF (average cophenetic correlation 89% compared to 83%). The best intrinsic clustering quality, averaged over the 4 control networks, is measured based on graphlet adjacency $\widetilde{A}_{G_8}$ using PNMTF at 96%. We conclude that as PNMTF based pathway embeddings show the clearest clustering structure in the embedding space.



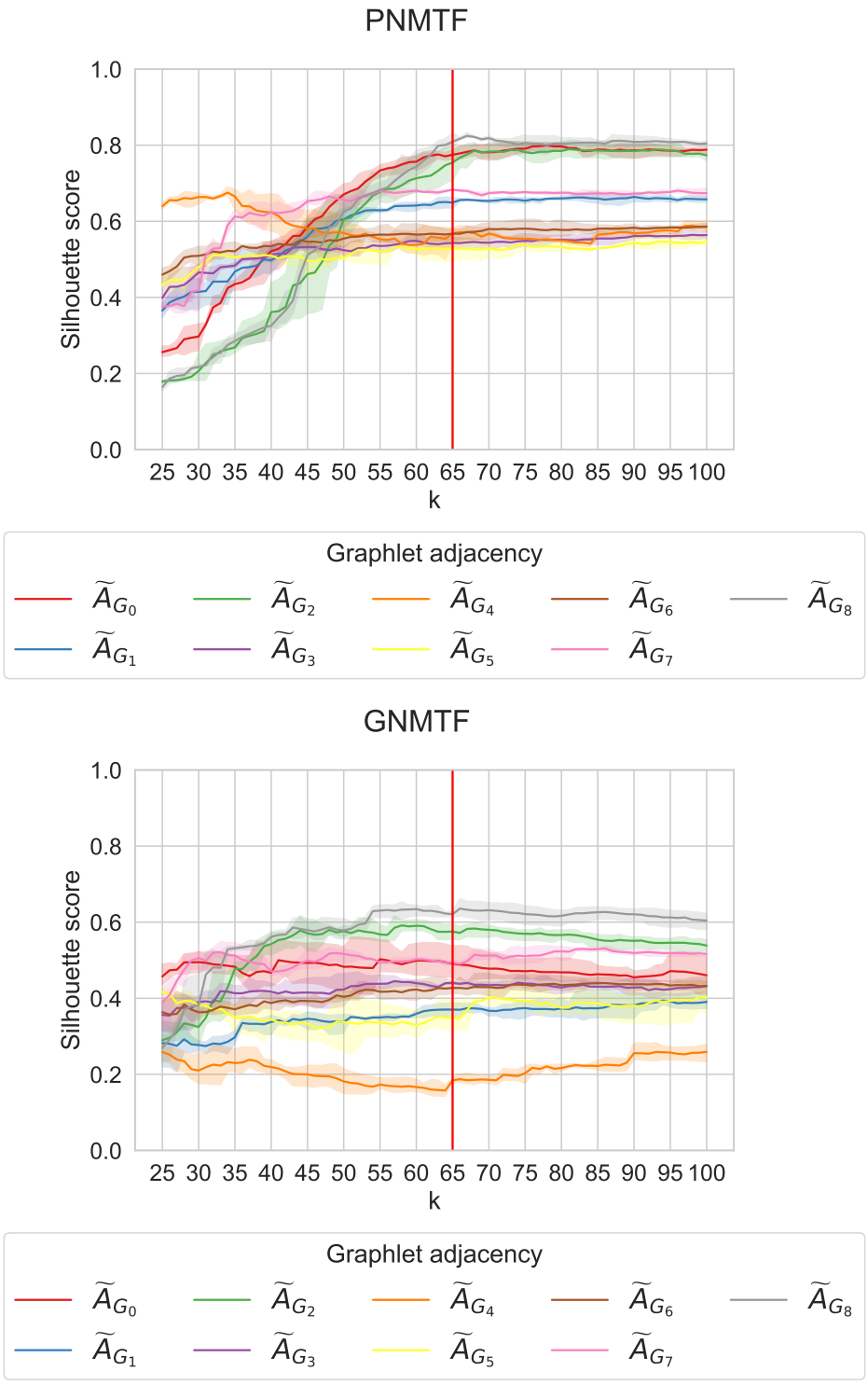
Supplementary Figure 1: **PNMTF best captures the functional organisation of pathways in the cell.** For our GNMTF and PNMTF models, we present the cophenetic correlation averaged over our four tissues (y-axis) for our pathway embeddings. The distance between pathways is measured using 'cosine distance'. Hierarchical clustering is performed using 'average linkage'.

3.1.2 The optimal number of clusters to extract

Next, we want to extract clusters of pathways from our PNMTF and GNMTF based hierarchical clusterings of pathways so that subsequently we can validate that pathways embedded/clustered together are functionally related (see Supplementary Section **??**). To that end, for both GNMTF and PNMTF, we first need to threshold the hierarchical tree to extract clusters of pathways. To determine the optimal threshold, we cut the tree at different heights, such that the number of exacted clusters, k, varies from 25 to 100. For each value of k, we compute the

corresponding silhouette score, which measures how well separated the extracted clusters are as a measure of intrinsic clustering quality. Results for PNMTF and GNMTF based on the different graphlet adjacencies are shown in Supplementary Figure **??**.

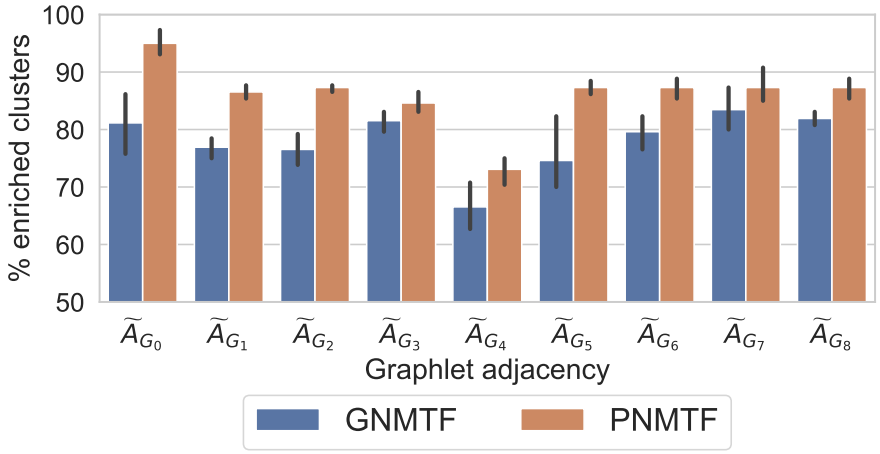
For PNMTF, we find that the silhouette scores are non-decreasing and plateau from 65 clusters onwards for all graphlet adjacencies, except for $\widetilde{A}_{G_4}$ and $\widetilde{A}_{G_5}$. For GNMTF, we find that the silhouette scores do not climb and plateau as for PNMTF but rather remain stable regardless of the value of k (except for graphlet adjacencies $\widetilde{A}_{G_2}$ and $\widetilde{A}_{G_8}$). This is because the pathway embeddings do not form easily separable clusters in space with GNMTF, as is visible in Figure 2 in the main manuscript. Therefore, for both methods, we cut the tree such that we achieve 65 clusters.



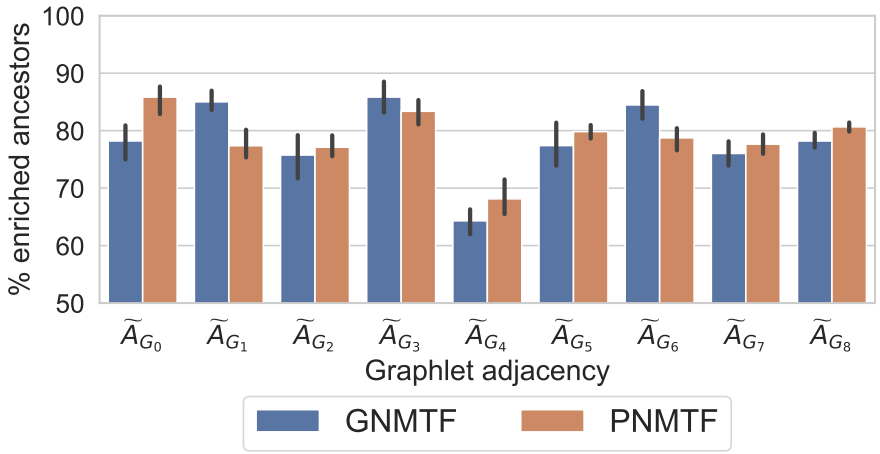
Supplementary Figure 2: **Determining the optimal number of pathways to extract.** For PNMTF (top) and GNMTF (bottom), the silhouette scores (y-axis) extracting k (x-axis) clusters of pathways from the embedding spaces, based on different graphlet adjacencies (colour, legend).

3.1.3 Extrinsic quality of pathway-clusters in space: enrichment analysis

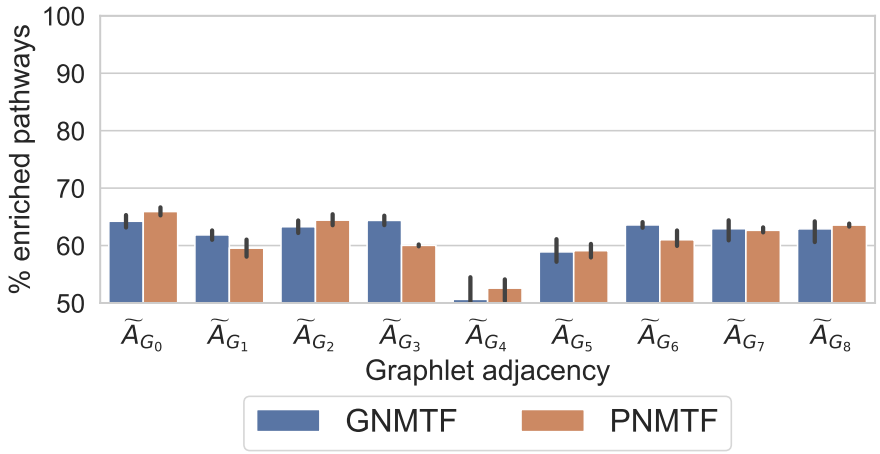
For a given network and graphlet adjacency, to measure if pathways clustered together in space are also functionally related, we perform ancestor-pathway enrichment analysis (defined in Section 2.7). The results averaged over the four control networks across the different graphlet adjacencies are presented in Supplementary Figure **??**. Averaged over the four control networks, we observe that pathway clusters in the shared space trained using PNMTF are at least as much or more enriched in ancestor annotations as pathway clusters using GNMTF; this is true across all graphlet adjacencies in terms of percentage of clusters enriched (on average 94% compared to 89%), percentage of ancestor pathway annotations enriched (on average 94% compared to 84%) and percentage of pathways with at least one ancestor enriched (on average 76% compared to 72%). This means that PNMTF outperforms GNMTF in capturing the functional organisation of pathways as described by Reactome.



(A)



(B)



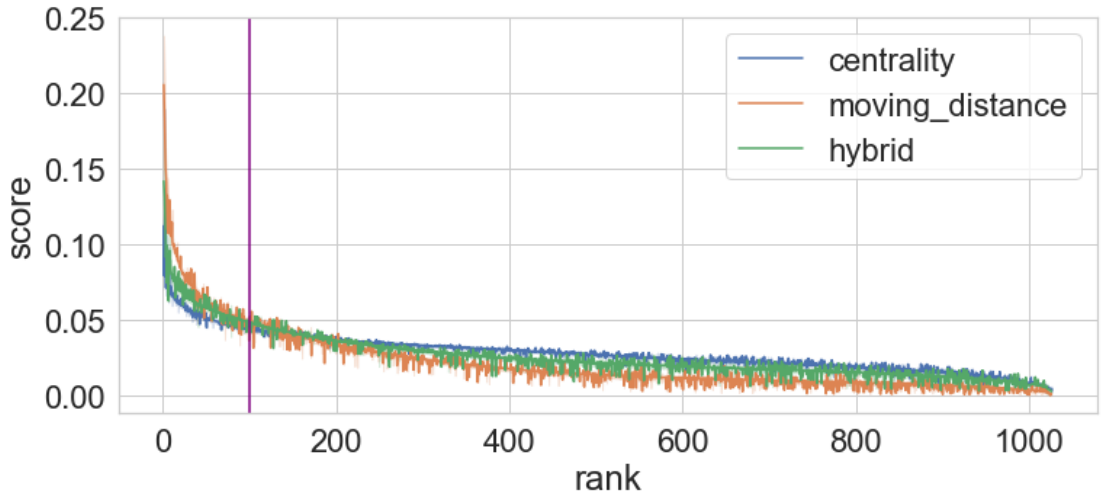
(C)

Supplementary Figure 3: Pathway clustering ancestor enrichment analysis. For our GNMTF model and our PNMTF model, we apply ancestor enrichment analysis on our pathway clusters and report: the percentage of clusters that contain at least one enriched ancestor annotation, the percentage of ancestor annotations that are enriched in at least one cluster and the percentage of pathways that have at least one ancestor annotation enriched.

3.2 Identifying the set of top-scoring pathways: defining a threshold

Here, we want to define for each of our different NMTF scores applied at the pathway level, a threshold to select top-scoring pathways. To do so, for each of our three different NMTF-scores, graphlet adjacency and tissue, we compute the pathway scores and sort them in descending order. We present the results per NMTF-score, i.e., averaged over the different tissues and graphlet adjacencies, in Supplementary Figure **??**.

The narrow error bands, representing the 95% confidence interval across different tissues and graphlet adjacencies, indicate that the trends across different graphlet adjacencies are similar, allowing us to pick a single threshold that holds for all graphlet adjacencies. We choose to consider the top 100 pathways as our set of pathways predicted to be implicated in cancer, as that is where the centrality score, which also underlies our hybrid score, flattens out (see purple line).



Supplementary Figure 4: **Determining a threshold for identifying top-scoring pathways through an elbow method.** For each of our three different measures (legend), we sort the pathway scores (y-axis) in descending order (x-axis), averaged over four cell types and graphlet adjacencies (represented as error bands). A vertical purple line indicates the top 100 highest ranked pathways.

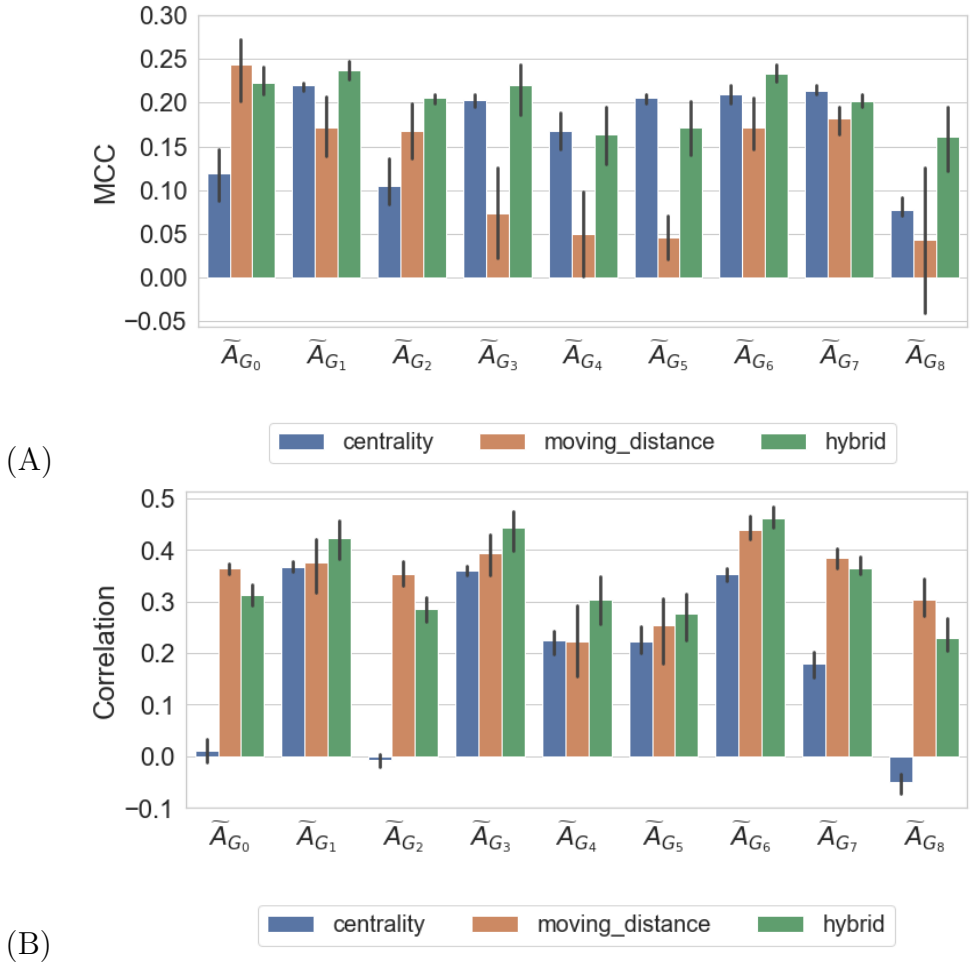
3.3 Pathway level prediction accuracy across graphlet adjacencies

Here, we identify what graphlet adjacencies lead to the best performance when using our NMTF-scores to prioritise pathways implicated in cancer. We compare

the performance of our NMTF-scores based on these best performing graphlet adjacencies against the state of the art in Section 3.3.

Specifically, for a given type of NMTF-score, tissue and graphlet-adjacency, we measure the cancer pathway prediction performance of the given score by computing the Mathew Correlation Coefficient (MCC) using known cancer pathways in Reactome as a gold standard and the top 100 highest scoring pathways as predictions for pathways implicated in cancer (see Supplementary Section **??** for details). Additionally, we also consider the ratio of driver genes in a pathway as an indication of a its engagement in cancer. Then, to evaluate a given pathway prediction method, we measure the Spearman's rank correlation between this ratio and a pathway's score. We present the results for our three NMTF-scores, for all graphlet adjacencies and averaged over the four tissue types, in Supplementary Figure **??**

We observe that in terms of MCC (Suppl. Fig.**??**A), the best performance is achieved when using our moving distance and regular adjacency (0.244), just outperforming our hybrid score with graphlet adjacencies $A_{G_1}, A_{G_3},$ and A_{G_6} (0.237, 0.220 and 0.233, respectively). Looking at our correlation results (Suppl. Fig. **??**B), we find that our hybrid-score with graphlet adjacencies A_{G_1}, A_{G_3}, and A_{G_6} greatly outperform the moving distance with regular adjacency (0.421, 0.443 and 0.461, compared to 0.312). Based on these results, we chose to focus on graphlet adjacencies AG_0, AG_1, AG_3 and AG_6 for our comparison against the state of the art in the main paper.



Supplementary Figure 5: **Comparing cancer pathway prediction accuracy for different NMTF-scores across graphlet adjacencies.** Sub-plots (A) and (B) respectively show the MCC and rank-correlation scores for predicting Reactome cancer pathways. From left to right, we compare results for PNMTF based on different graphlet adjacencies (x-axis) and different NMTF-scores (legend).

3.4 Pathway level validation

To further validate our method captures cancer-implicated pathways, we present the top 10 highest scoring pathways in each of our tissues, using our hybrid-score based on graphlet adjacency A_{G_1}, in Supplementary Tables **??** to **??**. We observe, for each of our four tissues, that between 5 and 7 out of 10 pathways are cancer pathways. For all four tissues, we observe that all top 10 pathways are related to the RAS-MAPK pathway, which transduces extracellular signals to the cell nucleus, regulating cell growth, cell division and cell repair. The RAS-MAPK

pathway has also been related to promoting angiogenesis. As such, the RAS-MAPK pathway has been frequently associated with tumour genesis (**?**) and is a frequent subject of therapeutic studies (**?**). Given that a high ratio of prioritised pathways is a known cancer pathway and the fact that the prioritised pathways are involved in mechanisms known to be rewired in cancer, we are confident that the remaining prioritised pathways that are not known as cancer pathways are indeed cancer related.

Rank	Pathway	Known cancer pathway
1	Constitutive Signaling by EGFRvIII	✓
2	Signaling by EGFRvIII in Cancer	
3	Signaling by ERBB2 ECD mutants	✓
4	GAB1 signalosome	
5	PI3K events in ERBB2 signaling	
6	HSF1 activation	
7	Constitutive Signaling by Ligand-Responsive EGFR Cancer Variants	✓
8	Signaling by Ligand-Responsive EGFR Variants in Cancer	✓
9	SHC1 events in EGFR signaling	
10	Signaling by ERBB2 KD Mutants	✓

Supplementary Table 5: **Top 10 highest scoring pathways in lung cancer.** The table ranks pathways according to their hybrid score using PNMTF based on graphlet adjacency A_{G_1}. The final column indicates if the pathway (2nd column) is a known cancer pathway in Reactome.

Rank	Pathway	Known cancer pathway
1	Constitutive Signaling by EGFRvIII	✓
2	Signaling by EGFRvIII in Cancer	✓
3	HSF1 activation	
4	Signaling by ERBB2 ECD mutants	✓
5	Constitutive Signaling by Ligand-Responsive EGFR Cancer Variants	✓
6	Signaling by Ligand-Responsive EGFR Variants in Cancer	✓
7	Signaling by FGFR4 in disease	✓
8	Signaling by FGFR3 fusions in cancer	✓
9	Downstream signaling of activated FGFR4	
10	Role of LAT2/NTAL/LAB on calcium mobilization	

Supplementary Table 6: **Top 10 highest scoring pathways in colon cancer.** The table ranks pathways according to their hybrid score using PNMTF based on graphlet adjacency A_{G_1}. The final column indicates if the pathway (2nd column) is a known cancer pathway in Reactome.

Rank	Pathway	Known cancer pathway
1	Signaling by EGFRvIII in Cancer	✓
2	Constitutive Signaling by EGFRvIII	✓
3	HSF1 activation	
4	Signaling by ERBB2 ECD mutants	✓
5	Constitutive Signaling by Ligand-Responsive EGFR Cancer Variants	✓
5	Signaling by Ligand-Responsive EGFR Variants in Cancer	✓
7	GAB1 signalosome	
8	Constitutive Signaling by Overexpressed ERBB2	✓
9	Downstream signaling of activated FGFR4	
10	Signalling to RAS	

Supplementary Table 7: **Top 10 highest scoring pathways in prostate cancer.** The table ranks pathways according to their hybrid score using PNMTF based on graphlet adjacency A_{G_1}. The final column indicates if the pathway (2nd column) is a known cancer pathway in Reactome.

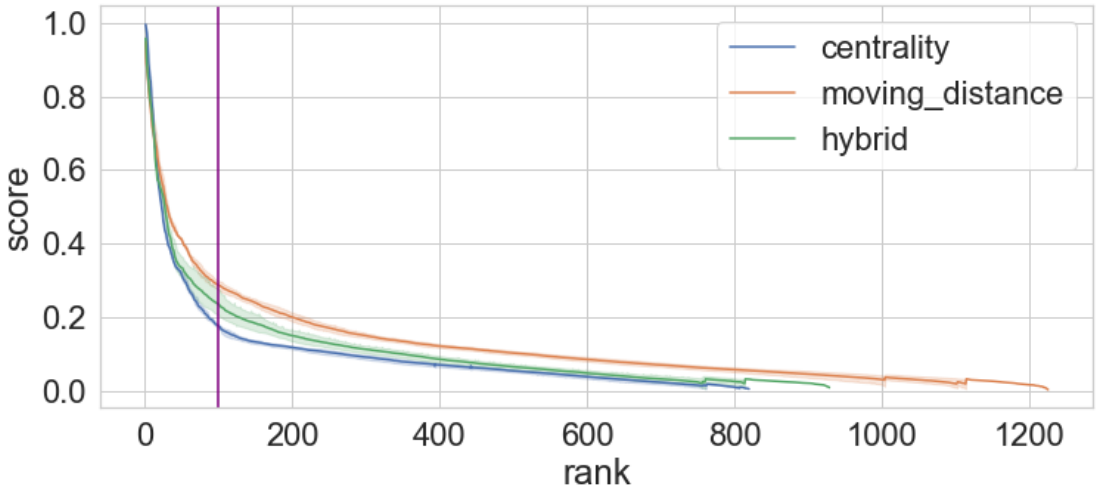
Rank	Pathway	Known cancer pathway
1	Signaling by EGFRvIII in Cancer	✓
2	Constitutive Signaling by EGFRvIII	✓
3	GAB1 signalosome	
4	Signaling by ERBB2 ECD mutants	✓
5	Transcriptional regulation by the AP-2 (TFAP2) fam. of TF.	
6	Signaling by EGFR in Cancer	✓
7	Downstream signaling of activated FGFR3	
8	Retrograde neurotrophin signalling	
9	Constitutive Signaling by Ligand-Responsive EGFR Cancer Variants	✓
10	Signaling by Ligand-Responsive EGFR Variants in Cancer	✓

Supplementary Table 8: **Top 10 highest scoring pathways in ovarian cancer.** The table ranks pathways according to their hybrid score using PNMTF based on graphlet adjacency A_{G_1}. The final column indicates if the pathway (2nd column) is a known cancer pathway in Reactome.

3.5 Identifying the set of top-scoring genes: defining a threshold

Here, we aim to define for each of our NMTF scores applied at the gene level, a threshold to select the top-scoring genes that participate in pathways that are in our set of 100 top-scoring pathways. We focus on graphlet adjacency A_{G_1}, as we show it best captures cancer mechanisms at the pathway level in Section 3.3.

Specifically, for each of our three NMTF-scores and four different tissues, we compute the gene scores for the genes participating in the top 100 highest scoring pathways and we rank them in descending order. Results averaged over the four tissues are presented in Supplementary Figure **??**. We choose to consider the top 100 highest ranked genes (purple vertical line) as our set of predicted cancer-related genes, as from there all three NMTF-scores plateau.



Supplementary Figure 6: **Determining a threshold for identifying top-scoring genes using an elbow method.** Limiting ourselves to the top 100 highest scoring pathways, for each of our three different NMTF-scores (legend) based on graphlet adjacency A_{G_1}, we sort the gene scores (y-axis), averaged over four cancer types (represented as error bands), in descending order (x-axis).

3.6 Gene-level validation

Here we validate in the literature our top 15 genes predicted as implicated in cancer for each of our four cancers (see Section 3.4 of the paper). We consider a gene to have a known role in cancer if it is listed in the COSMIC cancer driver database or if any wet-lab experiment demonstrates it has one of the following properties in the cancer:

1. enhances/diminishes susceptibility to anti-cancer agents,
2. promotes cell proliferation and cell survivability,
3. promotes migration and invasion,
4. inhibits tumour genesis (i.e. has a suppressor role).

Note that differential expression is not accepted as evidence.

For genes that we can not validate, we discuss their potential role in cancer, considering whether they are prognostic in the given cancer per the Pathology Atlas (**?**) and if they are implicated in other cancer(s) based on wet-lab experiments or the COSMIC cancer driver database.

Rank	Symbol	Pathway	Validation in Lung Adenocarcinoma	Prognostic	Validation in non-lung cancer	Drugability	Immune System
1	GRB2	SHC-mediated cascade: FGFR2	PMID: 26693065 (2, 3, mouse model)			Approved	
2	CSK	Phosphorylation of CD3 and TCR zeta chains			PMID: 11054667 (colon cancer, 4, in vitro)	Trial	✓
3	PTPN11	PD-1 signaling	PMID: 25730908 (2,3, mouse model)				✓
4	FYN	Dectin-2 family	PMID: 21371426 (3, in vitro)			Trial	✓
5	HSP90AA1	Attenuation phase			COSMIC (non-Hodgkins lymphoma)	Trial	
6	PIK3R1	RHOF GTPase cycle	PMID: 24550137 (1, mouse model)			Trial	
7	EGFR	Transcriptional reg. by the AP-2 family of TF	PMID: 20979469 (1, patient data)			Approved	
8	SRC	FCGR activation	PMID: 17200208 (1, in vitro)			Approved	✓
9	GNB1	G beta:gamma signalling through PI3Kgamma					
10	MYC	Transcriptional reg. by the AP-2 family of TF	PMID: 19551151 (2,3, mouse model)			Trial	
11	CUL1	Prolactin receptor signaling	PMID: 33478195 (4, in vitro)				✓
12	XPO1	Extra-nuclear estrogen signaling	PMID: 27680702 (1, in vitro)			Trial	
13	HNRNPH1	Signaling by FGFR2			PMID: 29362363 (Rhabdomyosarcoma, 2, mouse model)	Trial	
14	GNG2	G-protein beta:gamma signalling			PMID: 24660107 (melanoma, 4, in vitro)		
15	LYN	Dectin-2 family	PMID: 23866081 (1, in vitro)			Approved	✓

Supplementary Table 9: **Validation of top-scoring genes in lung cancer.** The table ranks the top 15 genes and corresponding pathways according to their hybrid score in lung cancer using PNMTF based on graphlet adjacency A_{G_1}. Genes in black have literature support for their role in Lung Adenocarcinoma (PubMed IDs or references to COSMIC are given in the column 'Validation in Lung Adenocarcinoma'). Genes without literature validation, i.e. predictions, are highlighted in blue. For predicted genes, additional literary evidence highlights their potential role in cancer. Firstly, we provide the p-values for genes that statistically significantly impact patient survival in Lung Adenocarcinoma based on Kaplan Meier survival curves in the column 'Prognostic'. Secondly, when available, we provide PubMed IDs or references to COSMIC for 'predicted' genes that have a proven role in cancers other than Lung Adenocarcinoma in the column 'Validation non-lung cancer'. In the column 'Druggability', we indicate which genes are a known drug target for at least one approved drug or one drug in clinical trial according to DrugBank (**?**). In the column 'Immune System', we indicate which pathways are immune system pathways according to Reactome.

Rank	Symbol	Pathway	Validated in colon cancer	Prognostic	Validation in non-colon cancer	Drugability	Immune System
1	GRB2	CD28 dependent Vav1 pathway	PMID: 12134161 (3, in vitro)			Approved	✓
2	PTPN11	PD-1 signaling	PMID: 32467571 (4, in vitro)				✓
3	PIK3R1	RND3 GTPase cycle	COSMIC			Trial	
4	CSK	Phosphorylation of CD3 and TCR zeta chains	PMID: 20010872 (3, in vitro)			Trial	✓
5	TRAF2	TNF receptor superfamily mediating non-canonical NF-kB pathway		9.31e-4	PMID: 28667915 (1, in vitro)		✓
6	SRC	FCGR activation	COSMIC			Approved	✓
7	HSP90AA1	HSF1 activation			COSMIC (non-Hodgkins lymphoma)	Trial	
8	LMNA	Meiotic synapsis			PMID: 22301279 (prostate, 2,3, in vitro)		
9	EGFR	GRB2 events in EGFR signaling	PMID: 15863375 (1, in vivo)			Approved	
10	XPO1	Extra-nuclear estrogen signaling	PMID: 26603256 (1, mouse model)			Trial	
11	CUL1	Prolactin receptor signaling	PMID: 29475926 (2, patient data)				✓
12	PTPRJ	Phosphorylation of CD3 and TCR zeta chains	PMID: 12089527 (2, mouse model)				✓
13	FN1	p130Cas linkage to MAPK signaling for integrins	PMID: 29274284 (2, 3, in vitro)			Approved	
14	BIRC3	TNF receptor superfamily mediating non-canonical NF-kB pathway		9.35e-4	COSMIC (leukemia, non-Hodgkins lymphoma, mantele cell lymphoma, multiple myeloma)		✓
15	GNB1	G beta:gamma signalling through PI3Kgamma			PMID: 25485910 (leukemia, 1, in vitro)		

Supplementary Table 10: **Validation of top-scoring genes in colon cancer.** The table ranks the top 15 genes and corresponding pathways according to their hybrid score in colon cancer using PNMTF based on graphlet adjacency A_{G_1}. Genes in black have literature support for their role in colon cancer (PubMed IDs or references to COSMIC are given in the column 'Validation in colon cancer'). Genes without literature validation, i.e. predictions, are highlighted in blue. For predicted genes, additional literary evidence highlights their potential role in cancer. Firstly, we provide the p-values for genes that statistically significantly impact patient survival in colon cancer based on Kaplan Meier survival curves in the column 'Prognostic'. Secondly, when available, we provide PubMed IDs or references to COSMIC for 'predicted' genes that have a proven role in cancers other than colon cancer in the column 'Validation non-colon cancer'. In the column 'Druggability', we indicate which genes are a known drug target for at least one approved drug or one drug in clinical trial according to DrugBank (**?**). In the column 'Immune System', we indicate which pathways are immune system pathways according to Reactome.

Rank	Symbol	Pathway	Validated in prostate cancer	Prognostic	Validated in non-prostate cancer	Drugability	Immune System
1	HSP90AA1	Binding and Uptake of Ligands by Scavenger Receptors			COSMIC (non-Hodgkins lymphoma)	Trial	
2	GRB2	SHC-mediated cascade:FGFR2	PMID: 17372910 (1, mouse model)			Approved	
3	SHC1	Interleukin-2 signaling	PMID: 29462661 (3, in vitro)				✓
4	PTPN11	PD-1 signaling	PMID: 21442024 (2, mouse model)				✓
5	PIK3R1	Interleukin-7 signaling			PMID: 20530665 (liver cancer, 4, mouse model) + downregulated in prostate cancer		✓
6	TRAF2	TNF receptor superfamily mediating non-canonical NF-kB pathway	PMID: 28667915 (1, in vitro)				✓
7	SRC	Nuclear signaling by ERBB4	PMID: 14662770(2, mouse model)	1.63e-4		Approved	
8	LMNA	Diseases of programmed cell death	PMID: 22301279 (2,3, in vitro)				
9	EGFR	Transcriptional regulation by the AP-2 family of transcription factors	PMID: 32574928 (1, in vitro)			Approved	Approved
10	NTRK1	Signalling to RAS	PMID: 17143529 (1, in vitro)			Approved	
11	HMGB1	Regulation of TLR by endogenous ligand	PMID: 31410208 (3, in vitro)			Trial	✓
12	UBE2I	SUMOylation of intracellular receptors	PMID: 30631151 (3, in vitro)				
13	PRC1	RHO GTPases activate CIT	PMID: 31327655 (3, in vitro)				
14	TLR4	Regulation of TLR by endogenous ligand	PMID: 18092352 (2, in vitro)			Approved	✓
15	BIRC3	TNF receptor superfamily mediating non-canonical NF-kB pathway	PMID: 31511829 (2, in vitro)				✓

Supplementary Table 11: **Validation of top-scoring genes in prostate cancer.** The table ranks the top 15 genes and corresponding pathways according to their hybrid score in colon cancer using PNMTF based on graphlet adjacency A_{G_1}. Genes in black have literature support for their role in prostate cancer (PubMed IDs or references to COSMIC are given in the column 'Validation in prostate cancer'). Genes without literature validation, i.e. predictions, are highlighted in blue. For predicted genes, additional literary evidence highlights their potential role in cancer. Firstly, we provide the p-values for genes that statistically significantly impact patient survival in prostate cancer based on Kaplan Meier survival curves in the column 'Prognostic'. Secondly, when available, we provide PubMed IDs or references to COSMIC for 'predicted' genes that have a proven role in cancers other than prostate cancer in the column 'Validation non-prostate cancer'. In the column 'Druggability', we indicate which genes are a known drug target for at least one approved drug or one drug in clinical trial according to DrugBank (**?**). In the column 'Immune System', we indicate which pathways are immune system pathways according to Reactome.

Rank	Symbol	Pathway	Validated in ovarian cancer	Prognostic	Validated in non-ovarian cancer	Drugability	Immune System
1	GRB2	Signal attenuation	PMID: 32754300 (1, mouse model)			Approved	
2	TRIM25	Ovarian tumor domain proteases	PMID: 32826889(2, mouse model)				
3	PTPN11	Signaling by Leptin	PMID: 28814887 (3, in vitro + mouse model)				
4	HSP90AA1	Attenuation phase	PMID: 23135731 (1, 2, in vitro)			Trial	
5	SRC	RUNX2 regulates osteoblast differentiation	PMID: 27526105 (1, in vitro)			Approved	
6	EGFR	GRB2 events in ERBB2 signaling	PMID: 22416774 (1, review paper)			Approved	
7	PIK3R1	RHOF GTPase cycle	PMID: 30755611 (1, mouse model)			Trial	
8	YWHAB	Frs2-mediated activation	PMID: 30535456 (3, mouse model)			Trial	
9	LMNA	Diseases of programmed cell death	PMID: 30384980 (4, in vitro)			Trial	
10	XPO1	Extra-nuclear estrogen signaling	PMID: 27649553 (1, mouse model)				
11	UBE2I	SUMOylation of intracellular receptors			PMID: 30631151 (prostate, 3, in vitro)		
12	HNRNPH1	Signaling by FGFR2			PMID: 34295818 (leukemia, 2, mouse model)	Trial	
13	MYC	Transcriptional regulation by the AP-2 (TFAP2) fam. of TFs	PMID: 8314536 (1, in vitro)	7.52e-4		Trial	
14	FN1	p130Cas linkage to MAPK signaling for integrins	PMID: 34093898 (3, in vitro)			Approved	
15	SYK	Interleukin-2 signaling	PMID: 29643476 (1, 3, in vitro)			Trial	

Supplementary Table 12: **Validation of top-scoring genes in ovarian cancer.** The table ranks the top 15 genes and corresponding pathways according to their hybrid score in colon cancer using PNMTF based on graphlet adjacency A_{G_1}. Genes in black have literature support for their role in ovarian cancer (PubMed IDs or references to COSMIC are given in the column 'Validation in ovarian cancer'). Genes without literature validation, i.e. predictions, are highlighted in blue. For predicted genes, additional literary evidence highlights their potential role in cancer. Firstly, we provide the p-values for genes that statistically significantly impact patient survival in ovarian cancer based on Kaplan Meier survival curves in the column 'Prognostic'. Secondly, when available, we provide PubMed IDs or references to COSMIC for 'predicted' genes that have a proven role in cancers other than ovarian cancer in the column 'Validation non-ovarian cancer'. In the column 'Druggability', we indicate which genes are a known drug target for at least one approved drug or one drug in clinical trial according to DrugBank (**?**). In the column 'Immune System', we indicate which pathways are immune system pathways according to Reactome.

References

Kondor, R. I. and Lafferty, J. (2002). Diffusion kernels on graphs and other discrete input spaces. In *ICML*, volume 2, pages 315–322.

Molina, J. R. and Adjei, A. A. (2006). The Ras/Raf/MAPK Pathway. *Journal of Thoracic Oncology*, **1**(1), 7–9.

Santarpia, L., Lippman, S. M., and El-Naggar, A. K. (2012). Targeting the MAPK–RAS–RAF signaling pathway in cancer therapy. *Expert Opinion on Therapeutic Targets*, **16**(1), 103–119.

Uhlen, M., Zhang, C., Lee, S., Sjöstedt, E., Fagerberg, L., Bidkhori, G., Benfeitas, R., Arif, M., Liu, Z., Edfors, F., and Others (2017). A pathology atlas of the human cancer transcriptome. *Science*, **357**(6352).

Windels, S. F. L., Malod-Dognin, N., and Pržulj, N. (2019). Graphlet Laplacians for topology-function and topology-disease relationships. *Bioinformatics*, **35**(24), 5226–5234.

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., and Others (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, **46**(D1), D1074—-D1082.