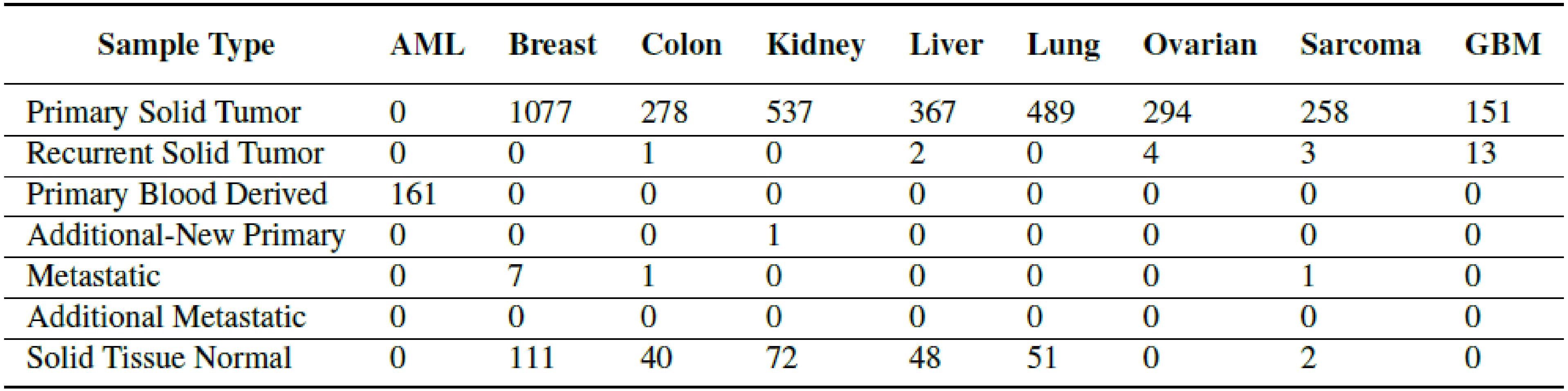
Identifying Cross-Cancer Similar Patients via a Semi-Supervised Deep Clustering Approach

Ping-Han Hsieh

Method
Input Data

- Feature Matrix (**X**)
  - Gene Expression (from TCGA) *
  - Age and Gender
- Analysis
  - Copy number variation
  - Somatic mutation

- Target
  - Cancer type (classification) (**y**)
  - Survival time (Cox partial likelihood) (**h**)
  - Clustering (K-means loss)

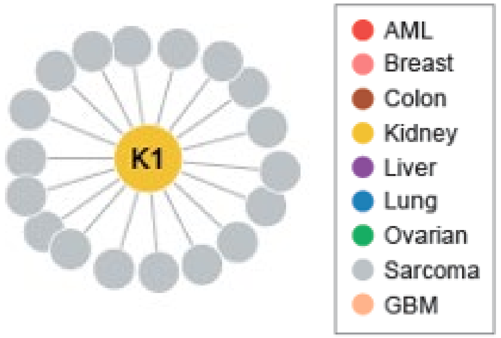


Sample Type	AML	Breast	Colon	Kidney	Liver	Lung	Ovarian	Sarcoma	GBM
Primary Solid Tumor	0	1077	278	537	367	489	294	258	151
Recurrent Solid Tumor	0	0	1	0	2	0	4	3	13
Primary Blood Derived	161	0	0	0	0	0	0	0	0
Additional-New Primary	0	0	0	1	0	0	0	0	0
Metastatic	0	7	1	0	0	0	0	1	0
Additional Metastatic	0	0	0	0	0	0	0	0	0
Solid Tissue Normal	0	111	40	72	48	51	0	2	0

Method
Cross Cancer Similar Patients

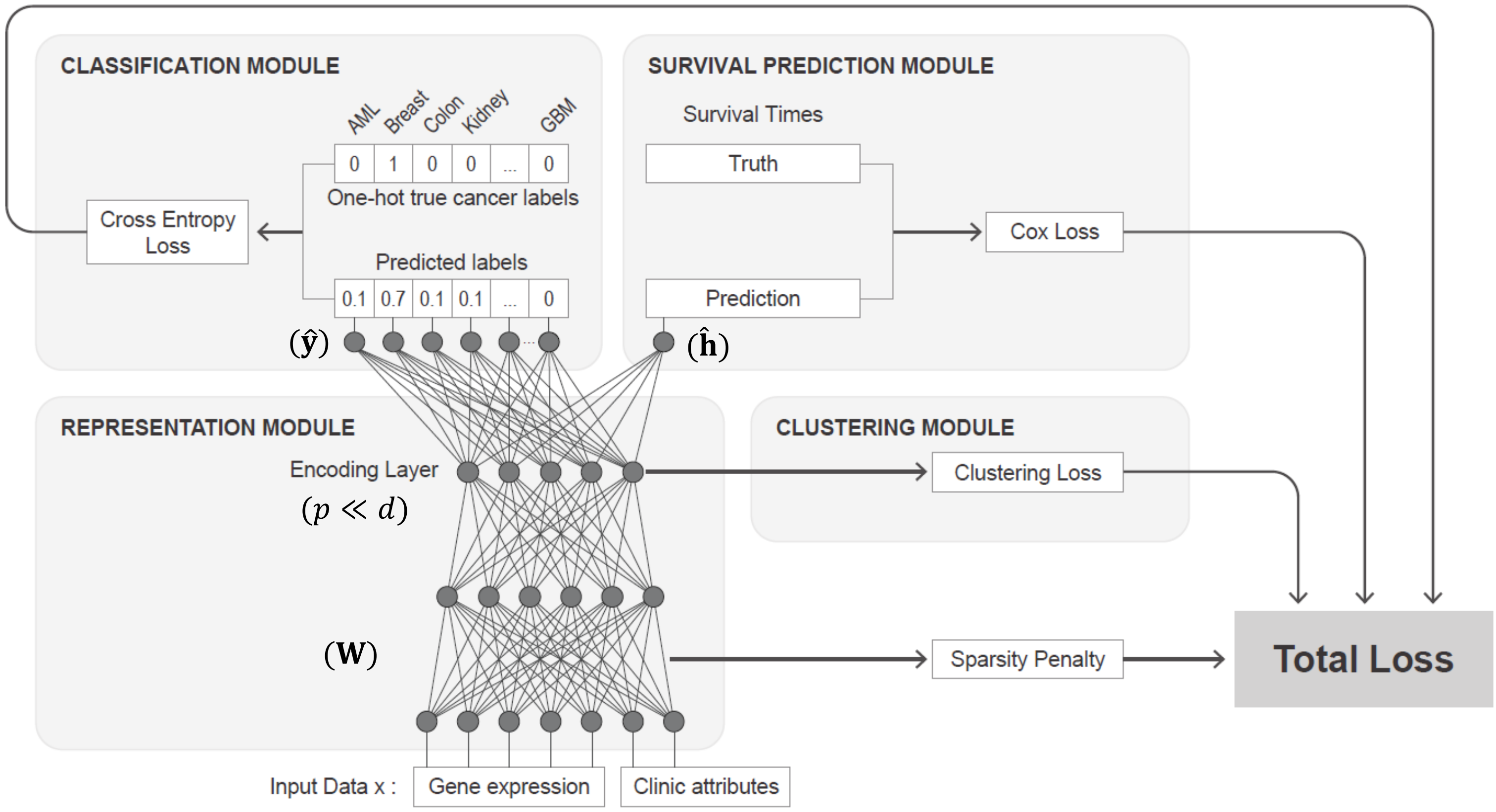
- Patient *i* is defined to be a cross cancer patient if:
  - The patient is co-cluster with another patient *j* with different cancer type *C* over multiple runs of clustering (with different number of cluster). *
  - The patient is closer to the patients with cancer type *C* than the patients of its own cancer type.

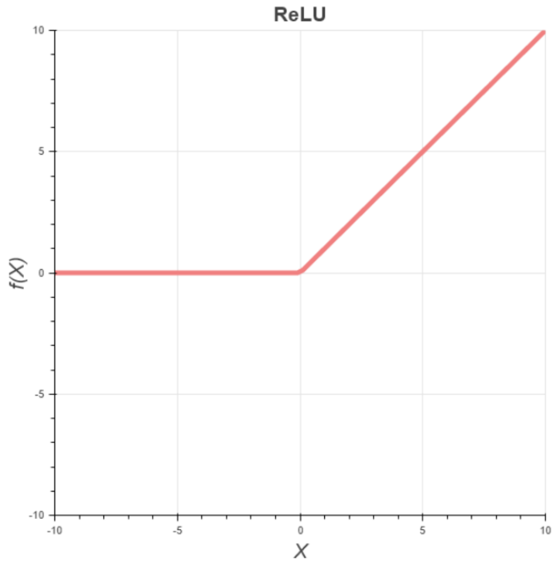
- The distance between two patient is defined to be the proportion of the two being co-cluster in multiple runs of clustering.



* The threshold is defined to yield maximum precision.

Method
Model design (1)

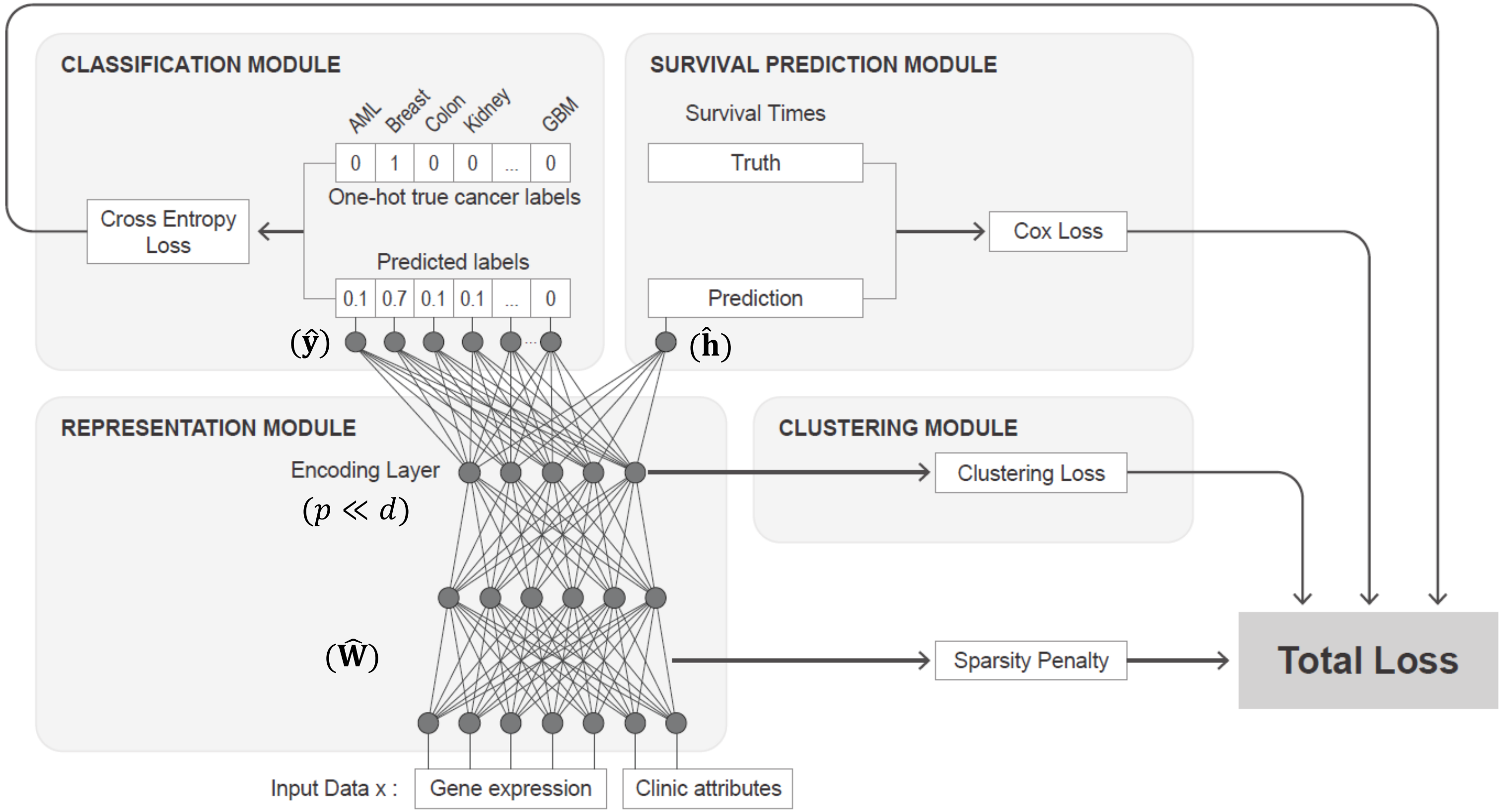




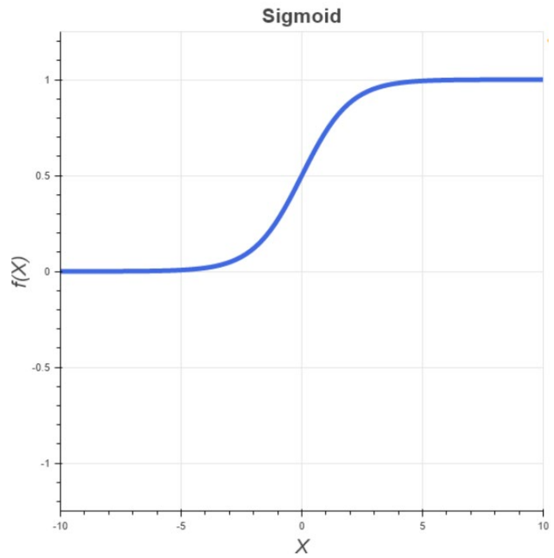
$$
\begin{aligned}
\mathbf{o}_1 &= \mathrm{ReLU}\left(\mathbf{W}_1\mathbf{X} + \mathbf{b}_1\right), \\
\mathbf{o}_i &= \mathrm{ReLU}\left(\mathbf{W}_i\mathbf{o}_{i-1} + \mathbf{b}_i\right), 2 \le i \le M, \\
\hat{\mathbf{y}} &= \mathrm{softmax}\left(\mathbf{W}_{M+1}\mathbf{o}_M + \mathbf{b}_{M+1}\right), \\
\hat{\mathrm{h}} &= \mathrm{sigmoid}(\mathbf{W}_{M+1}\mathbf{o}_M + \mathbf{b}_{M+1}).
\end{aligned}
$$

# of neurons in hidden layer: 32, 16

Method
Model design (2)



# of neurons in hidden layer: 32, 16



$$\mathbf{o}_1 = \mathrm{ReLU}\left(\mathbf{W}_1\mathbf{X} + \mathbf{b}_1\right),$$
$$\mathbf{o}_i = \mathrm{ReLU}\left(\mathbf{W}_i\mathbf{o}_{i-1} + \mathbf{b}_i\right), 2 \leq i \leq M,$$
$$\hat{\mathbf{y}} = \mathrm{softmax}\left(\mathbf{W}_{M+1}\mathbf{o}_M + \mathbf{b}_{M+1}\right),$$
$$\hat{\mathbf{h}} = \mathrm{sigmoid}(\mathbf{W}_{M+1}\mathbf{o}_M + \mathbf{b}_{M+1}).$$

Method
Loss function

- Objective Function

$$\min_{\{\Theta,\mathbf{Q},\mathbf{U}\}} L_{\text{classification}} + \alpha L_{\text{clustering}} + \beta L_{\text{survival}} + \lambda L_{\text{sparsity}}$$

- Classification Loss

$$L_{\text{classification}} = -\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ji} \log \hat{y}_{ji}$$

- Negative log likelihood of multinomial distribution
- Cross entropy

- Sparsity Loss

$$L_{\text{sparsity}} = \|\mathbf{W}_1^{\top}\|_1$$

- L1 regularization

- Clustering Loss

$$L_{\text{clustering}} = \sum_{i=1}^{n} \|z_i - \mathbf{U}\mathbf{q}_i\|_2^2, \text{ subject to } \sum_{j=1}^{k} q_{ji} = 1, q_{ji} \in \{0,1\}, \forall j, \forall i$$

- Distance to the cluster centroid **U**

- Survival Loss

$$L_{\text{survival}} = \sum_{i:c^{(i)}=1} \left(\log \hat{\mathrm{h}}^{(i)} - \log \sum_{j:t^{(j)} \geq t^{(i)}} e^{\hat{\mathrm{h}}^{(j)}} \right)$$

- The interpretable variable act proportionally to the risk:

$$h(t|x) = h_0(t)e^{\hat{h}} \qquad \hat{h} = f(x)$$

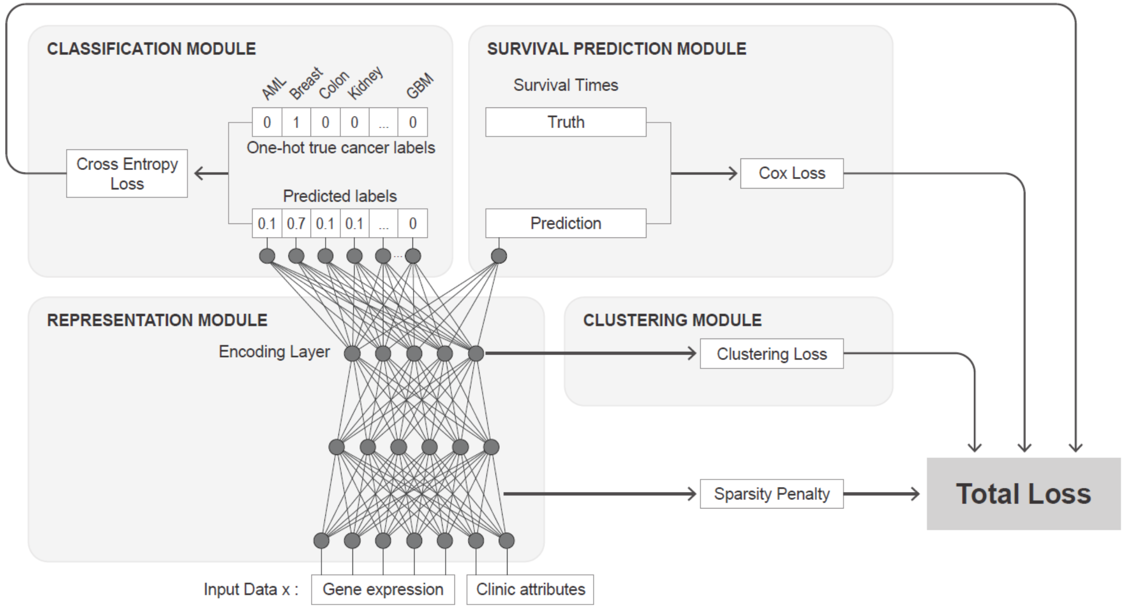
Define risk set R_j (the samples that haven't occur the event right before time t_j). The conditional probability for a sample occurring the event at t_j is:

$$\frac{h_0(t_j)exp(h_i)}{\sum_{k\in R_j} h_0(t_j)exp(h_k)} = \frac{exp(h_i)}{\sum_{k\in R_j} exp(h_k)}$$

The negative log likelihood is the Cox partial likelihood (survival loss)

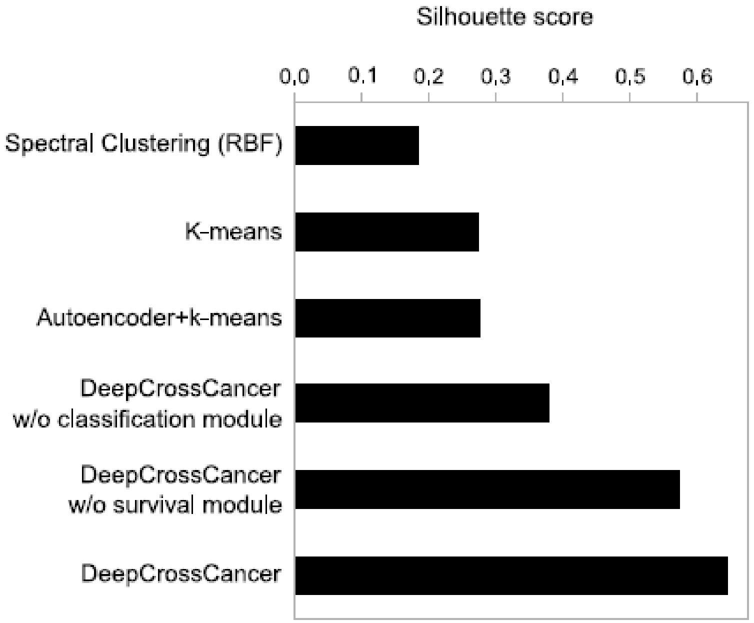
Method
Optimization (training)

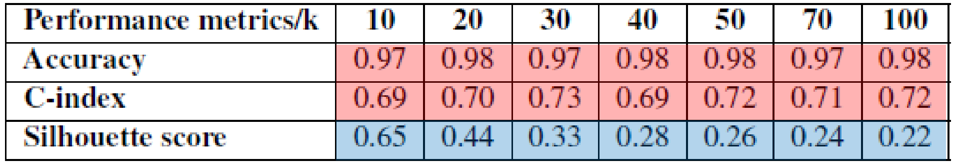
- Parameter
  - Ignore clustering loss first. $(\alpha = 0)$
  - Iterate (pretrain, find optimal $\mathbf{W}, \mathbf{b}, \beta, \lambda$):
    - Forward propagation.
    - Backpropagation in classification and survival modules, tune β, λ.
    - Fix $(\mathbf{W}, \mathbf{b})$, run K-means and acquire the cluster and centroid for each point.
    - Fix centroid and cluster, reiterate.
  - Fix $\mathbf{W}, \mathbf{b}, \beta, \lambda$, find α.
- Hyperparameter (α, β, λ)
  - 10-fold cross validation $(\alpha \leftarrow \ \beta \leftarrow \lambda)$
    - Random Search
    - Probability Reduction



Mini-batch gradient descent with Adam and SGD

Result
Cluster Evaluation (1)





Performance metrics/k	10	20	30	40	50	70	100
Accuracy	0.97	0.98	0.97	0.98	0.98	0.97	0.98
C-index	0.69	0.70	0.73	0.69	0.72	0.71	0.72
Silhouette score	0.65	0.44	0.33	0.28	0.26	0.24	0.22

- Silhouette score

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \text{ if } |C_i| > 1$$

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i,j)$$ (mean intra-cluster distance)

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j)$$ (minimum inter-cluster distance)

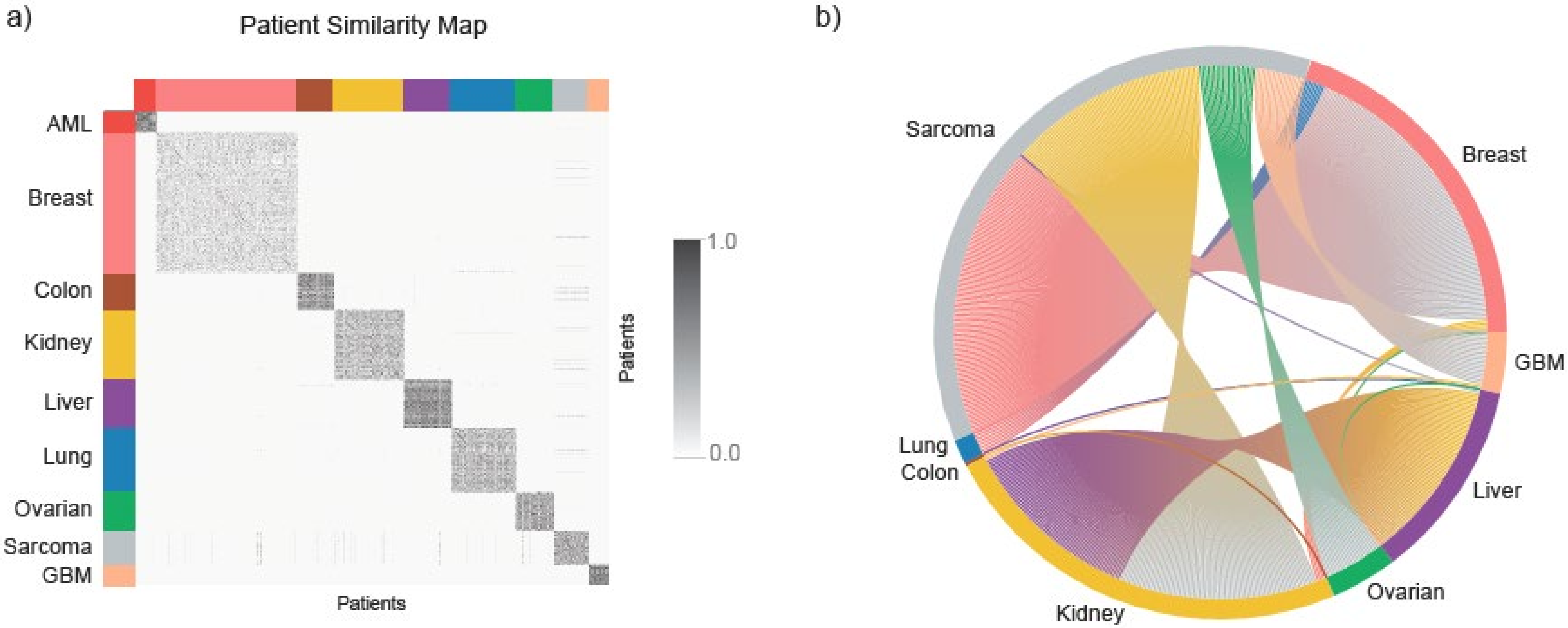
the cluster is defined to be the diagnosed cancer type

- Concordance index

$$\text{C-index} = \frac{\sum_{i,j} 1_{T_j < T_i} \cdot 1_{\eta_j > \eta_i} \cdot \delta_j}{\sum_{i,j} 1_{T_j < T_i} \cdot \delta_j}$$

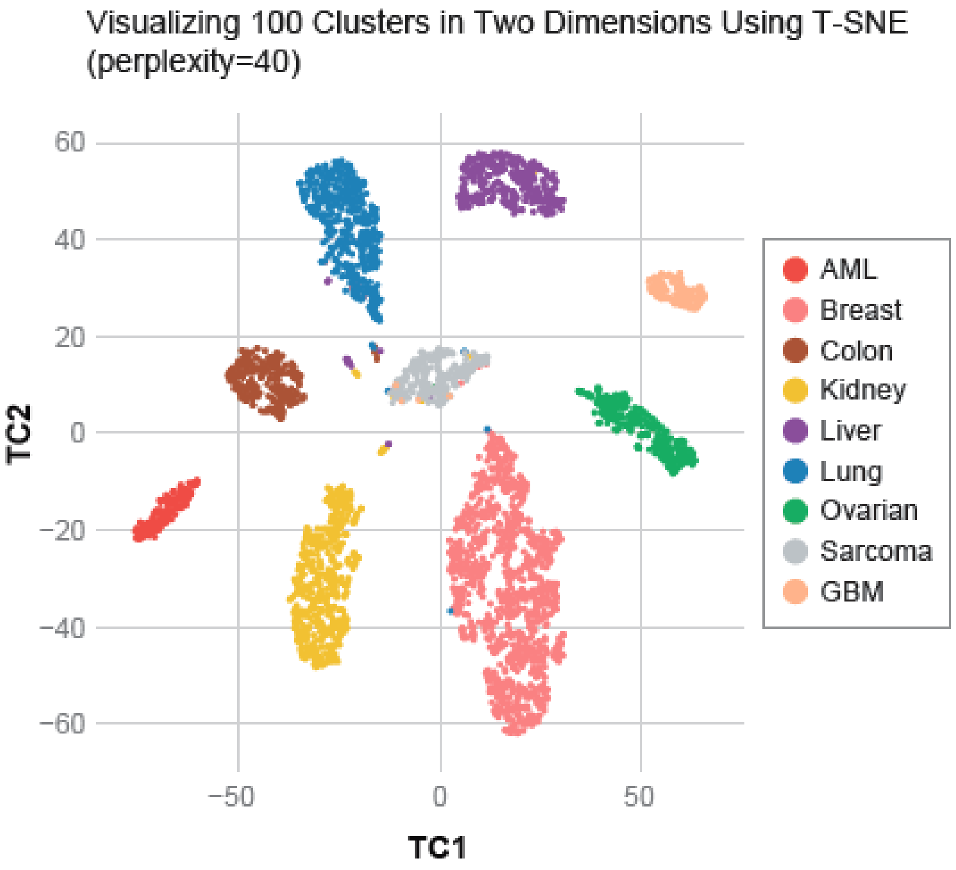
whether or not the prediction and the target show the same trend (if rank is the same)

Result
Cluster Evaluation (2)

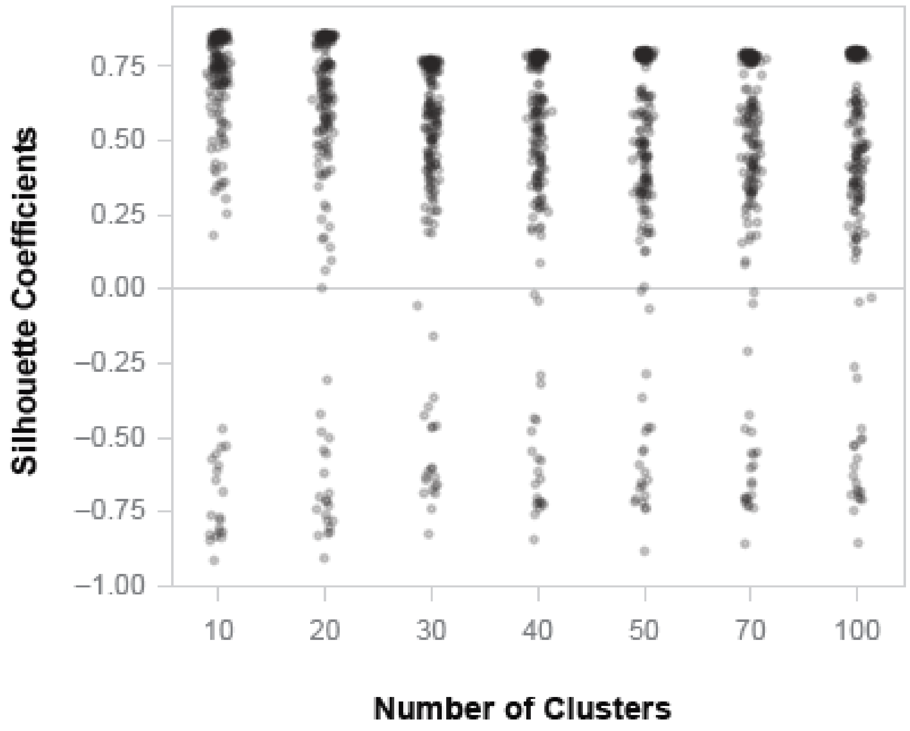


Result
Cluster Evaluation (3)

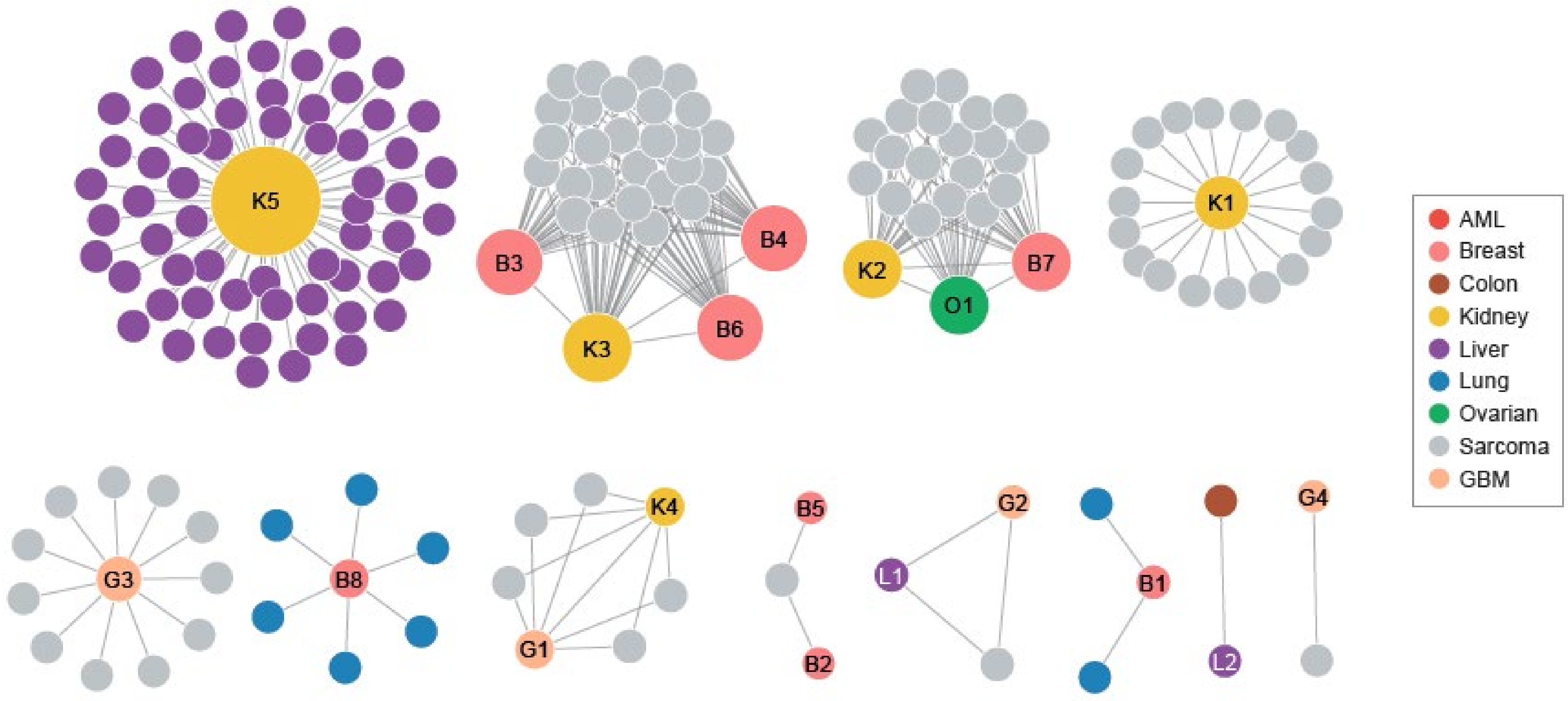
c)



d)



Result
Cross Cancer Patient



Result
Shared Predictive Genes

- Infer the contributions of genes (expression) for clustering
  - SHAP (DeepExplainer)
- Permutation test
  - t-statistics on the distribution of the number of common genes.
  - Patients similar to a cross cancer patient *i.*
  - Patients with the same diagnosed cancer type of patient *i*.
  - Benjamini and Hochberg (B&H) correction.
- Result
  - The kidney patient (K5) shares 13 common predictive genes with 63 liver patients.
  - The number of common genes is always bigger compared to randomly selected 63 kidney patients (p-value: 0.0001).
  - 8 cross-cancer patients sharing a significantly large number of genes (p-value: 0.05)

Common genes are not clearly defined (Algorithm 4 is missing).
Detail list of the shared predicted genes is missing (Supplementary file 4).