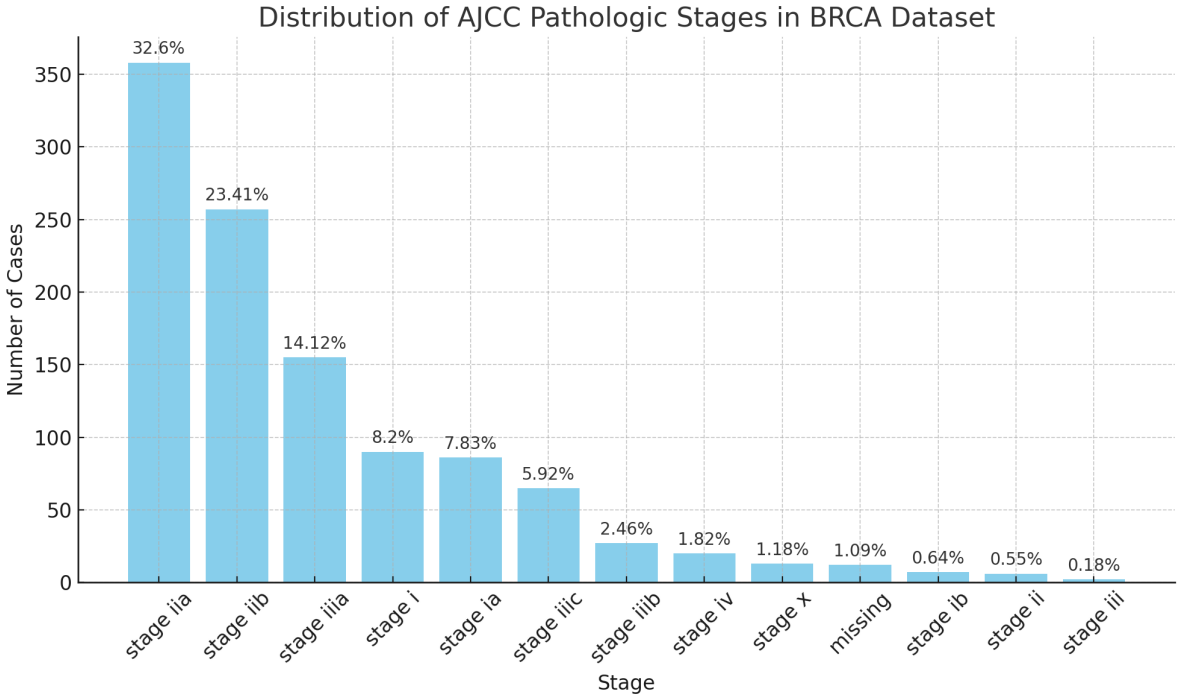
Supplementary Information for Exploiting Common Patterns in Diverse Cancer Types via Multi-Task Learning

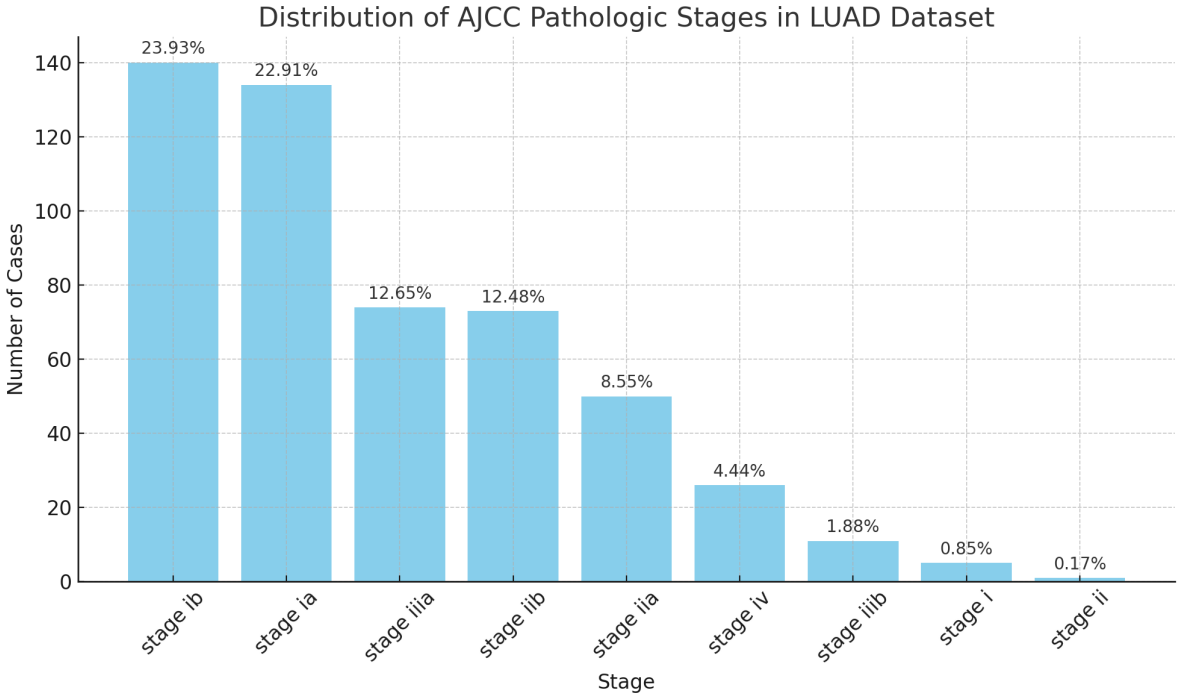
Supplementary Section 1. AJCC TNM Classification of analyzed tissues

The following figures show the frequency of various cancer stages as defined by the American Joint Committee on Cancer (AJCC) in our four datasets (BRCA, LUAD, COAD, and SCLC). Each bar represents a different cancer stage, including subdivisions for some stages (like IIA and IIB). The bars are labeled with the number of cases and the percentage of the total cases that each stage represents for a particular dataset. For all three cancers, we see a higher rate of samples of early cancer stages. As early-stage cancers are typically associated with a better prognosis and a higher likelihood of successful treatment outcomes, it is natural that most of our samples correspond to the "good prognosis" group.

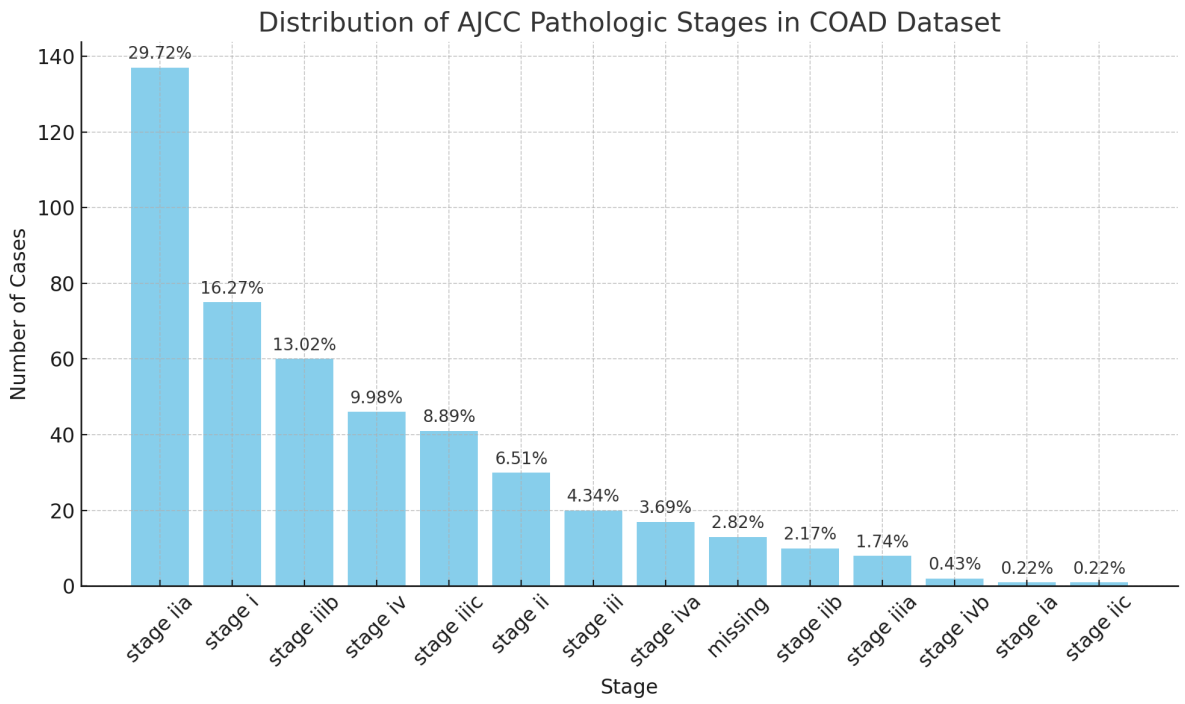
Supplementary Section 1.0.1. Overall AJCC TNM distribution for the three main datasets



Supplementary Figure 1: AJCC Pathologic Stages Distribution for BRCA.



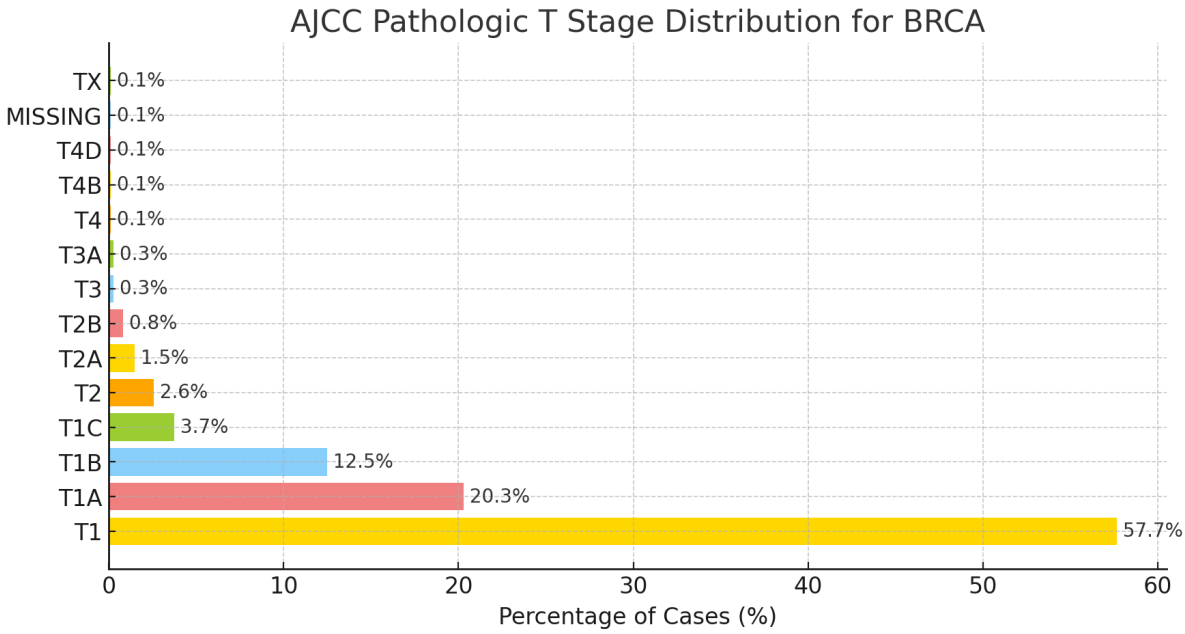
Supplementary Figure 2: AJCC Pathologic Stages Distribution for LUAD.



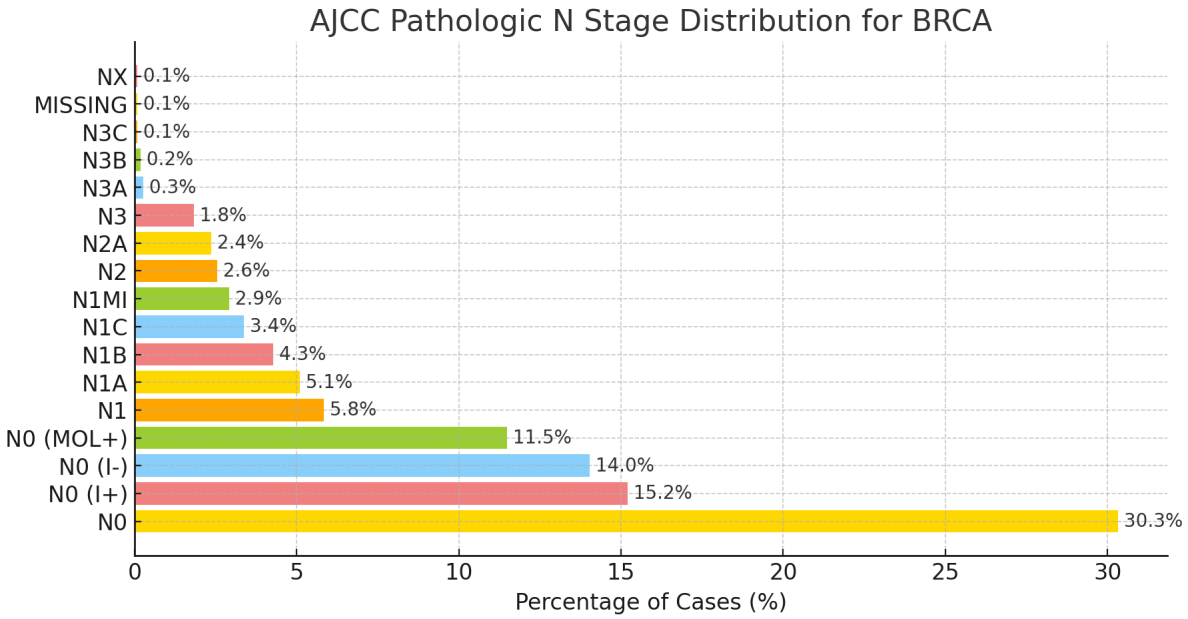
Supplementary Figure 3: AJCC Pathologic Stages Distribution for COAD.

A more detailed breakdown of the AJCC TNM distribution of each type of cancer can be found in the following subsections.

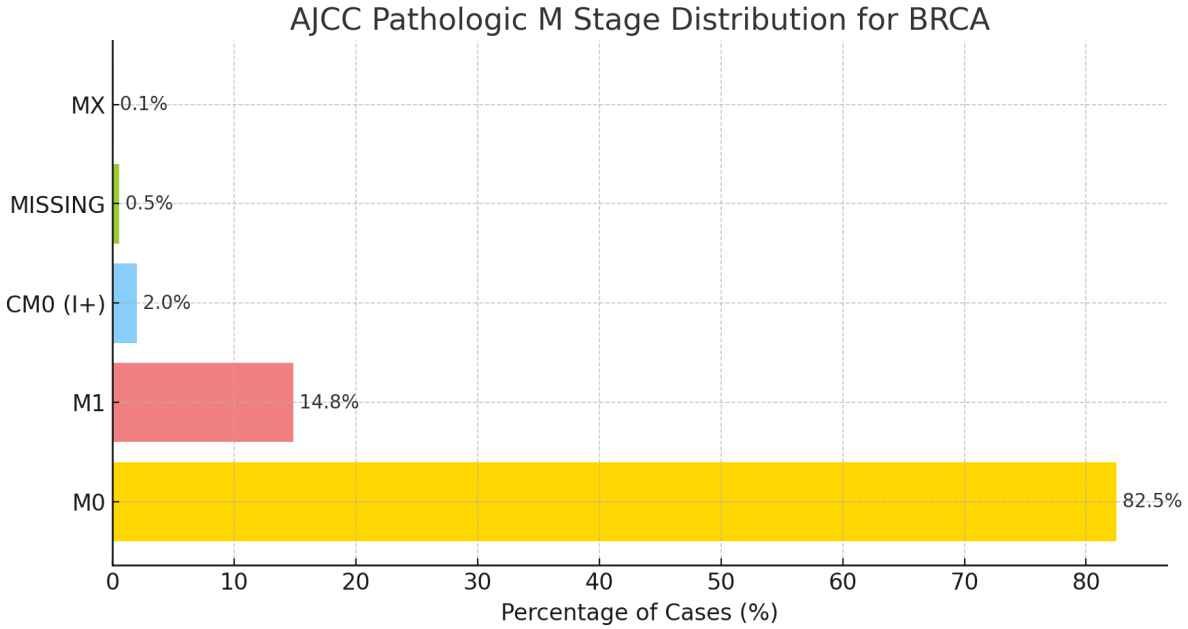
Supplementary Section 1.0.2. BRCA AJCC TNM distribution



Supplementary Figure 4: AJCC Pathologic T Stage Distribution for BRCA.

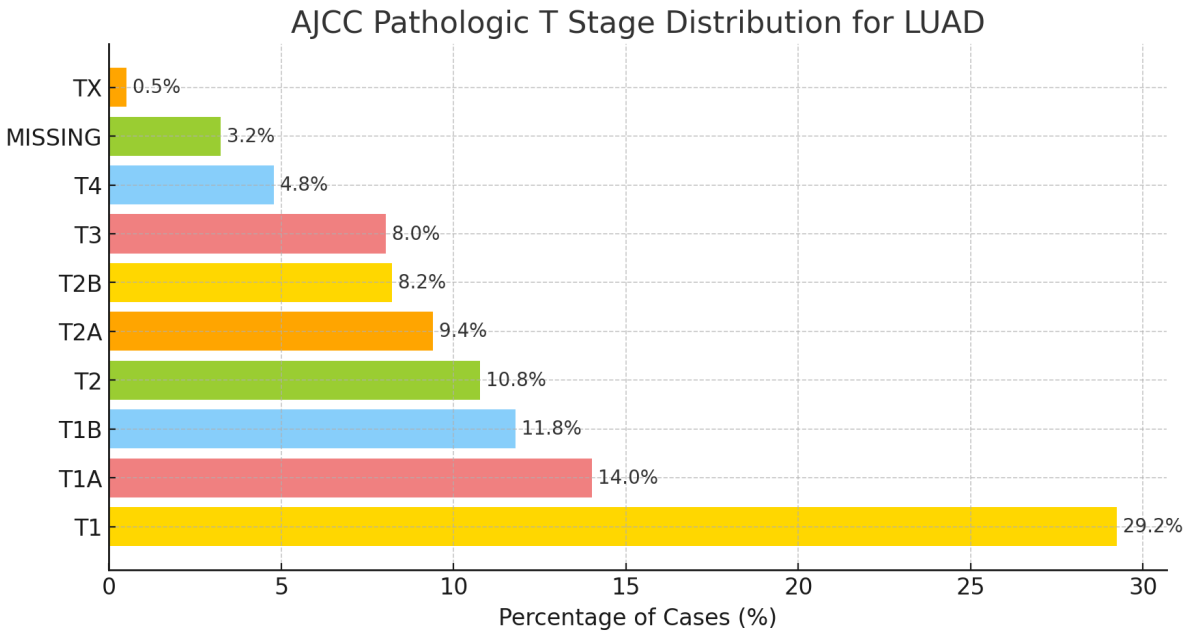


Supplementary Figure 5: AJCC Pathologic N Stage Distribution for BRCA.

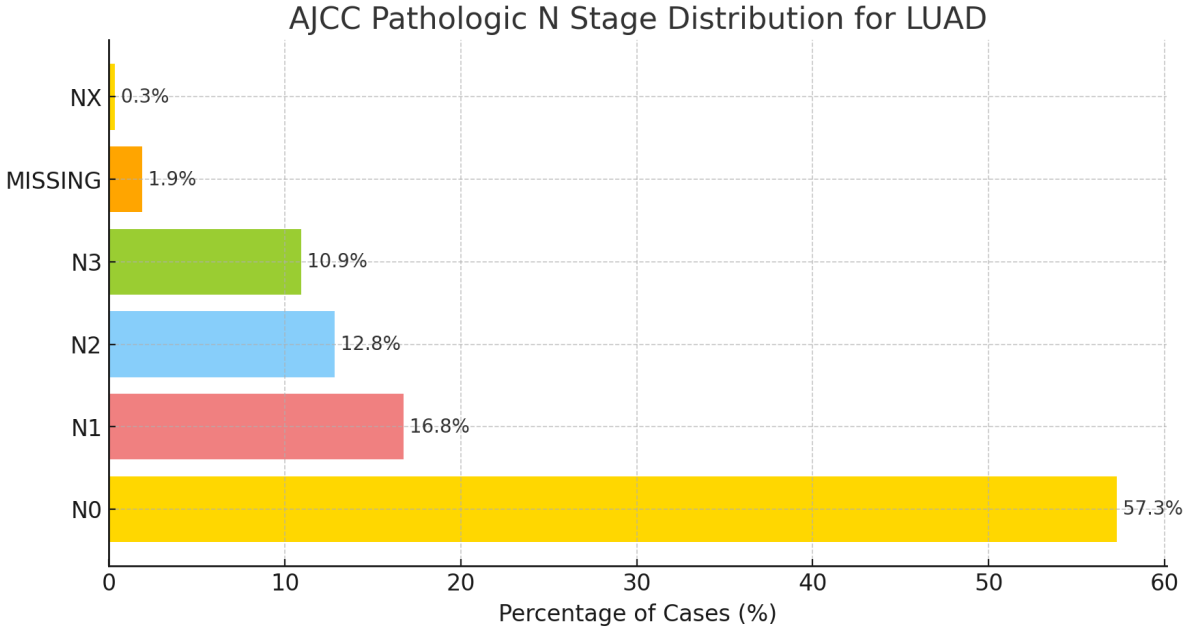


Supplementary Figure 6: AJCC Pathologic M Stage Distribution for BRCA.

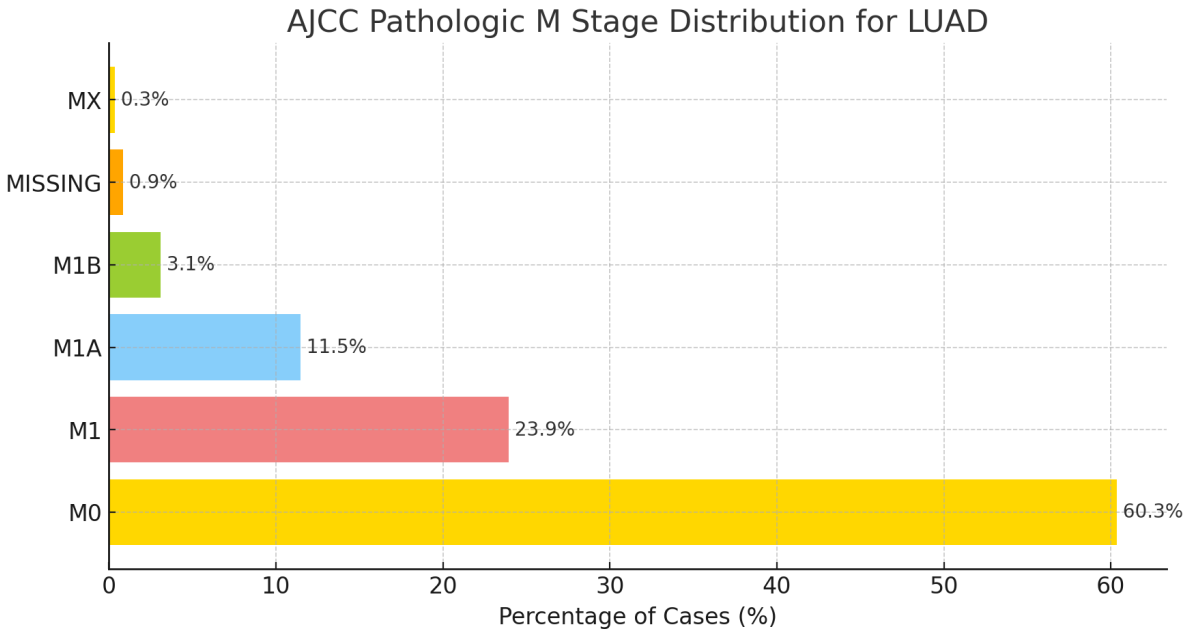
Supplementary Section 1.0.3. LUAD AJCC TNM distribution



Supplementary Figure 7: AJCC Pathologic T Stage Distribution for LUAD.

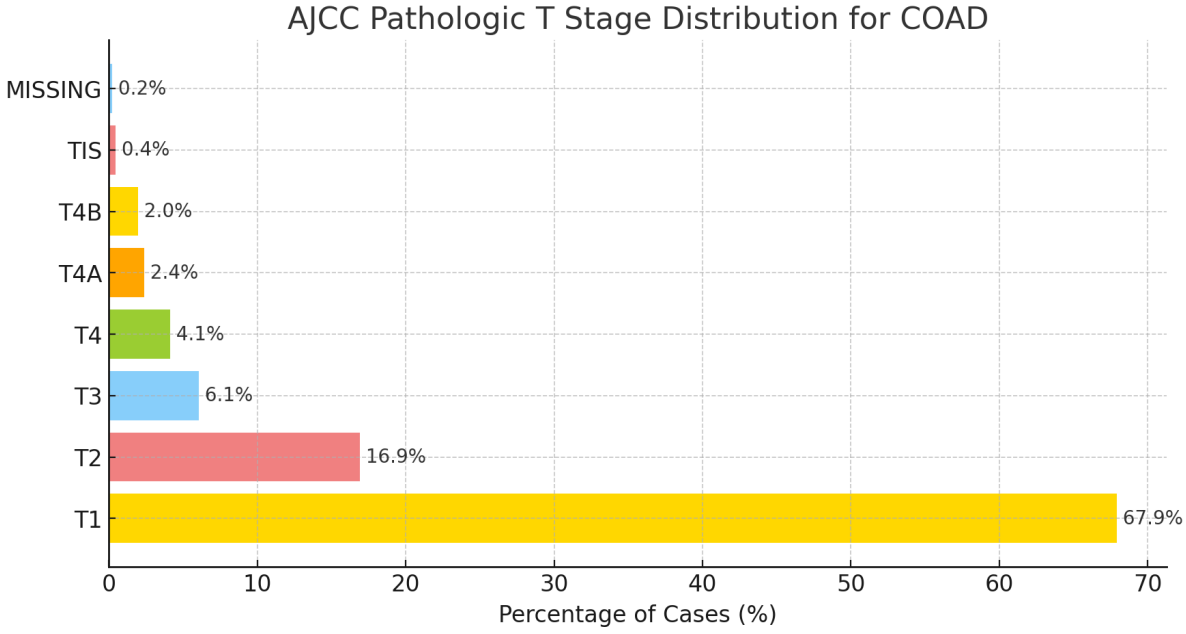


Supplementary Figure 8: AJCC Pathologic N Stage Distribution for LUAD.

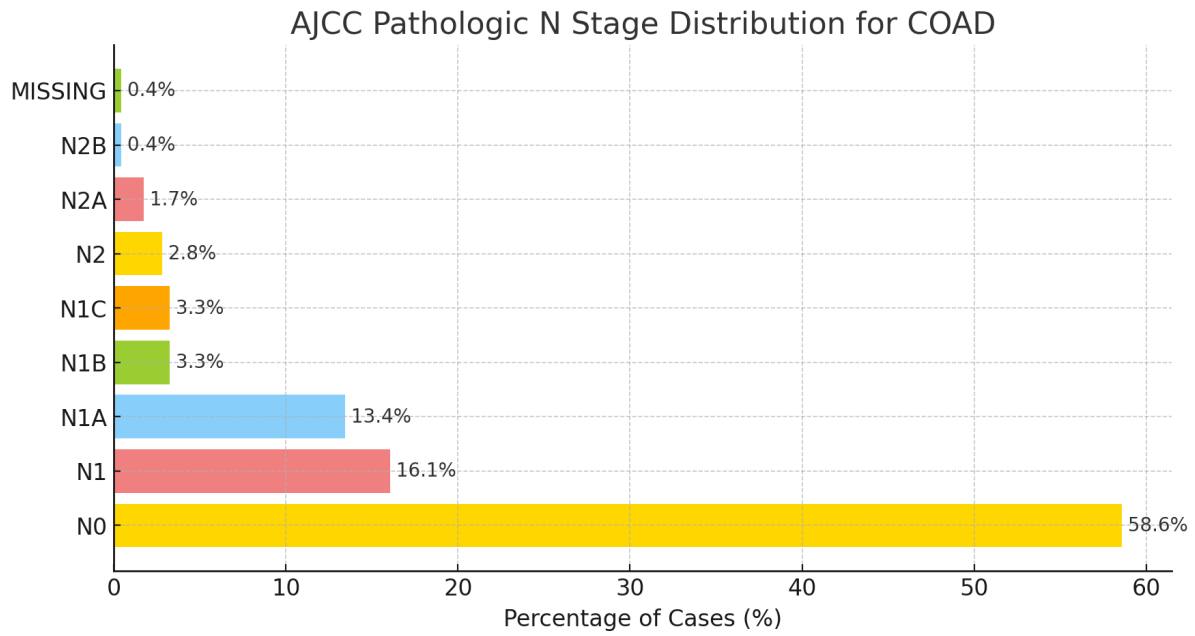


Supplementary Figure 9: AJCC Pathologic T Stage Distribution for LUAD.

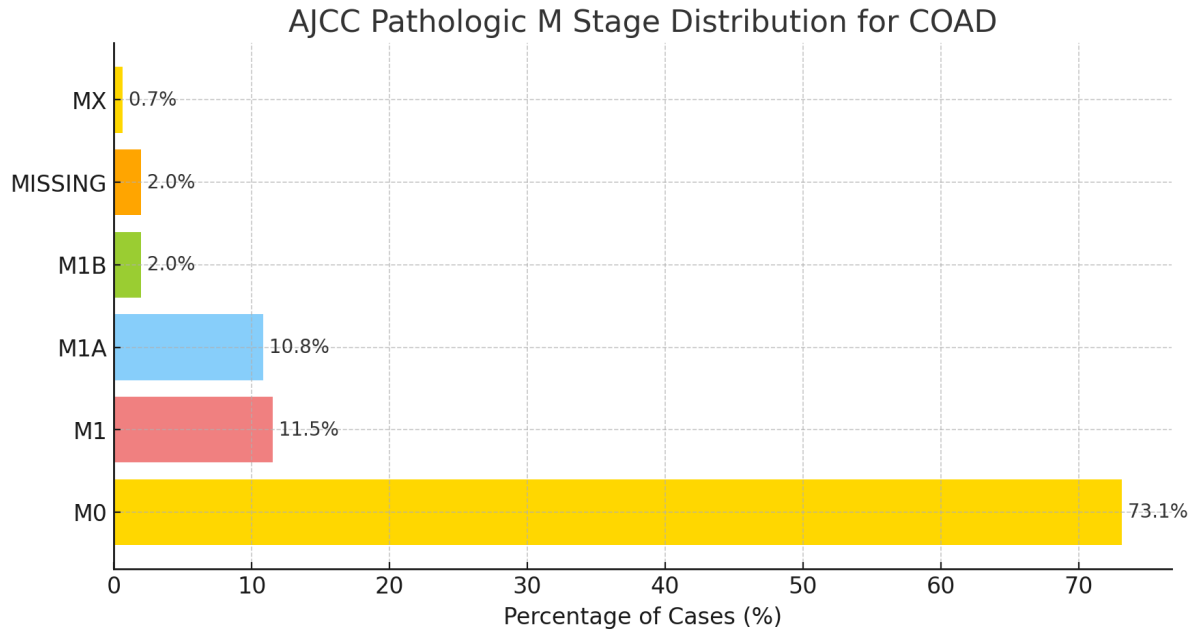
Supplementary Section 1.0.4. COAD AJCC TNM distribution



Supplementary Figure 10: AJCC Pathologic T Stage Distribution for COAD.



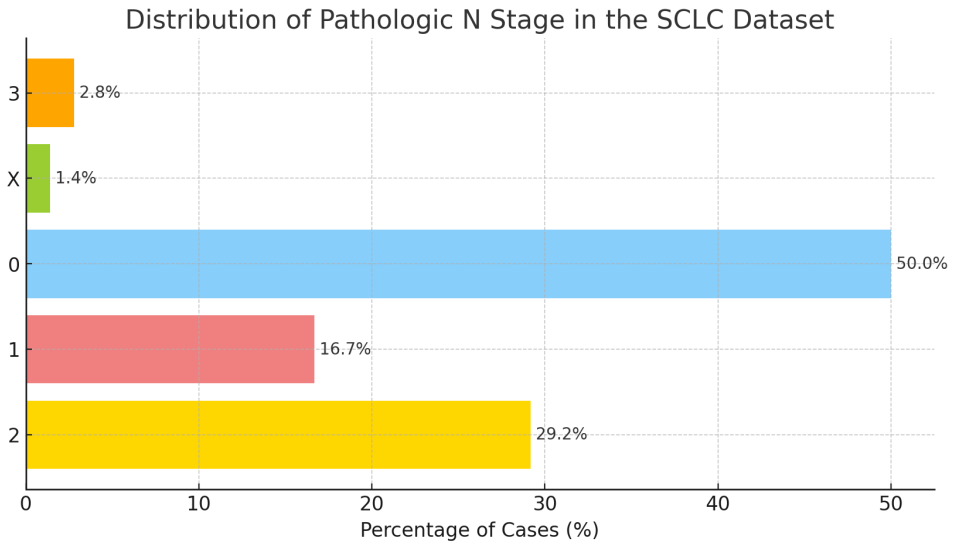
Supplementary Figure 11: AJCC Pathologic N Stage Distribution for COAD.



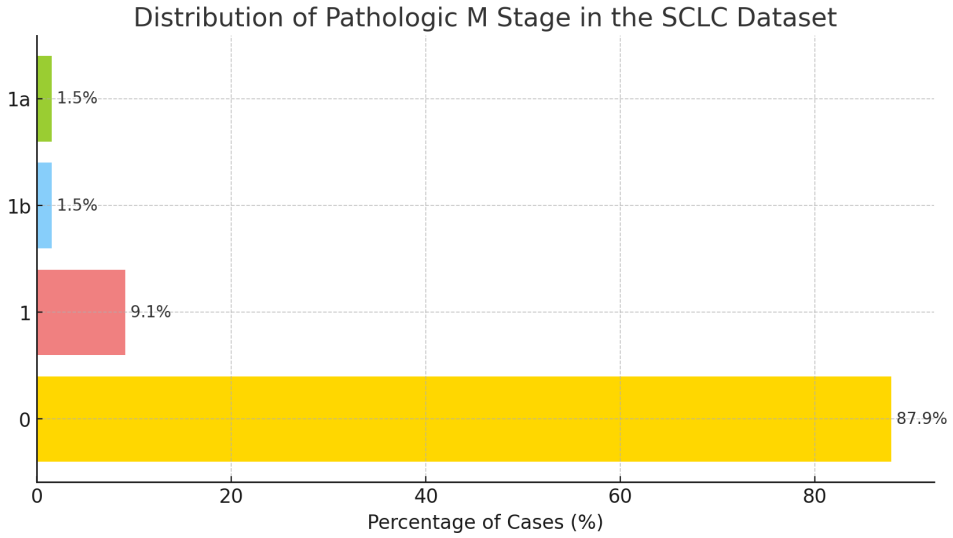
Supplementary Figure 12: AJCC Pathologic M Stage Distribution for COAD.

Supplementary Section 1.0.5. External dataset: SCLC TNM distribution

Only the N and M stage distribution information was available for the SCLC dataset. This distribution is shown below.



Supplementary Figure 13: Pathologic N Stage Distribution for SCLC.



Supplementary Figure 14: Pathologic M Stage Distribution for SCLC.

Supplementary Section 2. Cross-Validation Results

Supplementary Section 2.1. 4-fold cross-validation

Table 1 illustrates the results obtained during the 4-fold cross-validation used for hyperparameter tuning. One fold was used for validation (25% of the training data), and three were used for training the model. The data for all folds was preprocessed in the same fashion to facilitate the comparison between folds. We performed a grid search initially to find the optimal hyperparameters. These parameters are shown in Table 2.

Validation Set 4-fold Cross-Validation Results				
Dataset	**Learning Paradigm**	**AUPRC**	**AUROC**	**C-index**
BRCA	STL	**0.24626 ± 0.02672**	**0.73795 ± 0.06471**	**0.71560 ± 0.06270**
	MTL	0.21760 ± 0.02997	0.69847 ± 0.01888	0.68190 ± 0.01679
LUAD	STL	0.54837 ± 0.06076	0.70491 ± 0.05681	**0.65440 ± 0.06245**
	MTL	**0.58313 ± 0.07779**	**0.70696 ± 0.05114**	0.64196 ± 0.04364
COAD	STL	0.31190 ± 0.05596	0.56211 ± 0.08845	0.54905 ± 0.08274
	MTL	**0.37425 ± 0.07848**	**0.61926 ± 0.07862**	**0.60019 ± 0.07179**

Supplementary Table 1: Validation Set 4-fold Cross-Validation Results with Single and Multi-Task Learning Models for all three cancers.

Learning Rate	**Momentum**	**Batch Size**
0.001	0.8	64
0.001	0.9	128
0.001	0.95	256
0.01	0.8	64
0.01	0.9	128
0.01	0.95	256
0.1	0.8	64
0.1	0.9	128
0.1	0.95	256

Supplementary Table 2: Parameters used for the grid search during cross-validation

Supplementary Section 2.2. Nested cross-validation

Table 3 shows the results for a 5 by 5 nested cross-validation carried out in the following manner, following previous studies using a similar approach [**?**].

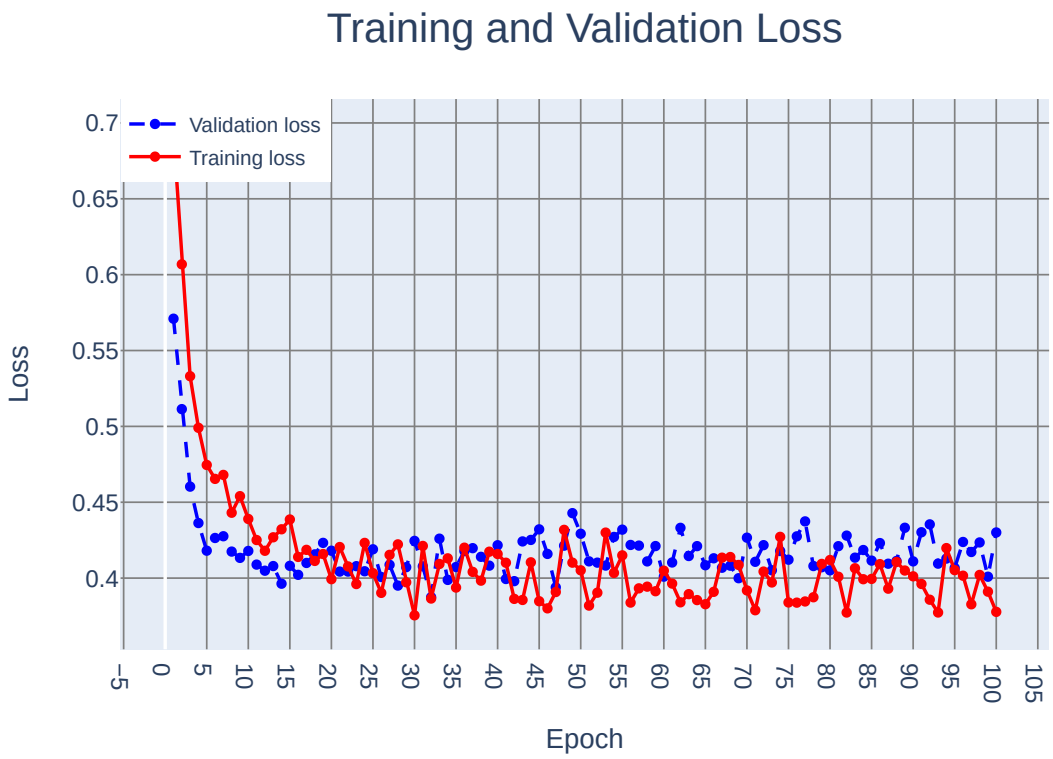
The dataset was split into two levels of cross-validation: outer and inner loops, each containing 5 folds. Within each outer fold, we conducted a 5-fold stratified inner cross-validation using the training and validation subsets. The outer training/validation set was then divided into inner training and validation sets for each inner fold. The model was trained on four of the inner training sets and validated on one inner validation set. Following the inner cross-validation, the best-performing model/configuration was chosen and tested on the outer loop.

5x5 Nested Cross-Validation Results				
Dataset	**Learning Paradigm**	**AUPRC**	**AUROC**	**C-index**
BRCA	STL	**0.27543 ± 0.11888**	**0.70779 ± 0.07496**	**0.69206 ± 0.06854**
	MTL	0.25909 ± 0.08423	0.73082 ± 0.05976	0.71273 ± 0.05837
LUAD	STL	0.57382 ± 0.12313	0.71029 ± 0.08026	**0.64967 ± 0.07134**
	MTL	**0.61475 ± 0.06920**	**0.73118 ± 0.05595**	0.66907 ± 0.04750
COAD	STL	0.35827 ± 0.08188	0.61725 ± 0.08638	0.60501 ± 0.07844
	MTL	**0.38061 ± 0.11741**	**0.60822 ± 0.04716**	**0.59511 ± 0.04561**

Supplementary Table 3: 5x5 Nested Cross-Validation Results with Single and Multi-Task Learning Models for all three cancers.

Supplementary Section 3. Training and Validation Loss

The following graph illustrates the training and validation loss fluctuations during 100 epochs. The x-axis shows the number of epochs, while the y-axis represents the loss. The red line refers to the training loss, and the blue line represents the validation loss. Initially, both losses decrease sharply. Using early-stopping, we chose 50 epochs as our cut-off mark, as the loss starts to plateau at about 45 epochs. Other metrics also show stagnation or decrease close to this mark. The graph indicates a good fit, as both losses remain close and show no divergence or overfitting.

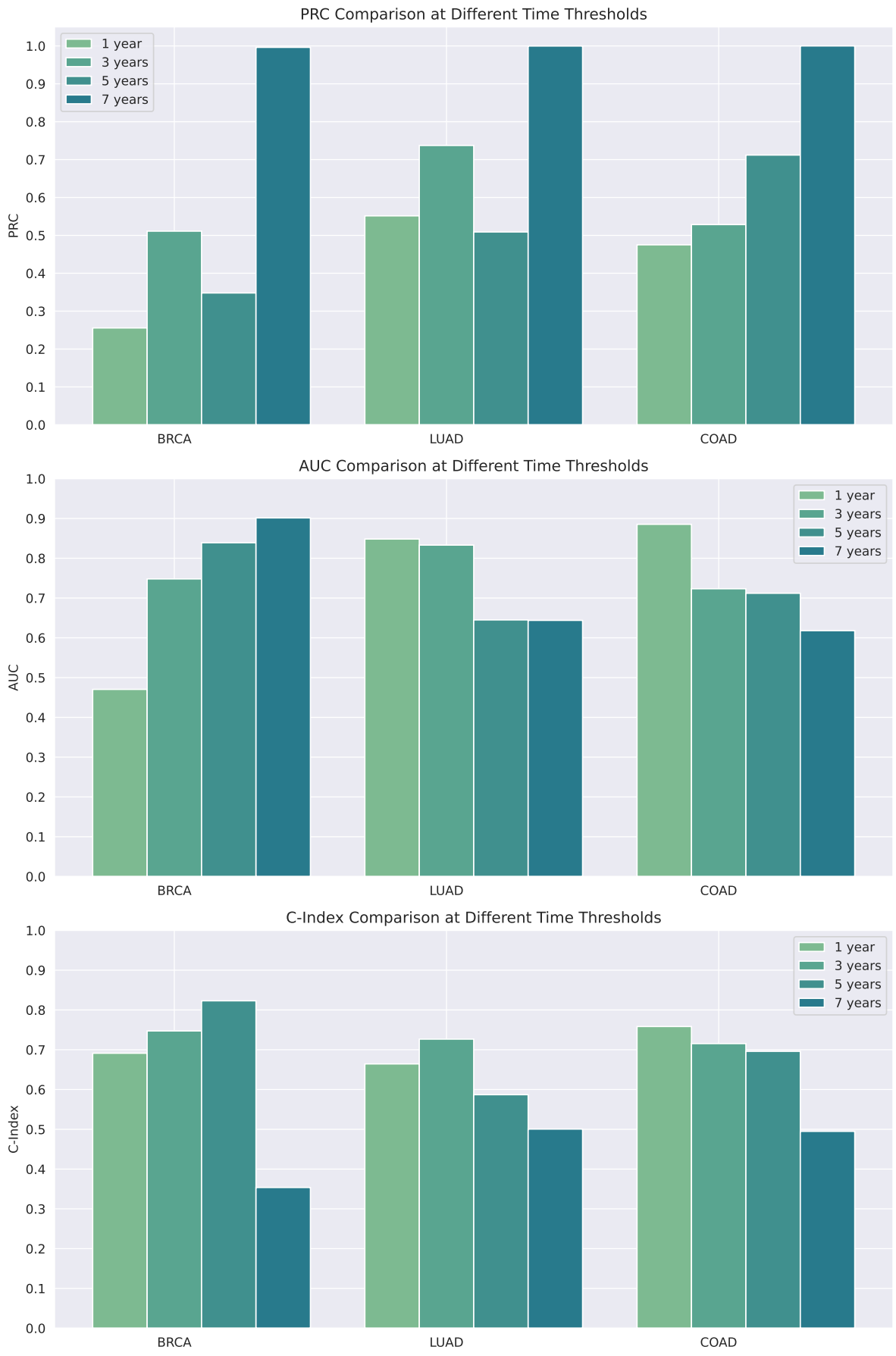


Supplementary Figure 15: Convergence of Training and Validation loss for the MTL model for 100 epochs.

Supplementary Section 4. Additional Ablation Studies

Supplementary Section 4.1. Changing Survival Cutoff Thresholds

We evaluated our model by varying the survival cutoff thresholds for the TCGA datasets. For each experiment, we trained our model using the specified threshold and then made predictions on the testing set with the same threshold. The results for each metric and dataset can be seen in the table below.



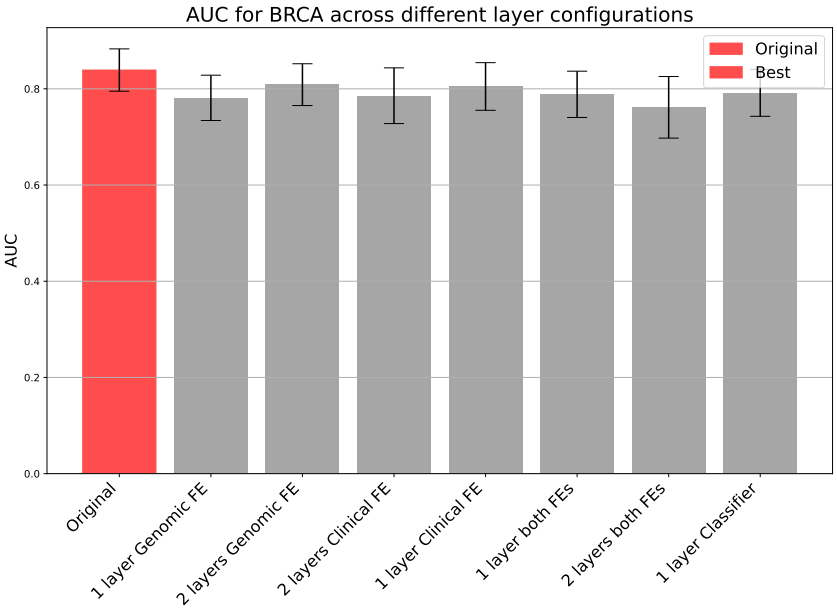
Supplementary Figure 16: **Metrics for the TCGA datasets at different month cutoff thresholds for determining their survival status.** The first subplot shows the variation of the AUPRC for different cutoffs. The second one shows the AUROC, and the third one displays the C-index.

The AUROC and C-index for LUAD and COAD indicate strong initial performance at 1 and 3 years, owing to more complete information and a higher number of surviving patients. However, there is some decline in predictive performance over longer durations, such as the 7-year interval. These results suggest that while the model performs well in the early stages, maintaining predictive accuracy over extended periods is challenging for cancers with higher recurrence rates and aggressiveness. This decline could be attributed to various factors, such as changes in patient follow-up or disease progression over time. Therefore, we conclude that the most reliable models use 3- or 5-year thresholds.

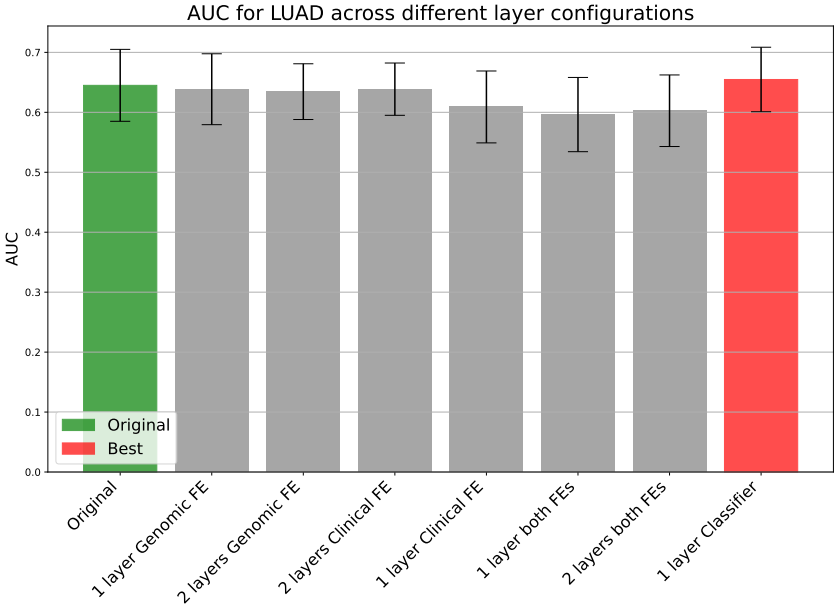
Supplementary Section 4.2. Removing layers from the Feature Extractors and Classifier

In this ablation study, we examined the impact of reducing the number of layers in the genomic and clinical feature extractors and the classifier. "Original" denotes the configurations depicted in Figures 5 and 6 in the main article. Alternative configurations are labeled according to their modifications while keeping the rest of the architecture unchanged from the "Original." For example, "1 layer Genomic FE" means a single-layer genomic feature extractor, with all other components remaining the same as in the "Original" setup.

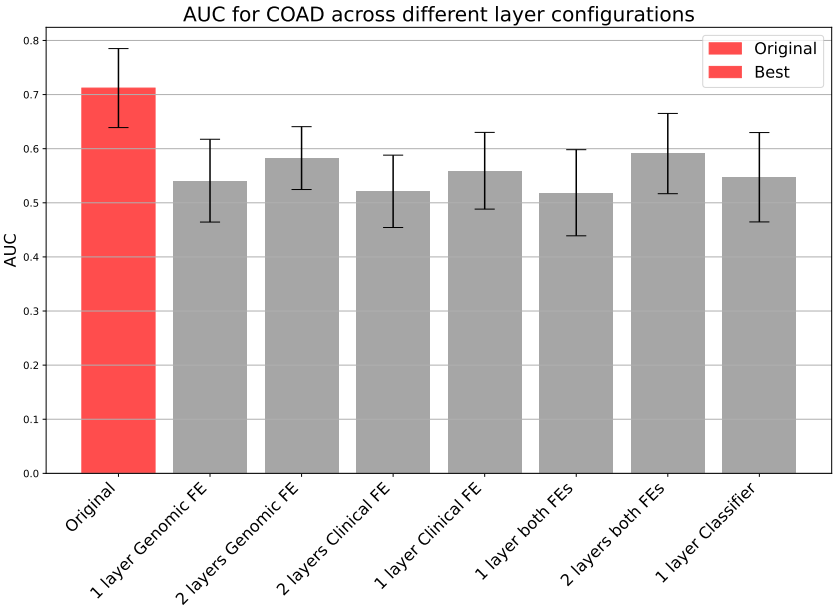
For BRCA and COAD, the "Original" configuration generally yielded the best results. However, for LUAD, both the AUROC and C-index benefited from reducing the number of layers in the classifier. In the case of BRCA, fewer layers in the Clinical FE improved the AUPRC. These findings suggest that while the "Original" setup suits this specific set of cancers, optimal layer configurations may vary across different combinations, indicating the need for tailored layer adjustments based on the cancers selected for training.



[a]

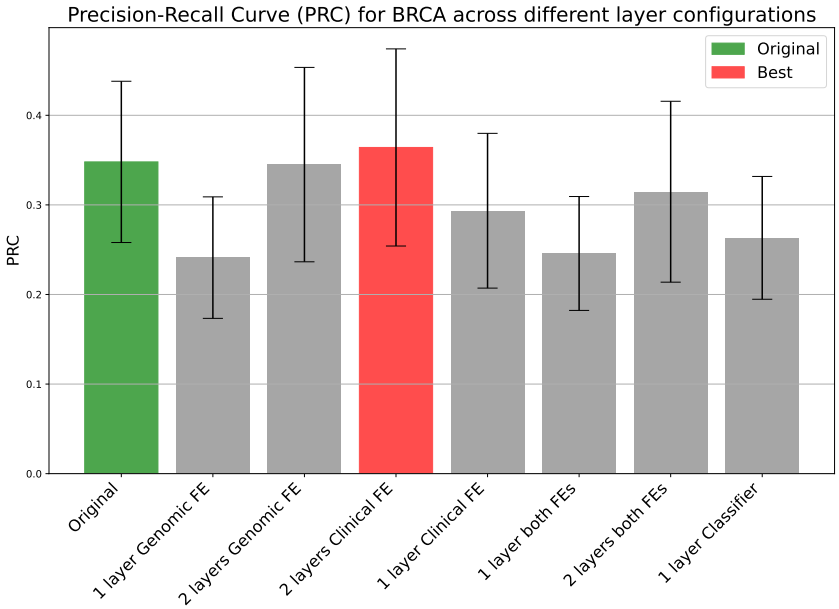


[b]

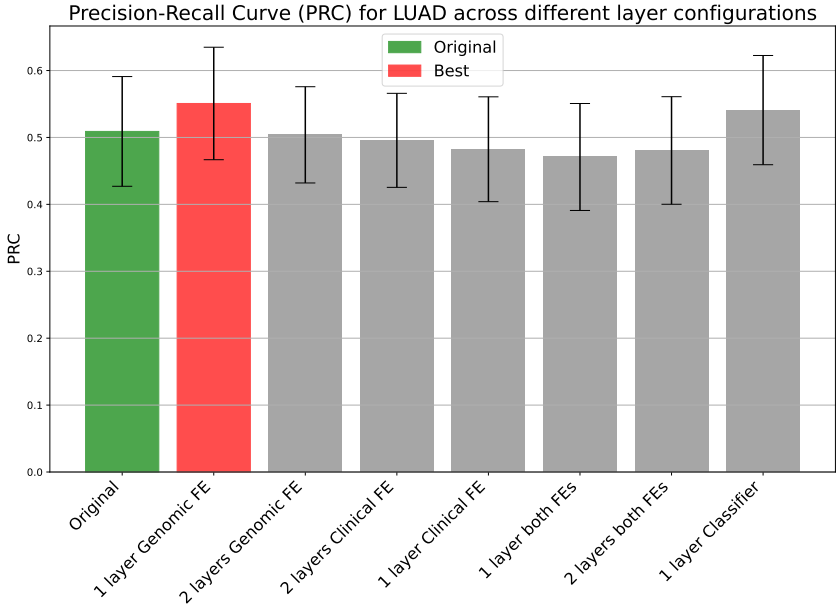


[c]

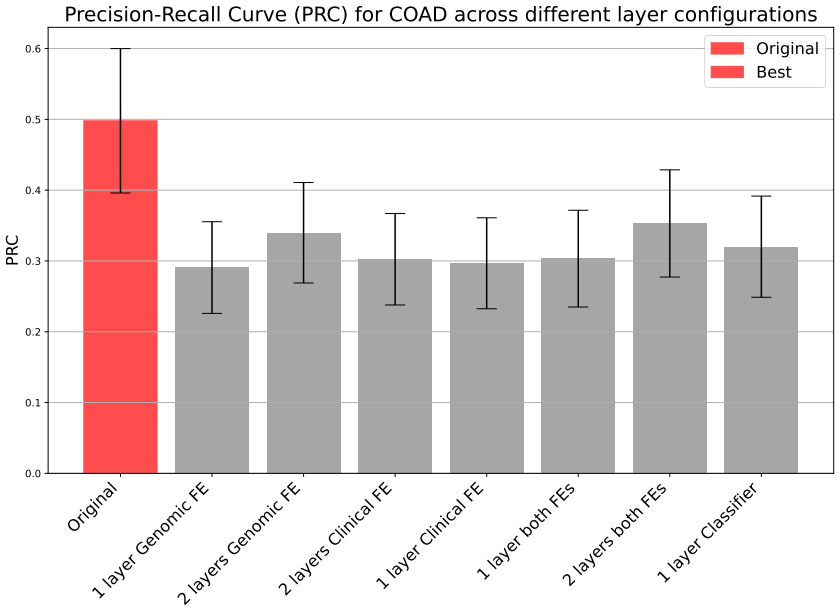
Supplementary Figure 17: AUROC variation for BRCA, LUAD, and COAD using different layer configurations.



[a]

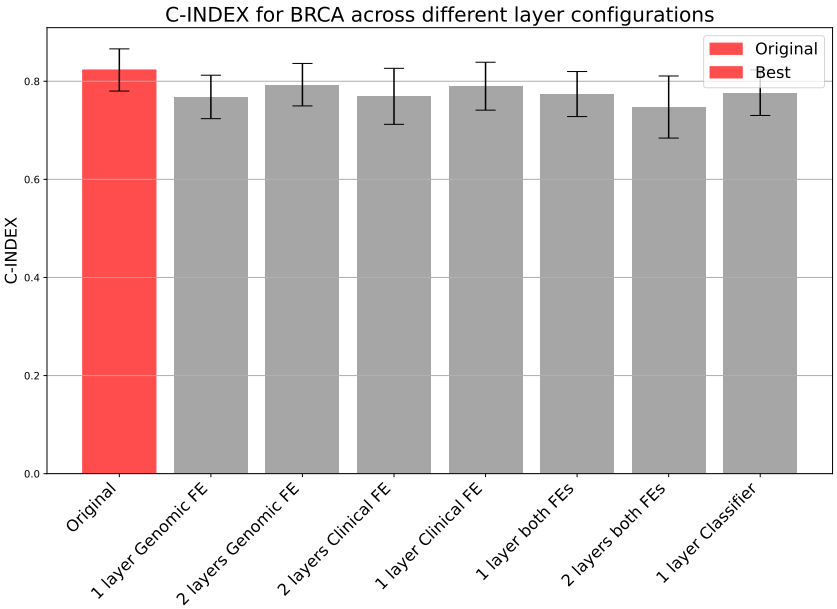


[b]

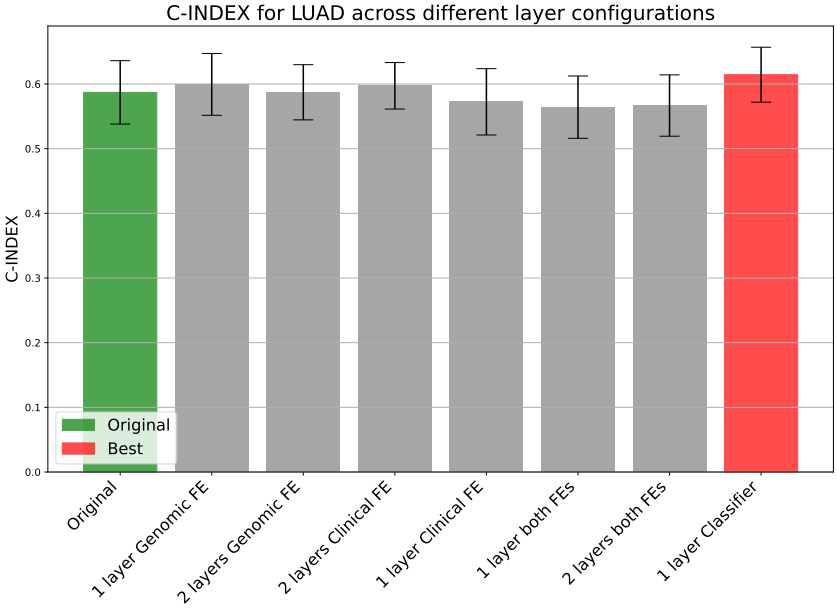


[c]

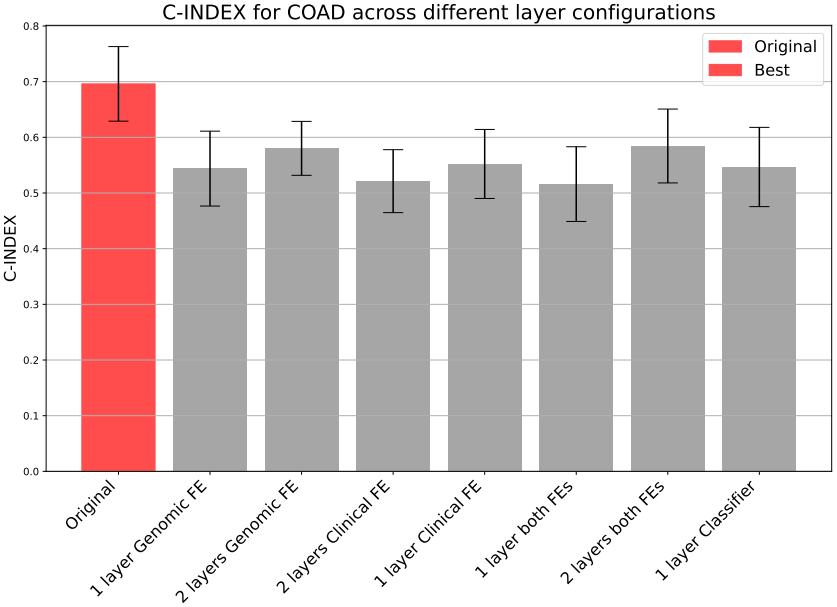
Supplementary Figure 18: AUPRC variation for BRCA, LUAD, and COAD using different layer configurations.



[a]



[b]



[c]

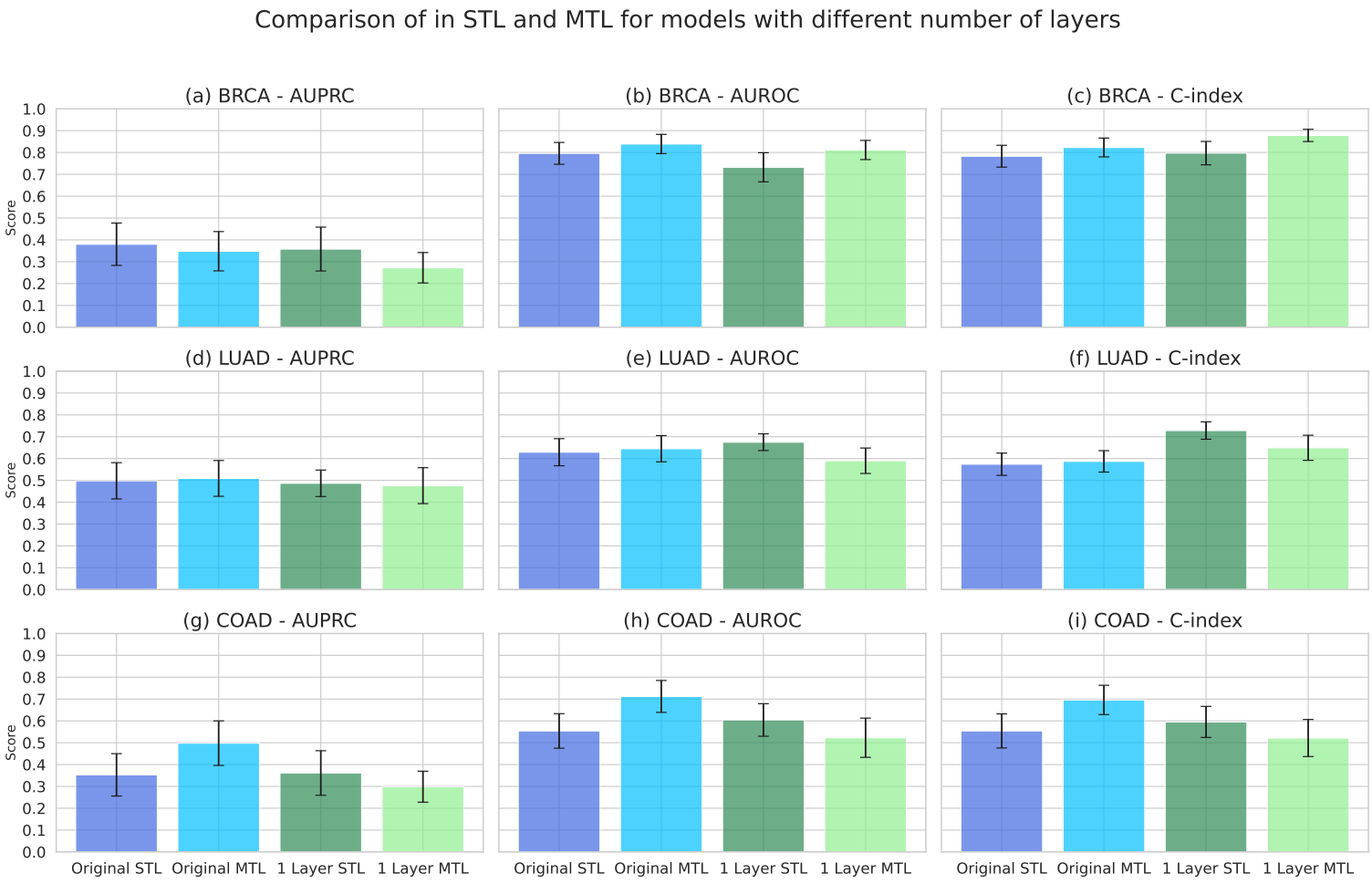
Supplementary Figure 19: C-index variation for BRCA, LUAD, and COAD using different layer configurations.

To test our model further, we reduced the number of layers to a minimum. The next table showcases the effects of this simplified configuration.

Datasets	**Learning Paradigms**	**AUPRC**	**AUROC**	**C-index**
BRCA	STL	**0.35790 ± 0.10046**	0.73241 ± 0.06667	0.79714 ± 0.05355
BRCA	MTL	0.27229 ± 0.06966	**0.81142 ± 0.04384**	**0.87828 ± 0.02802**
LUAD	STL	**0.48666 ± 0.06058**	**0.67481 ± 0.03825**	**0.72807 ± 0.03980**
LUAD	MTL	0.47578 ± 0.08226	0.58984 ± 0.05821	0.64920 ± 0.05745
COAD	STL	**0.36142 ± 0.10198**	**0.60417 ± 0.07446**	**0.59528 ± 0.07106**
COAD	MTL	0.29871 ± 0.07092	0.52288 ± 0.08976	0.52162 ± 0.08481

Supplementary Table 4: Performance metrics for STL and MTL for a BNN model with 1 layer in both feature extractors and 1 in the Classifier.

The following figure highlights the impact of model complexity on predictive performance, showing how the number of layers affects the evaluation metrics for each cancer type. In most cases, the original configuration provided the best results. Hence, we chose it as our global configuration.



Supplementary Figure 20: **Performance Comparison of STL and MTL Models with Varying Layers Across Cancer Datasets.** Comparison of STL and MTL models with different numbers of layers across BRCA, LUAD, and COAD datasets. The plots display the AUPRC, AUROC, and C-index performance metrics across different configurations: Original STL, Original MTL, 1 Layer STL, and 1 Layer MTL. Each bar represents the mean score, with error bars indicating the standard deviation. Subfigures (a), (b), and (c) show the AUPRC, AUROC, and C-index for BRCA, respectively. Subfigures (d), (e), and (f) show the same metrics for LUAD. Subfigures (g), (h), and (i) display the metrics for COAD.

Supplementary Section 4.3. Excluding one cancer from MTL

We extended the initial ablation studies to evaluate how training on fewer cancer types affected model performance. We trained an MTL BNN using only two cancer training samples for this. The results are shown in Tables 5, 6, and 7, where the "Datasets" category indicates which data was used to train the model. "All" means that BRCA, LUAD, and COAD were used as training and test sets.

This ablation study examines the performance of the MTL model when trained with fewer datasets, as well as the impact of various dataset combinations on performance. When BRCA and LUAD datasets are combined, there is a positive effect on the LUAD predictions, while the improvement for BRCA is restricted to the AUPRC. A similar pattern emerges when combining the LUAD and COAD, where the primary benefits are

seen in COAD, and LUAD sees enhancements exclusively in its AUPRC. Surprisingly, combining BRCA and COAD consistently underperforms compared to integrating all three cancer types.

The findings indicate that the advantages of the MTL model over the STL model likely stem from the number of shared genes, their predictive value, and other common patterns across cancers. Furthermore, these results imply that the performance boost observed in MTL is attributed to the total number of training samples and the similarities in genomic expression and clinical data across different cancers. This suggests a more significant impact of choosing cancers with shared genes and similar data distributions than initially anticipated.

Datasets		AUROC	AUPRC	C-index
ALL	BRCA	**0.839 ± 0.044**	0.348 ± 0.090	**0.823 ± 0.043**
	LUAD	0.645 ± 0.060	0.509 ± 0.082	0.587 ± 0.049
BRCA & LUAD	BRCA	0.814 ± 0.056	**0.387 ± 0.103**	0.795 ± 0.055
	LUAD	**0.650 ± 0.063**	**0.560 ± 0.081**	**0.600 ± 0.055**

Supplementary Table 5: Results of a test set bootstrap evaluation of an MTL BNN trained only on BRCA and LUAD data.

Datasets		AUROC	AUPRC	C-index
ALL	BRCA	**0.839 ± 0.044**	**0.348 ± 0.090**	**0.823 ± 0.043**
	COAD	**0.712 ± 0.073**	**0.498 ± 0.102**	**0.696 ± 0.067**
BRCA & COAD	BRCA	0.789 ± 0.056	0.301 ± 0.087	0.772 ± 0.055
	COAD	0.533 ± 0.080	0.325 ± 0.090	0.534 ± 0.073

Supplementary Table 6: Results of a test set bootstrap evaluation of an MTL BNN trained only on BRCA and COAD data.

Datasets		AUROC	AUPRC	C-index
ALL	LUAD	**0.645 ± 0.060**	0.509 ± 0.082	**0.587 ± 0.049**
	COAD	0.712 ± 0.073	0.498 ± 0.102	0.696 ± 0.067
LUAD & COAD	LUAD	0.600 ± 0.064	**0.536 ± 0.080**	0.562 ± 0.053
	COAD	**0.771 ± 0.068**	**0.597 ± 0.107**	**0.748 ± 0.063**

Supplementary Table 7: Results of a test set bootstrap evaluation of an MTL BNN trained only on LUAD and COAD data.

Supplementary Section 5. Additional External Validation Details

Supplementary Section 5.1. Ovarian Cancer (OV)

We conducted additional analyses using TCGA's OV dataset. This dataset was preprocessed and partitioned in the same manner as the other TCGA datasets. Regarding the Learning Paradigms category, the models including "STL" correspond to an STL model trained exclusively using the specified cancer genes (BRCA, LUAD, or COAD) but trained and tested using only OV data. For the MTL models, "BRCA & OV" means the model was trained and tested using only BRCA and OV data. The model was trained and tested for "MTL (All cancers)" using BRCA, LUAD, COAD, and OV data. Finally, the "MTL External Validation" model was trained on BRCA, LUAD, and COAD but tested solely on OV. We tested OV using the genes selected for BRCA due to their similarity.

Surprisingly, for OV, the STL models outperformed their MTL counterparts. This is likely due to the OV dataset's more even label distribution. OV has an imbalance rate of 51.54%, compared to the lower rates in BRCA, LUAD, and COAD (9.241%, 33.333%, and 20.485%, respectively). This discrepancy indicates that our model, optimized for MTL with a higher imbalance rate, tends to overfit when applied to more equal imbalance rates, like those found in OV. Consequently, to maximize the benefits of MTL, future studies should consider grouping and training on cancers with analogous data distributions and imbalance ratios, enabling more compatible extrapolations and better predictive performance. Therefore, in addition to considering biological similarities such as shared genes or risk factors when grouping cancers for an optimal MTL model, cancers should also be grouped based on similar death rates and survival time distributions to enhance predictive accuracy.

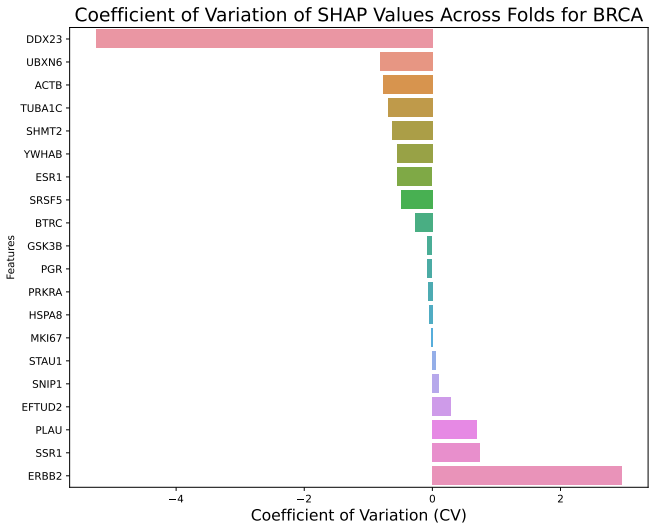
Learning Paradigms	AUROC	AUPRC	C-index
STL-BRCA	**0.80631** ± 0.04735	0.77459 ± 0.06945	**0.72725** ± **0.03744**
STL-LUAD	0.75630 ± 0.05169	0.73584 ± 0.07057	0.67586 ± 0.04043
STL-COAD	0.79025 ± 0.04889	**0.79838** ± 0.06380	0.69859 ± 0.03726
MTL (BRCA & OV)	0.78166 ± 0.06435	0.76487 ± 0.06953	0.71789 ± 0.05028
MTL (All cancers)	0.78565 ± 0.05312	0.76497 ± 0.07353	0.71461 ± 0.03356
MTL External validation	0.75993 ± 0.05418	0.76575 ± 0.06797	0.69867 ± 0.03943

Supplementary Table 8: Different STL and MTL model performance tested using only the TCGA Ovarian Cancer dataset. Regarding the Learning Paradigms, the models including "STL" correspond to an STL model trained using only the specified cancer's genes (BRCA, for instance), though using OV data. For the MTL models, "BRCA & OV" means the model was trained and tested using only BRCA and OV data.

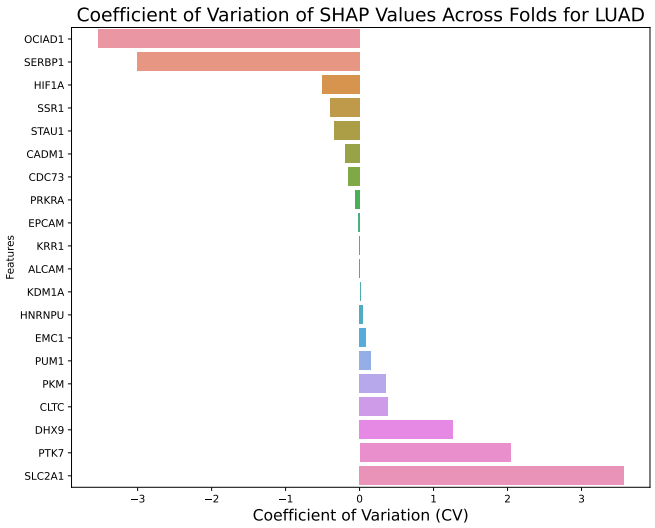
Supplementary Section 6. Additional SHAP value analyses

Supplementary Section 6.1. Coefficient of variation (CV)

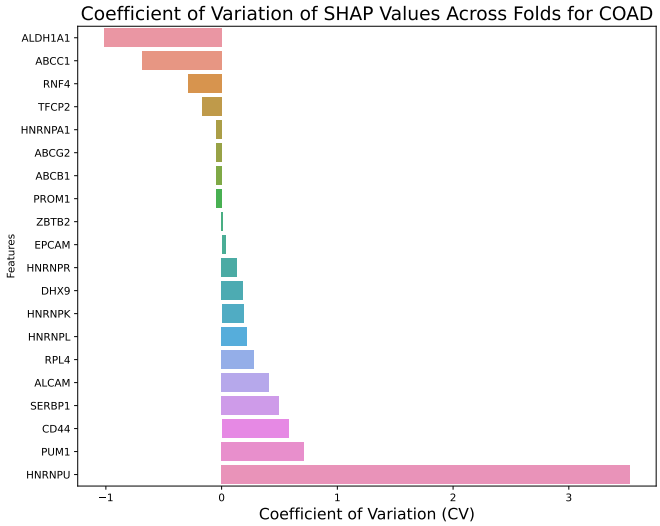
We calculated the coefficient of variation, which is equal to the ratio between the standard deviation and mean of the SHAP values across different folds. This coefficient symbolizes a normalized measure of variability relative to the size of the feature importance.



[a]



[b]

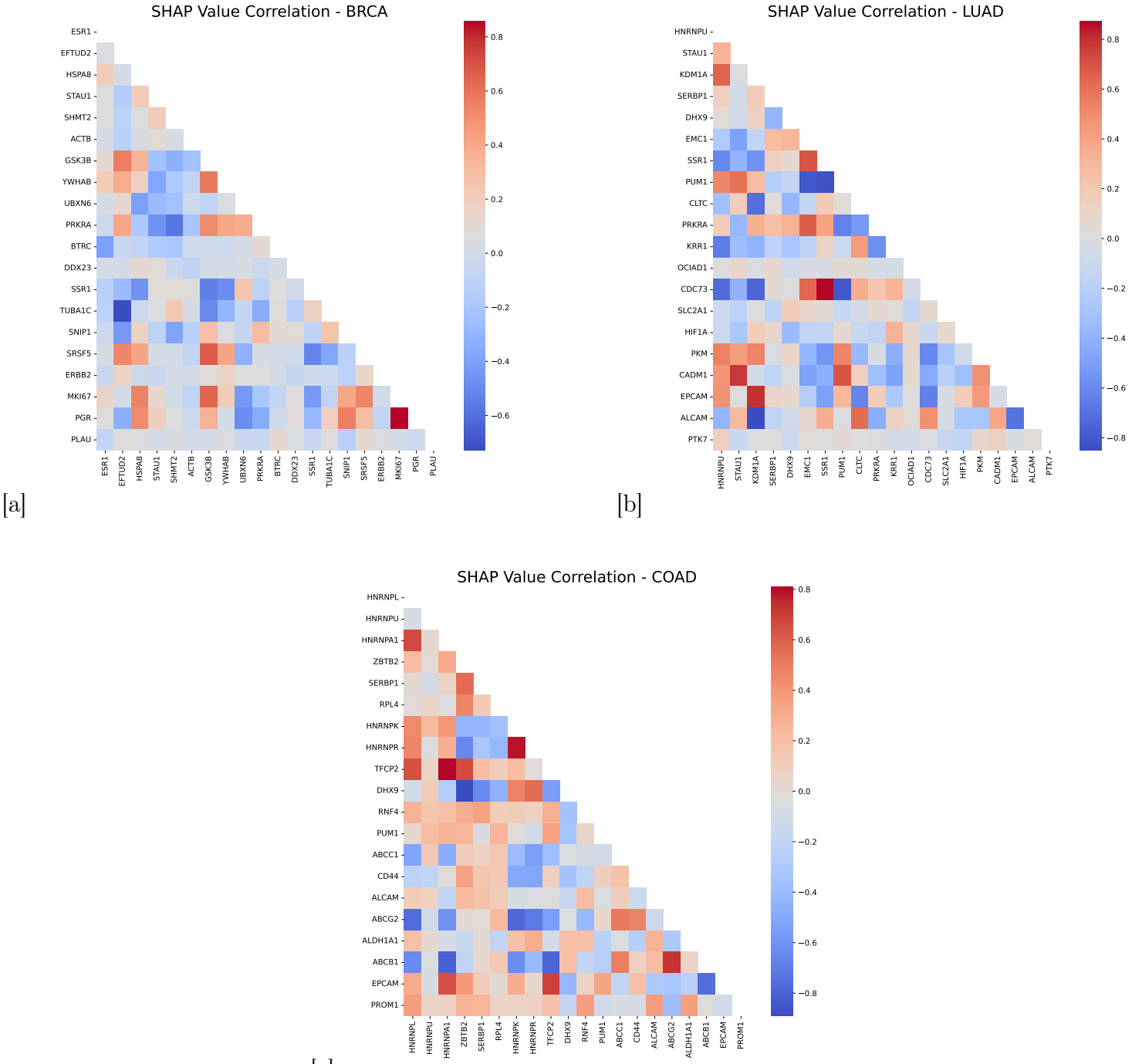


[c]

Supplementary Figure 21: **Coefficient of variation of SHAP values across folds for BRCA, LUAD, and COAD.** This chart illustrates the CV of SHAP values across four folds for BRCA (a), LUAD (b), and COAD (c). Each bar represents a different gene, with the length indicating the degree of variation in SHAP values, quantifying each feature's importance in the Genomic FE. A high CV suggests that the feature's influence on the model's predictions varies greatly across different folds, while a low CV shows a consistent influence.

Supplementary Section 6.2. Correlation between SHAP values for different genes

The matrix below showcases the interdependencies between gene expression levels and their contribution to the predictive model's output. For all datasets, most features don't show more than three strongly correlated features. The majority of features lie on the neutral to weakly correlated territory. As such, we cannot infer a robust linear relationship between features. This could be explained because it is common for this kind of high-dimensional data to have sparse correlations, primarily when different genes and pathways operate independently.



Supplementary Figure 22: **SHAP value correlation matrix.** The subfigures show the correlation matrices for BRCA (a), LUAD (b), and COAD (c). Each cell reflects the degree of correlation between the SHAP values of two genes, varying from -1 (blue) to 1 (red), where 1 indicates perfect positive correlation, 0 indicates no correlation, and -1 indicates perfect negative correlation.

Supplementary Section 7. Python reusable code

We developed a Python code base to interact with the REST API provided by the GDC Portal for data filtering and download. Due to the limited functionality of Python packages compared to those in other languages, like the R package TCGAbiolinks, we resorted to using this custom Python code base. We filtered outpatients lacking survival status, time, complete RNA-Seq, and clinical data. We downloaded the data of qualified patients using GDC API version 33.1, released on May 31, 2022. This code, along with our models and documentation, can be found in our Github repository.