Article

Graph Convolutional Networks for Predicting Cancer Outcomes and Stage: A Focus on cGAS-STING Pathway Activation

Mateo Sokač [1,*] , Borna Skračić [1], Danijel Kučak [1] and Leo Mršić [2]

1 Software Engineering Department, Algebra University, 10000 Zagreb, Croatia; borna.skracic@algebra.hr (B.S.); danijel.kucak@algebra.hr (D.K.)
2 Algebra University, 10000 Zagreb, Croatia; leo.mrsic@algebra.hr
* Correspondence: msokac@algebra.hr

Abstract: The study presented in this paper evaluated gene expression profiles from The Cancer Genome Atlas (TCGA). To reduce complexity, we focused on genes in the cGAS–STING pathway, crucial for cytosolic DNA detection and immune response. The study analyzes three clinical variables: disease-specific survival (DSS), overall survival (OS), and tumor stage. To effectively utilize the high-dimensional gene expression data, we needed to find a way to project these data meaningfully. Since gene pathways can be represented as graphs, a novel method of presenting genomics data using graph data structure was employed, rather than the conventional tabular format. To leverage the gene expression data represented as graphs, we utilized a graph convolutional network (GCN) machine learning model in conjunction with the genetic algorithm optimization technique. This allowed for obtaining an optimal graph representation topology and capturing important activations within the pathway for each use case, enabling a more insightful analysis of the cGAS–STING pathway and its activations across different cancer types and clinical variables. To tackle the problem of unexplainable AI, graph visualization alongside the integrated gradients method was employed to explain the GCN model's decision-making process, identifying key nodes (genes) in the cGAS–STING pathway. This approach revealed distinct molecular mechanisms, enhancing interpretability. This study demonstrates the potential of GCNs combined with explainable AI to analyze gene expression, providing insights into cancer progression. Further research with more data is needed to validate these findings.

Keywords: cGAS–STING; graph-convolutional-network; graphs; cancer; pan-cancer; machine learning; NGS

1. Introduction

The recent advancements in next-generation sequencing (NGS) have brought about a paradigm shift in research, particularly in the realm of gene expression analysis [1,2]. This technology has enabled researchers to delve deeply into the intricate mechanisms governing gene regulation and expression. With the decreasing cost of sequencing, it has become commonplace to employ NGS techniques to explore the various facets of biological processes within the same samples [3,4]. However, this comes at cost of the total storage required for a single whole-genome sequence can range from 100 GB to 1 TB, including backup and redundancy requirements [5]. Furthermore, traditional statistical methods often struggle to fully utilize the vast and complex data generated by NGS due to several inherent limitations. NGS data are characterized by high dimensionality, noise, and heterogeneity, presenting challenges for classical statistical approaches, which typically rely on predefined hypotheses and linear relationships. These methods often require simplifying assumptions, such as the normal distribution of data or independence of features, which do not hold in the intricate biological systems captured by NGS [6]. Additionally, the need for multiple testing corrections in traditional statistical analyses

can lead to a loss of statistical power, resulting in the dismissal of potentially significant findings [7,8]. As a result, traditional statistics may fail to capture the full spectrum of biological variation and subtle interactions present in NGS data [9]. This includes delving into DNA sequencing to unravel genomic architecture and detect single-nucleotide variants, analyzing RNA to decipher gene expression patterns, and investigating methylation to unveil gene regulation dynamics and chromatin structure [10–12]. Multi-omics datasets are rich in information, allowing for integrated analyses that consider multiple layers of biological data simultaneously. However, traditional storage and commonly used methods often rely on data in tabular formats, fail to capture the inherent spatial organization of the genome, and lose crucial connectivity information [5]. As a consequence, the full potential of these rich data remains untapped [13–15].

Alternative data structures are essential to address this challenge effectively. Ideally, data structures that preserve the spatial organization of the genome should be employed [16,17]. This could involve utilizing graph-based representations that maintain the genomic architecture and allow efficient traversal and analysis [16,18–20]. Additionally, considering the vast feature space inherent in genomic data, specialized data structures capable of handling high-dimensional data are warranted. These may include sparse matrices or compressed data representations that optimize storage and computation [21]. Currently, there is a great effort in developing machine learning models that will capture the relationships of gene interactions used to predict the variables of interest regarding cancer development and treatment [16]. However, more complex models tend to have less initial interpretability which is a crucial aspect of model development since effective interpretations lead to discoveries of novel biomarkers and a better understanding of biological processes.

The convolutional method proved to be a strong and influential improvement over existing methods that were based on tabular data. Based on the previously mentioned intuition of representing genes as images to use convolutional neural networks (CNNs) to capture their spatial dependency, a new method of presenting genetic data is presented [16]. The novel approach to presenting data would be to project them as graphs, which should in turn be able to capture complex and non-linear relationships in biological gene interaction networks. Furthermore, using message passing as a basis of the inner workings of neural network models based on graph neural networks (GNNs) should be able to mimic pathway signaling found in gene pathways. This approach should enhance prediction precision, improving the inference and feature discovery of cancer biomarkers [22]. In essence, the analysis of multi-omics data demands a paradigm shift towards more sophisticated data structures and analytical frameworks that acknowledge the spatial complexity of the genome and mitigate the pitfalls associated with multiple hypothesis testing [23]. Recent studies also managed to utilize graphs as mathematical data structures to model biological pathways [22,24,25]. Thus, machine learning models that exploit such premises can be used. Graph convolutional network (GCN) architecture was chosen for its strong expressive power when equivalent graph topology is used for each data point representation.

While these models offer substantial advancements in predictive accuracy and the ability to model complex biological processes, their interpretability remains a significant challenge. The intricate nature of advanced machine learning techniques often results in a "black-box" scenario where the internal workings and decision-making processes are hidden from the users. This lack of transparency can be problematic in the medical field, where understanding the rationale behind predictions is crucial for clinical trust and application. Therefore, developing methods or relying on methods that elucidate how these models process and prioritize different features are needed. Tools like Integrated Gradients by Captum [26], DeepLIFT [27,28], SHAP (SHapley Additive exPlanations) [27,28], LIME (Local Interpretable Model-agnostic Explanations) [29], and Grad-CAM (Gradient-weighted Class Activation Mapping) [27] provide a non-parametric assessment of input importance. Such approaches can help demystify the decision pathways within the model, making it possible to identify and validate key biomarkers and gene interactions, ultimately bridging

the gap between high predictive power and actionable biological insights, leading towards precision medicine application [30,31]. When combined with topological inference and the visualization of a gene interaction network, this approach can highlight the most influential nodes and connections within a specific use case's model.

To tackle the complexities of genomic data, graphs provide an intuitive and novel approach for representing and analyzing gene interactions [32]. By projecting genomic data onto graph structures, we can capture the intricate dependencies and signaling pathways inherent in biological systems (Figure 1A). The cGAS–STING pathway serves as an exemplary case due to its pivotal and dual role in cancer biology [33]. As a crucial component of the innate immune system, it functions to detect the presence of cytosolic DNA and trigger the expression of immune genes [18]. This pathway also plays a critical role in mediating immune defense against viruses [34,35]. In the context of cancer research, the cGAS–STING pathway has been identified as a promising target for cancer research [36]. It participates in regulating cancer, autoimmune and inflammatory diseases, microbial and parasitic infectious diseases, and other diseases [37]. However, the cGAS–STING pathway is often referred to as a double-edged sword in cancer research [33,38]. On one hand, it plays a tumor-suppressive function. On the other hand, cGAS is often associated with high genomic instability, a well-known hallmark of cancerogenesis [39]. Therefore, understanding the dual role of the cGAS–STING pathway is crucial for developing effective cancer therapies [39–41].
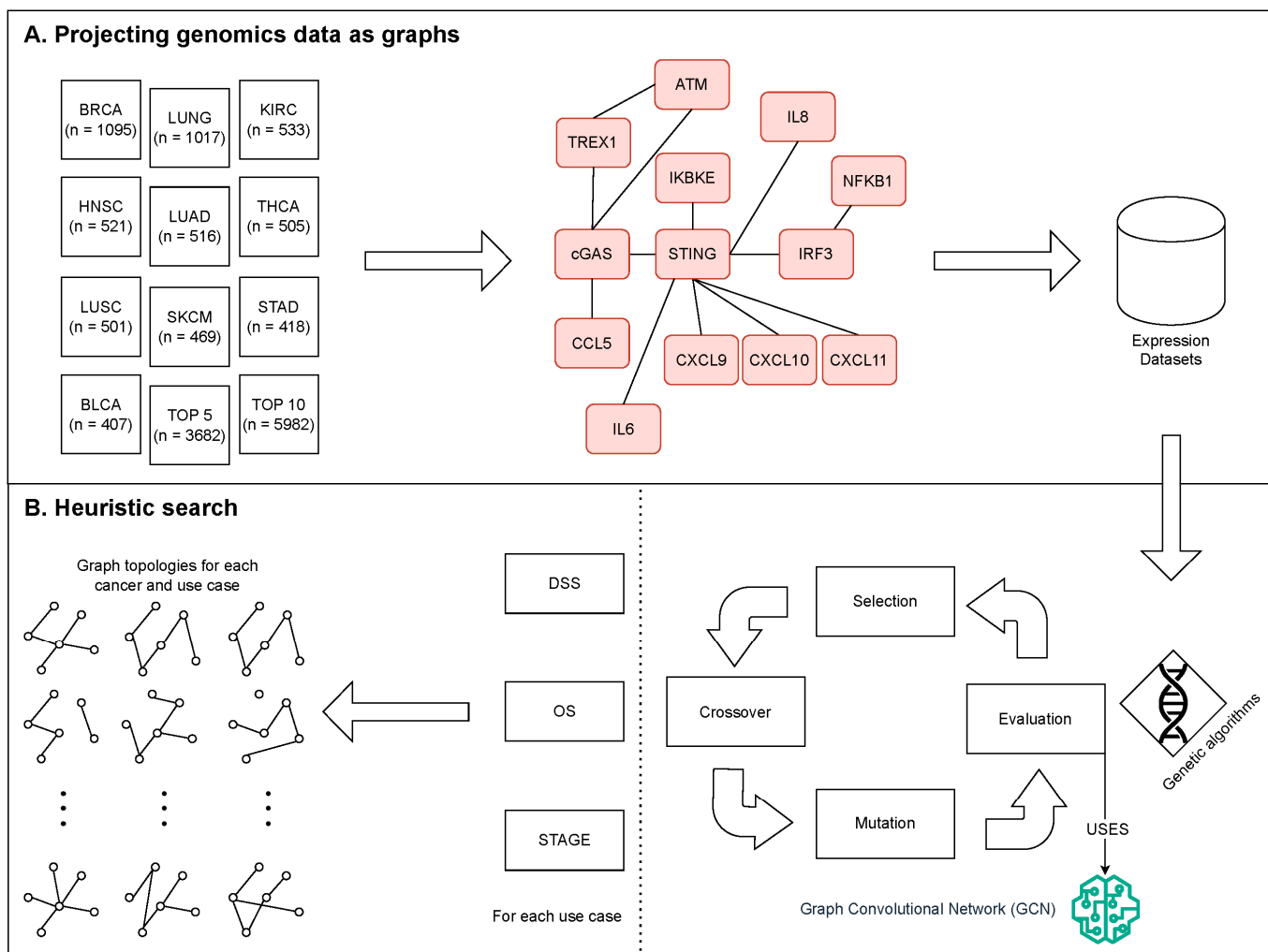

Figure 1. *Cont.*
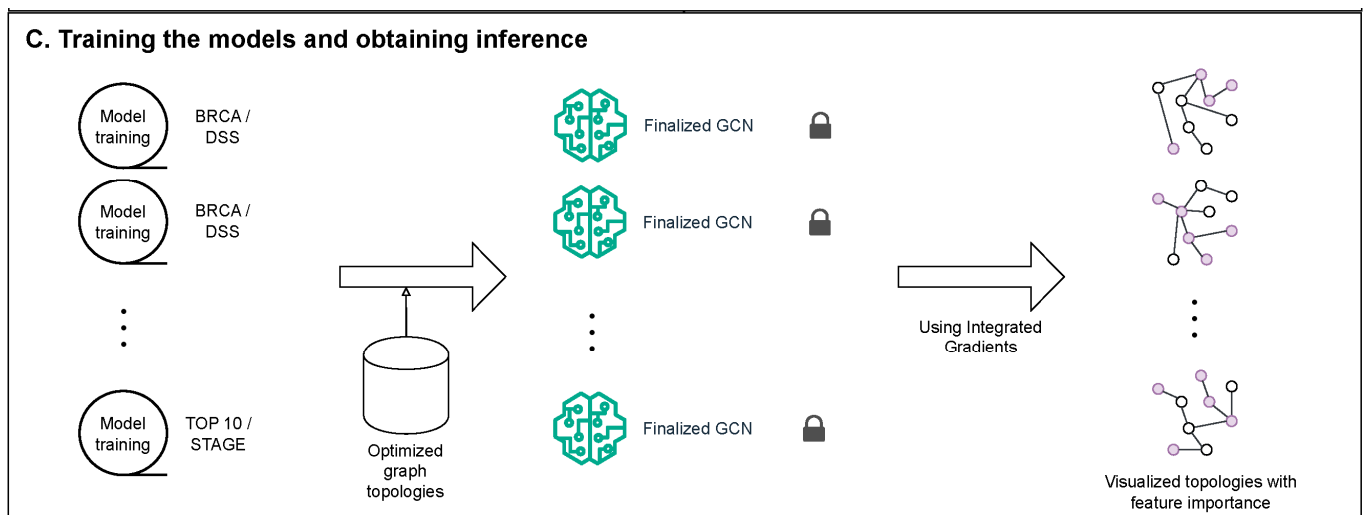
Figure 1. Study overview. (**A**) Projecting gene expression data across cancer cohorts as graphs. (**B**) Heuristic search using the genetic algorithm optimization technique for each use case. (**C**) Training the models and visualizing inference by highlighting the most important gene activations in the pathway.

In this paper, we present a novel method of projecting gene pathway expression data using graphs as well as a complete framework for finding an optimal graph representation of gene interactions in the cGAS–STING gene pathway, named GENIE (Graph gEnetic Network Inference and Evaluation). After projecting the data as graphs, a heuristic approach was used to find the optimal hyperparameters for machine learning models, most notably, graph topology for specific use cases (Figure 1B). After the identification of optimal graph topology for each use case in selected cancers, the GCN model was trained with the appropriate genomic data with gene expressions being the feature of interest. By utilizing the integrated gradients method, we were able to discover important nodes in the decision-making model for each use case. By utilizing the intuitive expressive power of visualized graph topology alongside integrated gradient inference, the interpretability of our method was compared and enhanced (Figure 1C). This enhancement provided clearer insights into important gene interactions in the cGAS–STING pathway as key biomarkers.

2. Methods

2.1. Samples and Training Data

The gene expression of cancer types was from The Cancer Genome Atlas (TCGA) (https://www.cancer.gov/tcga, accessed on 3 March 2024) [39]. Based on the sample count, the top 10 cancer datasets were chosen. Based on a preliminary analysis of the gene expression dataset, we observed that certain cancer datasets lacked adequate information regarding pathological or clinical stages. Consequently, these samples were excluded from further analysis. To ensure the robustness of our study, we replaced these incomplete samples with those from other leading cancer types that provided comprehensive pathological and clinical data. The mentioned approach resulted in the top 10 cancers being chosen as described in the table. Besides those datasets, two datasets were created for the purpose of this study by merging the top 5 and top 10 cancer datasets, respectively. All TCGA sample data used in the study were obtained from Xena Browser [42], which were processed by the UCSC (University of California, Santa Cruz) Toil pipeline [43], and RNAseq summarized transcript per million (TPM) on the gene level (Table 1, Supplementary Figure S1).

Table 1. Cancer types included in the study and their corresponding number of samples.

Cancer	Name	Count
TOP 10	Top 10 cancer datasets merged	5982
TOP 5	Top 5 cancer datasets merged	3682
BRCA	Breast cancer	1095
LUNG	Lung cancer	1017
KIRC	Kidney renal clear cell carcinoma	533
HNSC	Neck squamous cell carcinoma	521
LUAD	Lung adenocarcinoma cancer	516
THCA	Thyroid cancer	505
LUSC	Lung squamous cell carcinoma	501
SKCM	Skin cutaneous melanoma	469
STAD	Stomach adenocarcinoma	418
BLCA	Bladder urothelial carcinoma	407

2.2. cGAS—STING Pathway Definition

Since datasets contain expressions of all sequence genes, to reduce computational complexity, we defined the list of genes in the cGAS–STING pathway based on the available literature. These genes encode proteins that play critical roles in the detection of cytosolic DNA, signal transduction, and the resulting immune responses. C6orf150 encodes cGAS, a DNA sensor that produces cGAMP upon binding to cytosolic DNA, thereby initiating the cGAS–STING pathway [44]. The CGAS gene detects not only cytosolic DNA from viruses but also double-stranded DNA found in micronuclei, small extra-nuclear structures that contain whole chromosomes, or chromosome fragments that did not integrate into the main nucleus after mitosis [45–47]. When chromosomal instability causes the micronucleus membrane to rupture, the double-stranded DNA becomes exposed, enabling cGAS to bind to it and trigger pathway activation [38]. IRF3 is a transcription factor activated by the cGAS–STING pathway that induces the expression of type I interferons and other antiviral genes. TMEM173 encodes STING, a central adaptor protein that, upon activation by cGAMP, activates IRF3 and NF-κB, leading to the production of type I interferons and proinflammatory cytokines (IL6 and IL8) [48,49]. NFKB1 and IKBKE are integral to NF-κB signaling, which drives the expression of inflammatory cytokines such as IL6 and IL8 (CXCL8), which are crucial for immune response modulation [44,49]. CCL5, CXCL9, CXCL10, and CXCL11 are chemokines that are regulated by NF-κB and play roles in recruiting immune cells to infection sites, thus enhancing the immune response [47,50]. TREX1 and ATM are involved in the DNA damage response and repair mechanisms, with TREX1 degrading cytosolic DNA to prevent unwarranted activation of the immune response, and ATM signaling DNA damage to promote repair and modulate immune signaling [51]. In the cancer microenvironment under the burden of high chromosomal instability, cGAS is associated with ATM, however, this connection is not clearly understood [52,53]. In summary, based on the literature research, C6orf150 (cGAS), CCL5, CXCL10, TMEM173 (STING), CXCL9, CXCL11, NFKB1, IKBKE, IRF3, TREX1, ATM, IL6, and IL8 (CXCL8) were chosen as representative genes of the cGAS–STING pathway to be used in the study (Figure 1A).

2.3. Use Cases Included in the Study

To describe the difference between canonical and non-canonical activation of the cGAS–STING pathway in different patients, three variables of interest were chosen: disease-specific survival (DSS), overall survival (OS), and clinical tumor stage. Overall survival describes whether the patient survived or not, not necessarily related to their tumor disease. DSS, on the other hand, describes if the patient died from the specific tumor or not. Stage describes the current stage of the cancer for the specified patient: stage I, stage II, stage III, and stage IV, as well as various substages and variations that were grouped under one of the four main stages. Since the distribution of stage classes was not balanced, stage I and stage II were grouped under the 'early' class while stage III and stage IV were grouped under the 'late' class. In our analysis, the OS use case exhibited the poorest distribution in THCA and BRCA datasets. Similarly, for the DSS use case, the distribution was most unfavorable in THCA and BRCA. Furthermore, the stage-specific analysis revealed that the LUNG, LUSC, and HNSC cohorts demonstrated the worst distribution (Supplementary Figure S2).

The following summarizes the use cases evaluated in this study:

- DSS, a binary variable representing disease-specific survival of a patient;
- OS, a binary variable representing the overall survival of a patient;
- Stage is split as a binary variable where "early" is defined as stages 0, I, and II, while "late" is defined as stages III and IV.

2.4. Graph Convolutional Neural Network Architecture

The GCN multi-layer model used in this study is designed to leverage the power of graph neural networks to analyze gene expression data. The implementation of this model was written in Python (version 3.10.13); more specifically, the model was generally defined as an inherited Module class, which is a part of the PyTorch (version 2.1.2) library. The model's input consisted of gene expression features, which were processed through a graph structure defined by an adjacency matrix. This matrix captured the connections between genes, allowing the model to aggregate information based on gene interactions. The first layer of the model was a graph convolutional network layer, implemented using the GCNConv class as a part of the PyTorch Geometric library (version 2.4.0). This layer performs convolution operations on the graph, transforming node features into a higher dimensional space. The hidden size of 64 denoted hidden channels, which determined the output dimensionality of this layer. The graph convolution operation utilized the edge indices derived from the adjacency matrix, enabling the model to gather and process information from neighboring nodes. The output from the graph convolution layer was passed through a ReLU (Rectified Linear Unit) activation function. Following this, the transformed features were fed into a linear layer, which mapped the features to the two possible classes, allowing for binary classification. This final layer served as the classifier, producing predictions for each data point (Supplementary Figure S3). The model was trained on the train set and validated on the test set, split in the 75:25 ratio, respectively, using the binary cross-entropy loss function, which is well-suited for binary classification tasks. Adam optimizer was employed to adjust the model's weights with a learning rate of 0.0001 and weight decay of 0.1. The model was trained for 10 epochs for objective function evaluation in the genetic algorithms heuristic search process and 100 epochs for final model training. Before training the classification model for each use case, hyperparameters (layers, node count, and batch size) of such models were optimized using Optuna (version 3.4.0).

2.5. Optimization of Graph Structure Topology

A specific hyperparameter applicable to the GCN scenario was the graph adjacency matrix representing graph topology, or in other words, the order of gene interactions based on their protein expression values. The optimization process in this work essentially involved hyperparameter tuning of a classification model that must be applied for each model based on the specific cancer cohort and target variable combination, resulting in 36 different models and optimization processes. To optimize this specific parameter, a heuristic method known as genetic algorithms (GAs) was chosen. Due to the versatile nature of the objective function rules for this algorithm and the discrete nature of the problem's search space, genetic algorithms were used. Using genetic algorithms, we designed a heuristic optimization for each of the tasks/scenarios. In summary, possible adjacency matrices were initially encoded and generated as a population that evolves over multiple generations. The objective function is defined as the F1 metric of the GCN model trained using the solution as an adjacency matrix hyperparameter (Figure 2B, Supplementary Figure S4). The GA search was performed using the PyGad library (version 3.2.0). The rationale for choosing GAs over other heuristic search algorithms lies in their ability to effectively explore complex and high-dimensional search spaces, which are often characterized by non-linear relationships and numerous local optima. Furthermore, the choice of GA was driven by the flexibility of the algorithm but also the flexibility of the PyGad framework. The GA can handle various types of data representations, including binary, real-valued, and permutations, which makes them applicable to a wide range of problems, which was important for this project. The inherent flexibility of GAs allows researchers to easily integrate additional modalities or types of data into the framework, making it a versatile tool for future applications.

Upon completing the GA optimization for each use case, the resulting optimal graph topology, represented as an adjacency matrix, was used as a hyperparameter to train the final classification model for the corresponding cancer and variable combination. This model will be used in obtaining inference based on the graph visualization and integrated gradients method.
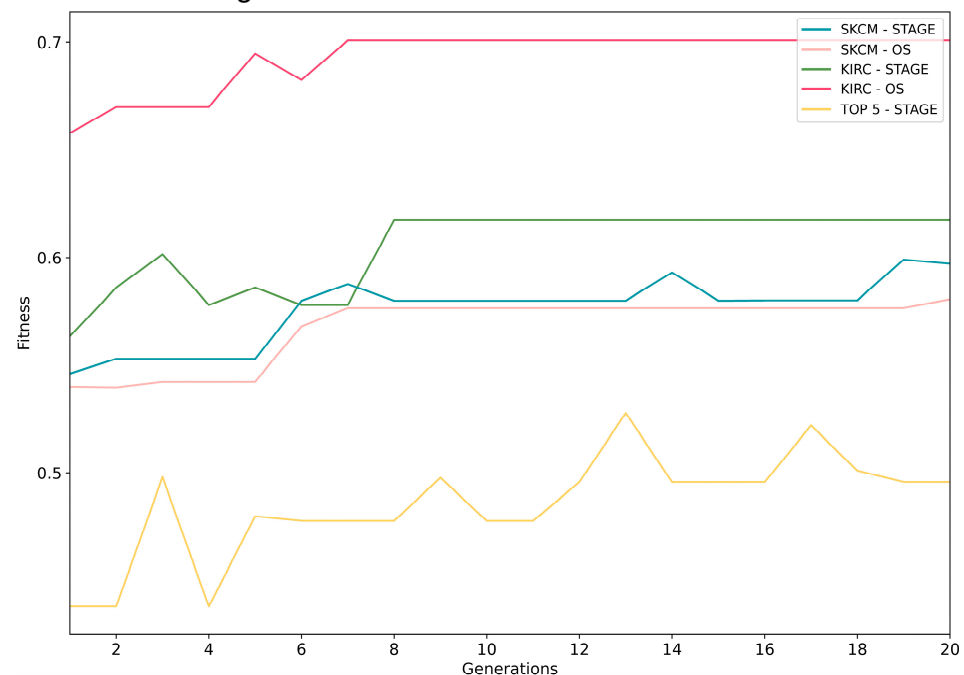
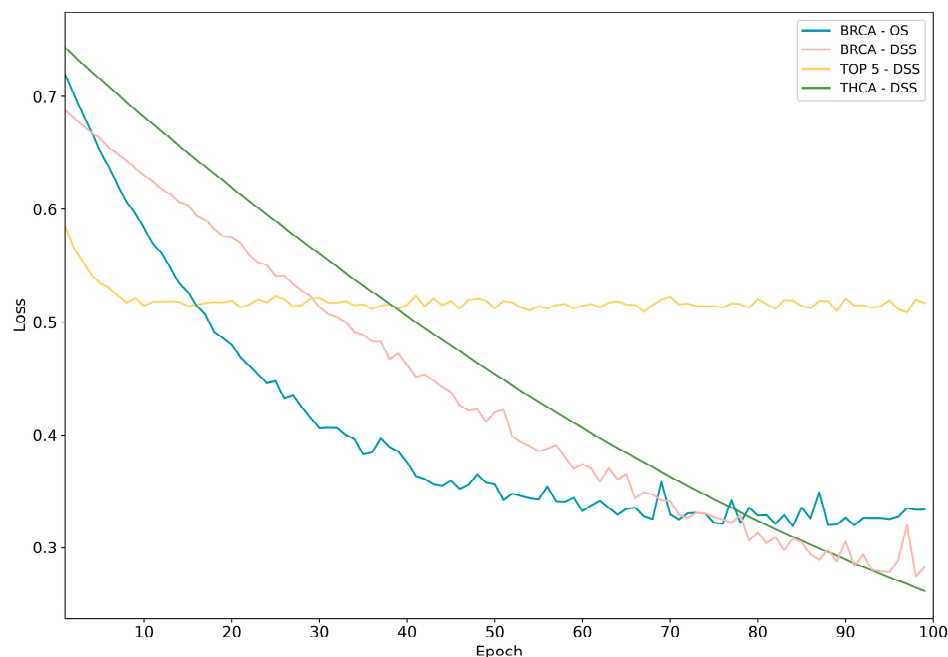## A. Fitness over generations

Figure 2. *Cont.*

Figure 2. Models' heuristic search and training. The figure illustrates the optimization process and training performance of the models. (**A**) The panel displays the fitness values over generations during the heuristic search, indicating an improvement in model performance as the search progresses. (**B**) The panel presents the categorical cross-entropy loss over training epochs, demonstrating how the loss decreases as the model trains, signifying better model fitting to the data.

2.6. Using Integrated Gradients to Evaluate cGAS-STING Pathway Activation

The integrated gradients method leverages the gradients of a trained model with fixed weights to ascertain the significance of each node in the graph convolutional network (GCN) for the classification process. By computing the gradients while keeping the model's weights constant, this approach quantifies the contribution of each node to the final prediction, thereby identifying the most influential nodes in the GCN network. To compute integrated gradients for networks, captum (version 0.7.0) was used.

2.7. Computational Requirements

For the simpler use cases in this study, computations were conducted using a setup comprising an Intel i5 8300H (4 cores) processor, Nvidia GeForce GTX 1050Ti with 4 GB of VRAM, and 8 GB of RAM. To handle more computationally complex use cases, Cuda (version 12.1) parallelization was executed on the HPC with the following specifications: Intel Xeon Silver 4216 CPU (64 cores) CPU alongside NVIDIA Quadro RTX 6000 (24 GB VRAM) GPU and 200GB RAM.

2.8. Code Availability

The source code for implementation used in this study is openly available as a GitHub repository at https://github.com/bskracic/genie-nextflow (accessed on 20 August 2024). The repository contains jobs implemented as a Nextflow (version 23.10.1.5891) pipeline which allows for reproducible runs across different computational environments.

3. Results

3.1. Finding Optimal Graph Structure

The genetic algorithm heuristic search in this study is configured with specific parameters to optimize the solutions iteratively. A population of 10 solutions is initially generated, which evolves over 20 generations. During each generation, four parent solutions are

selected for mating based on their fitness, utilizing crossover and mutation operations to produce offspring solutions (Figure 2A). Each solution's fitness is assessed using an objective function, and the resulting value guides the selection process. The optimization process yielded varying degrees of success across the 12 cancer cohorts and three clinical variables evaluated. Notably, the LUAD cohort with the OS variable exhibited a modest improvement in the final generations; however, the consistently low F1 scores suggest that the cGAS–STING pathway may not adequately capture the underlying biological processes associated with cancer survival in this context. In the Top 5 cancer cohort, the stage variable showed steady, albeit limited, progress throughout the generations, reflected by a relatively low overall F1 score. In contrast, the other variables in the Top 5 as well as all variables in the Top 10 cohort and the LUAD stage use case demonstrated no significant improvement, indicating potential limitations in the model's ability to optimize for these scenarios. Conversely, the BLCA and BRCA cohorts, when evaluated for the stage variable, exhibited minimal progress, suggesting that the GA optimization may have reached a plateau early in the generational cycle. Interestingly, the KIRC and SKCM cohorts for the stage variable performed significantly better, with relatively high overall F1 scores and consistent progress across generations, where in KIRC, the graph topology was found quite early, resulting in minor improvements over generations, while SKCM showed improvements as epochs and generations passed by (Figure 2A,B, Supplementary Figure S5).

3.2. Training Classification Model

To evaluate graph topology fitness for each use case, the final value of the F1 test metric was used. We trained a GCN model with the aforementioned architecture for each use case in each cancer entry. Each model was trained for 100 epochs, with the appropriate dataset split into 75:25 for training and test sets, resulting in the performance measured with the F1 measure (Supplementary Figure S5, Table 2). The following table shows the F1 metric of each classification model after completing the optimization process. For the DSS use case, BRCA showed the highest F1 score at 0.9445, indicating the best performance, while STAD had the lowest F1 score at 0.5234, marking the worst performance. In the OS use case, THCA had the highest F1 score of 0.9921, demonstrating superior performance, whereas STAD had the lowest at 0.4619. Regarding the stage use case, HNSC achieved the highest F1 score of 0.8154, while STAD again had the lowest performance with an F1 score of 0.2755. On average, the DSS use case performed the best with a mean F1 score of 0.7264 (Figure 3A). When all use cases were averaged, the best-performing cancer was BRCA (breast cancer) with an average F1 of 0.8663 over all three use cases (Figure 3B).

Table 2. Summary of classification results.

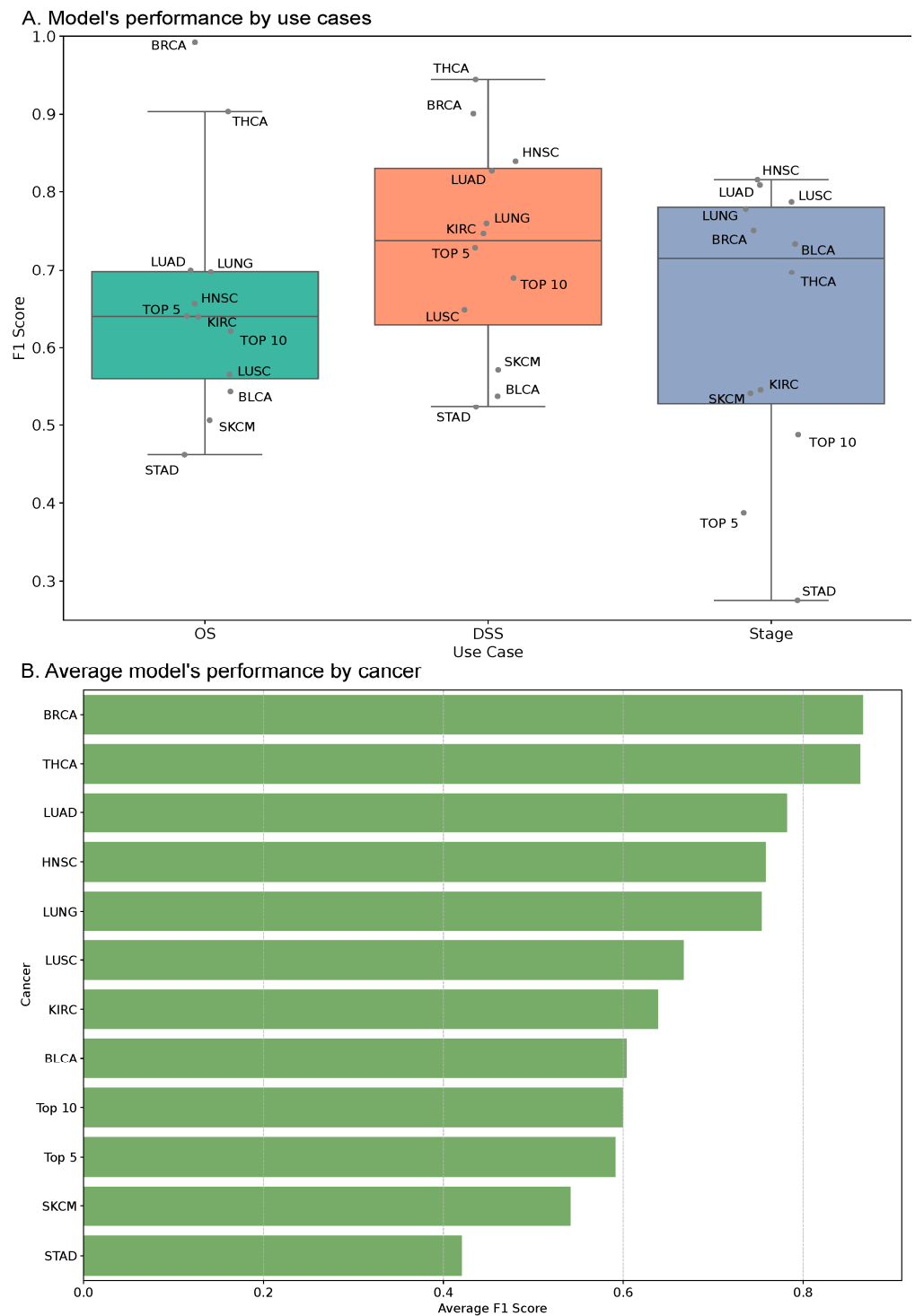| Cancers	OS (F1)	DSS (F1)	Stage (F1)
Top 10	0.6213	0.6892	0.4876
Top 5	0.6395	0.747	0.3875
SKCM	0.544	0.5378	0.5417
HNSC	0.7	0.7598	0.8154
STAD	0.4619	0.5234	0.2755
LUSC	0.5658	0.6485	0.7872
THCA	0.9921	0.9008	0.697
LUAD	0.698	0.8388	0.8087
KIRC	0.6406	0.7289	0.5461
LUNG	0.6564	0.8268	0.7781
BRCA	0.9035	0.9445	0.7508
BLCA	0.506	0.5716	0.7337

Figure 3. Models' performance. This figure compares the performance of the models across different scenarios. (**A**) Figure shows the F1 scores for each use case (OS, DSS, and stage) across all cancer types included in the study. (**B**) Figure illustrates the average F1 scores for each cancer type, providing a summary of model performance by cancer cohort.

3.3. Integrated Gradients Show Important Genes in the cGAS–STING Pathway

To comprehensively visualize the acquired graphs, the consensus method was used. For each use case, adjacency matrices were summed, and the final topology was determined using a 95th percentile filter. This approach ensured that only node connections with a sum greater than the 95th percentile were included in the final adjacency matrix.

Figure 4 illustrates the distinct patterns of cGAS–STING pathway activations, categorized by key clinical variables: OS, DSS, and cancer stage (early and late). Each graph topology depicting gene interactions results from a heuristic optimization process utilizing genetic algorithms and GCN models, as previously described. For each cancer type and specific use case, a combination of adjacency matrices was employed, integrating a method to filter connections that represent the top 5% in terms of frequency across all cancers within each variable category. This approach yielded consensus graphs that emphasized the prominent network structures associated with each clinical class, providing insights into the differential molecular mechanisms in the cGAS–STING pathway. Therefore, each node within the graphs is colored to reflect its impact on the model's decision-making process, with purple nodes indicating the least impactful and pink nodes representing the most impactful.

A. OS = false vs. OS = true
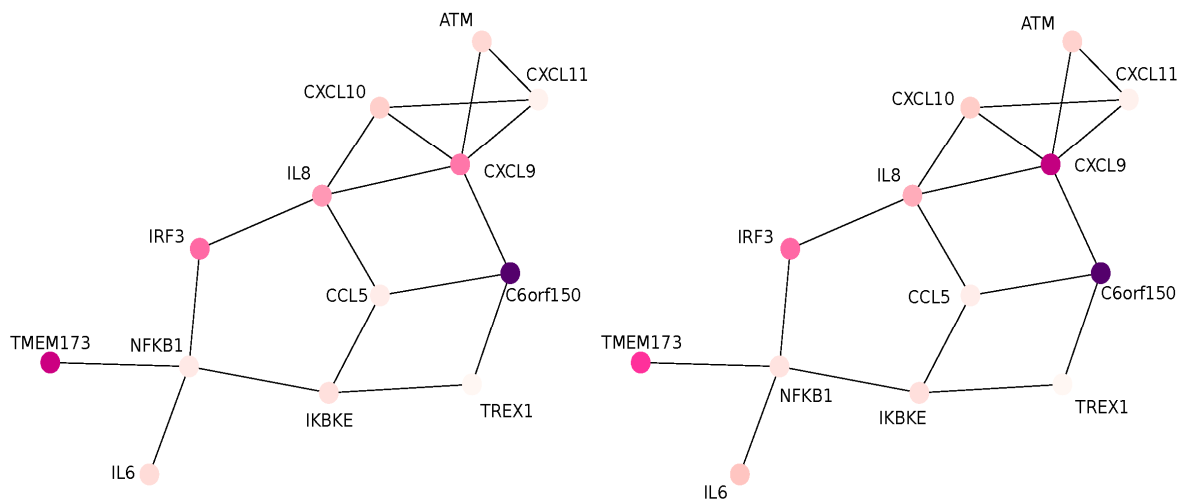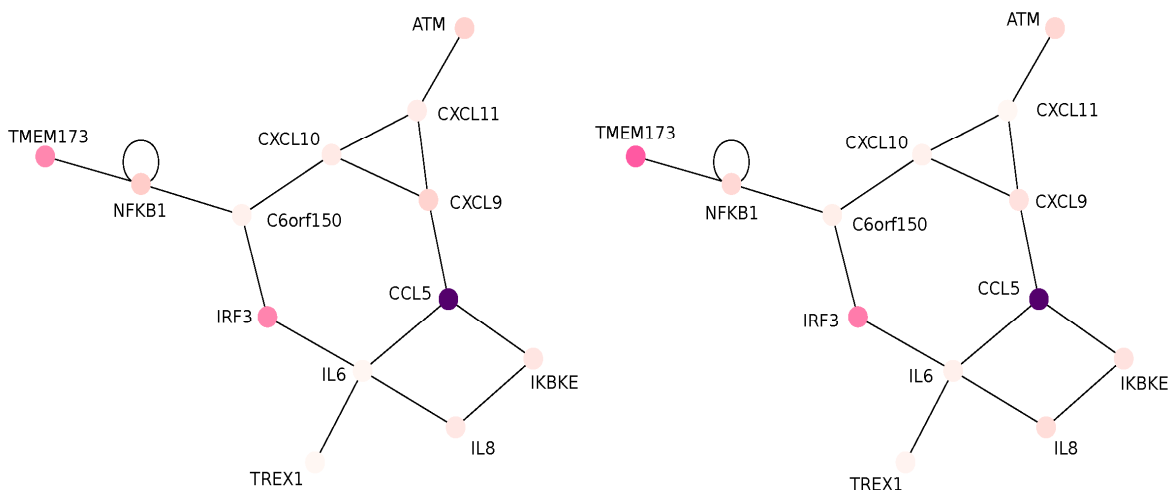

B. DSS = false vs. DSS = true

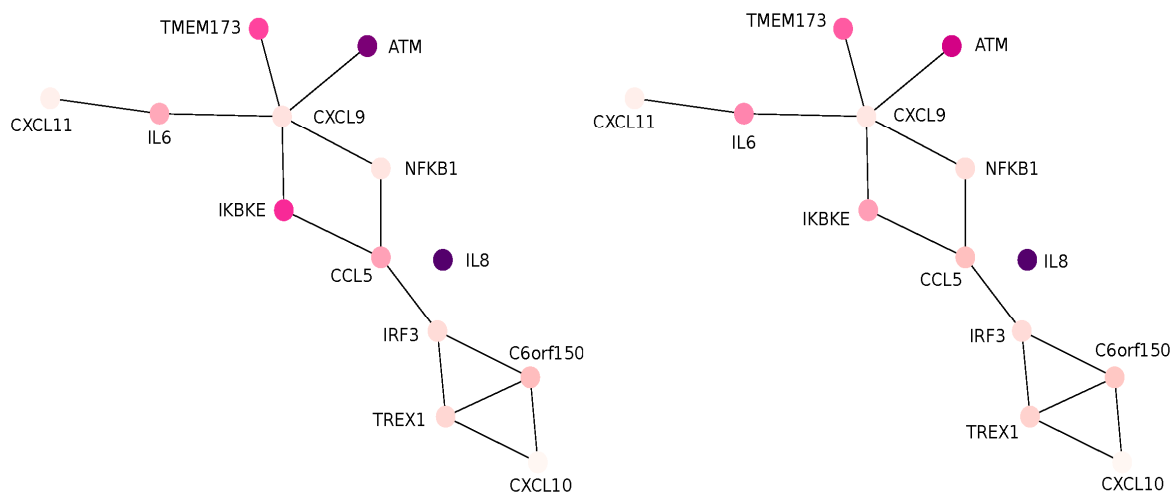
Figure 4. *Cont.*

C. Stage = early vs. Stage = late

Figure 4. Consensus feature-importance graphs. This figure visualizes the consensus feature importance across different clinical variables. (**A**) Figure compares feature importance between patients with OS = false and OS = true, highlighting key genes involved in overall survival. (**B**) Panel contrasts feature importance between patients with DSS = false and DSS = true, showcasing significant genes in disease-specific survival. (**C**) Figure examines the differences in gene activation between the early and late stages of cancer, identifying critical genes involved in tumor progression.

3.4. Integrated Gradients Distinguish between Non-Canonical and Canonical Sting Activation

Based on the evaluation of graph topology fitness using the F1 test metric for each cancer use case, a theoretical and empirical distinction between canonical and non-canonical cGAS–STING pathway activations becomes apparent. The F1 metric, derived from training each GCN model on specific datasets, reflects the model's predictive accuracy in distinguishing between different clinical variables such as OS, DSS, and tumor stage. Figure 4 illustrates the diverse activations of the cGAS–STING pathway, delineated by these clinical variables. Comparing the differences between OS, DSS, and stage graph topologies, we observed expected similarities between the OS and DSS topologies. The connection between TMEM173 (STING), NFKB1, IRF3, and interleukin (IL6 or IL8) shows consistent signaling. CXCL9, CXCL10, CXCL11, and ATM are connected in both cases, suggesting a close association. In the OS scenario, the CGAS (c6orf150) is the most impactful gene (according to the gradients) while in the DSS scenario, CCL5 showed the highest enrichment of gradients (Figure 4A,B). When predicting the stage, TMEM173 (STING), ATM, IKBKE, and IL8 showed the highest impact (Figure 4C). Interestingly, IL8 was excluded from the graph topology by heuristic search but was marked as highly important based on gradients, suggesting the importance of the IL8 for prediction but not as a part of the cGAS–STING pathway (Figure 4C). Finally, the graph topology of the stage is drastically different from the graph topology for OS and DSS, suggesting a change in the cGAS–STING pathway activation as a tumor becomes more aggressive (stage III and stage IV).

4. Discussion

This study evaluated the performance and explainability of a graph convolutional network in predicting outcomes based on gene expression data across various cancer types. Specifically, we focused on the cGAS–STING pathway's role in tumor immunity and its impact on patient survival. We evaluated a total of 36 scenarios, encompassing different cancer types and survival metrics, to comprehensively understand the model's capabilities. The study revealed notable variability in model performance across different cancer types and use cases. THCA for the OS use case (F1 = 0.9921) and BRCA for the DSS use case (F1 = 0.9445) emerged as the best-performing models, whereas STAD for the stage use

case (F1 = 0.2755) and SKCM for the DSS use case (F1 = 0.5378) demonstrated the lowest performance. These variations can be attributed to the underlying biological complexity, heterogeneity of these cancers, and dependency on cGAS–STING pathway activation. Non-canonical cGAS–STING pathway activation is associated with a known hallmark of cancer, DNA damage, and chromosomal instability. The DNA damage is associated with the DNA damage sensor, the ATM gene [53]. Our results show that ATM is one of the most important genes when predicting tumor stage in the context of cGAS–STING pathway topology (Figure 4C), suggesting non-canonical cGAS–STING activation through ATM DNA damage sensing [53]. Moreover, the best result for predicting tumor stage was achieved in LUNG (LUAD and LUSC) and HNSC (Figure 3A,B). These results are further strengthened by past studies as LUAD and LUSC are both primarily induced by smoking, which is associated with DNA damage and chromosomal instability, while HNSC and BRCA tumors are associated with copy number abnormalities [54,55]. Furthermore, in the graph topology for predicting tumor stage, TMEM173 (STING) and ATM are closely located, while in OS and DSS graph topologies, they are on the opposite sides of the graphs, suggesting little if no interaction between these two genes. For predicting tumor stage, IL8 was excluded from the graph topology during the heuristic search, yet it was marked as highly important based on gradients. This suggests that while IL8 is crucial for prediction, it may not play a direct role in the non-canonical cGAS–STING pathway activation as IL8 often has the role of a downstream activator and provides feedback to the secreting cells to reinforce senescence signaling [56–59].

This study showcases promising results in the application of graphs for projecting biological data, especially the genomic pathways, and applying GCN models for transforming and interpreting complex gene expression data in cancer research. Incorporating integrated gradients and graph visualization with feature importance into our study was instrumental in addressing the commonly perceived issue of AI models functioning as "black boxes", which is a significant problem in many machine learning applications. By leveraging this technique, we were able to identify which features were most influential in driving the model's predictions, and intuitively display graph visualizations which further complemented the explainability with a clear and interpretable representation of the relationships and interactions among genes in the cGAS–STING pathway. By employing GCNs, we were able to not only predict outcomes but also visualize the complex interactions within the cGAS–STING pathway. The use of GCNs allowed us to mathematically represent these interactions as graphs, providing a more nuanced understanding of how the pathway behaves across different cancer types and stages. We also aimed to distinguish between the canonical and non-canonical activation of the cGAS–STING pathway, particularly in the context of cancer progression and tumor stage prediction. Our findings offer some intriguing clues that suggest differences in pathway activation when predicting tumor stage, specifically, the activation patterns of ATM, TMEM173 (STING), and IL8 genes observed through our analysis. Overall, our approach demonstrates the utility of graph-based methods in cancer research, particularly in elucidating complex biological pathways like cGAS–STING. To strengthen the validity and generalizability of our approach, further data are needed to validate the methodology across multiple cohorts. Expanding the dataset to include additional modalities, such as mutation data, gene amplifications and deletions, and methylation profiles, could significantly enhance the model's robustness. More importantly, integrating these diverse data types into a comprehensive graph structure will enable a more holistic representation of the underlying biological processes. This approach would be advantageous as it bypasses the need for the classical statistical analyses that often involve multiple testing corrections, which can lead to the exclusion of potentially important findings [16]. By incorporating various data sources directly into the model, we can utilize advanced interpretability methods such as integrated gradients, DeepLIFT [60], SHAP (SHapley Additive exPlanations) [27,28], LIME (Local Interpretable Model-agnostic Explanations) [29], Grad-CAM (Gradient-weighted Class Activation Mapping) [27], and similar frameworks, to provide a non-parametric assessment of input importance. This

would not only improve the model's predictive power but also offer deeper insights into the contributions of different data types, facilitating more precise and meaningful biological interpretations, which are crucial for precision medicine. Future research should also focus on integrating additional omics data and exploring the multiple pathways and their interactions with other immune mechanisms to explore capabilities of this framework on other use cases. The ability to visualize and quantify these interactions opens up new avenues for understanding tumor biology and could eventually lead to more targeted therapeutic strategies that consider the distinct activation states of critical immune pathways.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/make6030100/s1. Figure S1. The plot shows the training process of the models for each cancer type and use case after acquiring specific optimized graph topologies. Figure S2. This plot depicts the distribution of target variables (OS, DSS, and stage) across all cancer cohorts included in the study. Figure S3. The bar plot shows the sample count of each cancer cohort included in the study. Figure S4. The plots display the fitness values over generations, divided into three groups of 12 scenarios, demonstrating the progress and convergence of the heuristic search process for different scenarios. Figure S5. Panel (A) shows the training loss over epochs, split into three groups of 12 scenarios, indicating how well the model learns from the training data over time. Panel (B) presents the validation loss over epochs, also split into three groups of 12 scenarios, providing a measure of the model's performance on unseen data and its ability to generalize.

Author Contributions: Lead author M.S. was responsible for the conceptualization and design of the study. Methodology development and experimental design were carried out by B.S., M.S. and D.K.; B.S. implemented the graph convolutional network models while L.M. managed computational resources. Data collection and investigation were performed by B.S. and L.M., while data curation was managed by M.S. The original draft of the manuscript was written by M.S. and B.S., and all authors contributed to the review, editing, and revision of the manuscript. Visualization of data was handled by B.S and D.K. Supervision and project administration were managed by M.S. and L.M. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The data is public (Xena browser and TCGA). The github link with scripts is stated in Methods section.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Satam, H.; Joshi, K.; Mangrolia, U.; Waghoo, S.; Zaidi, G.; Rawool, S.; Thakare, R.P.; Banday, S.; Mishra, A.K.; Das, G.; et al. Next-Generation Sequencing Technology: Current Trends and Advancements. *Biology* **2023**, *12*, 997. [CrossRef]
2. Jain, M. Next-generation sequencing technologies for gene expression profiling in plants. *Briefings Funct. Genom.* **2011**, *11*, 63–70. [CrossRef]
3. Muir, P.; Li, S.; Lou, S.; Wang, D.; Spakowicz, D.J.; Salichos, L.; Zhang, J.; Weinstock, G.M.; Isaacs, F.; Rozowsky, J.; et al. The real cost of sequencing: Scaling computation to keep pace with data generation. *Genome Biol.* **2016**, *17*, 53.
4. Cost of NGS. Available online: https://emea.illumina.com/science/technology/next-generation-sequencing/beginners/ngs-cost.html (accessed on 16 February 2024).
5. Dong, Y.; Sun, F.; Ping, Z.; Ouyang, Q.; Qian, L. DNA storage: Research landscape and future prospects. *Natl. Sci. Rev.* **2020**, *7*, 1092–1107. [CrossRef] [PubMed]
6. Libbrecht, M.W. Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* **2015**, *16*, 321–332. [CrossRef]
7. Yi, N.; Xu, S.; Lou, X.-Y.; Mallick, H. Multiple Comparisons in Genetic Association Studies: A Hierarchical Modeling Approach. *Stat. Appl. Genet. Mol. Biol.* **2014**, *13*, 35. [CrossRef]
8. Groenwold, R.H.H.; Goeman, J.J.; Cessie, S.L.; Dekkers, O.M. Multiple testing: When is many too much? *Eur. J. Endocrinol.* **2021**, *184*, E11–E14. [CrossRef]

9. Zou, J.; Huss, M.; Abid, A.; Mohammadi, P.; Torkamani, A.; Telenti, A. A primer on deep learning in genomics. *Nat. Genet.* **2018**, *51*, 12–18. [CrossRef]

10. Geraci, F.; Saha, I.; Bianchini, M. RNA-Seq Analysis: Methods, Applications and Challenges. *Front. Genet.* **2020**, *11*, 220. [CrossRef]

11. Toro-Domínguez, D.; Villatoro-García, J.A.; Martorell-Marugán, J.; Román-Montoya, Y.; Alarcón-Riquelme, M.E.; Carmona-Sáez, P. A survey of gene expression meta-analysis: Methods and applications. *Brief. Bioinform.* **2021**, *22*, 1694–1705. [CrossRef] [PubMed]

12. Chatterjee, A.; Ahn, A.; Rodger, E.J.; Stockwell, P.A.; Eccles, M.R. A Guide for Designing and Analyzing RNA-Seq Data. *Methods Mol. Biol.* **2018**, *1783*, 35–80. [PubMed]

13. Song, M.; Greenbaum, J.; Luttrell, J., IV; Zhou, W.; Wu, C.; Shen, H.; Gong, P.; Zhang, C.; Deng, H.W. A Review of Integrative Imputation for Multi-Omics Datasets. *Front. Genet.* **2020**, *11*, 570255. [CrossRef]

14. Dong, X.; Liu, C.; Dozmorov, M. Review of multi-omics data resources and integrative analysis for human brain disorders. *Brief. Funct. Genomics.* **2021**, *20*, 223–234. [CrossRef]

15. Hardiman, G. *Systems Analytics and Integration of Big Omics Data*; MDPI: Basel, Switzerland, 2020.

16. Sokač, M.; Kjær, A.; Dyrskjøt, L.; Haibe-Kains, B.; Jwl Aerts, H.; Birkbak, N.J. Spatial transformation of multi-omics data unlocks novel insights into cancer biology. *Elife* **2023**, *12*, RP87133. [CrossRef]

17. Sokač, M.; Ahrenfeldt, J.; Litchfield, K.; Watkins, T.B.; Knudsen, M.; Dyrskjøt, L.; Jakobsen, M.R.; Birkbak, N.J. Classifying cGAS-STING Activity Links Chromosomal Instability with Immunotherapy Response in Metastatic Bladder Cancer. *Cancer Res. Commun.* **2022**, *2*, 762–771. [CrossRef]

18. Brandon, M.C.; Wallace, D.C.; Baldi, P. Data structures and compression algorithms for genomic sequence data. *Bioinformatics* **2009**, *25*, 1731–1738. [CrossRef] [PubMed]

19. Baaijens, J.A.; Bonizzoni, P.; Boucher, C.; Della Vedova, G.; Pirola, Y.; Rizzi, R.; Sirén, J. Computational graph pangenomics: A tutorial on data structures and their applications. *Nat. Comput.* **2022**, *21*, 81–108. [CrossRef]

20. Andreace, F.; Lechat, P.; Dufresne, Y.; Chikhi, R. Comparing methods for constructing and representing human pangenome graphs. *Genome. Biol.* **2023**, *24*, 274. [CrossRef]

21. Woolson, R.F.; Clarke, W.R. *Statistical Methods for the Analysis of Biomedical Data*; John Wiley & Sons: Hoboken, NJ, USA, 2011.

22. Pavlopoulos, G.A.; Secrier, M.; Moschopoulos, C.N.; Soldatos, T.G.; Kossida, S.; Aerts, J.; Schneider, R.; Bagos, P.G. Using graph theory to analyze biological networks. *BioData Min.* **2011**, *4*, 10. [CrossRef] [PubMed]

23. Sedgwick, P. Pitfalls of statistical hypothesis testing: Multiple testing. *BMJ* **2014**, *349*, g5310. [CrossRef]

24. Milano, M.; Agapito, G.; Cannataro, M. Challenges and Limitations of Biological Network Analysis. *BioTech* **2022**, *11*, 24. [CrossRef]

25. Koutrouli, M.; Karatzas, E.; Paez-Espino, D.; Pavlopoulos, G.A. A Guide to Conquer the Biological Network Era Using Graph Theory. *Front. Bioeng. Biotechnol.* **2020**, *8*, 34. [CrossRef] [PubMed]

26. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic Attribution for Deep. Networks. 2017. Available online: http://arxiv.org/abs/1703.01365 (accessed on 23 June 2024).

27. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Available online: https://ieeexplore.ieee.org/document/8237336 (accessed on 3 September 2024).

28. Lundberg, S.; Lee, S.-I.A. Unified Approach to Interpreting Model Predictions. 2017. Available online: http://arxiv.org/abs/1705.07874 (accessed on 23 July 2024).

29. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust. You?": Explaining the Predictions of Any Classifier. 2016. Available online: http://arxiv.org/abs/1602.04938 (accessed on 24 June 2024).

30. Johnson, K.B.; Wei, W.; Weeraratne, D.; Frisse, M.E.; Misulis, K.; Rhee, K.; Zhao, J.; Snowdon, J.L. Precision Medicine, AI, and the Future of Personalized Health Care. *Clin. Transl. Sci.* **2021**, *14*, 86–93. [CrossRef] [PubMed]

31. Sahu, M.; Gupta, R.; Ambasta, R.K.; Kumar, P. Artificial intelligence and machine learning in precision medicine: A paradigm shift in big data analysis. *Prog. Mol. Biol. Transl. Sci.* **2022**, *190*, 57–100. [PubMed]

32. Paten, B.; Novak, A.M.; Eizenga, J.M.; Garrison, E. Genome graphs and the evolution of genome inference. *Genome Res.* **2017**, *27*, 665. [CrossRef]

33. Du, J.-M.; Qian, M.-J.; Yuan, T.; Chen, R.-H.; He, Q.-J.; Yang, B.; Ling, Q.; Zhu, H. cGAS and cancer therapy: A double-edged sword. *Acta Pharmacol. Sin.* **2022**, *43*, 2202–2211. [CrossRef]

34. Wang, M.M.; Zhao, Y.; Liu, J.; Fan, R.R.; Tang, Y.Q.; Guo, Z.Y.; Li, T. The role of the cGAS-STING signaling pathway in viral infections, inflammatory and autoimmune diseases. *Acta Pharmacol. Sin.* **2024**, 1–14. [CrossRef]

35. Cheng, Z.; Dai, T.; He, X.; Zhang, Z.; Xie, F.; Wang, S.; Zhang, L.; Zhou, F. The interactions between cGAS-STING pathway and pathogens. *Signal Transduct. Target. Ther.* **2020**, *5*, 1–15. [CrossRef]

36. Ou, L.; Zhang, A.; Cheng, Y.; Chen, Y. The cGAS-STING Pathway: A Promising Immunotherapy Target. *Front. Immunol.* **2021**, *12*, 795048. [CrossRef]

37. He, W.; Mu, X.; Wu, X.; Liu, Y.; Deng, J.; Liu, Y.; Han, F.; Nie, X. The cGAS-STING pathway: A therapeutic target in diabetes and its complications. *Burns Trauma.* **2024**, *12*, tkad050. [CrossRef]

38. Hong, C.; Tijhuis, A.E.; Foijer, F. The cGAS Paradox: Contrasting Roles for cGAS-STING Pathway in Chromosomal Instability. *Cells* **2019**, *8*, 1228. [CrossRef] [PubMed]

39. Gan, Y.; Li, X.; Han, S.; Liang, Q.; Ma, X.; Rong, P.; Wang, W.; Li, W. The cGAS/STING Pathway: A Novel Target for Cancer Therapy. *Front. Immunol.* **2021**, *12*, 795401. [CrossRef] [PubMed]

40. Li, X.-J.-Y.; Qu, J.-R.; Zhang, Y.-H.; Liu, R.-P. The dual function of cGAS-STING signaling axis in liver diseases. *Acta Pharmacol. Sin.* **2024**, *45*, 1115–1129. [CrossRef]

41. Khoo, L.T.; Chen, L. Role of the cGAS–STING pathway in cancer development and oncotherapeutic approaches. *EMBO Rep.* **2018**, *19*, e46935. [CrossRef]

42. Goldman, M.J.; Craft, B.; Hastie, M.; Repečka, K.; McDade, F.; Kamath, A.; Banerjee, A.; Luo, Y.; Rogers, D.; Brooks, A.N.; et al. Visualizing and interpreting cancer genomics data via the Xena platform. *Nat. Biotechnol.* **2020**, *38*, 675–678. [CrossRef]

43. Vivian, J.; Rao, A.A.; Nothaft, F.A.; Ketchum, C.; Armstrong, J.; Novak, A.; Pfeil, J.; Narkizian, J.; Deran, A.D.; Musselman-Brown, A.; et al. Toil enables reproducible, open source, big biomedical data analyses. *Nat. Biotechnol.* **2017**, *35*, 314–316. [CrossRef]

44. Huang, Y.; Liu, B.; Sinha, S.C.; Amin, S.; Gan, L. Mechanism and therapeutic potential of targeting cGAS-STING signaling in neurological disorders. *Mol. Neurodegener.* **2023**, *18*, 79. [CrossRef]

45. Bai, J.; Liu, F. The cGAS-cGAMP-STING Pathway: A Molecular Link Between Immunity and Metabolism. *Diabetes* **2019**, *68*, 1099. [CrossRef]

46. Patel, D.J.; Yu, Y.; Xie, W. cGAMP-activated cGAS–STING signaling: Its bacterial origins and evolutionary adaptation by metazoans. *Nat. Struct. Mol. Biol.* **2023**, *30*, 245–260. [CrossRef]

47. Decout, A.; Katz, J.D.; Venkatraman, S.; Ablasser, A. The cGAS–STING pathway as a therapeutic target in inflammatory diseases. *Nat. Rev. Immunol.* **2021**, *21*, 548–569. [CrossRef]

48. Hopfner, K.-P.; Hornung, V. Molecular mechanisms and cellular functions of cGAS–STING signalling. *Nat. Rev. Mol. Cell. Biol.* **2020**, *21*, 501–521. [CrossRef] [PubMed]

49. Pan, J.; Fei, C.-J.; Hu, Y.; Wu, X.-Y.; Nie, L.; Chen, J. Current understanding of the cGAS-STING signaling pathway: Structure, regulatory mechanisms, and related diseases. *Zool. Res.* **2023**, *44*, 183. [CrossRef]

50. Withers, S.S.; Moeller, C.E.; Quick, C.N.; Liu, C.-C.; Baham, S.M.; Looper, J.S.; Subramanian, R.; Kousoulas, K.G. Effect of stimulator of interferon genes (STING) signaling on radiation-induced chemokine expression in human osteosarcoma cells. *PLoS ONE* **2023**, *18*, e0284645. [CrossRef]

51. Motwani, M.; Pesiridis, S.; Fitzgerald, K.A. DNA sensing by the cGAS–STING pathway in health and disease. *Nat. Rev. Genet.* **2019**, *20*, 657–674. [CrossRef] [PubMed]

52. Banerjee, D.; Langberg, K.; Abbas, S.; Odermatt, E.; Yerramothu, P.; Volaric, M.; Reidenbach, M.A.; Krentz, K.J.; Rubinstein, C.D.; Brautigan, D.L.; et al. A non-canonical, interferon-independent signaling activity of cGAMP triggers DNA damage response signaling. *Nat. Commun.* **2021**, *12*, 1–24. [CrossRef]

53. Dunphy, G.; Flannery, S.M.; Almine, J.F.; Connolly, D.J.; Paulus, C.; Jønsson, K.L.; Jakobsen, M.R.; Nevels, M.M.; Bowie, A.G.; Unterholzner, L. Non-canonical Activation of the DNA Sensing Adaptor STING by ATM and IFI16 Mediates NF-κB Signaling after Nuclear DNA Damage. *Mol. Cell.* **2018**, *71*, 745–760.e5. [CrossRef]

54. Lim, S.M.; Hong, M.H.; Kim, H.R. Immunotherapy for Non-small Cell Lung Cancer: Current Landscape and Future Perspectives. *Immune Netw.* **2020**, *20*, e10. [CrossRef]

55. Taluri, S.; Oza, V.H.; Soelter, T.M.; Fisher, J.L.; Lasseigne, B.N. Inferring chromosomal instability from copy number aberrations as a measure of chromosomal instability across human cancers. *Cancer Rep.* **2023**, *6*, e1902. [CrossRef]

56. Acosta, J.C.; O'Loghlen, A.; Banito, A.; Guijarro, M.V.; Augert, A.; Raguz, S.; Fumagalli, M.; Da Costa, M.; Brown, C.; Popov, N.; et al. Chemokine signaling via the CXCR2 receptor reinforces senescence. *Cell* **2008**, *133*, 1006–1018. [CrossRef] [PubMed]

57. Kuilman, T.; Michaloglou, C.; Vredeveld, L.C.; Douma, S.; van Doorn, R.; Desmet, C.J.; Aarden, L.A.; Mooi, W.J.; Peeper, D.S. Oncogene-induced senescence relayed by an interleukin-dependent inflammatory network. *Cell* **2008**, *133*, 1019–1031. [CrossRef]

58. Dou, Z.; Ghosh, K.; Vizioli, M.G.; Zhu, J.; Sen, P.; Wangensteen, K.J.; Simithy, J.; Lan, Y.; Lin, Y.; Zhou, Z.; et al. Cytoplasmic chromatin triggers inflammation in senescence and cancer. *Nature* **2017**, *550*, 402–406. [CrossRef] [PubMed]

59. Glück, S.; Guey, B.; Gulen, M.F.; Wolter, K.; Kang, T.-W.; Schmacke, N.A.; Bridgeman, A.; Rehwinkel, J.; Zender, L.; Ablasser, A. Innate immune sensing of cytosolic chromatin fragments through cGAS promotes senescence. *Nat. Cell Biol.* **2017**, *19*, 1061–1070. [CrossRef] [PubMed]

60. Shrikumar, A.; Greenside, P.; Kundaje, A. Learning Important Features Through Propagating Activation Differences. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017.