



Randomly Wired Neural Networks for IMRT 3D dose prediction in oropharyngeal cancer

Pakshootra D.J.¹, Pakshootra K.J.², Lee Y.H.³

Submitted: September 11, 2025, Revised: version 1, April 10, 2026, version 2, July 7, 2026, version 3, July 19, 2026

Accepted: August 5, 2026

Abstract

Intensity-modulated radiation therapy (IMRT) requires patient-specific three-dimensional dose distributions that maximize target coverage while limiting irradiation of surrounding organs at risk (OARs). Although deep-learning approaches can accelerate dose prediction, commonly used architectures such as U-Nets rely on manually designed connectivity patterns that constrain exploration of the architectural search space. This study investigates randomly wired neural networks (RWNNs) for oropharyngeal cancer dose prediction. Architectures were generated from three stochastic graph families—Erdős-Rényi, Barabási-Albert, and Watts-Strogatz—and optimized using Bayesian optimization. Performance was evaluated using five-fold cross-validation and five random seeds against a U-Net baseline across voxel-wise, spatial, target-volume, and OAR metrics. RWNNs achieved comparable or significantly improved performance across several metrics, including dose error, spatial agreement, PTV₇₀D₉₅, homogeneity, and spinal-cord Dmax, although the U-Net retained superior brainstem dose accuracy. These results were achieved with substantially fewer trainable parameters, with the Watts-Strogatz architecture containing nearly 30-fold fewer parameters than the U-Net, although RWNNs required greater activation memory and inference time. Seed analysis indicated generally stable performance, while node ablation revealed no significant deterioration after removal of individual nodes, suggesting architectural redundancy and opportunities for further compression. These findings demonstrate that stochastic graph-based architecture search can identify compact networks capable of competitive IMRT dose prediction. Determining how graph properties such as connectivity, clustering, path length, and hub structure influence dose-prediction performance may provide a mechanistic basis for designing task-specific neural architectures.

Keywords

Randomly Wired Neural Networks, Radiotherapy dose prediction, Intensity-Modulated Radiation Therapy, Neural architecture search, Random graph neural networks, Deep learning, Radiation treatment planning, Oropharyngeal cancer, Medical image analysis, Dose distribution prediction

¹Dev Jay Pakshootra, pakshootra.dev.jay@gmail.com, ²Krithin Jay Pakshootra, krithinjayp@gmail.com, Nexus International School, 1 Aljunied Walk, Singapore 387293.

³Corresponding author: Yong Heng Lee, National University of Singapore, 21 Lower Kent Ridge Road, Singapore 119077. garyleeyh@physics.org

1. Introduction

In 2022 alone, there were close to 20 million new cancer cases and 9.7 million cancer-related deaths (1). As global cancer incidence rates rise, so does the need for precise treatments. One such technique is Intensity-Modulated Radiotherapy (IMRT), an advanced form of 3D Conformal Radiotherapy (2). As the name implies, radiation doses are modulated to focus on tumours and target volumes, minimizing damage to surrounding organs (3). For this reason, IMRT has been considered one of the most successful developments in radiation oncology since the introduction of computed tomography (CT) into treatment planning (4,5).

To precisely control where radiation is delivered in IMRT, a 3-dimensional dose distribution is generated for each patient. This defines the radiation dosage for a target area by splitting it into small voxels (3D pixels) and assigning each voxel a radiation dose (6). Oropharyngeal cancer is particularly complex due to the proximity of vital organs such as the spinal cord and brainstem, which are extremely sensitive to radiation (7). IMRT employs concave dose distributions to avoid damaging vital organs such as the spinal cord. This method has proved to be effective, with 2-year overall survival rates of 91-96% for oropharyngeal cancer (8).

High-quality dose distributions are essential to allow effective delivery of IMRT, especially in oropharyngeal cancer treatment. If tumour cells are missed, the chances for relapse increase greatly (9). Conversely, overdosing can cause damage to vital organs such as the spinal cord, which can significantly decrease patient quality of life. Therefore, the role of creating dose distributions is assigned to a specialist called a dosimetrist. Before a dosimetrist can create a

dose distribution for oropharyngeal cancer, a lengthy planning process needs to take place. As part of diagnosis, doctors must visualise the head and neck area using medical imaging technologies such as CT scans and MRIs. These scans are then contoured to delineate the important regions, which can be separated into two distinct categories, [1] Planned Target Volumes (PTVs): a tissue volume that contains the tumour and surrounding afflicted areas, with margins added to account for variability in treatment (10) and [2] Organs At Risk (OARs) (otherwise known as critical normal structures): tissues that if irradiated could suffer significant morbidity and thus might influence the treatment planning and/or the absorbed-dose prescription (10).

Dosimetrists input these contoured scans into a Treatment Planning System (TPS). This software aims to maximize radiation delivered to the PTV, and minimize radiation to OARs (11). By repeatedly tweaking weights and priorities, an optimal dose distribution is created. The final treatment parameters can then be calculated from the dose distribution using a technique called Inverse Treatment Planning (ITP): the radiation beam setup can be derived from the existing dose distribution (12).

However, the iterative nature of this process means that creating a dose distribution can take up to several hours (13). With hospitals often being under-resourced in terms of personnel and equipment, such time pressures can lead to sub optimal planning, long patient wait times, and low scalability for higher volumes of patients (14). In the UK, patients wait 63 days on average to receive treatment (15). This is concerning because tumour progression during treatment delays may compromise treatment accuracy. As

the tumour enlarges, its margins may become less well defined and extend toward or into adjacent OARs, reducing the accuracy of the original treatment plan. Such treatment delays have also been associated with poorer patient outcomes (16).

Deep learning models have significant potential to accelerate the generation of radiotherapy dose distributions, thereby increasing patient throughput and allowing dosimetrists to devote more time to quality assurance and complex cases (17). Faster dose prediction may improve planning efficiency and, provided that the predicted distributions can be accurately translated into deliverable treatment plans, could ultimately contribute to improved tumour control. Deep learning may also facilitate adaptive radiotherapy by enabling more rapid replanning as tumour anatomy changes over the course of treatment. However, these benefits depend critically on the ability of deep learning models to generate dose distributions with sufficient accuracy and reliability for clinical use.

Current research has investigated U-Net and ResNet architectures for radiotherapy dose prediction (18,19). Although these models have shown promising performance, their architectures and connectivity patterns are largely predefined by human researchers. Consequently, the range of architectures explored is constrained by human design choices and may exclude unconventional configurations that are better suited to dose-distribution prediction. Manually designed architectures may also embed researchers' assumptions and prior expectations, potentially limiting objective exploration of the broader architectural search space. Stochastic graph-

based neural architectures offer an alternative approach in which network connectivity is generated with reduced reliance on predefined human design choices. However, their application to radiotherapy dose prediction, and their potential to identify effective architectures beyond conventional human-designed models, remains underexplored.

Therefore, this study investigates the use of Randomly Wired Neural Networks (RWNNs) for radiotherapy dose-distribution prediction. By introducing stochastic network connectivity, RWNNs expand the architectural search space beyond conventional human-designed configurations and reduce reliance on researcher-defined architectural priors. This broader search space may enable the identification of network structures better suited to modeling the complex spatial relationships between tumors, surrounding tissues, and organs at risk. The rationale for randomly wired architectures is supported by Xie, Kirillov, Girshick, and He (2019), who demonstrated unexpectedly strong performance on image-recognition tasks using the CIFAR-10 dataset (20). However, the extent to which such architectures generalize across domains and tasks remains uncertain, motivating their evaluation in the distinct and clinically relevant context of radiotherapy dose prediction.

2. Related work

2.1 Deep Learning in dose distribution prediction

Recently, deep learning has emerged as a promising approach for predicting radiotherapy dose distributions from imaging and contour data. Nguyen et al. (21) first applied a U-Net architecture to dose prediction for prostate

cancer patients undergoing IMRT. Subsequent studies investigated ResNet (22) and Res-U-Net (23) architectures, achieving further reductions in Mean Absolute Error (MAE). Despite these advances, prediction accuracy remains a critical consideration because clinical radiotherapy requires highly precise dose delivery, and model performance can vary substantially across anatomical sites and cancer types. Dose prediction is generally more challenging in head-and-neck cancers than in prostate cancer because of the greater number and anatomical complexity of surrounding OARs. Accordingly, this study evaluates randomly wired neural architectures using oropharyngeal cancer cases, providing a more challenging setting in which to assess their potential for radiotherapy dose prediction.

2.2 Neural architectural design

Neural network connectivity has evolved considerably over time. Beginning with early feed-forward perceptrons (24), neural architectures were initially dominated by relatively sequential, feed-forward structures, including conventional Convolutional Neural Networks (CNNs) (25). Over time, architectures became increasingly complex and less strictly sequential. For example, Inception networks introduced multiple parallel computational branches to address computer vision tasks (26). As networks grew deeper, difficulties associated with vanishing and exploding gradients became increasingly important. ResNets (27) addressed the degradation of information and gradients in deep networks through residual skip connections, allowing information from earlier layers to bypass intermediate transformations. U-Nets (28) subsequently combined an encoder-decoder structure with skip connections between corresponding resolution levels, an

architecture particularly well suited to biomedical image segmentation.

Despite their success, these architectures remain fundamentally human-designed. Manual architectural exploration is necessarily constrained by the time and cognitive limitations of researchers, restricting the number and diversity of connectivity patterns that can feasibly be conceived and evaluated. Consequently, potentially effective architectures may exist outside the range of structures that appear intuitive to human designers. Exploring non-intuitive or stochastically generated connectivity patterns may therefore reveal neural architectures capable of outperforming conventional human-designed configurations.

More recently, machine-generated neural architectures have gained increasing attention. Zoph and Le (29) used reinforcement learning (RL) to train a recurrent neural network (RNN) that acted as a “network generator,” automatically producing and evaluating candidate neural architectures. The resulting architectures achieved performance competitive with human-designed models on the CIFAR-10 dataset. This work helped establish the field of Neural Architecture Search (NAS), in which computational methods systematically explore a predefined architectural search space under specified constraints. By evaluating a substantially broader range of connectivity patterns than is feasible through manual design, NAS provides a framework for discovering architectures that can compete with or potentially outperform conventional human-designed networks.

2.3 Randomly Wired Neural Networks

Xie, Kirillov, Girshick and He (2019) further loosened the constraints of Neural Architectural Search to explore completely random networks called Randomly Wired Neural Networks (RWNNs). They used classical random graph generation models, mapping network architectures to Directed Acyclic Graphs (DAGs) (20). Their RWNNs were competitive against both hand-designed and optimized

models from NAS on the CIFAR-10 datasets. Similarly, Du et al. applied RWNNs to brain tumour image classification and attained comparable results to other models using the same dataset (30).

3. Materials and Methods

3.1 Dataset

The 2020 OpenKBP challenge dataset (31) was

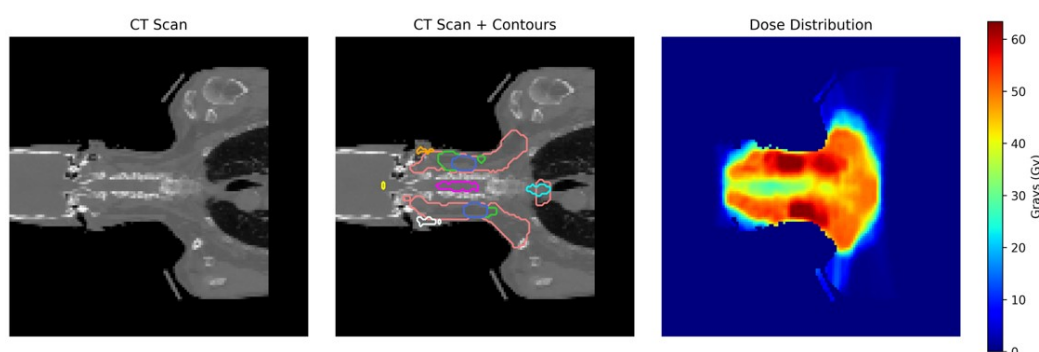


Figure 1. 2D slice of patient data, showing CT scan (HU), contours, and dose distribution (Gy)

used to train the model. It contains data for 340 oropharyngeal (head and neck) cancer patients. This data was initially acquired from The Cancer Imaging Archive (TCIA) before being cleaned and standardized. Each patient is associated with a 3D CT scan, binary structure masks for the PTVs and OARs and a feasible dose mask - this defines where radiation can be applied. Finally, a planned dose distribution was attached to each patient profile.

The provided dose distributions were derived from fluence-based treatment plans with comparable levels of fluence complexity, ensuring consistency across patient data (32). All plans used nine equally spaced coplanar fields positioned from 0° to 320° in 40° increments. Each plan typically prescribed a

maximum dose of 70 Gy delivered in 35 fractions using 6 MV step-and-shoot IMRT beams. Dose delivery was modeled as a collection of individual beamlets, with the contribution of each beamlet to the patient dose represented by a dose-deposition matrix. To maintain consistent treatment physics across cases, all dose-deposition matrices were generated using identical parameters within the Computational Environment for Radiotherapy Research (CERR) (33).

The original dataset contained multiplanar reconstructions with dimensions of $128 \times 128 \times 128$ voxels. Due to computational constraints, each scan and corresponding dose distribution was downsampled to $64 \times 64 \times 64$ voxels, resulting in 262,144 voxels per volume. The

resulting voxels had an average size of approximately $7 \times 7 \times 4$ mm, with exact patient-specific voxel dimensions provided in the dataset. To enable consistent comparison across models, five-fold cross-validation was used. In each fold, 20% of the dataset was reserved as the test set. The remaining 80% was further divided, with 80% used for training and 20% for validation, corresponding to approximately 64% and 16% of the full dataset, respectively. Each patient was assigned to only one test fold per random seed, ensuring that the training, validation, and test partitions remained patient-independent.



Figure 2. K-fold cross validation diagram

The binary structure masks required no normalization. Dose distributions were min-max normalized to a range of 0 to 1 and subsequently converted back to physical dose values during post-processing by multiplying the predicted values by 80 Gy, the maximum dose represented in the dataset. The CT scans were provided in a mixture of 12-bit and 16-bit formats. To standardize their intensity range, CT values were clipped to 0–4095, corresponding to a 12-bit representation, and then min-max normalized to a range of 0 to 1. The preprocessed data were subsequently stacked to form a 12-channel model input with dimensions of $[12, 64, 64, 64]$, comprising the CT volume, ROI masks arranged in alphabetical order, and the feasible-dose mask.

Table 1. Variance Inflation Factor (VIF) for each feature

Feature	VIF	Feature	VIF
CT	3.311	Left Parotid	1.004
PTV ₅₆	1.050	Right Parotid	1.004
PTV ₆₃	1.013	Larynx	1.003
PTV ₇₀	1.058	Spinal Cord	1.004
Brainstem	1.003	Mandible	1.018
Esophagus	1.001	Feasible Dose Mask	3.398

To assess redundancy among the input channels, variance inflation factors (VIFs) were calculated

across all voxels pooled from the training set (Table 1). All VIF values were below 3.40, substantially below the conventional threshold of 5 for problematic multicollinearity, indicating that no channel exhibited substantial linear dependence on the combined set of remaining channels. Pairwise correlation analysis identified one strong association ($r = 0.83$), between the CT and feasible-dose-mask channels. This relationship was expected because the feasible dose mask is derived from the CT data, with both channels delineating the patient volume from the surrounding air. All other pairwise correlations were below 0.23, while correlations among the structure masks were below 0.03. Consistent with these findings, anatomical overlap among the binary structure masks was minimal, with Jaccard indices of $J < 0.005$ for all pairwise comparisons. Together, these analyses indicate limited redundancy among the model's input channels despite the expected association between the CT and feasible-dose-mask inputs.

3.2 Network generators

A network generator can be defined as a mapping from a parameter space to a space of neural network architectures (20). Given a specified set of parameters, the generator produces a corresponding neural network architecture. Conventional network generators are deterministic, meaning that identical input parameters will consistently produce the same architecture. Stochastic network generators extend this framework by incorporating randomness into the architecture-generation process. This is typically achieved by introducing a random seed as an additional parameter, resulting in reproducible pseudo-random architectures. For a fixed set of generator parameters, changing the random seed

can therefore produce different network architectures, while retaining the same seed allows a particular architecture to be reproduced exactly.

3.2.1 Random graph models

At its core, the stochastic network generator used in this research is a random graph generation model. In graph theory, a graph refers to a set of nodes connected by edges. Every node V_i has a degree $deg(V_i)$, which represents the total number of connections. Any graph contains paths - a sequence of nodes and edges that starts at a node, moves to an adjacent node, and continues without repeating any nodes or edges. These ideas form the basis of graph theory. There are several methods to create random graphs, each of which uses distinct probabilistic processes. These are separated into different random graph families. This research uses the Watts-Strogatz, Erdős-Rényi and Barabási-Albert models, three of the most common and most popular random graph models currently. When first creating a random graph, there are no restrictions on the graph, which allows for a wider search space. Every graph model takes in a parameter n , the number of nodes in the graph, and additional parameters that vary by model.

The Erdős-Rényi model is the oldest and simplest model. Given n nodes, the model takes an extra parameter p - the probability that two nodes are connected by an edge, independent of all other nodes and edges. After generating n nodes, the model considers every pair of nodes and connects them with probability p (34).

The Barabási-Albert model is used to generate scale-free networks; these networks tend to have high variance in the connectivity of nodes and

contain nodes called hubs with unusually high degrees as compared to the rest of the graph. Along with the graph size n , the model takes in a parameter m : the initial number of nodes (35). Barabási-Albert models are generated through a mechanism known as preferential attachment. The construction begins with an initial connected graph of m nodes. At each subsequent time step, a new node is added to the network and forms m edges to existing nodes. The probability that in graph G of size $V(G)$, a new node connects to an existing node V_i is proportional to the current degree $\deg(V_i)$ of that

node, such that: $P(V_i) = \frac{d(V_i)}{\sum_{n=0} d(V_n)}$ (eq.1). The

nature of this formula is such that nodes with a high degree tend to make more connections while nodes with a low degree are likely to remain low degree.

Watts–Strogatz graphs take in two additional parameters, k and p . The construction of a WS random graph begins with a regular ring lattice of n nodes, where each node is connected to its k nearest neighbours, with $\frac{k}{2}$ neighbors on either side (36). This initial configuration exhibits high local clustering but large average path lengths.

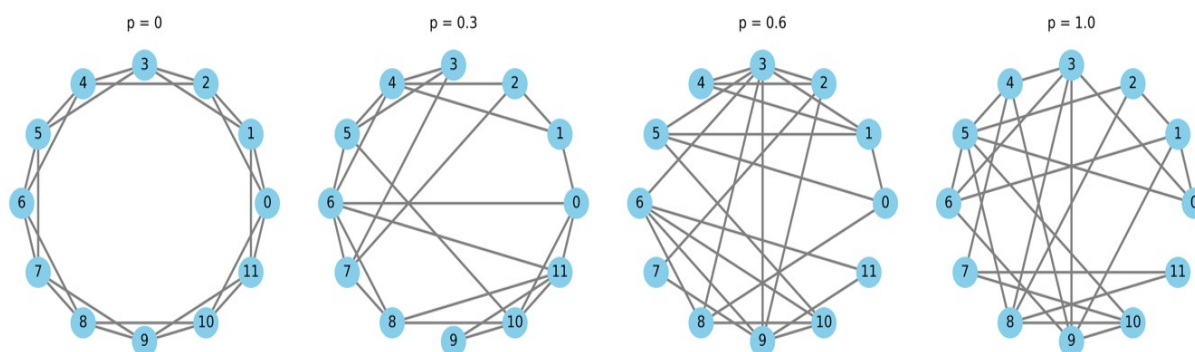


Figure 3. Diagram showing the process of creating a Watts-Strogatz graph. From the left, the graph progresses from the original ring lattice to the final random graph

To introduce stochasticity to the graph, each edge in the lattice is rewired with probability p . During this rewiring process, an edge is removed and replaced by a new edge connecting the original node to a randomly chosen node in the graph, excluding self-loops and duplicate edges. When $p = 0$, the graph remains a regular lattice, while $p = 1$ produces a structure similar to a random graph. For intermediate values of p , the model exhibits high clustering

coefficients and short average path lengths. This leads to the small-world phenomenon: most nodes can reach other nodes using short paths. Due to this property, Watts–Strogatz graphs are able to capture key structural properties observed in large social networks. They also resemble the neural connections within certain species of roundworm (36), motivating interest in their topologies for use in neural networks.

3.2.2 Converting to neural networks

There are two requirements to map a random graph to a neural network architecture. First, the graph must not contain any cycles. A cycle is a path that starts and ends at the same node, forming a closed loop; second, the graph must have a direction of data flow. When adding direction to a graph, vertices can have an incoming degree (incoming connections) and

outgoing degree (outgoing connections), as well as a total degree.

These two conditions can be satisfied by converting each model to a Directed Acyclic Graph (DAG). Each edge specifies a direction (an input node and an output node), and all cycles are removed from the graph.

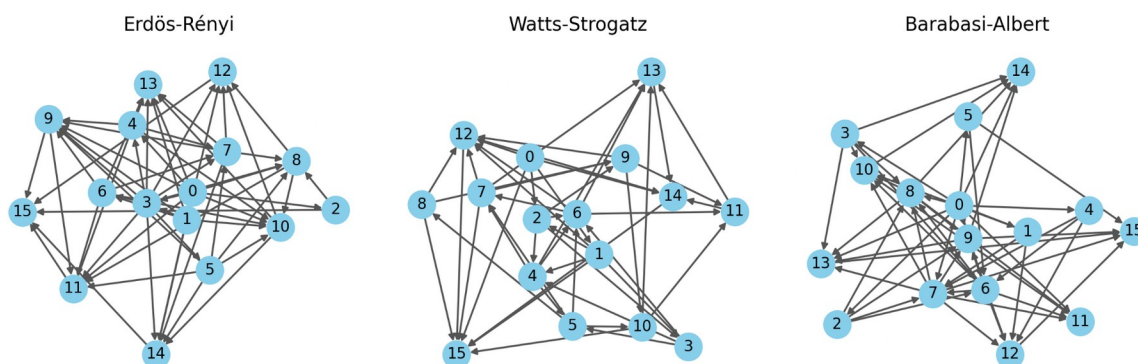


Figure 4. Examples of Erdős-Rényi, Watts-Strogatz and Barabási-Albert graphs with 16 nodes each converted to DAGs

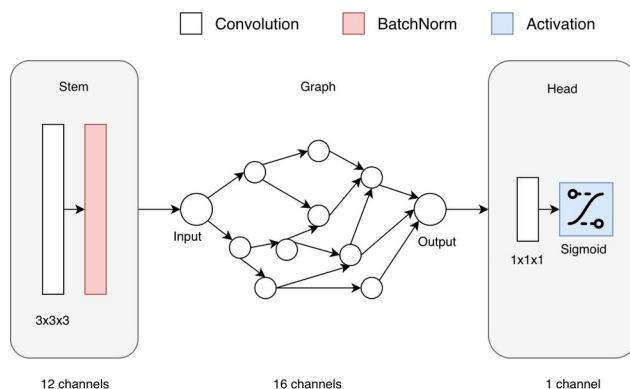


Figure 5. Structure of network generated from a graph. The graph is sampled from a random network family, and wrapped with a stem and head

Every node is then given a functional computation. For all DAGs, there are three main types of computations. First, *aggregation*; the process of combining inputs from multiple

nodes. In our model, the information from incoming edges is aggregated as a weighted sum. Then, all inputs are stacked and a softmax activation is applied to ensure that the weights

sum up to 1. Second, *transformation*; the processing of aggregated input. Our model uses a Conv-BatchNorm-ReLU triplet as a baseline, then explores different options during graph and architectural search. Third, *distribution*; the sending of information to the next nodes. Our model only applies a ReLU activation function here.

To create a final model output, all of the output nodes of the graph (nodes with an outgoing degree of 0) are connected to the final output node, which takes a weighted sum of all the output nodes. The graph model is then wrapped with a stem and head architecture. The stem applies an initial 3x3x3 convolution and normalization, and resamples data to increase the number of channels. The head contains a final 1x1x1 convolution and a sigmoid activation function to ensure outputs are between 0 and 1.

3.3 Graph search

3.3.1 Graph search strategy

We employed Bayesian Optimization to identify an optimal network architecture by exploring

the graph and node-operation design space. The search was performed independently for each graph model, with the objective of minimizing validation loss.

For the Erdős–Rényi (ER) model, the edge probability parameter p was sampled from a continuous range between 0.1 and 0.9. In the Barabási–Albert (BA) model, the parameter m , representing the number of edges attached by each new node, was varied between 2 and 10. For the Watts–Strogatz (WS) model, two parameters were optimized: the number of nearest neighbours k , sampled from the discrete set [2,4,6,8,10], and the rewiring probability p , sampled from the range 0.1 to 0.9. Below 0.1, connectivity is extremely low, so many parameters and convolution operations are not used. Above 0.9, the Watts–Strogatz topology increasingly approaches a random-network regime and loses the characteristic balance of local clustering and short path lengths associated with the small-world regime. The graph size n , corresponding to the number of nodes in the network, was also treated as a tunable parameter. n was restricted to integers between 16 and 32.

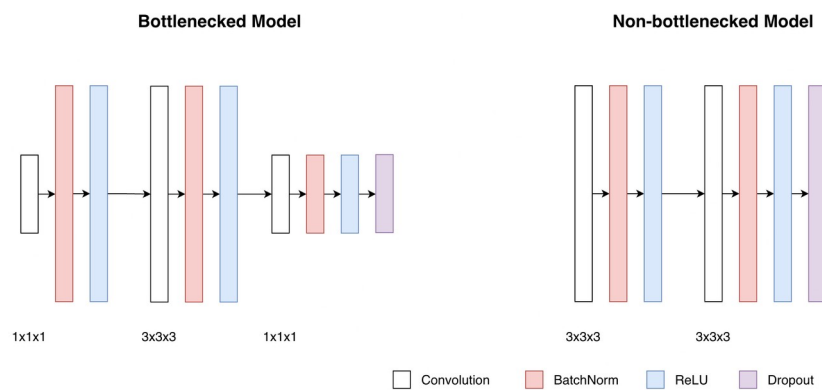


Figure 6. Comparison between bottlenecked and non-bottlenecked models

In addition to graph topology, some node-level operations were optimized. The kernel size was optimized between 3 and 5 due to computational constraints. We also tested BatchNorm, InstanceNorm and GroupNorm. Each model also had two architectural options: bottlenecked or non-bottlenecked. Bottlenecked models had a 1x1x1 convolution, followed by a 3x3x3 convolution, and then another 1x1x1 convolution. The 1x1x1 convolutions compress the number of channels and reduce the 3D size of data. This helps the model encode coarser details in CT scans and binary masks. The non-bottlenecked models have two successive 3x3x3 convolution, which maintains the channel count and creates a wider receptive field. This means the model wouldn't compress data, allowing it to reproduce finer details in the output dose distributions.

Optimization was performed using Optuna (37), with the search direction set to minimization. Models were evaluated based on their validation performance, using a composite objective defined as $L_{obj} = \mu(L_{val}) + 0.5\sigma(L_{val})$ (eq.2), where μ and σ denote the mean and standard deviation of the validation loss across repeated training runs. This formulation penalizes architectures with high variance, favoring more stable training behaviour.

For each graph model, a total of 100 optimization trials were conducted. In each trial, a model was sampled and trained three times with different random seeds to account for stochasticity in graph generation and model training. Random seeds were set for the NetworkX graph generation, PyTorch, NumPy, and Python's random module.

To reduce computational cost during the search phase, a reduced dataset was used. This consisted of 32 training patients and 8 validation patients, corresponding to $\sim 15\%$ of the full training and validation folds, respectively. Training was performed with a batch size of 8 due to computational constraints. To maintain fair comparison between models, we used a fixed learning rate of 10^{-3} instead of performing hyperparameter tuning. No learning rate scheduler was applied in order to maintain a linear training regime. Using an aggressive learning rate scheduler in the search stage would yield models that performed well in the initial stages of training, but would fail to improve in longer training. All models were optimized using the AdamW optimizer, with a weight decay of 10^{-5} , which is standard for medical imaging tasks.

3.4 Final model training

After the graph search stage, the best models were selected and used for final model training, along with a baseline model. The complete dataset was used for training.

To train the models, we used a cosine-annealing learning rate scheduler, starting at a maximum learning rate of 10^{-3} . The initially high learning rates allow the model to explore the parameter space. The scheduler subsequently decreases the learning rate so model parameters can be fine-tuned. This reduces overfitting and helps the models escape early local minima in the cost function.

To appropriately penalize errors in the predicted dose distributions, a composite loss function was used. Preliminary testing showed that, without masking, models tended to predict values close to zero because the large number of

voxels outside the patient volume contain little or no dose, resulting in an artificially low MAE. To prevent these voxels from dominating the optimization objective, the feasible dose mask was applied so that the loss was calculated only within regions in which radiation dose was relevant, excluding areas such as the surrounding air. Within this region, prediction error was quantified using L1 loss (MAE). Errors within regions of interest (ROIs) were assigned additional weight to more strongly penalize clinically relevant deviations, particularly underdosage of PTVs and overdosage of OARs.

Each model was trained for a maximum of 150 epochs. Early stopping was employed with a patience of 20 epochs, terminating training if the validation loss did not improve for 20 consecutive epochs. The model with the best recorded validation loss was then saved. Similarly to the graph search stage, models were optimized using the AdamW optimizer and a weight decay of 10^{-5} for stable model convergence. The models were trained on two T4 GPUs, using the DataParallel API in PyTorch. The models were trained on 5 folds, and each training was repeated using 5 random seeds to measure the effect of stochastic parameters. Each seed yields exactly one prediction per patient, giving 5 seed-specific predictions per patient. The folds were fixed across seeds to allow for fair comparison.

The baseline model used was a simple U-Net model (28). The U-Net was initialised with a depth of 4 and configured to accept 12 input channels and output 1 channel. All convolutions had kernel size 3 and a padding of 1. No other modifications were made to the model so as to allow for a fair comparison.

3.5 Evaluation

Models were evaluated with a set of industry standard statistical and clinical metrics. As a general measure of a model's performance, we used the dose score of the output distribution. To compute a dose score, a feasible dose mask was applied to the model predictions, after which voxel-wise MAE was calculated. This approach is preferred over unmasked MAE because voxels outside the patient volume are zero or near zero, which would artificially reduce the error. Consequently, an unmasked MAE value in gray (Gy) would not provide a fair comparison. The dose score is defined as

$$\text{Dose Score} = \frac{\sum_{i=1}^N M_i |D_i^{\text{pred}} - D_i^{\text{GT}}|}{\sum_{i=1}^N M_i}$$

(eq.3), where D_i^{pred} and D_i^{GT} denote the predicted and ground-truth dose at voxel i , $M_i \in [0,1]$ is the feasible dose mask, and N is the total number of voxels.

We used the Dice coefficient to assess the similarity between model predictions and ground truth dose distributions. This metric measures the structural conformity between both dose distributions in a given isodose surface. These are 3D volumes that connect all points receiving the same radiation dose, similar to an equipotential area. Dice coefficients closer to 1 indicate better similarity between distributions. Two metrics were used to evaluate accuracy within the PTVs. Underdosing was quantified using D_{95} , defined as the minimum dose received by 95% of the target volume. This metric reflects the minimum effective dose delivered to the PTV. D_{95} is preferred over $D_{\min}$ because it is less sensitive to local fluctuations in the dose distributions that may arise in

complex treatment plans. D_{95} was computed in both the model predictions and ground truth to assess discrepancies in target coverage.

Dose homogeneity was measured using the homogeneity index (HI). This measures the uniformity of dosage across the dose distribution. Deep learning models tend to follow one pattern when producing outputs, which would result in either PTV underdosing or OAR overdosing. Evaluating on uniformity allows both targets to be met, which is essential for multiple-objective optimization tasks such as dose distribution prediction. Several formulations exist for HI in the literature; in this work, we adopted the following definition of the Dose Homogeneity Index (HI):

$$HI = \frac{D_2 - D_{98}}{D_{50}} \quad (\text{eq.4}).$$

For this definition, HI values approaching 0 indicate greater dose uniformity within the target volume.

For OARs, accuracy was primarily assessed using $D_{\max}$, which is defined as the maximum dose delivered to the structure. This metric was selected because excessive local dose peaks can cause significant damage to healthy tissue, killing non-tumor cells. Alternative metrics such as D_2 can sometimes be used to mitigate sensitivity to outliers, but $D_{\max}$ provides a conservative estimate of worst-case exposure. Although clinical guidelines specify maximum dose constraints for many OARs, the ground truth dose distributions were used as the reference in this study. Future work may incorporate larger datasets and integrate established clinical constraints with deep learning-based predictions.

For testing, the predicted dose distributions from all 5 seeds per model were compiled together. All metrics were computed once per seed-specific prediction, then averaged across all 5 predictions per patient. The predicted dose distributions were not averaged together, as that would constitute a 5-model ensemble prediction rather than a single model's evaluation. For metrics involving structure masks, patients without data for a given structure were excluded from the metric calculation. The overall standard deviation for the metric was then calculated across the per-patient values.

95% confidence intervals were calculated using non-parametric percentile bootstrap, with 10,000 resamples of patients around the sample mean. We used Friedman testing along with a Nemenyi post-hoc, computed on a matrix of patient-level dose distributions across all 4 models. This was due to the non-parametric nature of our data, which restricted us from using ANOVA. We then used Cliff's delta to measure effect sizes: this refers to the probability for a given metric that a randomly selected value from one model would exceed a random value from another model, minus the reverse probability. This statistic ranges from -1 to 1, with the sign indicating the direction of effect. Effect sizes close to 0 indicate negligible relationships between models.

We also conducted a seed variability analysis to establish the stability of the model training for both the RWNNs and U-Net model. This included calculating the variance across all 5 seeds for each metric as a percentage of the value for each metric.

4. Results and Discussion

4.1 Graph and architectural search results

Following the graph and architectural search, the best-performing configuration from each

random graph family was selected for subsequent training and evaluation.

Table 2. Optimal graph parameters identified by Bayesian optimization

Model	n	m	p	k	Score
Erdős-Rényi	28	-	0.1908	-	0.1630
Barabási-Albert	30	8	-	-	0.1448
Watts-Strogatz	24	-	0.1654	10	0.1341

For both the Erdős–Rényi and Barabási–Albert models, Bayesian optimization selected relatively large values of n (28 and 30, respectively), approaching the maximum search bound of 32 (Table 2). This suggests that larger graph sizes may be advantageous for these two graph families. However, the optimal Watts–Strogatz model contained only 24 nodes, indicating that performance is not determined by graph size alone and may instead depend on interactions between the number of nodes and other topological properties.

Bayesian optimization also selected low values of p (< 0.2) for both graph families in which p was a tunable parameter. Because lower p values produce sparser connectivity, this finding suggests that relatively sparse graph structures may be advantageous. For Watts–Strogatz

models, Spearman correlation analysis identified a moderate association between lower p values and better performance ($\rho = 0.443$, $p < 0.05$). In contrast, no statistically significant association between p and performance was observed for Erdős–Rényi models, suggesting that the effect of connectivity may depend on interactions with other architectural parameters.

The optimal Barabási–Albert model, however, had relatively high connectivity, with $m = 8$ edges added per new node. No statistically significant association between m and model performance was detected. Taken together, these findings suggest that sparsity alone does not determine performance; rather, the effect of connectivity may depend on the underlying graph topology and its interaction with other architectural parameters.

Table 3. Optimal architectural parameters identified by Bayesian optimization

Model	Bottlenecked?	Kernel Size	Normalisation	Score
Erdős-Rényi	Yes	3x3x3	BatchNorm	0.1630
Barabási-Albert	No	3x3x3	BatchNorm	0.1448
Watts-Strogatz	No	5x5x5	BatchNorm	0.1341

For the Erdős–Rényi and Barabási–Albert models, which contained 28 and 30 nodes respectively, Bayesian optimization selected a 3

$\times 3 \times 3$ convolutional kernel (Table 3). Smaller kernels require fewer parameters than larger kernels for a given number of input and output

channels, which may be advantageous in larger graph architectures by limiting model complexity and reducing the potential for overfitting. However, this mechanism was not directly tested in the present study.

In contrast, the smaller 24-node Watts–Strogatz architecture favored a larger $5 \times 5 \times 5$ kernel. The larger receptive field provided by this

kernel may allow individual nodes to integrate information across a broader spatial region, potentially compensating for the smaller number of nodes. Together, these findings suggest a possible interaction between graph size and convolutional receptive field, although dedicated ablation studies would be required to determine whether this relationship is causal.

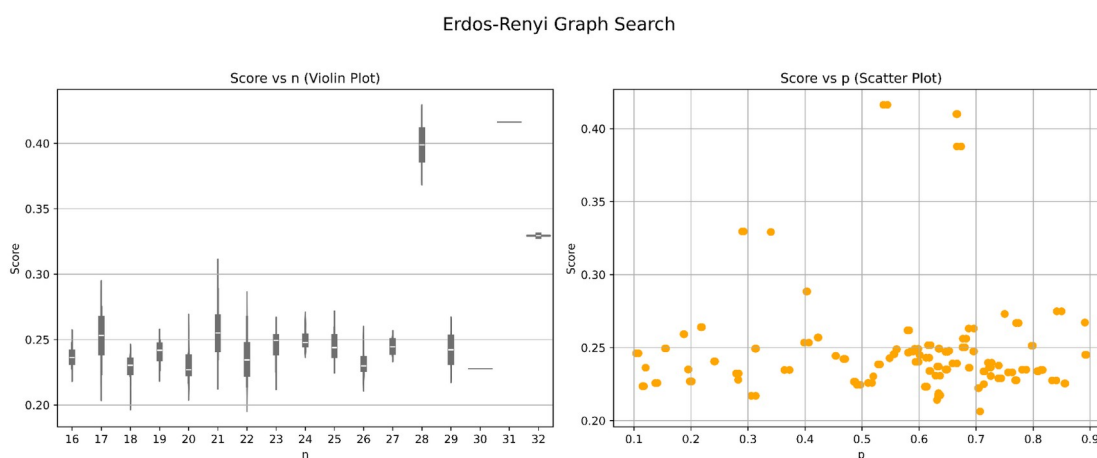


Figure 7. Parameter space searched for Erdős-Rényi graphs against final scores

As illustrated in Figure 7, no clear visual relationship emerged between individual graph parameters and model performance across the explored parameter space. Rather than indicating that these parameters have no effect, the absence of a simple trend may reflect interactions among multiple graph and architectural parameters, such that the influence of any single parameter depends on the values of the others. Future work could systematically vary individual parameters while holding other architectural features constant to isolate their independent effects on model performance.

4.2 Final Model

4.2.1 Training curves

The final models exhibited relatively stable convergence during training (Figure 8). Training loss decreased rapidly during the early epochs and showed comparatively little improvement thereafter. Validation loss followed a similar pattern, although modest oscillations were observed during later epochs. The absence of a sustained divergence between the training and validation curves provides no clear evidence of substantial overfitting during training, although generalization is more appropriately assessed using performance on the held-out test data. Overall, the convergence behavior suggests that most learning occurred during the earlier epochs and that additional training produced diminishing improvements. Future work could evaluate alternative learning-

rate schedules to determine whether convergence and final predictive performance can be further improved.

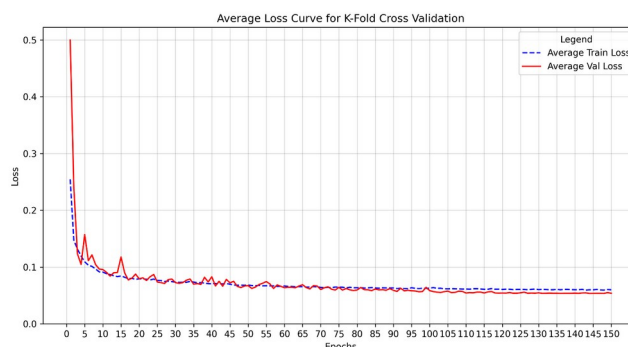


Figure 8. Training loss and validation loss curve of the Erdős-Rényi model

4.2.2 Clinical and technical metrics

Statistical testing using the Shapiro-Wilk test showed that the data were not normally distributed ($p < 0.05$). Since the normality assumption was violated, we used the Friedman test instead of ANOVA to measure significant statistical differences between pairwise observations for multiple models. The testing was conducted on all 340 patients as samples, confirming that all patients were used exactly once for testing. The Friedman test revealed significant differences between the three different graph models and the U-Net baseline model in all metrics shown in Table 4 ($p < 0.05$). Post-hoc Nemenyi tests confirmed a number of findings within each group of metrics.

For the Dose Score metric, the Barabási-Albert model achieved the lowest voxel-wise dose error among the four models. Post-hoc Nemenyi testing showed that both the Barabási-Albert and Watts-Strogatz models significantly improved upon the U-Net baseline ($p < 0.05$). Effect-size analysis supported these findings, with Cliff's delta values of 0.735 (95% CI [0.695, 0.794]) for Barabási-Albert and 0.716 (95% CI [0.679, 0.748]) for Watts-Strogatz,

indicating large differences in favor of the randomly wired architectures. Because the Dose Score measures voxel-wise error within the feasible dose volume, lower values indicate more accurate reproduction of the reference dose distribution. These findings therefore demonstrate improved dose-prediction accuracy for the Barabási-Albert and Watts-Strogatz architectures relative to the U-Net baseline, although the clinical consequences of these improvements would require validation in deliverable treatment plans.

A similar pattern was observed for the Dice coefficient, with all three random graph models achieving higher spatial agreement than the U-Net baseline. Post-hoc analysis showed statistically significant improvements for the Barabási-Albert and Erdős-Rényi models relative to the U-Net ($p < 0.05$), with effect sizes of 0.628 (95% CI [0.589, 0.663]) and 0.599 (95% CI [0.544, 0.643]), respectively. These effect sizes indicate substantial differences in favor of the randomly wired architectures. Higher Dice coefficients reflect greater spatial agreement between the predicted and ground-truth isodose volumes, suggesting that these

models more accurately reproduced the spatial dose prediction rather than the accuracy with characteristics of the reference dose which a treatment plan can ultimately be delivered. However, this metric evaluates distributions.

Table 4. Results of the U-Net baseline and the three random graph models. Mean value $\pm$ standard deviation is reported, as well as 95% bootstrapped confidence intervals. The Friedman test revealed significant differences ($p < 0.05$) in all metrics. All values are patient-level means, averaged across the 5 seeds per patient

Metric	U-Net	ER	BA	WS
Dose Score (Gy)	4.15 ± 1.38 [4.00, 4.30]	3.61 ± 1.19 [3.48, 3.74]	3.50 ± 1.21 [3.37, 3.63]	3.53 ± 1.22 [3.40, 3.66]
Dice _{50%}	0.897 ± 0.041 [0.893, 0.901]	0.921 ± 0.023 [0.919, 0.923]	0.921 ± 0.024 [0.918, 0.924]	0.919 ± 0.023 [0.917, 0.921]
PTV ₇₀ D ₉₅ (Gy)	49.12 ± 22.28 [46.75, 50.49]	52.96 ± 23.72 [50.43, 55.48]	53.06 ± 23.77 [50.53, 55.59]	53.01 ± 23.75 [50.49, 55.53]
PTV ₇₀ D ₉₅ Error (Gy)	-3.50 ± 3.78 [-3.90, -3.10]	0.33 ± 1.910 [0.13, 0.53]	0.44 ± 1.91 [0.24, 0.64]	0.39 ± 1.97 [0.18, 0.60]
PTV ₇₀ HI	0.184 ± 0.116 [0.172, 0.196]	0.083 ± 0.048 [0.078, 0.088]	0.079 ± 0.046 [0.074, 0.083]	0.084 ± 0.050 [0.079, 0.089]
Brainstem D _{max} (Gy)	63.68 ± 2.21 [63.45, 63.91]	63.25 ± 1.70 [63.07, 63.43]	63.20 ± 1.70 [63.02, 63.38]	63.58 ± 1.76 [63.39, 63.77]
Brainstem D _{max} Error (Gy)	-0.97 ± 1.95 [-1.18, -0.76]	-1.46 ± 2.04 [-1.68, -1.24]	-1.51 ± 2.03 [-1.73, -1.29]	-1.08 ± 2.03 [-1.30, -0.86]
Spinal Cord D _{max} (Gy)	64.54 ± 12.31 [63.23, 65.84]	60.38 ± 12.73 [59.03, 61.73]	60.35 ± 12.68 [59.00, 61.70]	60.54 ± 12.81 [59.18, 61.90]
Spinal Cord D _{max} Error (Gy)	5.13 ± 6.44 [4.45, 5.81]	0.98 ± 3.70 [0.59, 1.37]	0.95 ± 3.74 [0.55, 1.35]	1.14 ± 3.78 [0.74, 1.54]

As illustration: 4.15 ± 1.38 [4.00, 4.30] indicates a metric value of 4.15 and a standard deviation of 1.38, with a lower bound 95% confidence interval of 4.00 and an upper bound of 4.30

All three random graph models demonstrated improved PTV metrics relative to the U-Net baseline. The U-Net underpredicted the PTV₇₀ D₉₅ by ~ 3.5 Gy on average, whereas the random graph models produced mean D₉₅ errors much closer to zero. In a clinical setting, inadequate PTV coverage is important because insufficient dose to portions of the target may increase the risk of locoregional treatment failure (38). The random graph models showed significant improvements in PTV₇₀ D₉₅ relative to the U-Net ($p < 0.05$). Post-hoc analysis demonstrated that both the Barabási–Albert and Watts–Strogatz models performed significantly better than the U-Net baseline ($p < 0.05$), with effect sizes of 0.699 (95% CI [0.688, 0.710]) and 0.619 (95% CI [0.608, 0.629]), respectively. These findings indicate that the randomly wired architectures reproduced the reference target-dose coverage more accurately than the U-Net, although

whether this improvement translates into improved tumor control or reduced recurrence would require validation using deliverable treatment plans and clinical outcomes.

The direction of prediction error was also structure dependent. All models tended to underpredict brainstem D_{max} and overpredict spinal-cord D_{max}, whereas the pronounced PTV₇₀ underprediction was largely specific to the U-Net. This distinction suggests that some systematic errors may arise from characteristics of the prediction task, dataset, or optimization objective rather than network topology alone.

The random graph models also outperformed the U-Net baseline in HI, which indicates that the dose distributions generated are more uniform. The Barabási–Albert model showed a statistically significant improvement against the

U-Net with a large effect size ($p < 0.05$; $\delta = 0.847$ [95% CI 0.812, 0.864]).

OAR prediction accuracy varied across the four models. For the spinal cord, all three random graph models reproduced the reference D_{max} more accurately than the U-Net baseline. The Barabási–Albert and Erdős–Rényi models showed statistically significant improvements ($p < 0.05$), with moderate-to-large effect sizes of $\delta = 0.637$ (95% CI [0.607, 0.663]) and $\delta = 0.613$ (95% CI [0.579, 0.645]), respectively. Mean spinal-cord D_{max} error was approximately 1 Gy for each random graph model, compared with 5.13 Gy for the U-Net, indicating substantially closer agreement with the reference dose distributions. Although the predicted mean spinal-cord D_{max} was below 61 Gy for all three random graph models, these values should not be interpreted directly as demonstrating a reduced risk of radiation-induced myelopathy because the present study evaluates dose prediction rather than delivered treatment. Nevertheless, accurate prediction of spinal-cord dose is clinically important because excessive spinal-cord exposure is associated with radiation-induced myelopathy (39).

However, the U-Net baseline predicted brainstem D_{max} more accurately than all three random graph models, representing an important limitation of the RWNN architectures. Accurate brainstem dose prediction is clinically important because excessive radiation exposure can produce neurological complications ranging from fatigue and headache to cognitive impairment (40), with higher doses associated with increased risk of toxicity (41). The significant difference between the Barabási–Albert and Watts–Strogatz models ($p < 0.05$) further suggests that performance is

influenced by the choice of graph generator. Because each generator imposes distinct topological characteristics on the sampled networks, these results raise the possibility that graph-specific wiring priors influence dose-prediction accuracy. Further controlled comparisons of graph properties would be required to identify which topological features account for these differences.

Despite their distinct graph-generation mechanisms, the three RWNN architectures achieved remarkably similar performance across most metrics. This convergence suggests that the observed advantage may not depend on a single specific graph family, but may instead arise from architectural properties shared across randomly wired networks. Determining which shared topological features are responsible for this convergence represents an important direction for future work.

Overall, these results suggest that incorporating stochastic graph connectivity into the network architecture can improve radiotherapy dose prediction relative to a conventional U-Net, particularly in terms of voxel-wise dose accuracy and spatial agreement with the reference distributions. However, this advantage was not consistent across all clinically relevant structures: the U-Net predicted brainstem dose more accurately than the randomly wired models. These findings therefore suggest that randomly wired architectures offer promising improvements in overall dose prediction while also identifying structure-specific limitations that require further optimization.

This comparison is limited to only a simple baseline architecture and does not consider a broader range of modern dose distribution

prediction models. Furthermore, robustness to image perturbations, contour uncertainty, scanner variation, and institution shift were not validated as all data came from the same dataset. These changes are highly clinically relevant.

Therefore, further work comparing RWNNs

against other contemporary architectures, on larger datasets, is required to determine whether stochastic connectivity offers a general advantage in radiotherapy dose prediction.

4.2.3 Seed variability

Table 5. Within-patient variability across random seeds for each model and evaluation metric

Metric	U-Net	ER	BA	WS
Dose Score	2.34	2.66	2.09	1.34
Dice ₅₀ %	1.49	2.27	2.50	2.53
PTV ₇₀ D ₉₅	0.34	0.03	0.02	0.03
PTV ₇₀ D ₉₅ Error	12.07	3.77	2.95	4.03
PTV ₇₀ HI	4.41	2.75	1.85	2.84
Brainstem D _{max}	2.22	0.18	0.21	0.33
Brainstem D _{max} error	9.51	0.67	0.79	1.23
Spinal Cord D _{max}	1.38	0.56	0.39	0.46
Spinal Cord D _{max} error	5.56	7.07	4.92	5.59

The RWNN models demonstrated greater stability than the U-Net across six of the nine evaluated metrics (Table 5), indicating that the stochastic nature of their architectures did not generally result in greater seed-dependent variability. However, variability differed across both models and metrics, with no single RWNN architecture consistently exhibiting the greatest stability. Spinal cord D_{max} error showed comparatively high seed variability across all four models, suggesting that predictions for this metric are particularly sensitive to stochastic variation during training. Future work should therefore evaluate RWNN robustness across a larger number of random seeds and alternative training configurations to determine the reproducibility and sources of this variability.

4.2.4 DVH comparison

The models were also assessed using Dose Volume Histogram (DVH) curves. These

summarize the 3D dose distributions by showing the percentage of different target volumes that receive a certain dose. In this work, the solid lines show the ground truth DVH curves while the dashed lines show the DVH curves of model output. Increased conformity between the two curves corresponds to increased accuracy in a target volume.

Figure 9 provides a visual comparison of the dose-volume histograms (DVHs) generated by the U-Net and Barabási–Albert models against the ground-truth distributions. Both models reproduce the overall shape of the reference DVH curves, demonstrating that they capture the principal dose-volume relationships in the treatment plans. Below ~ 50 Gy, the predicted curves from both models closely follow the ground truth across most structures, indicating good agreement in the lower-dose range. Above ~ 50 Gy, however, the U-Net predictions begin

to deviate more noticeably from the reference curves, particularly in the higher-dose target regions. In contrast, the Barabási–Albert predictions remain closely aligned with the ground truth up to approximately 70 Gy. This

qualitative comparison is consistent with the quantitative PTV results and suggests that the Barabási–Albert architecture more accurately reproduces the high-dose characteristics of the reference treatment plans.

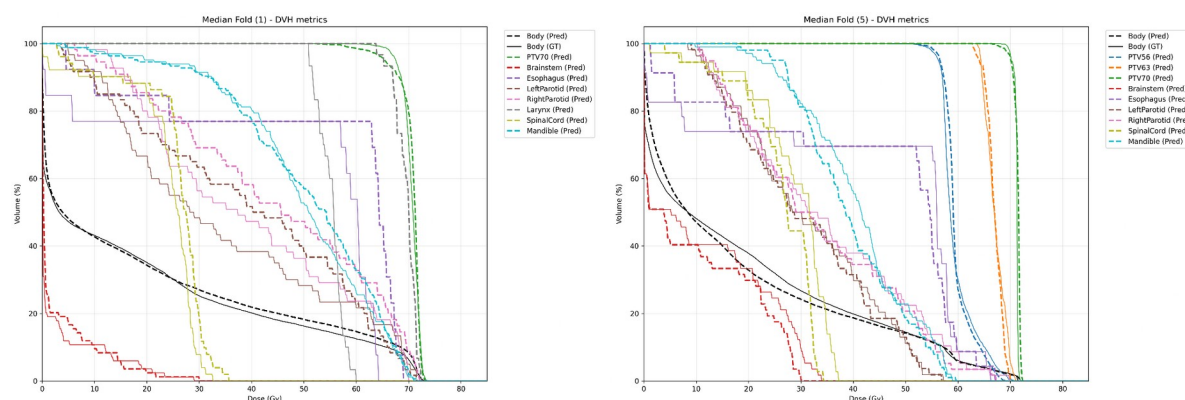


Figure 9. DVH curve comparison. The U-Net DVH curve is on the left, the Barabási–Albert DVH curve is on the right

4.2.5 Inference time and efficiency

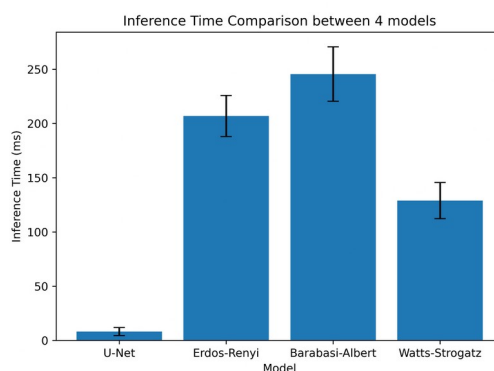


Figure 10. Inference time comparison between models

The U-Net baseline demonstrated substantially lower inference latency than all three random graph models (Figure 10), with inference times ~ an order of magnitude lower. Among the RWNN architectures, Watts–Strogatz was the most computationally efficient, achieving ~40–50% lower inference latency than the Erdős–

Rényi and Barabási–Albert models. Across the three RWNNs, higher node counts were associated with longer inference times, although the limited number of architectures evaluated prevents this relationship from being established independently of other topological differences. Overall, despite their improvements in several

dose-prediction metrics, the RWNN improve their suitability for time-sensitive architectures require further computational clinical applications. optimization to reduce inference latency and

Table 6. Model complexity and computational memory requirements

Model	Parameter Count	Storage (MB)	Activation memory (MB)
U-Net	2319457	4.42	398.73 $\pm$ 0.20
Barabási-Albert	422849	0.81	872.33 $\pm$ 16.31
Watts-Strogatz	76865	0.15	811.53 $\pm$ 11.97
Erdős-Rényi	395009	0.75	817.93 $\pm$ 59.35

Despite achieving the lowest inference time, the U-Net had the highest parameter count (Table 6). All three random graph models were substantially more parameter-efficient, with the Watts–Strogatz architecture containing nearly 30-fold fewer trainable parameters than the U-Net. Their smaller model storage sizes further reflect this reduction in parameter count, demonstrating that RWNNs can achieve competitive or improved dose-prediction performance using considerably fewer learned parameters. However, this parameter efficiency did not translate into lower runtime memory requirements: all three RWNNs required $\sim$ twice the activation memory of the U-Net. This may result from graph-based execution requiring multiple intermediate tensors to be retained across node-level operations and connections rather than from the number of trainable parameters itself. The current implementation was not specifically optimized for tensor reuse or memory efficiency. Future work should therefore investigate graph-execution and memory-management strategies that preserve the parameter efficiency of RWNNs while reducing their activation-memory requirements.

4.2.6 Node ablation study

We performed a node-ablation study to assess

the contribution of individual nodes to network performance. Two ablation strategies were evaluated. In the first, the selected node was removed while its incoming connections were rewired to its outgoing connections, thereby eliminating the node-level transformation and associated aggregation weights while preserving overall graph connectivity. In the second, the node and all of its incoming and outgoing connections were removed entirely, thereby altering both the computational operations and graph connectivity. For each patient, 10 nodes were randomly sampled and independently ablated using each strategy. The resulting changes in model performance were evaluated using a focused set of metrics, including Dose Score and Dice coefficient.

Surprisingly, removal of a single node produced no statistically significant change in performance across the evaluated metrics ($p > 0.05$) under either ablation strategy. This indicates that the RWNN architectures are robust to individual-node removal and suggests a degree of functional redundancy within the networks. Importantly, the absence of measurable performance degradation raises the possibility that the architectures could be further compressed without sacrificing dose-prediction

accuracy. Future work could therefore investigate systematic pruning or progressive node ablation to determine the minimum graph complexity required to maintain performance across the full dose distribution.

4.3 Perspectives: from random architecture search to mechanistic graph design

The present findings raise a broader question beyond whether randomly wired neural networks can predict radiotherapy dose distributions: why do some graph topologies perform better than others? In the current study, Bayesian optimization identifies high-performing architectures empirically, but does not establish which structural properties of those architectures are responsible for their performance. The differing behaviour of the Erdős-Rényi, Barabási-Albert, and Watts-Strogatz models suggests that network topology itself may represent an important determinant of predictive performance rather than merely a means of generating architectural diversity.

Each graph family provides a different form of information connectivity. Erdős-Rényi graphs distribute connections probabilistically throughout the network, potentially providing diverse pathways for feature propagation without imposing a strong structural hierarchy. Barabási-Albert graphs instead generate highly connected hub nodes, which may facilitate integration of information arriving through multiple computational pathways. Watts-Strogatz graphs combine local clustering with relatively short paths between distant regions of the graph, potentially allowing local and global information to be integrated efficiently. These characteristics provide plausible explanations for differences among the architectures, but the present study does not directly test these

mechanisms.

This distinction may be particularly relevant to three-dimensional radiotherapy dose prediction. Predicting dose distributions requires simultaneous representation of local anatomical relationships, such as dose gradients near tumour boundaries and adjacent organs at risk, and more global relationships among spatially separated targets and dose-limiting structures. A topology that efficiently combines local feature processing with long-range information transfer could therefore be advantageous. Conversely, excessive connectivity may introduce unnecessary parameters and computational cost, whereas insufficient connectivity may restrict information propagation. The optimal architecture may consequently reflect a balance among connectivity, information-transfer distance, and architectural redundancy rather than simply network size.

The node-ablation findings provide preliminary support for investigating this possibility. Removal of individual nodes did not significantly impair the evaluated performance metrics, suggesting that some computational pathways may be redundant. This raises the possibility that the relevant determinant of performance is not the presence of individual nodes but the preservation of particular routes through which information can propagate across the network. Future ablation experiments could therefore target nodes according to graph properties such as degree, betweenness centrality, or participation in short paths rather than removing nodes randomly. If removal of structurally important nodes produces greater deterioration than removal of peripheral nodes, this would provide direct evidence linking graph topology to network function.

Future studies could extend this approach by calculating graph-theoretic properties—including clustering coefficient, average shortest-path length, degree distribution, graph density, betweenness centrality, and measures of hub structure—for every sampled architecture and relating these properties quantitatively to validation and test performance. Such analysis could determine whether performance can be predicted from graph structure independently of graph family. Controlled experiments could then vary individual topological properties while holding network size and node operations approximately constant, allowing hypotheses concerning the mechanisms underlying performance to be tested directly.

Ultimately, the objective would not necessarily be to identify a single universally optimal random graph. Different prediction tasks may favour different information-flow structures. Instead, sufficiently consistent relationships between graph properties and performance could transform random architecture generation from a largely empirical search problem into a more principled design strategy. Rather than generating many networks and retaining whichever performs best, future models could be constructed or sampled preferentially from regions of graph space whose structural properties are known to be advantageous for a particular task. The present study therefore provides not only evidence that RWNNs can be applied to radiotherapy dose prediction, but also a foundation for investigating how neural-network topology itself influences predictive function.

5. Conclusion

This study demonstrates the feasibility of using

Randomly Wired Neural Networks (RWNNs) for three-dimensional IMRT dose prediction in oropharyngeal cancer. Rather than restricting network connectivity to conventional human-designed architectures, stochastic network generators based on Erdős-Rényi, Barabási-Albert, and Watts-Strogatz graphs were combined with Bayesian optimization to explore a broader architectural search space. The resulting RWNNs achieved performance comparable to or significantly better than the U-Net baseline across several voxel-wise, spatial, target-volume, and organ-at-risk metrics, including improvements in PTV_{70} , D_{95} , homogeneity, and spinal-cord D_{max} . However, the U-Net retained superior brainstem dose accuracy, demonstrating that the advantages of random connectivity are metric- and structure-dependent rather than universal.

The results also indicate that predictive performance does not necessarily require greater architectural complexity. The optimized RWNNs contained substantially fewer trainable parameters than the U-Net, with the Watts-Strogatz model using nearly 30-fold fewer parameters. Seed analysis indicated generally stable performance across stochastic training runs, while node-ablation experiments found no significant deterioration following removal of individual nodes. Together, these findings suggest that randomly generated architectures may contain considerable redundancy and that further simplification could yield even more compact networks without sacrificing predictive accuracy. These advantages must nevertheless be considered alongside the greater activation-memory requirements and inference times observed for the current RWNN implementations.

More broadly, this work suggests a progression from empirical architecture search toward understanding why particular network topologies are suited to particular prediction tasks. The three random-graph families impose fundamentally different connectivity structures, and their differing performance raises the possibility that properties such as clustering, path length, sparsity, degree distribution, and hub structure influence how anatomical information propagates through a network during dose prediction. Future studies should therefore relate these graph-theoretic properties directly to predictive performance and determine whether specific topological features can prospectively guide architecture design. Such work could ultimately replace increasingly large searches over arbitrary network configurations with more principled, task-

specific selection of graph topology.

The present findings remain limited by comparison with a single U-Net baseline and evaluation on a single dataset. External validation across institutions, imaging protocols, and patient populations, together with comparison against additional state-of-the-art dose-prediction architectures, will therefore be necessary before conclusions regarding clinical generalizability can be drawn. Nevertheless, the results establish stochastic graph-based architecture generation as a viable approach to radiotherapy dose prediction and provide a foundation for investigating the mechanistic relationship between neural-network topology, computational efficiency, and predictive performance.

6. References

1. Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 74(3), 229–263. <https://doi.org/10.3322/caac.21834>
2. Zhang, B., Mo, Z., Du, W., Wang, Y., Liu, L., & Wei, Y. (2015). Intensity-modulated radiation therapy versus 2D-RT or 3D-CRT for the treatment of nasopharyngeal carcinoma: A systematic review and meta-analysis. *Oral Oncology*, 51, 1041–1046. <https://doi.org/10.1016/j.oraloncology.2015.08.005>
3. Teh, B. S., Woo, S. Y., & Butler, E. B. (1999). Intensity modulated radiation therapy (IMRT): A new promising technology in radiation oncology. *The Oncologist*, 4(6), 433–442. <https://doi.org/10.1634/theoncologist.4-6-433>
4. Bentzen, S. M. (2005). Radiation therapy: Intensity modulated, image guided, biologically optimized and evidence based. *Radiotherapy and Oncology*, 77, 227–230. <https://doi.org/10.1016/j.radonc.2005.11.001>
5. Nutting, C. (2003). Intensity-modulated radiotherapy (IMRT): The most important advance in

radiotherapy since the linear accelerator? *British Journal of Radiology*, 76, 673.

<https://doi.org/10.1259/bjr/14151356>

6. Webb, S. (2001). *Intensity-modulated radiation therapy*. Institute of Physics Publishing.

7. Owosho, A. A., Tsai, C. J., Lee, R. S., Freymiller, H., Kadempour, A., Varthis, S., Sax, A. Z., Rosen, E. B., Yom, S. K., Randazzo, J., Drill, E., Riedel, E., Patel, S., Lee, N. Y., Huryn, J. M., & Estilo, C. L. (2017). The prevalence and risk factors associated with osteoradionecrosis of the jaw in oral and oropharyngeal cancer patients treated with intensity-modulated radiation therapy (IMRT): The Memorial Sloan Kettering Cancer Center experience. *Oral Oncology*, 64, 44–51.

<https://doi.org/10.1016/j.oraloncology.2016.11.015>

8. Bhide, S. A., Newbold, K. L., Harrington, K. J., & Nutting, C. M. (2012). Clinical evaluation of intensity-modulated radiotherapy for head and neck cancers. *British Journal of Radiology*, 85, 487–494. <https://doi.org/10.1259/bjr/85942136>

9. Ohnleiter, T., Antoni, D., Lefebvre, F., Truntzer, P., Schultz, P., Burgy, M., Borel, C., & Noël, G. (2018). Factors improving the outcome of patients re-irradiated with intensity-modulated radiotherapy (IMRT) for relapse or new head and neck cancer developed in irradiated areas. *Chinese Clinical Oncology*, 7(6), 60. <https://doi.org/10.21037/cco.2018.06.07>

10. International Commission on Radiation Units and Measurements. (2010). Prescribing, recording, and reporting photon-beam intensity-modulated radiation therapy (IMRT) (ICRU Report No. 83). <https://doi.org/10.1093/jicru/ndq001>

11. American Association of Medical Dosimetrists. (2023). Practice standards for the medical dosimetrist. <https://pubs.medicaldosimetry.org/pub/0D4F14CE-F3FE-DC5B-3329-9A7658229EEB>

12. Bortfeld, T., & Thieke, C. (2006). Optimization of treatment plans, inverse planning. In W. Schlegel, T. Bortfeld, & A.-L. Grosu (Eds.), *New technologies in radiation oncology* (pp. 207–220). Springer. https://doi.org/10.1007/3-540-29999-8_17

13. Sheng, Y., Zhang, J., Ge, Y., Li, X., Wang, W., Stephens, H., Yin, F.-F., Wu, Q., & Wu, Q. J. (2021). Artificial intelligence applications in intensity modulated radiation treatment planning: An overview. *Quantitative Imaging in Medicine and Surgery*, 11(12), 4859–4880. <https://doi.org/10.21037/qims-21-208>

14. Osman, A. F. I., & Maalej, N. M. (2021). Applications of machine and deep learning to patient-specific IMRT/VMAT quality assurance. *Journal of Applied Clinical Medical Physics*, 22, 20–36. <https://doi.org/10.1002/acm2.13375>

15. McPhail, S., Barclay, M. E., Swann, R., Johnson, S. A., Alvi, R., Barisic, A., Bucher, O., Creighton, N., Denny, C. A., Dewar, R. A., Donnelly, D. W., Dowden, J. J., Downie, L., Finn, N., Gavin, A. T., Habbous, S., Huws, D. W., Kumar, S. E., May, L., ... Lyratzopoulos, G. (2024). Use of radiotherapy in patients with oesophageal, stomach, colon, rectal, liver, pancreatic, lung, and ovarian cancer: An International Cancer Benchmarking Partnership (ICBP) population-based study. *The Lancet Oncology*, 25(3), 352–365. [https://doi.org/10.1016/S1470-2045\(24\)00032-9](https://doi.org/10.1016/S1470-2045(24)00032-9)
16. Su, W. W.-Y., Lee, Y.-H., Yen, A. M.-F., Chen, S. L.-S., Hsu, C.-Y., Chiu, S. Y.-H., Fann, J. C.-Y., Lee, Y.-C., Chiu, H.-M., Hsiao, S.-C., Hsu, T.-H., & Chen, H.-H. (2021). Impact of treatment delay on survival of oral/oropharyngeal cancers: Results of a nationwide screening program. *Head & Neck*, 43(2), 473–484. <https://doi.org/10.1002/hed.26504>
17. Jarrett, D., Stride, E., Vallis, K., & Gooding, M. J. (2019). Applications and limitations of machine learning in radiation oncology. *British Journal of Radiology*, 92, 20190001. <https://doi.org/10.1259/bjr.20190001>
18. Osman, A. F. I., & Tamam, N. M. (2022). Attention-aware 3D U-Net convolutional neural network for knowledge-based planning 3D dose distribution prediction of head-and-neck cancer. *Journal of Applied Clinical Medical Physics*, 23, e13630. <https://doi.org/10.1002/acm2.13630>
19. Zimmermann, L., Faustmann, E., Ramsel, C., Georg, D., & Heilemann, G. (2021). Technical note: Dose prediction for radiation therapy using feature-based losses and one cycle learning. *Medical Physics*, 48, 5562–5566. <https://doi.org/10.1002/mp.14774>
20. Xie, S., Kirillov, A., Girshick, R., & He, K. (2019). Exploring randomly wired neural networks for image recognition. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 1284–1293). IEEE. <https://doi.org/10.1109/ICCV.2019.00137>
21. Nguyen, D., Long, T., Jia, X., Lu, W., Gu, X., Iqbal, Z., & Jiang, S. (2019). A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning. *Scientific Reports*, 9, 1076. <https://doi.org/10.1038/s41598-018-37741-x>
22. Chen, X., Men, K., Li, Y., Yi, J., & Dai, J. (2019). A feasibility study on an automated method to generate patient-specific dose distributions for radiotherapy using deep learning. *Medical Physics*, 46, 56–64. <https://doi.org/10.1002/mp.13262>
23. Osman, A. F. I., Tamam, N. M., & Yousif, Y. A. M. (2023). A comparative study of deep learning-based knowledge-based planning methods for 3D dose distribution prediction of head and neck. *Journal of Applied Clinical Medical Physics*, 24, e14015.

<https://doi.org/10.1002/acm2.14015>

24. Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65, 386–408. <https://doi.org/10.1037/h0042519>
25. LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4), 541–551. <https://doi.org/10.1162/neco.1989.1.4.541>
26. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2014). Going deeper with convolutions [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.1409.4842>
27. He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.1512.03385>
28. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.), *Medical image computing and computer-assisted intervention—MICCAI 2015* (pp. 234–241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28
29. Zoph, B., & Le, Q. V. (2017). Neural architecture search with reinforcement learning [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.1611.01578>
30. Du, X., Chen, L., Liu, Z., Li, S., Liu, M., Yang, J., & Jin, L. (2020). Brain tumor image classification by randomly wired neural networks with a modified method. In *2020 IEEE 9th Data Driven Control and Learning Systems Conference (DDCLS)* (pp. 1141–1146). IEEE. <https://doi.org/10.1109/DDCLS49620.2020.9275242>
31. Babier, A., Zhang, B., Mahmood, R., Moore, K. L., Purdie, T. G., McNiven, A. L., & Chan, T. C. Y. (2021). OpenKBP: The open-access knowledge-based planning grand challenge and dataset. *Medical Physics*, 48(9), 5549–5561. <https://doi.org/10.1002/mp.14845>
32. Craft, D., Süß, P., & Bortfeld, T. (2007). The tradeoff between treatment plan quality and required number of monitor units in intensity-modulated radiotherapy. *International Journal of Radiation Oncology, Biology, Physics*, 67, 1596–1605. <https://doi.org/10.1016/j.ijrobp.2006.11.034>
33. Deasy, J. O., Blanco, A. I., & Clark, V. H. (2003). CERR: A computational environment for radiotherapy research. *Medical Physics*, 30, 979–985. <https://doi.org/10.1118/1.1568978>

34. Erdős, P., & Rényi, A. (1960). On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5, 17–61.
35. Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286, 509–512. <https://doi.org/10.1126/science.286.5439.509>
36. Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of ‘small-world’ networks. *Nature*, 393, 440–442. <https://doi.org/10.1038/30918>
37. Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2623–2631). Association for Computing Machinery. <https://doi.org/10.1145/3292500.3330701>
38. Eisbruch, A., Harris, J., Garden, A. S., Chao, C. K. S., Straube, W., Harari, P. M., Sanguineti, G., Jones, C. U., Bosch, W. R., & Ang, K. K. (2010). Multi-institutional trial of accelerated hypofractionated intensity-modulated radiation therapy for early-stage oropharyngeal cancer (RTOG 00-22). *International Journal of Radiation Oncology, Biology, Physics*, 76(5), 1333–1338. <https://doi.org/10.1016/j.ijrobp.2009.04.011>
39. Kirkpatrick, J. P., van der Kogel, A. J., & Schultheiss, T. E. (2010). Radiation dose–volume effects in the spinal cord. *International Journal of Radiation Oncology, Biology, Physics*, 76(3 Suppl.), S42–S49. <https://doi.org/10.1016/j.ijrobp.2009.04.095>
40. Alrasheed, A. S., Aleid, A. M., Alharbi, R. A., Alhodibi, M. H., Alhussain, A. A., Alessa, A. A., & Almalki, S. F. (2024). Brainstem toxicity following proton beam radiation therapy in pediatric brain tumors: A systematic review and meta-analysis. *Cancers*, 16(21), 3655. <https://doi.org/10.3390/cancers16213655>
41. Mayo, C., Yorke, E., & Merchant, T. E. (2010). Radiation associated brainstem injury. *International Journal of Radiation Oncology, Biology, Physics*, 76(3 Suppl.), S36–S41. <https://doi.org/10.1016/j.ijrobp.2009.08.078>