medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

Automatic radiotherapy treatment planning with deep functional reinforcement learning

Bin Liu[1], Yu Liu[1], Zhiqian Li[1], Jianghong Xiao *[2], Guosheng Yin[3], and Huazhen Lin[1]

[1] *Center of Statistical Research, School of Statistics, Southwestern University of Finance and Economics, Chengdu, China*

[2] *Radiotherapy Physics & Technology Center, Cancer Center, West China Hospital, Sichuan University, Chengdu, Sichuan, China*

[3] *Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong, China*

*corresponding author:xiaojh@scu.edu.cn

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

Abstract

Background and purpose: Intensity-modulated radiation therapy (IMRT) is a crucial radiotherapy technique, often formulated as an optimization problem. However, when the constraints are too tight to provide a feasible solution, human planners resort to relaxing the optimization parameters and re-evaluating until an acceptable solution is obtained. This process is laborious and time-consuming which has prompted attempts to automate radiotherapy through inverse planning studies using reinforcement learning. Unfortunately, these studies face two major limitations. Firstly, a separate sub-network must be designed for each organ, rendering them difficult to apply to patients with an inconsistent number of structures. Secondly, the low signal-to-noise inputs and discrete action space result in low training efficiency. To address these limitations, this study proposes a novel and effective model.

Methods: This study proposes an organ-sharing network called *F*unctional *a*utomatic *t*reatment *P*lann*I*ng *N*etwork (FatPIN), which contains a functional embedding layer to extract curve features of the dose-volume histogram (DVH). It outputs continuous actions that adjust the optimization parameters, thereby automating the radiotherapy planning process.

Results: Experiments were conducted on the cervical cancer dataset

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

and the results show that the FatPIN is feasible and effective in real-world radiotherapy. With automatic iteration, FatPIN gradually increased the PTV dose, while reducing the dose levels of the OARs. Specifically, at step 50, the D_{95} of the PTV reached 51.68 Gy, exceeding the clinical standard of 50.40 Gy, the V_{30}, V_{40} and V_{50} of all OARs were within clinical requirements.

Conclusion: We proposed FatPIN to implement automatic radiotherapy treatment planning. Experimental assessments conducted on cases of cervical cancer demonstrate significant improvements in patient metrics facilitated by FatPIN, thus confirming its practical applicability in clinical contexts.

1 Introduction

Cancer is a leading cause of death worldwide, accounting for approximately 10 million deaths in 2020 (nearly one in six deaths worldwide) [1]. Among the various treatment modalities available, radiation therapy plays a pivotal role in effectively managing cancer. It utilizes concentrated beams of high-energy radiation to target and control the planning target volume(s) (PTVs). However, a significant challenge arises due to the presence of organs-at-risk (OARs) in close proximity to the targeted area. The beams employed in radiation therapy not only aim to eradicate cancer cells but also inad-

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

vertently affect adjacent healthy tissues, which comprise the OARs. To address this challenge, inverse treatment planning techniques, such as intensity-modulated radiation therapy (IMRT), have emerged as crucial components of modern radiation therapy strategies.

In practice, achieving IMRT involves a complex optimization process as shown in eq. (12). The objective of the optimization typically contains several dose-level constrains to maximize the delivery of radiation of targets and minimize the hurts for OARs simultaneously. However, the formulation of the optimization problem of IMRT is typically based on a set of empirical parameters, such as relative importance weights and constraint levels for PTVs and OARs, which aim to accommodate various clinical considerations. These empirical parameters are initially initialized by human planners but can be adjusted as needed. To determine the optimal solution for a given set of parameters, a modern treatment planning system (TPS) is employed. However, it is not uncommon for the resulting solution to be inconsistent with the desired clinical considerations. In such cases, the planner must manually adjust these parameters and re-optimize the treatment plan using the TPS. This trial-and-error process is often repeated multiple times until a satisfactory result is achieved. Furthermore, the quality of the resulting treatment plan heavily relies on the experience and expertise of the planners. Consequently, there is a significant need to develop automatic methods that

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

can produce high-quality and efficient treatment plans, thereby reducing the reliance on manual intervention.

Over the past years, significant efforts have been made to develop automatic planning methods. A typical solution is to add an outer-loop optimization for parameter searching on top of the inner optimization. Several notable examples of this double-optimization approach can be found in [2, 3, 4]. More recently, attention has turned towards the potential of deep learning in automatic treatment planning [5]. Deep neural networks offer greater flexibility in the parameter searching compared to traditional methods. Particularly, the trial-and-error process of human planners can be effectively emulated using deep reinforcement learning techniques. Indeed, studies have demonstrated that deep reinforcement learning can be employed to adjust treatment plan parameters[6, 7, 8].

For instance, Shen et al.[9] propose a deep reinforcement learning-based virtual therapy planner network to model the decision-making behavior of human planners when adjusting treatment planning parameters (TPPs) for prostate cancer. However, existing methods suffer from several limitations, 1) Firstly, they typically design separate sub-networks for each organ [10, 7], making it challenging to generalize their models to different cases. This is because the number of organs involved in treatment planning can vary among patients, even for those with the same disease. 2) Secondly, these methods

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

neglect the functional characteristics of the dose-volume histogram (DVH) curve and often take it as an image, resulting in a significant reduction in the signal-to-noise ratio of the input. Consequently, the learning process becomes more challenging, and the number of organs that can be effectively planned is limited. For example, as discussed in [6, 10], these methods only consider the bladder and rectum as OARs, while clinical settings frequently involve adjusting more than a dozen organs. 3) Thirdly, existing methods model the process of dose adjustment as a classification problem, for example, setting five possible TPPs adjustment options {+2%, +1%, 0, -1%, -2%}, corresponds to increase or decrease TPP values by 2% or 1%, or keep them unchanged [9]. However, this classification approach severely restricts the search capability of the model. Therefore, there is substantial potential for advancing the generalization of multi-parameter TPP adjustment across multiple organs.

In this study, we introduce a deep reinforcement learning framework that takes into account the functional characteristics of the dose-volume histogram (DVH). We refer to this framework as the *F*unctional *a*utomatic *t*reatment *P*lann*I*ng *N*etwork (FatPIN). The primary objective of FatPIN is to train a model capable of manipulating the TPS and adjusting treatment planning parameters to generate high-quality treatment plans automatically. This automated approach aims to enhance the efficiency and accuracy of treatment

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

planning by leveraging the capabilities of deep reinforcement learning. The ultimate goal is to deliver treatment plans that effectively adhere to clinical standards while reducing the reliance on manual intervention.

In the realm of clinical practice, the condition of patients displays notable variability. Even among individuals diagnosed with the same type of cancer, the specific considerations for radiation therapy by medical practitioners may diverge due to unique factors such as tumor stage and organ size. The presence of these individualized requirements presents challenges when implementing a model predicated on a singular organ corresponding to a sub-network in a clinical context.

To mitigate this issue, FatPIN incorporates a neural network architecture tuned for organ-sharing. This innovative design facilitates the incorporation of DVHs from all tissue structures (organs) into the same network during the training phase. Subsequently, in testing, the DVHs of diverse structures for a particular patient can be directly input into the network to yield adjusted action prediction results. Essentially, this approach removes the constraint on the number of organs involved in treatment planning imposed by the model.

To bolster the learning efficiency of FatPIN, a functional decomposition layer is employed to initially extract features from the DVH curves. This functional mapping layer markedly enhances the signal-to-noise ratio

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

of the input data. Additionally, departing from the conventional method of modeling dose adjustment as a classification problem, we propose learning the distribution of dose adjustment. Specifically, we establish a connection between the output of FatPIN and the mean and variance of the dose adjustment distribution. Ultimately, parameterization techniques are utilized to generate continuous dose adjustments based on the acquired mean and variance. This strategic approach enhances the overall efficiency of model training, facilitating its application in real-world scenarios with heightened effectiveness and reliability.

2 Continuous Functional Treatment Parameters Adjustment Networks

Manual adjustment of dosage and weight to achieve a high-quality treatment plan is a cumbersome and time-consuming process in clinical radiotherapy. Therefore, automatic radiotherapy regimens are required with automatic dose and weight adjustments by planners. The overall framework will be introduced in next section.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

2.1 Overall Procedure

Figure 1 depicts the overall framework of our proposed method, *F*unctional *a*utomatic *t*reatment *P*lann*I*ng *N*etwork (FatPIN). It comprises three essential modules: a functional embedding layer, a continuous treatment planning neural network, and an environment for deep reinforcement learning (DRL). The overall process of automatic planning can be divided into the following steps:

1. The functional embedding layer accepts DVHs (state) $\boldsymbol{s}_t = (s_t(1), s_t(2), \cdots, s_t(200))$ at time t as input and outputs the functional embedding of DVHs $\boldsymbol{x}_t$.

2. The treatment planning neural network take functional embedding of DVH $\boldsymbol{x}_t$ as input and yield a compressed feature $\mathbf{z}_t$;

3. Estimates the mean $\boldsymbol{\mu}_t \in \mathbb{R}^M$ and the logarithmic form of variance $diag(\log \boldsymbol{\sigma}_t^2) \in \mathbb{R}^{M \times M}$ of the TPP action $\boldsymbol{a}_t \in \mathbb{R}^M$ for the M tissue structures with $\mathbf{z}_t$;

4. Randomly samples TPP actions: $\boldsymbol{a}_t \sim \mathcal{N}(\boldsymbol{\mu}_t, diag(\log \boldsymbol{\sigma}_t^2))$;

5. Updates the constraints of the optimization with $\boldsymbol{a}_t$ (see eq. (15)).

6. Solves the new optimization problem with TPS and yields a corresponding new state $\boldsymbol{s}_{t+1}$ (updated DVH) and calculates the corresponding

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

reward r_t;

7. Stores $(\boldsymbol{a}_t, \boldsymbol{s}_t, r_t, \boldsymbol{s}_{t+1})$ into sample set C;

8. Updates the parameters of FatPIN and repeats the above steps until convergence.

The following sections will introduce the details of the modules involved in the above process.

2.2 Functional Embedding Layer

2.2.1 Dose-volume Histogram

In radiotherapy planning, dose-volume histograms (DVH) $\boldsymbol{s}$ is a curve that depicts the distribution of the dose for each structure present in the treatment area, including the OAR and the target. Figure 2 presents a typical example of a DVH for an OAR and a target in a treatment plan. The horizontal axis in the DVH represents the dose level, while the vertical axis represents the fraction of the volume. For instance, the point P in the target DVH in Figure 2 indicates that at least 60% of the volume of the target voxels receive a dose level of 1.8 Gy or less. Ideally, the DVH in the target area should be vertical at the prescribed dose, indicating that all the voxels in PTV receive the prescribed dose, while the OAR should be vertical at a

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

relative dose of 0, indicating that it receives a dose of 0. Oncologists may be willing to sacrifice a certain portion of the OAR close to the target area to ensure optimal tumor control probability. Thus, OAR must be required to have at least a certain percentage of the dose below the specified level to reduce the risk of damaging the nearby healthy tissue.

2.2.2 Functional Embedding of DVH

Instead of directly feeding the DVH curves into the DRL model, FatPIN embeds them into finite-dimensional vectors. We assert that these embedding vectors are a more suitable input for subsequent adaptation to standard deep neural networks. The details are as following.

Let $s_t(u)$ denotes a single input DVH curve at timestamp t with horizontal axis u, the calculation of a hidden neuron in the neural network is

$$h = \sigma\left(\int_{\mathcal{U}} w(u)s_t(u)du + b\right) \tag{1}$$

where $w(u)$ are functional weights and b is the intercept term. We further decompose $w(u)$ using its basis representation,

$$w(u) = \sum_{k=1}^{K} c_k\phi_k(u) \tag{2}$$

where $\phi_k(u)$ represents the basis function of the functional weights $w(u)$. This basis function can be a Fourier basis or a spline basis, and we choose

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

K as the number of basis functions. In our approach, we select $\phi_k(u)$ to be the Fourier basis,

$$\begin{cases} \phi_0(u) = \frac{\sqrt{2}}{\sqrt{2}} \\ \qquad \vdots \\ \phi_{2k-1}(u) = \frac{\sin(\frac{2\pi k}{T}u)}{\sqrt{\frac{T}{2}}} \\ \phi_{2k}(u) = \frac{\cos(\frac{2\pi k}{T}u)}{\sqrt{\frac{T}{2}}} \end{cases}$$

and set $K = 5$ as shown in Figure 3.

Therefore, we can reformulate eq. (1) as,

$$\begin{aligned} h &= \sigma\left(\int_{\mathcal{U}} \sum_{k=1}^{K} c_k \phi_k(u) s_t(u) du + b\right) \\ &= \sigma\left(\sum_{k=1}^{K} c_k \int_{\mathcal{U}} \phi_k(u) s_t(u) du + b\right) \\ &= \sigma\left(\boldsymbol{c}^{\mathrm{T}} \boldsymbol{x}_t + b\right) \end{aligned} \tag{3}$$

where

$$\boldsymbol{x}_t = \begin{bmatrix} \int_{\mathcal{U}} \phi_1(u) s_t(u) du \\ \int_{\mathcal{U}} \phi_2(u) s_t(u) du \\ \vdots \\ \int_{\mathcal{U}} \phi_K(u) s_t(u) du \end{bmatrix}$$

is the functional embedding of DVH what we try to get. In practice, the integral $\int_{\mathcal{U}} \phi_k(u) s_t(u) du$ in eq. (3) can be approximated with numerical integration methods such as the composite Simpson's rule.

As for effectively handling the DVH curve $s_t(u)$, we employ a B-spline basis expansion approach. That is, we express each DVH as a linear combi-

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

nation of J B-spline bases, which can be represented as follows:

$$s_t(u) = \sum_{j=1}^{J} \alpha_j \varphi_j(u)$$

where $\alpha_j, j = 1, 2, \cdots, J$, is the corresponding coefficients of $s_t(u)$ with respect to the basis $\varphi_j(u)$, $\varphi_j(u)$ is the B-spline basis, J is the number of the B-spline basis, we set $J = 35$ as shown in fig. 4.

In our problem, we often encounter scenarios where multiple organs are involved, resulting in a set of input DVHs. Let's assume that M is the number of organs that need to be adjusted, then we will get functional embedding of DVHs $X_t \in R^{M \times K}$ in matrix form.

2.3 Treatment Planner Neural Networks

We use a similar framework of Deep Deterministic Policy Gradient [11] to implement the virtual treatment planner neural networks. The planner network can output actions $\boldsymbol{a} \in \mathbb{R}^M$ for M structures. Suppose a is i-th component of $\boldsymbol{a}$, which corresponds to the action of the i-th tissue structure. The process of estimate a with the virtual treatment planner neural networks is as follows.

In the context of deep reinforcement learning, our virtual treatment planner neural network acts as an agent that aims to interact with the environment. This network comprises two parts: the actor network $m(\boldsymbol{s}|\Theta)$ is

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

tasked with learning a policy for the agent and the critic network $Q(\boldsymbol{s}, a|\Phi)$ aims to estimate the action-value function $Q_\pi(\boldsymbol{s}_t, a_t)$.

To fully explore the environment, we add two linear layers to respectively output the mean μ_t and the logarithmic form of variance $\log \sigma_t^2$ of action a_t based on the original actor network output $\mathbf{z}_t$ as follows,

$$\mu_t := \mu_t(\mathbf{z}_t) = \mathbf{c}^\top \mathbf{z}_t + d \tag{4}$$

$$\log \sigma_t^2 := \log \sigma_t^2(\mathbf{z}_t) = \mathbf{e}^\top \mathbf{z}_t + f \tag{5}$$

where $\mathbf{c}, d, \mathbf{e}, f$ are parameters of the linear layers. Then, we use them to generate a new action a_t with

$$a_t = \mu_t + \sigma_t^2 \cdot \varepsilon_t, \tag{6}$$

where $\varepsilon_t \sim \mathcal{N}(0, 1)$. The action a_t is exactly what our new actor network outputs.

To make the critic network's action estimate as close as possible to the reward-based action estimate, the loss function of original critic network $Q(\boldsymbol{s}, a|\Phi)$ is

$$L(\Phi) = \mathbb{E}\left[\left(Q(\boldsymbol{s}_t, a_t|\Phi) - y_t\right)^2\right], \tag{7}$$

where

$$y_t = r_t + \gamma \cdot Q'(\boldsymbol{s}_{t+1}, m'(\boldsymbol{s}_{t+1}|\Theta')|\Phi') \tag{8}$$

is TD (Temporal Difference) target [12], γ is a discount parameter, $m'(\boldsymbol{s}|\Theta')$

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

and $Q'(\boldsymbol{s}, a|\Phi')$ are two target networks, which are copies of the learned networks $m(\boldsymbol{s}|\Theta)$ and $Q(\boldsymbol{s}, a|\Phi)$. We update the target networks in a soft way, that is, $\theta' \leftarrow \delta\theta + (1-\delta)\theta'$ with $\delta \ll 1$, where θ and θ' are the parameters corresponding to the learned network and the target network respectively.

Based on the preceding statement, we hope to make the optimization of the action a more precise, so it is necessary to control the magnitude of its variance σ_t^2. Therefore, we have added a penalty term for the variance to regularize the loss function as following,

$$L(\Phi) \triangleq \mathbb{E}\left[\left(Q(\boldsymbol{s}_t, a_t|\Phi) - y_t\right)^2\right] + \sigma_t^2. \tag{9}$$

2.4 Model Training

The DVH $\boldsymbol{s}_t(u)$ is a curve that can be represented by coordinates. In this work, we record the DVH using 100 2-D points. Therefore, the curve $\boldsymbol{s}_t(u)$ can be expressed as a 200-dimensional vector $\boldsymbol{s}_t(u) = (s_{t1}, s_{t2}, \cdots, s_{t200})$, where the first 100 dimensions correspond to the horizontal coordinates, and the subsequent 100 components represent the corresponding vertical ordinates.

At iteration step t, the functional embedding layer firstly accepts the $\boldsymbol{s}_t$ (DVH of a structure at time t) as input and outputs its functional embedding $\boldsymbol{x}_t$ (section 2.2). Next, we feed the functional embedding of DVH $\boldsymbol{x}_t$ for the

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

actor network $m(\boldsymbol{s}|\Theta)$ to predict the mean μ_t and the logarithmic variance $\log \sigma_t^2$ of the dose adjustment action a_t. Then, we sample a random action a_t using eq. (6).

Then the agent interacts with the environment by updating the optimization problem with a_t (as shown in eq. (15)). Upon solving the new optimization problem, we can calculate the corresponding reward r_t and obtain the updated state $\boldsymbol{s}_{t+1}$. The sequence of transitions $(\boldsymbol{s}_t, a_t, r_t, \boldsymbol{s}_{t+1})$ is stored in a replay buffer R. During model training, a batch of transitions is randomly selected from R to calculate $Q(\boldsymbol{s}_t, a_t|\Phi)$ using the critic network. Then the temporal difference (TD) target is computed as in eq. (8). With y_t, the loss can be calculated using eq. (9). That is, we can update parameters of the critic network $Q(\boldsymbol{s}, a|\Phi)$ with the TD algorithm [12]. For the actor network $m(\boldsymbol{s}|\Theta)$,we calculate the gradient of objective function $J(\Theta)$ as follows,

$$\begin{aligned} \nabla J(\Theta) &= \mathbb{E}\left[\nabla_\Theta Q(\boldsymbol{s}, a|\Phi)|_{\boldsymbol{s}=\boldsymbol{s}_t, a=a_t}\right] \\ &= \mathbb{E}\left[\nabla_a Q(\boldsymbol{s}, a|\Phi)|_{\boldsymbol{s}=\boldsymbol{s}_t, a=m(\boldsymbol{s}_t)} \nabla_\Theta m(\boldsymbol{s}|\Theta)|_{\boldsymbol{s}=\boldsymbol{s}_t}\right], \end{aligned} \tag{10}$$

the parameters of the actor network can then be updated using policy gradient ascent. Finally, soft target updates are used to update the parameters of the two target network. We set training stop conditions based on clinical expertise. For example, the algorithm stops updating when the D95 in the target area exceeds its prescribed dose, and when the V50 of the OAR

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

exceeds its prescribed dose.

2.5 Environment

The interactive environment for reinforcement learning, as depicted in Figure 1, receives adjustment actions $\boldsymbol{a}_t$ derived from the agent (virtual treatment planner network) as input. These actions are utilized to create a new optimization problem by modifying its original constraints. The new optimization problem is then solved by the TPS, resulting in a new DVH. Rewards are determined by comparing the new DVH with the previous one. Further details regarding the reward computation can be available in section 2.6.

2.5.1 Optimization Formulating

Suppose there are M structures that need to be planned, including the PTV and OARs. For our case, which involves cervical cancer, $M = 12$, and the OARs include the bladder, rectum, small intestine, bilateral femoral heads, and five ring structures.

Let the m-th organ be partitioned into N_m cubic volume elements or voxels of equal size [13](Typically greater than 2 mm in side length). Hence the dose for the m-th organ is represented by a vector $\mathbf{d}^m = [d_1^m, d_2^m, \cdots, d_{N_m}^m]^\top \in \mathbb{R}^{N_m}$, and $\sum_m N_m = N$, for $m = 1, \cdots, M$, where N is the total number of

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

voxels of the patient. Then the dose distribution for the M organs can be represented as $\mathbf{d} = [\mathbf{d}^1; \mathbf{d}^2; \cdots; \mathbf{d}^M] \in \mathbb{R}^N$.

In order to achieve the required dose depicted by $\mathbf{d}$, we will employ L beamlets at various angles to deliver radiation therapy to the M structures. Let $\Omega_{ij} \geq 0$ represent the dose deposition, which depicts the i-th voxel element by the j-th beamlet at unit beam intensity [13]. $\Omega \in \mathbb{R}^{N \times L}$ can be calculated by TPS, such as matRad [14]. It depends on the patient's personal structure parameters.

Then the goal of IMRT is to find, for each beamlet and according to the optimization goals of the respective model, a suitable nonnegative beamlet weight defining its radiation intensity. Specifically, the dose d_i of voxel i can be computed as follows,

$$d_i = \sum_{j}^{L} \Omega_{i,j} w_j \tag{11}$$

where w_j is the weight of beamlet j. The parameter w_j is the variable to be optimized. However, in this study, the optimization of w_j is indirectly achieved through the matRad optimization engine. In other words, we optimizes the dose vector $\mathbf{d}$ instead of w.

In the context of IMRT, the planning process involves solving a series of optimization problems that consist of well-defined objective functions and constraints. In our work, we adopt an overall optimization objective ex-

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

pressed as a weighted sum of various individual components as shown in eq. (12),

$$\begin{aligned} \min_{\mathbf{d}\in\mathbb{R}^N} \quad & \beta^{PTV} f^{PTV}(\mathbf{d}^1) + \sum_{m=2}^{M} \beta_m{}^{OAR} f^{OAR}(\mathbf{d}^m) \\ \text{subject to} \quad & c^{PTV}(\mathbf{d}^1) > 95\%, \\ & c^{OAR}(\mathbf{d}^m) < 30\%, m = 2, \cdots, M \\ & \mathbf{d} \succ 0 \end{aligned} \tag{12}$$

where $f^{PTV}(\mathbf{d}^1)$ is the objective function of PTV with a relative weighting (penalty) denoted as β^{PTV}, and $f^{OAR}(\mathbf{d}^m)$ denotes the objective function of m-th OAR with relative weighting β_m^{OAR}. The corresponding constraint $c^{PTV}(\mathbf{d}^1) > 95\%$ě indicates that 95% of the PTV voxels are required to receive a dose no lower than the prescription dose. The constraint $c^{OAR}(\mathbf{d}^m) < 30\%$ě indicates that no more than 30% OAR voxels are required to receive the prescription dose. The positivity constraint $\mathbf{d} \succ 0$ ensures that only positive radiation fluences are considered. In the implementation, we set $\beta^{PTV} = 80$ and $\beta_m^{OAR} = 40$.

In this work, $f^{PTV}(\mathbf{d}^1)$, $f^{OAR}(\mathbf{d}^m)$, and the corresponding constraints

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

are chosen as follows [14],

$$
\begin{aligned}
f^{PTV}(\mathbf{d}^1) &= \frac{1}{N_1}\sum_{i=1}^{N_1}(d_i^1 - d_p^1)^2 I(d_i^1 - d_p^1) I(d_i^1 - \tilde{d}^1) \\
f^{OAR}(\mathbf{d}^m) &= \frac{1}{N_m}\sum_{i=1}^{N_m}(d_p^m - d_i^m)^2 I(d_p^m - d_i^m) I(\tilde{d}^m - d_i^m) \\
c^{PTV}(\mathbf{d}^1) &= \frac{1}{N_1}\sum_{i=1}^{N_1} I(d_i^1 - d_p^1) \\
c^{OAR}(\mathbf{d}^m) &= \frac{1}{N_m}\sum_{i=1}^{N_m} I(d_p^m - d_i^m),
\end{aligned}
\tag{13}
$$

where N_1 is the number of PTV voxels, N_m is the number of voxels of m-th OAR, d_i^m represents the dose in voxel i of the m-th structure, $\tilde{d}^m$ is prescribed volume, and d_p^m is the prescribed dose of m-th organ. $I(x)$ is Heaviside function defined as,

$$
I(x) = \begin{cases} 0, & x < 0, \\ 1, & x \geq 0 \end{cases}
\tag{14}
$$

and $\tilde{d}^m, d_p^m, m = 1, \cdots, M$ are prescribed volume and prescribed dose for m-th structure [14]. Typically, $\tilde{d}^m$ and d_p^m are parameters determined by the doctor. Different values for $\tilde{d}^m$ and d_p^m correspond to distinct constraints for the optimization problems in eq. (12). In our study, we fix $\tilde{d}^m$ and adjust $\mathbf{d}_p = [d_p^1, d_p^2, \cdots, d_p^M]^T \in \mathbb{R}^M$ using the proposed method as follows,

$$
(\mathbf{d}_p)^{t+1} = (\mathbf{d}_p)^t \odot \boldsymbol{a}_t
\tag{15}
$$

where $\odot$ denotes the element-wise product. It suggests that action $\boldsymbol{a}_t$ updates $\mathbf{d}_p$ and consequently gives rise to a new optimization problem as we stated in

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

section 2.3. The updated objective function and constraint equations initiate a new round of optimization to obtain the results of this iteration as shown in Figure 1.

DVH is a criterion for evaluating the performance of the treatment plan. To drive the deep reinforcement learning process, we input an initial DVH into the ATPAN as a start. Instead of using the original DVH (an image that plotting the original DVH curves), we proposed to use its functional embedding as introduced in section 2.2. The input DVHs involve multiple organs. Suppose there are N organs require to adjust. A linear basis expansion is used with i basis functions defining each of these N DVHs. The embedding vectors of these DVHs are as defined follows,

2.6 Reward function

Based on the updated DVHs, we can compute the reward for the current iteration in the proposed model. The learning objective of our model is to maximize the dose delivered to the target while minimizing the dose received by the OARs. Consequently, we consider the change in the area between the DVH curves of the target A_{Target} and OARs A_{OARs} as a component of the reward. Additionally, the change in the D95 metric, which represents the dose received by 95% of the target volume, is also considered an important

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

indicator that captures the attention of clinicians. Therefore, we incorporate the change in D95 as another source of reward. The final reward can be expressed as follows,

$$r = \Delta D_{95} + \Delta(A_{Target} - A_{OARs}). \tag{16}$$

2.7 Cohort

In this study, a cohort of cervical cancer patients who underwent treatment with IMRT at our institution was selected. The treatment planning process involved defining the PTV as well as the organs at risk (OARs). The prescription dose for all patients was set at 50.40 Gy, delivered in fractions of 1.8 Gy each. The treatment plans for IMRT were optimized using the matRad software. Specifically, a configuration of equally spaced 5 coplanar photon beams was employed for all patients. These beam arrangements were then optimized using the IPOPT optimization algorithm within the matRad framework. The goal of the optimization process was to achieve an optimal dose distribution that effectively targeted the PTV while minimizing the dose to the surrounding OARs.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

3 Results

3.1 Visualizations of Planning

Figure 5(a) and Figure 5(b) visualize the process of parameter adjustment. In these representations, the dash line corresponds to the Planning Target Volume (PTV), while the solid line represents the five Organs at Risk (OARs): Bladder, Femoral Head R/L, Rectum, and Small Intestine. As shown in Figure 5(a), during the initial five steps, FatPIN trys to reduce the dose levels of the five OARs while concurrently increasing the PTV dose to 50.4 Gy. We observe that the Small Intestine registers the lowest dose level, while the Bladder exhibits the highest dose level among the five OARs. The initial dose of the Small Intestine is approximately 45 Gy, decreasing to 20.42 Gy by the end of the iteration, aligning with the prescribed dose for OARs.

Commencing from the initial adjustment, FatPIN initiates the process of adjusting the PTV dose level, gradually increasing it while simultaneously reducing the dose level of the OARs until convergence is achieved around step 50. It is noteworthy that FatPIN employs an alternating approach in which the dose levels of both the PTV and OARs are adjusted. Figure 5(b) illustrates the changes in the area under the DVH curves (DVH AUC) for both the target and the five OARs. The decreasing solid lines represent the area change of the five OARs. We observe that the DVH AUC of the

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

target continually increases, while the DVH AUC of the OARs decreases accordingly, which demonstrates the efficiency of the proposed model.

Figure 6 depicts the treatment planning process. The heat maps in the top panel represent the changes in dose distribution, while the corresponding DVH curves are depicted in the bottom panel at three stages: the initial step (left), step 50 (middle), and step 100 (right). The quality of the plans can be assessed based on the dose distributions of each structure, including the PTV and the five OARs. The bottom panel demonstrates the corresponding DVHs of each structure. For instance, the bottom left panel visualizes the initial DVHs in solid lines, while the top left represents the corresponding initial dose distribution, with segmentation curves delineated by the clinician. Changes in dose distribution in these areas help understand the dynamic planning process. The bottom middle and bottom right panels show the results of steps 50 and 100, respectively. At step 50, the DVH changes are compared by plotting both the initial step results (solid lines) and step 50 results (dashed lines). It is observed that the PTV and OAR curves shift at step 50, indicating a change in the dose distribution. At step 100, the dose level gap between PTV and OARs becomes more pronounced. The overall dose distribution in the top right panel maintains the expected distribution as labeled by the clinician. The result indicates that 95% volume of PTV would receive the prescription dose at stage three.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

3.2 Clinical Results

Our experiment covers 12 structures, including both targets and OARs. We have omitted the results of the 6 PTV rings and reported only the results of the other 6 most important structures. Table 1 illustrates the key DVH parameters (with standard deviation) that clinicians most interested in during planning process of the 6 structures including one PTV and five OARs. For PTV, we evaluate D_{98}/ D_{95} (the minimum absorbed dose that covers 98%/95% of the volume of the region of interest) and mean dose (D_{mean}). The D_{98} of the Planning Target Volume (PTV) increased from the initial 44.91 Gy to 51.12 Gy, and the D_{95} increased from the initial 45.24 Gy to 51.68 Gy. It is noteworthy that the final D_{95} already meets the clinician standard 50.40 Gy.

For the Organs at Risk (OARs), most DVH parameters, other than V_{50} at step 50, are significantly lower than the initial values. After 100 steps of adjustment, some DVH parameters still showed a certain improvement. Although the V_{50} of most organs has slightly increased, it remains within clinical requirements. For example, compared with the results at step 50, the initial V_{30} and V_{40} of the bladder, rectum, small intestine, and the right and left femoral heads decreased by 43.71% and 64.54%, 40.45% and 63.77%, 54.91% and 68.05%, 38.63% and 85.92%, 31.52% and 79.67%, respectively.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

Compared with the results of step 100, V_{30} and V_{40} in step 50 of the bladder and small intestine decreased by 21.23% and 17.91%, 13.81% and 90.25%, respectively.The V_{30} of rectum decreased by 5.51% during the adjustment from step 50 to step 100.

In comparison with the results at initial step, the V_{30} and V_{40}, and mean dose of all OARs at step 50 were reduced. More specifically, in comparison with the initial mean dose, the single mean dose of the bladder, rectum, small intestine, and the right and left femoral heads at step 50 were reduced by 9.65 Gy (-22.50%), 11.61 Gy (-26.91%), 5.29 Gy (-24.96%), 9.86 Gy (-24.96%), and 8.66 Gy (-21.75%), respectively.

4 Discussion

The proposed automatic treatment planning framework was developed to mimic the behavior of human planners. It addresses two key challenges in automatic treatment planning. The first one is we broke through the restriction that the number of organs to be planned must be the same for different patients. In radiotherapy planning, even for patients with the same type of cancer, the structures (organs) of radiotherapy to be considered will be different. However, traditional DRL-based models design an exclusive sub-network for each organ to be planned, which implies the model to be

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

consistent with the training patient in the test patient, severely reducing the scalability of the model [15, 10, 7]. To improve the scalability, we only use one network that is shared by all structures in FatPIN. During training, the functional features of DVH curves of different structures are fed into the network sequentially, and the output rewards of each organ are accumulated in one round of iterations before the gradient is passed back. In this way, the proposed model can be applied to real clinical scenarios where the number of organs planned for radiotherapy need not be the same for different patients. That is, personal planning protocols can be implemented for different patients.

Secondly, to enhance the efficiency of model learning, we introduced modifications to the input and output of FatPIN. On the input side, we devised a functional data embedding layer to extract features from the DVH curves. The existing models choose to process DVH curves directly with convolution [7, 10, 16], which ignore the functional data characteristics of DVH itself, making the signal-to-noise ratio of the processed features low and leading to low model learning efficiency. Hence,the proposed functional data embedding layer enables FatPIN to provide efficient computation for complex tumor patients. Regarding the treatment planning parameters to be output, we posited that they follows a normal distribution, rather then the traditional treatment of considering it as discrete classification problems. Consequently,

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

we link the final output features with the mean and variance of a normal distribution. Continuous parameter adjustment was then obtained through sampling techniques. It is noteworthy that we can obtain the adjustment values with high confidence by constraining the variance.

Currently, automated radiotherapy planning algorithms proposed in the existing literature are far from clinical needs and are generally considered to be solved under a simplified framework only. For example, some of the existing methods require additional human intervention to narrow the state-action space [10]. And in [10], the authors only evaluate the case of one PTV and two OARs. This is because as the number of organs to be considered increases, the process of solving the optimization problem becomes more complex or even infeasible. In contrast, our algorithm makes the computational process efficient due to the two aforementioned strategies, thus simplifying the solution process. For example, the results demonstrate in Figure 5, Figure 6 and Table 1 are the treatment planning for cervical carcinoma patients which involve 12 areas. In the example of Figure 6, the results of 12 organs are actually processed, and only the results of the most critical 6 of them are kept in order to facilitate the presentation of the DVH curves. In Table 1, we give the results of a patient with 12 structures to be considered during treatment planning.

There are several limitations of our work. First, we know that radiother-

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

apy planners are able to make planning across tumor categories. Although we have designed the organ-sharing TPPs network for patients from a same kind of cancer but with different number of structures, it is poorly trained in cases with different tumors. Second, the computational speed of our system is not fully satisfactory, as it takes 3 to 4 hours for a single test, which is comparable to the time taken by a human planner. This undermines the advantage of employing machine automation. The most time-consuming computational module is the interaction with the environment matRad [14], a MatLab-based TPS. During this interaction, a new optimization problem based on the action selection results predicted by DRL, needs to be addressed. Our model then transmits the constraint parameters of this new optimization problem to matRad to perform the next step of the optimization solution.

5 Conclusion

We present FatPIN, an organ-sharing network specifically devised for automatic radiotherapy treatment planning. FatPIN integrates a functional embedding layer to extract features from DVH curves and introduces a novel methodology for learning dose adjustment distribution, eschewing the reliance on predefined discrete adjustment steps. Experimental assessments carried out on cervical cancer cases showcase notable enhancements in pa-

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

tient metrics facilitated by the proposed network, thereby validating its direct utility in clinical contexts.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

References

[1] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 71(3):209–249, 2021.

[2] L Xing, J G Li, S Donaldson, Q T Le, and A L Boyer. Optimization of importance factors in inverse planning. Physics in Medicine & Biology, 44(10):2525–2536, sep 1999.

[3] Renzhi Lu, Richard J Radke, Laura Happersett, Jie Yang, Chen-Shou Chui, Ellen Yorke, and Andrew Jackson. Reduced-order parameter optimization for simplifying prostate IMRT planning. Physics in Medicine & Biology, 52(3):849–870, jan 2007.

[4] Huan Wang, Peng Dong, Hongcheng Liu, and Lei Xing. Development of an autonomous treatment planning strategy for radiation therapy with effective use of population-based prior data. Medical Physics, 44(2):389–396, 2017.

[5] Jiawei Fan, Jiazhou Wang, Zhi Chen, Chaosu Hu, Zhen Zhang, and Weigang Hu. Automatic treatment planning based on three-dimensional

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

dose distribution predicted from deep learning technique. Medical physics, 46(1):370–381, 2019.

[6] Chenyang Shen, Yesenia Gonzalez, Peter Klages, Nan Qin, Hyunuk Jung, Liyuan Chen, Dan Nguyen, Steve B Jiang, and Xun Jia. Intelligent inverse treatment planning via deep reinforcement learning, a proof-of-principle study in high dose-rate brachytherapy for cervical cancer. Physics in Medicine & Biology, 64(11):115013, 2019.

[7] Yingzi Liu, Chenyang Shen, Tonghe Wang, Jiahan Zhang, Xiaofeng Yang, Tian Liu, Shannon Kahn, Hui-Kuo Shu, and Zhen Tian. Automatic inverse treatment planning of gamma knife radiosurgery via deep reinforcement learning. Medical Physics, 49(5):2877–2889, 2022.

[8] William Thomas Hrinivich and Junghoon Lee. Artificial intelligence-based radiotherapy machine parameter optimization using reinforcement learning. Medical physics, 47(12):6140–6150, 2020.

[9] Chenyang Shen, Dan Nguyen, Liyuan Chen, Yesenia Gonzalez, Rafe McBeth, Nan Qin, Steve B. Jiang, and Xun Jia. Operating a treatment planning system using a deep-reinforcement learning-based virtual treatment planner for prostate cancer intensity-modulated radiation therapy treatment planning. Medical Physics, 47(6):2329–2336, 2020.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

[10] Chenyang Shen, Liyuan Chen, Yesenia Gonzalez, and Xun Jia. Improving efficiency of training a virtual treatment planner network via knowledge-guided deep reinforcement learning for intelligent automatic treatment planning of radiotherapy. Medical Physics, 48(4):1909–1920, 2021.

[11] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971, 2015.

[12] Richard S Sutton and Andrew G Barto. Reinforcement learning: An introduction. MIT press, 2018.

[13] Rembert Reemtsen and Markus Alber. Continuous optimization of beamlet intensities for intensity modulated photon and proton radiotherapy. Handbook of optimization in medicine, pages 1–40, 2009.

[14] Hans-Peter Wieser, Eduardo Cisternas, Niklas Wahl, Silke Ulrich, Alexander Stadler, Henning Mescher, Lucas-Raphael Müller, Thomas Klinge, Hubert Gabrys, Lucas Burigo, et al. Development of the open-source dose calculation and optimization toolkit matrad. Medical physics, 44(6):2556–2568, 2017.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

[15] Chenyang Shen, Dan Nguyen, Liyuan Chen, Yesenia Gonzalez, Rafe McBeth, Nan Qin, Steve B Jiang, and Xun Jia. Operating a treatment planning system using a deep-reinforcement learning-based virtual treatment planner for prostate cancer intensity-modulated radiation therapy treatment planning. Medical physics, 47(6):2329–2336, 2020.

[16] Chenyang Shen, Liyuan Chen, and Xun Jia. A hierarchical deep reinforcement learning framework for intelligent automatic treatment planning of prostate cancer intensity modulated radiation therapy. Physics in Medicine & Biology, 66(13):134002, jun 2021.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

Figure Caption

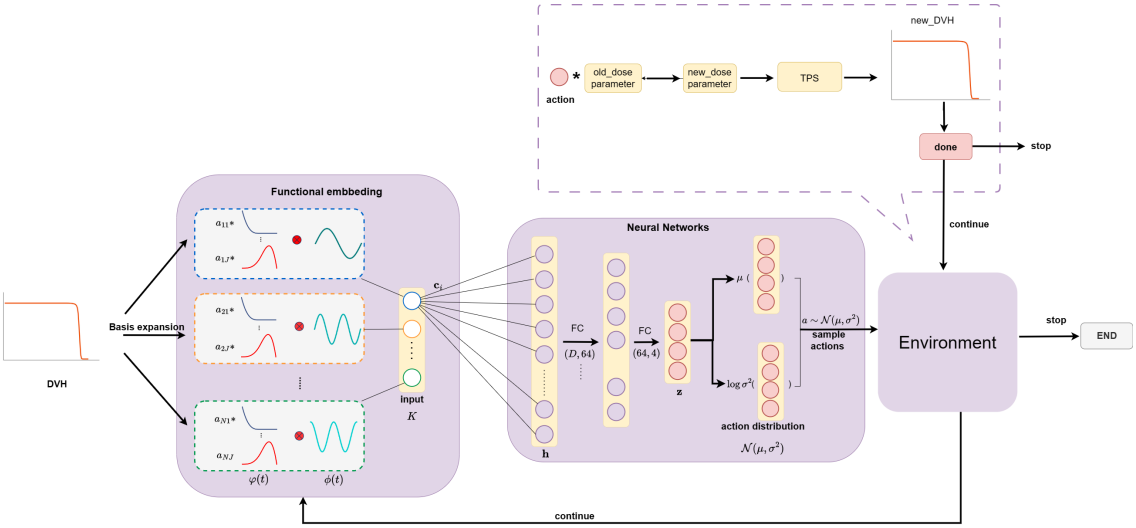


Figure 1: The overall framework of the proposed FatPIN.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

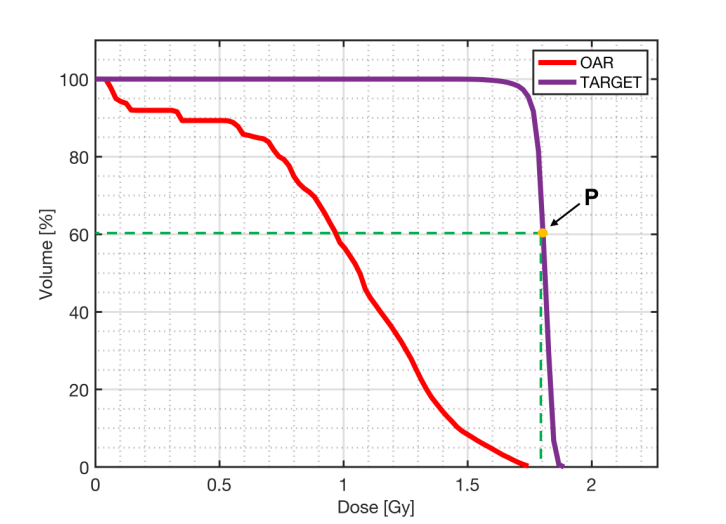


Figure 2: Example of DVH curves for an OAR and a target. The DVH observations $s_t(u) \in [0, 1]$ correspond to a functional relationship between the dose level u [Gy] (or relative dose level) of the horizontal axis and the fraction of volume $s_t(u)$ receiving that dose level of the vertical axis. This functional relationship provides a comprehensive representation of the radiation dose distribution within the target and the OARs.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

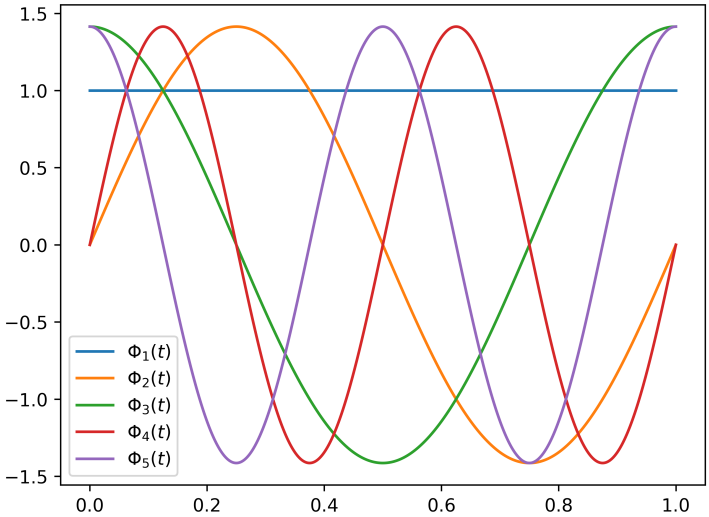


Figure 3: The 5 Fourier basis $\phi(u)$.

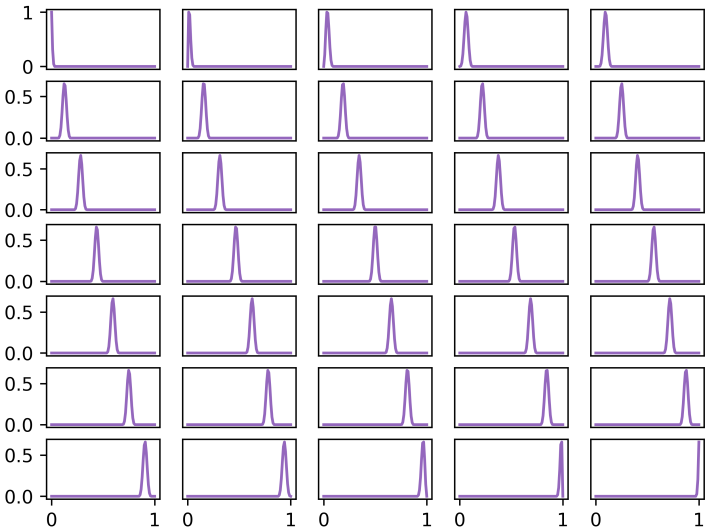


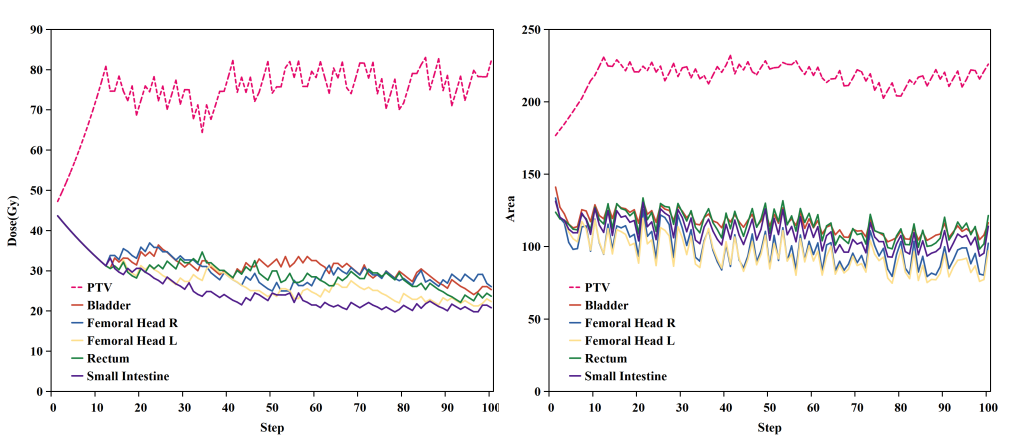
Figure 4: The 35 B-spline basis $\varphi(u)$.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

Table 1: Summary of Interested DVH Parameters Changes (with standard deviation) to the PTV, and OARs.

ROI	Parameters	Init	Step50	Step100
PTV	D_{98}(Gy)	44.91±0.65	51.12±0.31	50.42±0.43
	D_{95}(Gy)	45.24±0.98	51.68±0.18	51.35±0.21
	D_{2}(Gy)	53.48±0.32	59.12±0.30	61.32±0.43
	D_{mean}(Gy)	46.08±0.37	56.34±0.62	57.71±0.15
Bladder	V_{30}(%)	99.73±0.11	56.14±2.32	44.22±2.32
	V_{40}(%)	60.44±3.91	21.43±7.51	17.59±9.93
	V_{50}(%)	0	11.23±1.39	11.32±2.57
	D_{mean}(Gy)	42.89±1.23	33.24±1.42	30.80±1.72
Rectum	V_{30}(%)	99.90±0.10	59.49±3.26	56.21±0.35
	V_{40}(%)	78.23±8.14	28.34±3.71	29.74±0.82
	V_{50}(%)	0	9.01±8.42	10.12±7.60
	D_{mean}(Gy)	43.14±1.26	31.53±1.37	33.48±0.84
Small intestine	V_{30}(%)	45.91±7.38	20.70±2.14	17.84±6.68
	V_{40}(%)	17.34±6.01	5.54±2.90	0.54±0.11
	V_{50}(%)	0	2.66±0.70	2.32±0.41
	D_{mean}(Gy)	21.19±0.87	15.90±1.27	15.32±1.03
Femoral Head R	V_{30}(%)	97.10±0.41	59.59±2.32	60.62±0.31
	V_{40}(%)	40.69±7.54	5.73±2.43	19.96±0.23
	V_{50}(%)	0	0	0.43±0.09
	D_{mean}(Gy)	39.51±0.71	29.65±0.73	30.57±0.73
Femoral Head L	V_{30}(%)	90.57±5.42	62.02±0.88	63.31±7.60
	V_{40}(%)	46.34±6.35	9.42±8.50	25.65±6.31
	V_{50}(%)	0	0	0.21±0.13
	D_{mean}(Gy)	39.82±1.39	31.16±1.33	32.67±1.43

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309060; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .



(a) dose adjustment (b) DVH AUC of target and OARs

Figure 5: The process of parameter adjustment of FatPIN.

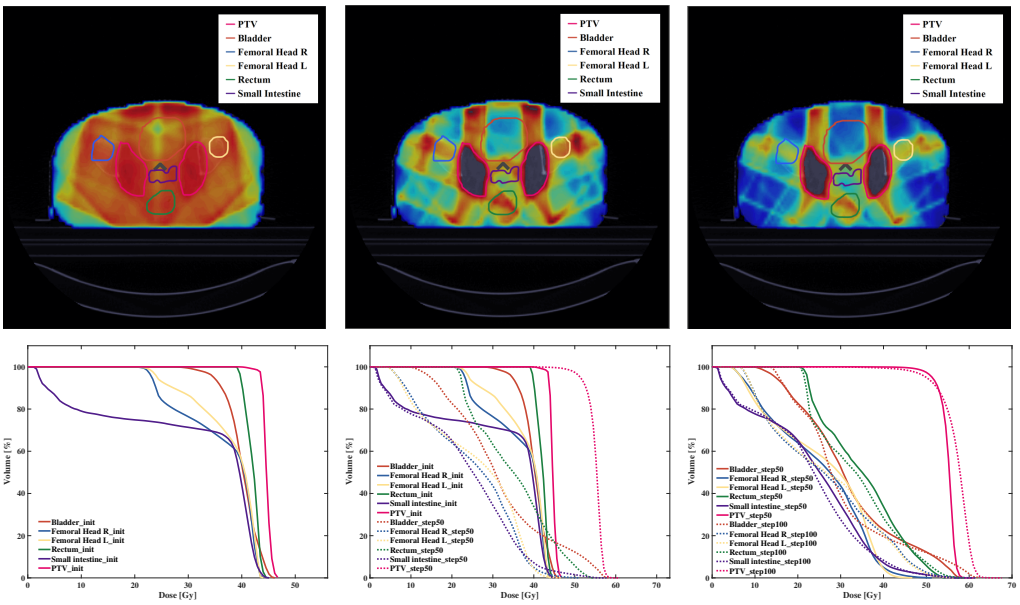


Figure 6: Variations of DVHs and dose distributions.