Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

*Phys Med Biol*. Author manuscript; available in PMC 2021 December 23.

Published in final edited form as:

*Phys Med Biol*. ; 66(13): . doi:10.1088/1361-6560/ac09a2.

# A Hierarchical Deep Reinforcement Learning Framework for Intelligent Automatic Treatment Planning of Prostate Cancer Intensity Modulated Radiation Therapy

**Chenyang Shen**[1,2], **Liyuan Chen**[1], **Xun Jia**[1,2]

[1]Medical Artificial Intelligence and Automation (MAIA) Laboratory, Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

[2]innovative Technology Of Radiotherapy Computation and Hardware (iTORCH) Laboratory, Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

## Abstract

**Purpose:** We have previously proposed an intelligent automatic treatment planning (IATP) framework that builds a virtual treatment planner network (VTPN) to operate a treatment planning system (TPS) to generate high-quality radiation therapy (RT) treatment plans. While the potential of IATP in automating RT treatment planning has been demonstrated, its poor scalability caused by an almost linear growth of network size with the number of treatment planning parameters (TPPs) is a bottleneck, preventing its application in complicate, but clinically relevant treatment planning problems. The decision-making behavior of the trained network is hard to understand. Motivated by the decision-making process of a human planner, this study proposes a hierarchical IATP framework.

**Methods and Materials:** The hierarchical VTPN (HieVTPN) consists of three networks, i.e. Structure-Net, Parameter-Net, and Action-Net. When interacting with a TPS, the networks are employed in a sequential order in each step to decide the structure to adjust, the TPP to adjust for the selected structure, and the specific adjustment manner for the parameter, respectively. We developed an end-to-end hierarchical deep reinforcement learning (DRL) scheme to simultaneously train the three networks. We then evaluated the effectiveness of the proposed framework in the treatment planning problems for prostate cancer intensity modulated RT (IMRT) and stereotactic body RT (SBRT). We benchmarked the performance of our approach by comparing plans made by VTPN of a parallel architecture, and the human plans submitted for competition in the 2016 American Association of Medical Dosimetrist (AAMD)/Radiosurgery Society (RSS) Plan Study. We analyzed scalability of the network size with respect to the number of TPPs. Numerical experiments were also performed to understand the rationale of the decision-making behaviors of the trained HieVTPN.

**Results:** Both HieVTPNs for prostate IMRT and SBRT were trained successfully using 10 training patient cases and 5 validation cases. For IMRT, HieVTPN was able to generate high-quality plans for 59 testing patient cases that were not included in training process, achieving an

Chenyang.Shen@UTSouthwestern.edu .

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

average plan score of 8.62 (±0.83), with 9 being the maximal score. The score was comparable to that of the VTPN, 8.45 (±0.48). For SBRT planning, HieVTPN achieved an average plan score of 139.07 on five testing patient cases compared to the score of 132.21 averaged over the human plans summited for competition in AAMD/RSS plan study. Different from VTPN with network size linearly scaling with the number of TPPs, the network size of HieVTPN is almost independent of the number of TPPs. It was also observed that the decision-making behaviors of HieVTPN were understandable and generally agreed with the human experience.

**Conclusions:** With the scalability and explainability, the hierarchical IATP framework is more favorable than the previous framework in terms of handling treatment planning problems involving a large number of TPPs.

## 1. Introduction

Treatment planning is one of the most critical steps for modern cancer radiation therapy (RT) (Oelfke and Bortfeld, 2001; Webb, 2003). It is often performed by an experience human planner using a treatment planning system (TPS). For a given patient case, the human planner defines a set of treatment planning parameters (TPPs), based on which the TPS solves an optimization problem to generate a plan. Typical TPPs include weighting factors, dose limits, and volume constraints defined for treatment targets and organs at risk (OARs), and often other structures created for optimization purpose. The values of these parameters critically affect the resulting plan quality. Due to the patient-specific nature of the treatment planning process, the optimal TPP values vary from patient to patient, and the human planner often needs to repeatedly interact with TPS to adjust TPPs. After the adjustment is applied, the TPS runs the plan optimization step again to update the plan. Such an interaction between the human planner and the TPS is continued, until a satisfactory plan is generated. The whole planning process is usually time consuming and labor intensive, and the resulting plan quality is affected by a number of human factors, such as experience level of planner and available time for planning (Das *et al.*, 2008; Nelms *et al.*, 2012). Hence, fully automated treatment planning approaches to automatically generate patient-specific high-quality plans are strongly desired.

To date, extensive research efforts have been devoted to solving this problem, and a number of methods have been successfully developed, such as greedy approaches (Xing *et al.*, 1999; Lu *et al.*, 2007; Wu and Zhu, 2001; Wang *et al.*, 2017), heuristic approaches (Yang and Xing, 2004; Wahl *et al.*, 2016; Yan and Yin, 2008), fuzzy inference (Yan *et al.*, 2003a; Yan *et al.*, 2003b; Holdsworth *et al.*, 2012; Holdsworth *et al.*, 2010), and statistics-based methods (Lee *et al.*, 2013; Boutilier *et al.*, 2015; Chan *et al.*, 2014). Recently, deep learning based methods (Shen *et al.*, 2020c) have been widely applied in the context of automatic treatment planning (Nguyen *et al.*, 2020; Nguyen *et al.*, 2019; Shen *et al.*, 2019; Fan *et al.*, 2019; Mahmood *et al.*, 2018; Shen *et al.*, 2020b; Shen *et al.*, 2020a; Zhang *et al.*, 2020; Hrinivich and Lee; Li *et al.*, 2020). In particular, an intelligent automatic treatment planning framework (IATP) (Shen *et al.*, 2020b; Shen *et al.*, 2020a; Shen *et al.*, 2019) has been proposed. Within this framework, a virtual treatment planner network (VTPN) was built to model the intelligent human behaviors of interacting with the TPS in the treatment planning process. Trained via an end-to-end deep reinforcement learning (DRL) technique, the VTPN was able to operate

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

a TPS in lieu of a human planner to generate high-quality plans. Specifically, similar to a human planner, the VTPN takes a state, i.e. the dose-to-volume histogram (DVH) of a plan generated by the optimization engine with given TPPs, as input and determines the action of adjusting TPPs to improve plan quality. Such a process is repeated, until the plan quality reaches a satisfactory level. The feasibility of IATP has been demonstrated in exemplary problems of high-dose-rate (HDR) brachytherapy for cervical cancer (Shen *et al.*, 2019) and intensity modulated RT (IMRT) for prostate cancer (Shen *et al.*, 2020b).

Despite its success, there are two issues in the IATP framework. The first one is the scalability of VTPN. Under the current formulation, VTPN consists of a number of subnetworks with each being responsible for one TPP in the TPS. Such a VTPN architecture is feasible, when operating a TPS for relatively simple treatment planning problems with only a small number (e.g. ~5) of adjustable TPPs, such as those in the in-house developed TPSs in previous studies (Shen *et al.*, 2019; Shen *et al.*, 2020b). However, to handle a more complicated treatment planning problem in a more sophisticated TPS, such as the planning for head and neck cancer volumetric modulated arc therapy (VMAT) using Varian Eclipse TPS (Varian Medical System, Palo Alto, CA), tens of TPPs may be involved, yielding a ~10 times of growth in the number of subnetworks in VTPN. Training such a VTPN may become infeasible due to the huge computation and memory requirement. The second issue is explainability. While it was observed that the trained network was able to decide which TPPs to adjust to improve plan quality, the reasons behind this behavior were still not clear. Understanding the reasons would help us to confidently use the developed system in clinical applications(Jia *et al.*, 2020).

Motivated by the hierarchical decision-making behavior of human planner in the treatment planning process, in this paper, we explored the feasibility of building a hierarchical VTPN (HieVTPN) to address the poor scalability and explainability of VTPN model. We developed a novel end-to-end hierarchical DRL (HieDRL) scheme to jointly train the three networks. We first focused on a proof-of-principle planning problem for seven-beam prostate cancer IMRT planning with the conventional fractionation scheme as the testbed allowing comparison between HieVTPN and the VTPN to benchmark the performance. We will also analyze the behaviors of the trained network to interpret the decision-making capability. Then, we applied the HieVTPN to a more realistic planning task for prostate cancer stereotactic body radiation therapy (SBRT), i.e. a similar seven-beam IMRT with a hypofractionation scheme and higher dose in each fraction compared to the conventional fractionation, on an Eclipse-like in-house TPS to demonstrate its feasibility to handle complex clinical treatment planning tasks. To simplify the notation, in the rest of the paper IMRT will be used to refer the conventional fractionation scheme, while SBRT will be used for hypofractionation unless otherwise specified.

## 2. Methods and Materials

### 2.1 Problem overview

IATP framework (Shen *et al.*, 2020b; Shen *et al.*, 2020a; Shen *et al.*, 2019) generally follows the typical inverse treatment planning workflow. In contrast to the conventional planning

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

process where a human planner is needed to interact with a TPS, IATP builds and applies a VTPN, an intelligent computer agent, to operate the TPS automatically.

In this paper, we would like to investigate the feasibility of training HieVTPN with a novel hierarchical architecture to perform automatic treatment planning in the IATP framework. We considered two example treatment planning problems, i.e. a proof-of-principle prostate IMRT planning and a realistic prostate SBRT planning, as testbeds. Due to the practical challenges interfacing the IATP framework with a commercial TPS efficiently, we developed two in-house TPSs for training and evaluation purposes. For the proof-of-principle prostate IMRT planning task, we considered exactly the same optimization engine used in (Shen *et al.*, 2019; Shen *et al.*, 2020b) to benchmark the performance of HieVTPN since it allows direct comparison against the previously developed VTPN of a parallel architecture. This optimization problem formulated for the prostate IMRT planning can be explicitly give as follows

$$\begin{aligned} \min_{x \geq 0} & \frac{1}{2}\|Mx - d_p\|_-^2 + \frac{\lambda}{2}\|Mx - d_p\|_+^2 + \frac{\lambda_{bla}}{2}\|M_{bla}x - \tau_{bla}d_p\|_+^2 + \frac{\lambda_{rec}}{2}\left\|M_{rec}x - \tau_{rec}d_p\right\|_+^2, \\ \text{s.t.} & \; D_{95\%}(Mx) = d_p \end{aligned} \tag{1}$$

$x \geq 0$ denotes the beam fluence map to be determined, while $M$, $M_{bla}$, and $M_{rec}$ indicates the dose deposition matrix for PTV, bladder, and rectum, respectively. $d_p$ denotes prescription dose and $\tau_{bla}$ and $\tau_{rec}$ are threshold values controlling the dose limits to bladder and rectum. $\|\cdot\|_+$ and $\|\cdot\|_-$ are $l_2$ norms computed for only positive and negative elements to penalize on the overdose and under-dose, respectively. For a clear presentation, we name them as upper objective and lower objective respectively in the rest of the paper. Note that the treatment targets, such as the PTV, may have both upper and lower objectives in planning, while the OARs typically have only the upper objectives as the goal is to spare the OARs as much as possible. $\lambda$, $\lambda_{bla}$, and $\lambda_{rec}$ are the weighting factors reflecting the priorities of the dosimetric structures of interest. The optimization problem in Eq. (1) consists of four planning objectives (two for PTV, one for bladder and one for rectum). Under such a formulation, this TPS involves five TPPs, i.e. the weighting factor $\lambda$, $\lambda_{bla}$, and $\lambda_{rec}$ to penalize overdose to PTV, bladder, and rectum, and the dose limits $\tau_{bla}$ and $\tau_{rec}$ to adjust dose to bladder and rectum.

For the realistic prostate SBRT planning problem, we built an Eclipse-like TPS for the planning purpose. The inverse planning optimization implemented in this in-house TPS exactly followed that of the Eclipse TPS, based on the detailed documentation of plan optimization method in (Eclipse, 2015). More specifically, we incorporated a DVH-based objective function which can be explicitly given as

$$\min_{x \geq 0} \sum_i \frac{\lambda_i}{2}\left\|[M_i x]_{V_i} - \tau_i d_p\right\|_{+/-}^2 \tag{2}$$

$M_i$ denotes the corresponding dose deposition matrix of the $i$-th planning objective involved in planning. In addition to the weighting factors $\lambda_i$ and dose limits $\tau_i$, this optimization

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

problem also involves volumes $V_i$ as tunable TPPs. The impact of these TPPs to the final plan quality is the same as they are in the Eclipse TPS. Compared to the proof-of-principle IMRT case, this planning problem is more realistic and clinically relevant. It involves more planning objectives, and thereby more TPPs to adjust in planning process, and hence is more complicated from the planning point of view. Table 1 lists all the planning structures and objectives considered. We included in total 13 commonly used planning structures. Most of these structures have only one planning objective set up for plan optimization purpose except for the PTV (one upper and one lower objectives), bladder (two lower objectives), and rectum (two lower objectives). Consequently, the planning optimization problem for prostate SBRT involves 16 planning objectives (two lower and 14 upper) while each one of them is controlled by three TPPs, i.e. one for weighting factor, one for volume and one for dose limits, resulting in 48 TPPs in total.

Note that for both TPSs, the TPPs involved critically affect the final plan quality, and hence need to be fine-tuned for clinically acceptable plans. The optimization problems of prostate IMRT and SBRT were solved using a gradient-based optimization algorithm, which iteratively updates the fluence map by enforcing the gradient of the objective function to be close to zero in each step.

### 2.2 Hierarchical formulation of virtual treatment planner network

VTPN established previously (Shen *et al.*, 2020b; Shen *et al.*, 2020a) utilized a parallel architecture to determine a TPPs to adjust, see Fig. 1(a). Depending on $N$, the number of TPPs involved in the optimization problem, a number of $N$ subnetworks was needed with each being dedicated to one TPP. As a consequence, the size of VTPN grew proportionally with $N$. Its practical value was hence limited by the extensive number of computations and load memory it may require, especially for complicated treatment planning problems in real patient cases solved by a commercial TPS.

In contrast to this parallel form, when a human planner operates the TPS and decides how to adjust TPPs, the planner in fact tackles the problem via a hierarchical decision-making process. More precisely, based on the observed plan generated by the TPS, the planner first determines which structure needs further improvement. Then among all the TPPs affecting this structure, one of them is chosen. Finally, the specific way of adjusting this TPP is determined, such as increasing or decreasing the value of this parameter. Motivated by such a hierarchical decision-making behavior of a human planner in the planning process, we proposed to build a hierarchical VTPN (HieVTPN). The detailed architecture of HieVTPN can be found in Fig. 1(b). It consists of three networks, i.e. Structure-Net $s(d, s; \theta_S)$, Parameter-Net $P(d, s, p; \theta_P)$, and Action-Net $A(d, s, p, a; \theta_A)$, with $\theta_S$, $\theta_P$, and $\theta_A$ representing their network parameters. These three networks are responsible to make decisions at the structure, parameter, and adjustment action levels, respectively, and are applied sequentially to improve the plan quality, each time when HieVTPN is interacting with the TPS. Specifically, $S(d, a; \theta_S)$ first takes $d$, the DVH of a plan, as input, and outputs the maximal accumulated future gains in plan quality associated to adjusting different structures for the current TPP adjustment step. Once this network is trained, the structure to choose for the current step can be decided by selecting the one that maximize the output,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

i.e. $s^* = \text{argmax}_s S(d, s; \theta_S)$. Then with the structure fixed, $P(d, s^*, p; \theta_P)$ further predicts the maximal accumulated future gains in plan quality corresponding to adjusting different parameters controlling this objective, such as dose limit and weighting factors of the selected structure $s^*$. Similarly, the parameter to be adjusted at this step is determined via $p^* = \text{argmax}_p P(d, s^*, p; \theta_P)$. Finally, given the structure and the corresponding parameter to adjust, $A(d, s^*, p^*, a; \theta_A)$ outputs the maximal accumulated future gain in plan quality of applying each adjustment action. The optimal action to improve the plan quality is determined as $a^* = \text{argmax}_a A(d, s^*, p^*, a; \theta_A)$. As such, a parameter adjustment action can be uniquely determined as $\{s^*, p^*, a^*\}$.

The detailed architectures of the three networks in HieVTPN are presented in Fig. 2. In general, each network consists of four convolutional blocks followed by three fully connected blocks. A convolutional block ($m$, $n$) is built with two 1D convolutional layers, each followed by a LeakReLU ($\alpha = 0.1$) as the activation function, and a 1D Maxpooling layer. $m$ and $n$ specify the convolutional filter size and number employed in the 1D convolution layer, respectively. A fully connected block ($k$) contains a fully-connected layer and a LeakReLU ($\alpha = 0.1$) activation with $k$ being the number of nodes output from the fully connected layer. Structure-Net takes DVH of a plan as input and predicts the gain in plan quality associated to adjusting TPPs of each structure. Based on its output, a structure-coded DVH (SC-DVH) as an indicator of the selected structure will be computed and fed to Parameter-Net as input. The SC-DVH has the same dimensionality as the original DVH. The DVH of the structure selected by Structure-Net is kept exactly the same in SC-DVH, while DVHs of all other structures are set to be zero. Parameter-Net takes both DVH and SC-DVH as input to predict the gain in plan quality associated with each parameter. After that, a one-hot vector is generated as the indicator of the selected parameter. The entry corresponding to the selected parameter is set to 1 and all the rest are set to 0. This vector, together with DVH and SC-DVH are fed to Action-Net to predict the gain in plan quality associated to each parameter adjustment action.

HieVTPNs established for both IMRT and SBRT in this study consist of three networks regardless of the large difference in the number of TPPs involved (5 vs. 48) due to the hierarchical architecture employed. However, their network sizes are slightly different. More specifically, planning for prostate IMRT involves only three planning structures while SBRT handles 13. In addition, for the IMRT case, the number of TPPs is one for PTV and two for each of the bladder and rectum, while in SBRT, the number of TPPs is six for each of the PTV, bladder, and rectum (three parameters for each planning objectives and two planning objectives for each structure) and three for each of the rest structures (one planning objectives for each). The number of planning structure determines the detailed architecture of Structure-Net while it affects the architectures of Parameter-Net and Action-Net together with the number of TPPs of each planning structure. In the next section, we will provide more detailed description on how the numbers of planning structures and TPPs impact the network sizes of HieVTPN with comparison made against the previously developed VTPN framework.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

### 2.3 Scalability analysis of HieVTPN

Consider a treatment planning task with $m_s$ structures, each with $m_p$ parameters, and the total number of TPPs to be adjusted is $N$. To construct a VTPN to handle this treatment planning problem, the same number of subnetworks are needed with each being dedicated to one TPP. All the subnetworks follow a similar architecture, with the number of trainable network parameters in the first convolution layer and the last fully-connected layer dependent on $m_s$ and $m_p$, respectively. Compared to the variations in the number of parameters in the first and the last layer, the growth in the number of subnetwork dominants the total number of network parameters in VTPN. As a consequence, its size grows in approximately a linear fashion with the number of TPPs involved. A huge VTPN is therefore needed for complicated treatment planning problems. In contrast, the number of networks in HieVTPN is independent of $N$. It always consists of three subnetworks. Only certain layers of Structure-Net, Parameter-Net, and Adjust-Net need to be adjusted accordingly. More specifically, the numbers of parameters in the first convolution layer and the final fully-connected layer of Structure-Net changes with $m_s$. In Parameter-Net, $m_s$ determines the number of network parameters in the first convolution layer, while $m_p$ affects the final fully-connected layer. For Action-Net, $m_s$ affects the number of network parameters in the first convolution layer while $m_p$ additional neurons are added to the input of all the fully connected layers, affecting the number of network parameters in these layers. Such impacts of $m_s$ and $m_p$ to the total number of parameters in each subnetwork are actually negligible since the number of network parameters in majority of the layers remain unchanged. Compared to VTPN, the network size of HieVTPN scales much weaker with the changing number of TPPs. In Section 3.4, we will provide the exact numbers of network parameters calculated for VTPN and HieVTPN, respectively, when different numbers of TPPs are involved in the treatment planning problem for a direct and clear comparison.

### 2.4 Hierarchical Q-learning framework for HieVTPN

We employed the Q-learning framework (Watkins and Dayan, 1992) in this study to train HieVTPN. Specifically, we considered the following optimal action-value function:

$$Q^*(d, C) = \max_{\pi} \left[ r^l + \gamma r^{l+1} + \gamma^2 r^{l+2} + \cdots \mid d^l = d, C^l = C, \pi \right]. \tag{3}$$

$Q^*$ is a function of state, i.e. the observed DVH $d$, and action set $C = \{s, p, a\}$, i.e. the decision regarding TPP adjustment policy formulated in a hierarchical manner. $d^l$ and $C^l$ stand for the state and action set at the $l$-th TPP adjustment step, respectively. $r^l$ is the reward obtained at step $l$, which is given by a predefined reward function related to clinical objectives. A positive reward is obtained, if the clinical objectives are better met by applying the action $C^l$ on the state $d^l$, and negative otherwise. The detailed definition of the reward function in this study will be provided in Section 2.5.1. $\gamma \in [0, 1]$ is a discount factor used to emphasize more on the current reward as opposed to future rewards. $\pi = P(C|d)$ denotes the policy of TPP adjustment: taking an action $C$ based on the observed state $d$. Note that the optimal action-value function satisfies the well-known Bellman equation (Bellman and Karush, 1964), which can be expressed in the form of

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$Q^*\left(d^l, C^l\right) = r^l + \gamma Q^*\left(d^{l+1}, C^{l+1}\right). \tag{4}$$

The general form of the $Q^*$ function is unknown and hence we train HieVTPN to parametrize it with deep neural networks. More precisely, Structure-Net is employed to represent the maximal accumulated future gain in plan quality for adjusting different structures and therefore we require.

$$S(d, s; \theta_S^*) = \max_{a,\, p \in C} Q^*(d, C), \tag{5}$$

where $\theta_S^*$ indicates the optimal network parameters of the fully trained Structure-Net.

Parameter-Net outputs the maximal accumulated gain in plan quality associated to adjusting different TPPs with a structure given. In other words, the training goal of Parameter-Net is to find a set of optimal network parameters $\theta_P^*$, such that

$$P(d, s, p; \theta_P^*) = \max_{a \in C} Q^*(d, C), \tag{6}$$

Furthermore, with the structure and TPP specified, the Action-Net is set up to predict the maximal accumulated gain in plan quality associated to different adjustment action. In this regard, let $\theta_A^*$ represent the optimal parameters of Action-Net, we have

$$A(d, s, p, a; \theta_A^*) = Q^*(d, C). \tag{7}$$

Eq. (5)–(7) explicitly describe the way of formulating HieVTPN into the Q-learning framework to parametrize $Q^*$ function on the structure, parameter, and action levels, respectively. It also reveals a hierarchical relationship among Structure-Net, Parameter-Net, and Action-Net, which can be represented in a sequential form

$$S(d, s; \theta_S) = \max_p P(d, s, p; \theta_P) = \max_{a,\, p} A(d, s, p, a; \theta_A) \tag{8}$$

By taking into account the property of $Q^*$ function characterized by Bellman equation in (4), we are able to combine equations (5)–(8) and reorganized them as

$$\begin{aligned} S\left(d^l, s^l; \theta_S^*\right) &= \max_p P\left(d^l, s^l, p; \theta_P^*\right) = \max_{a,\, p} A\left(d^l, s^l, p, a; \theta_A^*\right), \\ P\left(d^l, s^l, p^l; \theta_P^*\right) &= \max_a A\left(d^l, s^l, p^l, a; \theta_A^*\right), \\ A\left(d^l, s^l, p^l, a^l; \theta_A^*\right) &= r^l + \gamma \max_s S\left(d^{l+1}, s; \theta_S^*\right) \end{aligned} \tag{9}$$

This set of equations derived in (9) motivate the training scheme for HieVTPN, which will be introduced in the following section.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

## 2.5 Training HieVTPN

**2.5.1. Reward function—**Before going into the detailed training scheme of HieVTPN, we will first present the way of computing reward $r$ reflecting the impact of TPPs adjustment action in plan quality for prostate cancer IMRT and SBRT, respectively. The key is to define a plan quality evaluation function $\psi$ (higher is better) to quantify the plan quality, such that the reward function can be calculated as $r = \psi(d') - \psi(d)$. Such a reward explicitly measures the change in plan quality between the two states $d$ to $d'$ before and after TPP adjustment. A positive reward implies improvement in plan quality, and negative otherwise.

For the prostate cancer IMRT, we followed our previous works (Shen *et al.*, 2020b; Shen *et al.*, 2020a) to use the scoring system of ProKnow (ProKnow Systems, Sanford, FL, USA) for prostate cancer IMRT plan as $\psi$ to quantify plan quality. In this scoring system, the score of a plan is calculated as equally-weighted sum over 9 selected scores for different clinical criteria including $D_{PTV}(0.03cc)$, $V_{bladder}(80Gy)$, $V_{bladder}(75Gy)$, $V_{bladder}(70Gy)$, $V_{bladder}(65Gy)$, $V_{rectum}(75Gy)$, $V_{rectum}(70Gy)$, $V_{rectum}(65Gy)$, and $V_{rectum}(60Gy)$. The score for each criterion is computed using carefully defined piece-wise linear function, which can be found in Table A1 of Appendix. Consequently, the score of any plan ranges between 0 and 9, with a higher score indicating better plan quality.

For the prostate cancer SBRT, we incorporated the scoring system used as plan evaluation criteria in the ProwKnow 2016 American Association of Medical Dosimetrist (AAMD) / Radiosurgery Society (RSS) Plan Study, an international planning challenge for prostate cancer SBRT. This scoring system consists of 15 metrics to evaluate the plan quality based on the dose to different dosimetric structures of interest. All the metrics and the detailed ways to compute them have been listed in Table A2 in Appendix. Final score of a prostate SBRT plan is obtained by the summation of the scores from all the metrics. It falls in the range between 0 and 150, and the higher the score is, the better the plan quality is.

**2.5.2 Hierarchical deep reinforcement learning scheme—**The goal of training HieVTPN is to determine the values of network parameters $\theta_S$, $\theta_P$, and $\theta_A$ for Structure-Net, Parameter-Net, and Action-Net, respectively. According to Equations (5)–(9), the three networks are coupled with each other in the formulation of the proposed framework, and existing training schemes developed for DRL fail to handle such a scenario. To address this problem, we have developed a hierarchical DRL (HieDRL) scheme to split the problem and sequentially tackle the training of each individual network in an iterative manner. Note that if $S(d, s; \theta_S)$ is fixed, the training of $A(d, s, p, a; \theta_A)$ can be performed using the standard Q-learning scheme (Watkins and Dayan, 1992; Shen *et al.*, 2020b). Specifically, with randomly initialized TPPs, an initial state $d$ can be obtained for each training patient case by executing plan optimization in TPS. Then an $\epsilon$-greedy algorithm can be employed to sample actions $C$ at structure, parameter, and adjustment levels sequentially to modify values of TPPs in the TPS. Plan optimization will be performed again to generate a new plan $d'$. Comparing its quality score $\psi(d')$ with $\psi(d)$ results in the reward $r$ corresponding to applying the sampled TPP adjustment action to the plan $d$. A training pair, i.e. $\{d, d', C, r\}$ is then generated for $A(d, s, p, a; \theta_A)$ and stored in a training pool $\Omega$. Repeating such a process

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

for all the training patients will produce a large number of training pairs, such that $A(d, s, p, a; \theta_A)$ can be updated by solving the problem

$$\min_{\theta_A} \sum_{\{d, d', C, r\} \in \Omega} \left\| A(d, s, p, a; \theta_A) - r - \gamma \max_{s} S(d', d; \theta_S) \right\|_2^2, \tag{10}$$

which is derived based on the last equation in (9). Such an optimization problem can be solved via the commonly used stochastic gradient descent algorithm (LeCun *et al.*, 1998). On the other hand, with $A(d, s, p, a; \theta_A)$ fixed, the training of $S(d, s; \theta_S)$ and $P(d, s, p; \theta_P)$ can be formulated as the standard supervised learning scheme. As such, the optimization problems for updating $\theta_S$ and $\theta_P$ can be expressed respectively as

$$\begin{aligned} &\min_{\theta_S} \sum_{\{d, C\} \in \Omega} \left\| S(d, s; \theta_S) - \max_{p, a} A(d, s, p, a; \theta_A) \right\|_2^2, \\ &\min_{\theta_P} \sum_{\{d, C\} \in \Omega} \left\| P(d, s, p; \theta_P) - \max_{a} A(d, s, p, a; \theta_A) \right\|_2^2. \end{aligned} \tag{11}$$

Similarly, we can use the stochastic gradient descent algorithm to tackle these two optimization problems. These three networks will be updated alternatively following the aforementioned strategies in the HieDRL, until convergence is established. The detailed algorithm has been summarized in Algorithm 1.

**2.5.3 Implementation details—**We followed the scheme in Algorithm 1 to perform training of HieVTPN for both MRT and SBRT planning. For the proof-of-principle planning problem of prostate IMRT, A cohort of 74 patient cases were collected for training, validation, and testing purposes. Among them, 10 patient cases were randomly picked for training, while another five patients were chosen as validation data. All the patients left were employed as testing data to evaluate our model. For prostate SBRT, we collected 20 patient cases. We randomly selected 10 for training, five for validation, and the rest five for testing purpose. Note that we could not use the patient cases collected for IMRT planning for SBRT since only the contours of bladder and rectum were recorded as OARs for the simple proof-of-principle IMRT planning problem, while other critical structures considered in SBRT, such as penile bulb and neurovascular bundles, were missing. DRL automatically generates a huge number of training data in the ϵ-greedy process, and it was found that 10 patients were sufficient to train a working model as demonstrated by (Shen *et al.*, 2020b; Shen *et al.*, 2020a). The maximal training episode number $N_E$ was set to 300 for IMRT and 1,000 for SBRT treatment planning. In each episode, we started with all TPPs randomly set for all training cases. For each training case, a sequence of TPPs adjustment steps were performed following the Algorithm 1. We terminated the TPPs adjustment and moved on to the next patient if the number of adjustment steps reached the maximum of $N_{T1} = 30$ for IMRT, or $N_{T1} = 100$ for SBRT. For both IMRT and SBRT, in each step, we selected $\{s, p, a\}$ to adjust a TPP using the $\epsilon$-greedy algorithm. Specifically, with a probability of $\epsilon$ (initial $\epsilon = 0.99$), we randomly picked the structure, parameter, and adjustment action uniformly among all possible choices; otherwise, $\{s, p, a\}$ that attained the highest output value of Structure-Net, Parameter-Net, and Action-Net, respectively was chosen. The probability $\epsilon$ decayed with a decay rate of 0.99/episode along the training process, because we gained

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

more and more confidence about the trained HieVTPN. Once $\{s, p, a\}$ were fixed, the TPPs were adjusted accordingly and then were fed to the in-house developed TPS for plan optimization. DVHs of the plans before and after the TPP adjustment were used to compute the reward function $r$. The states prior to and after TPP adjustment, the chosen action, and the reward $r$ computed based on the states together formed a training sample which was stored in the training pool $\Omega$. For training Action-Net, the experience replay strategy was employed to update the network using batches of 16 randomly selected training sample in order to overcome the strong correlation among sequentially generated training samples (Mnih *et al.*, 2015). In addition, Structure-Net and Parameter-Net were trained using states stored in the training pool via the standard supervised training scheme while the number of training steps $N_{T2}$ in each episode was set to be 1,000.

The HieDRL framework was implemented using Python with TensorFlow (Abadi *et al.*, 2016) on a desktop workstation with eight Intel Xeon 3.5 GHz CPU processors, 32 GB memory and two Nvidia Quadro M4000 GPU cards.

**Algorithm 1.**

HieDRL algorithm to train HieVTPN.

Initialize network coefficients $\theta_S$, $\theta_P$, and $\theta_A$;

**for** episode = 1, 2, …, $N_E$

  **for** $k$ = 1, 2, …, $N_P$ **do**

    2. Initialize TPPs

      Solve optimization problem (1)/(2) with initial TPPs for $d^1$;

    **for** $l$ = 1, 2, …, $N_{T1}$ **do**

      3. Select an action $C^l = \{s^l, p^l, a^l\}$ with $\epsilon$-greedy:

        **Case 1:** with probability $\epsilon$, select $s^l$, $p^l$, and $a^l$ randomly;

        **Case 2:** otherwise $s^l = \underset{s}{\operatorname{argmax}} S\left(d^l, s; \theta_S\right)$,

        $p^l = \underset{p}{\operatorname{argmax}} P\left(d^l, s^l, p; \theta_P\right)$, and

        $a^l = \underset{a}{\operatorname{argmax}} A\left(d^l, s^l, p^l, a; \theta_A\right)$;

      4. Update TPPs using $C^l$;

      5. Solve optimization problem (1)/(2) with updated TPPs for $d^{l+1}$;

      6. Compute reward $r^l = \Phi(d^{l+1}) - \Phi(d^l)$;

      7. Store state-action pair $\{d^l, C^l, r^l, d^{l+1}\}$ in training data pool $\Omega$;

      8. Train Action-Net with experience replay:

        Randomly select $N_{53678}$ training data from training data pool $\Omega$;

        Update $\theta_A$ by solving the optimization problem in (10);

    **end for**

  **end for**

  **for** $k$ = 1, 2, …, $N_{T2}$ **do**

    9. Train $\theta_S$ and $\theta_P$:

      Randomly select $N_{batch}$ states from training data pool $\Omega$;

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Forward evaluate Action-Net using the selected states as input:

Compute $A(d, s, p, a; \theta_A)$ for all possible choices of $s$, $p$, and $a$;

Compute $\max_{p,a} A(d, s, p, a; \theta_A)$ and $\max_{a} A(d, s, p, a; \theta_A)$;

Update $\theta_S$ and $\theta_P$, respectively, by solving the optimization problems in (11)

**end for**

**end for**

**Output** $\theta_S$, $\theta_P$ and $\theta_A$.

### 2.6 Evaluations

For both IMRT and SBRT cases, we chose the models achieving the best performance on validation data and evaluated them on the corresponding testing cases. For model evaluation, we set all TPPs to be unity at the beginning of the treatment planning process for each case and used the trained HieVTPNs to operate the Eclipse-like in-house TPS. The repetitive TPP adjustment was continued, until one of the following three criteria was met: the plan reached the maximal score (9 for IMRT and 150 for SBRT), HieVTPNs decided to keep all TPPs unchanged, or a maximal number of adjustment steps (50 for IMRT and 150 for SBRT) was reached. We compared the quality of the final plans with that of the initial plans generated using the initial TPP values. For prostate IMRT, we also compared with the scores of the plans generated by VTPN of the parallel architecture trained via the standard DRL (Shen *et al.*, 2020b) to benchmark the performance of HieVTPN. We didn't try VTPN for SBRT since the computational demands exceed the capability of the workstation.

### 2.7 Understanding the decision-making behaviors of HieVTPN

In addition to evaluating the performance of HieVTPN in terms of quality of the generated plans, we analyzed its decision-making behaviors in the treatment planning process for interpretability of the trained network model. We focused on the HieVTPN established for the proof-of-principle prostate cancer IMRT planning task for clear presentation and easy understanding of the results. For each patient case in the testing dataset, we generated 500 treatment plans using the in-house developed TPS with randomly selected TPPs values. As the TPPs were not purposely set to ensure plan quality, the corresponding plans under these set of TPPs would cover a wide range of scenarios in terms of quality. These treatment plans were then fed to HieVTPN one by one, and we observed the way HieVTPN determined to adjust TPPs based on the observed plan. The plan scores and the corresponding TPP adjustment actions were recorded and analyzed to understand the rationale of the decision-making behaviors of the established HieVTPN in the treatment planning process.

## 3. Results

### 3.1 Training results

Training of HieVTPNs for prostate IMRT and SBRT were successfully performed using the HieDRL algorithm (Algorithm 1). In Fig. 3, we depict the rewards obtained during the training process, respectively. The effectiveness of the proposed training framework is indicated by the overall increasing trend in rewards along the training steps for both cases.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Note that we selected the model established at the episode 288 for IMRT and episode 872 for SBRT, as the final models (highlighted by red circle in Fig. 3), since they achieved the best performance on the corresponding validation datasets among all the models saved along training steps. For IMRT, the selected HieVTPN obtained an average plan score of 8.76 (out of 9) on validation dataset, while the HieVTPN chosen for SBRT achieved a score of 137.89 (out of 150).

### 3.2 HieVTPN guided intelligent automatic treatment planning

We let the trained HieVTPNs for IMRT and SBRT to operate the TPS to automatically generate treatment plans for their corresponding training and testing patients. Using IMRT planning as an example, we show shows the complete planning process of HieVTPN for one representative case in the testing dataset in Fig. 4. The whole planning process involved 9 steps. At the beginning of the planning process, the plan was produced by setting all TPPs to 1 and performed the optimization. After that, HieVTPN first decided to decrease the threshold values of rectum in the first step to spare more dose to the rectum. With dose to rectum significantly decreased, it then focused on adjusting PTV weighting factor in the next 5 steps to improve PTV coverages. In step 7, $\tau_{bladder}$ controlling the dose limit to balder, was adjusted, which successfully decreased the bladder dose. Afterwards, HieVTPN continued to increasing weighting factor of PTV to further enhance the PTV homogeneity, until in step 10 it decided to keep all TPPs unchanged which concluded the planning process. In this process, the ProKnow score of the generated plan was increased from 2 to 8.35, which was close to the maximal score of 9.

In Table 2, we report the average performance of HieVTPN in prostate IMRT and SBRT planning, and compare the achieved average plan scores with those of the initial plans. For IMRT planning, we also compared the performance with VTPN established using the standard DRL approach (Shen *et al.*, 2020b). For IMRT planning, both HieVTPN and VTPN were able to automatically operate the TPS to generate high-quality plans, as evidenced by the high average plan quality scores of 8.62 and 8.43, respectively on testing data while in SBRT planning, the HieVTPN successfully improved the plan score from 95.56 of the initial plans to 139.07.

To further benchmark the performance, for IMRT treatment planning, we asked a human planner with a good understanding of the optimization engine and extensive experience in treatment planning to plan for the testing cases. The human planner was able to achieve an average plan score of ~8.5, which is comparable to the performance of HieVTPN, and VTPN. On average, the planning time needed for each patient using HieVTPN was around 3 min, similar to that of VTPN, while it took around the same time for human planner to complete the planning process.

For SBRT case, we compared the achieved plan score with that reported in the results of the 2016 AAMD/RSS Plan Study for prostate SBRT (Nelms and Mobile). The proposed HieVTPN was able to outperform the averaged plan score of 132.21. Note that these plans were made by experienced human planners and were submitted for competition in the planning challenge. This further illustrates the effectiveness of the prosed framework. The average planning time needed for HieVTPN to perform SBRT planning is ~8 min. Note

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

that most of the planning time for both IMRT and SBRT was spent on plan optimization, as determining the way to adjust TPPs only required forward evaluation of the established network, which can be achieved in almost real-time.

### 3.3 Decision-making behaviors of HieVTPN for prostate cancer IMRT

Using IMRT planning as an example, we studied what the trained HieVTPN decided to perform in response to a plan. As such, we randomly generated a number of plans by feeding random TPPs to the TPS and then observed how the HieVTPN responded to the plans. The results are shown in Fig. 5. In Fig. 5(a), we plot the frequency of actions that adjust PTV-related parameters or OAR-related parameters in response to observed scores in the plan. Note that there are multiple adjustable parameters for PTV or each of the OARs. When counting the frequency, we did not further differentiate among parameters for the PTV or each OAR. Among all the scores for different dosimetric measures for PTV and OARs, all the scores for the OARs (bladder and rectum) add up to maximally 8, which hence sets the range of the OAR score. The PTV score ranges from 0 to 1. If we ignore the color of the bars in both Fig. 5(a) and (b), the value of the vertical axis of each bar shows how many among the randomly generated plans obtained the plan scores specified by the coordinates of the other two axis. In general, the value of Z-axis is quite uniform with across the different plan scores mainly due to the random TPPs employed in the plan generation process. This actually highlights the need of careful TPPs adjustment, since random TPPs may end up to a treatment plan with arbitrary quality. It has a large chance to generate unacceptable plans with low plan scores. The key information in Fig. 5 is encoded by the color of the bars. Note that the color on the top of each bar shows the dominant actions taken by the HieVTPN by observing those plans receiving the specific plan scores, while the ratio between the two colors in each bar reflects the ratio of the actions taken by HieVTPN. In Fig. 5(a), it was apparent that HieVTPN tended to adjust the TPPs of OARs for plans with poor OAR sparing. Once the OAR score reached a reasonable level ~5, it focused on tuning the parameter of PTV. Fig 5(b) shows the frequency of decisions made by HieVTPN in responses to bladder and rectum scores, when it decided to adjust TPPs for these OARs. In general, HieVTPN tended to pick the one receiving the lower score among the two organs. There is visually a diagonal line in the rectum-bladder score plane. On the side corresponding to higher bladder scores, HieVTPN preferred to adjust rectum TPPs to improve scores for the rectum, and bladder TPPs for the other side. These behaviors indicated that the HieVTPN was trained to make reasonable decisions to improve plan quality.

We further investigated the rationale of HieVTPN's decision-making behaviors of in terms of choosing between adjusting dose limit and weighting factors. Fig. 6 gives the histograms showing the frequencies of HieVTPN taking different actions for each TPPs with respect to the score of the structure. The strategy learnt by the HieVTPN appeared to be in general agreement with a planner's intuition. For instance, when the PTV or OAR receives a low score, HieVTPN increased the weighting factor of the structure to enhance its importance in the objective function, as indicated by the blue bars in in Fig. 6(a), (c1) and (c2). On the other hand, when a high score had already been achieved for the structure, e.g. a score close to 1 for PTV or close to 4 for each of the OARs, HieVTPN decided to reduce the weights.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

As for adjusting the dose limits to an OARs, as indicated by Fig. 6(b1) and (b2), when better OAR sparing was desired, the decision of reducing the dose limit was made. When the scores for OARs were high enough, it allowed increasing the dose limits to sacrifice dose to the OAR to improve on other dosimetric metrics.

### 3.4 Comparisons of network parameter numbers of VTPN and HieVTPN

In this section, we compare the number of learnable network parameters in HieVTPN architecture with that of VTPN constructed in (Shen *et al.*, 2020b; Shen *et al.*, 2020a). In Table 3, we considered not only the scenarios of the proposed study which involved 5 (highlighted using *) in IMRT planning and 48 TPPs (highlighted using **) in SBRT planning, but also three others when 3, 30, and 60 TPPs involved in different treatment planning problems. We found that the network sizes of VTPN and HieVTPN are actually very close to each other, when there are only three TPPs involved. This is expected since in this case both networks consist of three subnetworks. To handle the IMRT planning, building VTPN needs ~66.6% more network parameters than constructing HieVTPN, while the number of network parameters needed by VTPN for SBRT planning is ~16 times of that needed by HieVTPN. In general, as the number of TPPs increases, the number of network parameters in VTPN grows in a linear fashion with the number of TPPs, while the network size of HieVTPN does not have any substantial changes. For instance, with the number of TPPs increasing from three in IMRT to 48 in SBRT, the network size of VTPN grows 16 times, while there is only a very small increment of ~0.1% in the number of trainable parameters in HieVTPN. This clearly illustrates the advantage of HieVTPN in scalability.

## 4. Discussion

Due to the hierarchical formulation, HieVTPN possesses a much better scalability compared to VTPN. The overall model size of HieVTPN was not significantly affected by the number of TPPs in a TPS, as opposed to the VTPN where the size of network grows approximately linearly with the number of TPPs. In this regard, the proposed HieVTPN has a great potential to accomplish the intelligent automatic treatment planning task for more complicated, but clinically relevant treatment planning problems in modern RT, which may often involve tens of TPPs to adjust.

In addition to numerical experiments performed for the purpose of demonstrating the performance of HieVTPN, we also studied the rationale of the decision-making behaviors of the established HieVTPN. By feeding a large number of treatment plans optimized with random TPPs into HieVTPN, we investigated the correlations between the plan scores and the decisions made by HieVTPN. Based on the results, it appeared that the behaviors learnt by the HieVTPN were in general agreement with the human experience. In most of the cases, it selected appropriate parameters and actions to address the issues in the treatment plan, and hence improve the plan quality. We would like to emphasize that these behaviors were spontaneously discovered by the HieVTPN in the HieDRL training process. Such reasonable decision-making behaviors gained us confidence in the successful deployment of HieVTPN framework for real clinical applications in future.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The current study has several limitations. First, for the purpose of demonstrating the feasibility of building HieVTPN via the proposed HieDRL, this paper used simple reward functions derived from the ProKnow scoring systems for prostate cancer IMRT and SBRT, respectively. They may not fully represent the planning objectives used in real clinical practice. Model criteria of more clinical relevance, e.g. physician's judgement, as reward function is necessary for building a clinically translatable HieVTPN. One possible solution is to incorporate the inverse deep reinforcement learning (Wulfmeier et al., 2015) to jointly learn the physicians judgement and TPS operating policy simultaneously in a unified framework. Second, we considered a relatively simple treatment planning problem of prostate cancer IMRT using an in-house TPS. To bring clinical impact, it is needed to develop intelligent automatic treatment planning systems to handle planning problems of more complicated treatment sites using commercially available TPSs. Future work will be performed along this direction. In addition, the proposed framework dose not incorporate any roll-back mechanism, which is common in human practice. This is mainly because that our formulation trains HieVTPN would pick the action that maximizes the total gain in plan quality in all subsequent TPPs adjustment steps, as we focused on the overall future reward with a discount rate (0.99 in our experiments) in the Q-learning process, see Eq. (3). However, from a practical point of view, a roll-back mechanism can be valuable as it allows to return to the previous plan if a serial of TPPs adjustment actions taken by HieVTPN fail to improve, or continuously degrade the plan quality. Effective integration of roll-back mechanism into the proposed framework will be an interesting future direction to explore.

## 5. Conclusion

This paper introduced a new hierarchical formulation of VTPN, i.e. HieVTPN, for the purpose of having a scalable network structure in the IATP framework to automatically operate a TPS for high-quality treatment planning. Compared to the conventional VTPN, which grows linearly with the number of TPPs involved in treatment planning problem, the network size of HieVTPN does not change significantly as the TPP number increases. In this regard, HieVTPN is more suitable to handle complicate clinical treatment planning tasks. Using prostate cancer IMRT and SBRT treatment planning as the testbeds, we showed that HieVTPNs can be successfully trained to automatically generate high-quality treatment plans by operating an in-house developed Eclipse-like TPS. The resulting IMRT plans were comparable to those made by conventional VTPN of parallel architecture, while for SBRT planning, the average plan score achieved by HieVTPN slightly outperformed that of the human plans submitted for 2016 AAMD\RSS Plan Study for prostate SBRT. The success of HieVTPN in these two planning tasks illustrated the effectiveness of the proposed scheme.

## Acknowledgement

This work was supported by the National Institutes of Health grant number R01CA237269 and Cancer Prevention and Research Institute of Texas grant number RP160661.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

# Appendix

**Table A1.**

Criteria in the ProKnow scoring system for prostate IMRT plan quality evaluation.

| Quantity of interest | Scoring Criterion |
|---|---|
| PTV D[0.03cc] (Gy) ($D_{PTV}$ [0.03cc]) | $\text{Score} = \begin{cases} 1, & \text{if } D_{PTV}[0.03cc] < 84.4\ \text{Gy} \\ \frac{D_{PTV}[0.03cc] - 87.12\text{Gy}}{84.4\text{Gy} - 87.12\text{Gy}}, & \text{if } 84.4\text{Gy} \le D_{PTV}[0.03cc] \le 87.12\text{Gy} \\ 0, & \text{if } D_{PTV}[0.03cc] > 87.12\text{Gy} \end{cases}$ |
| Bladder V[80Gy] (%) [$V_{bla}$[80Gy] ] | $\text{Score} = \begin{cases} 1, & \text{if } V_{bla}[80\text{Gy}] < 15\% \\ \frac{\text{Bladder } V_{bla}[80\text{Gy}] - 20\%}{15\% - 20\%}, & \text{if } 15\% \le V_{bla}[80\text{Gy}] \le 20\% \\ 0, & \text{if } V_{bla}[80\text{Gy}] > 20\% \end{cases}$ |
| Bladder V[75Gy] (%) ($V_{bla}$[75Gy]) | $\text{Score} = \begin{cases} 1, & \text{if } V_{bla}[75\text{Gy}] < 25\% \\ \frac{V_{bla}[75\text{Gy}] - 30\%}{25\% - 30\%}, & \text{if } 25\% \le V_{bla}[75\text{Gy}] \le 30\% \\ 0, & \text{if } V_{bla}[75\text{Gy}] > 30\% \end{cases}$ |
| Bladder V[70Gy] (%) ( $V_{bla}$[70Gy]) | $\text{Score} = \begin{cases} 1, & \text{if } V_{bla}\,[70\text{Gy}] < 35\% \\ \frac{V_{bla}\,[70\text{Gy}] - 40\%}{35\% - 40\%}, & \text{if } 35\% \le V_{bla}\,[70\text{Gy}] \le 40\% \\ 0, & \text{if } V_{bla}\,[70\text{Gy}] > 40\% \end{cases}$ |
| Bladder V[65Gy] (%) ($V_{bla}$[65Gy]) | $\text{Score} = \begin{cases} 1, & \text{if } V_{bla}\,[65\text{Gy}] < 50\% \\ \frac{V_{bla}\,[65\text{Gy}] - 55\%}{50\% - 55\%}, & \text{if } 50\% \le V_{bla}\,[65\text{Gy }] \le 55\% \\ 0, & \text{if } V_{bla}\,[65\text{Gy}] > 55\% \end{cases}$ |
| Rectum V[75Gy] (%) ($V_{rec}$[75Gy]) | $\text{Score} = \begin{cases} 1, & \text{if } V_{rec}\,[75\text{Gy }] < 15\% \\ \frac{V_{rec}\,[75\text{Gy}] - 20\%}{15\% - 20\%}, & \text{if } 15\% \le V_{rec}\,[75\text{Gy}] \le 20\% \\ 0, & \text{if } V_{rec}\,[75\text{Gy}] > 20\% \end{cases}$ |
| Rectum V[70Gy] (%) ($V_{rec}$[70Gy] ) | $\text{Score} = \begin{cases} 1, & \text{if } V_{rec}[70\text{Gy}] < 25\% \\ \frac{V_{rec}[70\text{Gy}] - 30\%}{25\% - 30\%}, & \text{if } 25\% \le V_{rec}[70\text{Gy}] \le 30\% \\ 0, & \text{if } V_{rec}[70\text{Gy}] > 30\% \end{cases}$ |
| Rectum V[65Gy] (%) ($V_{rec}$[65Gy]) | $\text{Score} = \begin{cases} 1, & \text{if } V_{rec}\,[65\text{Gy}] < 35\% \\ \frac{V_{rec}\,[65\text{Gy}] - 40\%}{35\% - 40\%}, & \text{if } 35\% \le V_{rec}\,[65\text{Gy}] \le 40\% \\ 0, & \text{if } V_{rec}\,[65\text{Gy}] > 40\% \end{cases}$ |
| Rectum V[60Gy] (%) ($V_{rec}$[60Gy]) | $\text{Score} = \begin{cases} 1, & \text{if } V_{rec}[60\text{Gy}] < 50\% \\ \frac{V_{rec}\,[60\text{Gy}] - 55\%}{50\% - 55\%}, & \text{if } 50\% \le V_{rec}[60\text{Gy}] \le 55\% \\ 0, & \text{if } V_{rec}[60\text{Gy}] > 55\% \end{cases}$ |

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

**Table A2.**

Criteria in the 2016 AAMD/RSS Plan Study for prostate SBRT.

| Quantity of interest | Scoring Criterion |
|---|---|
| PTV V[36.25Gy] (%)) ($V_{PTV}$[36.25Gy]) | $\text{Score} = \begin{cases} 35, & \text{if } V_{PTV}\,[36.25\text{Gy}] \geq 95\% \\ \frac{5 \times (V_{PTV}[36.25\text{Gy}] - 93\%)}{95\% - 93\%} + 30, & \text{if } 93\% \leq V_{PTV}[36.25\text{Gy}] < 95\% \\ \frac{30 \times (V_{PTV}[36.25\text{Gy}] - 90\%)}{93\% - 90\%}, & \text{if } 90\% \leq V_{PTV}[36.25\text{Gy}] < 93\% \\ 0, & \text{if } V_{PTV}[36.25\text{Gy}] < 90\% \end{cases}$ |
| Prostate V[40Gy] (%) ($V_{pros}$[40Gy]) | $\text{Score} = \begin{cases} \frac{2 \times \left(V_{pros}\,[40\text{Gy}] - 95\%\right)}{100\% - 95\%} + 18, & \text{if } V_{pros}[40\text{Gy}] \geq 95\% \\ \frac{18 \times \left(V_{pros}[40\text{Gy}] - 90\%\right)}{95\% - 90\%}, & \text{if } 90\% \leq V_{pros}[40\text{Gy}] < 95\% \\ 0, & \text{if } V_{pros}[40\text{Gy}] < 90\% \end{cases}$ |
| PTV D[0.03cc] (Gy) ($D_{PTV}$[0.03cc]) | $\text{Score} = \begin{cases} 10, & \text{if } D_{PTV}[0.03\text{cc}] \geq 36.25\text{Gy} \\ \frac{2 \times (D_{PTV}[0.03\text{cc}] - 32.625\text{Gy})}{36.25\text{Gy} - 32.625\text{Gy}} + 8, & \text{if } 32.625\text{Gy} \leq D_{PTV}[0.03\text{cc}] < 36.25\text{ Gy} \\ \frac{8 \times (D_{PTV}[0.03\text{cc}] - 29\text{Gy})}{32.625\text{Gy} - 29\text{Gy}}, & \text{if } 29\text{Gy} \leq DD_{PTV}[0.03\text{cc}] < 32.625\text{Gy} \\ 0, & \text{if } D_{PTV}[0.03\text{cc}] < 29\text{Gy} \end{cases}$ |
| Conformation number (Conf) | $\text{Score} = \begin{cases} \frac{10 \times (\text{Conf} - 0.6)}{1 - 0.6}, & \text{if Conf} \geq 0.6 \\ 0, & \text{if Conf} < 0.6 \end{cases}$ |
| Rectum V[36Gy] (cc)) ($V_{rec}$[36Gy]) | $\text{Score} = \begin{cases} 15 - \frac{1.5 \times V_{rec}\,[36\text{Gy}]}{1\text{cc}}, & \text{if } V_{rec}\,[36\text{Gy}] \leq 1\text{cc} \\ 13.5 - \frac{13.5 \times (V_{rec}\,[366\text{y}] - 1\text{cc})}{1\text{cc}}, & \text{if } 1\text{cc} < V_{rec}\,[36\text{Gy}] \leq 2\text{cc} \\ 0, & \text{if } V_{rec}\,[36\text{Gy}] > 2\text{cc} \end{cases}$ |
| Bladder V[37Gy] (cc) ) ($V_{bla}$[37Gy]) | $\text{Score} = \begin{cases} 15 - \frac{1.5 \times V_{bla}\,[37\text{Gy}]}{3\text{cc}}, & \text{if } V_{bla}\,[37\text{Gy}] \leq 3\text{cc} \\ 13.5 - \frac{13.5 \times (V_{bla}[37\text{Gy}] - 3\text{cc})}{5\text{cc} - 3\text{cc}}, & \text{if } 3\text{cc} < V_{rec}[36\text{Gy}] \leq 5\text{cc} \\ 0, & \text{if } V_{bla}[37\text{Gy}] > 5\text{cc} \end{cases}$ |
| Rectum D[40%] (Gy) ($D_{rec}$ [40%]) | $\text{Score} = \begin{cases} 12, & \text{if } D_{rec}[40\%] \leq 10\text{Gy} \\ 12 - \frac{2 \times (D_{rec}[40\%] - 10\text{Gy})}{15\text{Gy} - 10\text{Gy}}, & \text{if } 10\text{Gy} < D_{rec}[40\%] \leq 15\text{Gy} \\ 10 - \frac{10 \times (D_{rec}[40\%] - 15\text{Gy})}{20\text{Gy} - 15\text{Gy}}, & \text{if } 15\text{Gy} < D_{rec}[40\%] \leq 20\text{Gy} \\ 0, & \text{if } D_{rec}[40\%] > 20\text{Gy} \end{cases}$ |
| Urethra D[20%] (Gy) ($D_{ure}$ [20%]) | $\text{Score} = \begin{cases} 10, & \text{if } D_{ure}\,[20\%] \leq 40\text{Gy} \\ 10 - \frac{10 \times (D_{ure}\,[20\%] - 40\text{Gy})}{44\text{Gy} - 40\text{Gy}}, & \text{if } 40\text{Gy} < D_{ure}\,[20\%] \leq 44\text{Gy} \\ 0, & \text{if } D_{ure}\,[20\%] > 44\text{Gy} \end{cases}$ |

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

| Quantity of interest | Scoring Criterion |
|---|---|
| Bowel D[1cc] (Gy) ($D_{bow}$[1cc]) | $\text{Score} = \begin{cases} \frac{5 \times (30\text{Gy} - D_{bow}\,[1\text{cc}])}{30\text{Gy}}, & \text{if } D_{bow}\,[1\text{cc}] \le 30\text{Gy} \\ 0, & \text{if } D_{bow}\,[1\text{cc}] > 30\text{Gy} \end{cases}$ |
| Penile bulb D[0.1cc] (Gy) ($D_{PB}$[0.1cc]) | $\text{Score} = \begin{cases} 3, & \text{if } D_{PB}[0.1\text{cc}] \le 10\text{Gy} \\ 3 - \frac{3 \times (D_{PB}[0.1\text{cc}] - 10\text{Gy})}{29.5\text{Gy} - 10\text{Gy}}, & \text{if } 10\text{Gy} < D_{PB}[0.1\text{cc}] \le 29.5\text{Gy} \\ 0, & \text{if } D_{PB}[0.1\text{cc}] > 29.5\text{Gy} \end{cases}$ |
| Neurovascular Bundles D[50%] (Gy) ($D_{NB}$[50%]) | $\text{Score} = \begin{cases} 3, & \text{if } D_{NB}[50\%] \le 37.5\text{Gy} \\ 3 - \frac{3 \times (D_{NB}[50\%] - 37.5\text{Gy})}{40\text{Gy} - 37.5\text{Gy}}, & \text{if } 37.5\text{Gy} < D_{NB}[50\%] \le 40\text{Gy} \\ 0, & \text{if } D_{NB}[50\%] > 40\text{Gy} \end{cases}$ |
| Right femoral head D[max] (Gy) ($D_{RFH}$[max]) | $\text{Score} = \begin{cases} 3, & \text{if } D_{RFH}[\text{max}] \le 10\text{Gy} \\ 3 - \frac{0.3 \times (D_{RFH}[\text{max}] - 10\text{Gy})}{14\text{Gy} - 10\text{Gy}}, & \text{if } 10\text{Gy} < D_{RFH}[\text{max}] \le 14\text{Gy} \\ 2.7 - \frac{2.7 \times (D_{RFH}[\text{max}] - 14\text{Gy})}{27.5\text{Gy} - 14\text{Gy}}, & \text{if } 14\text{Gy} < D_{RFH}[\text{max}] \le 27.5\text{Gy} \\ 0, & \text{if } D_{RFH}[\text{max}] > 27.5\text{Gy} \end{cases}$ |
| Left femoral head D[max] (Gy) ($D_{LFH}$ [max]) | $\text{score} = \begin{cases} 3, & \text{if } D_{LFH}[\text{max}] \le 10\text{Gy} \\ 3 - \frac{0.3 \times (D_{LFH}[\text{max}] - 10\text{Gy})}{14\text{Gy} - 10\text{Gy}}, & \text{if } 10\text{Gy} < D_{LFH}[\text{max}] \le 14\text{Gy} \\ 2.7 - \frac{2.7 \times (D_{LFH}[\text{max}] - 14\text{Gy})}{27.5\text{Gy} - 14\text{Gy}}, & \text{if } 14\text{Gy} < D_{LFH}[\text{max}] \le 27.5\text{Gy} \\ 0, & \text{if } D_{LFH}[\text{max}] > 27.5\text{Gy} \end{cases}$ |
| Skin D[max] (Gy) ($D_{skin}$ [max]) | $\text{Score} = \begin{cases} 3, & \text{if } D_{skin}\,[\text{max}] \le 10\text{Gy} \\ 3 - \frac{3 \times (D_{skin}\,[\text{max}] - 10\text{Gy})}{30\text{Gy} - 10\text{Gy}}, & \text{if } 10\text{Gy} < D_{skin}\,[\text{max}] \le 30\text{Gy} \\ 0, & \text{if } D_{skin}[\text{max}] > 30\text{Gy} \end{cases}$ |
| Testes D[max] (Gy) ($D_{tes}$[max]) | $\text{Score} = \begin{cases} \frac{3 \times (2\text{Gy} - D_{tes}[\text{max}])}{30\text{Gy}}, & \text{if } D_{tes}[\text{max}] \le 2\text{Gy} \\ 0, & \text{if } D_{tes}[\text{max}] > 2\text{Gy} \end{cases}$ |

## References


Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G and Isard M 2016 TensorFlow: A System for Large-Scale Machine Learning. In: OSDI, pp 265–83

Bellman R and Karush R 1964 Dynamic programming: a bibliography of theory and application: RAND CORP SANTA MONICA CA)

Boutilier JJ, Lee T, Craig T, Sharpe M B and Chan TCY 2015 Models for predicting objective function weights in prostate cancer IMRT Medical Physics 42 1586–95 [PubMed: 25832049]

Chan TCY, Craig T, Lee T and Sharpe MB 2014 Generalized Inverse Multiobjective Optimization with Application to Cancer Therapy Operations Research 62 680–95

Das I J, Cheng C-W, Chopra KL, Mitra RK, Srivastava S P and Glatstein E 2008 Intensity-Modulated Radiation Therapy Dose Prescription, Recording, and Delivery: Patterns of Variability Among Institutions and Treatment Planning Systems JNCI: Journal of the National Cancer Institute 100 300–7 [PubMed: 18314476]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Eclipse2015Eclipse Photon and Electron Algorithms Reference Guide. Varian Medical SystemsPalo Alto, CA)

Fan J, Wang J, Chen Z, Hu C, Zhang Z and Hu W 2019 Automatic treatment planning based on three - dimensional dose distribution predicted from deep learning technique Med Phys 46 370–81 [PubMed: 30383300]

Holdsworth C, Kim M, Liao J and Phillips M 2012 The use of a multiobjective evolutionary algorithm to increase flexibility in the search for better IMRT plans Med Phys 39 2261–74 [PubMed: 22482647]

Holdsworth C, Kim M, Liao J and Phillips MH 2010 A hierarchical evolutionary algorithm for multiobjective optimization in IMRT Med Phys 37 4986–97 [PubMed: 20964218]

Hrinivich WT and Lee J Artificial intelligence-based radiotherapy machine parameter optimization using reinforcement learning Med Phys n/a

Jia X, Ren L and Cai J 2020 Clinical implementation of AI technologies will require interpretable AI models Medical physics 47 1–4 [PubMed: 31663612]

LeCun Y, Bottou L, Bengio Y and Haffner P 1998 Gradient-based learning applied to document recognition Proceedings of the IEEE 86 2278–324

Lee T, Hammad M, Chan TCY, Craig T and Sharpe MB 2013 Predicting objective function weights from patient anatomy in prostate IMRT treatment planning Medical Physics 40 121706-n/a [PubMed: 24320492]

Li X, Zhang J, Sheng Y, Chang Y, Yin F-F, Ge Y, Wu QJ and Wang C 2020 Automatic IMRT planning via static field fluence prediction (AIP-SFFP): a deep learning algorithm for real-time prostate treatment planning Physics in Medicine & Biology 65 175014 [PubMed: 32663813]

Lu R, Radke R J, Happersett L, Yang J, Chui C-S, Yorke E and Jackson A 2007 Reduced-order parameter optimization for simplifying prostate IMRT planning Physics in Medicine & Biology 52 849 [PubMed: 17228125]

Mahmood R, Babier A, McNiven A, Diamant A and Chan TC 2018 Automated treatment planning in radiation therapy using generative adversarial networks arXiv preprint arXiv:1807.06489

Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, Petersen S, Beattie C, Sadik A, Antonoglou I, King H, Kumaran D, Wierstra D, Legg S and Hassabis D 2015 Human-level control through deep reinforcement learning Nature 518 529–33 [PubMed: 25719670]

Nelms B and Mobile K 2016 2016 AAMD/RSS Plan Study SBRT Prostate (https://blog.proknowsystems.com/wp-content/uploads/2016/06/2016-AAMD-RSS-Plan-Study.pdf).

Nelms BE, Robinson G, Markham J, Velasco K, Boyd S, Narayan S, Wheeler J and Sobczak ML 2012 Variation in external beam treatment plan quality: An inter-institutional study of planners and planning systems Practical Radiation Oncology 2 296–305 [PubMed: 24674168]

Nguyen D, Barkousaraie AS, Shen C, Jia X and Jiang S 2019 Generating Pareto optimal dose distributions for radiation therapy treatment planning arXiv preprint arXiv:1906.04778

Nguyen D, McBeth R, Sadeghnejad Barkousaraie A, Bohara G, Shen C, Jia X and Jiang S 2020 Incorporating human and learned domain knowledge into training deep neural networks: A differentiable dose-volume histogram and adversarial inspired framework for generating Pareto optimal dose distributions in radiation therapy Med Phys 47 837–49 [PubMed: 31821577]

Oelfke U and Bortfeld T 2001 Inverse planning for photon and proton beams Medical dosimetry 26 113–24 [PubMed: 11444513]

Shen C, Gonzalez Y, Klages P, Qin N, Jung H, Chen L, Nguyen D, Jiang SB and Jia X 2019 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 115013 [PubMed: 30978709]

Shen C, Liyuan C, Gonzalez Y and Jia X 2020a Improving Efficiency of Training a Virtual Treatment Planner Network via Knowledge-guided Deep Reinforcement Learning for Intelligent Automatic Treatment Planning of Radiotherapy arXiv:2007.12591

Shen C, Nguyen D, Chen L, Gonzalez Y, McBeth R, Qin N, Jiang S B and Jia X 2020b Operating a treatment planning system using a deep-reinforcement learning-based virtual treatment planner for

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

prostate cancer intensity-modulated radiation therapy treatment planning Med Phys 47 2329–36 [PubMed: 32141086]

Shen C, Nguyen D, Zhou Z, Jiang S B, Dong B and Jia X 2020c An introduction to deep learning in medical physics: advantages, potential, and challenges Physics in Medicine & Biology 65 05TR1

Wahl N, Bangert M, Kamerling CP, Ziegenhein P, Bol GH, Raaymakers BW and Oelfke U 2016 Physically constrained voxel-based penalty adaptation for ultra-fast IMRT planning Journal of Applied Clinical Medical Physics 17 172–89 [PubMed: 27455484]

Wang H, Dong P, Liu H and Xing L 2017 Development of an autonomous treatment planning strategy for radiation therapy with effective use of population-based prior data Medical Physics 44 389–96 [PubMed: 28133746]

Watkins CJ and Dayan P 1992 Q-learning Machine learning 8 279–92

Webb S2003The physical basis of IMRT and inverse planningBritish Journal of Radiology76678–89

Wu X and Zhu Y 2001 An optimization method for importance factors and beam weights based on genetic algorithms for radiotherapy treatment planning Physics in Medicine & Biology 46 1085 [PubMed: 11324953]

Wulfmeier M, Ondruska P and Posner I 2015 Maximum entropy deep inverse reinforcement learning arXiv preprint arXiv:1507.04888

Xing L, Li JG, Donaldson S, Le QT and Boyer AL 1999 Optimization of importance factors in inverse planning Physics in Medicine and Biology 44 2525 [PubMed: 10533926]

Yan H and Yin F-F 2008 Application of distance transformation on parameter optimization of inverse planning in intensity-modulated radiation therapy Journal of Applied Clinical Medical Physics 9 30–45 [PubMed: 18714279]

Yan H, Yin F-F, Guan H-q and Kim JH 2003a AI-guided parameter optimization in inverse treatment planning Physics in Medicine & Biology 48 3565 [PubMed: 14653563]

Yan H, Yin F-F, Guan H and Kim JH 2003b Fuzzy logic guided inverse treatment planning Medical Physics 30 2675–85 [PubMed: 14596304]

Yang Y and Xing L 2004 Inverse treatment planning with adaptively evolving voxel-dependent penalty scheme Medical Physics 31 2839–44 [PubMed: 15543792]

Zhang J, Wang C, Sheng Y, Palta M, Czito B, Willett C, Zhang J, Jensen PJ, Yin F-F and Wu Q 2020 An interpretable planning bot for pancreas stereotactic body radiation therapy arXiv preprint arXiv:2009.07997

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript



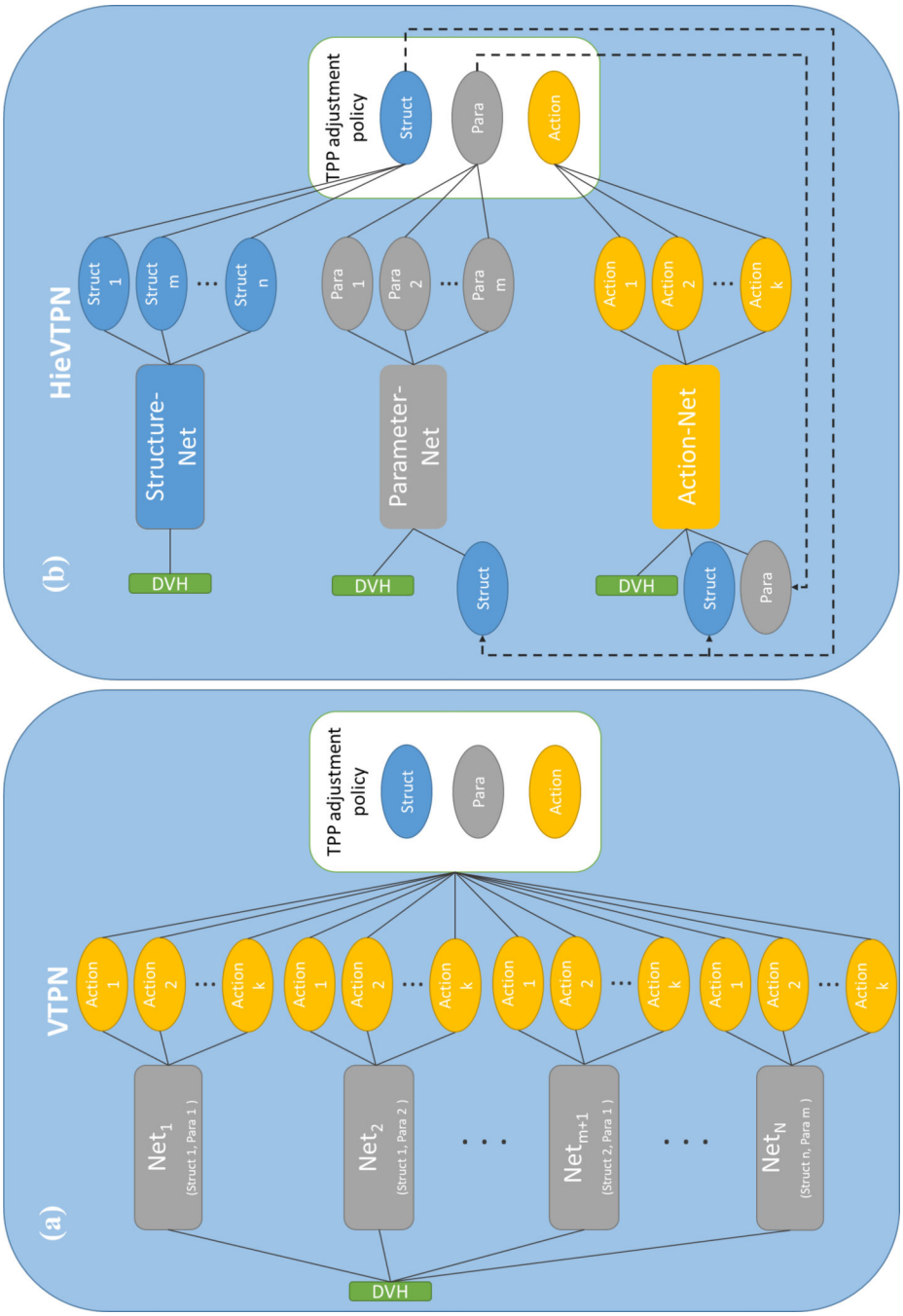


**Figure 1.**
Architecture comparison between VTPN and HieVTPN. (a) VTPN. (b) HieVTPN.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

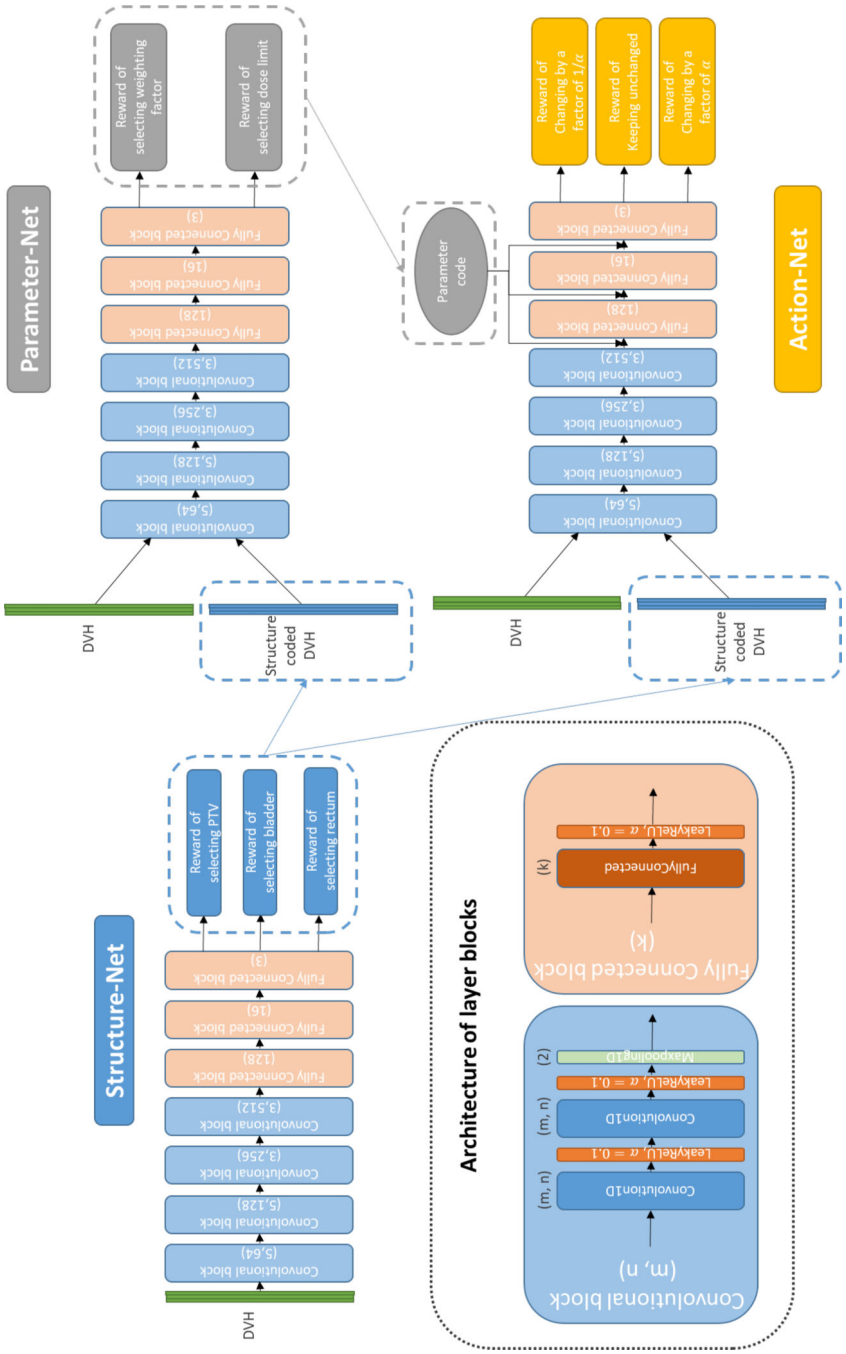


**Figure 2.**
Network structure of the Structure-Net, Parameter-Net, and Action-Net.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

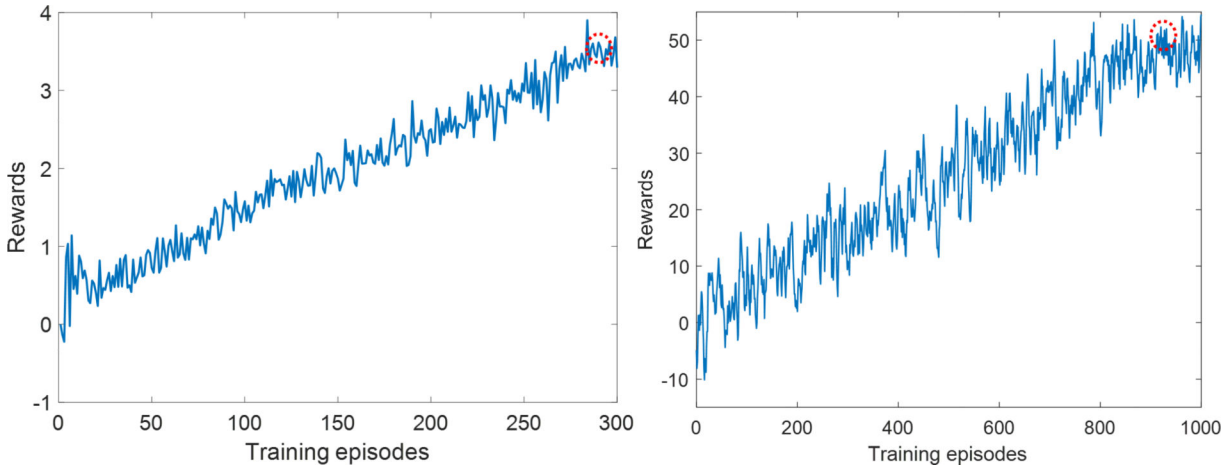


**Figure 3.**

Rewards of HieVTPNs for prostate IMRT (a) and SBRT (b) along training episodes. Red circles highlight the reward of the selected final model established at episode 288 for IMRT and 872 for SBRT.

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript



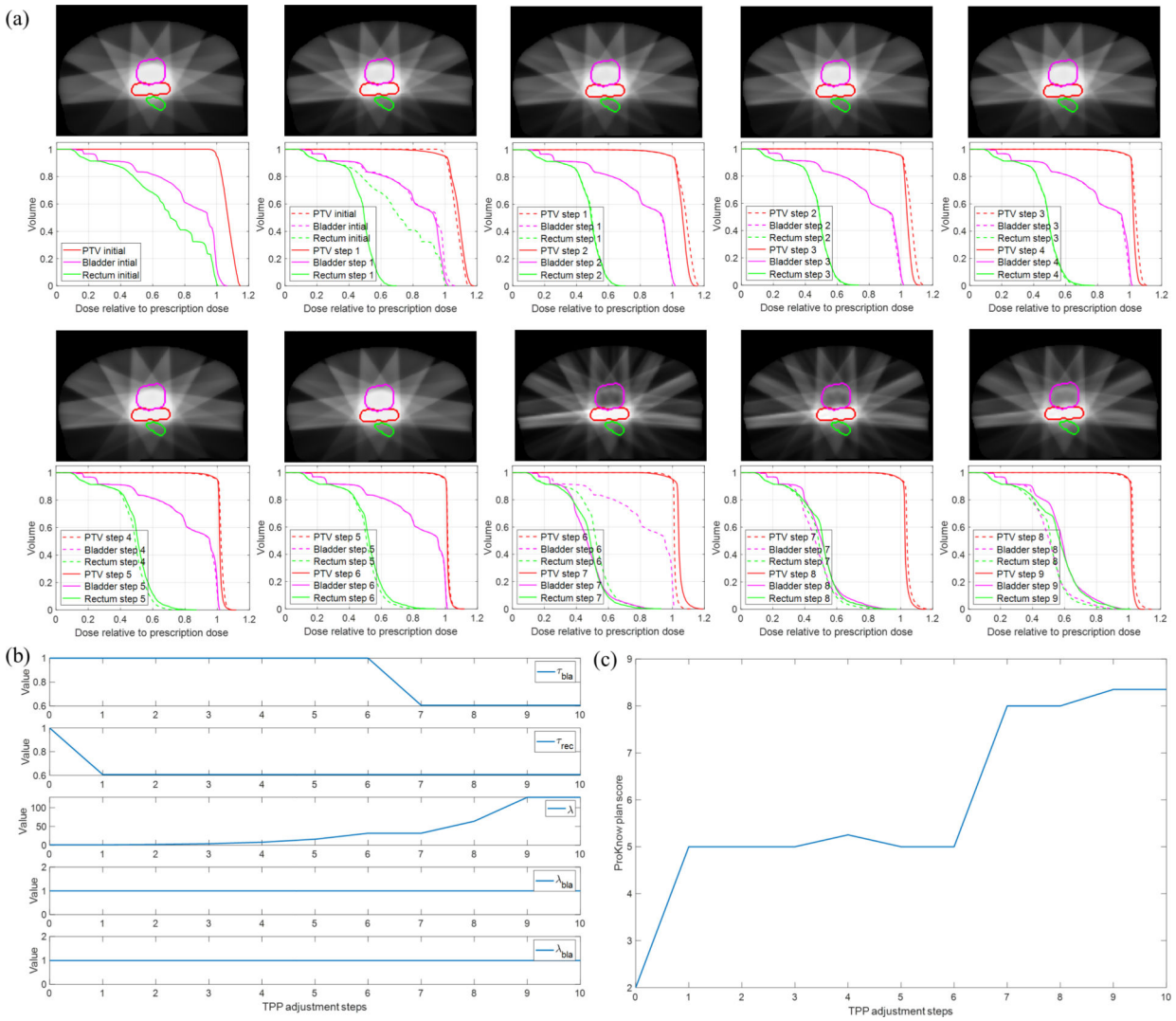


**Figure 4.**

(a) Evolution of DVHs and dose distributions (Top: initial and steps 1–4; Bottom: steps 5 to 9). (b) Evolution of TPP values. (c) Evolution of ProKnow scores in the planning process of a testing patient case.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

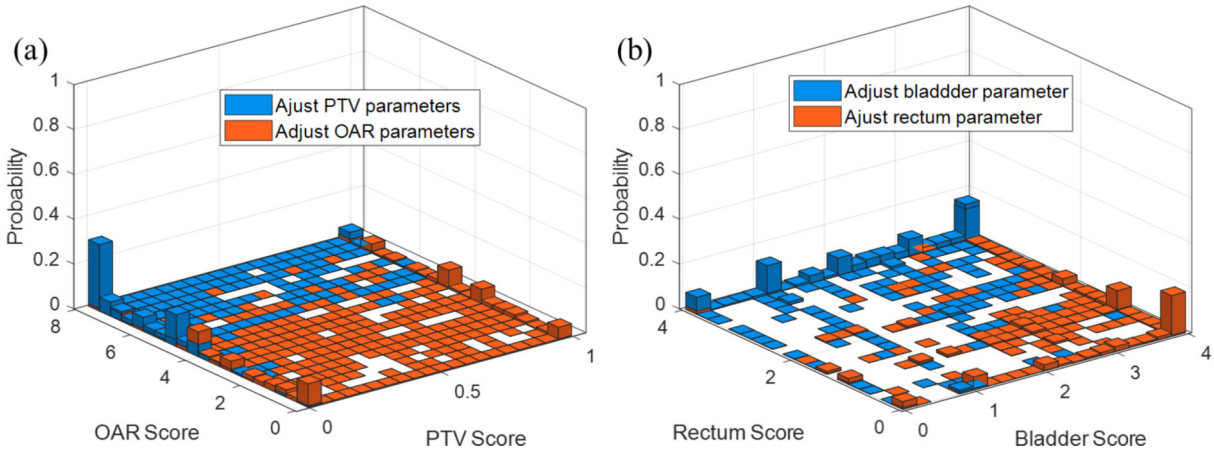


**Figure 5.**

Histograms of decisions made by HieVTPN at structure level. (a) Adjusting PTV parameters vs. adjusting OAR parameters. (b) Adjusting bladder parameters vs. adjusting rectum parameters.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

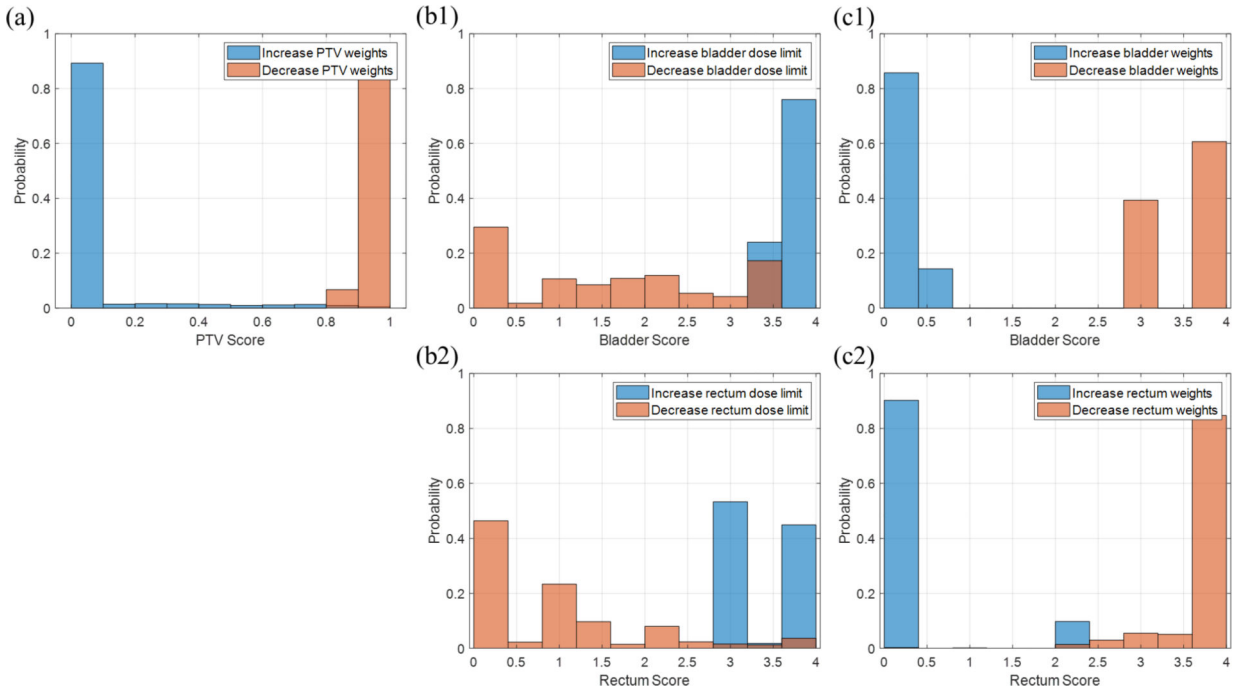


**Figure 6.**

Histograms of decisions made by HieVTPN on action level. Blue histograms: increase parameter values. Red: decrease parameter values. (a) Histograms of the PTV weighting factor. (b) Histograms of the bladder dose limit (b1) and weighting factor (b2), respectively. (c) Histograms of the rectum dose limit (c1) and weighting factor (c2), respectively.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

**Table 1.**

The planning structures and objectives set for prostate cancer SBRT treatment planning.

| **Planning structure** | **PTV** | **Prostate** | **Ring structure** | **Bladder** | **Rectum** | **Urethra** | **Penile bulb** |
|---|---|---|---|---|---|---|---|
| Upper objective | 1 | 0 | 1 | 2 | 2 | 1 | 1 |
| Lower objective | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| Planning structure | Bowel | Left femoral head | Right femoral head | Testes | Neurovascular bundle | Skin | **Total** |
| Upper objective | 1 | 1 | 1 | 1 | 1 | 1 | **14** |
| Lower objective | 0 | 0 | 0 | 0 | 0 | 0 | **2** |

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

**Table 2.**

Average scores (±standard deviation) of plans generated by HieVTPNs.

| | | Plan score | | |
|---|---|---|---|---|
| | | **Initial** | **HieVTPN** | **VTPN** |
| Prostate IMRT | Training | 4.04 (±2.36) | 8.47 (±0.90) | 8.46 (±0.50) |
| | Testing | 4.97 (±2.02) | 8.62 (±0.83) | 8.45 (±0.48) |
| | Overall | 4.84 (±2.07) | 8.60 (±0.84) | 8.45 (±0.48) |
| Prostate SBRT | Training | 97.43(±10.62) | 139.88((±3.19) | - |
| | Testing | 95.56(±9.74) | 139.07(±3.35) | - |
| | Overall | 96.81(±10.02) | 139.61(±3.15) | - |

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

**Table 3.**

Comparisons of total numbers of network parameters in VTPN and HieVTPN when different numbers of TPPs are involved in a treatment planning task

| | **TPPs #** | **3** | **5*** | **30** | **48**** | **60** |
|---|---|---|---|---|---|---|
| Network Parameter # | VTPN | 9,022,473 | 15,037,455 | 90,243,930 | 144,390,288 | 180,583,860 |
| | HieVT PN | 9,024,506 | 9,024,670 | 9,028,560 | 9,033,411 | 9,036,645 |

(* and ** indicate the TPPs setup for prostate IMRT and SBRT, respectively).