Volumetric Modulated Arc Therapy Machine Parameter Optimization for Localized Prostate Cancer Using Tandem Reinforcement Learning Networks

Nathan Shaffer[1] · Avinash Reddy Mudireddy[2] · Joel St-Aubin[1]

[1] University of Iowa Department of Radiation Oncology · [2] University of Iowa Initiative for Artificial Intelligence

2026 JOINT AAPM | COMP MEETING
JULY 19–22 | VANCOUVER, BC

Introduction

- **VMAT machine parameter optimization (MPO)**, or setting MLC positions and monitor units (MUs) at every control point of an arc, is a high-dimensional, constrained problem solved by inverse planning that is sensitive to objective and weight choices, driving long, variable planning times.
- **These limits are acute** in online adaptive and real-time radiotherapy, where plans must be re-optimized rapidly to daily anatomical change.
- **Supervised learning** still relies on a conventional optimizer and is bounded by training-plan quality. Reinforcement learning (RL) instead maximizes a reward, giving it the ability to discover novel solutions and generalize across clinical goals.
- **Prior RL work** either tuned TPS parameters as a "virtual planner" or still invoked a commercial TPS in the final stage to reach clinical quality.

Purpose

- **Develop and validate** a deep-RL VMAT MPO algorithm that automatically generates clinically comparable prostate cancer plans meeting machine constraints while being entirely independent of a commercial TPS optimizer.

Methods & Materials

- **Cohort:** 100 prostate SBRT patients (PACE-B), 36.25 Gy / 5 fx, 5 mm PTV margin, 70/10/20 training / validation / testing split. Elekta Unity MR-Linac VMAT in a research Monaco build; single arc, 2° control points, 4–23° cryostat gap; in-house developed dose calculation[1].
- **MDP:** VMAT MPO cast as a finite Markov decision process (MDP) over N control points; the policy maps state s_t → next machine parameters, earning a DVH-based reward r_t.
- **Tandem CNNs:** a spatial MLC network (per control point, BEV inputs) and a global MU network (all control points). Decoupling MLC / MU exploration stabilized training.
- **Reward:** per-structure mean square error (MSE) to ideal dose, scaled down each time a constraint is met, driving the agent to meet clinical objectives.
- **Training:** Proximal policy optimization (PPO) actor–critic (clip ε = 0.15) + GAE (γ = 0.99, λ = 0.95); imitation-learning warm start; TensorFlow 2.12 on 8× Tesla V100.

joel-st-aubin@uiowa.edu nathan-shaffer@uiowa.edu

[1] Shaffer N, Snyder J, St-Aubin J. Validation of a rapid algorithm for repeated intensity modulated radiation therapy dose calculations. *Biomed Phys Eng Express.* 2024;11(1):015046.

[2] Shaffer N, Mudireddy AR, St-Aubin J. A tandem reinforcement learning framework for localized prostate cancer treatment planning and machine parameter optimization. *Medical Physics.* 2026;53(2):e70306.

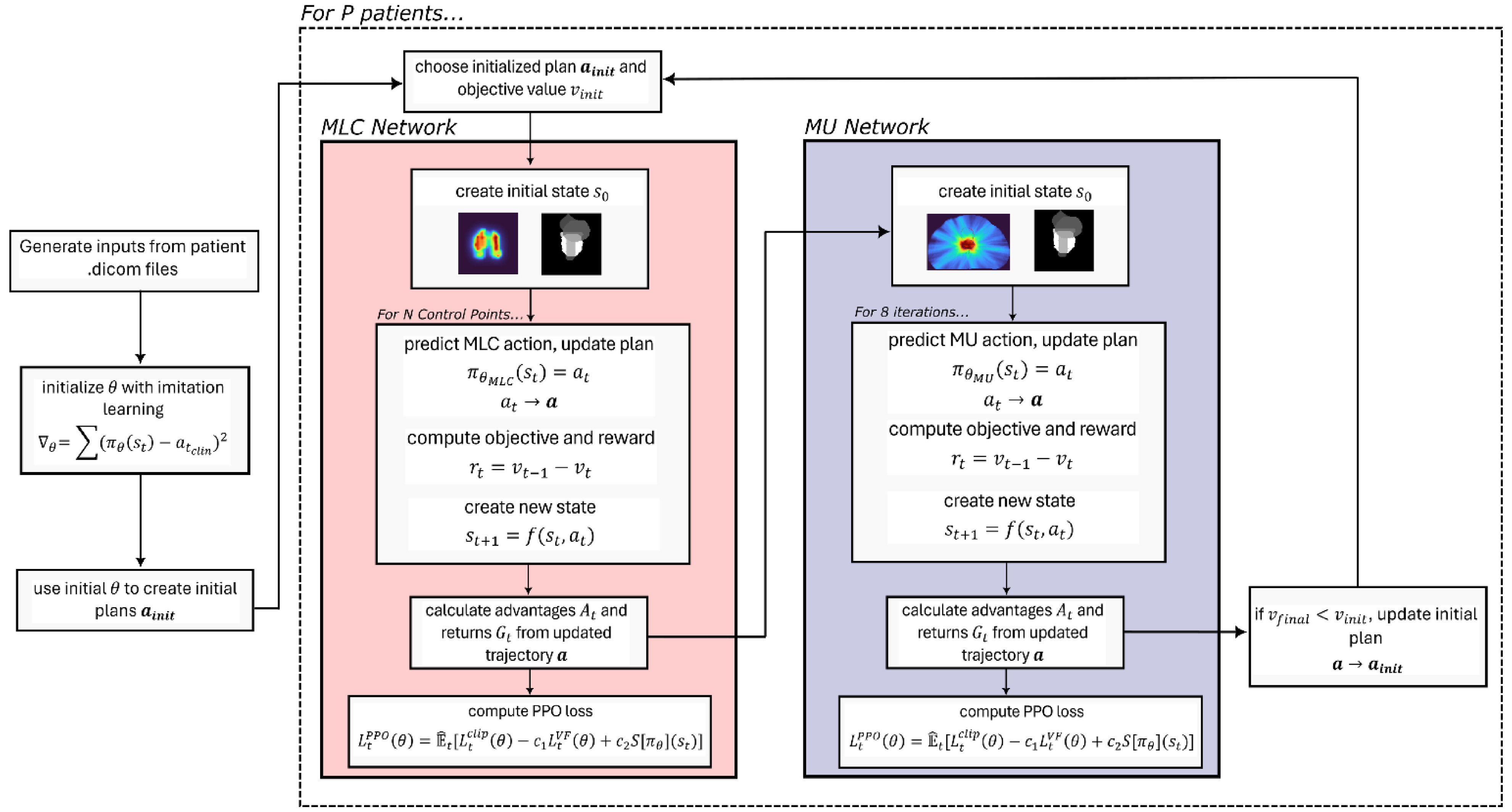

Figure 1. Training workflow with sequential RL updates of the tandem MLC and MU networks.

Results

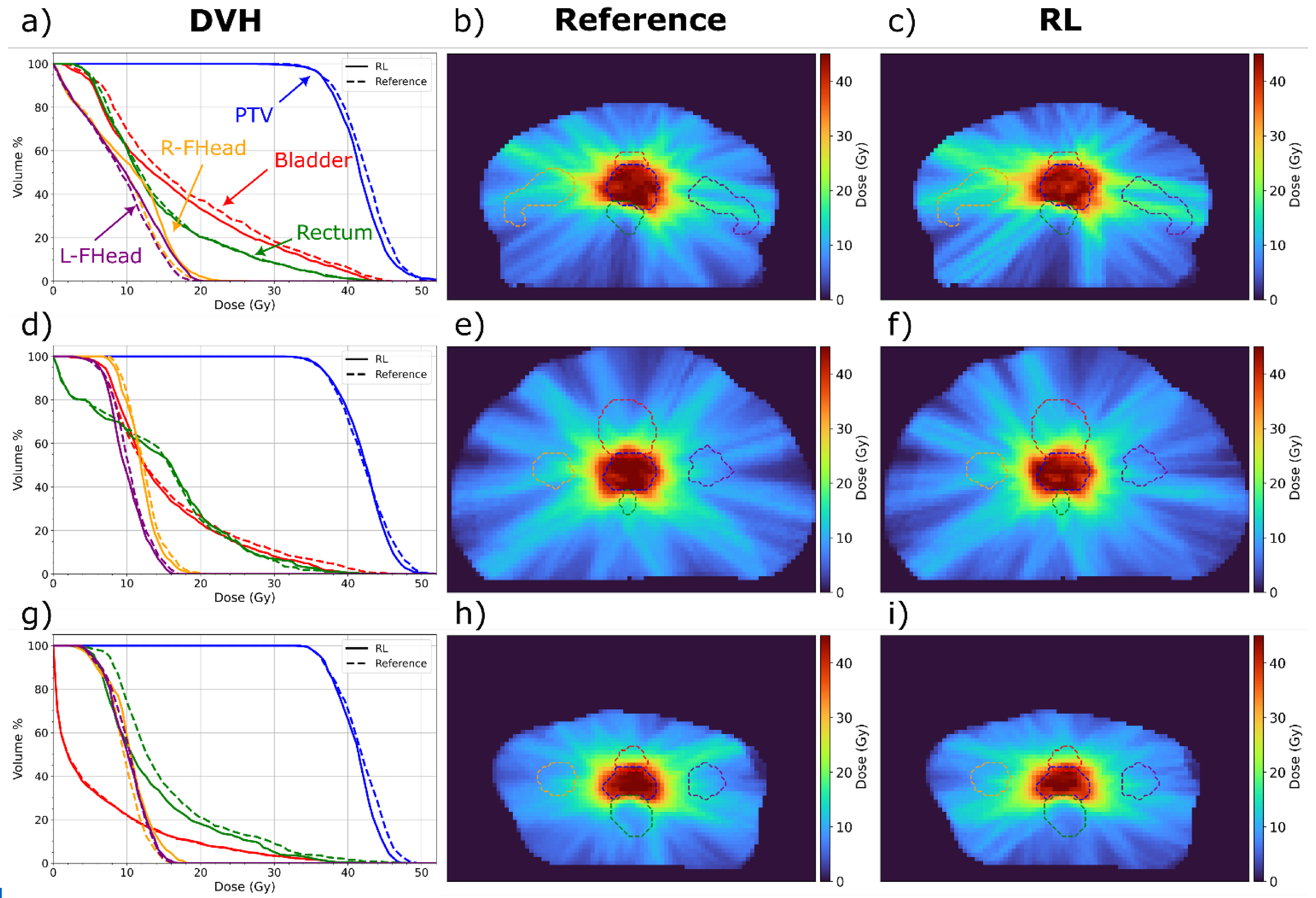

Figure 2. DVHs (a, d, g) with axial reference (b, e, h) and RL (c, f, i) dose distributions for three sample patients. Contours: PTV (blue), bladder (red), rectum (green), R/L femoral heads (orange/purple).

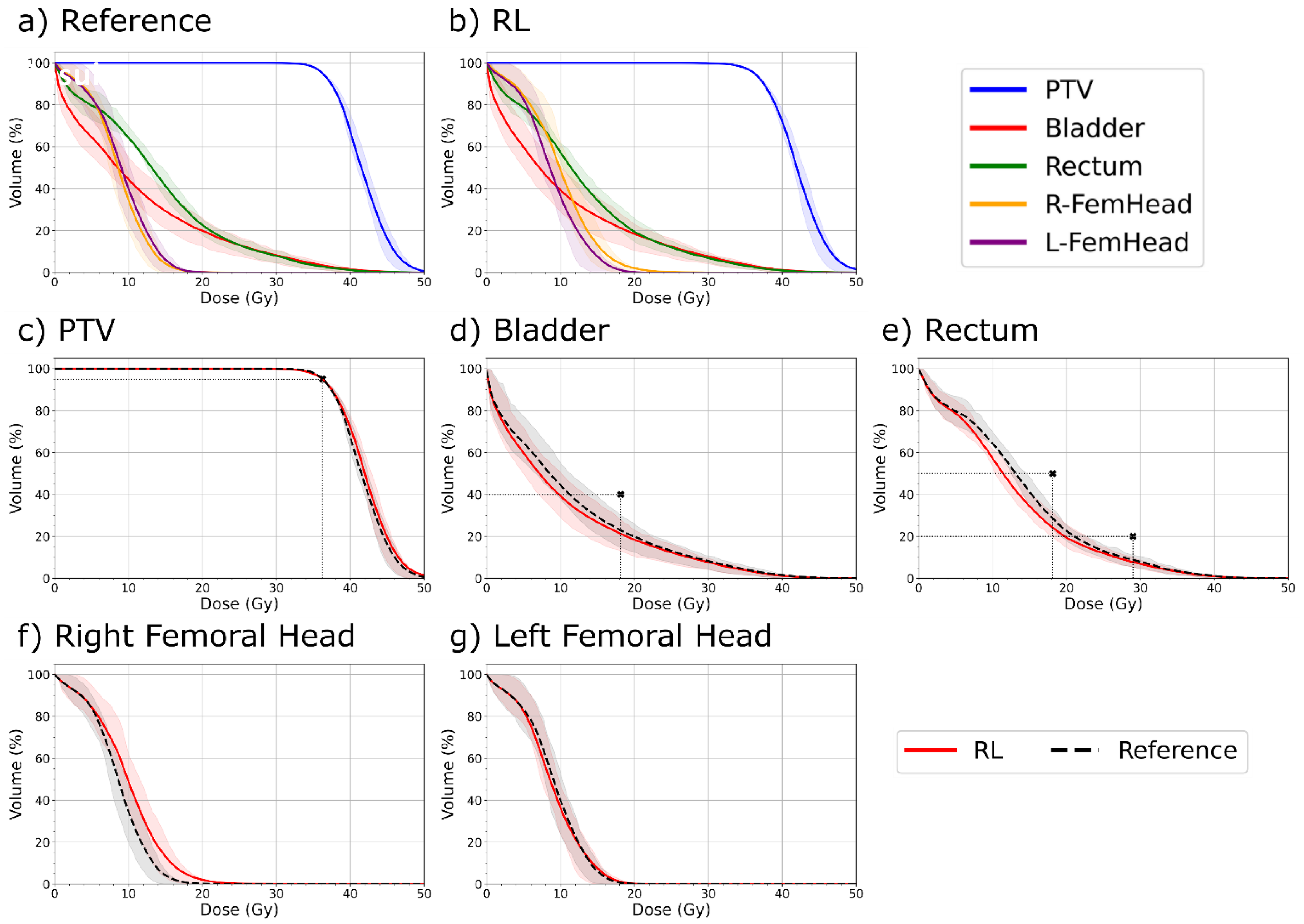

Figure 3. Mean DVH for reference (a) and RL (b) plans, with per-structure curves (c–g). Black markers denote reward constraints; shaded bands show the interquartile range.

Table 1. Mean ± SD dose metrics for reference vs. RL plans (test cohort). Significant differences (p < 0.05) in bold.*

Structure	Metric	Reference	RL	p
PTV	Dmean (Gy)	41.64 ± 1.23	42.01 ± 1.22	0.304
	D2% (Gy)	47.35 ± 2.61	48.10 ± 2.50	**0.048***
	D98% (Gy)	34.94 ± 0.60	34.59 ± 0.67	0.224
	CI	1.40 ± 0.11	1.42 ± 0.07	0.251
Bladder	Dmean (Gy)	11.41 ± 3.99	10.58 ± 3.75	0.417
	V50% (Gy)	8.29 ± 3.79	7.71 ± 3.45	0.607
Rectum	Dmean (Gy)	14.04 ± 2.13	12.96 ± 1.85	**0.020***
	V50% (Gy)	10.41 ± 3.14	8.79 ± 2.74	0.130
	V80% (Gy)	3.18 ± 1.53	2.83 ± 1.32	0.417
L Fem. Head	Dmean (Gy)	8.92 ± 1.65	8.72 ± 1.87	0.626
	V40% (Gy)	2.85 ± 2.48	3.39 ± 3.64	0.892
R Fem. Head	Dmean (Gy)	8.57 ± 1.74	9.87 ± 2.24	**0.045***
	V40% (Gy)	1.97 ± 2.62	5.59 ± 5.11	**0.020***

6.3s
avg. plan generation

100%
of RL plans met all clinical constraints

- **Fast auto-planning:** RL plans were generated in 6.3 ± 4.7 seconds
- **Adaptive inference:** plans used variable amount of inference passes, continuing while total plan quality continued to improve
- **Clinically comparable:** all plans met every pre-defined planning objective
- **Improved OAR sparing:** RL plans had improved bladder and rectum sparing vs. reference plans
- **All plans are physical:** plans are all in agreement with the machine's mechanical limits (425 MU/min, 6 cm/s leaf speed).

Discussion

- **End-to-end RL VMAT MPO** produces clinically acceptable prostate SBRT plans in ~6 s with no commercial optimizer
- **Constraint-driven reward** taught the agent to meet clinical goals without using clinical plans for training; OAR weighting may be slightly over-prioritized vs. PTV homogeneity but is tunable for future iterations of this work.
- **Future Work:** multi-arc plans for higher-risk disease; geometric rewards for avoidance angles; variable dose rate; finer 2–3 mm dose grids; GPU-side pre-processing; DICOM-RT export once Unity import is supported.

Conclusion

- **Built a deep-RL framework** for end-to-end VMAT optimization rapidly generates prostate SBRT plans that are dosimetrically comparable to conventional plans and meet all machine constraints without a commercial TPS optimizer.
- **Its speed, adaptability, and independence** from prior plan quality make it a promising auto-planning tool for adaptive radiotherapy.

*Since submission, this work was published in *Medical Physics* [2]