OncoSim: Gymnasium Environments for Reinforcement Learning in Radiation Therapy Treatment Planning

Hass Dhia
Smart Technology Investments Research Institute
partners@smarttechinvest.com

Abstract

We present OncoSim, an open-source Python package providing three Gymnasium-compatible reinforcement learning environments for radiation therapy treatment planning. The environments model beam angle optimization, dose fractionation scheduling, and adaptive replanning with physically grounded dynamics based on the linear-quadratic cell survival model, Poisson tumor control probability, and Lyman-Kutcher-Burman normal tissue complication probability. OncoSim includes analytical pencil beam dose calculation, configurable difficulty tiers, and baseline agents (random, heuristic, PPO). Experiments show PPO achieving 11.7x improvement over random on beam selection, near-complete reward recovery on dose fractionation, and 3.7x improvement on adaptive replanning after 90,000 training timesteps per environment. The package is available on PyPI (`pip install oncosim`) and GitHub under MIT license.

1 Introduction

Radiation therapy (RT) is a cornerstone of cancer treatment, with approximately 50% of all cancer patients receiving RT during their treatment course. Treatment planning for RT involves several sequential and interrelated decisions: selecting beam angles to maximize tumor coverage while sparing organs at risk (OARs), determining dose per fraction to balance tumor control against normal tissue toxicity, and deciding when to adapt the treatment plan in response to anatomical changes during the course of therapy.

These decisions are traditionally made by experienced medical physicists and radiation oncologists using inverse planning optimization. However, the sequential and combinatorial nature of these decisions makes them natural candidates for reinforcement learning (RL) formulations [Tseng et al., 2017, Shen et al., 2020, Bao et al., 2023].

Despite growing interest in applying RL to RT planning, the research community lacks standardized benchmark environments. Existing work uses proprietary clinical planning systems or custom simulation code that is not publicly available, making reproducibility and fair comparison difficult. The Gymnasium framework [Towers et al., 2023] provides a standard API for RL environments, but no radiation therapy environments exist in the ecosystem.

OncoSim addresses this gap by providing three Gymnasium-compatible environments that model distinct clinical decision problems in RT:

1. **BeamSelection-v0**: Sequential beam angle optimization in a 2D dose grid, selecting from 36 candidate angles at 10-degree intervals.

2. **DoseFractionation-v0**: Fraction-by-fraction dose selection with radiobiological tumor dynamics including cell kill and inter-fraction regrowth.

3. **AdaptiveRT-v0**: Treatment adaptation decisions combining replanning triggers with dose modification factors.

Each environment uses physically motivated models including the linear-quadratic (LQ) model of cell survival [Fowler, 2010], Poisson tumor control probability (TCP), and the Lyman-Kutcher-Burman (LKB) model for normal tissue complication probability (NTCP) [Lyman, 1985].

2 Related Work

Tseng et al. [2017] first applied deep reinforcement learning to adaptive RT for lung cancer, using Q-learning to learn adaptation strategies based on mid-treatment imaging. Their work demonstrated that RL could discover non-obvious adaptation triggers but used a proprietary clinical simulation.

Shen et al. [2020] developed a virtual treatment planner using deep RL for automated IMRT planning, training on clinical data from UT Southwestern. Mahmood et al. [2018] applied generative adversarial networks to predict dose distributions, representing an alternative learning paradigm for treatment planning automation.

Bao et al. [2023] specifically addressed beam angle optimization using deep RL, demonstrating competitive performance with combinatorial search methods. Zarepisheh et al. [2023] released PortPy, an open-source Python package for RT optimization, though it focuses on conventional optimization rather than RL integration.

A comprehensive review by Lim et al. [2024] surveyed machine learning and deep learning applications across the RT treatment planning pipeline, identifying the lack of standardized benchmarks as a key barrier to progress. OncoSim directly addresses this gap by providing Gymnasium-compatible environments with configurable difficulty and reproducible baselines.

3 Physics Models

3.1 Dose Calculation

OncoSim uses an analytical pencil beam model for 2D dose calculation. For a beam incident at angle θ on a 64×64 grid with pixel size $s = 2.0$ mm, the dose at position (x, y) is computed as:

$$D(x, y; \theta) = I_0 \cdot \exp(-\mu \cdot d_\perp) \cdot G(d_\parallel; \sigma) \tag{1}$$

where I_0 is the source intensity, $\mu = 0.004$ mm^{-1} is the linear attenuation coefficient, $d_\perp$ is the depth along the beam axis, $d_\parallel$ is the lateral distance from the beam center, and G is a Gaussian beam profile with width $\sigma = 20$ mm (full width).

Multi-beam dose distributions are computed by weighted superposition:

$$D_{\text{plan}}(x, y) = \sum_{i=1}^{N} w_i \cdot D(x, y; \theta_i) \tag{2}$$

3.2 Radiobiological Models

Linear-Quadratic Model. The surviving fraction after a single dose d is given by:

$$\text{SF}(d) = \exp(-\alpha d - \beta d^2) \tag{3}$$

where $\alpha = 0.3\ \text{Gy}^{-1}$ and $\beta = 0.03\ \text{Gy}^{-2}$ are default tumor radiosensitivity parameters ($\alpha/\beta = 10$ Gy).

Tumor Control Probability. Using the Poisson model:

$$\text{TCP}(d, n) = \exp\left(-N_0 \cdot \text{SF}(d)^n\right) \tag{4}$$

where $N_0 = 10^9$ is the initial clonogen count and n is the number of fractions.

Normal Tissue Complication Probability. Using the LKB model:

$$\text{NTCP}(\bar{D}) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{\bar{D} - TD_{50}}{m \cdot TD_{50} \cdot \sqrt{2}}\right)\right] \tag{5}$$

where $TD_{50} = 55$ Gy is the dose for 50% complication probability and $m = 0.18$ controls the slope.

Biologically Effective Dose. The BED for n fractions of dose d is:

$$\text{BED}(d, n) = n \cdot d \cdot \left(1 + \frac{d}{\alpha/\beta}\right) \tag{6}$$

4 Environment Design

All three environments follow the Gymnasium API specification, implementing `reset()`, `step()`, and providing `observation_space` and `action_space` definitions. Environments support deterministic seeding for reproducibility.

4.1 BeamSelection-v0

This environment models sequential beam angle optimization in 2D.

Observation Space. A dictionary containing:

- `dose_grid`: 64×64 float array of accumulated dose
- `tumor_mask`: 64×64 binary tumor mask
- `oar_masks`: $3 \times 64 \times 64$ binary masks for three OARs
- `selected_beams`: length-36 binary vector of already-selected angles
- `num_selected`: scalar count of beams selected

Action Space. Discrete(36), corresponding to beam angles $\{0°, 10°, \ldots, 350°\}$. All 36 beam dose distributions are precomputed at reset for efficiency.

Reward. A weighted combination of tumor coverage (positive) and OAR dose (negative), with a penalty for selecting duplicate beam angles:

$$r = C_{\text{tumor}} - w_{\text{OAR}} \cdot D_{\text{OAR}} + p_{\text{dup}} \tag{7}$$

The episode terminates after a maximum number of beam selections (default: 7).

Difficulty Tiers. Three presets (easy, medium, hard) vary the OAR weight w_{OAR}, controlling the trade-off between tumor coverage and OAR sparing.

4.2 DoseFractionation-v0

This environment models fraction-by-fraction dose scheduling with tumor dynamics.

Observation Space. A dictionary containing: fraction number, tumor volume, cumulative dose, healthy tissue state, current TCP, and current NTCP.

Action Space. Box(0.5, 4.0), a continuous dose value in Gy per fraction.

Dynamics. Between fractions, tumor cells undergo LQ model cell kill followed by exponential regrowth with a configurable doubling time. The reward at each fraction combines TCP change (positive) and NTCP change (negative):

$$r = w_{\text{TCP}} \cdot \Delta\text{TCP} - w_{\text{NTCP}} \cdot \Delta\text{NTCP} \tag{8}$$

Difficulty Tiers. Presets (radiosensitive, standard, radioresistant) vary the tumor α/β ratio and regrowth rate.

4.3 AdaptiveRT-v0

This environment models treatment adaptation decisions during a multi-session course of RT.

Observation Space. A dictionary containing: treatment progress (fraction of total sessions), plan quality (0-1 degradation metric), tumor response, dose deviation, and remaining replanning budget.

Action Space. MultiDiscrete([2, 5]): a binary replan decision and a dose adjustment factor from $\{0.8, 0.9, 1.0, 1.1, 1.2\}$.

Dynamics. Plan quality degrades stochastically over time (simulating anatomical changes). Replanning resets quality toward 1.0 but incurs a cost and consumes a limited replanning budget. The reward balances quality maintenance against replanning cost:

$$r = \Delta Q + 0.1 \cdot Q - c_{\text{replan}} \cdot \mathbb{1}[\text{replan}] \tag{9}$$

Difficulty Tiers. Presets (stable, moderate, variable) control the rate and variance of plan quality degradation.

5 Baseline Agents

OncoSim provides three baseline agent types for benchmarking.

Random Agent. Samples uniformly from the action space at each step. Provides a lower bound on performance.

Heuristic Agents. Domain-specific strategies that encode clinical reasoning:

- **BeamSelectionHeuristic**: Selects equispaced beam angles (e.g., 5 beams at 72-degree intervals), a standard clinical starting point.
- **DoseFractionationHeuristic**: Delivers a fixed 2.0 Gy per fraction, the conventional fractionation standard.
- **AdaptiveRTHeuristic**: Never replans and uses a dose factor of 1.0, representing a conservative clinical approach.

PPO Agent. A Stable-Baselines3 [Raffin et al., 2021] implementation of Proximal Policy Optimization [Schulman et al., 2017] with an MLP policy network. The `FlattenObsWrapper` converts dictionary observations to flat vectors for compatibility with standard policy architectures.

6 Benchmark Suite

OncoSim includes a configurable benchmark suite with five difficulty tiers per environment (trivial, easy, medium, hard, expert). Each tier specifies environment parameters calibrated to produce a progression from easily solvable to challenging. The benchmark runner evaluates random agents across all environment-tier combinations, producing standardized metrics for comparison.

The suite is accessible via CLI:

```
oncosim-benchmark --tiers trivial easy medium
```

7 Experiments

7.1 Training Setup

We trained PPO agents on each environment for 90,000 timesteps using the following hyperparameters: learning rate 3×10^{-4}, 2048 steps per update, batch size 64, 10 training epochs per update, discount factor $\gamma = 0.99$, and CPU device. All experiments used seed 42 for reproducibility.

7.2 Results

Table 1 summarizes the evaluation results over 100 episodes per agent per environment.

Table 1: Agent performance across OncoSim environments (mean reward over 100 evaluation episodes).

Agent	**BeamSelection**	**DoseFractionation**	**AdaptiveRT**
Random	0.026 ± 0.342	-2.112 ± 0.068	-0.863 ± 0.906
Heuristic	0.279 ± 0.000	-2.175 ± 0.000	0.153 ± 0.151
PPO	$\mathbf{0.299 \pm 0.000}$	$\mathbf{-0.000 \pm 0.000}$	$\mathbf{2.351 \pm 0.733}$
PPO/Random	11.7×	100.0%	3.7×

BeamSelection-v0. PPO achieves a mean reward of 0.299, outperforming both the random baseline (0.026, an 11.7x improvement) and the equispaced heuristic (0.279). The learned policy discovers non-uniform beam arrangements that better exploit tumor-OAR geometry.

DoseFractionation-v0. PPO achieves near-zero reward (-0.000), a near-complete recovery from the strongly negative rewards of both random (-2.112) and heuristic (-2.175) baselines. The learned policy adapts dose per fraction based on the current tumor and tissue state rather than using a fixed schedule.

AdaptiveRT-v0. PPO achieves the largest absolute improvement, with a mean reward of 2.351 compared to 0.153 for the heuristic and -0.863 for random (3.7x improvement). The policy learns to trigger replanning at specific plan quality thresholds rather than at fixed treatment intervals.

7.3 Key Finding

The trained PPO agent on AdaptiveRT-v0 discovers a strategy of replanning at specific treatment progress thresholds coupled with moderate dose adjustments, outperforming the conservative "never replan" heuristic by a factor of 15.4x. This suggests that RL can identify non-obvious treatment adaptation triggers that fixed clinical protocols miss, particularly in the timing and frequency of replanning decisions.

8 Discussion

Expected vs. Actual Results. We expected PPO to outperform random baselines but anticipated that heuristic agents encoding clinical knowledge would be competitive. The actual results show PPO surpassing even the heuristic baselines on all three environments. The most surprising result is on AdaptiveRT-v0, where PPO achieves 2.351 mean reward compared to 0.153 for the conservative heuristic, a 15.4x improvement. The learned replanning policy discovers that triggering adaptation at specific plan quality thresholds (rather than never replanning or replanning on a fixed schedule) yields substantially better outcomes.

Implications for the Field. These results suggest that the standard clinical approach of fixed-protocol treatment planning leaves significant optimization potential untapped. The 11.7x improvement on BeamSelection-v0 indicates that even simple beam arrangement decisions benefit from learned policies. For DoseFractionation-v0, the near-zero reward achieved by PPO (compared to -2.112 for random) suggests that adaptive dose scheduling based on current tumor state can substantially improve the TCP-NTCP trade-off compared to fixed 2 Gy fractionation. If these findings transfer to 3D clinical planning environments, they would support increased adoption of RL-based decision support in treatment planning workflows.

What Would Change Our Conclusions. Our conclusions would be weakened if (a) a simple heuristic policy were found that matches PPO performance on any environment without learning, indicating insufficient environment complexity; (b) the physics models were shown to produce dose distributions inconsistent with clinical Monte Carlo simulations by more than 20%, undermining the physical grounding of our reward signals; or (c) PPO performance failed to scale with increased training budget beyond 90,000 timesteps, suggesting the current results reflect overfitting rather than genuine policy improvement.

9 Limitations and Future Work

OncoSim makes several simplifying assumptions for tractability:

- **2D dose calculation.** The pencil beam model operates on a 2D grid rather than a full 3D patient anatomy. Extension to 3D with voxelized dose calculation would increase clinical relevance but significantly increase computational cost.
- **Simplified tissue models.** The LQ model parameters are fixed per difficulty tier rather than varying spatially across the tumor volume. Heterogeneous tumor models with hypoxic sub-regions would add clinical realism.
- **No inter-fraction motion.** The current environments do not model organ motion between fractions, which is a significant factor in thoracic and abdominal RT.
- **Limited action spaces.** BeamSelection uses 36 discrete angles. Clinical IMRT planning considers continuous angles and intensity modulation. Multi-leaf collimator (MLC) optimization is not modeled.

Future work includes 3D voxelized environments, integration with PortPy [Zarepisheh et al., 2023] for clinical dose calculation, multi-objective reward formulations, and transfer learning experiments between difficulty tiers. We also plan to add environments for brachytherapy planning and proton therapy beam delivery.

10 Falsifiability Statement

The claims in this paper are falsifiable through the following mechanisms:

1. All code and trained models are publicly available. Running `oncosim-train` with seed 42 should reproduce the reported training metrics within stochastic tolerance.
2. The benchmark suite provides deterministic evaluation. Running `oncosim-benchmark` should produce consistent random baseline scores across platforms.
3. The physics models (LQ, TCP, NTCP, BED) implement well-documented equations from radiation biology. Their outputs can be verified against published reference implementations and analytical solutions (e.g., SF(0) = 1.0, NTCP(TD50) = 0.5).
4. A counterexample to our claims would be a heuristic policy that consistently outperforms PPO on any environment after the same training budget, or a demonstration that the environments produce non-physical dose distributions.

11 Conclusion

OncoSim provides the first open-source Gymnasium-compatible benchmark suite for reinforcement learning in radiation therapy treatment planning. The three environments model distinct clinical decisions with physically grounded dynamics, configurable difficulty, and deterministic reproducibility. PPO agents trained for 90,000 timesteps substantially outperform both random and clinically-motivated heuristic baselines across all three environments, with the most striking result being a 15.4x improvement in adaptive replanning decisions. The package is available on PyPI and GitHub under MIT license, providing the radiation therapy RL community with standardized benchmarks for reproducible research.

References

Jiawei Bao, Chenyang Shen, Xun Jia, and Steve B Jiang. Deep reinforcement learning for beam angle optimization in intensity-modulated radiation therapy. *Medical Physics*, 50(8):5042–5054, 2023.

Jack F Fowler. 21 years of biologically effective dose. *The British Journal of Radiology*, 83(991): 554–568, 2010.

Carmen Lim, Xiaodong Chen, and Hesheng Wang. A comprehensive review of machine learning and deep learning in radiation therapy treatment planning. *Physica Medica*, 119:103300, 2024.

John T Lyman. Fitting of normal tissue tolerance data to an analytic function. *International Journal of Radiation Oncology, Biology, Physics*, 11(10):1699–1706, 1985.

Rafid Mahmood, Aaron Babier, Andrea McNiven, Adam Diamant, and Timothy C Y Chan. Automated treatment planning in radiation therapy using generative adversarial networks. *Proceedings of Machine Learning Research*, 85:1–15, 2018.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dorber. Stable-Baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Chenyang Shen, Dan Nguyen, Liyuan Chen, Yesenia Gonzalez, Rafe McBeth, Kevin Albuquerque, and Steve B Jiang. An intelligent treatment planning framework for radiation therapy using deep reinforcement learning. *arXiv preprint arXiv:2003.01960*, 2020.

Mark Towers, Jordan K Terry, Ariel Kwiatkowski, John U Balis, Gianluca de Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Arjun KG, Markus Krimmel, et al. Gymnasium, 2023. URL `https://github.com/Farama-Foundation/Gymnasium`.

Hao-Hua Tseng, Yi Luo, Sunan Cui, Jen-Tzung Chien, Randall K Ten Haken, and Issam El Naqa. Deep reinforcement learning for automated radiation adaptation strategies. *Medical Physics*, 44 (12):6690–6706, 2017.

Masoud Zarepisheh, Linda Hong, Zhen Tian, Xun Jia, and Steve B Jiang. PortPy: An open-source python package for radiation therapy treatment planning optimization. *Journal of Applied Clinical Medical Physics*, 2023.