Gamification of Proton Beam Therapy Treatment Planning

Hayden Tibbals
10624576
School of Physics and Astronomy
The University of Manchester
MPhys Report Semester 2
May 2024

This project was performed in collaboration with *Oliver Lloyd - 10630429*

Abstract

Proton beam therapy is a form of cancer radiotherapy that utilises the sharp dose deposition of energised protons to accurately treat tumour cells within a patient and spare normal tissue. The joint complexity of proton dose patterns and radiotherapy makes the process of creating a patient treatment plan difficult and time consuming.

This project aims to tackle this problem by creating an AI model using the Proximal Policy Optimisation reinforcement learning algorithm that can design near fully optimised treatment plans. This can be done by gamifying the task and training the agent with positive or negative reinforcement, through a reward system.

The key results of this semester are within the environment setup, containing a fast analytical dose model, computing in 3.55 ± 0.47ms, with a χ^2 of 0.11, and an accurate representation of patient data. A hyperparameter test is also conducted, to identify the optimal set of algorithm parameters to train with.

1 Introduction

Photon radiation therapy is one of the most prominent modern cancer treatments, serving 50% of cancer patients and contributing to the care of 29% of cancer survivors. While advancements in planning and treatment techniques have improved its efficiency and minimised collateral patient damage, the inherent nature of photon dose distribution presents limitations for tumours in close proximity to Organs at Risk (OAR). Proton Beam Therapy (PBT) emerges as a promising alternative, addressing this challenge more directly. With heavy charged particles exhibiting a sharp dose deposition peak (Bragg Peak), PBT offers a more precise dosage delivery with reduced peripheral tissue damage. These peaks can be combined to create a flat conformal dose, called a spread out bragg peak (SOBP), which provides tangible clinical advantages, particularly in treating anatomically complex tumours and pediatric patients, who face heightened risks of secondary tumours.

Despite these benefits of PBT over photon radiotherapy, it is not a direct replacement. The cost of building a new proton beam facility can exceed £125 million (1), compared to £6 million for radiotherapy and individual patient costs are £45,000 to £3,672 (2). These high costs and fewer treatment rooms force PBT to be reserved for patients with the most to gain over photon therapy.

Another drawback of PBT is its sensitivity to uncertainties, which if handled incorrectly severely undermine its benefits. Unlike photon radiotherapy the dose is very localised, meaning small fluctuations in distance can heavily influence the dose pattern supplied. The uncertainties stem from both calculations within the plan and potential movement of the patient during treatment. To mitigate these effects stringent plans, accounting for variations in patient position and bodily composition are required (3). This creates a lengthy process full of iterative changes to ensure each patient has a plan which doses all tumours appropriately while preserving OAR from potential overdoses. Consequently, there is a heavy demand for Medical Physicists and treatment planners, who can each spend up to a week working on an individual treatment plan. A process capable of creating an almost complete treatment plan, with patient scan data as input, is therefore in demand. It would have the potential to free up valuable healthcare resources, increase the availability of PBT treatment and potentially improve the quality of treatment.

The solution to this problem appears to be within the new age of healthcare emerging. AI has begun to take part in most aspects of a patient's treatment path, from diagnosis to recovery and is producing tools for healthcare providers to "facilitate and enhance human work".(4) Currently its largest and most successful application is within radiology, with 53% of all AI and ML devices being "marked for Radiological use".(5) Studies on these applications have shown the tools' abilities to meet or exceed the performance of experts in image-based detection of diabetic retinopathy (6), pnuemonia(7) and appendicitus(8). Particularly, convolutional neural networks have outperformed professionals at detecting pneumonia with labelled frontal X-rays (9) and diagnosed heart attacks with a performance similar to cardiologists (10).

These successes place AI, particularly reinforcement learning (RL) with neural networks, as a candidate for solving the tedious process of creating PBT treatment plans. Creating a plan involves working with labelled CT scans of a patient, in a similar style to current success in the field, suggesting it is a sensible direction to take. RL is a type of machine learning which teaches an agent to solve a particular task by gamifying it with a set of actions it can take and rewarding or penalising specific choices, such that it learns the optimal way to solve a problem, by maximising its reward.

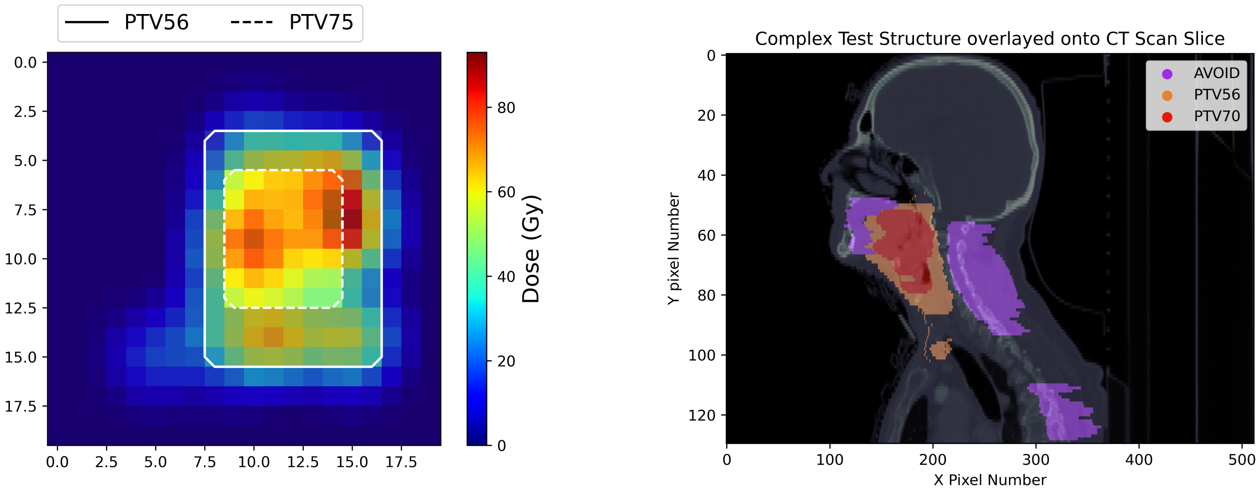
This project aims to follow the field's trend of using AI to recognise patterns in patient data, by using a performance metric to train an RL AI agent which can design near-usable treatment plans in a fraction of the time currently required.

This report will detail the physics related to Proton Beam Therapy and how this can be integrated into our model, along with an explanation of our setup to train an AI agent using reinforcement learning and how this project can progress further. It follows our past semester's work and the previous work of MPhys students Huw Mcnally and Robert Chambers, whose results are saved in a public GitHub PBT-GYM (11).

This project was carried out alongside partner Oliver Lloyd. Most sections were collaborative, but Oliver specialised in reinforcement learning, while I focused on data processing and the beam model.

1.1 Previous Work

The preliminary work in the previous semester of this project focused on understanding the best approach to training an agent to dose a tumour and how patient data can be utilised to create accurate environments. This was done within a 'toy environment', which served as a mock test environment and allowed experimentation of different RL algorithms and action sets on a small fast scale before applying techniques to more complex and computationally intensive systems. An example of a 'toy' and complex environment is shown in Figure 1



(a) Simple environment displaying a two leveled 8 x 12 tumour within a 20 x 20 grid. The dose applied is from a model trained for 200,000 steps.

(b) Complex environment with no dose applied, mapped onto a CT scan. The avoid regions consist of the brain stem, spinal cord and oral cavity

Figure 1: Two types of environments used in the first semester to experiment with training processes and dose models. The PTV (Planned Tumour Volume) define the regions with tumours and the dose required for treatment in Gy. (12)

Key takeaways were that policy gradient algorithms such as PPO were the most effective, the array of actions the agent can select should be as simple as possible and the need for external metrics to validate performance.

The largest limitations were the lack of a sophisticated reward system that could deter the agent from overdosing patient tumours and an accurate dose model. The previous dose was a simple Gaussian mask, which crucially had no dose tail, preventing an understanding of how the agent would deal with a Bragg peak beam.

2 Theory

2.1 Interactions of radiation with matter

The benefits of proton beam therapy stem from the distinct interactions of heavy charged particles (protons) with matter, differing from photon interactions. Photons immediately reach their dose peak when entering matter, decreasing with penetration depth. In contrast, protons initially deliver a consistently low dose, gradually depositing more until they reach a sharp dose peak (known as the Bragg peak), concentrating most of their energy in one localised area. This allows the majority of the dose to be absorbed by the tumour by placing the peak over it, minimising absorption in tissue before and eradicating any past the tumour region.

Protons interact with matter in three ways: *stopping* due to collisions with atomic electrons, *scattering* due to deflections from atomic nuclei and *nuclear interactions* from ricochet with atomic nuclei.(13) The first two interactions are most influential and occur through the electromagnetic force, while nuclear interactions are more infrequent and assumptions in calculations of proton paths can absolve considerations. Therefore to model a proton beam's stopping power through a patient, scattering and stopping are the main considerations. These interactions result in a loss of proton kinetic energy, either distributed to surrounding material or carried away by neutrons and γrays.

2.1.1 Dose Rate and Bragg Peak

To calculate energy deposition by a proton beam within a patient the energy loss of individual particles is required. The mass stopping power for individual particles is required;

$$\frac{S}{\rho} \equiv -\frac{1}{\rho}\frac{dE}{dx} \quad (\frac{MeV}{g/cm^2}), \tag{1}$$

describing the energy loss dE (MeV) per distance dx (cm) of local density ρ (g/cm^3) as the proton traverses matter. Including the fluence of beam ϕ, which describes the number of protons passing through each infinitesimal area dA, the total energy deposition (Dose) can be calculated by;

$$D = \phi \frac{S}{\rho} \tag{2}$$

in units of MeV/cm, which are commonly converted to J/kg in medical physics, also known as Gray (Gy).

The slowing of protons due to atomic electrons is the defining factor in calculating the stopping power of a proton within a patient. It can be described by the *continuous slowing down approximation* (CSDA) developed by Bethe Bloche in 1933 for fast-charged particles (14). While this approximation was derived for electrons previous papers have determined corrections are negligible for protons in the therapeutic treatment energy range (3 - 300MeV). (15)(16)

For a material of constant atomic mass A, mass stopping power S/ρ[MeV/gcm^{-2}] is approximated by:

$$\frac{S}{\rho} = \frac{1}{\rho}\frac{dE}{dx} = 0.3072\frac{Z}{A}\frac{1}{\beta^2}(ln\frac{W_m}{I} - \beta^2) \quad MeVg^{-1}cm^2 \tag{3}$$

where

$$\beta = \frac{v}{c}, \quad \text{and} \quad W_m = \frac{2m_e c^2 \beta^2}{1-\beta^2} \quad \text{and} \quad 0.3072MeVg^{-}1cm^2 = 4\pi N_A r_e^2 m_e c^2. \tag{4}$$

In this equation v [cms^{-1}] is velocity, Z is the material atomic number and W_m [eV] is the energy loss of a head on collision between a proton and electron (maximum energy loss). I [eV] is the *mean excitation energy* of the material, which cannot be directly calculated and instead fitted with data (13).

Together this equation reveals dose deposition from individual protons has four dependencies; inverse of velocity, charge of the material ions, inverse of mean excitation energy and material density. Though when I is fitted it is approximately proportional to Z (16), canceling out that relation leaving dose deposition solely dependent on $1/v^2$ and ρ. As a result material density becomes the dominant external factor influencing proton energy loss.

Throughout a patient density varies massively, often with changes on the order of 60, making proton paths inconsistent between patients and even angles of entry on the same patient. The result is areas of high density, such as the mandible causing considerable uncertainty to the path of the beam, with small unexpected variations of dense material, altering the position of the Bragg peak by a significant distance.

The final $1/v^2$ dependency is the cause of the Bragg peak observed for a proton beam. As velocity decreases the deposition will be larger, due to a larger interaction time with each electron, increasing energy lost with a quadratic dependency and forming a sharp peak. An example beam is depicted in Figure 2, describing the source of specific features of the beam. The primary contribution is electromagnetic interactions with atomic electrons but atomic interactions contribute to the initial build-up and tail energy loss. The width of the peak is due to variations in proton energies leaving the device.

The peripheral spread of this distribution is caused by scattering from atomic nuclei. Continuous deflections create a jagged proton path, which for enough particles can be modeled by a Gaussian distribution. The theoretical equation of this spread is:

$$D(r,\phi;r_0)r\,dr\,d\phi = D_0\frac{1}{2\pi r_0^2}e^{-\frac{1}{2}\left(\frac{r}{r_0}\right)^2} r\,dr\,d\phi \tag{5}$$

where D_0 is the dose deposition [MeV/g] at the central axis and r [cm] is the distance perpendicular to the axis. The r_0 [cm] is a parameter defined by Molière theory (17) with $r_0 = L\theta_0$, where L [cm] is the distance along the beam and θ_0 is a theoretical spread angle. Details of this derivation are in Appendix B but for this project, the required understanding is that the spread follows a Gaussian distribution of which the parameters can be fitted using data.

2.1.2 Biological Interactions of Protons

Despite dose distributions following a different pattern, the therapeutic effects of protons and photons are comparable, with damage to genetic material disrupting the cell cycle and causing cell death.Damage to the DNA occurs from two key pathways, direct deposition of energy in the DNA and chemical reactions through reactive free radicals from ionisation in water (13). Direct structural damage to the DNA is uncommon and the primary interaction occurs in chemical processes.

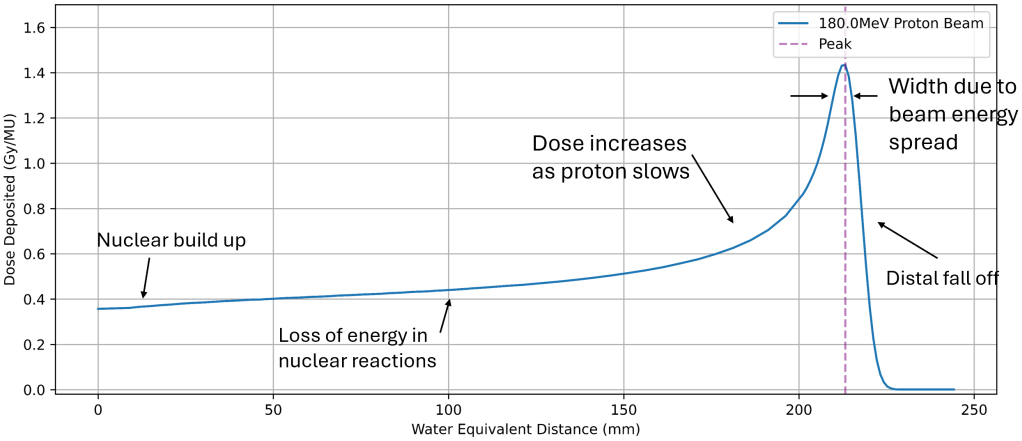


Figure 2: Labeled proton dose deposition distribution per water equivalent thickness within a material. The sharp Bragg peak is a result of a proton's stopping power dependence on $1/v^2$.

Radiation interactions with water release free radicals such as hydroxyl (•OH) which react with DNA bases to cause the lesions to DNA strands (18). These breaks can take form of double or single strand breaks, double the least common but causing the most damage (19). Once lesions are formed the cells undergo repair mechanisms, but if not fully repaired (more likely in double strands) it leads to cellular dysfunction and lost of genetic material which will results in cell death (no reproductive capability).

These interactions are consistent between radiation types, but the effectiveness is not, with heavy ions like protons and carbon ions causing biological damage more effectively. This presents issues in comparing the dose required between different radiations, which is solved by assigning each radiation an effectiveness variable. Relative Biological effectiveness (RBE) is the ratio of the absorbed dose of reference radiation (^{60}Co gamma rays) and different radiation to attain the same biological effect (20). PBT radiation clinically has an RBE of 1.1 throughout the Bragg peak, signalling increased efficiency at killing cells than traditional photon therapy. A primary component of the calculation of this value is the Linear Energy Transfer (LET) describing energy loss along the path of the proton (21), yet throughout the SOBP, LET increases. The result is recorded RBE values of "~1.1 in the entrance, to ~1.2 in the center, ~1.4 at the distal edge, and ~1.7 in the distal fall-off the SOBP"(22). Consequently, this deviation in the values and adoption of 1.1 in clinical use increases uncertainty in dose delivered to specific locations, particularly at the distal fall-off which will likely be in close proximity to a vulnerable region.

2.2 Treatment Planning

Treatment planning involves creating the exact plan for beam locations and energies required to fully treat a tumour. It is designed using specialist software, with patient CT scans, serving as the base to plan on. Before planning begins CT scans of the patient must be annotated, creating contours for specific regions. These are done by a specialist radiologist using the information in the CT scan and in cases of ambiguity, additional MRI scan data. Regions of interest (ROI) will range from spinal cord and brain stem (regions to avoid) to the different extents of the tumour, assigning different cancer regions dose levels, anticipated to completely kill it and mitigate risk of secondary malignancies.

Given the breadth of uncertainties within PBT, tumours have different associated contours. Commonly each CT will be annotated with a Gross Tumour Volume (GTV), Clinical Tumour Volume (CTV) and Planned Treatment Volume (PTV). GTV describes the exact contour of the tumour, encompassing only the cancerous regions. Yet time differences between imaging and treatment, allow for a spread which cannot be accurately imaged and accounted for. To combat any potential spread a CTV is drawn, which adds a margin onto a GTV to account for any spread. Therefore the CTV is the suitable region to place dose on and for photon radiotherapy with less influential uncertainties may be sufficient (Usually will still have a PTV) but the high dependence of dose position on proton range, requires PBT plans to have a stringent additional PTV. This contour will build on the CTV, accounting for uncertainties in range, stemming in both planning and delivery. (23)

With the associated tumour targets and ROI identified, planning can begin, with oncologists using SOBP from various angles to provide a conformal dose, and limiting dose in peripheral tissue. The "holy grail" is to provide 100% dose to the tumour and 0% to normal tissue (24), but in practise this is unattainable, making planning a task of optimisation. It becomes a monotonous time consuming task of making minor alterations to beam strength and position in order to maximise its performance.

Once principally completed the plan must be stress tested using uncertainties in CT scan conversions and beam ranges, to ensure the plan will pass in even the worst case scenario. The inability to test this during the planning stage is a major limitation of the work, which may require additional alterations before implementation.

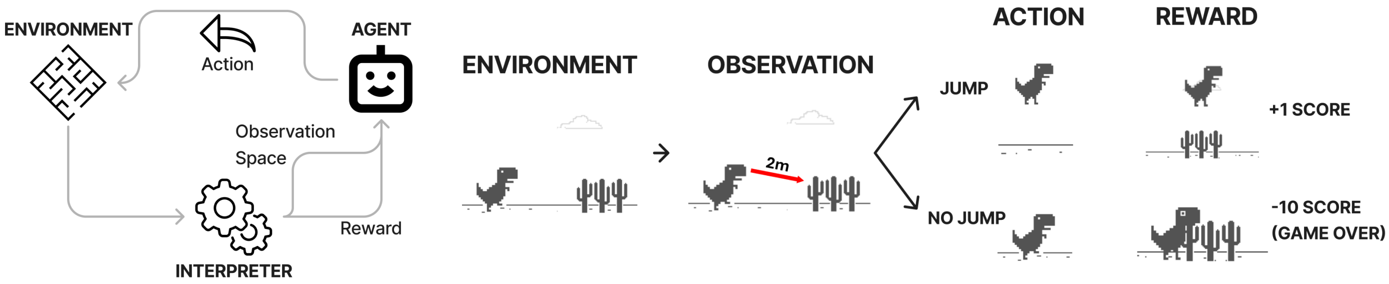
The issue of the long planning process and the considerations of errors, therefore place the task as one in need of an alternative method. An AI agent capable of completing these plans would present itself as a significant benefit to the field. It would reduce the strain on vital healthcare professional resources and potentially account for the uncertainty ranges during its work, creating robust implementable plans in fractions of the current time.

A complete PBT and photon treatment plan for the same patient is made available within Appendix C.

2.3 Reinforcement Learning

Reinforcement learning is a type of machine learning where an AI agent learns to accomplish a particular task by receiving penalties and rewards, which guide its behaviour in a continuous feedback loop. Figure 3a illustrates this cyclical process. As the agent interacts with its environment, each action generates an observation and a corresponding reward. These observations and rewards are then used to adjust the agent's policy in an iterative process that continues indefinitely.

The environment is an active space which changes over time under influences from the agent. The agent perceives this through the observation space, which is a set of information fed back, varying from a few vector values to large multidimensional arrays. Using this information it picks an action based on its policy (prior experience) from the action space. This consists of either a set of finite binary actions or a continuous set of real-valued actions. The reward fed back to the agent from the interpreter is a single float value quantifying its performance, allowing it to adjust its policy.



(a) Reinforcement learning feedback loop

(b) Example environment setup for a dinosaur game with the objective to jump over obstacles while continuously moving forward (25).

Figure 3: Overview and example of cyclical reinforcement learning loop. The agent receives an observation, makes an action and the interpreter feeds back a new observation and reward so the agent can undergo its next action. The dinosaur game panels represent one RL loop.

This can be exemplified by a simple game of a dinosaur jumping over obstacles while continuously moving forward, shown in Figure 3b. The environment is the setting the dinosaur is moving within, observation is the distance to any visible obstacles and action space will be two binary options: jump (1) and nothing (0). The agent will then choose an action at equal time intervals and receive a reward. This reward will be 0 if nothing happens as a result, +1 if it vaults an obstacle and -10 if it collides. The agent will then receive a new observation of the next obstacle and continue choosing actions to maximise its reward. The training episode will end either at a maximum number of steps or if it collides with an obstacle.

RL encompasses various compatible algorithms, all grounded in the cyclic principle illustrated in Figure 3a. These algorithms span from policy-based techniques employing gradient descent, wherein a strategy is crafted by linking specific actions to various states through a neural network, to value-based methods that assess the potential reward associated with each action. Additional, model-based approaches, which simulate the environment to facilitate efficient planning and decision-making, are also employed but primarily rely on trial and error, lacking an attempt to comprehend the environment.

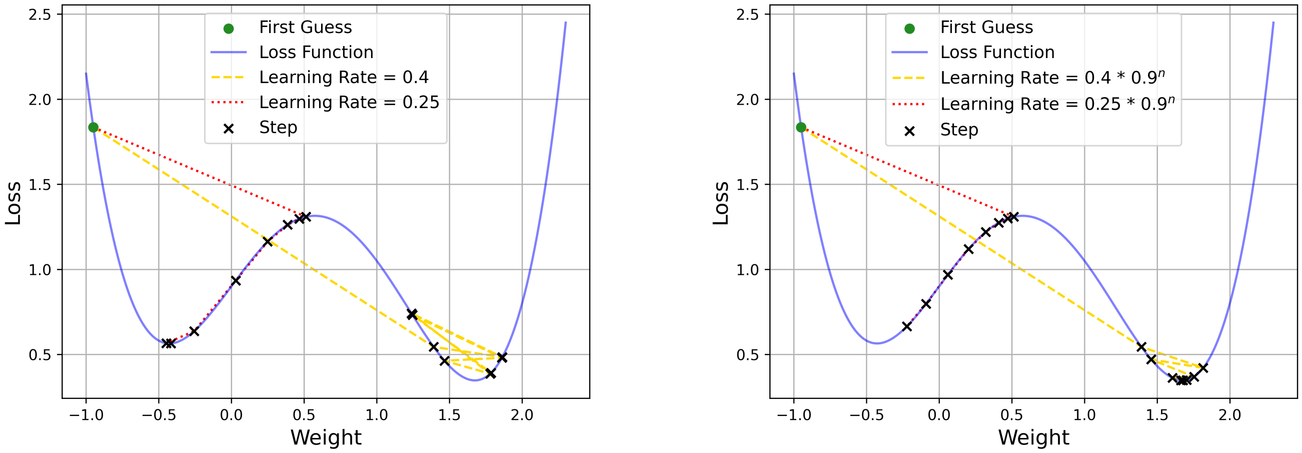
Each of these algorithms presents distinct advantages, excelling in addressing problems of varying complexities and in discrete or continuous spaces.

2.3.1 Proximal Policy Optimisation Algorithm

Proximal Policy Optimisation (PPO), developed by OpenAI in 2017 (26) is a policy gradient method for training an agent's policy network. These networks drive an agents decision making process by estimating the action that will yield the best reward given a certain observation. To train these policies a gradient descent method is employed to adjust the weights of network nodes to maximise the expected reward. Gradient descent operates by adjusting parameters by subtracting a fraction of the gradient to minimise a function. In this case, it is a loss function which determines the difference between expected reward based on the policy and the actual reward. PPO operates on this principle but diverges from traditional gradient descent methods by employing a specialised Clipped Surrogate Objective function. This prevents dramatic policy updates by limiting the size of change to within a 'clip range'. Adjusted policies therefore remain "proximal" to the previous policy, allowing PPO training to be more simplistic, stable and efficient than alternatives (27).

2.3.2 Hyperparameters

PPO is a complex algorithm guided by hyperparameters that adjust the approach and speed of policy refinement. While it is a robust algorithm, adjustments in these parameters can significantly improve both training efficiency and overall performance (28). The four key parameters utilised within this project are learning rate, entropy coefficient, discount factor and clip range.



(a) Gradient descent with two static learning rates and the same initial position.

(b) Gradient descent with two decreasing learning rates. For each step the learning rate is multiplied by 0.9

Figure 4: Loss function which the agent must minimise to reach an optimal policy. The learning rate determines the step size when reducing the loss function. Large values allow more exploration, while smaller values allow more convergence

Learning Rate (α): is the most critical value and determines the step size at which the neural network will adjust its weights. When adjusting weights it is aiming to minimise the loss function, which is the difference between expected and received reward.

An example of this adjustment is depicted in Figure 4, comparing the efficiency of a static rate and a decreasing one. The benefits of both high and low values are seen in Figure 4a. The lower static rate is efficient in converging on the first minima but fails to find the global minimum, while The larger rate locates the global minima but given its larger step size moves around the minima without converging at the bottom. In Figure 4b the two values are reduced with every step. This is beneficial for the larger 0.4 value, allowing it to find and then efficiently converge on the minima with its smaller step. However, the smaller value requires is hindered by its range, requiring additional steps to reach the minima. Overall it requires a balance between a large initial size to find the optimal minima, but also a reduction in size to allow convergence.

Entropy Coefficient (β): Determines the strength of the entropy regularisation term in the loss function. This encourages exploration within the environment by choosing more varied actions. Using a larger value will increase the randomness of the agents, allowing more of the environment to be observed, while a lower value will utilise the agent's current policy and stick with what it has previously done.

Discount Factor (γ): This value determines the extent to which the agent values future rewards relative to immediate rewards. The value ranges 0-1, with a higher value, close to 1, heavily prioritising long-term rewards and focusing on actions which set up the environment for future success. A lower value will prioritise actions which give the highest immediate reward, irrespective on if it will dampen progress later on.

Clip Range (ϵ): This is the only value specific to PPO. It determines the range at which the policy ratio can be clipped when updating. Updates are limited to the range $1 - \epsilon < x < 1 + \epsilon$. Larger ϵ values allow for more aggressive updates that may lead to faster convergence but also increase the risk of divergence and instability. Lower values only allow conservative updates and a gradual refining of the policy.

3 Method

3.1 Implementations of Patient Data

For an agent to accurately design patient treatment plans it must undergo training on a substantial dataset of similar real patient scans alongside measured dose values. Data is embedded in the training in two key ways: patient CT scans to create environments the agent trains in, and clinical beam data that can be interpolated so the agent can place dose using any specific beam energy. All data used is preprocessed offline and stored as dictionaries within `pickle` files, that can be imported into each training run. The data structure of the two `pickles` used are shown in Figure 5.

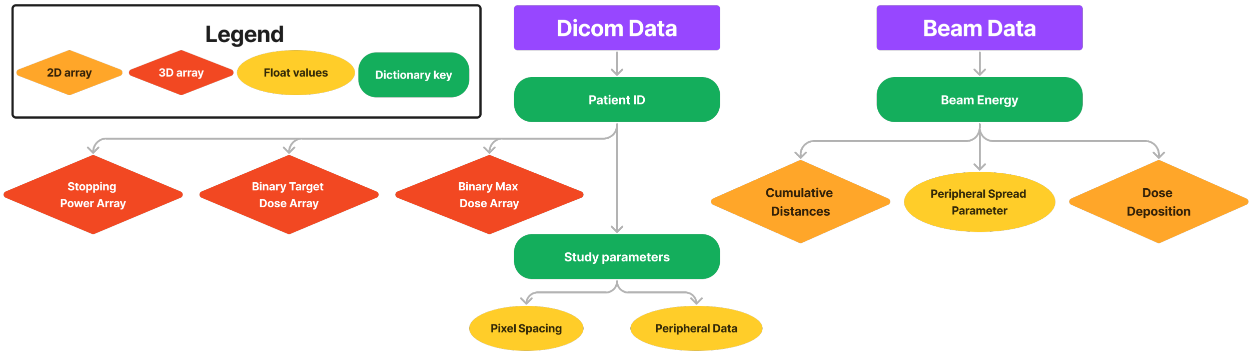


Figure 5: Depicts the breakdown of the two data dictionaries which store the extracted and preprocessed data from the patient DICOM datasets. They are stored offline and imported into each of the training runs as .pkl files to build specific environments for the agent to train in.

Creating accurate patient environments requires embedding contours for different structures and using density data specific to each patient. This data is stored as DICOM data files(29) and handled using the `pydicom` package(30). Data sets comprise of CT scans, manually annotated with contours specify the location of tumours and key structures such as the brainstem or spinal cord.

The same method for extracting structure locations within a scan and stacking them to create 3D arrays has remained from the first semester of this project. This involved using computer vision to create binary indexed maps of different structures from contour coordinates, assigning each structure a max and target dose value and then merging all arrays with a priority. The final output is two 3D arrays containing maximum dose and target dose for every voxel within the patient scan.

Changes to data handling this semester occur for CT density values. Within the DICOM files, the density of CT scans are stored in Hounsfield Units (HU). These units are a linear transformation of linear attenuation coefficients of material, such as the radiodensity of distilled water at standard pressure and temperature is 0HU, with air defined as -1000HU. The conversion is done using;

$$HU = 1000 \times \frac{\mu - \mu_{water}}{\mu_{water} - \mu_{air}} \tag{6}$$

with μ as the average linear attenuation coefficient of each voxel measured and μ_{air}, μ_{water} constants (31). The initial data handling script adjusted the HU to water attenuation but did not recognise that often stored values are adjusted by +1000 such that air is 0HU. This meant correcting the data within the CT scans before applying the HU conversion to create an accurate water equivalent density map. Conversions were done using linear interpolations of a clinical data set, that followed the linear relation set out in Equation 6.

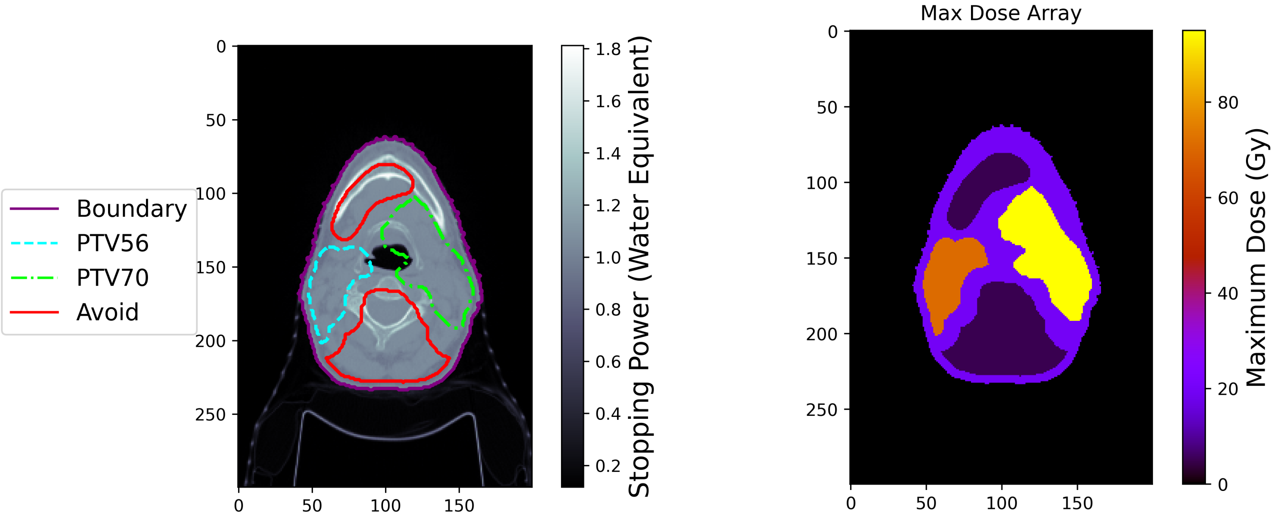
The other aspect of data handling, that was adjusted for the second semester of the project, is the creation of maximum and target arrays. As the environments became more complex they required a more sophisticated reward system that relied on each individual pixel having an assigned target and maximum dose in Gy. This meant within the data extraction, each contour was assigned a target, maximum and priority number. By adjusting all contour names within different datasets to a universal naming convention a map can be created of voxels with values that are defined based on the contour with the highest priority present. Saving these as 3D arrays within `dicom` file, allows each slice to be individually called to form the part of the observation and reward system for specific training runs.

3.2 Environment Design

Having previously worked with simplified 'toy' environments, it offered an understanding of environment features that will work best when implementing patient data. Simplistic action and observation space designs were the key takeaways, yielding the best results by allowing the agent to train faster and produce more consistent results. The final setup in this project consists of an environment with two 2D arrays depicting the maximum and target dose at each voxel and an observation space feeding back normalised dose values against the maximum value. The action space, allows the agent to place the peak of Bragg peak at any voxel and define one of five angles evenly spaced around a circle for its entry path. The reward metric is a negative quadratic function called at every voxel, with the reward increasing until the target dose and decreasing beyond.

Example environment slices are seen in Figure 6 with 6a detailing the contours for visualisation and the water equivalent density values along with Figure 6b depicting the maximum dose values. These maps have been created for every slice within each patient, over six different patient datasets, to produce 260 different environments the agent can train on. Data slices are stored as 512x512 arrays but to improve the efficiency of the training slices are cut to (300,200), removing empty space around the patient.

Agent training was done in `python`, using the `stablebaselines3` package (32) to import the PPO algorithm and `gymnassium` (33) to build environments spaces the agent can train in. Using these prebuilt packages allowed the freedom to train the agent using multiple environments within each run, either vectorised or by switching data after a set amount of training. This introduced the decision of which and how many datasets should be used for training. While random slice combination bodes well for the generality of the model and increases the variety of situations the agent is exposed to, external successes with curriculum learning prompted a more carefully designed sequence of slices (34). Curriculum learning aims to improve agent training by exposing it to initially simple environments, increasing difficulty with training and allowing it to hone an initial basic policy by avoiding poor habits. To utilise this technique, initial slices contained large tumour volumes and either small or non-existent avoid regions. This allowed the agent freedom to explore all angles and dose positions without the restriction of avoid regions that more complex slices contained. After this general learning, the latter slices trained the agent to learn dosing a tumour from specific angles to preserve dosing avoid regions. The initial simple slices were trained in loops for a large number of time steps with a decreasing learning rate to create a base policy, which was 'finetuned' by training after with an increased learning rate on loops of complex slices.



(a) Patient data CT scan, depicting two tumour levels and regions to avoid. Avoid regions consisting of Spinal Cord, Brain stem and Oral Cavity.

(b) Corresponding max array for CT scan. Depicts the dose values in Gy that should not be exceeded.

Figure 6: An example of contours drawn onto a patient CT scan slice and the corresponding maximum dose array fed into the training environment.

3.2.1 Action Space

The action space is a predefined range of numbers within a multidimensional array which in combination corresponds to a specific action an agent can take. Optimised spaces have the minimum number of combinations possible for the required functionality of the model.

The initial action space implemented consisted of four components; choice of place or remove dose, position of dose peak placement within the 512x512, angle of beam entry and an index to define which dose to remove. This space was highly complex and contained values that were potentially irrelevant depending on the initial choice of action (dose removal index irrelevant if first choice is to place dose). The total space became a minimum of 1.89×10^8 different combinations, multiplied further by the number of dose indexes to remove.

Results using this space were inconsistent and required a lengthy amount of training. To improve, the action space was shrunk and adjusted to retain core capabilities but lose unnecessary options. The final action space is two-dimensional, consisting only coordinates of dose placement and entry angle. This eliminates the option for removing a dose, which should be unnecessary given a refined enough policy. These arrays were also diminished, such the choice of coordinates was one out of every three voxels within a 300x200 grid and the angle options were five evenly values. This creates an action space with only 34700 options, 5400 times smaller than previously.

3.2.2 Observation Space

The observation is a form feedback providing the agent with the relevant information to make a decision on which action to take following a step. Similar to the action space it should be as simplistic as possible while providing all necessary information. In this setup it needs details on the position and quantity of dose and how it compares to the maximum and target. To reduce the dimensionality of the space, all information is conveyed in one array, as seen in Figure 7.

The value of each voxel follows these set of formula, where L_T is the target limit, L_M is the maximum limit and D_A is the applied dose.

$$\text{Observation} = \begin{cases} -2, & L_M = 0 \\ \dfrac{D_A}{L_T} - 1, & D < L_T \\ \dfrac{D_A - L_T}{L_M - L_T}, & L_T < D_A < L_M \\ \dfrac{D_A}{L_M} - 1, & D_A > L_M \end{cases}$$

There was experimentation with providing additional information in the form of arrays such as the target array and maximum array in a 3D format. Yet these yielded no clear improvements in training results and only increased the time taken to complete each step.

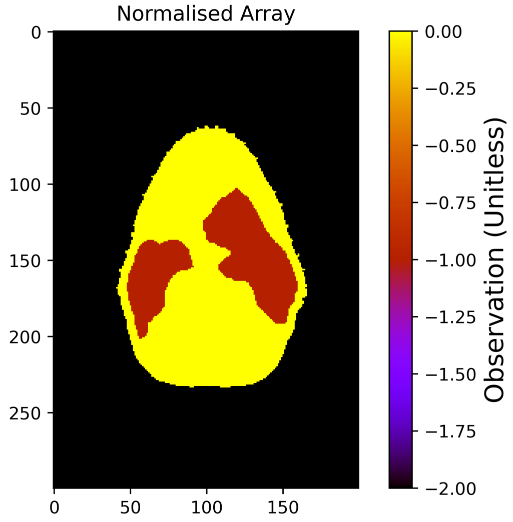


Figure 7: Observation space within environment. Values at -2 are empty space, -1 are tumour regions and 0 are avoid and normal regions within the patient.

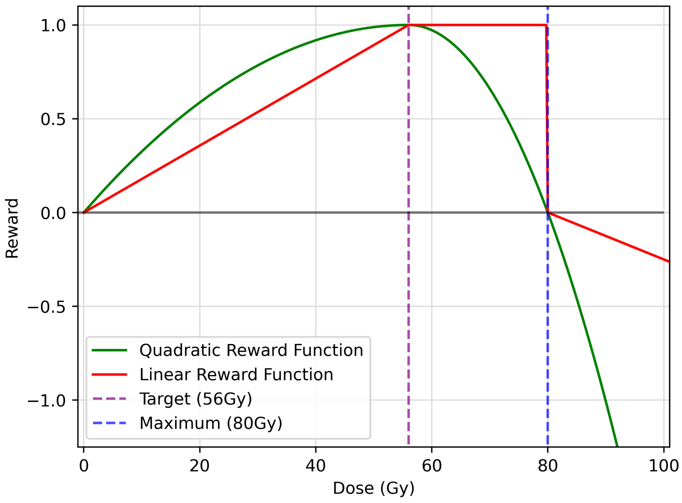


Figure 8: Reward system function. This specific function is for a tumour with a target of 56Gy and a max of 70Gy. All other regions have the same function but are adjusted for target and maximum values. normal and avoid tissue have a target of zero so the function is only the second half.

3.2.3 Reward System

The aim of training a model is to maximise the total reward which can be achieved over an episode and is depicted by a plateau in episode reward against step number during training. Therefore, the metric for calculating reward needs to promote dosing the tumour to its target and avoid going over the maximum. The final value should also be normalised, making 1 the largest reward available for each step, allowing performance to be easily compared between different environments and training data.

Initial systems consisted of normalised linear values, increasing from zero to one up to the target dose, constant at one between the target and max and decreasing beyond to form a negative penalty above the maximum. This system did not yield optimal results, due to no difference in reward just above the target and below the maximum. The agent struggled to determine that the optimal dose was just above the target and instead continuously overdosed tumours leaving a high average dose within peripheral material.

To prevent this, a non-linear reward system was implemented, penalising overdosing heavily and promoting a conformal dose over the tumour around the target. The function can be seen in Figure 8. Formed of two negative quadratic equations pieced together at the target, it ensures the reward increase is initially fast, leading to a maximum at the target and a relatively consistent reward in the range surrounding. Results of this implementation were a reduction in overdosing given the stricter penalisation from the quadratic function and a more conformal dose around the tumour, derived from the largest reward occurring around the target.

3.3 Fast Analytical Dose Model

One key limitation in the progress of the first semester stems from the absence of a dose model. Previously, a simple Gaussian grid sufficed, providing a basic understanding of dose implementation. However, it significantly differed from a Bragg Peak due to the absence of a dose tail. Therefore, a key advancement in this semester involves developing an analytical dose model that accurately portrays the proton's dose pattern

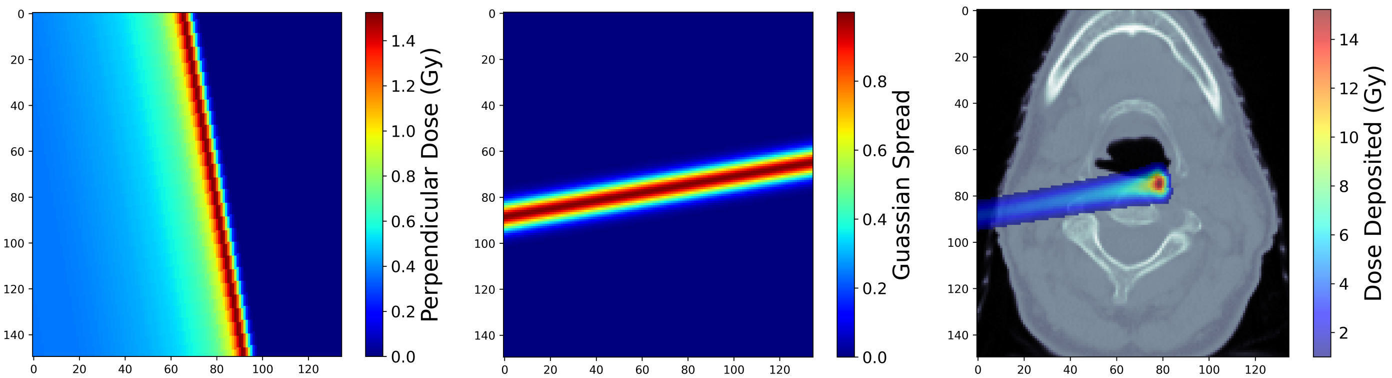
To match the action space the dose model is defined by a peak position, angle and beam energy. This succeeded a model defined by angle, beam energy and an entry point on the edge of the environment. Issues arose in that setup, with the agent struggling to understand which beam energy to reach certain depths in the patient given the inhomogeneity in density. Therefore, defining the model by peak location and angle instead allowed the backtracking of the path to calculate the beam energy required for that placement. This significantly reduces the action space, as only the position and angle must be defined.

The process for calculating the dose map for a specific beam given a peak position and angle is as follows:

1. Using the peak position and angle a straight line is defined through the environment along this path
2. Intercepts with the environment voxels are calculated, creating an array of δx distances through each voxel
3. Multiplying the voxel crossover distances by the Water equivalent densities found within the CT scan creates an array of Water Equivalent Path Lengths (WEPL)
4. Beam energy required for this depth is calculated using a relation between WEPL to peak and beam energy.
5. A linear interpolation for this specific energy is calculated using clinical beam dose data and is then mapped onto the WEPL δx to create a linear pencil beam.
6. To add peripheral spread, for each voxel the perpendicular distance to the beam and the dose value at the intercept is calculated
7. A Gaussian function is then applied using the distance, peak value and corresponding Gaussian standard deviation for that energy to create a full dose map.

The components of this model can be seen in Figure 9. Gaussian spread and dose perpendicular to the beam line are multiplied to create a single proton beam, specific for the density map and beam position.

This technique was the product of research into the optimal dose calculation methods in treatment planning. Commonly used methods are either Monte Carlo numerical methods(35) or algorithms based on look-up table values (36)(37). The most accurate of the two is a numerical Monte Carlo method using the Bethe Bloch equation. This provides high precision within heterogeneous material (38) and allows for a significant decrease in uncertainty of plans. This is shown to reduce the uncertainty of beam ranges in plans relating to lung tumours from 6.3% ± 1.2 mm (look-up table techniques) to 2.4% ± 1.2 mm (39). Drawbacks of this method exist within compute time and resource demand, with fast calculations taking “0.82 to 4.54 seconds” for an accurate brag peak (40). This method would limit agent training to a maximum of 4390 training steps an hour, making a training run on the order of millions of steps completely unfeasible. The second method of using lookup tables to interpolate data depending on the scenario is more applicable to this project, with linear interpolations taking the order of milliseconds and placing the maximum training steps per hour at 3.6 million. The implementation of this technique, as explained previously, was inspired by the work of Da Silva et al (41) who used calculations of water equivalent distances and mapped offline calculations of dose deposition data to create an analytical pencil beam model.



(a) Map with voxels assigned the dose value of their perpendicular intercept with the beam line

(b) Gaussian spread map along the beam line. Spread parameter is fit from data and energy dependent.

(c) Final Dose map produced by multiplying the Gaussian spread and perpendicular dose map

Figure 9: The creation of a dose map for a single Bragg peak in a specific patient density map. The first two maps are the bulk of the necessary model calculations and are multiplied together to create the final dose beam. For multiple beams each map will be summed together to create a total dose map.

4 Results

4.1 Dose Model

The dose model created within this project utilises linear approximations of clinical data, allowing the easy input of data values but limiting the accuracy against numerical Monte Carlo methods. The benefits of the analytical method preside in the fast compute time and the avoidance of preliminary offline calculations with each change of data input.

Understanding the limitations of the model's accuracy requires analysis of different aspects of its calculation. This section will focus on examining deviations of linear approximations from the data set and the impact of voxelising the dose values.

4.1.1 Accuracy of Model

To compare the dose model with data, linear approximations need to be created for each set of beam data energy, yet modelling beam energies that are within the data set, extracts the exact interpolation. Instead, the intermediary energies without predefined data must be examined. To account for this, in each comparison between model and data, the specific beam energy data was removed from the dataset and the linear approximation is a product of the two energy datasets on either side of it.

The first set of results are a comparison between the model and data, done by inputting the necessary Bragg peak distance and set of distance values from the data into the model. Results are depicted in Figure 10 for ten different energies within the 70 - 245MeV range, recording the chi squared for each fit compared to the data. The figure exhibits strong fits for all energies except 70MeV, which hosts a chi-squared an order of magnitude larger than the others. This deviation is to be expected though, with the 70MeV being calculated only with the 80MeV data as a reference, due to a lack of lower energy data. The rest, all with two comparison beams boast strong fits, suggesting an accurate model.

Delving into more detail, the residuals of the beams can be analysed to understand the weakest sections of the fit. Figure 11 depicts the residuals for the best and worst fit chi-squared values (excluding 70MeV). The largest deviations of the fit from the data occur around the Bragg peak, in particular along the distal fall-off. This creates an issue of undervaluing the dose before the peak and overvaluing beyond it, creating a larger spread of dose and suggesting the total dose delivered is larger is the case.

These results simply recreating the dose deposition value within the data. To understand its accuracy within the voxelised environment the agent trains in, the value assigned to each voxel must be examined. To increase the compute speed of the model, each voxel's dose value is assigned using the distance along the beam at the midpoint of each crossover. Initially,

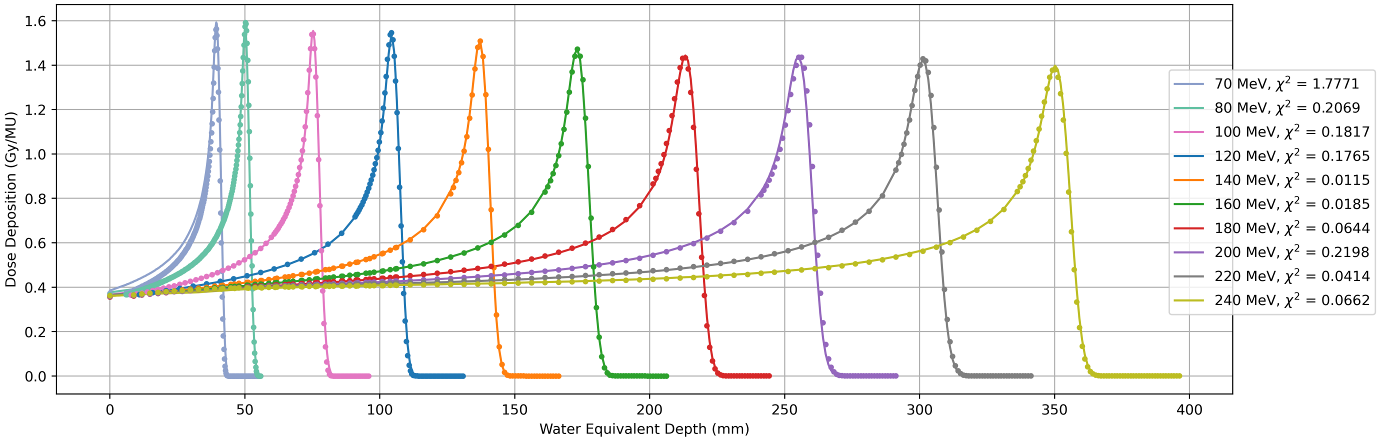


Figure 10: Comparison of the linear beam model against the actual beam data set. Each set of comparison data was removed from the dataset during calculations. Chi-squared values comparing the model and data are displayed in the legend.

the model used an integral function to calculate an average deposition value over the entire range of distances the voxel encompasses. Yet this calculation demanded significantly more compute time and was 98.7% slower. This technique, while unimplementable, can serve as a baseline for a comparison between the midpoint approximation and the true average across each voxel.

The variations between the two are highly dependent on dx lengths across the voxel, as smaller voxel crossovers will have a smaller dose deviation across it. Within the environment, the largest possible WEPL crossover, 2.52mm occurs when the real distance is 1.4mm and the water equivalent density ratio of 1.8, while the smallest (within the patient), 0.2mm, will have a real distance 0.2mm and a density ratio of 1. To understand how these variations impact accuracy and the worst-case scenario, chi-squared values are calculated for a range of different beam energies and δx values in Table 1.

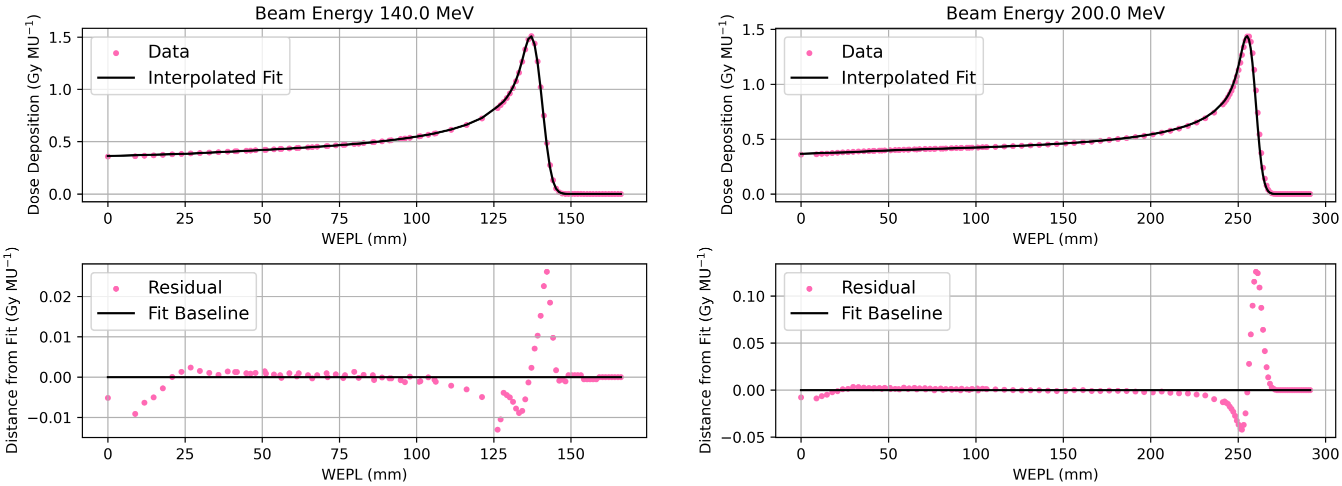
		Beam Energy (MeV)				
δx WEPL distance (mm)	χ^2	80	120	160	200	240
0.2	(10^{-6})	1.9123	0.4341	0.1141	0.0822	0.0647
1.1	(10^{-4})	2.8155	0.7218	0.1390	0.1408	0.1080
2.52	(10^{-2})	1.6255	0.7150	0.2179	0.1598	0.1150

Table 1: Chi-squared comparison of calculating voxel dose deposition using a midpoint assumption and using an accurate integral average method. The comparison are done for different voxel crossover distances δx and at a range of beam energies

The largest deviations from the true value occur for small beam energies and large δx, with 80MeV at $\delta x = 2.52$mm bearing a chi-squared of 0.00163, while the smallest, using a 240MeV beam and 0.2mm δx the chi-squared is 6.47×10^{-8}. These results signal a strong fit between the two, validating the midpoint assumption for the case of prioritising compute speed.

A residual plot of δx = 2.52, 0.8mm for the 140MeV beam is depicted in Figure 12. A similar residual pattern to the linear approximation model emerges, underestimating around the peak and overestimating the fall-off. This is consistent throughout the energies and δx.

As stated, the accuracy of the model is not of utmost importance at this stage of work, but these results present the need for additional consideration of its calculation. The largest deviations occur around the peaks, where uncertainties have the largest impact on the suitability of the plan and combining both sets of approximations only exacerbates this difference. While the linear approximation from the data is complicated to solve the midpoint approximation can be reconsidered in a simpler manner. Given the deviations are solely around the peak, it raises the possibility of implementing the integral average calculation over just this set of values, retaining the approximation for values without significant deviation. This would greatly improve the accuracy when voxelising the dose, but would not consider the larger deviation occurring in the



(a) Residual of 140MeV beam with χ^2 = 0.0115.

(b) Residual of 200MeV beam with χ^2 = 0.2198.

Figure 11: Residual plots comparing the best and worst chi-squared value fits with their respective beam data. The largest deviations occur at the nuclear build-up, Bragg peak and distal fall-off.

linear approximation. These are notably larger, particularly for the 200MeV beam displaying differences above 0.1 Gy/MU, an order of magnitude larger than the largest differences from a δx = 2.52mm. Therefore, this questions the necessity of improving the voxelisation at the cost of compute speed if the primary source of uncertainty remains unsolved, making the smaller uncertainty redundant.

4.1.2 Computational Speed of Model

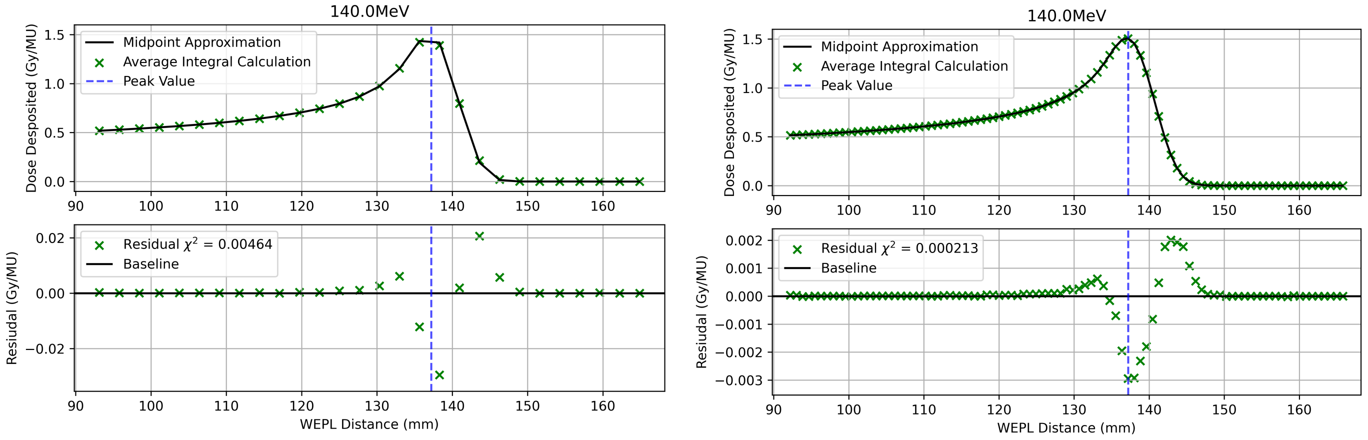
Compute time is the primary consideration for the beam model within this project. While a highly accurate beam is desired, the trade-off between accuracy and compute time is too large at this stage of the work. The benefits of additional experimentation, only capable with a faster, less accurate dose model are greater than the limited testing otherwise.

Multiple adjustments to the beam calculation were made during this project, all impacting the speed of calculation. Most influential was the size of the environment, with arrays of fewer indices performing much faster. To understand the effect of environment size and other changes, 63200 beams were calculated at a range of distances and average times measured. The mean times for different array sizes and calculation changes are displayed in Table 2.

Environment Size ($n_y \times n_x$)	Mean Dose Calculation Time (ms)
300 x 200 *	351 (90)
512 x 512	20.59 (4.73)
412 x 412	10.60 (1.23)
362 x 362	8.16 (0.99)
300 x 200	3.55 (0.47)
300 x 200 †	4.44 (0.55)
200 x 200	2.45 (0.40)

Table 2: Average beam calculation times for 63200 beams of varying length and angle of entry for different environment array sizes. † Calculates Gaussian function for all voxels, irrespective of distance. * Accurate average voxel dose calculation

Between different environment sizes a notable drop in compute time is observed. Shrinking the number of array elements by 6.5 times, from 262144 (512 x 512) to 40000 (200x200), reduces the compute time by 88.1%. This change is significant



(a) Residual of 80MeV beam with $\delta x = 2.52$mm. $\chi^2 = 0.0163$.

(b) Residual of 80MeV beam with $\delta x = 1.1$mm. $\chi^2 = 2.82 \times 10^{-4}$.

Figure 12: Comparison of midpoint approximation fit with the true average dose deposition value for each voxel along a beam line. Different voxel cross over distances are examined for a 140MeV beam. The largest deviations are around the Bragg peak and the model using larger deviations exhibits larger deviations.

for the training of a model, increasing the rate of training by a comparable percentage. These observations therefore led to the decision to reduce the environment array sizes. The drawback of reducing the size is the loss of data, meaning the size could not be shrunk such that it loses density data from the CT scan. Removing empty space is not an issue, the water equivalent density is insignificant enough that the beam path is not altered. Alternatively, if patient regions were removed the beam range would be altered, excluding water equivalent distance it must traverse.

300x200 was the final decision as throughout data sets encompassed all of the dense patient regions, while reducing the number of voxels the model must consider by 77%. This change presented additional benefits within the environment spaces, reducing the number of actions required to allow the placement of dose in any voxel, further increasing training speeds and performance.

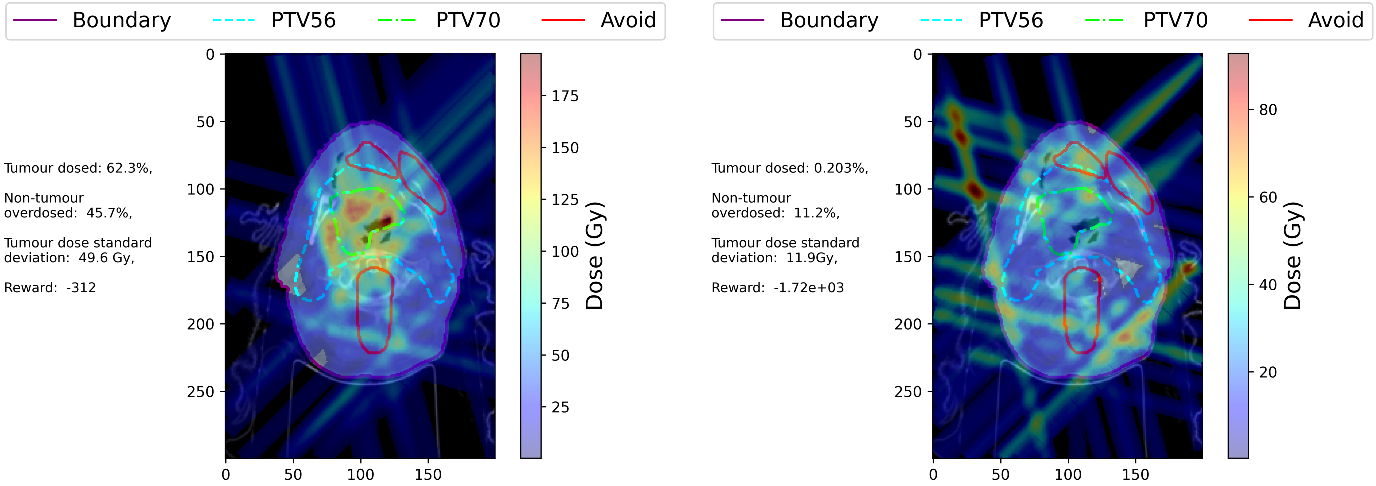
Along with adjusting the array sizes, tweaks in the calculation technique were implemented throughout to improve calculation times. Two changes are shown in Table 2. An example of a large adjustment is the deviation in average voxel dose, discussed in the previous results section which reduced calculation times by 98.7%. Minor code-based tweaks such as limiting the amount of Gaussian calculations still had relevant improvements improving the speed by 20% without sacrificing any accuracy.

4.2 Effects of Hyperparameters on Training

In order to optimise the performance of the model, hyperparameters must also be optimised. Therefore, to understand the best combination, a training run of 6 training slices looping every 100,000 steps, for a total of 400,000 total steps was trained with a combination of values for learning rate (α), clip range (ϵ), entropy coefficient (β) and discount factor (γ). Three different values, a high, standard and low were used for each, creating 81 different combinations of hyperparameters. The learning rate refers to the initial value as it was decreased stepwise by 5% every 60,000 steps. These different combinations were also trained on the same set of environments but with a small SOBP placed instead of a standard single Bragg peak, to ensure these successful combinations of hyperparameters are not independent of the exact environment setup. The trend of results between the two was the same with a slight underperformance for SOBP. Only results for the single Bragg peak model will be discussed.

To validate the performance of the model, supplementary metrics have been used, describing the models ability to dose the tumour conformally and avoid the normal tissue. These values are the standard deviation on dose within the tumour voxels, the percentage of tumour voxels dosed within the target and maximum and the amount of normal tissue voxels over the maximum.

The episode performance of models trained with the best and worst parameter combinations are shown in Figure 13. Despite both having been trained for 400,000 time steps they display a vast difference in ability. The worse model shows no



(a) Highest performing model. α = 0.001 (std), β = 0.01 (high), ϵ = 0.3 (std), γ = 0.99 (high)

(b) Highest performing model. α = 0.01 (high), β = 0.001 (low), ϵ = 0.2 (low), γ = 0.90 (low)

Figure 13: Two models with the highest and lowest final episode reward, placing 125 doses within a patient slice. The maximum dose values of the tumours have been increased to 200Gy to further discourage placing dose peaks in normal tissue

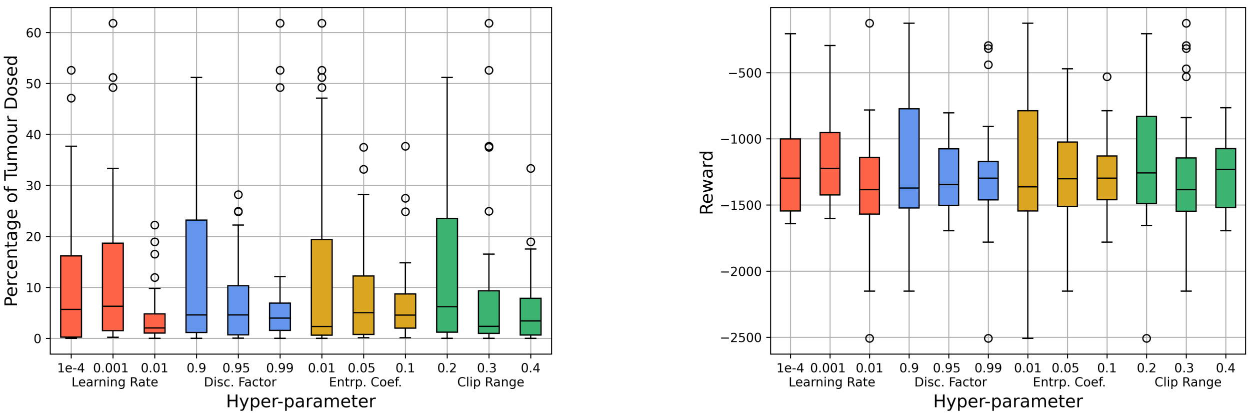
sign of a refined policy, seemingly randomly placing dose with no consideration to where the tumour or the patient is, while the best model performs much better, placing the majority of dose within the high-level tumour. The reward from both supports the visible observations with the worst receiving an almost minimum reward and the best, a value in the highest range any models have produced (-400, -200). Additional metrics provide context to the visible dose pattern but must be considered collectively. Ignoring the tumour dosed value, the worst model appears to perform better, with minimal dose in normal tissue, and less variation across the tumour, yet this is only caused by its lack of relevant dose placements. To extract insight from the metrics, the first consideration is tumour dosed, and only if that value is above 30% can the other metrics inform on performance.

The performance of individual hyperparameters is shown in Figure 14. Each parameter forms part of 27 combinations and the metrics of each combination's model are depicted within a box plot, showing the spread for each specific parameter. Figures for the standard deviation and non-tumour overdose are shown in Appendix E. Most model combinations trained poorly, with only ten reaching 30% tumour dosed and five obtaining a reward larger than -400. Observing the plots corroborates this, as 'good' performances of tumour dose (Figure 14a) and reward (Figure 14b) are outliers within their own dataset.

Trends between individual hyperparameters are also observed. Notably is the poor performance of the high 0.01 learning rate, boasting the lowest spread of percentage tumour dosed and the lowest mean reward. Revisiting Figure 4a it is clear having too large a value has prevented it from converging on an accurate minimum and is instead bouncing around the function edges. This is despite the learning rate gradually decreasing as training went on suggesting any value above 0.001 is insufficient for this setup.

Interestingly the worst-performing discount factor on average, produced the best overall results. γ = 0.99 had the lowest quartile range for both the percentage of tumour dosed and reward, yet was a part of 3/4 of the best performing models. This presents the discount factor as a variable heavily susceptible to its adjacent parameters, which is understandable given its influence on prioritising long-term rewards. If the policy never focuses on an optimal strategy the model will generate poor rewards initially in the hope of future benefits but lack the refined policy to obtain those.

Entropy coefficient results also match up with its definition. The spread of rewards decreases as the coefficient increases and the actions become more random. This is counter-intuitive but the randomness of actions allows the agent to explore more of the environment and place dose in more locations, creating a more consistent policy over all the combinations. A lower value decreases the randomness, prompting the agent to continue placing dose where it initially sees a positive effect. Depending on the location this can either severely overdose avoid regions or continuously dose the tumour, leading



(a) Percentage of tumour voxels with dose between the target and max.

(b) Total cumulative reward after 125 agent actions.

Figure 14: Shows the spread of performance metrics of models trained for 400,000 steps. Models completed 125 steps within an environment ten times and metric average were taken. Each boxplot includes data from the 27 hyperparameter combinations containing that specific parameter.

to a large spread of rewards. Similar to the discount factor its strong dependency on other parameters is clear, as the less explorative 0.01 is part of the best tumour-dosed values observed, suggesting large values inhibit a policy from refining by including excess randomness.

The clip range results are more ambiguous, with both 0.2 and 0.3 appearing good choices. The larger 0.3 shows larger variations in reward, and an extremely low average tumour percentage, but also produced the best models. 0.2 on the other hand produced more consistently high results, but did not peak as high as 0.3. This leads to the conclusion that while low values like 0.2 will allow for a steadier policy refinement avoiding the agent getting stuck in bad policies, it is not sufficient enough like 0.3 to make the large changes when necessary to converge on an optimal strategy.

5 Conclusion

The initial plan for this semester involved creating a 3D environment and transitioning into treatment planning for an entire patient, yet difficulties in progression within the 2D environment necessitated further research and experimentation to understand what the best approach would be. The AI agent models produced in this project have also not been successfully at creating a patient treatment plan or accurately dosing a tumour within an environment. Training has allowed it to understand that dose should be applied within the tumour, but has not deterred it from dosing avoid regions or placing a conformal dose. Therefore the positives reside in the significant progress towards creating an accurate environment to achieve a treatment planning model. Particularly the dose model which computes in the order of milliseconds and an automated environment modelled entirely from patient data that has laid the foundation for future progress.

6 Discussion and Future Considerations

6.1 Discussion of Results

Key results of this semester are analysing the accuracy of the beam model, in relation to original data and voxelisation, along with the effect of hyperparameters on the performance of trained models. Within the dose model, adjusting a linear function to discrete voxel values will always lose accuracy, but a continuous model is not possible. The current midpoint assumption in place has proven to be accurate within 1.3% for the most extreme cases and less than 1% for the majority of cases. The model's comparison to real clinical data is also examined. This deviation is far more significant, with specific voxels around the Bragg peak fall of, exhibiting a 10% change from the expected dose. Throughout the rest of the beam, the difference is minimal, but the edge is often the most influential, so is of higher cancer. Overall for the beam the chi-squared values were 0.11, over a large number of data points. This makes the voxelisation uncertainty currently redundant until the larger deviation is solved, but suggests with improvements to the model around the peak, it will become sufficient for training an agent to create treatment plans.

The results of the hyperparameter tests are harder to quantify. By conducting these adjustments the highest performing model of the project was achieved, suggesting the best combination for this current setup. Yet with adjustments to the observation and reward system, which are required, this will likely change. Therefore understanding the impact of the parameters on different aspects of performance are more beneficial. Key observations are the variability of performance due to changes in parameters, with the combinations producing the best results becoming far worse with one small adjustment, preventing it from converging on a policy. This means for future setups, parameters with large variations, such as 0.01 entropy coefficient or 0.9 discount factor, should be experimented with to witness the range of convergences throughout trained models.

Overall the underlying results not explicitly discussed in the results, present themselves as this project's largest success. While the models did not train optimally, the tools to conduct the training such as the comprehensive environment slices extracted from patient datasets or a dose beam modelled off clinical data, have been optimal. These allowed extensive experimentation with RL methods and would form the basis for future work given the accurate treatment plan test environment.

6.2 Discussion of Uncertainties

Uncertainties are a critical consideration in the field of proton therapy. The nature of treating human patients automatically warrants a high degree of accuracy but the precarious nature of a protons beam's localised high dose exacerbates this consideration, with millimetre deviations in Bragg peak positioning potentially resulting in serious overdosing of vital organs. Despite this, the nature of this project warrants a lesser consideration, with priorities on examining RL's ability to plan dose placements within a patient dataset and experimenting with training techniques. To achieve the best results, compute speed can be prioritised over precision to allow for a larger quantity of results at this early stage in development. Then once a high-performing model is obtained, uncertainties will become the primary focus.

The largest uncertainty in the delivery of proton therapy arises not within the treatment planning but in the positioning and anatomy of the patient. Treatment periods are stressful for patients and the increased distress in their lives leads to breaks in habits. Often leading to patients experiencing weight loss or gain that can influence their anatomy significantly, increasing uncertainty and in certain cases requiring an adjusted treatment plan. The setup uncertainty can therefore range from 1mm for cranial treatments to several mm around the centre of the body, before considering anatomic movements such as breathing (42). To account for this typical plans will have an assigned margin of around 7% on beam ranges that encompass all uncertainties. Then, if the plan passes all validity tests at both extreme cases the plan can pass. Though recent suggestions in the field have been of creating "robust plans". These plans consider uncertainties in dose calculation, imaging, biological effect and patient setup from the outset (43), resulting in less sharp dose gradients and lower dose around OARs. Yet this type of plan is hard to manually create, leading to the possibility of this projects final goal being a model which creates a robust plan by considering all possible uncertainties in its training.

The voxelised nature of this project's environment grid is a major source of uncertainty, limiting the precision of any plan produced to regions of 1.1x1.1mm in the xy plane and 3.1mm in z by constraining the accuracy of tumour contours and the dose model. While small deviations along the beam line have proved efficient at modeling linear dose through a voxelised environment by using average dose values, the Gaussian spread of the beam is more complicated, with the centre of voxels used as distance for the spread. Given the angle of the beam relative to specific voxels, using the centre can occasionally not be representative of the average distance of the whole voxel. Therefore to account for that, each voxel should have a maximum, minimum and average distance to the beam line calculated using the voxel edges, to create a range of possible Gaussian parameters. Once a plan is created these values can then be used to create three different plans with maximum and minimum spread of the beam, to give a best and worse case scenario. This would be a similar approach to what is taken now in PBT planning, so to get closer towards a robust plan the dose beams with the different spreads can be integrated into the reward system, so the agent ensures the plan is robust to all variations.

Another significant uncertainty revolves around converting density data from HU to water equivalent distance. This conversion relies on linear interpolations of clinical data, detailed in Appendix D. During beam calculations, minor fluctuations in stopping power can lead to substantial variations in dose deposition in voxels surrounding the Bragg peak.

The extent of peak fall-off varies from 3 to 10mm, contingent upon beam energy, with typical distances over voxels ranging from 0.4 to 1.2mm. Consequently, water equivalent distances in the densest areas can approach 2.4mm. Occasionally, such small distances can represent a significant portion of the distal fall-off. In such cases, a voxel containing the very apex of the Bragg Peak, receiving maximum dose, may register as receiving negligible dose if its midpoint aligns with a lower dose. While this underscores an issue with voxel dose calculation, it also exemplifies how slight variations in water equivalent distance can profoundly impact the received dose, particularly around OAR commonly situated near a peak fall-off. Hence, the uncertainty in HU to Water Equivalent Stopping Power conversion warrants consideration similar to the approach described for Gaussian spread. This involves integrating the maximum and minimum conversion values into the model to assess the plan's suitability under varying scenarios.

6.3 Future Considerations

The process of creating an AI agent capable of creating a near-perfect PBT treatment plan is very complex and requires extensive consideration of all potential uncertainties given its intended purpose. This project has not reached this goal but has instead provided insight and tools to achieve it in the future.

Progress can be made towards the goal by first adjusting the observation space. Current observations are large arrays which overload the agent with information. Instead a refined approach using an array of simple parameters such as previous dose's distance to the tumour and the total dose in the voxel previously dosed. These would represent a completely different approach to how the agent learns without requiring a rework of the underlying functionally of the code. This change would require additional hyperparameter testing though, but by using the parameters shown to have the largest spread in this report, the performance potential of the setup can be easily seen, within a few combinations.

Uncertainties will also need a new consideration and an implementation into the agents feedback loop. A solution would be introducing a method of informing the agent about the best and worst case scenarios of each action. By including them in new observations or rewards it will guide the agent to create a 'robust' plans. This should be principally done for the range of the proton beam, using a combinations of uncertainties on the HU to Water equivalent fit and deviation of dose model from the data to create a set of potential ranges. Yet can also be applicable to the Gaussian spread of the beam, by adjusting the spread parameter and perpendicular distance used for calculation.

Generality must also examined within the model. The aim of the agent is not just to solve its training environments but future unobserved ones. This requires an extensive dataset filled with a large variety of scenarios. It also needs implemented tactically, involving the principles of curriculum learning to ensure a steady growth and prevent exceptionally unusual environments from preventing an optimal policy convergence.

A final consideration of uncertainties and generality could be through the introduction of a break in training test. This will examine the models performance using a Monte Carlo method beam model on a validation dataset and stress testing with uncertainty ranges. This can feed back additional rewards, ensuring the policy is valid with a more accurate beam and it is being trained for general treatment planning, rather than to solve individual cases.

Ethical considerations must also be contemplated. Both about the use of AI in patient treatment and its potential bias. Current datasets consist entirely of patients from wealthy western nations, tailoring the model and potentially disadvantaging those not from these backgrounds with less optimal results. Overall these topics are not yet relevant, but must not be ignored at later stages.

References

[1] Radiology Board, "National costs and resource requirements of radiotherapy: costing estimate for England from the ESTRO-HERO project," *Royal College of Radiologists*, May 2024. [Online]. Available: https://www.rcr.ac.uk/media/yxkjbjfr/rcr-policy_hero-radiotherapy-report_may-2024.pdf

[2] Brainstrust, "Proton Beam Therapy - FAQs," Brainstrust, July 2020. [Online]. Available: https://brainstrust.org.uk/wp-content/uploads/2020/07/proton-beam-therapy-faqs.pdf

[3] M. Lowe, A. Gosling, O. Nicholas, Y. Tsang, N. Sisson, and S. Gulliford, "Comparing proton to photon radiotherapy plans: Uk consensus guidance for reporting under uncertainty for clinical trials," *Clinical Oncology*, vol. 32, no. 7, April 2020.

[4] A. Bohr and K. Memarzadeh, "The rise of artificial intelligence in healthcare applications," in *Artificial Intelligence in Healthcare*. Elsevier, 2020, pp. 25–60. [Online]. Available: https://doi.org/10.1016/B978-0-12-818438-7.00002-2

[5] U. J. Muehlematter, P. Daniore, and K. N. Vokinger, "Approval of artificial intelligence and machine learning-based medical devices in the usa and europe (2015–20): a comparative analysis," *The Lancet Digital Health*, vol. 3, no. 3, pp. e195–e203, 2021.

[6] S. Li, R. Zhao, and H. Zou, "Artificial intelligence for diabetic retinopathy," *Chinese Medical Journal*, vol. 135, no. 3, pp. 253–260, 2022.

[7] J. Becker, J. A. Decker, C. Römmele, M. Kahn, H. Messmann, M. Wehler, F. Schwarz, and T. Kroencke, "Artificial intelligence-based detection of pneumonia in chest radiographs," *Diagnostics*, vol. 12, no. 6, p. 1465, 2022.

[8] M. M. Mijwil and K. Aggarwal, "A diagnostic testing for people with appendicitis

using machine learning techniques," *Multimedia Tools and Applications*, vol. 81, pp. 7011–7023, 2022.

[9] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2097–2106.

[10] N. Strodthoff and C. Strodthoff, "Detecting and interpreting myocardial infarction using fully convolutional neural networks," *Physiological measurement*, vol. 40, no. 1, p. 015001, 2019.

[11] R. Chambers, H. McNally, and S. Ingram, "PBT-Gym." [Online]. Available: https://github.com/PRECISE-RT/PBT-Gym

[12] H. Tibbals, "Gamification of proton beam therapy," January 2024.

[13] H. Paganetti, Ed., *Proton Therapy Physics*. CRC Press LLC, 2011. [Online]. Available: https://ebookcentral-proquest-com.manchester.idm.oclc.org/lib/manchester/detail.action?docID=827022

[14] B. Gottschalk, "Radiotherapy proton interactions in matter," 2018.

[15] J. Janni, "Calculations of energy loss, range, pathlength, straggling, multiple scattering, and the probability of inelastic nuclear collisions for 0.1-to 1000- mev protons," p. 451, 09 1966.

[16] M. Berger, M. Inokuti, H. Andersen, H. Bichsel, D. Powers, S. Seltzer, D. Thwaites, and D. Watt, "Stopping powers for protons and alpha particles," 1993-01-01 1993.

[17] H. A. Bethe, "Molière's theory of multiple scattering," *Phys. Rev.*, vol. 89, pp. 1256–1266, Mar 1953. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRev.89.1256

[18] M. M. Greenberg, "Reactivity of Nucleic Acid Radicals," *Advances in physical organic chemistry*, vol. 50, pp. 119–202, 2016. [Online]. Available: https://doi.org/10.1016/bs.apoc.2016.02.001

[19] E. T. Vitti and J. L. Parsons, "The radiobiological effects of proton beam therapy: Impact on dna damage and repair," *Cancers (Basel)*, vol. 11, no. 7, p. 946, Jul 2019.

[20] E. L. Alpen, "Chapter 1 - quantities, units, and definitions," in *Radiation Biophysics (Second Edition)*, second edition ed., E. L. Alpen, Ed. San Diego: Academic Press, 1998, pp. 1–10. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B978012053085450003X

[21] H. Paganetti and P. van Luijk, "Biological considerations when comparing proton therapy with photon therapy," *Seminars in Radiation Oncology*, vol. 23, no. 2, pp. 77–87, 2013, controversies in Proton Therapy. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053429612001051

[22] J. J. Wilkens and U. Oelfke, "A phenomenological model for the relative biological effectiveness in therapeutic proton beams," *Physics in Medicine Biology*, vol. 49, no. 13, p. 2811, jun 2004. [Online]. Available: https://dx.doi.org/10.1088/0031-9155/49/13/004

[23] N. G. Burnet, S. J. Thomas, K. E. Burton, and S. J. Jefferies, "Defining the tumour and target volumes for radiotherapy," *Cancer imaging : the official publication of the International Cancer Imaging Society*, vol. 4, no. 2, pp. 153–161, 2004.

[24] T. Bortfeld, "Optimized planning using physical objectives and constraints," *Seminars in Radiation Oncology*, vol. 9, no. 1, pp. 20–34, 1999.

[25] Chrome dino game. Accessed on 2024-05-04. [Online]. Available: https://chrome-dino-game.github.io/

[26] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," 2017.

[27] T. Simonini, "Proximal Policy Optimization (PPO)," Hugging Face – The AI community building the future, 2024, https://huggingface.co/blog/deep-rl-ppo.

[28] AurelianTactics. (2018, July 25) PPO Hyperparameters and Ranges. [Online]. Available: https://medium.com/aureliantactics/ppo-hyperparameters-and-ranges-6fc2d29bccbe

[29] National Electrical Manufacturers Association, "Digital Imaging and Communications in Medicine (DICOM) Standard," http://www.dicomstandard.org/, Rosslyn, VA, USA, eMA PS3 / ISO 12052.

[30] D. L. Mason *et al.*, "pydicom: An open source dicom library," https://github.com/pydicom/pydicom, [Online; accessed 2023-12-20].

[31] International Atomic Energy Agency, *Diagnostic Radiology Physics: A Handbook for Teachers and Students*. Vienna: International Atomic Energy Agency, 2014.

[32] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-baselines3: Reliable reinforcement learning implementations," *Journal of Machine Learning Research*, vol. 22, no. 268, pp. 1–8, 2021. [Online]. Available: http://jmlr.org/papers/v22/20-1364.html

[33] M. Towers, J. K. Terry, A. Kwiatkowski, J. U. Balis, G. de Cola, T. Deleu, M. Goulão, A. Kallinteris, A. KG, M. Krimmel, R. Perez-Vicente, A. Pierré, S. Schulhoff, J. J. Tai, A. J. S. Tan, and O. G. Younis, "Gymnasium." [Online]. Available: https://github.com/Farama-Foundation/Gymnasium

[34] S. Narvekar, B. Peng, M. Leonetti, J. Sinapov, M. E. Taylor, and P. Stone, "Curriculum Learning for Reinforcement Learning Domains: A Framework and Survey," *CoRR*, vol. abs/2003.04960, 2020. [Online]. Available: https://arxiv.org/abs/2003.04960

[35] H. Jiang and H. Paganetti, "Adaptation of geant4 to monte carlo dose calculations based on ct data." *Medical physics*, vol. 31 10, pp. 2811–8, 2004. [Online]. Available: https://api.semanticscholar.org/CorpusID:2997183

[36] H. Szymanowski and U. Oelfke, "Two-dimensional pencil beam scaling: an improved proton dose algorithm for heterogeneous media," *Physics in Medicine & Biology*, vol. 47, no. 18, pp. 3313–3330, 2002.

[37] J. O. Deasy, "A proton dose calculation algorithm for conformal therapy simulations based on molière's theory of lateral deflections," *Medical Physics*, vol. 25, no. 4, pp. 476–483, 1998. [Online]. Available: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1118/1.598222

[38] W. Ulmer and E. Matsinos, "Theoretical methods for the calculation of bragg curves and 3d distributions of proton beams," *The European Physical Journal Special Topics*, vol. 190, no. 1, p. 1–81, Dec. 2010. [Online]. Available: http://dx.doi.org/10.1140/epjst/e2010-01335-7

[39] J. Schuemann, S. Dowdell, C. Grassberger, C. H. Min, and H. Paganetti, "Site-specific range uncertainties caused by dose calculation algorithms for proton therapy," *Physics in Medicine & Biology*, vol. 59, no. 15, pp. 4007–4031, Aug 2014.

[40] S. Li, B. Cheng, Y. Wang, X. Pei, and X. G. Xu, "A gpu-based fast monte carlo code that supports proton transport in magnetic field for radiation therapy," *Journal of Applied Clinical Medical Physics*, vol. 25, no. 1, p. e14208, 2024. [Online]. Available: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/acm2.14208

[41] J. da Silva, R. Ansorge, and R. Jena, "Sub-second pencil beam dose calculation on gpu for adaptive proton therapy," *Physics in Medicine and Biology*, vol. 60, no. 12, pp. 4777–4795, 2015, epub 2015 Jun 4. [Online]. Available: https://doi.org/10.1088/0031-9155/60/12/4777

[42] H. Paganetti, P. Botas, G. C. Sharp, and B. Winey, "Adaptive proton therapy," *Phys Med Biol*, vol. 66, no. 22, Nov 2021.

[43] J. Unkelbach, T. C. Y. Chan, and T. Bortfeld, "Accounting for range uncertainties in the optimization of intensity modulated proton therapy," *Physics in Medicine Biology*, vol. 52, no. 10, p. 2755, apr 2007. [Online]. Available: https://dx.doi.org/10.1088/0031-9155/52/10/009

[44] B. Gottschalk, A. Koehler, R. Schneider, J. Sisterson, and M. Wagner, "Multiple coulomb scattering of 160 mev protons," *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, vol. 74, no. 4, pp. 467–490, 1993. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0168583X9395944Z

A Technical Overview of Proton Beam Delivery

Compared to traditional radiotherapy the setup of proton delivery is on a larger scale. With the requirement of protons exceeding energies of 230MeV, significant accelerators are required to bring the particles up to speed and sizable gantries to ensure delivery at accurate positioning and angles.

The first stage of treatment requires the acceleration of the particles. All protons are stripped from hydrogen gas and fed into a cyclotron where magnets perpendicular to the plane constrain particles to circular motion within two dees and Radio Frequency fields accelerate them across gaps. Initial designs of this nature were limited to 10MeV due to the protons relativistic mass, but adjustments to introduce a magnetic field with varying radius, allows energies significantly larger.

All protons are energised to the max 230MeV before having their energy adjusted for its specific purpose. To deliver a therapeutic dose to most tumours, energies must range from 100 - 230 MeV in a stepwise manner, but for tumours on the surface energies around 70-80MeV may be necessary.

Adjustments are made using the energy selection system, which drives mechanical carbon wedges in front of the beam. These gradually reduce the energy, allowing all energies in a SOBP to be delivered. Once a specific beam energy is extracted it follows along the beam line towards the gantry nozzle and the patient. On route the beam encounters two types of magnets, a quadrapole magnet focuses the beam into a 'pencil width' to remove outward spread and a dipole magnet directs the proton beam towards a specific gantry as typically a centre will have multiple treatment rooms.

The final component of the technology is the nozzle and gantry. The gantry is a large 360 degree circular component which can rotate around patient that lies on a table below and the nozzle is a device that aims the proton beam on the patient with a deflection mechanism to move the beam spot and spread it across the target area. Patients will typically receive near-daily treatment over a period of six weeks, allowing dose to be gradually applied to the tumour while peripheral regions have the opportunity to repair themselves.

B Derivation of Peripheral Dose Spread

Peripheral spread is can be modeled using Multiple Coulomb Scattering Theory, which can be theoretically expressed using Molière (17) theory. The derivation of this complete theory is complex, but by using Highlands approximation it can be expressed for electrons as:

$$\theta_0 = \theta_{\text{Highland}} = \frac{14.1 \text{ MeV}}{pv}\sqrt{\frac{t}{\rho X_0}}\left[1 + \frac{1}{9}\log_{10}\left(\frac{t}{\rho X_0}\right)\right]. \tag{7}$$

Where ρX_0 $[gcm^{-2}]$ is the mass radiation length, which electron will lose energy to Bremmstrahlung and ionsiation over , v $[cms^{-1}]$ is particle velocity and t $[cm]$ is the thickness of material.

A correction to express this heavier protons was derived by (B. Gottschalk, A. Koehler, R. Schneider, J. Sisterson, and M. Wagner) (44):

$$\theta_0 = 14.1 \text{ MeV}z\left[1 + \frac{1}{9}\log_{10}\left(\frac{t}{\rho X_0}\right)\right] \times \left(\int_0^t \left(\frac{1}{pv}\right)^2 \frac{dt'}{\rho X_0}\right)^{1/2}. \tag{8}$$

This evaluates θ_0 parameter by including particle charge z and removing the highland correction and evaluating it over the whole target, rather than in each integral step.

Combining this equation with the *Continuous slowing down approximation* (CSDA) creates an entire theoretical model for the path of protons in a material. The CSDA can be used to model a pencil beam of protons along a path and the θ_0 can be used to calculate the dose spread peripherally by placing it in this Gaussian function:

$$D(r, \phi; r_0) r\, dr\, d\phi = D_0 \frac{1}{2\pi r_0^2} e^{-\frac{1}{2}\left(\frac{r}{r_0}\right)^2} r\, dr\, d\phi \tag{9}$$

where D_0 is the dose deposition [MeV/g] at the central axis and r is the distance perpendicular to the axis. The r_0 is a product of the defined parameter, with $r_0 = L\theta_0$, where L is distance along the beam.

C Treatment Plan Comparisons

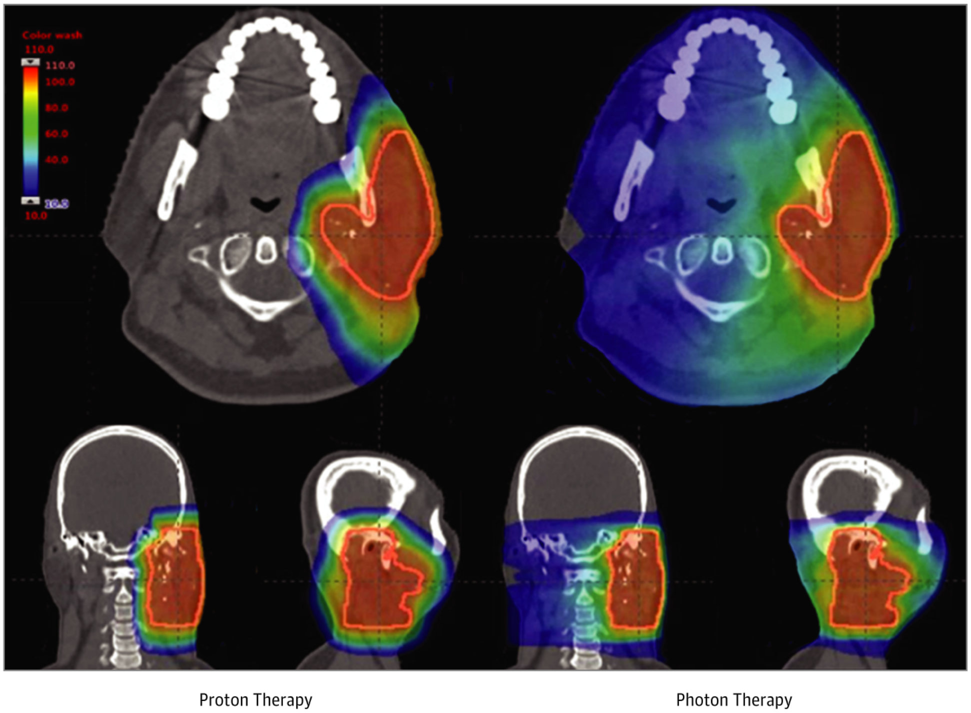


Figure 15: Shows a photon and proton therapy treatment plan for the same patient. The proton plan exhibits less dose to peripheral tissue, especially on the opposite side to the tumour and a more concise dose. Credit: Baumann BC, Mitra N, Harton JG, et al. Comparative Effectiveness of Proton vs Photon Therapy as Part of Concurrent Chemoradiotherapy for Locally Advanced Cancer. JAMA Oncol. 2020;6(2):237–246. doi:10.1001/jamaoncol.2019.4889

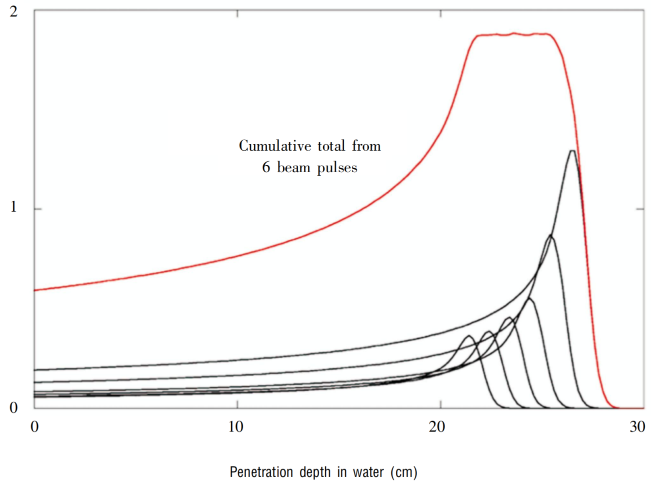


Figure 16: Spread out Bragg peak using 6 beams to create a flat dose. Credit: Liu, Hui & Chang, Joe. (2011). Proton therapy in clinical practice. Chinese journal of cancer. 30. 315-26. 10.5732/cjc.010.10529.

D Dose Model Figures

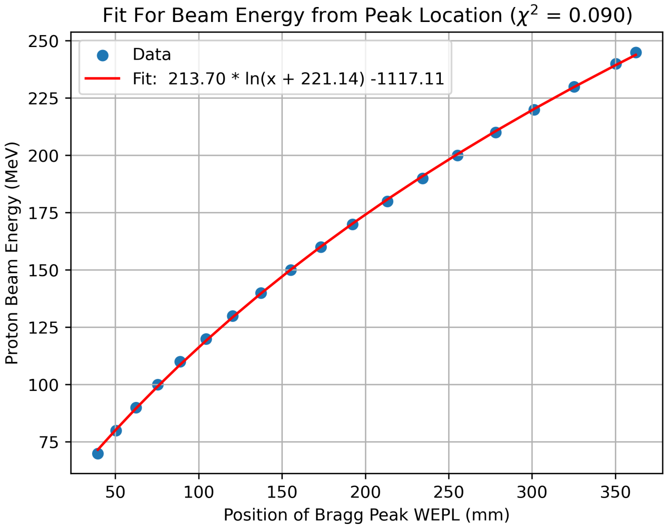


Figure 17: Fit for beam energy to distance to bragg peaks. Fit is a logarithmic function.

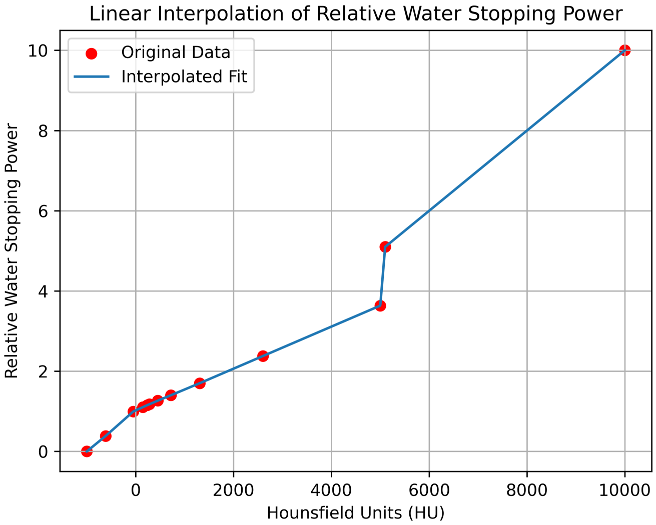
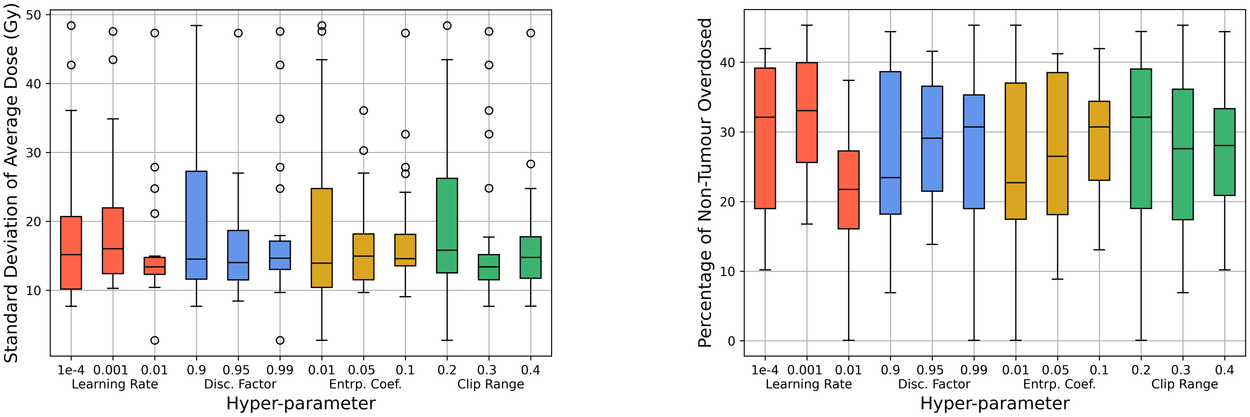


Figure 18: Conversion from Hounsfield Units to water equivalent stopping power, using clinical data in the Christie

E Additional Parameter Results



(a) Results for metric showing the standard deviation of dose within the tumour voxels

(b) Results for the metric showing percentage of non tumour voxels dosed above their maximum.

Figure 19: Results from the hyperperamter tests for the two metrics not shown within the results section.