PAPER • OPEN ACCESS

Uncertainty-aware spot rejection rate as quality metric for proton therapy using a digital tracking calorimeter

To cite this article: Alexander Schilling *et al* 2023 *Phys. Med. Biol.* **68** 194001

IOP Publishing Phys. Med. Biol. 68 (2023) 194001 https://doi.org/10.1088/1361-6560/acf5c2

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

OPEN ACCESS

RECEIVED
25 May 2023

REVISED
9 August 2023

ACCEPTED FOR PUBLICATION
31 August 2023

PUBLISHED
20 September 2023

PAPER

Uncertainty-aware spot rejection rate as quality metric for proton therapy using a digital tracking calorimeter

Alexander Schilling[1,2], Max Aehle[2], Johan Alme[3], Gergely Gábor Barnaföldi[4], Tea Bodova[3], Vyacheslav Borshchov[5], Anthony van den Brink[6], Viljar Eikeland[3], Gregory Feofilov[7], Christoph Garth[8], Nicolas R Gauger[2], Ola Grøttvik[3], Håvard Helstrup[9], Sergey Igolkin[7], Ralf Keidel[1,2], Chinorat Kobdaj[10], Tobias Kortus[1], Viktor Leonhardt[8], Shruti Mehendale[3], Raju Ningappa Mulawade[1], Odd Harald Odland[3,11], George O'Neill[3], Gábor Papp[12], Thomas Peitzmann[6], Helge Egil Seime Pettersen[11], Pierluigi Piersimoni[3,13], Maksym Protsenko[5], Max Rauch[3], Attiq Ur Rehman[3], Matthias Richter[3], Dieter Röhrich[3], Joshua Santana[1], Joao Seco[14,15], Arnon Songmoolnak[3,10], Ákos Sudár[4,16], Ganesh Tambave[17], Ihor Tymchuk[5], Kjetil Ullaland[3], Monika Varga-Kofarago[4], Lennart Volz[18,19], Boris Wagner[3], Steffen Wendzel[1], Alexander Wiebel[1], RenZheng Xiao[3,20], Shiming Yang[3] and Sebastian Zillien[1]

[1] Center for Technology and Transfer (ZTT), University of Applied Sciences Worms, D-67549 Worms, Germany
[2] Chair for Scientific Computing, University of Kaiserslautern-Landau, D-67663 Kaiserslautern, Germany
[3] Department of Physics and Technology, University of Bergen, NO-5007 Bergen, Norway
[4] Wigner Research Centre for Physics, Budapest, Hungary
[5] Research and Production Enterprise 'LTU' (RPELTU), Kharkiv, Ukraine
[6] Institute for Subatomic Physics, Utrecht University/Nikhef, Utrecht, Netherlands
[7] St. Petersburg University, St. Petersburg, Russia
[8] Scientific Visualization Lab, University of Kaiserslautern-Landau, D-67663 Kaiserslautern, Germany
[9] Department of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, NO-5020 Bergen, Norway
[10] Institute of Science, Suranaree University of Technology, Nakhon Ratchasima, Thailand
[11] Department of Oncology and Medical Physics, Haukeland University Hospital, NO-5021 Bergen, Norway
[12] Institute for Physics, Eötvös Loránd University, 1/A Pázmány P. Sétány, H-1117 Budapest, Hungary
[13] UniCamillus—Saint Camillus International University of Health Sciences, Rome, Italy
[14] Department of Biomedical Physics in Radiation Oncology, DKFZ—German Cancer Research Center, Heidelberg, Germany
[15] Department of Physics and Astronomy, Heidelberg University, Heidelberg, Germany
[16] Budapest University of Technology and Economics, Budapest, Hungary
[17] Center for Medical and Radiation Physics (CMRP), National Institute of Science Education and Research (NISER), Bhubaneswar, India
[18] Biophysics, GSI Helmholtz Center for Heavy Ion Research GmbH, Darmstadt, Germany
[19] Department of Medical Physics and Biomedical Engineering, University College London, London, United Kingdom
[20] College of Mechanical & Power Engineering, China Three Gorges University, Yichang, People's Republic of China

E-mail: aschilling@hs-worms.de

Keywords: machine learning, particle therapy, range verification, uncertainty

Abstract

Objective. Proton therapy is highly sensitive to range uncertainties due to the nature of the dose deposition of charged particles. To ensure treatment quality, range verification methods can be used to verify that the individual spots in a pencil beam scanning treatment fraction match the treatment plan. This study introduces a novel metric for proton therapy quality control based on uncertainties in range verification of individual spots. *Approach.* We employ uncertainty-aware deep neural networks to predict the Bragg peak depth in an anthropomorphic phantom based on secondary charged particle detection in a silicon pixel telescope designed for proton computed tomography. The subsequently predicted Bragg peak positions, along with their uncertainties, are compared to the treatment plan, rejecting spots which are predicted to be outside the 95% confidence interval. The such-produced spot rejection rate presents a metric for the quality of the treatment fraction. *Main results.* The introduced spot rejection rate metric is shown to be well-defined for range predictors with well-calibrated uncertainties. Using this method, treatment errors in the form of lateral shifts can be detected down to 1 mm after around 1400 treated spots with spot intensities of 1×10^7 protons. The range verification model used in this metric predicts the Bragg peak depth to a mean absolute error of

1.107 ± 0.015 mm. *Significance.* Uncertainty-aware machine learning has potential applications in proton therapy quality control. This work presents the foundation for future developments in this area.

1. Introduction

Proton therapy, first theorized by Wilson (1946), exploits the characteristics of charged particles to treat tumors with highly conformal dose distributions, while sparing healthy tissue much more effectively compared to x-ray therapy. Electromagnetic interactions cause charged particles to continuously slow down while travelling through matter, the rate of which is determined by the energy-dependent stopping power of the material traversed. The such-deposited energy increases with decreasing velocity and reaches its maximum towards the end of the particle's range, at the Bragg peak (BP).

Therefore, the deposited dose can be precisely targeted at the tumor. To cover the entire volume, one treatment strategy is pencil beam scanning (PBS). In this method, coverage is achieved by varying the direction and energy of a mono-energetic pencil beam in a way to evenly spread out the dose. Different energy levels along the same beam direction together form a spread-out Bragg peak (SOBP). Lower energies in an SOBP are irradiated with fewer primary particles, since they will additionally be covered by the dose of higher-range particles passing the spot along their path. With this strategy, we get a number of distinct treatment spots, which can individually be verified to get a quality measure over the treatment fraction (Knopf and Lomax 2013).

The Bragg curve makes particle therapy sensitive to range uncertainties, potentially causing the BP to be displaced compared to the computed treatment plan. The main causes are patient misalignment, organ motion, and inaccurate dose calculation (Paganetti 2012). To compensate for such uncertainties, it is common practice to utilize safety margins, which increase the target volume to ensure full tumor coverage. This in turn degrades the advantages of particle therapy by putting healthy tissue within the safety margins at increased risk. Alternatively, robust optimization techniques can be used to directly incorporate uncertainties into the treatment planning process and model treatment either with a probabilistic or a worst-case approach (Unkelbach *et al* 2007). Range uncertainties in proton therapy amount to approximately 2.7% + 1.2 mm on static targets (Paganetti 2012).

In the presence of such uncertainties, methods of quality control have been proposed to ensure correct dose delivery to the planned target volume. To reduce errors in patient positioning and in the conversion of Hounsfield unit to relative stopping power (RSP), proton radiography can be used right before a treatment fraction, as suggested by Schneider and Pedroni (1995). Other methods for quality control are performed shortly after the treatment fraction, such as positron emission tomography (PET) (Parodi and Enghardt 2000). However, the delay between treatment and PET measurements can potentially degrade verification ability.

In contrast, *in situ* range verification aims to assess treatment quality while it is being delivered by determining the BP positions of the individual beam spots and comparing them to the treatment plan. Such methods utilize interactions of protons traversing matter, causing the emission of secondary particles, mainly neutrons and photons, which can be detected outside the patient. The most prominent methods are based on prompt gamma (PG) detection, e.g. (Kurosawa *et al* 2012, Smeets *et al* 2012). There are also efforts to utilize neutrons (Clarke *et al* 2016, Marafini *et al* 2017), or a combination of modalities, e.g. PG and PET (Moteabbed *et al* 2011, Choi *et al* 2020). When using heavier ions for treatment, it is possible to use tracked secondary charged particles for *in situ* range verification through interaction vertex imaging (Amaldi *et al* 2010, Henriquet *et al* 2012, Gwosch *et al* 2013).

Since proton interactions do not produce secondary charged particles with sufficient residual range to exit the patient, *in situ* range verification with charged secondaries is limited to heavy ion therapy (Kraan 2015). However, a sufficiently large detector, containing converter material for neutrons and photons in addition to silicon detectors, can enable charged particle detection with enough readout yield to be usable for range verification in proton therapy.

One suitable detector for this task is the digital tracking calorimeter (DTC) designed by the Bergen pCT collaboration (Alme *et al* 2020). The readout of this high-granularity pixel detector is a large amorphous point cloud with little apparent relation to the originating treatment spot. However, there are notable differences, such as vastly different readout yield, that become visible when looking at aggregate properties. The high dimensionality and complexity of the problem is the ideal scenario to employ the help of machine learning (ML) techniques.

There are some efforts to incorporate ML into the range verification process. However, most studies are limited to improving readout by separating the signal from background noise (Lerendegui-Marco *et al* 2022,

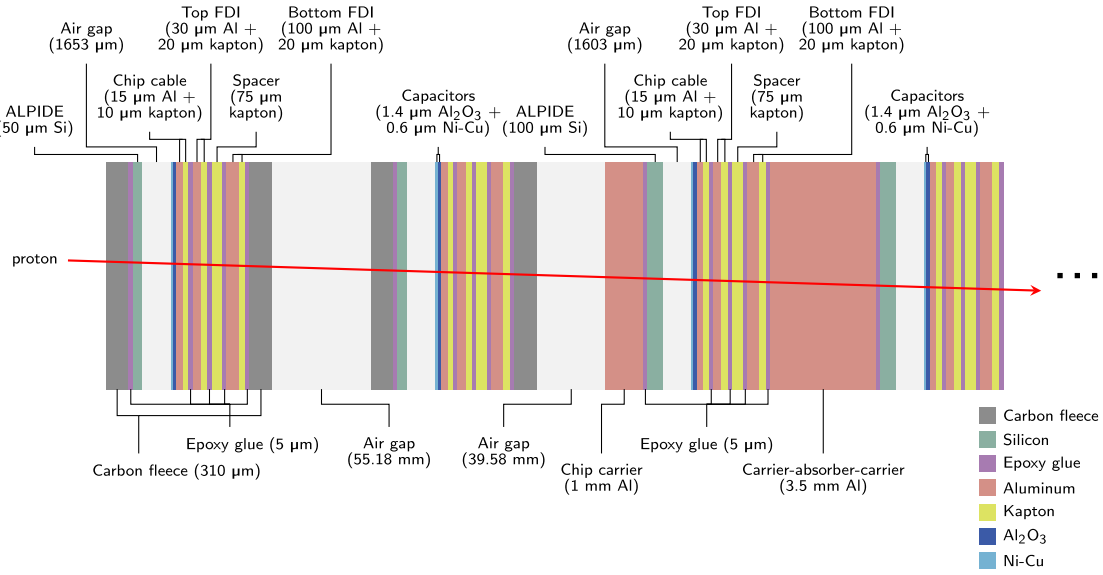

Figure 1. Material budget for the digital tracking calorimeter as implemented in the GATE Monte Carlo simulation. Figure reproduced and updated from Alme *et al* (2020) (CC BY).

Polf *et al* 2022) or ordering detector interactions (Polf *et al* 2022). More recently, methods of integrating deep learning into the range verification procedure itself have been introduced, e.g. by Jiang *et al* (2023).

This work aims to utilize uncertainty-aware ML for range verification with a detector built for charged particle detection. The range predictions are used to compute a metric assessing the quality of a treatment fraction. The efficacy of the developed methods is evaluated on Monte-Carlo-simulated data. In summary, the contributions of this work are:

(i) The development of a fully machine-learning-based range verification model for proton therapy utilizing the high-granularity silicon pixel telescope ('digital tracking calorimeter') designed by the Bergen pCT collaboration.

(ii) The introduction of a novel metric 'spot rejection rate' for quality control of proton therapy treatment fractions based on uncertainty-aware ML.

2. Material and methods

2.1. Digital tracking calorimeter

The DTC was designed for proton computed tomography (pCT) (Cormack 1963) by the Bergen pCT collaboration (Alme *et al* 2020). It is constructed with sufficient material to stop protons of 230 MeV, the target energy used in the pCT application. The energy of incoming particles is estimated by sampling the energy deposition values at 43 distinct depths along the path, reconstructing the individual tracks, and fitting the sampled values to the differentiated Bragg-Kleeman rule (Bortfeld and Schlegel 1996).

The detector consists first of two tracking layers with small material budget and 57.8 mm distance from layer to layer, to determine the incoming angle of the tracked particle with as little deflection as possible. Behind the tracker, separated by an air gap, is the 41-layer detector-absorber sandwich structure constituting the calorimeter. In between the individual calorimeter layers, aluminum absorbers are used to slow down incoming particles.

An absorber consists of the two chip carriers (1 mm aluminum) on either side, in addition to 1.5 mm aluminum solely used as additional absorber material. The first calorimeter layer is preceded only by 1 mm aluminum as carrier for the front chips of the first calorimeter layer, while all following layers are separated by the full 3.5 mm aluminum. In the tracker layers, 310 μm carbon fiber is used as a chip carrier to reduce scattering. The estimated material budget for the full detector can be seen in figure 1.

Each of the 43 sensitive layers of the DTC consists of 108 ALICE pixel detector (ALPIDE) chips (Mager 2016), which is a 3×1.5 cm^2 monolithic active pixel sensor developed by the ALICE collaboration at CERN. The chips are arranged in twelve horizontal staves, alternating between front and back-facing chips, with

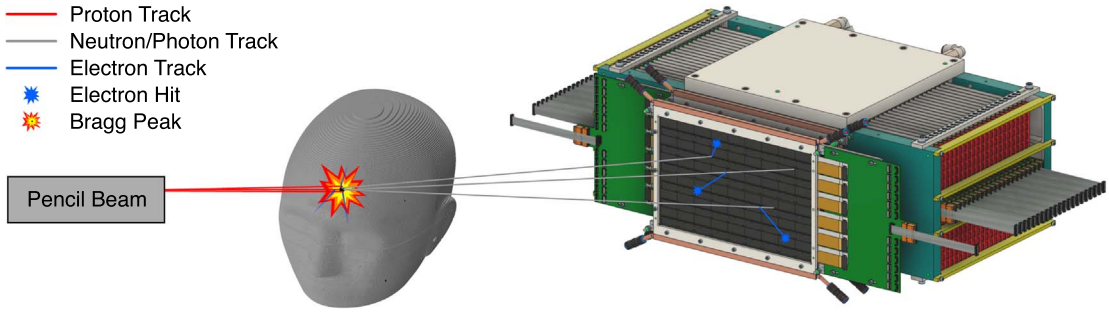

Figure 2. Schematic representation of the treatment setup as implemented in the MC simulation, with a pediatric head phantom between the pencil beam source and the DTC.

a 2 mm gap between the respective carrier plates. Each stave is a string of nine chips, creating a sensitive area of $27 \times 16.4\ \text{cm}^2$.

2.2. Monte Carlo simulation

To evaluate the expected performance of the proposed methods, we use Monte Carlo (MC) simulations with Geant4 Application for Tomographic Emission (GATE) (Jan *et al* 2004) version 9.2 and Geant4 (Agostinelli *et al* 2003, Allison *et al* 2006, 2016) version 11.0.0. We use the recommended physics list `QBBC_EMZ`.

The target at the isocenter of the simulation is the pediatric head and neck model 715-HN by CIRS Inc. (Norfolk, VA, United States), digitized into a voxel phantom by Giacometti *et al* (2017). This means the ground truth relative stopping power values of the individual materials are known and can be taken directly from table 2, 'mean experimental values' from the original paper (Giacometti *et al* 2017). It is possible to rotate the phantom around the vertical axis to simulate treatment from different angles.

To test generalizability, a second phantom can instead be used in the simulation: the visible human female (VHF) head phantom (Ackerman *et al* 1995), courtesy of the U.S. National Library of Medicine, resampled to a resolution of $1 \times 1 \times 1\ \text{mm}^3$ voxel size, scaled down to 80% to make it comparable in size to the pediatric head and to fit within the bounds of the DTC. When not otherwise specified, all tests are conducted with the dataset from the 715-HN phantom only.

As the particle source, we use a pencil beam located 500 mm from the isocenter, pointing parallel to the z-axis. To achieve total phantom coverage, the beam can be moved in the xy-plane arbitrarily, while maintaining the same beam direction. The beam shape is Gaussian with $\sigma_x = \sigma_y = 2$ mm, an angular divergence $\sigma_\theta = \sigma_\phi = 2.5$ mrad, and 3 mrad mm beam emittance.

Analogous to the pCT setup described by Alme *et al* (2020), the DTC is placed distal to the phantom starting at $z = 225$ mm with the tracker facing the beam. The DTC itself is modeled as homogenous slabs of $270 \times 164\ \text{mm}^2$ with thicknesses according to figure 1, representing a simplified geometry of the final design, where the sensitive layers will be divided into front and back-chips. Each ALPIDE layer is subdivided into three different silicon parts: the front electronics, the epitaxial layer (25 μm), and the substrate.

The epitaxial layer is the only sensitive volume in the simulation, recording hits in the form of deposited energy through creation of electron–hole pairs. For this to occur, neutral particles need to interact inside the sensitive volume, or close to it, to generate charged secondaries depositing energy to be detected. The simulation setup is schematically depicted in figure 2.

Possible simulated beam energy values range from 60.13 to 150.35 MeV (31 mm and 157 mm range in water, respectively) with 3 mm water range interval. These values are taken from the open-source treatment planning system (TPS) matRad (Wieser *et al* 2017). Relevant spots to consider should be inside the phantom and not too close to the distal edge, since those would most likely be treated from a different angle. To find those spots, we run a series of probing simulations with 1×10^5 primaries. They scan the entire xy-plane in 10mm intervals for different phantom rotations in 30 deg intervals with all available beam energies. For each lateral position in each rotation, the lowest and highest energy without any primary hits in the detector are determined, and all energies in the interval are classified as 'relevant' spots resulting in 36 258 and 35 673 data points for the 715-HN and the VHF phantom, respectively (Schilling et al 2023).

The number of primary particles per treatment spot is assumed to be in the order of 10^7 to 10^8. To get a realistic estimate for the worst-case scenario with the least detector readouts, the lower bound of 1×10^7 protons is used to run full simulations of all relevant spots.

In addition to the hits from energy deposition in the detector chips, there is a `ProductionAndStoppingActor` attached to the phantom's parent volume, which is a 200 mm cube with 1 mm voxel size, filtered to only primary particles. The actor records how many primary protons stopped in which voxel. From this data, the ground truth BP position is extracted by aggregating all voxels in each dimension and fitting a Gaussian function

$$f(x) = a \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{1}$$

along the respective dimension. The means μ_x, μ_y, μ_z constitute the BP coordinates in the simulation world, where only μ_z is considered for the range verification case, henceforth called z.

2.3. Feature extraction

Each of the simulated spots consists of an arbitrary number of hits, denoting energy deposition in sensitive volumes of the DTC, as well as the RSP image of the phantom according to its current rotation. Feature extraction is separated into 416 *detector features* and 201 *phantom features*, which are extracted from the detector hits and the RSP image, respectively.

2.3.1. Detector features

To apply ML methods, the point cloud, formed by an arbitrary number of hits, needs to be transformed into a representation that can be used as input for a neural network. This is done by extracting a fixed number of detector features from the raw data. However, the data from the simulation does not consider the properties of the used ALPIDE chip. Therefore, we perform two pre-processing steps before extracting features.

The first step eliminates duplicate hits from the data. Such hits occur for short-range electrons, which can produce multiple hit entries within the same pixel because the underlying simulation code records a hit at each step. These hits are binned together by grouping all hits from the same track in the same layer together to be able to determine the total deposited energy of the particle in this layer. The hits are aggregated by computing the mean of the x, y, and z-coordinates of the hit positions, and the sum of the energy deposition values. This procedure has been done similarly in prior work (Pettersen *et al* 2021).

Secondly, the deposited energy is discretized. The ALPIDE chip consists of binary pixels with charge sharing, leading to a cluster of pixels being activated around the incident particle. The cluster size is an indicator of the deposited energy. The reverse relationship from energy deposition E_D in keV μm^{-1} to cluster size n can be described by the following power fit from experimental data obtained from a previous experiment, rounded to the nearest integer (Pettersen *et al* 2019):

$$n = 4.23 \cdot E_D^{0.65}. \tag{2}$$

The resulting cluster size can then be turned into a discretized energy deposition by solving the equation for E_D and converting to MeV. If a hit activated a cluster of size $n = 0$, it is removed under the assumption that the deposited energy did not reach the electron threshold at the chip's diodes. This procedure emulates the energy resolution of the ALPIDE chip without computing pixel coordinates. We call this method of conversion 'pseudo-pixels'. Pseudo-pixels cannot simulate the overlap of two clusters from different hits. However, this problem is negligible, as the hit rate during treatment is too low for overlaps to occur with the readout frame duration of 10 μs used in the DTC.

The following 383 detector features are extracted:

- Total number of active pixels
- Total number of clusters (hits)
- Number of clusters over threshold (5, 20 pixels)
- Mean and standard deviation of cluster sizes
- The number of clusters of any given size (1–72)
- Mean and standard deviation of x- and y-coordinates over each layer (0–42), and the entire detector
- Number of active pixels in each layer (0–42)
- Number of clusters (hits) in each layer (0–42)
- Total energy deposition of the hits in each layer (0–42)

In addition, some of the aforementioned features are combined into 33 higher-level features through linear, cubic, and exponential ($f(x) = a \cdot \exp(-b \cdot x) + c$) fits to the histogram data, along with their mean squared residuals:

- Active pixels over layer
- Number of clusters over layer
- Total deposited energy over layer

2.3.2. Phantom features

The 201 phantom features are comprised of 200 RSP values from 1 mm slices of the phantom in the world in the direction of beam traversal symmetric to the isocenter, as well as the sum over these values. Each RSP value is assumed to represent a combined RSP value for the currently traversed millimeter slice of the phantom. This is achieved by forming the integral at each depth z over all RSP values at that depth across the phantom, weighted by a given beam model B:

$$\mathrm{RSP}(z) = \iint B(x, y, z) \cdot \mathrm{RSP}(x, y, z)\mathrm{d}x\ \mathrm{d}y \tag{3}$$

$B(x, y, z)$ describes a probability density function for the particle distribution at a given depth z (in water). The distribution depends on the traversed matter up to depth z, which is intractable for inhomogeneous targets, such as a voxelized phantom. Furthermore, it depends on the initial beam energy, which we use as one of the target variables in the multitask ML setting, making it impossible to include the quantity, to avoid a ground truth leak.

Consequently, the beam model is simplified to a fixed multivariate normal distribution matching the beam shape at the source to represent the general area which is expected to be traversed by the beam in this spot:

$$B(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{1}{2}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]\right) \tag{4}$$

with $\sigma_x = \sigma_y = 2$ mm and (μ_x, μ_y) the coordinates of the beam spot in the xy-plane.

Calculating this integral is computationally expensive because of the discontinuous nature of the RSP image. In practice, the values are computed by forming a weighted sum over the 1 mm voxels within the aperture of the detector (272×168 mm^2) with the value of the distribution in the voxel center as weight:

$$\mathrm{RSP}(z) = \sum_{y=-83.5}^{83.5} \sum_{x=-135.5}^{135.5} B(x, y) \cdot \mathrm{RSP}(x, y, z) \tag{5}$$

The final phantom feature set consists of RSP(z), $z \in \{0,...,199\}$ and the sum of the values $\mathrm{RSP}_{\mathrm{total}}$.

2.4. Machine learning

2.4.1. Model architecture and training regime

The extracted features and labels from above form the dataset used for ML. Basic outlier detection is applied, shrinking the datasets from 36 258 to 36 244 samples (715-HN) and from 35 673 to 35 670 samples (VHF). The data is split into training (70%), validation (10%), and test (20%) sets, separately for each phantom. All 617 features, as well as the labels, are scaled to $\mu = 0$ and $\sigma = 1$ before training a neural network implemented in PyTorch (Paszke *et al* 2019).

The network architecture is depicted in figure 3. The features extracted from a full simulation of a treatment spot constitute the input for three linear hidden layers with 1024, 512, and 128 units, respectively. All layers use sigmoid activation and 5% dropout (Srivastava *et al* 2014). The output is one neuron per regression task with no activation. This architecture can be used for a single task, predicting either the water range R, or the z-coordinate Z, as well as for multitask learning, where both quantities are predicted simultaneously.

In a multitask setting, the individual losses for the tasks $\mathcal{L}_R$ and $\mathcal{L}_Z$ have to be combined in order for both tasks to contribute to the training by influencing backpropagated gradients. One way to do this is to form a weighted sum of losses. The weights are set to 0.4 and 0.6 for R and Z, respectively, to ensure that the more important (and more difficult) task Z has more influence during training, resulting in slightly lower errors than equal weights.

Instead of manually setting weights for the tasks, we additionally use an alternative approach with automatically learned weights based on homoscedastic (task) uncertainty, introduced by Kendall *et al* (2018). The task uncertainty σ^2_{task} is an additional parameter for optimization, which in practice describes the log-variance for numerical stability, as recommended by the authors (Kendall *et al* 2018). The resulting loss term is

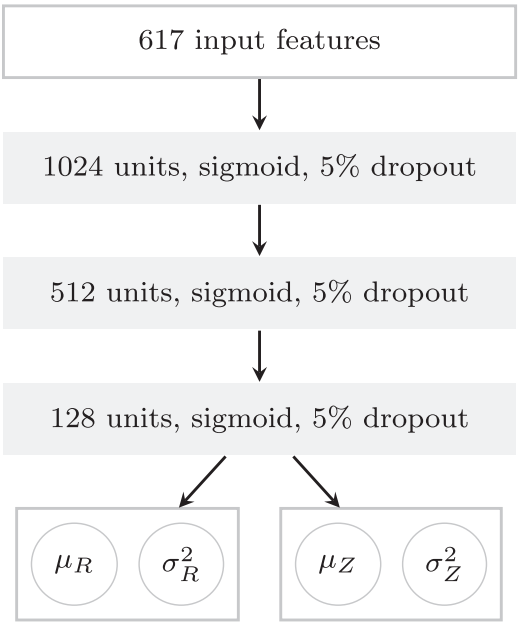

Figure 3. Neural network architecture for multitask learning, predicting both water range R and z-coordinate Z simultaneously, along with their respective variances.

given as

$$\mathcal{L} = \frac{1}{2\sigma^2_{\text{task},R}}\mathcal{L}_R + \frac{1}{2\sigma^2_{\text{task},Z}}\mathcal{L}_Z + \log \sigma_{\text{task},R}\sigma_{\text{task},Z}. \tag{6}$$

Both the weighted sum and the homoscedastic uncertainty-weighted loss are evaluated for the multitask setting.

The Adam optimizer (Kingma and Ba 2015) is used for training with learning rate 1×10^{-4} and weight decay 1×10^{-4}. The batch size is set to 32 with random shuffle in each epoch. The networks are trained for 500 epochs, which is sufficient for convergence in all learning scenarios.

2.4.2. Uncertainty

In safety-critical applications, it is paramount to have an accurate uncertainty estimate to make decisions with confidence, while discarding or manually re-evaluating cases where the machine is highly uncertain. Uncertainty can be decomposed into three parts: epistemic (model) uncertainty, heteroscedastic aleatoric (data) uncertainty, and homoscedastic aleatoric (task) uncertainty. Task-dependent uncertainty boils down to a constant in any given task. Therefore, we only model epistemic and heteroscedastic uncertainty. The total predictive variance σ^2_{total} is then given as

$$\sigma^2_{\text{total}} = \sigma^2_{\text{model}} + \sigma^2_{\text{data}}. \tag{7}$$

Model uncertainty for an architecture with dropout can be estimated through MC dropout (Gal and Ghahramani 2016). Dropout was originally conceived as a regularization technique during training. However, it can be used for estimating model uncertainty by using it during inference and sampling over multiple runs with different randomly dropped-out units. The mean and variance of the samples constitute the final prediction and their epistemic uncertainty. We use 100 MC forward passes for inference.

In addition to predicting the target value, for each task an additional output is used to interpret the result as a normal probability distribution with the target value constituting the mean, and the second output predicting the log-variance of the distribution (see figure 3), as described in (Kendall and Gal 2017). This variance is interpreted as heteroscedastic aleatoric uncertainty. The loss for each individual task needs to be adjusted to describe the likelihood of the predicted distribution. Hence, the Gaussian negative log-likelihood loss (Nix and Weigend 1994) is used:

$$\mathcal{L}_R = \mathcal{L}_Z = \frac{1}{2}\left(\log(\sigma^2) + \frac{\|y_i - \hat{y}_i\|^2}{\sigma^2}\right). \tag{8}$$

2.4.3. Uncertainty calibration

The epistemic uncertainty through MC dropout is innately uncalibrated (Gal and Ghahramani 2016). To evaluate the quality of calibration for the trained uncertainties, we can plot the observed confidence interval over the expected confidence interval and see how closely it follows the diagonal (Kuleshov *et al* 2018). Since the predictions follow a Gaussian assumption, we can compute the observed confidence interval by counting how many predictions fall within the corresponding σ-interval around the predicted mean, e.g. for the 95% confidence, we would compute the fraction of absolute errors less than $1.96\sigma_{\text{total}}$ to find the observed confidence interval.

If the observed confidence intervals do not match the expected confidence interval, Kuleshov *et al* (2018) proposed fitting an estimator R on a separate calibration dataset. R translates from the expected confidence to the point where this confidence level is actually observed. To this end, the points on the aforementioned curve are fit to an isotonic regression as recommended by the authors (Kuleshov *et al* 2018). Whenever confidence level p is evaluated thereafter, we evaluate $R(p)$ instead.

2.5. Simulated treatment errors

To properly evaluate the range verification method, we simulate error cases, which will be used to assess how large the treatment error must be for the proposed method to be able to detect it, and to investigate the consistency of a metric, which is supposed to increase proportional to the treatment error.

The full simulation runs are very time-consuming. However, to simulate lateral shifts from the planned spot to the actual spot, it is possible to re-use the detector readout of the previously conducted full simulations and compute a differently planned spot for it. For each spot in the test set, lateral shifts in each direction along the x- and y-axes are considered, while maintaining the beam direction parallel to the z-axis. The shift distance is between 1 mm and 10 mm in 1 mm intervals to cover the entire gap between two simulated spots.

The dataset then consists of the shifted spot, which is the desired spot according to the treatment plan, for which the phantom features are re-computed. Additionally, the detector features for the actual spot are kept as is. The new spot will, however, have a new ground truth BP location, which is determined by simulating treatment with 2×10^4 primary protons, yielding sufficient statistics to accurately determine the new BP coordinates.

2.6. Spot rejection rate

Uncertainty-aware predictions evoke a method to determine whether a spot was correctly treated: If the planned Z-position is outside the predicted 95% confidence interval ($\mu \pm 1.96\sigma_{\text{total}}$), treatment is considered suboptimal and the spot is rejected. However, assessing the quality of each individual spot is highly volatile due to the stochastic nature of the physical processes involved in producing detector hits. Additionally, if the 95% confidence interval is considered, it is expected to have 5% of the treatment spots rejected, even in case of successful treatment. To get a more meaningful measure of treatment quality, it is important to consider the entire treatment fraction.

Over all spots in a treatment fraction, we can define the rate of rejected spots. We call this *rejection rate* $rr \in [0, 1]$. It is defined for a collection of ground truth treated spot depths $z \in Z$, along with their detector readout $x \in X$ and the range verification predictor $f(x)$ estimating BP depth $\hat{z}$ and uncertainty σ^2. rr is then defined as the number of rejected spots over the number of treated spots:

$$rr = \frac{|\{z_t \in Z \mid 1.96\sigma < |z_t - f(x_t)|\}|}{|Z|}. \tag{9}$$

We call such a metric well-defined, if $rr \propto q^{-1}$ with treatment quality q. Therefore, rr is well-defined for any predictor $f(x)$ with calibrated uncertainties σ^2 at the 95% confidence interval, which we can achieve with the uncertainty calibration described in section 2.4.3. We then evaluate the confidence interval $R(0.95)$ instead, which is perfectly calibrated given enough i.i.d. data (Kuleshov *et al* 2018).

In real-world situations, we are missing the ground truth BP depth z. We use the planned depth from the treatment plan instead. As soon as treatment errors occur, the predicted range will generally be further away from the planned depth, while the input data is still from a similar distribution as the training data, producing similar uncertainties. This leads to more spots being rejected because their error now exceeds the predicted uncertainty. Hence, rr will increase from the baseline $rr = 0.05$, signifying decreasing treatment quality.

The rejection rate is innately a statistical quantity, which tends to its expected value with an increasing number of spots. As such, rr needs to be accompanied by a statistical confidence estimate to be able to discard values with insufficient significance. This can be achieved with a one-sample t-test (Student 1908) with $H_0 = 0.05$ and the alternative hypothesis $H_a > H_0$. The tested sample is a series of ones and zeros for each spot that was rejected (one) or not rejected (zero). A t-test yields a p-value, which can be compared to a desired significance value α to decide if the null-hypothesis is rejected, meaning the resulting rr measure is reliable.

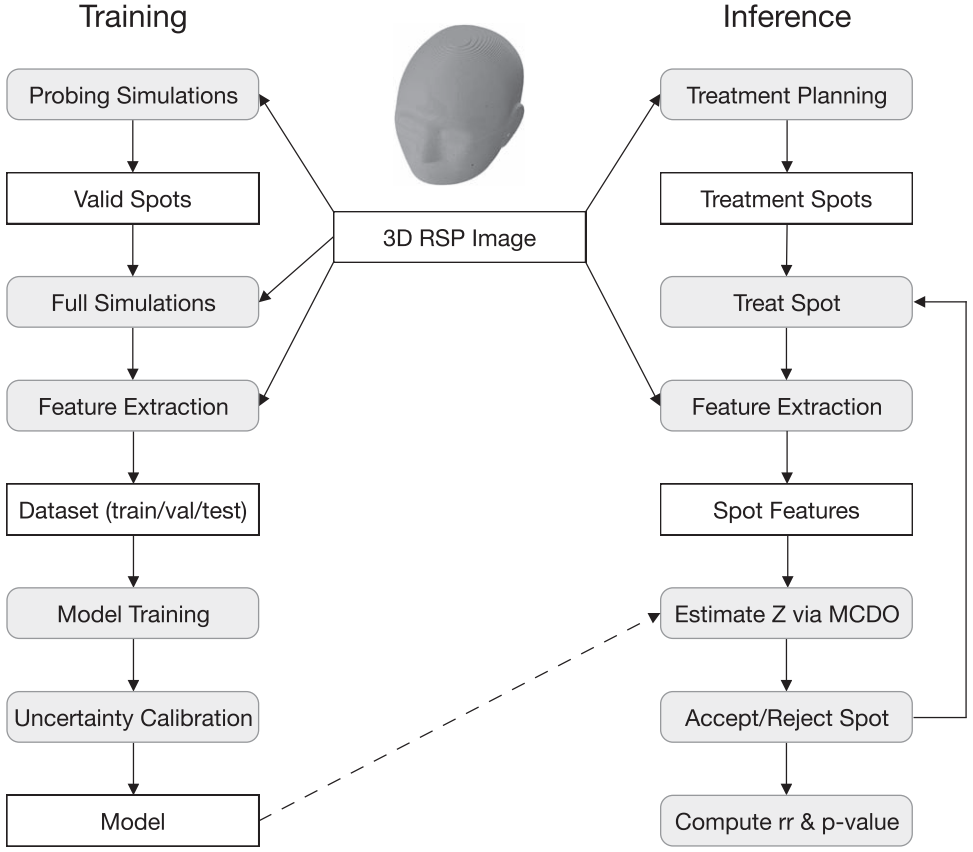

Figure 4. Illustration of the complete workflow presented in this work, starting with the training phase on the left to produce a range estimator model, which is applied in the inference phase on the right to ultimately result in the rejection rate *rr* as a quality metric for a treatment fraction.

For the choice of α, we propose using two values to signify 'potential' error ($\alpha = 0.05$), where additional imaging is recommended after the fraction, and 'certain' error ($\alpha = 0.01$), which could trigger mid-fraction treatment termination in future applications.

2.7. The complete workflow

Figure 4 illustrates how all the presented components interact. First, during the training phase, we use the probing simulations to determine potential treatment spots, for which to do full simulations of the detector response with 1×10^7 primaries each. From the full simulations and the 3D RSP image, we can extract detector and phantom features, respectively. Together, these constitute the dataset used to train, calibrate, and test the model for this patient.

The inference phase is what happens during treatment of the patient. It begins with treatment planning based on the 3D RSP image, which yields all the spots we want to treat, possibly from multiple different beam angles. Then treatment delivery starts, where the same steps are repeated for each individual spot. Features are extracted from the 3D RSP image and the detector readout to generate a dataset analogous to the training data. These features are used as input for the model repeatedly, using MC dropout to estimate the Bragg peak position and the predictive uncertainty. With that, it can be determined whether the spot is accepted or rejected.

Finally, with all spots either accepted or rejected, the rejection rate *rr* and the *p*-value are computed. Alternatively, it is possible to update *rr* and the *p*-value after each spot is delivered and the decision is made.

3. Results

3.1. Prediction error

The regression networks in the 4 different scenarios (single task *R*, single task *Z*, multitask weighted sum, and multitask homoscedastic) are run 10 times each with different random seeds and the resulting evaluation scores are recorded. Mean absolute error (MAE) and root mean squared error (RMSE) scores are listed in table 1 along with the standard deviation across runs. The results indicate that the multitask scenarios outperform the single-task setup for both, *R* and *Z*. Adding the auxiliary task *R* leads to a mean absolute error (MAE) of 1.087 mm and

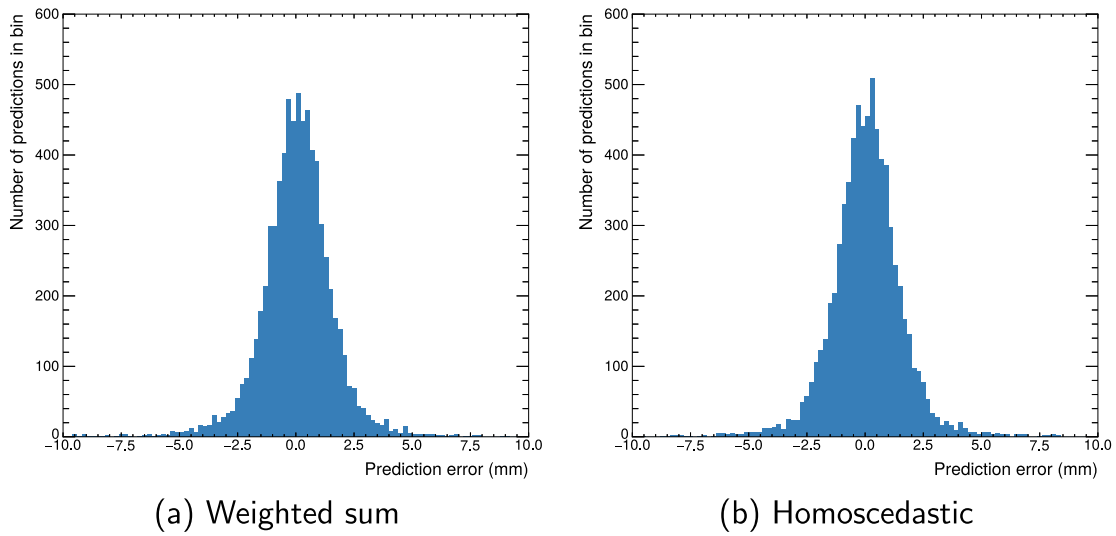

Figure 5. Histogram of prediction errors for Z in different multitask settings, (a) weighted sum of losses and (b) homoscedastic uncertainty-weighted loss.

Table 1. MAE and RMSE scores ± 1 standard deviation in mm computed on the test set over 10 runs with different seeds for different learning scenarios and both targets, R and Z.

Model	MAE_R	MAE_Z	$RMSE_R$	$RMSE_Z$
Single task	0.822 ± 0.023	1.254 ± 0.021	1.082 ± 0.029	1.745 ± 0.025
Weighted sum	0.763 ± 0.013	1.087 ± 0.020	0.990 ± 0.015	1.526 ± 0.030
Homoscedastic	0.782 ± 0.009	1.107 ± 0.015	1.020 ± 0.011	1.559 ± 0.023

Table 2. MAE and RMSE scores ± 1 standard deviation in mm computed with different combinations of training and test sets over 10 runs with different seeds for both targets, R and Z, trained with the weighted sum of losses.

Train	Test	MAE_R	MAE_Z	$RMSE_R$	$RMSE_Z$
715-HN	715-HN	0.763 ± 0.013	1.087 ± 0.020	0.990 ± 0.015	1.526 ± 0.030
VHF	VHF	0.826 ± 0.022	1.200 ± 0.017	1.072 ± 0.032	1.696 ± 0.029
715-HN	VHF	2.220 ± 0.047	5.712 ± 0.026	3.710 ± 0.034	7.270 ± 0.031
VHF	715-HN	1.666 ± 0.028	4.183 ± 0.037	2.250 ± 0.036	5.480 ± 0.040

1.107 mm for weighted sum and homoscedastic, respectively. However, the homoscedastic uncertainty-weighted loss appears slightly more stable, with a lower standard deviation across runs.

It is notable that the training loss in the homoscedastic setup converges around -4.5×10^{19}, which could indicate potential numerical instability for this approach. This could be a result of combining the Gaussian negative log-likelihood loss with the homoscedastic uncertainty-weights, since both of the components include a log-term for regularization.

Figure 5 shows the error distribution histograms for task Z in both multitask settings. The distribution is approximately Gaussian, centered around 0, indicating the absence of a systematic error in any direction. Similarly, the prediction plot in figure 6 shows a distribution around the diagonal.

3.2. Uncertainty calibration

Figure 7 shows the calibration plots for the epistemic σ^2_{model} and aleatoric σ^2_{data} uncertainties before and after calibration in both multitask settings. The plots show that even when calibrated, the individual uncertainty components alone deviate from the expected confidence level in the critical region for rr computation, around 95%, thereby confirming the necessity to model both, epistemic and aleatoric uncertainty, for a well-calibrated model. Either component tends to be over-confident, i.e. to underestimate the uncertainty, which is undesirable for safety-critical applications where the decision dictates further action.

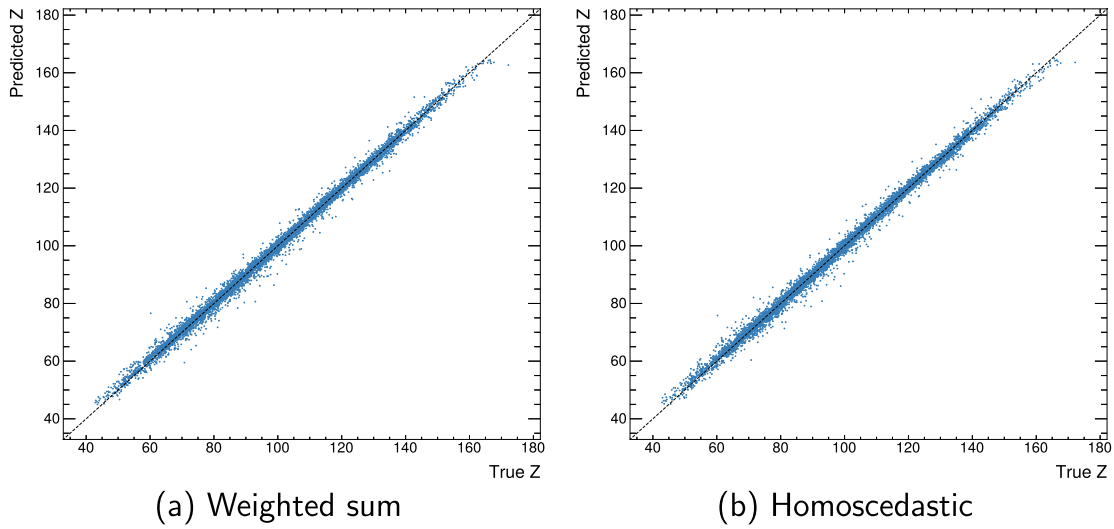

Figure 6. Predictions of Z over their true values in different multitask settings, (a) weighted sum of losses and (b) homoscedastic uncertainty-weighted loss.

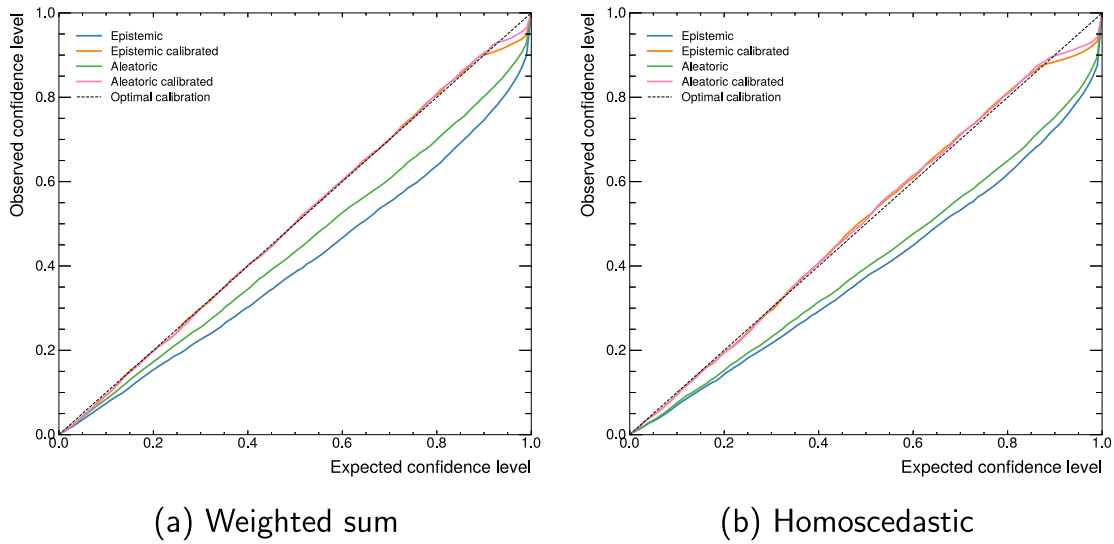

Figure 7. Uncertainty calibration curves for the individual components, epistemic and aleatoric uncertainty, for task Z in different multitask settings, (a) weighted sum of losses and (b) homoscedastic uncertainty-weighted loss, before and after calibration.

When both uncertainties are modeled and combined, the resulting uncertainty can be calibrated to follow the expected confidence levels closely for both tasks, R and Z, as depicted in figure 8. Both multitask settings perform similarly well when comparing the capability of estimating their uncertainties. Task R is generally slightly more under-confident than Z. However, after calibration, there is no notable difference between the tasks.

3.3. Spot rejection rate

Figure 9 shows rr over different degrees of treatment error for different predictors before and after uncertainty calibration. For a predictor with well-calibrated uncertainties, we expect $rr = 0.05$ for correct treatment (0 lateral shift) to reflect the 95% confidence interval used in the definition of rr. With uncalibrated uncertainties, this assumption does not hold for any of the predictors. After calibration, only the single task setting differs from 5% for correct treatment.

Increasing lateral shift compared to the treatment plan leads to monotonically increasing rejection rates. Hence, the condition for a well-defined metric $rr \propto q^{-1}$ holds for all tested learning scenarios. Both multitask

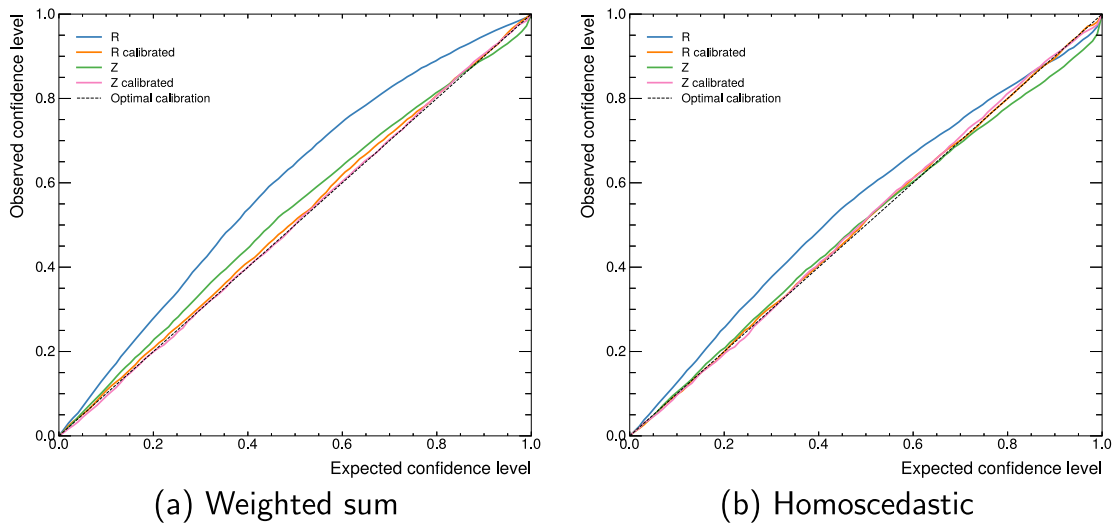

Figure 8. Uncertainty calibration curves for the total uncertainty for both tasks, *R* and *Z*, in different multitask settings, (a) weighted sum of losses and (b) homoscedastic uncertainty-weighted loss, before and after calibration.

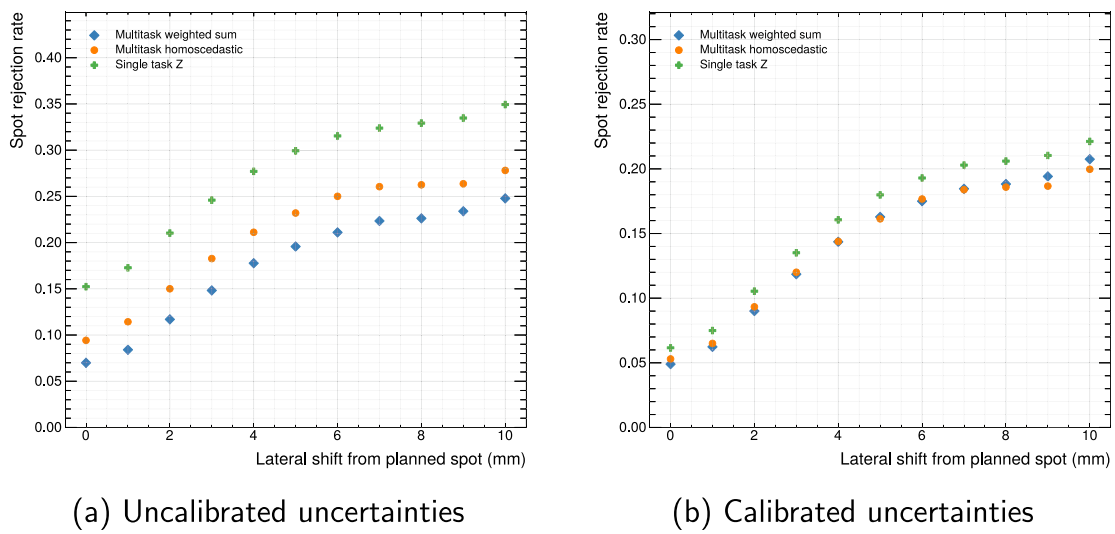

Figure 9. Rejection rates of different learning scenarios for the 95% confidence interval of predictions, (a) uncalibrated and (b) calibrated, with increasing lateral shift, representing treatment error.

settings show very similar *rr* across the entire lateral shift spectrum after calibration. We can observe that a shift of as little as 1 mm can be detected through the *rr* metric.

Further, the monotonic increase, even in cases where uncertainties are not well-calibrated, means that calibration can be deferred to the metric instead of the predictive uncertainty, by simply subtracting rr_0, the rejection rate predicted for correct treatment. The metric then describes poor treatment as soon as the value is positive, instead of the threshold being 0.05.

3.4. Required treatment spots

Using t-tests allows us to compute the minimum number of spots required for statistical significance ($\alpha = 0.05$ or $\alpha = 0.01$). Different spot counts are evaluated by repeatedly sampling randomly from all predicted spots of the test set with a certain lateral shift distance. Figure 10 shows the resulting mean *p*-values computed for 1 mm, 2 mm, and 3 mm shifts with 1×10^4 samples for each spot count.

The plot shows that for the weighted sum of losses, 1 mm shifts can be detected with significance $\alpha = 0.05$ starting at around 2000 spots, 2 mm shifts starting at around 250 spots, and 3 mm shifts starting at around 150 spots. Larger shifts follow the trend of requiring even fewer spots to reach statistical significance. With

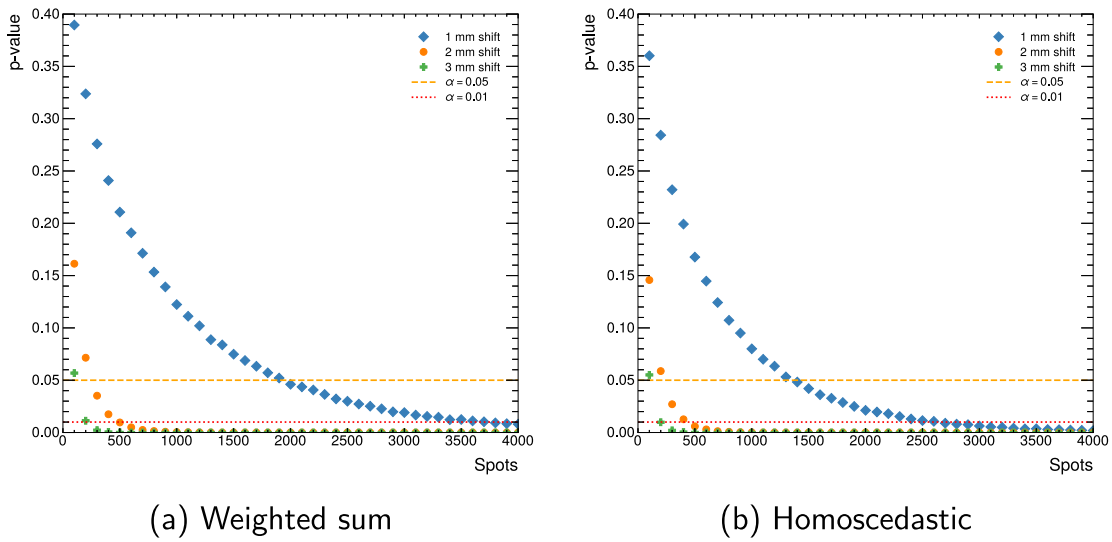

Figure 10. Average *p*-values for t-tests sampled 10 000 times for different spot counts with 1, 2, and 3 mm lateral shift in different multitask settings, (a) weighted sum of losses and (b) homoscedastic uncertainty-weighted loss.

homoscedastic uncertainty-weighted losses, the required spots for the detection of 1 mm shifts is much lower on average, starting at around 1400 spots.

For more confident predictions for the detection of 1 mm shifts with significance $\alpha = 0.01$, 3700 spots and 2700 spots are required for weighted sum of losses and homoscedastic uncertainty-weighted loss, respectively.

3.5. Generalizability

We differentiate between two different concepts of generalizability. First, the general applicability of the proposed workflow for different targets through patient-specific re-training. This process is time-consuming because of the large amount of simulations required to generate the training dataset. Nevertheless, proving viability of the workflow on a second phantom would underscore its effectiveness.

The second concept of generalizability is the transferability of a model trained on one patient to another, previously unseen patient. This approach is akin to one-shot learning, which aims to learn a general model from one or a few examples. To test the one-shot generalizability of the proposed model, it is trained on one phantom and evaluated on another. Table 2 shows the result for both cases, re-training completely on a patient-by-patient basis, as well as training on one phantom and evaluating on the other.

When the same phantom is used for training and evaluation, the error scores are close to each other, VHF being slightly higher than 715-HN with 1.200 ± 0.017 mm MAE on task Z. However, the uncertainty calibration works equally well, leading to similar *rr* curves for increasing treatment error as in the 715-HN phantom. Figure 11 shows the uncalibrated and calibrated rejection rates for the VHF phantom in different learning scenarios.

Concerning the one-shot generalizability from one phantom to another, the table shows a noticeable increase in errors, with MAE above 4.1 mm for both cases. Additionally, the uncertainties are no longer calibrated, yielding $rr \approx 72\%$ and $rr \approx 49\%$ without any treatment error for training on 715-HN and VHF, respectively.

4. Discussion

4.1. Uncertainty calibration

rr is a well-defined metric for perfect uncertainties, which is usually not a reachable goal in practice. However, the uncertainty calibration considers the entire range of confidence intervals, while the definition of *rr* only considers the 95%-interval. It may be beneficial to incorporate the entire range of confidence intervals, since the calibration is close to ideal across the board, as shown in section 3.2.

Furthermore, there may be a need for an additional metric to control the rejection rate. If the data is out-of-distribution for the training set, the epistemic uncertainty will increase. One possibility for out-of-distribution data could be cases where many primaries reach the detector because the beam is erroneously placed at a much thinner position of the patient. It needs to be ensured that in such cases, the uncertainty is not too large to accept

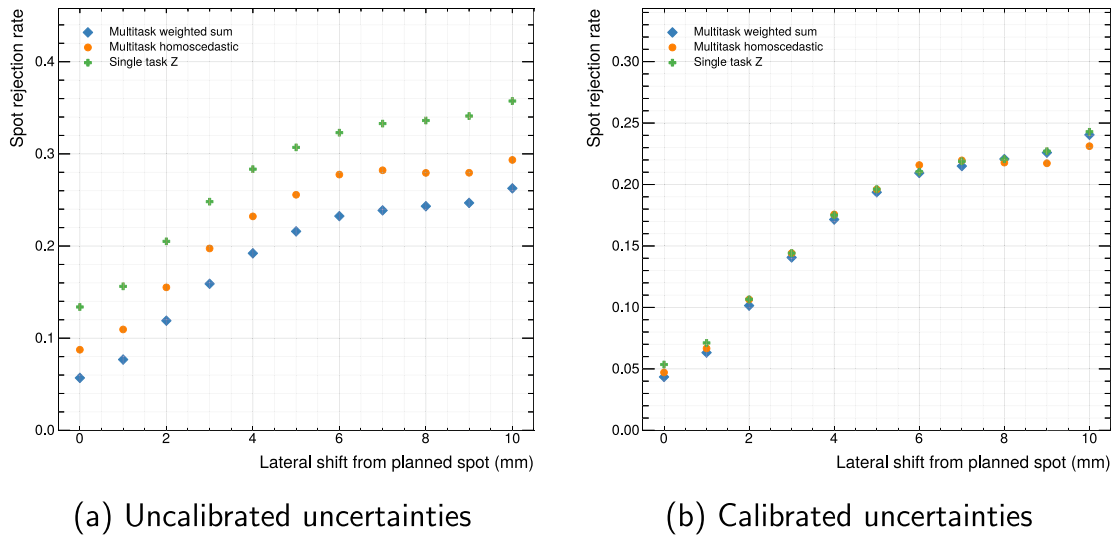

Figure 11. Rejection rates of different learning scenarios on the VHF phantom for the 95% confidence interval of predictions, (a) uncalibrated and (b) calibrated, with increasing lateral shift, representing treatment error.

more spots than it should. A foundation for such a metric could be the measure of sharpness, defined by Kuleshov *et al* (2018) as the mean of the predicted uncertainties σ^2.

4.2. Generalizability

The results in section 3.5 indicate viability for the proposed workflow with patient-specific re-training of the model. However, one-shot learning, with a single phantom providing the training data, does not generalize well to a second phantom with the current model architecture. It should be the focus of future work to either improve one-shot or few-shot generalizability, or to find a more easily generalizable workflow to make this method viable for clinical application. To this end, model architectures with translation invariance are worth evaluating. Additionally, transfer learning can be employed to fine-tune a model to new patients. This would require much less data than re-training from scratch and therefore reduce computation time significantly.

Another challenge for generalizability is the gap between simulation and reality. It needs to be investigated if MC data can be used for training or if it does not match experimental data well enough. Since the 715-HN phantom used in this work is modeled after a physical counterpart, it may be possible to use this as a foundation for calibrating a model trained on simulations to the real world.

4.3. Clinical applicability

If the generalization described in the previous section can be achieved without loss of precision, the *rr* metric could be applied to the clinical setting. Detection of errors of 1 mm lateral shift with $1 \cdot 10^7$ primaries per spot is comparable to previously introduced range verification methods in terms of the detected error distance, while requiring lower spot intensities. In particular, other studies report, e.g. 1–3 mm with 1×10^8 primaries (Lerendegui-Marco *et al* 2022), 2 mm with $1 \cdot 10^8$ primaries (Draeger *et al* 2018), and 1 mm with 1.35×10^8 (Tian *et al* 2018) primaries.

Another aspect of clinical applicability is the minimum number of spots required for statistical significance of the *rr* metric. The average number of spots required to detect 1 mm shifts with homoscedastic uncertainty-weighted losses is at around 1400. Many clinical treatment plans exceed this requirement. Some example plans can be found in Maradia *et al* (2022), where the minimum spot count is around 8400, well above the required threshold.

However, the number of spots is highly dependent on the size of the tumor. Furthermore, treatment plans can be created with a spot reduction technique introduced by van de Water *et al* (2020). This has the potential to reduce the number of spots by a factor of up to 20. The plans in the above example are reduced to around 800 (Maradia *et al* 2022). These plans will most likely still detect shifts of 2 mm or more, but not necessarily down to 1 mm.

The statistical significance levels in the *rr* metric are currently arbitrarily chosen to be $\alpha = 0.05$ and $\alpha = 0.01$. It is unclear, if these represent medically relevant thresholds, or if better values have to be found. Alternatively, it

is possible to simply compute the p-value of the t-test and display it to the physician to decide on further action on a case-by-case basis.

Additionally, the combination of the rejection rate with a statistical significance measure enables on-line evaluation with the potential of automated intervention in case an error is detected with confidence. To determine the feasibility for this approach, a computation time measurement was added to the current prototypical Python implementation. The entire inference workflow can be executed in 105 ms on average in all learning scenarios using a single core of an AMD EPYC 7F72 CPU and an NVIDIA A100 GPU. Assuming a spot delivery time of around 100 ms, it is well in the realm of possibility to apply the workflow on-line after code optimization.

4.4. R as input

The beam in a proton therapy facility is monitored by internal control mechanisms. We can, therefore, assume the beam energy and corresponding range in water R to be a known quantity. This enables us to use R as an input feature instead of using a multitask learning scheme with R as an output. While this has the potential of improving the MAE of the prediction, it may cause the network to learn to behave like a TPS and simply compute the expected BP location based on the beam energy as well as the RSP image, entirely ignoring detector readout. This means the prediction would always correspond to the treatment plan, ignoring what actually happened during treatment.

To avoid this scenario, R is instead used to improve the predictor by serving as an auxiliary task. However, there may be other possibilities to model a predictor preventing it from learning this behavior, which is a possible avenue for future work.

One advantage of including R as an input rather than an output is the potential to compute more accurate phantom features without a ground truth leak, as described in section 2.3.2. Because the beam diverges depending on its energy and thus the computed combined RSP value adjusts based on the current beam width at depth z, this information can lead to better features for a more accurate prediction, even without using R as an input directly.

5. Conclusion

In this work, we presented *spot rejection rate rr* as a quality metric for proton therapy, based on ML and predictive uncertainties for range verification in PBS. We show that the metric is well-defined for well-calibrated uncertainties, and that it is independent of the range predictor.

An MC study evaluating its efficacy shows promising results for the measurement of treatment quality, with lateral shifts introduced as error. The metric monotonically increases with increasing treatment error. Modeling uncertainties in deep neural networks with the prediction of a Gaussian variance and the usage of MC dropout proves to be a good architecture for a range predictor. With subsequent calibration, the model produces well-calibrated uncertainties and can be used for the *rr* metric.

Further, this study showed the feasibility of using a detector meant for charged particle detection for range verification in proton therapy, which does not produce any charged secondary particles directly.

Acknowledgments

This work was supported by the German federal state Rhineland-Palatinate (Forschungskolleg SIVERT) and by the Research Council of Norway (Norges forskningsråd) and the University of Bergen, grant number 250 858.

Simulations were executed on the high-performance cluster Elwetritsch at the University of Kaiserslautern-Landau, which is part of the Alliance of High-Performance Computing Rhineland-Palatinate (AHRP). We kindly acknowledge the support of the regional university computing center (RHRZ).

The ALPIDE chip was developed by the ALICE collaboration at CERN.

The open access publication was supported by the publications fund of the University of Applied Sciences Worms.

The authors have no conflicts of interest to disclose.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://doi.org/10.5281/zenodo.8192778.

ORCID iDs

Alexander Schilling https://orcid.org/0000-0001-8802-3247
Max Aehle https://orcid.org/0000-0002-6739-5890
Tobias Kortus https://orcid.org/0000-0002-0987-8544
Helge Egil Seime Pettersen https://orcid.org/0000-0003-4879-771X
Ákos Sudár https://orcid.org/0000-0001-6529-1636
Lennart Volz https://orcid.org/0000-0003-0441-4350
Alexander Wiebel https://orcid.org/0000-0002-6583-3092

References

Ackerman M J, Spitzer V M, Scherzinger A L and Whitlock D G 1995 The visible human data set: an image resource for anatomical visualization *Medinfo* **8** 1195–8
Agostinelli S *et al* 2003 Geant4a simulation toolkit *Nucl. Instrum. Methods Phys. Res.* A **506** 250–303
Allison J *et al* 2006 Geant4 developments and applications *IEEE Trans. Nucl. Sci.* **53** 270–8
Allison J *et al* 2016 Recent developments in geant4 *Nucl. Instrum. Methods Phys. Res.* A **835** 186–225
Alme J *et al* 2020 A high-granularity digital tracking calorimeter optimized for proton CT *Front. Phys.* **8** 568243
Amaldi U, Hajdas W, Iliescu S, Malakhov N, Samarati J, Sauli F and Watts D 2010 Advanced quality assurance for cnao *Nucl. Instrum. Methods Phys. Res.* A **617** 248–9
Bortfeld T and Schlegel W 1996 An analytical approximation of depth-dose distributions for therapeutic proton beams *Phys. Med. Biol.* **41** 1331–9
Choi H J, Jang J W, Shin W-G, Park H, Incerti S and Min C H 2020 Development of integrated prompt gamma imaging and positron emission tomography system for in vivo 3D dose verification: a monte carlo study *Phys. Med. Biol.* **65** 105005
Clarke S D, Pryser E, Wieger B M, Pozzi S A, Haelg R A, Bashkirov V A and Schulte R W 2016 A scintillator-based approach to monitor secondary neutron production during proton therapy *Med. Phys.* **43** 5915–24
Cormack A M 1963 Representation of a function by its line integrals, with some radiological applications *J. Appl. Phys.* **34** 2722–7
Draeger E, Mackin D, Peterson S, Chen H, Avery S, Beddar S and Polf J C 2018 3D prompt gamma imaging for proton beam range verification *Phys. Med. Biol.* **63** 035019
Gal Y and Ghahramani Z 2016 Dropout as a bayesian approximation: representing model uncertainty in deep learning *Int. Conf. on Machine Learning* pp 1050–9
Giacometti V, Guatelli S, Bazalova-Carter M, Rosenfeld A B and Schulte R W 2017 Development of a high resolution voxelised head phantom for medical physics applications *Phys. Med.* **33** 182–8
Gwosch K, Hartmann B, Jakubek J, Granja C, Soukup P, Jäkel O and Martišíková M 2013 Non-invasive monitoring of therapeutic carbon ion beams in a homogeneous phantom by tracking of secondary ions *Phys. Med. Biol.* **58** 3755–73
Henriquet P *et al* 2012 Interaction vertex imaging (ivi) for carbon ion therapy monitoring: a feasibility study *Phys. Med. Biol.* **57** 4655–69
Jan S *et al* 2004 GATE: a simulation toolkit for PET and SPECT *Phys. Med. Biol.* **49** 4543–61
Jiang Z, Polf J C, Barajas C A, Gobbert M K and Ren L 2023 A feasibility study of enhanced prompt gamma imaging for range verification in proton therapy using deep learning *Phys. Med. Biol.* **68** 075001
Kendall A and Gal Y 2017 What uncertainties do we need in bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems* 30
Kendall A, Gal Y and Cipolla R 2018 Multi-task learning using uncertainty to weigh losses for scene geometry and semantics *Proce. of the IEEE Conference on Computer Vision and Pattern Recognition* pp 7482–91
Kingma D P and Ba J 2015 Adam: a method for stochastic optimization *In Proceedings of 3rd International Conference on Learning Representations* (https://doi.org/10.48550/arXiv.1412.6980)
Knopf A-C and Lomax A 2013 In vivo proton range verification: a review *Phys. Med. Biol.* **58** R131–60
Kraan A C 2015 Range verification methods in particle therapy: underlying physics and monte carlo modeling *Front. Oncol.* **5** 150
Kuleshov V, Fenner N and Ermon S 2018 Accurate uncertainties for deep learning using calibrated regression *Int. Conf. on Machine Learning* pp 2796–804
Kurosawa S *et al* 2012 Prompt gamma detection for range verification in proton therapy *Curr. Appl Phys.* **12** 364–8
Lerendegui-Marco J, Balibrea-Correa J, Babiano-Suárez V, Ladarescu I and Domingo-Pardo C 2022 Towards machine learning aided real-time range imaging in proton therapy *Sci. Rep.* **12** 1–17
Mager M 2016 Alpide, the monolithic active pixel sensor for the alice its upgrade *Nucl. Instrum. Methods Phys. Res.* A **824** 434–8
Maradia V, van de Water S, Meer D, Weber D C, Lomax A J and Psoroulas S 2022 Ultra-fast pencil beam scanning proton therapy for locally advanced non-small-cell lung cancers: field delivery within a single breath-hold *Radiother. Oncol.* **174** 23–9
Marafini M, Mirabelli R, Pinci D, Patera V, Sciubba A, Spiriti E, Traini G and Sarti A 2017 Mondo: a neutron tracker for particle therapy secondary emission characterisation *Phys. Med. Biol.* **62** 3299–312
Moteabbed M, Espana S and Paganetti H 2011 Monte carlo patient study on the comparison of prompt gamma and pet imaging for range verification in proton therapy *Phys. Med. Biol.* **56** 1063–82
Nix D A and Weigend A S 1994 Estimating the mean and variance of the target probability distribution *Proc. of 1994 IEEE Int. Conf. on Neural Networks* vol 1 pp 55–60
Paganetti H 2012 Range uncertainties in proton therapy and the role of monte carlo simulations *Phys. Med. Biol.* **57** R99–R117
Parodi K and Enghardt W 2000 Potential application of pet in quality assurance of proton therapy *Phys. Med. Biol.* **45** N151–6
Paszke A *et al* 2019 Pytorch: an imperative style, high-performance deep learning library *Advances in Neural Information Processing Systems* 32, 8024–35
Pettersen H E S *et al* 2019 Design optimization of a pixel-based range telescope for proton computed tomography *Phys. Medica* **63** 87–97
Pettersen H E S *et al* 2021 Helium radiography with a digital tracking calorimetera monte carlo study for secondary track rejection *Phys. Med. Biol.* **66** 035004
Polf J C, Barajas C A, Peterson S W, Mackin D S, Beddar S, Ren L and Gobbert M K 2022 Applications of machine learning to improve the clinical viability of compton camera based in vivo range verification in proton radiotherapy *Front. Phys.* **10** 838273

Schilling A *et al* 2023 Proton Therapy Treatment Simulations with the Bergen DTC Prototype for Range Verification *Zenodo*
Schneider U and Pedroni E 1995 Proton radiography as a tool for quality control in proton therapy *Med. Phys.* **22** 353–63
Smeets J *et al* 2012 Prompt gamma imaging with a slit camera for real-time range control in proton therapy *Phys. Med. Biol.* **57** 3371–405
Srivastava N, Hinton G, Krizhevsky A, Sutskever I and Salakhutdinov R 2014 Dropout: a simple way to prevent neural networks from overfitting *J. Mach. Learn. Res.* **15** 1929–58
Student 1908 The probable error of a mean *Biometrika* **6** 1–25
Tian L, Landry G, Dedes G, Kamp F, Pinto M, Niepel K, Belka C and Parodi K 2018 Toward a new treatment planning approach accounting for in vivo proton range verification *Phys. Med. Biol.* **63** 215025
Unkelbach J, Chan T C Y and Bortfeld T 2007 Accounting for range uncertainties in the optimization of intensity modulated proton therapy *Phys. Med. Biol.* **52** 2755–73
van de Water S, Belosi M F, Albertini F, Winterhalter C, Weber D C and Lomax A J 2020 Shortening delivery times for intensity-modulated proton therapy by reducing the number of proton spots: an experimental verification *Phys. Med. Biol.* **65** 095008
Wieser H-P *et al* 2017 Development of the open-source dose calculation and optimization toolkit matrad *Med. Phys.* **44** 2556–68
Wilson R R 1946 Radiological use of fast protons *Radiology* **47** 487–91