Physics in Medicine & Biology

PAPER • **OPEN ACCESS**

Proton dose calculation with LSTM networks in presence of a magnetic field

To cite this article: Domagoj Radonic *et al* 2024 *Phys. Med. Biol.* **69** 215019

IOP Publishing Phys. Med. Biol. 69 (2024) 215019 https://doi.org/10.1088/1361-6560/ad7f1e

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

OPEN ACCESS

RECEIVED
26 April 2024

REVISED
18 September 2024

ACCEPTED FOR PUBLICATION
24 September 2024

PUBLISHED
21 October 2024

PAPER

Proton dose calculation with LSTM networks in presence of a magnetic field

Domagoj Radonic[1,2,9], Fan Xiao[1,9], Niklas Wahl[3,4], Luke Voss[3,4,5], Ahmad Neishabouri[4,6], Nikolaos Delopoulos[1], Sebastian Marschner[1], Stefanie Corradini[1], Claus Belka[1,7,8], George Dedes[2], Christopher Kurz[1,10] and Guillaume Landry[1,10,*]

[1] Department of Radiation Oncology, LMU University Hospital, LMU Munich, Munich, Germany
[2] Department of Medical Physics, LMU Munich, Munich, Germany
[3] Department of Medical Physics in Radiation Oncology, German Cancer Research Center (DKFZ), Heidelberg, Germany
[4] National Center for Radiation Oncology (NCRO), Heidelberg Institute for Radiation Oncology (HIRO), Heidelberg, Germany
[5] Ruprecht Karl University of Heidelberg, Institute of Computer Science, Heidelberg, Germany
[6] Clinical Cooperation Unit Radiation Oncology, German Cancer Research Center (DKFZ), Heidelberg, Germany
[7] German Cancer Consortium (DKTK), partner site Munich, a partnership between DKFZ and LMU University Hospital Munich, Munich, Germany
[8] Bavarian Cancer Research Center (BZKF), Munich, Germany
[9] Equal contribution.
[10] Shared senior authorship.
[*] Author to whom any correspondence should be addressed.

E-mail: Guillaume.Landry@med.uni-muenchen.de

Keywords: dose calculation, deep learning, MR-guided proton therapy, treatment planning, LSTM

Abstract

Objective. To present a long short-term memory (LSTM) network-based dose calculation method for magnetic resonance (MR)-guided proton therapy. *Approach.* 35 planning computed tomography (CT) images of prostate cancer patients were collected for Monte Carlo (MC) dose calculation under a perpendicular 1.5 T magnetic field. Proton pencil beams (PB) at three energies (150, 175, and 200 MeV) were simulated (7560 PBs at each energy). A 3D relative stopping power cuboid covering the extent of the PB dose was extracted and given as input to the LSTM model, yielding a 3D predicted PB dose. Three single-energy (SE) LSTM models were trained separately on the corresponding 150/175/200 MeV datasets and a multi-energy (ME) LSTM model with an energy embedding layer was trained on either the combined dataset with three energies or a continuous energy (CE) dataset with 1 MeV steps ranging from 125 to 200 MeV. For each model, training and validation involved 25 patients and 10 patients were for testing. Two single field uniform dose prostate treatment plans were optimized and recalculated with MC and the CE model. *Results.* Test results of all PBs from the three SE models showed a mean gamma passing rate (2%/2 mm, 10% dose cutoff) above 99.9% with an average center-of-mass (COM) discrepancy below 0.4 mm between predicted and simulated trajectories. The ME model showed a mean gamma passing rate exceeding 99.8% and a COM discrepancy of less than 0.5 mm at the three energies. Treatment plan recalculation by the CE model yielded gamma passing rates of 99.6% and 97.9%. The inference time of the models was 9–10 ms per PB. *Significance.* LSTM models for proton dose calculation in a magnetic field were developed and showed promising accuracy and efficiency for prostate cancer patients.

1. Introduction

Proton radiotherapy has the potential to achieve superior dose conformity compared with traditional photon therapy due to the Bragg peak dose deposition and distal dose fall-off (Paganetti *et al* 2021). In practice, however, proton therapy has limitations in achieving high-precision dose delivery due to its inherent range uncertainty caused by inaccuracies in photon-derived tissue stopping powers, increased sensitivity to inter-

and intra-fractional anatomical changes and patient setup differences (Lomax 2008a, 2008b, Unkelbach *et al* 2009, Paganetti 2012). These issues stress the need for online image guidance during proton dose delivery (Lane *et al* 2023).

Magnetic resonance imaging (MRI) has the advantages of high soft tissue contrast and using non-ionizing radiation. It additionally offers the potential for real-time anatomical and physiological imaging (Metcalfe *et al* 2013, Schmidt and Payne 2015). With the combination of in-room real-time MRI and proton therapy, the treatment plan could be adapted or re-optimized based on the latest patient geometry before dose delivery (and ultimately during dose delivery), which would improve the target dose coverage and reduce the dose to surrounding normal tissues (Moteabbed *et al* 2014, Kurz *et al* 2017, Matter *et al* 2019, Hoffmann *et al* 2020, Pham *et al* 2022). To make this real-time adaptive mode a reality, the plan adaptation and re-optimization require an accurate and fast proton dose calculation engine.

For proton dose calculation in magnetic fields, the Lorentz force causes charged primary protons to be deflected as they travel towards the patient and decelerate inside the patient's body (Raaymakers *et al* 2008, Wolf and Bortfeld 2012, Hartman *et al* 2015). Monte Carlo (MC) methods, which compute the trajectories of individual particles by simulating particle transport physics, can consider the impact of magnetic fields on dose distributions (Hartman *et al* 2015, Saini *et al* 2017, Luhr *et al* 2019, Padilla-Cabal *et al* 2020). By simulating a large number of particles, these methods can achieve state-of-the-art dose accuracy but also require high computation times. To offer a trade-off between speed and precision, several analytical methods based on correction factors have been proposed to estimate the proton beam deflection under the magnetic field and achieve accuracy close to MC methods in homogeneous tissues (Fuchs *et al* 2017, Schellhammer and Hoffmann 2017, Padilla-Cabal *et al* 2018, Teoh *et al* 2020). Although accelerated by graphics processing units (GPUs), the calculation time of recent MC methods is still in the order of seconds, while analytical algorithms can be closer to real-time (Fracchiolla *et al* 2021, Duetschler *et al* 2023, Li *et al* 2024). However, the analytical algorithms lack the accuracy of MC methods. The ultimate need for accurate real-time plan adaptation and re-optimization is still unmet, which requires dose engines to produce MC accuracy at sub-second speeds in the online iterative process.

Recently, several studies have shown the feasibility of deep learning techniques to realize sub-second proton dose calculations with MC accuracy. There are two main approaches: improvement of fast dose calculations and methods modeling beam sequences. The first methods focused on improving the accuracy of a low-cost dose estimation method in the patient coordinate system (Javaid *et al* 2021, Wu *et al* 2021, Zhang *et al* 2022). These methods are fast but depend on computationally affordable methods to provide the required physical input. This study is based on a second approach, which uses models that operate effectively on individual pencil beams (PB) (Neishabouri *et al* 2021, Zhang *et al* 2021, Pastor-Serrano and Perkó 2022a, 2022b). These models serve as the building blocks to generate the dose distribution of the entire field, enabling effective use for subsequent optimization and re-planning in real-time adaptive radiotherapy. Although the two kinds of methods achieved good accuracy and high speed, to the best of our knowledge, they have not been applied to proton dose calculation scenarios under magnetic fields.

In this work, based on a previous long short-term memory (LSTM) proton dose calculation study (Neishabouri *et al* 2021), we trained, validated and tested LSTM-based neural networks for single energy (SE) proton pencil-beam (PB) dose calculation in a 1.5 T magnetic field for prostate cancer patients. Further, energy dependence was introduced into extended LSTM models for multi-energy (ME) or continuous energy (CE) PB dose calculation.

2. Materials and methods

2.1. Proposed framework

The goal is to calculate the PB dose in the magnetic field using an LSTM model based on initial beam direction and energy in a given patient's anatomy. Therapeutic energy protons travel mainly in one direction, with additional directional deflection caused by the magnetic field, and stop after undergoing energy loss. We thus chose to frame the problem within a spatiotemporal context similar to Neishabouri *et al* (2021) to focus on how the protons interact with tissues over time and space as observed in the beam's eye view (BEV). Besides, considering the impact of the magnetic field, we adjusted the BEV context extraction method to accommodate the deflections of proton trajectories induced by the Lorentz force (more details described in 2.3.). The framework of this study is shown in figure 1, where the schematic of the ME and CE models includes the optional energy embedding (orange component) introduced on top of the SE model.

The SE model closely follows the framework presented by Neishabouri *et al* (2021). Given the fixed initial energy of all beams, no energy information is inputted into the model. Along the beam direction, a 3D relative stopping power (RSP) cuboid of interest containing the proton beam dose region is resampled from the patient's computed tomography (CT)-derived RSP. Then, in the BEV coordinate system, we treat the 3D

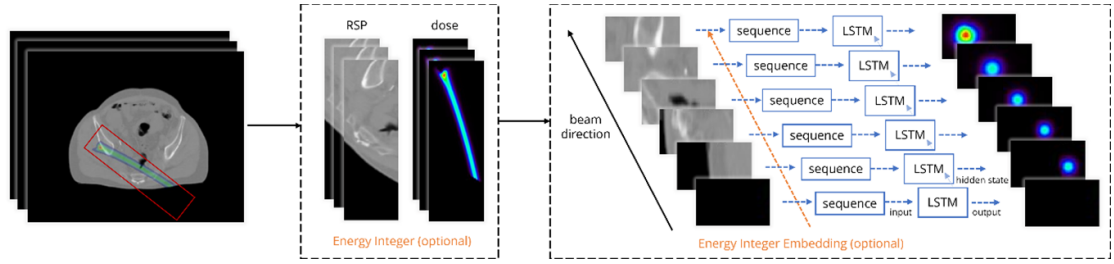


Figure 1. The framework of this study. For the SE model: (left) dose distribution from a single PB obtained from MC simulation with outline of the 3D cuboid of interest, (middle) resampling of input RSP and output dose to the cuboid of interest, (right) LSTM model converting a sequence of RSP slices to a sequence of dose slices for a single energy. For the ME and CE model: (middle and right) The 3D RSP cuboid is inputted into the model with an additional energy integer, and a single LSTM model converts the sequence of RSP slices and an energy vector to dose at variable energies.

cuboid of interest as a sequence of 2D slices traveling from upstream to downstream. By flattening these 2D slices into 1D sequences, the uni-directional LSTM unit can process the anatomy sequences and generate the internal hidden states and outputs. The internal hidden states can be used as the input information for the subsequent slices and the outputs are passed into a fully-connected-layer neural network to generate the predicted dose slices. Finally, the dose slices from the SE model are spliced back to the 3D dose cuboid size and compared with the ground truth MC dose.

For the ME and CE model, RSP cuboids extracted from beams of different initial energies are inputted into the model along with the corresponding energy integer. For each beam, the energy integer is mapped into an energy vector using a learnable weight matrix, which is then concatenated with the flattened 1D RSP sequences and passed to the LSTM unit (more details described in 2.4.). The subsequent steps are similar to those for the SE models. With the initial energy embedded, the proposed ME and CE LSTM models can output dose cuboids with different energies.

2.2. Dataset preparation

2.2.1. Patient data

The data used in this work come from CT scans of patients treated at the Department of Radiation Oncology of the LMU University Hospital. All the patients were treated in-house for prostate cancer with a 0.35 T MR-linac (MRIdian, ViewRay, USA) (Kluter 2019). The workflow of treating patients at the MR-linac involves obtaining a planning CT scan (Aquilion LB, Canon Medical Systems, NL) and an additional MRI scan before the CT scan. Then, a deformable image registration between CT and MRI is performed to adjust for anatomical deformations. In this study we used the deformed CT, which is the basis in the treatment planning system for plan optimization. The voxel size was 1.5 mm $\times$ 1.5 mm $\times$ 1.5 mm. We chose the deformed CT since it best represents images which would be used in practice. Furthermore, patients having artificial implants, mostly artificial hips, have been excluded from the study. In the end, 35 patients in total were selected who went through a prior anonymization to comply with data protection law.

2.2.2. MC simulation setting

The ground truth MC dose distributions were generated using Geant4 v11.00-patch-03 with the *QGSP_BERT_HP* physics list, which is commonly used for dose calculation and radiation protection. The voxel geometry of the CT scans was converted to elemental composition and mass density maps for Geant4 simulations based on the CT scanner-specific calibration curve from our department (Schmid *et al* 2015). RSP maps with respect to water were then created within the Geant4 framework and exported, utilizing a proton energy of 150 MeV and a mean excitation energy of water of 78 eV. We used RSP instead of Hounsfield Unit maps as the anatomy inputs of the LSTM model because they are consistent with the dose maps generated from Geant4, and to be independent of the CT scanner model used.

A 1.5 T homogeneous magnetic field was simulated in Geant4. This field was set to be aligned with the z-axis (parallel to the superior-inferior patient direction) and act within a cylinder of 30 cm radius. Outside of the cylinder the field strength was set to 0. The dose was scored on the same voxel geometry as the CT input, and all the proton PBs were simulated with 10^6 histories. By estimating the uncertainty of these statistics using a simple batch approach (Walters *et al* 2002) with a batch of $n = 10$ identical MC simulations, we ensured that the maximal relative standard error in all regions above 10% of the maximum dose is always below 1.15%. Each PB simulation in Geant4 was performed using a single CPU core on a cluster with 144 CPU cores.

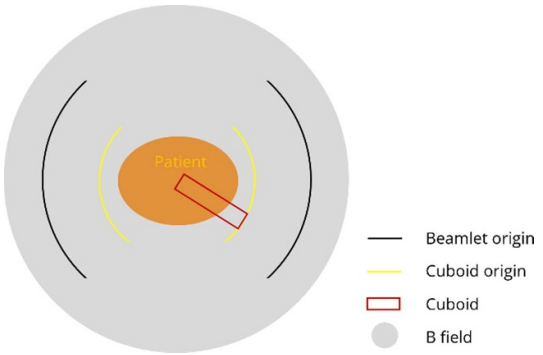


Figure 2. Schematic diagram for the proposed cuboid extraction method from a cross-sectional perspective.

Table 1. RSP-dose cuboid pair parameters for the three SE (150, 175 and 200 MeV), the ME (150/175/200 MeV), and the CE (125–200 MeV) datasets.

Energy (MeV)	Cuboid size ($I \times J \times K$)	Number of PBs for training & validation	Flip augmentation (True/False)
150	170 × 43 × 23	5400	True
175	188 × 49 × 25	5400	True
200	205 × 55 × 25	5400	True
150/175/200	205 × 55 × 25	5400 × 3	True
125–200	205 × 55 × 25	5400 × 4	False

2.2.3. Beam geometries

Given that the objective of this study is to predict the dose at the level of individual PBs, the dose distributions were simulated for single spots. PBs with energies 150 MeV, 175 MeV and 200 MeV were simulated to train three SE models and subsequently combined to train a ME model. Additional PBs within a CE range of 125–200 MeV with 1 MeV steps were simulated for the CE model. PBs had an energy spread of 0.83 MeV and a Gaussian distribution with a standard deviation of 4.2 mm. To optimize the use of patient data, PBs were isotropically sampled over two 100° arcs on either side of the patient. The sampling was confined within a predefined angular sector to avoid the anterior and posterior directions, ensuring no spots were placed directly in front or behind the patient. The propagation of the PBs was chosen to take place in the transverse plane and orthogonal to the magnetic (B) field. Defining the right lateral direction as 90°, the sampling was done within $\alpha \in [40°, 140°] \cup [220°, 320°]$ regions in $\alpha = 16.\overline{6}°$ steps, yielding 12 different angles per patient. At each of the 12 angles, a 3 × 6 PB grid with a spacing of 2.5 cm × 2.5 cm was generated in the plane orthogonal to the beam's direction, with the 6 PB grid side parallel to the superior-inferior direction. For the dataset with three energies, 216 PBs were generated per patient, yielding 7560 PBs in total for each of the 150 MeV, 175 MeV, 200 MeV energies. For the CE dataset, 216 × 4 PBs (4 energies per PB origin) with energy randomly selected from 125–200 MeV (76 energy integers) were simulated per patient, leading to around 284 PBs per energy on average.

2.3. Data preprocessing

For each patient, CT and dose values outside of the patient's skin were set to zero. Then, by the transformation relationship between the patient and BEV coordinate systems, the RSP-dose 3D cuboid pairs could be cropped and resampled based on the origin position and orientation of PBs. With the introduction of the B field, the PBs traveling longer in the air will experience greater deflections before they reach the patient. To allow the model to learn to relate the amount of air before the patient to the magnitude of deflections, we set the cuboid origin to a predefined distance from the PB origin rather than directly starting from the patient's surface, as shown in figure 2.

The size of the cropped volume $I \times J \times K$ was chosen depending on the energy with different cuboid sizes chosen for different energies, as summarized in table 1. The three dimensions of cuboids are the original beam direction x, the beam deflection direction y and the remaining direction z (parallel to the superior-inferior patient direction). The position of a single voxel within the cropped volume is denoted by $(i,j,k) \in Z^3$, where $0 \leqslant i < I, 0 \leqslant j < J, 0 \leqslant k < K$. The resolution of all cuboids was 2 mm × 2 mm × 2 mm, which is a common resolution for dosimetric evaluation.

After obtaining all the RSP-dose cuboid pairs, the cuboid datasets were split patient-wise to ensure separation between training, validation, and testing datasets. For each SE dataset, 4320 PBs from 20 patients were used to train the model, 1080 PBs from 5 patients were used as a validation set and the remaining 10 patients with their 2160 PBs were kept as a test set. For the ME dataset, the combination of all three SE datasets was used. The training datasets of SE and ME were augmented by randomly mirroring the cuboids along the *xy*-plane at each epoch. This transformation preserves the original direction of the Lorentz force in transformed cuboids and avoids interpolation to a new cuboid grid. Therefore, the original training dataset was essentially doubled without performing additional MC simulations. For the CE dataset, 4320 $\times$ 4 PBs with random energies were used for training and 1080 $\times$ 4 PBs were used for validation. Additionally, two test sets for the CE model were considered: the 2160 $\times$ 3 PBs with three fixed energies described above, and an additional 2160 PBs with random energies (125–200 MeV) from the same 10 test patients. A global normalization was performed for all the training, validation, and testing cuboid datasets using the maximum RSP and dose cuboid values of the training dataset.

Finally, to evaluate the CE model's ability to calculate a full plan, two patients were selected from the 10 test patients. One case was random, and one case was chosen because of an air cavity adjacent to the planning target volume (PTV). The contours of the PTV and organ-at-risk structures were delineated for these patients as part of the clinical workflow and were available for the study. Two single-field uniform dose (SFUD) plans were generated at a gantry angle of 90° and optimized using matRad (Wieser *et al* 2017, Ackermann *et al* 2020), which utilized PB doses that were pre-calculated using MC, with each PB dose simulated using 10^5 histories. The dose prescription was 74 Gy in 37 fractions, and MC plans were normalized to a $D_{95\%}$ of the PTV above 95% of the prescribed dose (in our case 70.4 Gy). Plan 1 required 7429 spots and an energy range of 155–200 MeV (46 discrete energies), while Plan 2 required 10332 spots and an energy range of 160–200 MeV (41 discrete energies). These PB lists were determined by trial and error by estimating the deflection and shifting the PBs accordingly. The CE model was used to recalculate the PB doses, which were subsequently used to reconstruct the SFUD plan dose using the same optimized weights and the same scaling factor applied for the normalization of the MC plan. The summed predicted and MC doses from the two plans were then evaluated.

2.4. Model training

The network architecture used for three SE models corresponds to the LSTM model used in (Neishabouri *et al* 2021) for proton dose calculation in the absence of a *B* field. The LSTM features one layer with 1000 neurons as an internal layer, followed by two fully connected layers with 100 neurons and ReLu activation layers. Each SE model was trained separately on the corresponding 150, 175 and 200 MeV energy training dataset.

For the ME model, a learnable embedding layer (Devlin *et al* 2018), which is widely used as a lookup table for mapping an index into a weight vector in the field of natural language processing, is applied to convert the energy integer from an energy value, e.g. 150 MeV, into an energy float vector. More specifically, the energy integer E takes a value from the initial energy set D_e containing all embedded energy integers and passes it through a learnable embedding layer.

In the embedding layer, E is first mapped to a one-hot encoded vector V_{in}:

$$V_{\text{in}} \in \{(1,0,\ldots,0),(0,1,\ldots,0),\ldots,(0,,\ldots,1,0),(0,0,\ldots,1)\} \in \mathbb{R}^{|D_e|}. \tag{1}$$

Then, a fully-connected layer with learnable weights W and bias b takes V_{in} as input and outputs a float vector V_{out} of dimension I, where I is the dimension of cuboid in the original beam direction x mentioned in 2.3:

$$V_{\text{out}} = V_{\text{in}} \cdot W + b, \quad W \in \mathbb{R}^{|D_e| \times I}, \quad b \in \mathbb{R}^I \ . \tag{2}$$

Afterwards, the sequence of 2D RSP slices (each slice is denoted as S_i) is flattened into the sequence of 1D RSP vectors F_{RSP}:

$$\text{RSP} = (S_i)_{i \leq I}, \quad S_i \in \mathbb{R}^{J \times K} \tag{3}$$

$$F_{\text{RSP}} = (\text{vector}(S_i))_{i \leq I} \in \mathbb{R}^{I \times l} : l = J \cdot K. \tag{4}$$

The energy embedding vector V_{out} is then concatenated with F_{RSP} to get the LSTM layer input F_{LSTM}:

$$F_{\text{LSTM}} = V_{\text{out}} \oplus F_{\text{RSP}} \in \mathbb{R}^{I\times(l+1)} \quad . \tag{5}$$

Therefore, each element of V_{out} is added to F_{RSP} at each depth of beam penetration, as shown in figure 1. F_{LSTM} passes through the LSTM layer and output layers, which are the same as for the SE model. The ME model was trained on the combined three energy datasets such that $D_e = \{150, 175, 200\}$. In addition, the same training process used for the ME model was applied to train the CE model using the CE dataset, but with an extended V_{in} where $|D_e| = 76$ (125–200 MeV).

For the model training, we used the Adam optimizer to minimize the mean square error (MSE) loss between the output dose and ground truth dose sequence. The initial learning rate was set to 10^{-5} and batch size to 8. Models were implemented based on Python 3.8 and Pytorch 2.0.1 with an NVIDIA® RTX A6000 GPU (48 GB memory). For validation loss convergence, the three SE models were trained for 15000 epochs, the ME model for 5000 epochs and the CE model for 3000 epochs, which took about 5 days. The models with the best overall MSE validation loss were obtained and used for the prediction in the test dataset. After training, the ME model and SE models were tested on the identical 150/175/200 MeV test datasets described in 2.3 and the results were compared. The CE model was tested on the same 150/175/200 MeV test dataset, plus the continuous 125–200 MeV test dataset and the two SFUD plans.

2.5. Post-processing

After obtaining all model predictions, the following changes have been applied before the final evaluation. Firstly, a mask was applied to assure that only dose inside the patient is evaluated and all dose voxels below a threshold of 0.01% of the global maximum were set to 0, similar to Pastor-Serrano and Perko (2022a). This is justified by the architectural limitations of neural networks where outputs of the last linear and activation layers are hardly ever truly 0. Secondly, 402 PBs that overshot the patient were filtered out in the 200 MeV test dataset; no PBs with the energy of 150 MeV or 175 MeV needed to be removed; and 44 PBs that overshot the patient in the 125–200 MeV test dataset were filtered out. Finally, the normalization defined in 2.3 was reversed once the model predictions were obtained.

2.6. Evaluation metrics

To evaluate the agreement of the predicted and MC doses, several radiotherapy-related metrics similar to Lysakovski *et al* (2021) were used. A 3D global gamma evaluation (Γ) was performed, and the gamma pass rate (γ_{PR}) was calculated to compare the 3D dose distributions using PyMedPhys (Biggs *et al* 2022). The distance-to-agreement threshold was set to 2 mm, the dose difference threshold to 2% and the dose cutoff to 10% of the maximum dose.

Laterally (over y and z) integrated depth-dose profiles $d_{yz}(x)$ were obtained to evaluate their agreement. The depth was defined as the initial direction of the PB (x), i.e. the effect of beam deflection was neglected, and we thus considered an effective projected range. The range difference between d_{yz}^{pred} and d_{yz}^{mc} was quantified using the depth of the distal dose falloff to 80% of the Bragg peak value, R_{D80}. The deviation between predicted R_{D80}^{pred} and MC-simulated R_{D80}^{mc} was evaluated through the absolute difference and relative difference:

$$\Delta R_{D80}\,[\text{mm}] = \left| R_{D80}^{\text{pred}} - R_{D80}^{\text{mc}} \right| \tag{6}$$

$$\Delta R_{D80}\,[\%] = \frac{\left| R_{D80}^{\text{pred}} - R_{D80}^{\text{mc}} \right|}{R_{D80}^{\text{mc}}} \times 100\%. \tag{7}$$

The relative error (ε_{rel}) was used to quantify the relative disagreement between d_{yz}^{pred} and d_{yz}^{mc}. To disentangle ε_{rel} from ΔR_{D80}, d_{yz}^{pred} was shifted by ΔR_{D80} prior to evaluation to highlight potential overall amplitude errors. At a given depth, ε_{rel} was defined as:

$$\varepsilon_{\text{rel}} = 200\frac{d_{yz}^{\text{pred}} - d_{yz}^{\text{mc}}}{\left(d_{yz}^{\text{pred}} + d_{yz}^{\text{mc}}\right)}\,[\%] \tag{8}$$

whereby a dose threshold for calculating ε_{rel} was set to 20% of the d_{yz}^{mc} at the Bragg Peak.

To evaluate the deflection of a beam in the magnetic field, the center of mass (COM) of the beam was calculated at each depth i:

$$\mathrm{COM}_i = \tfrac{1}{M_i}\sum_j^J\sum_k^K (i,j,k)^{\mathrm{T}}\, v_{ijk} \mid M_i \overset{\mathrm{def}}{=} \sum_j^J\sum_k^K v_{ijk} \tag{9}$$

where v_{ijk} is the corresponding voxel's dose. Slices i in which $M_i < 20\% \times \max\{M_i \forall\ 0 \leq i \leq I\}$ were filtered out.

The agreement of lateral profiles in the direction of beam deflection (y) was evaluated after integrating lateral dose in the direction of the B field (z) to obtain $d_z(x,y)$. Then the difference in the full width at half maximum (FWHM) was computed by:

$$\mathrm{FW}_{50i} = \mathrm{FWHM}_i^{\mathrm{pred}} - \mathrm{FWHM}_i^{\mathrm{mc}} \tag{10}$$

at each depth i (slices i in which $M_i < 20\% \times \max\{M_i \forall\ 0 \leq i \leq I\}$ were filtered out).

The differences in COM, FW_{50} and $\varepsilon_{\mathrm{rel}}$ were computed at each depth in the patient in steps of 2 mm. For each PB, the farthest difference in COM (F_{COM}) and FW_{50} up to the Bragg peak was reported along with the mean absolute $\varepsilon_{\mathrm{rel}}$ value ($\varepsilon^{\star}_{\mathrm{rel}}$) above the 20% threshold. Additionally, for selected PB examples, relative dose difference was also reported in the xy-plane by computing $\tilde{\varepsilon}_{\mathrm{rel}}$ from d_z^{pred} and d_z^{mc} analogous to equation (8) and the agreement of full $d_z(y)$ profiles was shown at depths of 20%, 50% and 80% of R_{D80} along the Bragg curve. To identify systematic dose offsets, a total integrated dose difference $\varepsilon_{\mathrm{int}}$ of each PB was computed after applying the threshold of 0.01% from the aforementioned post-processing steps, and deviation was assessed analogous to equation (8).

To assess the model accuracy, in each test dataset, the percentile rankings of all predicted PBs for each metric were recorded. The percentile rank was defined as the percentage of beamlets that achieved equally good or higher metric score. Thus, a low percentile indicates a better performance (high ranking). When several PBs had the same value, the lowest percentile rank was shared, e.g. all γ_{PR} of 100% shared the best rank. Besides, all PB examples shown in the next section were zoomed in from the original cuboid size to focus on the dose region. A Wilcoxon signed-rank test was performed on the γ_{PR} for test datasets at 150/175/200 MeV between the SE and ME models, and between ME and CE models.

To compare the SFUD plan dose distributions from the MC simulation and the CE model, the γ_{PR}, dose-volume histograms (DVH) and DVH indices were evaluated.

3. Results

3.1. SE models on 150/175/200 MeV test datasets

High-, mid- and lowest-ranking 175 MeV SE model PBs are shown in figures 3–5, respectively. The ranking is based on γ_{PR}. Mid-ranking cases from the SE models at 150 MeV and 200 MeV are additionally shown in figures A1 and A2 (appendix A). Qualitatively, the three SE models accurately predict PB deflection and proton range. The worst predicted PB for the γ_{PR} is shown in figure 5, with 98% of PBs being as good or better in terms of ΔR_{D80} and 94% in terms of F_{COM}. In the region of dose fall-off after the Bragg peak, there is a large portion of failed voxels. It can be assumed that the air cavity caused an underestimation in the PB's range by the model.

For mid-ranking cases at 150 MeV/175 MeV/200 MeV, laterally integrated profiles of the simulated and predicted PBs with high agreement are shown in figures A1(c), 4(c) and A2(c). The ΔR_{D80} and $\varepsilon^{\star}_{\mathrm{rel}}$ were all within 0.5 mm and 1%, respectively. The integrated profiles show that $\varepsilon_{\mathrm{rel}}$ is higher at the patient's skin and around the Bragg peak, while in the plateau it is comparatively flat for all three example cases. The mid-ranking cases had γ_{PR} above 99.9%. F_{COM} and FW_{50} differences, shown in figures A1(e), 4(e) and A2 (e), were all below 1 mm. The 2D distributions of integrated dose differences $\tilde{\varepsilon}_{\mathrm{rel}}$ over the z axis are displayed in figures A1(f), 4(f) and A 2(f), where it can be noticed that the low dose areas tend to have higher relative integrated dose deviation. Lateral PB dose profiles at different depths are shown in figures A1(g), 4(g) and A2(g). Finally, the total integrated dose differences $\varepsilon_{\mathrm{int}}$ (integrated over x, y and z) were all lower than 0.5%.

Figure 6 displays boxplots for the metrics of predicted PBs from the three SE models across 10 test patients in the 150/175/200 MeV test datasets and table 2 (SE) shows the worst and average values of each evaluation metric. The worst γ_{PR} was above 94%, and the maximum ΔR_{D80}, $\varepsilon^{\star}_{\mathrm{rel}}$, F_{COM} and FW_{50} differences were 3.4%, 13.2%, 4.7 mm, 4.2 mm, respectively. Overall, the SE models achieved good accuracy for all metrics.

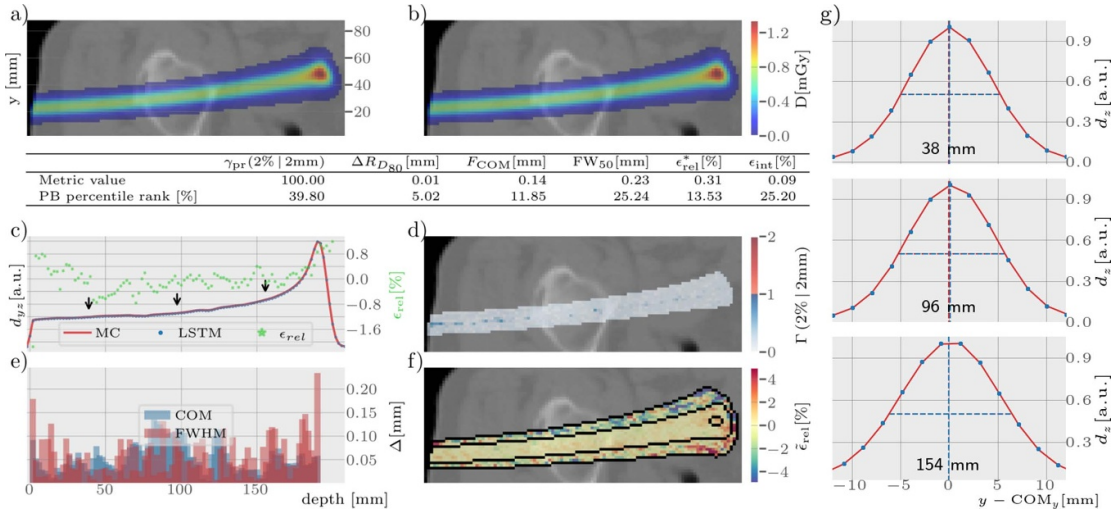


Figure 3. A high-ranking example of a predicted PB from the 175 MeV SE model (test case). An *xy*-slice of (a) MC and (b) LSTM predicted dose distribution is plotted over the RSP map. (c) Depth-dose profiles $d_{yz}(x)$ (integrated over y and z) and ε_{rel} of each slice. The black arrows indicate the depths at which lateral dose profiles $d_z(y)$ (integrated over z) are shown in (g). (d) $\Gamma(2\%|2\text{ mm})$ map. (e) Differences in COM and FW_{50} at each depth. (f) Relative dose difference of $d_z(x,y)$ capped at $\pm 5\%$. Black lines denote 80%, 10% and 1% isodose lines with respect to $d_z(x_{\text{BP}}, y_{\text{BP}})$ in the Bragg peak. (g) Lateral dose profiles $d_z(y)$ at depths of 20%, 50% and 80% of R_{D80} along the Bragg curve; color coded as in (c).

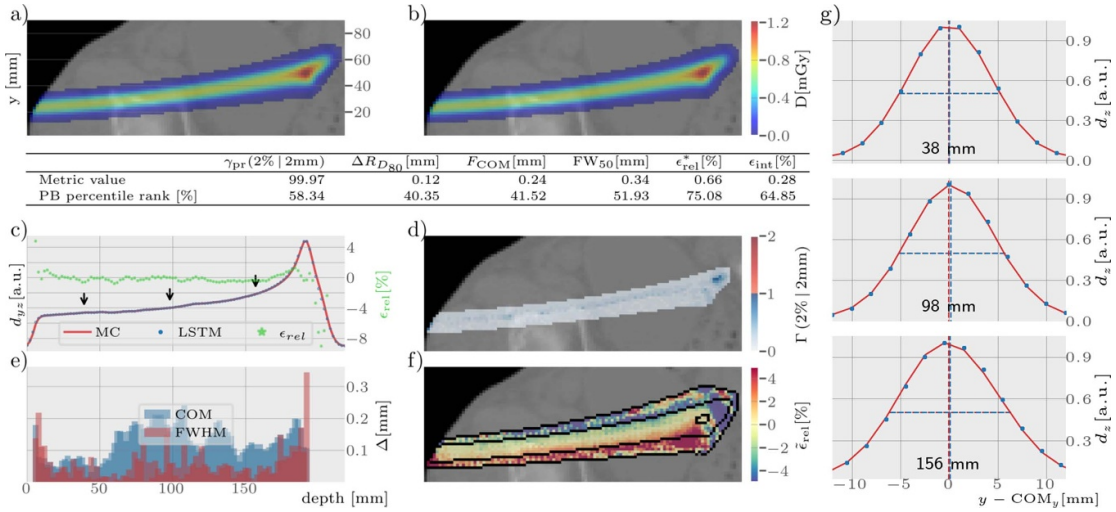


Figure 4. A mid-ranking example of a predicted PB from the 175 MeV SE model (test case). An *xy*-slice of (a) MC and (b) LSTM predicted dose distribution is plotted over the RSP map. (c) Depth-dose profiles $d_{yz}(x)$ (integrated over y and z) and ε_{rel} of each slice. The black arrows indicate the depths at which lateral dose profiles $d_z(y)$ (integrated over z) are shown in (g). (d) $\Gamma(2\%|2\text{ mm})$ map. (e) Differences in COM and FW_{50} at each depth. (f) Relative dose difference of $d_z(x,y)$ capped at $\pm 5\%$. Black lines denote 80%, 10% and 1% isodose lines with respect to $d_z(x_{\text{BP}}, y_{\text{BP}})$ in the Bragg peak. (g) Lateral dose profiles $d_z(y)$ at depths of 20%, 50% and 80% of R_{D80} along the Bragg curve; color coded as in (c).

3.2. The ME model on 150/175/200 MeV test datasets

For the ME model, the worst and average values of each evaluation metric at 150/175/200 MeV are reported in table 2 (ME). Differences between the SE and ME models were significant in the Wilcoxon signed-rank test. However, the differences in average metrics between the two models were all within 0.1% or 0.1 mm. Figure 7 shows the boxplots of γ_{PR} for SE models and the ME model. For patient case P03, the ME model exhibited a slightly wider interquartile range compared to the corresponding SE model, but both models demonstrated similar 5th to 95th percentile ranges above 98%. In the remaining nine test cases, the SE models and ME model demonstrate comparable 5th to 95th percentile ranges. The outlier of the ME model in P10 for 200 MeV showed the worst γ_{PR} of 87.8%, which is lower than the worst case in the three SE models (94.2%).

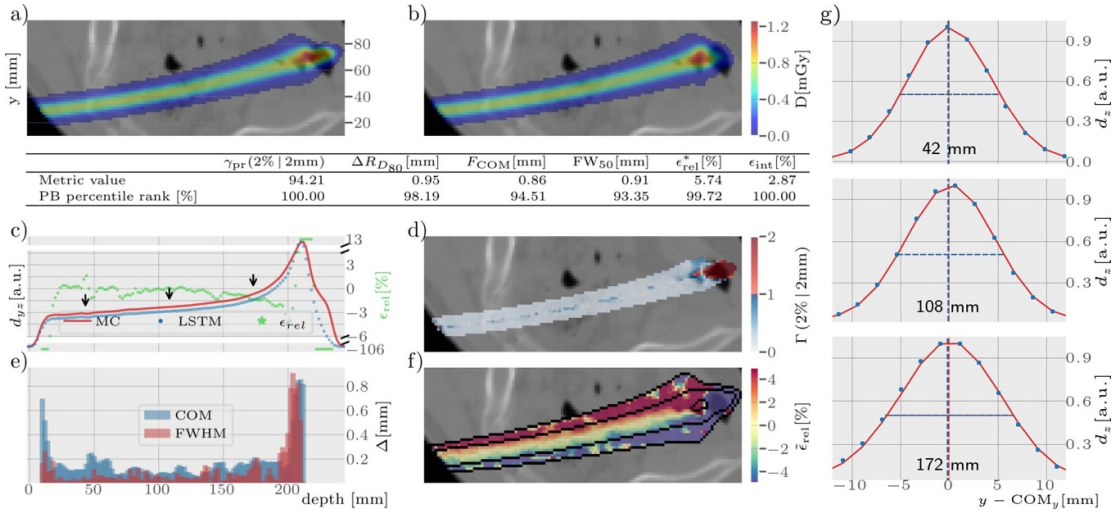


Figure 5. The lowest-ranking example of a predicted PB from the 175 MeV SE model (test case). An *xy*-slice of (a) MC and (b) LSTM predicted dose distribution is plotted over the RSP map. (c) Depth-dose profiles $d_{yz}(x)$ (integrated over y and z) and ε_{rel} of each slice. For better visibility the ordinate is broken at $\pm 7.5\%$ and only the value of the highest outlier is added to the tick labels. The black arrows indicate the depths at which lateral dose profiles $d_z(y)$ (integrated over z) are shown in (g). (d) $\Gamma(2\%|2\text{ mm})$ map. (e) Differences in COM and FW_{50} at each depth. (f) Relative dose difference of $d_z(x,y)$ capped at $\pm 5\%$. Black lines denote 80%, 10% and 1% isodose lines with respect to $d_z(x_{\text{BP}}, y_{\text{BP}})$ in the Bragg peak. (g) Lateral dose profiles $d_z(y)$ at depths of 20%, 50% and 80% of R_{D80} along the Bragg curve; color coded as in (c).

3.3. The CE model on 150/175/200/125-200 MeV PB and two SFUD plan test datasets

Table 2 (CE) presents the results of each evaluation metric in the 150/175/200 MeV and 125–200 MeV PB test datasets for the CE model. Differences between the ME and CE models were significant in the Wilcoxon signed-rank test. Figure 7 shows the boxplot of γ_{PR} for the same 150/175/200 MeV PB test datasets used with the SE and ME models. Similar performance was obtained when compared to the ME model only trained on 3 energies. The worst γ_{PR} example from the CE model, as shown in figure 8, indicates that the γ_{PR} of all PBs is higher than 90.4%, while the ΔR_{D80} for over 97% of PBs is lower than 0.9 mm. Similar to the worst predicted PB from the SE models, the beam passed through an air cavity, causing an underestimation in the PB's range by the model. Its laterally integrated profiles and lateral profiles can be seen in figures 8(c) and (g).

Furthermore, the γ_{PR} results for the two full SFUD plan dose distributions from the CE model prediction and MC simulations were 99.61% and 97.9%, as shown in figures 9(c) and 10(c), with the lower γ_{PR} for the case with an air cavity in the rectum. The DVH curves in figures 9(d) and 10(d) match well, and table 3 indicates that the largest differences of all DVH indices were less than 1.5 Gy or 1.5%.

3.4. Runtimes

The total runtime of our method consists of the geometrical calculation time to extract the RSP cuboid and the model inference time. The runtime measurement was conducted on a server with an Intel® Xeon® Gold 6354 3.00 GHz CPU and an NVIDIA® RTX A6000 GPU. The average computation time for the cuboid extraction was about 55 ms. Table 4 lists the average inference time for different models for a single PB. The inference time of the four LSTM models is between 9–10 ms and the addition of an embedding layer has a negligible impact on the inference speed. Overall, our method takes approximately 65 ms to calculate the PB dose. On the other hand, Geant4 simulations that were used to obtain the ground truths took between 65 and 114 min on a single CPU core to simulate the dose distribution of a PB containing 1×10^6 protons, depending on the initial proton energy.

4. Discussion

This study extended a previously developed architecture (Neishabouri *et al* 2021) to proton dose calculation in a magnetic field. We used more extensive evaluation criteria, and to our knowledge this work is the first to utilize an AI architecture to learn proton dose calculation in magnetic fields. A recent study proposing a transformer-based dose calculation method for proton PB dose calculation (without B fields) reported poorer performance for the LSTM architecture (Pastor-Serrano and Perkó 2022b). We did not observe poor performance in our study, and the LSTM has the advantage of having a limited set of parameters. The GPU memory requirements of our model never increased above 1 GB and our network is easy to train and converged with a constant learning rate of 10^{-5}. An extended LSTM model was also proposed for the

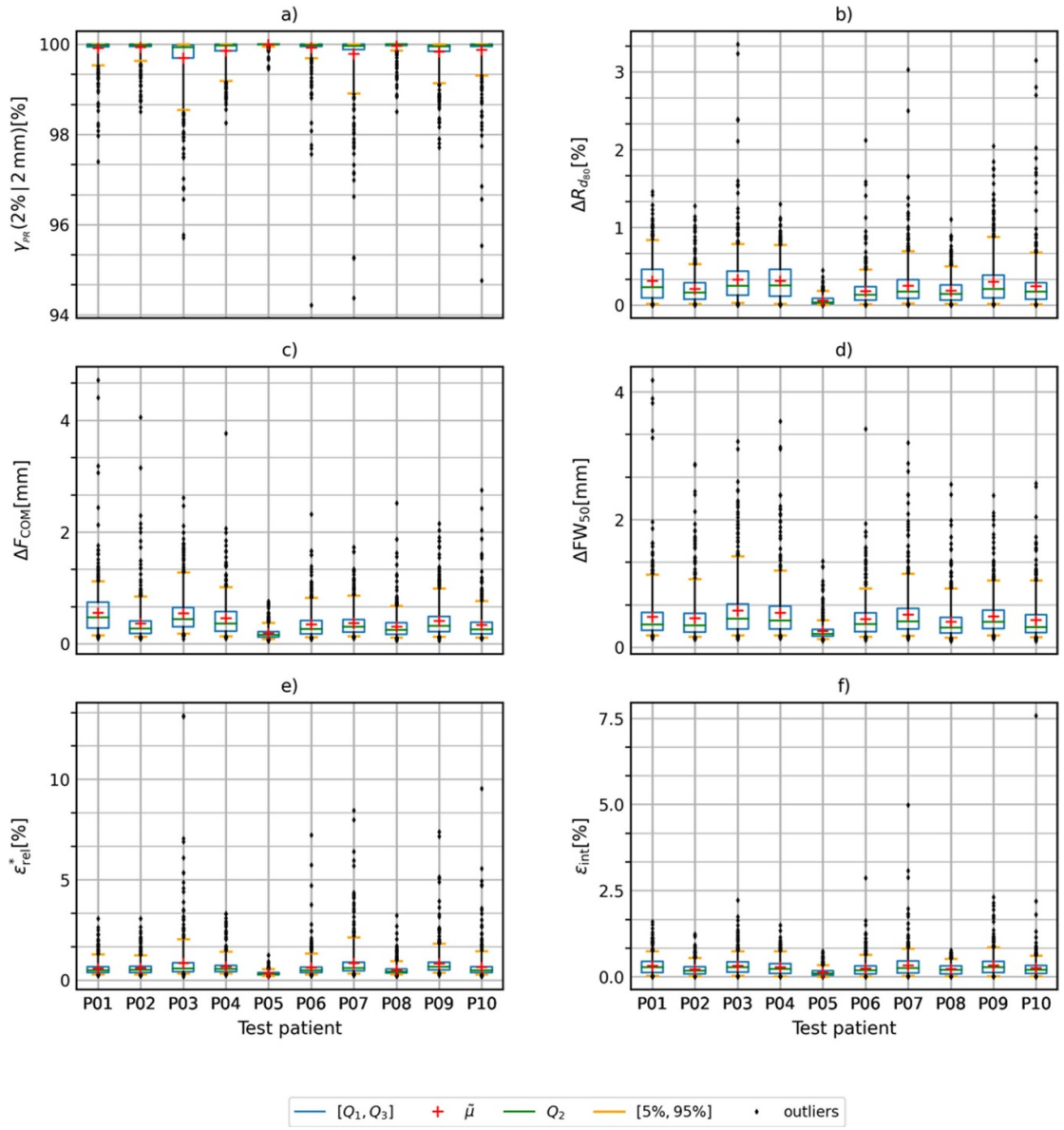


Figure 6. Boxplots for all metrics of predicted PBs from the three SE models combined. The red cross denotes the mean and the green line the median value. The blue box denotes the range between 25th and 75th percentiles. The orange lines denote the 5th and 95th percentiles, while the black points denote outliers outside of that range.

Table 2. Worst and mean values of metrics for the test sets. The top section of the table is for the 3 SE models. The middle section is for the ME model trained on 3 energies. The bottom section is for the CE model, which was tested on the 3 energies (first 3 rows of the bottom section), and on the energy range test datasets (last row of the bottom section). Values in bold indicate the worst PB per metric.

	Test energy	γ_{PR}(%)		ΔR_{D80} (%)		F_{COM} (mm)		FW_{50} (mm)		ε^*_{rel} (%)		ε_{int} (%)
Model	E(MeV)	min	mean	max	mean	max	mean	max	mean	max	mean	mean
SE	150	95.3	99.9	3.2	**0.3**	4.4	**0.4**	3.8	0.4	13.1	**0.7**	**0.3**
	175	**94.2**	**99.9**	**3.4**	0.2	**4.7**	0.4	3.9	0.4	**13.2**	0.6	0.3
	200	94.4	99.9	3.1	0.2	4.1	0.4	**4.2**	**0.5**	9.5	0.6	0.2
ME	150	94.8	99.8	**4.6**	**0.3**	2.9	**0.5**	**5.7**	**0.5**	**14.1**	**0.8**	**0.3**
	175	94.7	99.8	3.7	0.2	3.0	0.4	3.3	0.5	8.6	0.7	0.2
	200	**87.8**	**99.8**	3.9	0.2	**5.4**	0.4	4.5	0.5	11.2	0.7	0.3
CE	150	93.3	99.8	4.0	**0.4**	2.7	0.5	4.2	0.5	**16.6**	**1.2**	0.4
	175	91.8	99.6	3.3	0.3	2.6	0.6	2.8	0.6	11.5	1.1	0.4
	200	**90.5**	**99.5**	3.5	0.3	**5.1**	**0.7**	**4.8**	**0.7**	12.9	1.1	**0.5**
	125–200	92.0	99.7	**4.5**	0.3	3.5	0.6	4.5	0.6	14.5	1.2	0.4

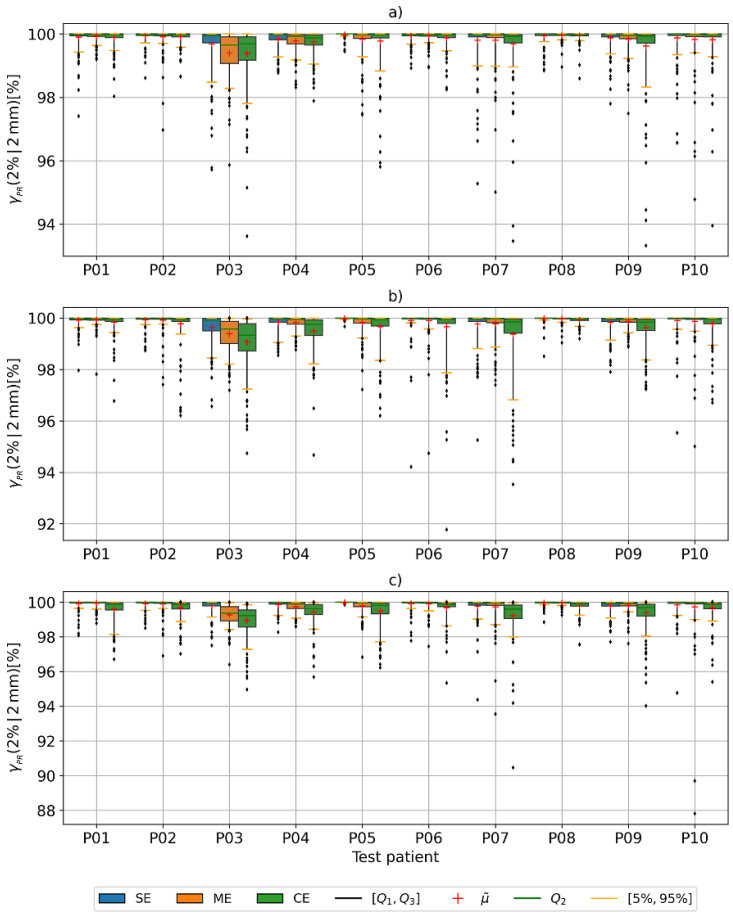
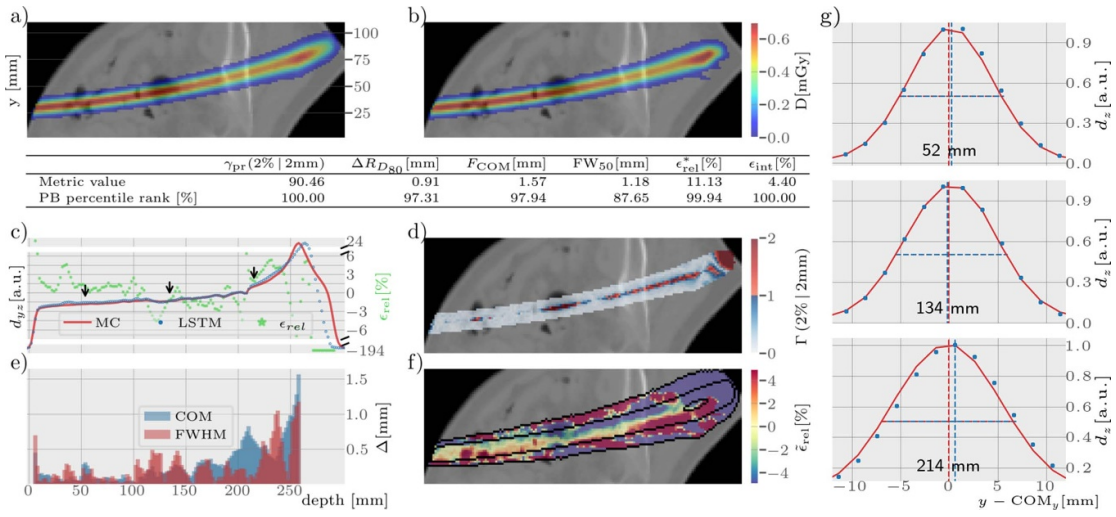


Figure 7. Boxplots of γ_{PR} for the SE, ME and CE models on (a) 150 MeV, (b) 175 MeV and (c) 200 MeV test datasets.



	$\gamma_{\mathrm{pr}}(2\%\,\vert\,2\mathrm{mm})$	$\Delta R_{D_{80}}$[mm]	F_{COM}[mm]	FW_{50}[mm]	$\epsilon^{*}_{\mathrm{rel}}$[%]	ϵ_{int}[%]
Metric value	90.46	0.91	1.57	1.18	11.13	4.40
PB percentile rank [%]	100.00	97.31	97.94	87.65	99.94	100.00

Figure 8. The worst CE model example (200 MeV PB). An *xy*-slice of (a) MC and (b) LSTM predicted dose distribution is plotted over the RSP map. (c) Depth-dose profiles $d_{yz}(x)$ (integrated over y and z) and $\varepsilon_{\mathrm{rel}}$ of each slice. For better visibility the ordinate is broken at $\pm 7.5\%$ and only the value of the highest outlier is added to the tick labels. The black arrows indicate the depths at which lateral dose profiles $d_z(y)$ (integrated over z) are shown in (g). (d) $\Gamma(2\%|2\,\mathrm{mm})$ map. (e) Differences in COM and FW_{50} at each depth. (f) Relative dose difference of $d_z(x, y)$ capped at $\pm 5\%$. Black lines denote 80%, 10% and 1% isodose lines with respect to $d_z(x_{\mathrm{BP}}, y_{\mathrm{BP}})$ in the Bragg peak. (g) Lateral dose profiles $d_z(y)$ at depths of 20%, 50% and 80% of R_{D80} along the Bragg curve; color coded as in (c).

multi-energy dose calculation task. As the training PB energies increase, the ME and CE LSTM models can achieve performance comparable to the SE LSTM models by integrating energy integer information into the model through an embedding layer. Besides, judging from the worst cases which included air cavities and the lower γ_{PR} for SFUD Plan 2, air cavities were problematic for our models. We searched the training set for air

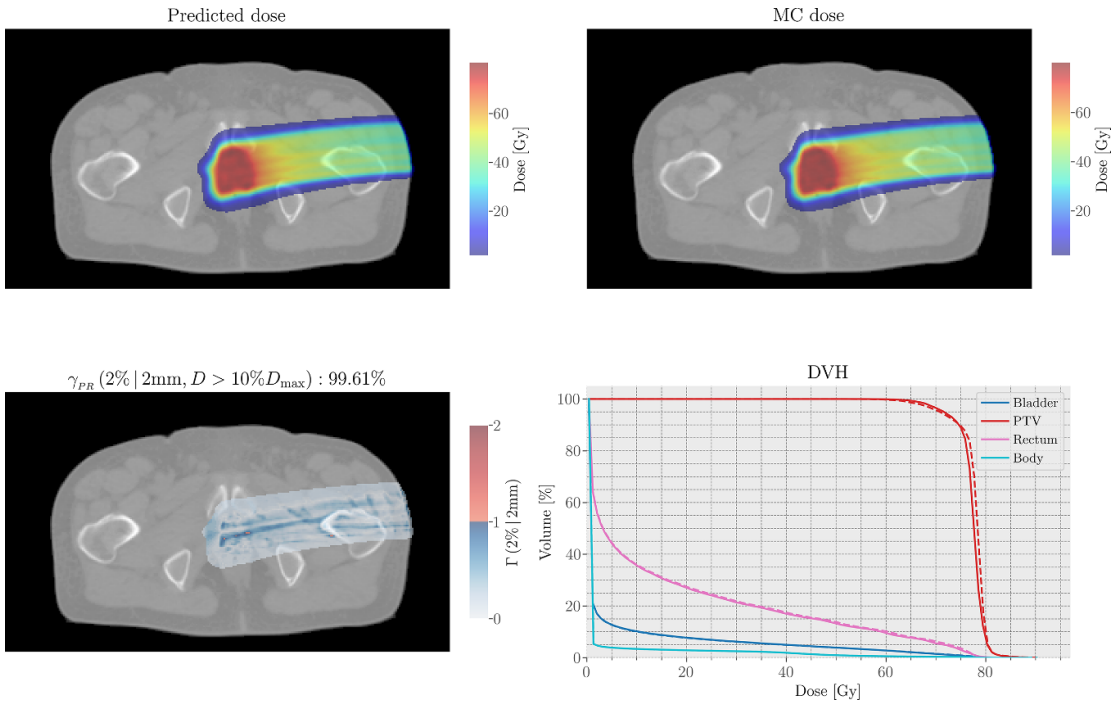


Figure 9. Transverse slice of the (*top left*) predicted dose, (*top right*) MC dose, (*bottom left*) Γ(2%|2 mm) map and (*bottom right*) DVH (predicted dose solid line and MC dose as dashed line) for Plan 1 (P08).

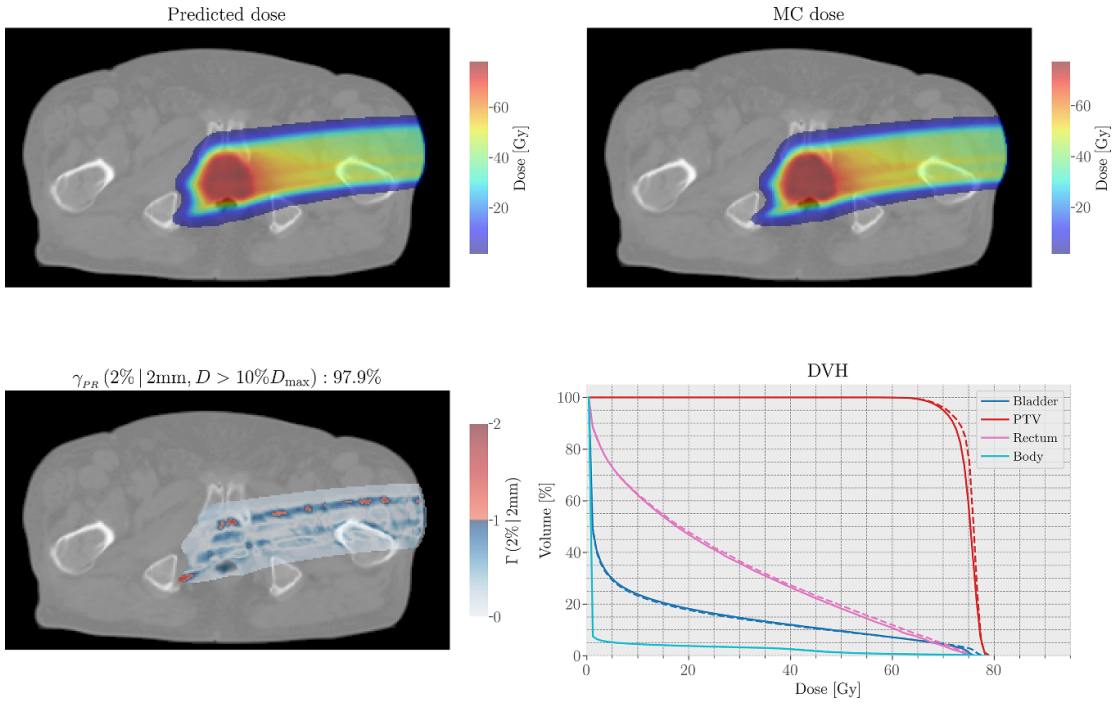


Figure 10. Transverse slice of the (*top left*) predicted dose, (*top right*) MC dose, (*bottom left*) Γ(2%|2 mm) map and (*bottom right*) DVH (predicted dose solid line and MC dose as dashed line) for Plan 2 (P09).

cavities in areas with doses higher than 10% based on an RSP threshold of 0.01 and found that 79 out of 4320 training beams (1.8%) had such cases in the 150 MeV training dataset. This under-representation suggests that more air cavity cases may be required for training, and we plan to address this issue in the future by potentially introducing artificial air cavities in the simulation of the ground truth training set.

We compared our results to literature on analytical and GPU-MC dose calculation methods. The first analytical method (Padilla-Cabal *et al* 2018) combined trajectory correction through a numerical solution of the relativistic Lorentz equation with a look-up table of dose deposition for different PB energies at each

Table 3. DVH indices and differences (CE-MC) of the full SFUD plan dose distributions from the CE model and MC simulation for two plans. Prescription dose 74 Gy.

		PTV		Bladder			Rectum			
		$D_{2\%}$ (Gy)	$D_{95\%}$ (Gy)	$D_{2\%}$ (Gy)	V_{60Gy} (%)	V_{65Gy} (%)	$D_{2\%}$ (Gy)	V_{50Gy} (%)	V_{60Gy} (%)	V_{65Gy} (%)
Plan 1	CE	80.9	71.2	65.5	2.7	2.1	76.0	13.0	9.3	7.6
	MC	80.8	70.4	66.0	2.8	2.1	76.5	13.0	10.1	7.9
	Difference	0.1	0.8	−0.5	−0.1	0.0	−0.5	0.0	−0.8	−0.3
Plan 2	CE	77.5	69.6	74.2	7.1	5.8	72.2	18.0	10.6	7.1
	MC	77.5	70.4	75.7	7.1	5.9	73.1	18.0	11.7	8.0
	Difference	0.0	−0.8	−1.5	0.0	−0.1	−0.9	0.0	−1.1	−0.9

Table 4. Average runtime taken by model inferences.

Model	Mean (SD) (ms)
SE (150 MeV)	9 (1)
SE (175 MeV)	9 (1)
SE (200 MeV)	10 (1)
ME	10 (2)
CE	10 (2)

depth. They validated their single PB results on a phantom. While they reported close to perfect agreement in homogenous water phantoms, the mean γ_{PR} (2%|2 mm, 0.1% dose cut off) they reported in bone material were only 91.4% and 85.4% for 150 and 240 MeV beams respectively. For mean ΔR_{D80}, they reported inaccuracies up to 1.0%. A computation time of 100 ms per PB was reported but an increase of up to 25 times was needed when beam splitting techniques were utilized to improve accuracy. Another analytical method (Duetschler *et al* 2023) incorporated a material-specific correction factor to account for different materials causing different beam trajectories in a magnetic field. Although they did not report γ_{PR} performance on a PB basis, they reported a γ_{PR} (2%|2 mm, 1% dose cutoff) of 81.3% for a full-plan lung patient, suggesting that their method may have been challenged by highly heterogeneous geometries. In addition, using a Gaussian fit, they reported extracting the center and width σ of the PB at a depth of 80% of its range. The maximal deviation in the beam center, which is closely related to F_{COM}, was reported to be 1.2 mm, while the maximal reported $\Delta\sigma$ of 2.3 mm corresponds to $FW_{50} \approx 5.4$ mm. Total calculation time of about 30 s for a full plan was reported without specifying the number of PBs in the plan. For the GPU-based MC algorithm ARCHER for MRI-guided proton therapy (Li *et al* 2024), very good agreement was reported when only considering electromagnetic processes in TOPAS simulations, but deviations appeared when all interaction processes were considered. For a 200 MeV beam in water, they reported achieving a $\Delta R_{D80} < 0.06$ mm and a mean ε_{rel} of 0.81%. For tissue and bone materials, a mean ε_{rel} of 1.12% and $F_{COM} < 0.06$ mm were achieved. They also calculated a full treatment plan for three prostate patients in a 1.5 T magnetic field and obtained γ_{PR} (2%|2 mm, 10% dose cut off) >99.5%. The computational time of ARCHER ranged from 0.82 to 4.54 s for 10^7 proton histories.

The performance of the LSTM model was evaluated on 10 test prostate patients, which covers more variety in test samples compared to the above-mentioned studies. It is challenging to compare the LSTM performance to results reported on phantoms. Even so, for some metrics, the performance of our method was noteworthy. For instance, for all PBs of three energies, the γ_{PR} (2%|2 mm, 10% dose cut off) from the SE model was always higher than 94%. Besides, our method showed good performance in terms of FW_{50}, with the worst FW_{50} of 4.2 mm for all PBs of three energies, in contrast to 3.9 and 5.4 mm achieved by the analytical methods. As for F_{COM}, the maximal deviation of the SE LSTM model was 4.7 mm, which is higher than the maximal reported deviation in the analytical methods. It should be noted that the FW_{50} and F_{COM} of the LSTM model were evaluated in stricter settings, with performance specifically reported at very shallow depths and at the Bragg peak, where most deviations occur. In contrast, the analytical studies only assessed a few selected depths in the phantoms before the Bragg peak. Compared with GPU-MC results, the mean γ_{PR} (2%|2 mm, 10% dose cut off) results for PBs in our method were about 99.9%, which is on par with the performance of MC methods in water, but other metrics were worse for LSTM. Additionally, our method is faster than the analytical methods (Padilla-Cabal *et al* 2018). For the runtime comparison with Geant4, it should be noted that the runtime for a PB MC simulation in Geant4 could be reduced by utilizing multi-threading. Furthermore, limiting the simulation to a specific region of interest, rather than the whole

CT, would also decrease computation time. We acknowledge the inherent limitations of this comparison with the cuboid-based deep learning method.

Despite the LSTM model performing very well on average, considerable outliers have been observed in terms of the evaluated metrics. While single outlier PBs may not have the most influence on a final treatment plan, quantifying the uncertainties of a method is an integral part of clinical decision-making. Bayesian neural network methods (Barragan-Montero *et al* 2022, Voss *et al* 2023) could be tested in future studies to assert the uncertainty of deep learning models. MC methods could be used to support deep learning dose calculation in outlier cases, where poor confidence is indicated by the Bayesian model.

We mainly conducted experiments on the variable of energy. For simplicity, we opted for a single spot size for all energies and did not consider range shifters. In general, the model could be trained to learn the usual correlation between energy and spot size. We think that it may be necessary to train a separate dose calculation model specifically for range shifters to account for the additional scattering.

5. Conclusion

We successfully developed an LSTM network-based proton dose calculation method in magnetic fields and confirmed its accuracy for prostate cancer patients. An extended LSTM model for multi-energy dose calculation was proposed and showed its feasibility. The developed deep learning-based sub-second dose engine has the potential to improve the efficiency of real-time plan optimization in the MR-guided proton therapy workflow.

Data availability statement

The data cannot be made publicly available upon publication because they contain sensitive personal information. The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

The work of Fan Xiao was supported by China Scholarship Council (No.202308440107).

Niklas Wahl acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project No. 443188743.

Ethical statement

This retrospective study was exempt from requiring ethics approval. Bavarian state law (Bayrisches Krankenhausgesetz/Bavarian Hospital Law§27 Absatz 4 Datenschutz (Data protection)) allows the use of patient data for research, provided that any personal related data are kept anonymous. German radiation protection laws request a regular analysis of outcomes in the sense of quality control and assurance, thus in the case of purely retrospective studies no additional ethical approval is needed under German law.

Conflict of interest

The Department of Radiation Oncology of the LMU University Hospital Munich has research agreements with Brainlab and Elekta.

Appendix A

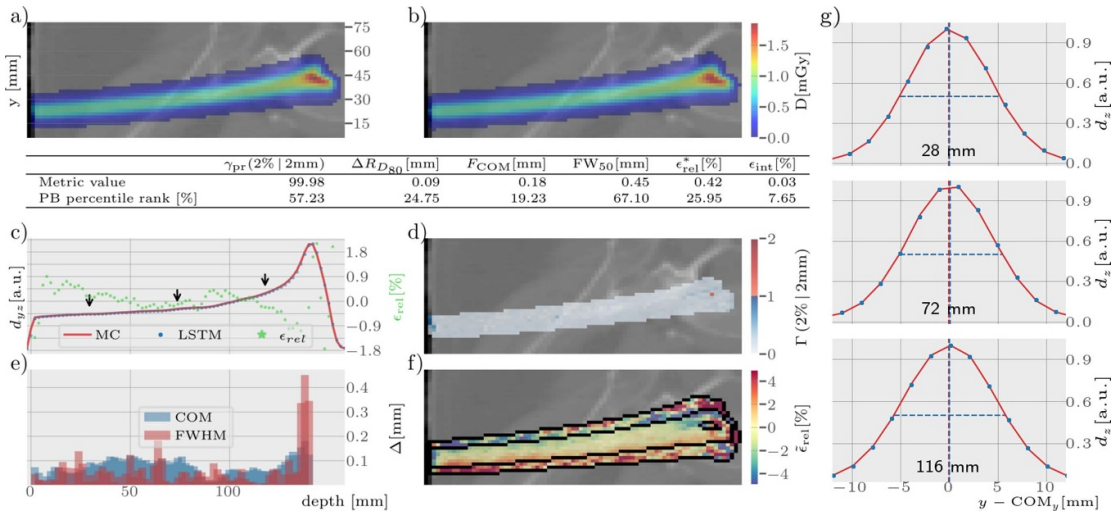


Figure A1. A mid-ranking example of a predicted PB from the SE model in the 150 MeV test dataset. An *xy*-slice of (a) MC and (b) LSTM predicted dose distribution is plotted over the RSP map. (c) Depth-dose profiles $d_{yz}(x)$ (integrated over y and z) and $\varepsilon_{\mathrm{rel}}$ of each slice. The black arrows indicate the depths at which lateral dose profiles $d_z(y)$ (integrated over z) are shown in (g). (d) $\Gamma(2\%|2\,\mathrm{mm})$ map. (e) Differences in COM and FW_{50} at each depth. (f) Relative dose difference of $d_z(x,y)$ capped at $\pm 5\%$. Black lines denote 80%, 10% and 1% isodose lines with respect to $d_z(x_{\mathrm{BP}}, y_{\mathrm{BP}})$ in the Bragg peak. (g) Lateral dose profiles $d_z(y)$ at depths of 20%, 50% and 80% of R_{D80} along the Bragg curve; color coded as in (c).

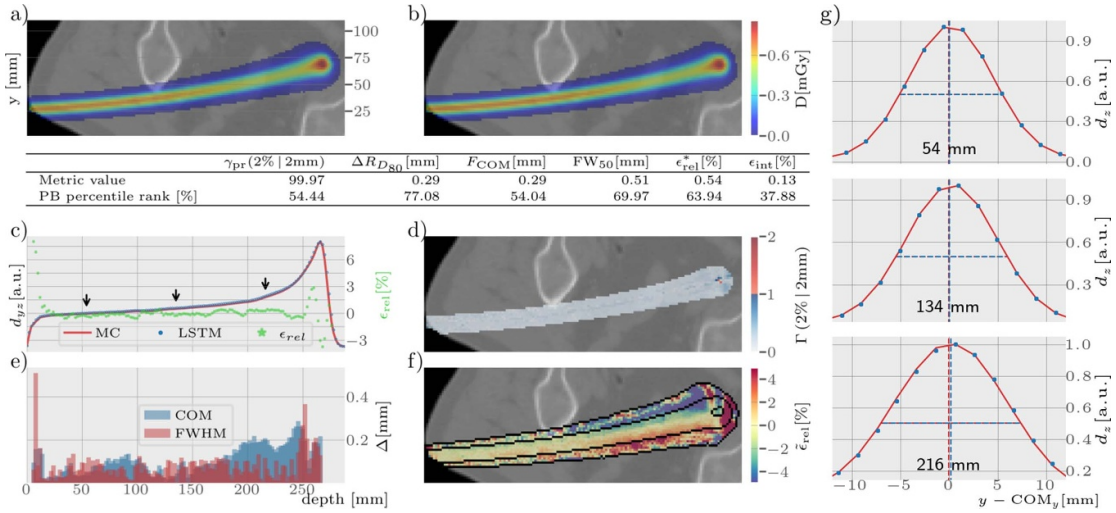


Figure A2. A mid-ranking example of a predicted PB from the SE model in the 200 MeV test dataset. An *xy*-slice of (a) MC and (b) LSTM predicted dose distribution is plotted over the RSP map. (c) Depth-dose profiles $d_{yz}(x)$ (integrated over y and z) and $\varepsilon_{\mathrm{rel}}$ of each slice. The black arrows indicate the depths at which lateral dose profiles $d_z(y)$ (integrated over z) are shown in (g). (d) $\Gamma(2\%|2\,\mathrm{mm})$ map. (e) Differences in COM and FW_{50} at each depth. (f) Relative dose difference of $d_z(x,y)$ capped at $\pm 5\%$. Black lines denote 80%, 10% and 1% isodose lines with respect to $d_z(x_{\mathrm{BP}}, y_{\mathrm{BP}})$ in the Bragg peak. (g) Lateral dose profiles $d_z(y)$ at depths of 20%, 50% and 80% of R_{D80} along the Bragg curve; color coded as in (c).

ORCID iDs

Domagoj Radonic https://orcid.org/0000-0002-0180-7995
Fan Xiao https://orcid.org/0000-0002-7502-0730
Niklas Wahl https://orcid.org/0000-0002-1451-223X
Ahmad Neishabouri https://orcid.org/0000-0001-8060-2957
Guillaume Landry https://orcid.org/0000-0003-1707-4068

References

Ackermann B *et al* 2020 e0404/matRad: blaise v2.10.1 *Zenodo*

Barragan-Montero A *et al* 2022 Towards a safe and efficient clinical implementation of machine learning in radiation oncology by exploring model interpretability, explainability and data-model dependency *Phys. Med. Biol.* **67** 11TR01

Biggs S *et al* 2022 PyMedPhys: a community effort to develop an open, Python-based standard library for medical physics applications *J. Open Source Softw.* **7** 4555

Devlin J, Chang M-W, Lee K and Toutanova K 2018 Bert: pre-training of deep bidirectional transformers for language understanding (arXiv:1810.04805)

Duetschler A, Winterhalter C, Meier G, Safai S, Weber D C, Lomax A J and Zhang Y 2023 A fast analytical dose calculation approach for MRI-guided proton therapy *Phys. Med. Biol.* **68** 195020

Fracchiolla F *et al* 2021 Clinical validation of a GPU-based Monte Carlo dose engine of a commercial treatment planning system for pencil beam scanning proton therapy *Phys. Med.* **88** 226–34

Fuchs H, Moser P, Gröschl M and Georg D 2017 Magnetic field effects on particle beams and their implications for dose calculation in MR-guided particle therapy *Med. Phys.* **44** 1149–56

Hartman J, Kontaxis C, Bol G H, Frank S J, Lagendijk J J, van Vulpen M and Raaymakers B W 2015 Dosimetric feasibility of intensity modulated proton therapy in a transverse magnetic field of 1.5 T *Phys. Med. Biol.* **60** 5955–69

Hoffmann A *et al* 2020 MR-guided proton therapy: a review and a preview *Radiat. Oncol.* **15** 129

Javaid U, Souris K, Huang S and Lee J A 2021 Denoising proton therapy Monte Carlo dose distributions in multiple tumor sites: a comparative neural networks architecture study *Phys. Med.* **89** 93–103

Kluter S 2019 Technical design and concept of a 0.35 T MR-Linac *Clin. Transl. Radiat. Oncol.* **18** 98–101

Kurz C, Landry G, Resch A F, Dedes G, Kamp F, Ganswindt U, Belka C, Raaymakers B W and Parodi K 2017 A Monte-Carlo study to assess the effect of 1.5 T magnetic fields on the overall robustness of pencil-beam scanning proton radiotherapy plans for prostate cancer *Phys. Med. Biol.* **62** 8470

Lane S A, Slater J M and Yang G Y 2023 Image-guided proton therapy: a comprehensive review *Cancers* **15** 2555

Li S, Cheng B, Wang Y, Pei X and Xu X G 2024 A GPU-based fast Monte Carlo code that supports proton transport in magnetic field for radiation therapy *J. Appl. Clin. Med. Phys.* **25** e14208

Lomax A J 2008a Intensity modulated proton therapy and its sensitivity to treatment uncertainties 1: the potential effects of calculational uncertainties *Phys. Med. Biol.* **53** 1027–42

Lomax A J 2008b Intensity modulated proton therapy and its sensitivity to treatment uncertainties 2: the potential effects of inter-fraction and inter-field motions *Phys. Med. Biol.* **53** 1043–56

Luhr A, Burigo L N, Gantz S, Schellhammer S M and Hoffmann A L 2019 Proton beam electron return effect: monte Carlo simulations and experimental verification *Phys. Med. Biol.* **64** 035012

Lysakovski P, Ferrari A, Tessonnier T, Besuglow J, Kopp B, Mein S, Haberer T, Debus J and Mairani A 2021 Development and benchmarking of a Monte Carlo dose engine for proton radiation therapy *Front. Phys.* **9** 741453

Matter M, Nenoff L, Meier G, Weber D C, Lomax A J and Albertini F 2019 Intensity modulated proton therapy plan generation in under ten seconds *Acta Oncol.* **58** 1435–9

Metcalfe P, Liney G P, Holloway L, Walker A, Barton M, Delaney G P, Vinod S and Tome W 2013 The potential for an enhanced role for MRI in radiation-therapy treatment planning *Technol. Cancer Res. Treat.* **12** 429–46

Moteabbed M, Schuemann J and Paganetti H 2014 Dosimetric feasibility of real-time MRI-guided proton therapy *Med. Phys.* **41** 111713

Neishabouri A, Wahl N, Mairani A, Kothe U and Bangert M 2021 Long short-term memory networks for proton dose calculation in highly heterogeneous tissues *Med. Phys.* **48** 1893–908

Padilla-Cabal F, Alejandro Fragoso J, Franz Resch A, Georg D and Fuchs H 2020 Benchmarking a GATE/Geant4 Monte Carlo model for proton beams in magnetic fields *Med. Phys.* **47** 223–33

Padilla-Cabal F, Georg D and Fuchs H 2018 A pencil beam algorithm for magnetic resonance image-guided proton therapy *Med. Phys.* **45** 2195–204

Paganetti H 2012 Range uncertainties in proton therapy and the role of Monte Carlo simulations *Phys. Med. Biol.* **57** R99–117

Paganetti H *et al* 2021 Roadmap: proton therapy physics and biology *Phys. Med. Biol.* **66** 05RM01

Pastor-Serrano O and Perkó Z 2022a Learning the physics of particle transport via transformers *Proc. AAAI Conf. on Artificial Intelligence* vol 36 pp 12071–9

Pastor-Serrano O and Perkó Z 2022b Millisecond speed deep learning based proton dose calculation with Monte Carlo accuracy *Phys. Med. Biol.* **67** 105006

Pham T T, Whelan B, Oborn B M, Delaney G P, Vinod S, Brighi C, Barton M and Keall P 2022 Magnetic resonance imaging (MRI) guided proton therapy: a review of the clinical challenges, potential benefits and pathway to implementation *Radiother. Oncol.* **170** 37–47

Raaymakers B W, Raaijmakers A J and Lagendijk J J 2008 Feasibility of MRI guided proton therapy: magnetic field dose effects *Phys. Med. Biol.* **53** 5615–22

Saini J, Maes D, Egan A, Bowen S R, St James S, Janson M, Wong T and Bloch C 2017 Dosimetric evaluation of a commercial proton spot scanning Monte-Carlo dose algorithm: comparisons against measurements and simulations *Phys. Med. Biol.* **62** 7659–81

Schellhammer S M and Hoffmann A L 2017 Prediction and compensation of magnetic beam deflection in MR-integrated proton therapy: a method optimized regarding accuracy, versatility and speed *Phys. Med. Biol.* **62** 1548–64

Schmid S, Landry G, Thieke C, Verhaegen F, Ganswindt U, Belka C, Parodi K and Dedes G 2015 Monte Carlo study on the sensitivity of prompt gamma imaging to proton range variations due to interfractional changes in prostate cancer patients *Phys. Med. Biol.* **60** 9329–47

Schmidt M A and Payne G S 2015 Radiotherapy planning using MRI *Phys. Med. Biol.* **60** R323–61

Teoh S, Fiorini F, George B, Vallis K A and Van den Heuvel F 2020 Is an analytical dose engine sufficient for intensity modulated proton therapy in lung cancer? *Br. J. Radiol.* **93** 20190583

Unkelbach J, Bortfeld T, Martin B C and Soukup M 2009 Reducing the sensitivity of IMPT treatment plans to setup errors and range uncertainties via probabilistic treatment planning *Med. Phys.* **36** 149–63

Voss L, Neishabouri A, Ortkamp T, Mairani A and Wahl N 2023 BayesDose: comprehensive proton dose prediction with model uncertainty using Bayesian LSTMs (arXiv:2307.01151)

Walters B R, Kawrakow I and Rogers D W 2002 History by history statistical estimators in the BEAM code system *Med. Phys.* **29** 2745–52

Wieser H P *et al* 2017 Development of the open-source dose calculation and optimization toolkit matRad *Med. Phys.* **44** 2556–68

Wolf R and Bortfeld T 2012 An analytical solution to proton Bragg peak deflection in a magnetic field *Phys. Med. Biol.* **57** N329–37
Wu C, Nguyen D, Xing Y, Montero A B, Schuemann J, Shang H, Pu Y and Jiang S 2021 Improving proton dose calculation accuracy by using deep learning *Mach. Learn. Sci. Technol.* **2** 015017
Zhang G, Chen X, Dai J and Men K 2022 A plan verification platform for online adaptive proton therapy using deep learning-based Monte-Carlo denoising *Phys. Med.* **103** 18–25
Zhang X, Hu Z, Zhang G, Zhuang Y, Wang Y and Peng H 2021 Dose calculation in proton therapy using a discovery cross-domain generative adversarial network (DiscoGAN) *Med. Phys.* **48** 2646–60