Millisecond speed deep learning based proton dose calculation with Monte Carlo accuracy

Pastor-Serrano, Oscar; Perkó, Zoltán

DOI
10.1088/1361-6560/ac692e

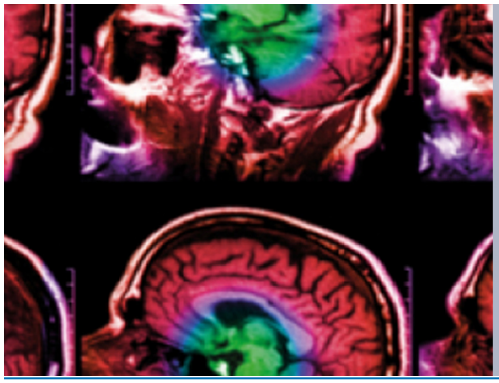
Publication date
2022

Document Version
Final published version

Published in
Physics in Medicine and Biology

PAPER • OPEN ACCESS

Millisecond speed deep learning based proton dose calculation with Monte Carlo accuracy



IOP Publishing Phys. Med. Biol. 67 (2022) 105006 https://doi.org/10.1088/1361-6560/ac692e

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

OPEN ACCESS

RECEIVED
5 February 2022

REVISED
5 April 2022

ACCEPTED FOR PUBLICATION
21 April 2022

PUBLISHED
9 May 2022

PAPER

Millisecond speed deep learning based proton dose calculation with Monte Carlo accuracy

Oscar Pastor-Serrano* and **Zoltán Perkó**

Delft University of Technology, Department of Radiation Science and Technology, Delft, The Netherlands

* Author to whom any correspondence should be addressed.

E-mail: o.pastorserrano@tudelft.nl and z.perko@tudelft.nl

Keywords: deep learning, dose calculation, online adaptation, proton therapy, Monte Carlo, pencil beam

Abstract

Objective. Next generation online and real-time adaptive radiotherapy workflows require precise particle transport simulations in sub-second times, which is unfeasible with current analytical pencil beam algorithms (PBA) or Monte Carlo (MC) methods. We present a deep learning based millisecond speed dose calculation algorithm (DoTA) accurately predicting the dose deposited by mono-energetic proton pencil beams for arbitrary energies and patient geometries. *Approach.* Given the forward-scattering nature of protons, we frame 3D particle transport as modeling a sequence of 2D geometries in the beam's eye view. DoTA combines convolutional neural networks extracting spatial features (e.g. tissue and density contrasts) with a transformer self-attention backbone that routes information between the sequence of geometry slices and a vector representing the beam's energy, and is trained to predict low noise MC simulations of proton beamlets using 80 000 different head and neck, lung, and prostate geometries. *Main results.* Predicting beamlet doses in 5 ± 4.9 ms with a very high gamma pass rate of $99.37 \pm 1.17\%$ (1%, 3 mm) compared to the ground truth MC calculations, DoTA significantly improves upon analytical pencil beam algorithms both in precision and speed. Offering MC accuracy 100 times faster than PBAs for pencil beams, our model calculates full treatment plan doses in 10–15 s depending on the number of beamlets (800–2200 in our plans), achieving a $99.70 \pm 0.14\%$ (2%, 2 mm) gamma pass rate across 9 test patients. *Significance.* Outperforming all previous analytical pencil beam and deep learning based approaches, DoTA represents a new state of the art in data-driven dose calculation and can directly compete with the speed of even commercial GPU MC approaches. Providing the sub-second speed required for adaptive treatments, straightforward implementations could offer similar benefits to other steps of the radiotherapy workflow or other modalities such as helium or carbon treatments.

1. Introduction

Radiotherapy (RT) treatments intimately rely on accurate particle transport calculations. In computed tomography (CT) image acquisition (Pereira *et al* 2014) simulations of the interaction between photons, tissues and detectors are used to obtain a detailed 3D image of the patient anatomy, which can be delineated to localize target structures and organs-at-risk. Modern intensity modulated treatments (Hussein *et al* 2018, Meyer *et al* 2018) require particle transport to compute the spatial distribution of physical dose delivered by thousands of individual electron, photon, proton or other heavy ion beamlets (aimed at the patient from a few different beam angles), based on which the beamlet intensities can be optimized. Treatment plans—especially sensitive proton and ion treatments–must also be repeatedly evaluated under uncertainties (e.g. setup and range errors, tumor motion or complex anatomical changes) to ensure sufficient plan robustness, requiring recalculating the dose distribution in many different scenarios (Perkó *et al* 2016, van der Voort *et al* 2016, Rojo-Santiago *et al* 2021). With RT practice steadily moving towards adaptive treatments, accurate, fast and general purpose dose (and particle transport) calculations represent an increasingly pressing, currently unmet need in most clinical settings.

We focus our attention specifically to proton dose calculations due to their more challenging nature caused by higher sensitivity and complexity compared to traditional photons. Current physics-based tools—by and large falling into 2 categories: analytical pencil beam algorithms (PBAs) (Hong *et al* 1996, Schaffner *et al* 1999) and Monte Carlo (MC) simulations—offer a trade-off between speed and precision. While PBAs yield results without the computational burden of MC engines, their accuracy is severely compromised in highly heterogeneous or complex geometries, making slow and clinically often not affordable MC approaches necessary (Teoh *et al* 2020, Schuemann *et al* 2015, Taylor *et al* 2017, Grassberger *et al* 2014, Saini *et al* 2017). The problem is most acute for online (and ultimately real-time) adaptive proton therapy aiming at treatment correction prior to (or even during) delivery to account for inter-fractional anatomical changes, motion due to breathing, coughs or intestinal movements. To become reality, such adaptive treatments require algorithms yielding MC accuracy with sub-second speed.

Reducing dose calculation times is an active area of research, with most works focusing on improving existing physics-based algorithms or developing deep learning frameworks. Several studies benefit from the parallelization capabilities of Graphics Processing Units (GPUs) to massively speed up MC simulations, reducing calculations times down to the range of few seconds (Fracchiolla *et al* 2021, Wan Chan Tseung *et al* 2015) to minutes (Ma *et al* 2014, Gajewski *et al* 2021, Pepin *et al* 2018, Wang *et al* 2016, Qin *et al* 2016), with simulation speeds up to 10^7 protons s^{-1}. Deep learning methods have also improved dose calculation times in several steps of the RT workflow (Meyer *et al* 2018), although usually paying the price of limited versatility and generalization capabilities. Some initial studies apply variants of U-net (Ronneberger *et al* 2015) and Generative Adversarial Networks (Goodfellow *et al* 2014) to aid treatment planning by approximating dose distributions from 'optimal' plans in very specific scenarios based on historical data. As input to these convolutional architectures, most works use organ and tumor masks (Chen *et al* 2019, Fan *et al* 2019, Nguyen *et al* 2019, Kajikawa *et al* 2019), CT images (Kearney *et al* 2018) or manually encoded beam information (Nguyen *et al* 2019, Barragán-Montero *et al* 2019) to directly predict full dose distributions, except for few papers predicting the required beam intensities needed to deliver such doses (Lee *et al* 2019, Wang *et al* 2020).

Regarding pure dose calculation, practically all deep learning applications rely on using computationally cheaper physics simulations as additional input apart from CTs. For photons, most works predict low noise MC dose distributions from high noise MC doses (Peng *et al* 2019, 2019, Bai *et al* 2021, Neph *et al* 2021) or simple analytical particle transport calculations (Xing *et al* 2020, Dong and Xing 2020), with some approaches also utilizing additional manually encoded beam/physics information such as fluence maps (Fan *et al* 2020, Xing *et al* 2020, Zhu *et al* 2020, Kontaxis *et al* 2020, Tsekas *et al* 2021). For protons, we are only aware of 3 papers (Wu *et al* 2021, Javaid *et al* 2021, Nomura *et al* 2020) that compute proton dose distributions via deep learning, using cheap physics models (noisy MC and PBA) or pre-calculated Bragg peak maps as input. While providing significant speed-up compared to pure physics-based algorithms, some even reaching sub-second speeds, all these works depend on secondary physics models to produce their output or are trained to predict only full plan or field doses for specific treatment sites. As a result, these methods do not qualify as generic dose algorithms and do not generalize to other steps of the RT workflow outside their original scope, e.g. to different plan or field configurations, treatment sites, or applications needing the individual dose distribution from each beamlet separately (such as treatment adaptation).

Instead, our study focuses on learning particle transport physics to substitute generic proton dose engines, providing millisecond speed and high accuracy, and is in principle applicable to all RT steps requiring dose calculations (e.g. dose-influence matrix calculation, dose accumulation, robustness evaluation). Our approach builds upon a previous study (Neishabouri *et al* 2021) using long short-term memory (LSTM) networks (Hochreiter and Schmidhuber 1997) to sequentially calculate proton pencil beam dose distributions from relative stopping power slices in sub-second times, but with the major disadvantage of requiring a separate model per beam energy. As shown in figure 1, we frame proton transport as modeling a sequence of 2D geometry slices in the beam's eye view, introducing an attention-based transformer backbone (Vaswani *et al* 2017) that dynamically routes information between elements of the sequence along beam depth. We extend on our previous work only focusing on lung cancer (Pastor-Serrano and Perkó 2021), training with a larger set of patients and treatment sites, and evaluating performance both for individual pencil beams and full treatment plans. The presented Dose Transformer algorithm (DoTA)—able to learn the physics of energy dependence in proton transport via a single model—can predict low noise MC proton pencil beam dose distributions purely from beamlet energy and CT data in ≈5 ms. Based on our experiments and available literature data, in terms of accuracy and overall speed DoTA significantly outperforms pencil beam algorithms and all other deep learning approaches (e.g. LSTM models (Neishabouri *et al* 2021) and 'denoising' networks (Wu *et al* 2021, Javaid *et al* 2021, Nomura *et al* 2020)), representing the current state-of-the-art in data-driven proton dose calculations and directly competing with (and even improving on) GPU Monte Carlo approaches.

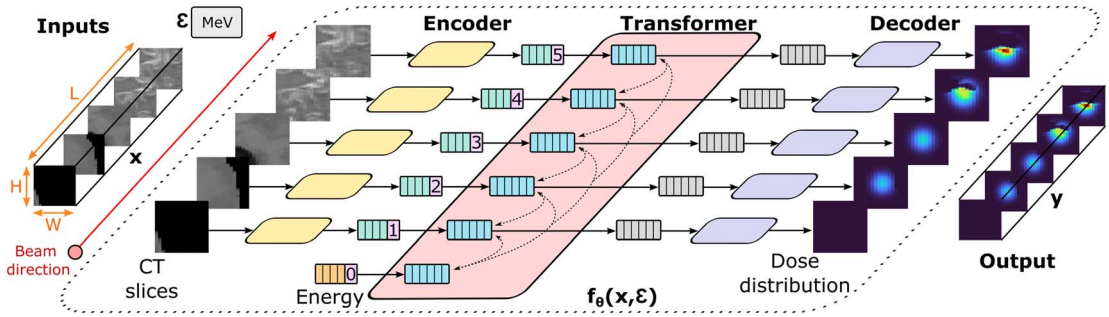


Figure 1. Dose transformer algorithm (DoTA). A data-driven model learns a mapping $\boldsymbol{y} = f_{\boldsymbol{\theta}}(\boldsymbol{x}, \varepsilon)$ between input CT cubes $\boldsymbol{x}$ and energies ε and output dose distributions $\boldsymbol{y}$. CT and dose distribution 3D volumes are both treated as a sequence of 2D slices in the beam's eye view. An encoder and a decoder individually transform each 2D slice into a feature vector and vice versa, whereas a transformer backbone routes information between different vectors along beam depth.

2. Methods and materials

The problem of dose calculation is common to many steps of RT workflow and ultimately involves estimating the spatial distribution of physical dose from thousands of pencil beams. A generic deep learning dose engine must be capable of calculating 3D dose distributions for arbitrary patient geometries purely from a list of beam directions and energies for a given beam model, without being conditioned on the type of treatment or task being solved. Therefore, our objective is to accurately predict dose distributions $\boldsymbol{y}$ from individual proton beamlets in sub-second speed, given patient geometries $\boldsymbol{x}$ and beam energies ε. We introduce DoTA, a parametric model that implicitly captures particle transport physics from data and learns the function $\boldsymbol{y} = f_{\boldsymbol{\theta}}(\boldsymbol{x}, \varepsilon)$ via a series of artificial neural networks with parameters $\boldsymbol{\theta}$.

In particular, DoTA learns a mapping between a 3D CT input voxel grid $\boldsymbol{x} \in \mathbb{R}^{L\times H\times W}$ and output dose distribution $\boldsymbol{y} \in \mathbb{R}^{L\times H\times W}$ conditioned on the energy $\varepsilon \in \mathbb{R}^{+}$, where L is the depth (in the direction of beam propagation), H is the height and W is the width of the grid. While traditional physics-based calculation tools process the entire geometry, we crop and interpolate the CT to the reduced sub-volume seen by protons as they travel through the patient, with a fixed 2 mm $\times$ 2 mm $\times$ 2 mm resolution and $L \times H \times W$ size. Framing proton transport as sequence modeling, DoTA processes the input volume as a series of L 2D slices in the forward beam direction. Ideally, the exchange of information between the different elements in the sequence should be dynamic, i.e. the contribution or impact of each 2D slice to the sequence depends on both its position and material composition. Unlike other types of artificial neural networks, the Transformer architecture (Vaswani *et al* 2017)—and specifically the self-attention mechanism—is notably well suited for this.

Recently, Transformer-based architectures have replaced their recurrent counterparts in many natural language processing (Devlin *et al* 2019, Brown *et al* 2020) and computer vision tasks (Ramachandran *et al* 2019, Dosovitskiy *et al* 2020, Touvron *et al* 2020, D'Ascoli *et al* 2021). For modeling the sequentiality in proton transport physics, the advantage of Transformers with respect to LSTM frameworks is two-fold. First, every element can directly access information at any point in the sequence without requiring an internal hidden state, which is crucial to include beam energy dependence. The routing of information—referred to as self-attention—is different for every element, allowing each geometry slice to be independently transformed based on the information it selectively gathers from other slices in the sequence. Second, Transformers allow manually encoding the mostly forward scattering nature of proton transport by restricting interaction to only previous slices via causal attention. Transformers typically run multiple self-attention operations in parallel (known as attention heads), with each head focusing on modeling separate features of the sequence. We provide a detailed description of the fundamentals of self-attention and the Transformer module in appendix A.

2.1. Model architecture and training

Figure 2 shows DoTA's architecture, which first applies the same series of convolutions to each 2D slice of the input sequence $\{\boldsymbol{x}_i | \boldsymbol{x}_i \in \mathbb{R}^{1\times H\times W}, \forall\, i = 1,...,L\}$ separately. This convolutional encoder contains two blocks—both with a convolution, a Group Normalization (GN) (Wu and He 2020) and a pooling layer, followed by a Rectified Linear Unit (ReLU) activation—which extract important features from the input, e.g. material contrasts and tissue boundaries. After the second block, the outputs of a final convolution with K filters are flattened into a vector of embedding dimension $D = H' \times W' \times K$, where H' and W' are the reduced height and width of the images after the pooling operations. The convolutional encoder applies the same operation to

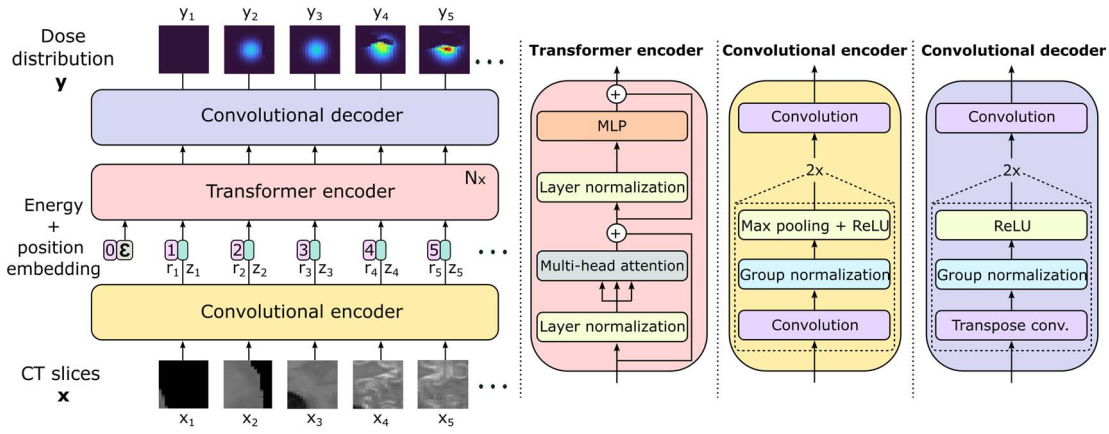


Figure 2. DoTA architecture. We treat the input and output 3D volumes as a sequence of 2D slices. A convolutional encoder extracts important geometrical from each slice into a feature vector. The particle energy is added at the beginning of the resulting sequence. A transformer encoder with causal self-attention subsequently combines information from the different elements of the sequence. Finally, a convolutional decoder individually transforms the low-dimensional vectors into output 2D dose slices.

every element $\boldsymbol{x}_i$, resulting in a sequence of L vectors $\{\boldsymbol{z}_i | \boldsymbol{z}_i \in \mathbb{R}^D, \forall\, i = 1,..., L\}$ referred to as tokens in the remainder of the paper.

A Transformer encoder models the interaction between tokens $\boldsymbol{z}_i$ via causal self-attention, resulting in an output sequence $\boldsymbol{z}' \in \mathbb{R}^D$. Since Transformers operate on sets and by default do not account for the relative position of the slices in the sequence, we add a learnable positional encoding $\boldsymbol{r}_i \in \mathbb{R}^D$ to each token $\boldsymbol{z}_i$, e.g. $\boldsymbol{r}_1$ is always added to the token $\boldsymbol{z}_1$ from the first slice seen by the proton beam. The energy dependence is included via a 0th token $\boldsymbol{z}_0 = \boldsymbol{W}_0 \varepsilon \in \mathbb{R}^D$ at the beginning of the sequence, where $\boldsymbol{W}_0 \in \mathbb{R}^{D\times 1}$ is a learned linear projection of the beam energy ε. We use the standard pre-Layer Normalization (LN) (Ba *et al* 2016) Transformer block (Xiong *et al* 2020), alternating LN and residual connections with a self-attention operation and a feed-forward block with two fully-connected layers, Dropout (Srivastava *et al* 2014) and a Gaussian Error Linear Unit activation (Hendrycks and Gimpel 2016).

Finally, a convolutional decoder independently transforms every output token to a 2D slice of the same size as the input $\{\boldsymbol{y}_i | \boldsymbol{y}_i \in \mathbb{R}^{1\times H\times W}, \forall\, i = 1,..., L\}$. The decoder's structure is identical to that of its encoder counterpart, but substituting the down-sampling convolution + pooling operation in the with an up-sampling convolutional transpose layer.

Dataset We train DoTA to predict low noise MC dose distributions calculated with MCsquare (Souris *et al* 2016), obtained using a set of 30 CT scans from prostate, lung and head and neck (H&N) cancer patients (Aerts *et al* 2014, 2015, Clark *et al* 2013) with 2 mm isotropic grid resolution. Given that proton beams have approximately 25 mm diameter and travel up to 300 mm through a small sub-volume of the CT, we crop blocks $\boldsymbol{x} \in \mathbb{R}^{150\times 24\times 24}$ covering a volume of approximately $48 \times 48 \times 300\ \text{mm}^3$. From each patient CT, we obtain $\approx$ 2500 of such blocks—corresponding to beamlets being shot at different angles and positions—by effectively rotating and linearly interpolating the CT scan in steps of 10° and by applying 10 mm lateral shifts.

For each block, we calculate 2 different dose distributions using 10^7 primary particles to ensure MC noise values around 0.3% and always below 0.5%, zeroing out dose values below noise levels. Both dose distributions correspond to a randomly sampled beam energy between 70 and 220 MeV, with a 140 MeV cap in lung and H&N geometries given the potential to overshoot the patient. As a result, we obtain ≈80,000 individual CT block–dose distribution input–output pairs. This amount is further quadrupled by rotating the CT and dose blocks in steps of 90° around the beam direction axis, yielding a final training dataset consisting of ≈320 000 samples, 10% of which are used as a validation set to prevent overfitting.

Our evaluation is based on an independent test set of 18 additional patients unseen during training, equally split into prostate, H&N and lung. Half of these patients (3 prostate, 3 H&N and 3 lung) are used to obtain 3888 test beamlet dose distributions (1386 lung, 1512 H&N and 990 prostate samples), with the other half serving to evaluate DoTA's performance in full plans.

Training details The model is trained end-to-end using Tensorflow (Abadi *et al*), with the LAMB optimizer (You *et al* 2019) and 8 samples per mini-batch, limited by the maximum internal memory of the Nvidia Tesla T4® GPU used during our experiments. We use a mean squared error loss function and a scheduled learning rate starting at 10^{-3} that is halved every 4 epochs, with a restart after 28 epochs. In total, we train the model for 56 epochs, saving the weights resulting in the lowest validation mean squared error. The best performing model consists of one transformer block with 16 heads and 12 convolutional filters in the last encoder layer, as obtained

from a hyperparameter grid search evaluating the lowest validation loss across all possible combinations of transformer layers $N \in \{1, 2, 4\}$, convolutional filters $K \in \{8, 10, 12, 16\}$ and attention heads $N_h \in \{8, 12, 16\}$. Given the two down-sampling pooling operations, the transformer processes tokens of dimension $D = H/4 \times W/4 \times K$, which in our case with initial height $H = 24$, width $W = 24$, and $K = 12$ kernels results in $D = 432$.

2.2. Model evaluation

Using the ground truth MC dose distributions in the test set, we compare DoTA to several data-driven dose engines, including LSTM models (Neishabouri *et al* 2021), and deep learning frameworks using noisy MC (Javaid *et al* 2021) and PBA (Wu *et al* 2021) doses as additional input. Since PBA is the analytical dose calculation method commonly used in the clinic and one of DoTA's competitors in terms of speed and accuracy, we include the PBA baseline from the open-source treatment planning software matRad (Wieser *et al* 2017) (https://e0404.github.io/matRad/).

Test set accuracy metrics In our evaluation, the main mechanism to compare predictions to ground truth 3D dose distributions from the test set is the gamma analysis (Low *et al* 1998), further explained in appendix B. To reduce the gamma evaluation to a single number per sample, we report the gamma pass rate as the fraction of passed voxels over the total number of voxels. All calculations are based on the PyMedPhys gamma evaluation functions (available at https://docs.pymedphys.com).

Additionally, the average relative error ρ is used to explicitly compare dose differences between two beamlet dose distributions. Given the predicted output $\boldsymbol{y}$ and the ground truth dose distribution $\hat{\boldsymbol{y}}$ with $n_v = L \times H \times W$ voxels, the average relative error can be calculated as

$$\rho = \frac{1}{n_v} \frac{\|\hat{\boldsymbol{y}} - \boldsymbol{y}\|_{L_1}}{\max \hat{\boldsymbol{y}}} \times 100. \tag{1}$$

Since the models are trained using a mean squared error (MSE) cost function, we also compute the root mean squared error (RMSE) between ground truth and predicted beamlet dose distributions, defined as

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} = \sqrt{\frac{1}{n_v} \sum_{i=1}^{n_v} .(\hat{y}_i - y_i)^2}. \tag{2}$$

Finally, as an alternative metric to the gamma pass rate for comparing full dose distributions, we calculate the relative dose error (RDE) (Nomura *et al* 2020) between the ground truth and predicted D_{95}, D_{90}, D_{50} and D_{20} values, where D_v is the dose received by v% of the tumor volume. The RDE is computed relative to the planned dose D_{pr} as

$$\mathrm{RDE}_v = \frac{D_v - \hat{D}_v}{D_{pr}} \times 100. \tag{3}$$

Experiments A generic data-driven dose engine must yield accurate predictions for both single beamlet and full plan dose distributions. To ensure DoTA's suitability for replacing conventional particle transport tools in dose prediction tasks, we assess its performance in two different settings:

- Individual beamlets. First, we evaluate the speed and accuracy in predicting single beamlet doses for 9 patients in the test set and compare gamma pass rate distributions and inference times of DoTA, the LSTM models and the PBA baseline. Given the 2 mm × 2 mm × 2 mm grid resolution, a gamma evaluation $\Gamma(3\text{ mm}, 1\%)$ using a distance-to-agreement criterion $\delta = 3$ mm ensures a neighborhood search of at least one voxel, while a dose criterion $\Delta = 1\%$ disregards any uncertainty due to MC noise. Since DoTA's outputs are hardly ever 0 due to numerical inaccuracies of the last convolutional linear layer, and to disregard voxels not receiving any dose, we exclude voxels with doses below 0.1% of the maximum dose for the gamma pass rate calculations, resulting in a stricter metric (as the many voxels with near 0 dose could artificially increase the passing rate). Additionally, we compute the relative error ρ and RMSE between PBA/DoTA predictions and MC dose distributions. For both ρ and the gamma pass rate, we compare probability densities across all test samples.
- Full plans. A treatment plan with 2 fields is obtained for the remaining 9 test set patients using matRad. Given the list of beam intensities and energies in the plan, we recalculate dose distributions using PBA, MCsquare (Souris *et al* 2016) and DoTA, and evaluate their performance via the gamma pass rate, masking voxels receiving a dose lower than 10% of the maximum dose. For each field angle in the treatment plan, we rotate the original CT, calculate the dose from each beamlet and rotate back the entire field dose its original angle for dose accumulation. To allow for a fair comparison with other data-driven models—referred to as baselines B1 (Javaid *et al* 2021) and B2 (Wu *et al* 2021)—we compute three gamma evaluations $\Gamma(1\text{ mm}, 1\%)$, $\Gamma(2\text{ mm}, 2\%)$ and $\Gamma(3\text{ mm}, 3\%)$ and compare the pass rate results to the available values in these baseline studies. Since the

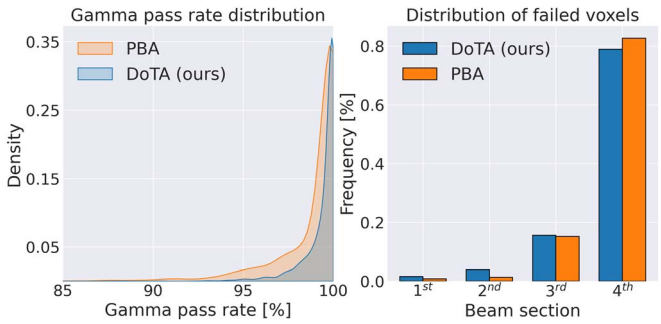


Figure 3. Gamma pass rate distribution. (Left) Distribution of the gamma pass rates Γ(3 mm, 1%) of the test samples for the pencil beam algorithm (PBA) and the presented DoTA model. (Right) Distribution of the failed voxels along the beam, where each bin is an equally-sized fragment (referred to as section) of the beam from dose entrance (1st) to Bragg peak and dose falloff (4th). Each bin shows the ratio of the number of test set voxels that fail the gamma evaluation within a section divided by the total number of failed voxels.

Table 1. Overview of experiments. Summary of the experiments, metrics and baselines used to evaluate DoTA's accuracy. $D_{\max}$ refers to the maximum dose value in a dose distribution and only voxels receiving dose above the cutoff level are included in the Γ calculations.

Experiment	Test data	Metric	Dose cutoff (Gy)	Baseline
Individual beamlets	3888 beamlets	Γ(3 mm, 1%)	0	LSTM
	1386 lung,		0.1% of $D_{\max}$	PBA
	990 prostate,	Error ρ	0	PBA
	1512 H&N	RMSE	0	PBA
Full plans	9 treatment plans	Γ(1 mm, 1%)	10% of $D_{\max}$	PBA, B2
		Γ(2 mm, 2%)	10% of $D_{\max}$	B1
		$\text{RDE}_{v\in\{20,50,90,95\}}$	Tumor doses	B3

third baseline B3 (Nomura *et al* 2020) does not report a gamma pass rate, we compare RDEs with the values reported in the paper. For more information about the experiments, table 1 contains a description of the metrics and evaluation settings.

3. Results

In this section, DoTA's performance and speed is compared to state-of-the-art models and clinically used methods. The analysis is three-fold: we assess the accuracy in predicting beamlet dose distributions and full dose distributions from treatment plans, and explore DoTAs' potential as a fast dose engine by evaluating its calculation runtimes.

3.1. Individual beamlets

For each individual beamlet in the test set, DoTA's predictions are compared to MC ground truth dose distributions using a Γ(3 mm, 1%) gamma analysis. In table 2, we report the average, standard deviation, minimum and maximum of the distribution of gamma pass rates across test samples. By disregarding voxels whose dose is below 0.1% of the maximum dose, our gamma evaluation approach is stricter than that of previous state-of-the-art studies (Neishabouri *et al* 2021), where only voxels with a gamma value of 0—which typically correspond to voxels not receiving any dose—are excluded from the pass rate calculation. Even with the stricter setting and including energy dependence, DoTA outperforms both the LSTM and PBA dose engines in all aspects: the average pass rates are higher, the standard deviation is lower, and the minimum is at least 5.5% higher. Similar results are observed for stricter gamma evaluation settings in appendix C. The left plot in figure 3 further demonstrates DoTA's superiority, showing a gamma pass rate distribution that is more concentrated towards higher values. We subsequently divide each beam dose distribution into 4 fragments of equal size between the entrance and the Bragg peak, where each fragment is referred to as *beam section* in the remainder of the paper. The right plot in figure 3 shows the proportion of voxels failing the gamma evaluation in each beam section, out of the total number of failed voxels, indicating for both PBA and DoTA that most of the failing

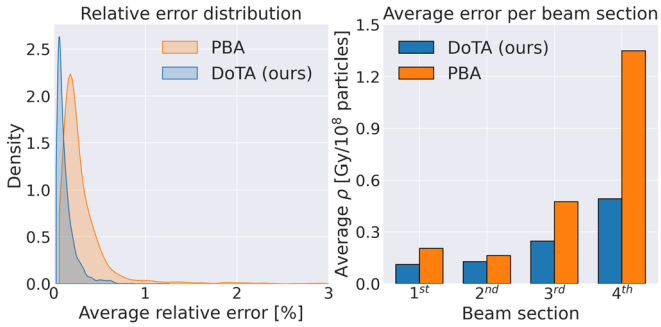


Figure 4. Average relative error ρ distribution. (Left) Distribution of the average relative error across the test samples for the pencil beam algorithm (PBA) and the presented DoTA model. (Right) Average relative error per beam section, where each bin is a section (4 equally-sized fragments) of the beam from dose entrance (1st) to Bragg Peak and dose falloff (4th). Each bin shows the average of the relative error values recorded within a section of the beam.

Table 2. Gamma pass rate of beamlet dose distributions. Gamma analysis results Γ(3 mm, 1%) for the presented DoTA, the pencil beam algorithm (PBA) from matRad (Wieser *et al* 2017) and the LSTM models are listed. Gamma pass rates are calculated using all test samples, with LSTM rates directly obtained from (Neishabouri *et al* 2021). The reported values include the mean, standard deviation (Std), minimum (Min) and maximum (Max) across the test set for different sites, and 'Multi-site' refers to computing statistics using all sites.

Model	Site	Energy (MeV)	Mean (%)	Std (%)	Min (%)	Max (%)
LSTM (Neishabouri *et al* 2021)	Lung	67.85	98.56	1.3	95.35	99.79
		104.25	97.74	1.48	92.57	99.74
		134.68	94.51	2.99	85.37	99.02
DoTA (ours)	Lung	[70, 140]	99.46	0.81	93.19	100
	H&N	[70, 140]	99.21	1.23	93.49	100
	Prostate	[70, 220]	99.51	1.46	94.06	100
DoTA (ours)	Multi-site	[70, 220]	99.37	1.17	93.19	100
PBA (matRad)	Multi-site	[70, 220]	98.68	3.14	87.53	100

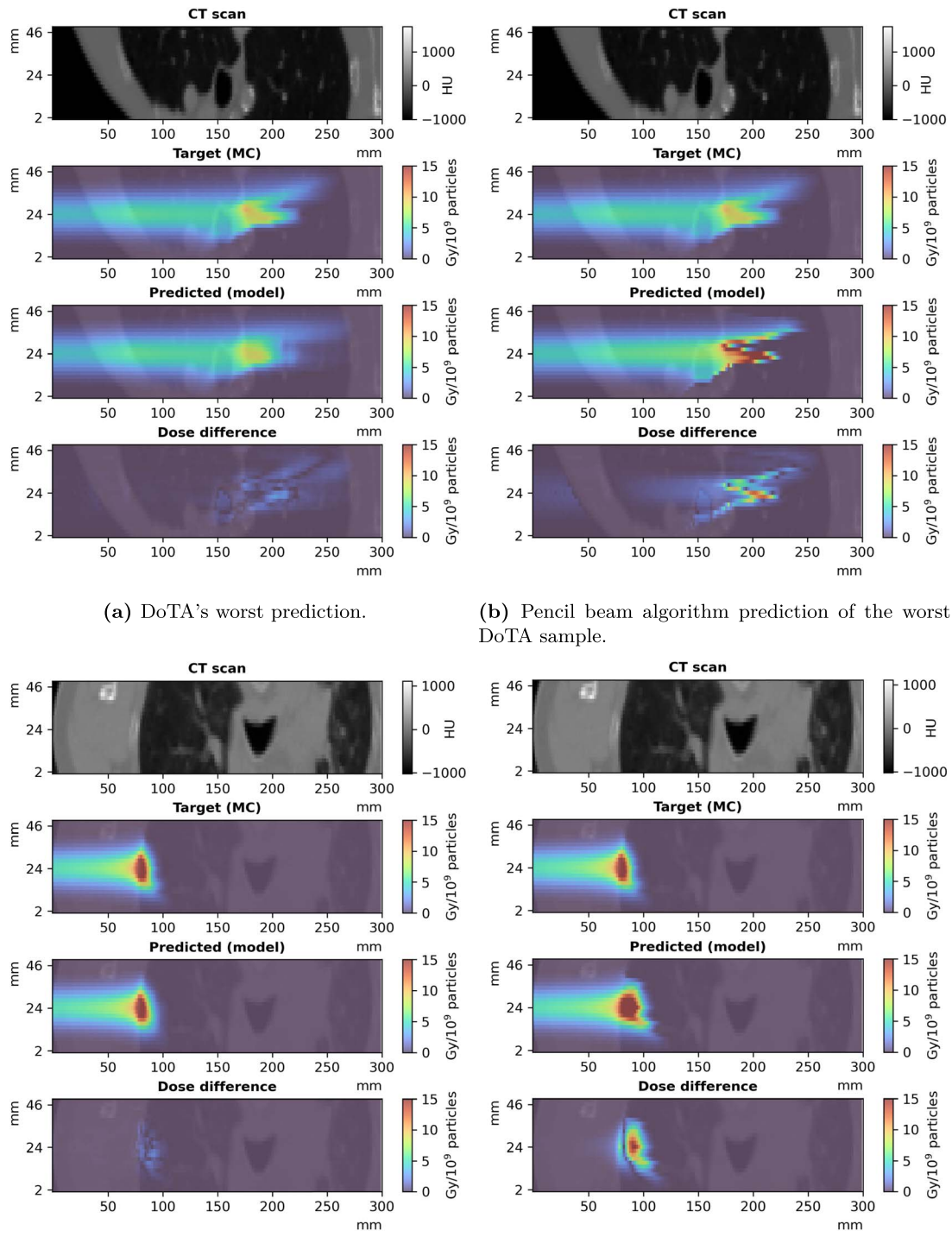
Table 3. Error of beamlet dose distributions. The reported values include the mean, standard deviation (Std), minimum (Min) and maximum (Max) values of the relative error ρ and root mean squared error (RMSE) between 3888 test predictions and reference MC dose distributions, for both the pencil beam algorithm (PBA) from matRad (Wieser *et al* 2017) and DoTA.

Model	Relative error ρ (%)				RMSE (Gy)			
	Mean	Std	Min	Max	Mean	Std	Min	Max
DoTA (ours)	0.126	0.109	0.025	1.258	0.083	0.041	0.024	0.277
PBA (matRad)	0.306	0.309	0.059	4.077	0.294	0.126	0.057	1.293

voxels belong to the 4th section, i.e. the high dose region around the Bragg peak where the effect of tissue heterogeneity is most evident.

As an additional measure of model performance, table 3 shows the mean and standard deviation of the relative error ρ and RMSE between predictions and ground truth MC dose distributions in the test set. The results confirm DoTA's improvement, with mean, maximum error and standard deviation less than half of PBA's. The left plot in figure 4 displays the distribution of ρ across all test samples, showing that values are smaller and closer to 0 for DoTA. As with the gamma pass rate, the beam is divided in 4 sections from entrance (1st) to the Bragg peak (4th), and the average relative error per section is shown in the right plot in figure 4. Although both models show a similar trend with errors increasing towards the beam's end, DoTA is on average twice better than PBA.

Finally, figure 5(b) shows DoTA's test sample with the lowest gamma pass rate, together with PBA's prediction of the same sample (figure 5(a)). Likewise, figures 5(d) and (c) show the predictions of the worst PBA sample from both models. In both cases, PBA results in errors as high as 80% of the maximum dose, severely overdosing parts of the geometry, while for DoTA errors are below 20% of the maximum dose.



(a) DoTA's worst prediction.

(b) Pencil beam algorithm prediction of the worst DoTA sample.

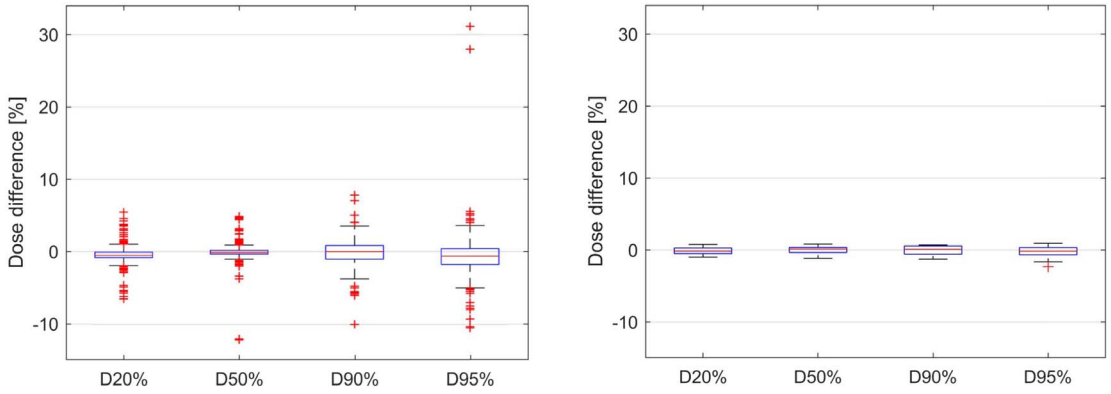
(c) DoTA's prediction of the worst pencil beam algorithm sample.

(d) Pencil beam algorithm worst prediction.

Figure 5. Worst performing DoTA and PBA test sample. (a) Worst performing test sample in the gamma evaluation for DoTA, with gamma pass rate of 93.19%, and (b) the pencil beam algorithm (PBA) prediction for the same sample. (d) Worst performing prediction in the gamma evaluation across the test set for PBA, with gamma pass rate of 87.53%, and (c) DoTA's prediction of the same sample. In descending order, all 4 subplots show: the central slice of the 3D input CT grid, the MC ground truth dose distribution, the model's prediction and the dose difference between the predicted and MC beams.

3.2. Full dose recalculation

To assess the feasibility of using DoTA as a dose engine in real clinical settings, we recalculate full dose distributions from treatment plans and compare them to MC reference doses via 3 different gamma analysis: Γ(1 mm, 1%), Γ(2 mm, 2%) and Γ(3 mm, 3%), in decreasing order of strictness. The resulting gamma pass rates for each of the 9 test patients are shown in table 4, showing values that are consistently high and similar across treatment sites, always at least 10% higher than PBA. We additionally compare DoTA to recently published



(a) B3 baseline. (b) DoTA.

Figure 6. Relative dose errors. Error between the ground truth and predicted D_{95}, D_{90}, D_{50} and D_{20} for (a) the B3 baseline (Reproduced from Nomura *et al* 2020. © 2020 Institute of Physics and Engineering in Medicine. All rights reserved.) and (b) the proposed DoTA model, relative to planned doses. Red crosses are outliers, red lines represent the median, and box boundaries denote the 25th and 75th percentiles.

Table 4. Gamma pass rate of planned dose distributions. Treatment plans of 9 test patients are recalculated using the presented DoTA model, and compared to ground truth MC dose distributions via 3 different gamma analysis: $\Gamma(1\text{ mm}, 1\%)$, $\Gamma(2\text{ mm}, 2\%)$ and $\Gamma(3\text{ mm}, 3\%)$. We additionally include the $\Gamma(1\text{ mm}, 1\%)$ pass rate for dose distributions recalculated with the pencil beam algorithm (PBA) from matRad (Wieser *et al* 2017). The baseline B1 corresponds to a MC-denoising U-net (Javaid *et al* 2021), while B2 is a U-net correcting PBA (Wu *et al* 2021), whose values are directly taken for their corresponding papers.

Site	Patient	Number of spots	DoTA (ours)			PBA	B1 (Javaid *et al* 2021)	B2 (Wu *et al* 2021)
			$\Gamma(1, 1\%)$	$\Gamma(2, 2\%)$	$\Gamma(3, 3\%)$	$\Gamma(1, 1\%)$	$\Gamma(2, 2\%)$	$\Gamma(1, 1\%)$
Lung	1	954	95.86	99.73	99.99	80.38	84.1	89.7±3.8
	2	2245	96.31	99.72	99.98	79.83		
	3	1646	95.63	99.64	99.97	78.92		
H&N	4	1554	95.02	99.39	99.81	68.32	76.5	92.8±2.9
	5	1064	94.71	99.62	99.97	76.63		
	6	708	96.93	99.88	99.99	83.02		
Prostate	7	1598	96.38	99.81	99.99	87.34	—	99.6±0.3
	8	2281	95.78	99.82	99.99	77.12		
	9	1518	96.18	99.71	99.98	83.64		

state-of-the-art deep learning approaches: a MC-denoising U-net (Javaid *et al* 2021) (B1), and a U-net correcting PBA (Wu *et al* 2021) (B2). Except for the prostate plans, DoTA outperforms both approaches, even without requiring the additional physics-based input.

Figure 6 shows the RDE of DoTA and the B3 baseline (a convolutional neural network predicting dose distributions from Bragg peak position maps). B3 results are taken directly from the paper Nomura *et al* 2020. Reproduced from Nomura *et al* 2020, © 2020 Institute of Physics and Engineering in Medicine. All rights reserved, while DoTA values are computed using all test set dose distributions. With a significantly lower spread and values much closer to 0%, the results further confirm DoTA's superiority and accuracy gains.

3.3. Runtime

Apart from high prediction accuracy, fast inference is critically important for clinical applications. Table 5 displays the mean and standard deviation runtime taken by each model to predict a single beamlet. Being particularly well-suited for GPUs, DoTA is on average faster than LSTM and physics-based engines, offering more than 100 times speed-up with respect to PBA. Additionally, although dependent on hardware, DoTA approximates doses four orders of magnitude faster than MC, providing millisecond dose calculation times without requiring any extra computations for real-time adaptive treatments.

Regarding full dose recalculation from treatment plans, figure 7 shows total runtimes for DoTA using both GPU and CPU hardware, including all steps from loading CT and beamlet weights from plan data files, necessary CT rotations and interpolations, DoTA dose inference time and reverse rotations and interpolation to assign dose on the original CT grid. Being optimized for GPU acceleration, DoTA is the fastest alternative, needing less than 15 s to calculate full dose distributions. For the baselines in this paper, we find that PBA runtimes oscillate between 100 and 150 s, while B1 and B2 report needing only few seconds to correct/denoise their inputs, but

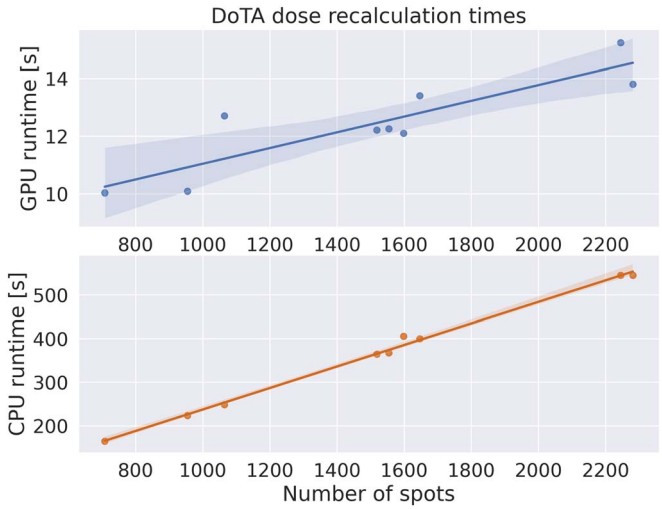


Figure 7. Full dose recalculation runtime. Time needed to recalculate planned dose distributions with DoTA using (top) a Nvidia® A100 GPU or (bottom) an intel Xeon® CPU. Estimates include time for loading CT and beam weights from plan data, for dose inference by DoTA and for the necessary CT and dose interpolations. Shaded areas denote the 95% confidence interval.

Table 5. Beamlet prediction runtime. The reported values include the mean inference time and standard deviation (Std) taken by each model to predict individual beamlet dose distributions. Both the DoTA and LSTM models run on GPU hardware, while the pencil beam algorithm (PBA) (Wieser *et al* 2017) and Monte Carlo (MC) dose engine use CPUs with multiple threads. LSTM inference times are taken directly from (Neishabouri *et al* 2021).

Model	Mean (ms)	Std (ms)
LSTM[a](Neishabouri *et al* 2021)	6.0	1.5
DoTA[b] (ours)	5.0	4.9
PBA[c] (matRad)	728.3	30.9
MC[c] (Souris *et al* 2016)	43 636.9	12 291.6

[a] Nvidia® Quadro RTX 6000 64 Gb RAM.
[b] Debian 10 4 vCPUs—Nvidia® A100 40 Gb RAM.
[c] CentOS 7 8 CPUs intel Xeon® E5-2620 16Gb RAM.

must add the runtime necessary to generate their respective PBA (123–303 s in Wu *et al* (2021)) or MC (≈10 s in Javaid *et al* (2021)) input doses, as well as data transfer times between the physics engine and the deep learning framework. Furthermore, B2 is a per beam network, hence its runtime scales linearly with the number of beams, in practice meaning 2–4 times higher total calculation times.

4. Discussion

In this study, we present a data-driven dose engine predicting dose distributions with high accuracy. The presented DoTA model builds upon previous work learning proton transport as sequence modeling task via LSTM networks (Neishabouri *et al* 2021), by introducing energy dependence and significantly improving its performance in a varied set of treatment sites. DoTA greatly outperforms analytical physics-based PBA algorithms in predicting dose distributions from individual proton pencil beams, achieving high accuracy even in the most heterogeneous patient geometries, demonstrated by the 6% improvement in the minimum gamma pass rate. With millisecond inference times, DoTA provides at least a factor 100 reduction in calculation time compared to the clinically still predominant analytical PBAs.

The drastic reduction in spot dose prediction times translates into the ability to calculate full dose distributions in 12 s on average and less than 15 s even for the plan with more than 2200 pencil beams, which times include the required time for all steps from loading CT and pencil beam weights from plan data (≈1 s on average), CT interpolation and beamlet geometry extraction (≈1 s), DoTA model and weights loading (≈2 s),

dose inference by DoTA (≈7.5 s) and interpolating the final dose distribution back to the original CT grid (≈1 s). Although publicly available deep learning frameworks are optimized for GPU architectures and may offer an advantage with respect to adapting MC and PBA to GPU hardware, we achieve this 10–15 s speed on a single GPU card, even without any optimization of GPU settings for inference, which can reportedly yield up to 9 times speed-ups depending on the task[1]. Without sacrificing accuracy, DoTA represents at least a factor 10 speed-up with respect to PBAs and a 33% speed-up (and ≈80% considering the difference in MC noise levels) with respect to the fastest GPU MC competitor we could find in the literature—clinically used GPU MC software Raystation® (Fracchiolla *et al* 2021), typically running in clusters or workstations with multiple GPUs and CPU cores. Moreover, DoTA offers a 10%–25% increase in the Γ(1 mm, 1%) gamma pass rate compared to PBA, and with a Γ(2 mm, 2%) gamma pass rate <99% it matches (Wang *et al* 2016) or outperforms (Wan Chan Tseung *et al* 2015, Qin *et al* 2016) the accuracy of GPU MC approaches. DoTA's accuracy is also on par with the agreement between commercial MC engines (Raystation®) and experimental measurements (Schreuder *et al* 2019, 2019). While the GPU-based PBA algorithm reported in da Silva *et al* (2015) calculates a full distribution in 0.22 s and is faster than DoTA, it was tested only on a single patient showing worse accuracy with a 3% lower Γ(2 mm, 2%) pass rate.

Our method is also substantially superior to the only 3 published deep learning approaches for proton full plan dose calculations (Javaid *et al* 2021, Wu *et al* 2021, Nomura *et al* 2020). We achieve 15% and 25% higher Γ(2 mm, 2%) pass rates compared to the MC-denoising U-net of Javaid *et al* (2021), and 6% and 2% higher Γ(1 mm, 1%) pass rates compared to the PBA correcting U-net of Wu *et al* (2021) in lung and H&N patients, respectively. With lower RDE values much more concentrated around 0, DoTA also improves upon the dose prediction U-net based on Bragg peak position maps (Nomura *et al* 2020). DoTA shows a slight inferiority in prostate patients, with a ≈3% lower Γ(1 mm, 1%) pass rates than (Wu *et al* 2021). However, this direct comparison is somewhat unfair to DoTA. In our work we evaluate performance on Intensity Modulated Proton Therapy plans with a small, 3–5 mm spot size, while in Wu *et al* (2021) double scattering proton therapy plans were used, which in general are less conformal and smoother, and therefore are expected to be easier to predict with data-driven approaches. We also use a finer voxel resolution of 2 mm × 2 mm × 2 mm compared to the 2 mm × 2 mm × 2.5 mm used in Wu *et al* (2021). Furthermore, Wu *et al* (2021) also reports site specific fine-tuning of their deep learning approach, unlike our method. Last, Wu *et al* (2021) has the further disadvantage of using per beam PBA calculations as input, thus the reported 2–3 s dose correction times easily translate to full treatment plan calculation times in the 5–10 min range depending on the number of beams (taking into account the >2 min PBA run times), even without accounting for the additional time for the necessary CT rotations and interpolations.

DoTA's accuracy may further be increased by training with larger datasets, as demonstrated by the improvement achieved when increasing training data from 4 lung patients in our earlier work (Pastor-Serrano and Perkó 2021) to 30 patients with varied anatomies in the current study. Using dose distributions with lower MC noise could further improve performance. Convincingly outperforming all recent works learning corrections for 'cheap' physics-based predictions (Wu *et al* 2021, Javaid *et al* 2021) both in terms of accuracy and speed, DoTA has the flexibility to be used in a great variety of treatment sites and clinical settings.

Application DoTA's accuracy and speed improvements outperform existing approaches and represent a new state-of-the-art that could benefit current RT practice in numerous aspects. The small number of potential geometries currently used to evaluate treatment plan robustness—whose size is limited by the speed of the dose calculation algorithm—can be extended with many additional samples, capturing a more diverse and realistic set of inter- and intra-fraction (Pastor-Serrano and Perkó 2021) geometrical variations. DoTA's capability to quickly and accurately estimate fraction dose distributions based on pre-treatment daily CT images could transform dosimetric quality assurance protocols, enabling direct comparison between the planned and estimated doses or even online adaptation of plans (Jagt *et al* 2017, 2018, Albertini *et al* 2020). Most crucially, by pre-computing the input volumes and updating their CT values in real time, the millisecond speed for individual pencil beam dose calculation makes our model well suited for real-time correction during radiation delivery.

Limitations The current version of DoTA is trained to predict MC ground truth dose distributions from a specific machine with unique settings and beam profiles, necessitating a specific model per machine. Likewise, range shifters—which are often dependent on treatment location and site—affect the dose delivered by some spots while inserted, thereby modifying the final dose distribution. Both problems could in principle be addressed by constructing a model that takes extra shape and range shifter specifications as input in the form of tokens at the beginning of the sequence, similar to our approach for treating the energy dependence.

DoTA is trained for a specific voxel grid resolution, requiring either an individual model per resolution level or an additional interpolation step that will likely negatively interfere with the gamma pass rate results, especially for gamma evaluations Γ(1, 1%) with a distance-to-agreement criterion lower than the voxel resolution level.

[1] Discussed in the non-peer-reviewed study in https://huggingface.co/transformers/v2.10.0/benchmarks.html.

While DoTA also works for finer nominal CT grids (Pastor-Serrano and Perkó 2021), an additional study testing the dose recalculation performance with more patients and finer grid resolution should confirm its suitability for direct clinical application needing such resolutions. MC noise may also affect the results of the gamma evaluation, as demonstrated in previous work (Cohilis *et al* 2020) showing that even 1% MC noise levels introduce significant under-estimation in the gamma pass rate. In our evaluation, we expect this detrimental effect to be limited given our lower noise levels of 0.3% in the ground truth MC doses (which level is considered as reference 'denoised' in Cohilis *et al* (2020)).

One of the main problems of deep learning algorithms is their limited generalization or extrapolation capability outside the domain of the used training dataset. In our evaluation, performed in an independent test set of patients with varied geometries unseen during training, DoTA is clearly superior to all other methods in all evaluated scenarios, showing strong evidence of high level of generalization. Nevertheless, just like any deep learning approach, DoTA may also yield unrealistic predictions for data that vastly differs from the training data (e.g. in the presence of metallic implants), contrarily to MC engines, which—when using enough particles—are certain to provide valid results. Whether or not 'more physics-based' PBAs perform better than DoTA in such cases is less straightforward. First, PBA clearly performed worse than DoTA in all our tests, and in particular showed worse performance in the examples of figure 5 exhibiting high heterogeneity (figures 5(a)–(b)) and the Bragg peak position coinciding with a sharp change in density (figures 5(c)–(d), further highlighted on the coronal views in figure 9 in appendix C). Second, the impact of approximations inherent to PBA approaches on the predicted dose in cases of unusual geometries (e.g. implants) is not easy to foresee without detailed analysis. The same holds for the error due to DoTA's potential generalization limitations in such cases. While we do not have direct evidence for it, physics-based approaches (even approximative ones) may maintain a higher level of accuracy when going far beyond the training dataset domain. For the specific case of radiotherapy however, to a large extent these problems could be mitigated by including geometries with metallic implants in the training data set and teaching DoTA to accurately predict dose distributions in such scenarios too and by limiting use to (the vast majority of) patients who do not have implants until such improved model is available.

Future work Besides the possibility to include shape, machine and beam characteristics as additional input tokens in the transformer, several extensions can widen its spectrum of applications, such as predicting additional quantities, e.g. particle flux, or estimating radiobiological weighted dose—potentially including simulating even DNA damage—typically significantly slower than pure MC dose calculation. A clinically highly relevant follow-up study is to include geometries with metallic implants in the training dataset and ensuring prediction accuracy in such challenging geometries too. Alternatively, future work adapting DoTA to learn photon physics would facilitate its use in conventional radiotherapy applications or provide CT/CBCT imaging reconstruction techniques with the necessary speed for real-time adaptation. Most importantly, DoTA offers great potential to speed up dose calculation times in heavy ion treatments with particles such as carbon and helium sharing similar, mostly forward scatter physics, whose MC dose calculation often take much longer to simulate all secondary particles generated as the beam travels through the patient.

5. Conclusion

We present DoTA: a generic, fast and accurate dose engine that implicitly learns proton particle transport physics and can be applied to speed up several steps of the radiotherapy workflow. Framing particle transport as sequence modeling of 2D geometry slices in the proton's beam travel direction, we use the power of transformers to predict individual beamlets with millisecond speed and close to MC precision. Our evaluation shows that DoTA has the right attributes to potentially replace the proton dose calculation tools currently used in the clinics for applications that critically depend on runtime. Predicting dose distributions from single pencil beams in milliseconds, DoTA offers 100 times faster inference times than widely used PBAs, yielding close to MC accuracy as indicated by the very high gamma pass rate $\Gamma(3\,\text{mm}, 1\%)$ of 99.37 ± 1.17, thus has the potential to enable next generation online and real-time adaptive radiotherapy cancer treatments. The presented model predicts MC quality full plan dose distributions with at least a 10% improvement in gamma pass rate $\Gamma(1\,\text{mm}, 1\%)$ with respect to current analytical approaches and reduces dose calculation times of planned doses to less than 15 s, representing a tool that can directly benefit current clinical practice too.

Acknowledgments

This work is supported by KWF Kanker Bestrijding [grant number 11 711] and is part of the KWF research project PAREL. Zoltán Perkó would like to thank the support of the NWO VENI grant ALLEGRO (016. Veni.198.055) during the time of this study.

Code availability

The code, weights and results are publicly available at https://github.com/opaserr/dota.

CRediT authorship contribution statement

Oscar Pastor-Serrano: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing—original draft, Visualization.

Zoltán Perkó: Conceptualization, Methodology, Formal Analysis, Resources, Writing—original draft, Writing—Review and editing, Supervision, Project Administration, Funding Acquisition.

Appendix A. Transformer and self-attention

Transformer DoTA's backbone is the Transformer (Vaswani *et al* 2017), based on the self-attention mechanism. Though originally introduced for sequential modeling applications in natural language processing such as machine translation, Transformers have recently achieved state-of-the-art performance across a wide variety of tasks, with large language (Devlin *et al* 2019, Brown *et al* 2020) or computer vision (Dosovitskiy *et al* 2020) models replacing and outperforming recurrent or convolutional architectures. One of the main reasons behind the success of attention-based models is the ability to model interactions between a large sequence of elements without needing an internal memory state. In Transformers, each sequence element is transformed based on the information it selectively gathers from other members of the sequence based on its content or position. In practice, however, the computational memory requirements scale quadratically with the length of the sequence, and training such large Transformers often requires a pre-training stage with a large amount of data.

Self-attention Given a sequence $\boldsymbol{z} \in \mathbb{R}^{L\times D}$ with L tokens, the self-attention (SA) mechanism (Vaswani *et al* 2017) is based on the interaction between a series of queries $\boldsymbol{Q} \in \mathbb{R}^{L\times D_h}$, keys $\boldsymbol{K} \in \mathbb{R}^{L\times D_h}$, and values $\boldsymbol{V} \in \mathbb{R}^{L\times D_h}$ of dimensionality D_h obtained through a learned linear transformation of the input tokens with weights $\boldsymbol{W}_{QKV} \in \mathbb{R}^{D\times 3D_h}$ as

$$[\boldsymbol{Q},\ \boldsymbol{K},\ \boldsymbol{V}] = \boldsymbol{z}\boldsymbol{W}_{QKV}. \tag{4}$$

Each token is transformer into a query, key and value vector. Intuitively, for an ith token $\boldsymbol{z}_i \in \mathbb{R}^{1\times D}$, the query $\boldsymbol{q}_i \in \mathbb{R}^{1\times D_h}$ represents the information to be gathered from other elements of the sequence, while the key $\boldsymbol{k}_i \in \mathbb{R}^{1\times D_h}$ contains token's information to be shared with other sequence members. The token $\boldsymbol{z}_i$ is then transformed into $\boldsymbol{z}_i{}'$ via a weighted sum of all values in the sequence $\boldsymbol{v}_j \in \mathbb{R}^{1\times D_h}$ as

$$\boldsymbol{z}_i{}' = \sum_{j=1}^{L} w_j \boldsymbol{v}_j, \tag{5}$$

where each weight is based on a the similarity between the ith query and the other keys in the sequence, measured as the dot product $w_j = \boldsymbol{q}_i^T \boldsymbol{k}_j$. The output sequence of transformed tokens $\boldsymbol{z} \in \mathbb{R}^{L\times D}$ is the result of the SA operation applied to all sequence elements, defined by the attention matrix containing all weights $\boldsymbol{A} \in \mathbb{R}^{L\times L}$ and the operations

$$\boldsymbol{A} = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{D_h}}\right), \tag{6}$$

$$\boldsymbol{z}' = \mathrm{SA}(\boldsymbol{z}) = \boldsymbol{A}\boldsymbol{V}. \tag{7}$$

A variant of SA called multi-head self-attention (MSA) runs N_h parallel SA operations focusing on different features or inter-dependencies of the data. Setting $D_h = D$, the outputs of the different SA operations, called *heads*, are first concatenated and then linearly projected with learned weights $\boldsymbol{W}_h \in \mathbb{R}^{N_h D_h\times D}$ as

$$\mathrm{MSA}(\boldsymbol{z}) = \underset{h\in\{N_h\}}{\mathrm{concat}}[\mathrm{SA}_h(\boldsymbol{z})]\,\boldsymbol{W}_h. \tag{8}$$

By definition, every token can attend to all previous and future tokens. Causal SA is a variant of SA applied to sequence modeling tasks restricting access to future information, where all elements above the diagonal in the attention matrix $\boldsymbol{A}$ are masked to 0. Additionally, since SA is invariant to the relative order of elements in the sequence, a fixed (Vaswani *et al* 2017) or learned (Dosovitskiy *et al* 2020) positional embedding $\boldsymbol{r} \in \mathbb{R}^{L\times D}$ is usually added or concatenated to the input tokens, where is element in the positional embedding sequence contains unique information about its position.

Transformer encoder The causal MSA Transformer backbone in DoTA is responsible of routing information between the geometry slices and the energy token. A learnable positional embedding $\boldsymbol{r}$ is added to

the sequence of tokens produced by the convolutional encoder, while we add the first 0th position embedding $\boldsymbol{r}_0$ in the sequence to the energy token. The transformer encoder is formed by alternating MSA and Multi-layer Perceptron (MLP) layers with residual connections, and applying Layer Normalization (LN) applied before every layer (Ba *et al* 2016). Therefore, the Transformer encoder blocks computes the operations

$$\boldsymbol{z} = [\boldsymbol{z}_e; \boldsymbol{z}] + \boldsymbol{r}, \tag{9}$$

$$\boldsymbol{s}_n = \boldsymbol{z} + \mathrm{MSA}(\mathrm{LN}(\boldsymbol{z})), \tag{10}$$

$$\boldsymbol{z}' = \boldsymbol{s}_n + \mathrm{MLP}(\mathrm{LN}(\boldsymbol{s}_n)), \tag{11}$$

where MLP denotes a two layer feed-forward network with Dropout (Srivastava *et al* 2014) and Gaussian Error Linear Unit (GELU) activations (Hendrycks and Gimpel 2016).

Appendix B. Gamma analysis

The gamma analysis is based on the notion that doses delivered in neighboring voxels have similar biological effects. Intuitively, for a set reference points—the voxel centers in the ground truth 3D volume—and their corresponding dose values, this method searches for similar predicted doses within small spheres around each point. The sphere's radius is referred to as distance-to-agreement criterion, while the dose similarity is usually quantified as a percentage of the reference dose, e.g. dose values are accepted similar if within 1% of the reference dose. Each voxel with coordinates $\boldsymbol{a}$ in the reference grid is compared to points $\boldsymbol{b}$ of the predicted dose grid and assigned a gamma value $\gamma(\boldsymbol{a})$ according to

$$\gamma(\boldsymbol{a}) = \min_{\boldsymbol{b}} \{\Gamma_{\boldsymbol{a},\boldsymbol{b}}(\delta, \Delta)\}, \tag{12}$$

$$\Gamma_{\boldsymbol{a},\boldsymbol{b}}(\delta, \Delta) = \sqrt{\frac{|\boldsymbol{a} - \boldsymbol{b}|^2}{\delta^2} + \frac{|\hat{y}_{\boldsymbol{a}} - y_{\boldsymbol{b}}|^2}{\Delta^2}}, \tag{13}$$

where $\hat{y}_{\boldsymbol{a}}$ is the reference dose at point $\boldsymbol{a}$, δ is the distance-to-agreement, and Δ is the dose difference criterion. A voxel passes the gamma analysis if $\gamma(\boldsymbol{a}) < 1$.

Appendix C. Additional results

Table C1 shows additional results for the accuracy of the beamlet dose predictions, using stricter gamma evaluation settings. DoTA's superiority over PBA is clearly demonstrated under these stricter conditions too, with significantly higher mean and minimum passing rates, as well as smaller standard deviation values. Figure 8 displays the beam eye's view of the worst PBA test sample (corresponding to the sample shown figure 5(d)) around the Bragg peak, showing the transition from lung tissue to air that results in an erroneous predicted dose distribution. The coronal view in figure 9 further confirms that denser bone tissue from the ribs near the lung air boundary is likely to exacerbate prediction errors.

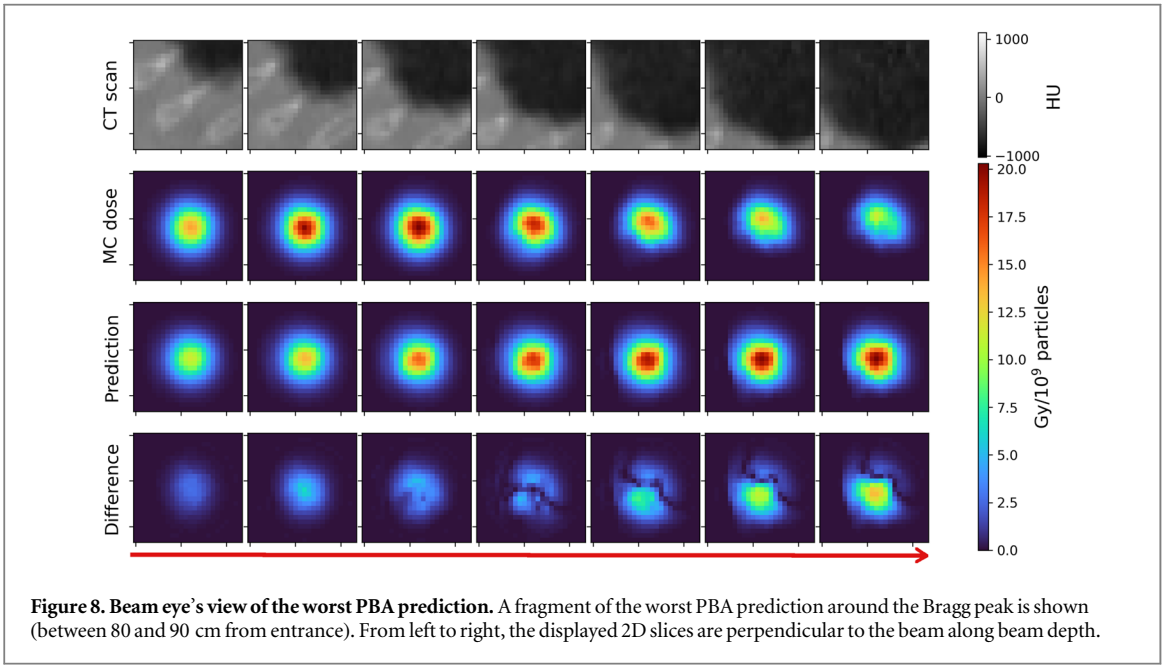
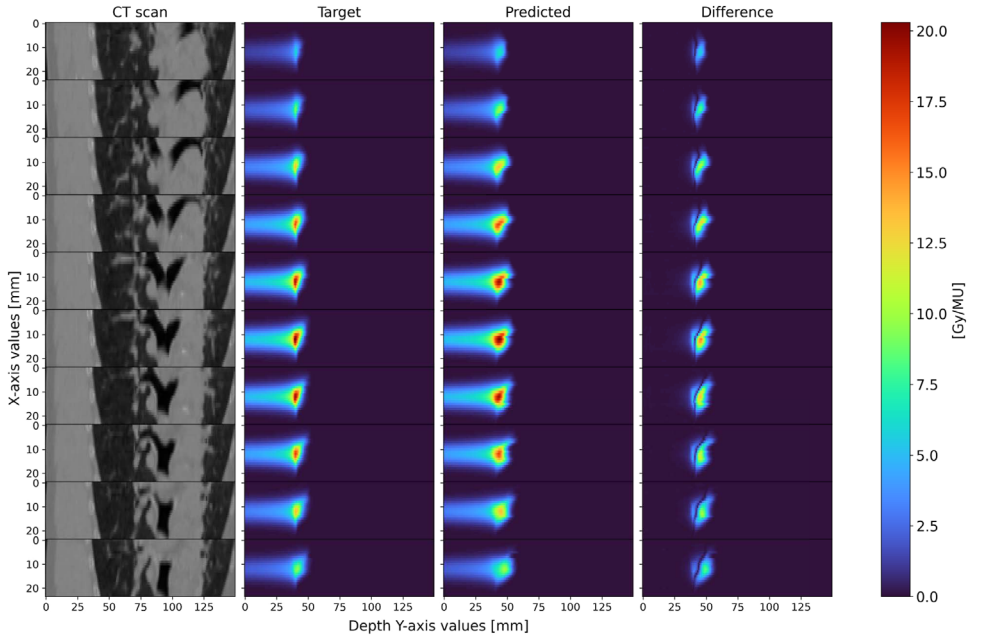
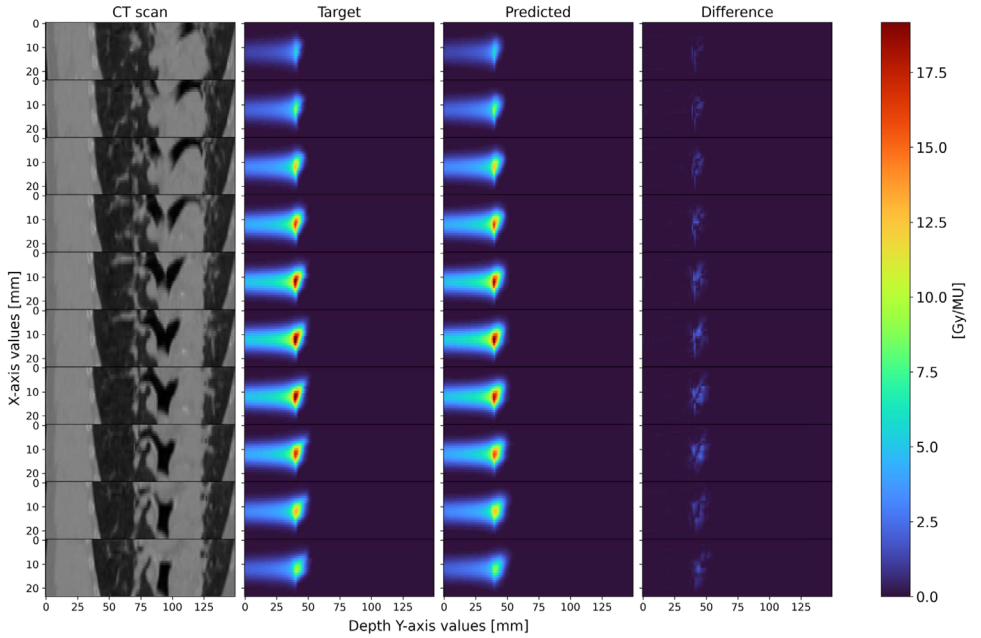


Figure 8. Beam eye's view of the worst PBA prediction. A fragment of the worst PBA prediction around the Bragg peak is shown (between 80 and 90 cm from entrance). From left to right, the displayed 2D slices are perpendicular to the beam along beam depth.



(a) PBA worst prediction.



(b) DoTA's prediction of the worst PBA sample.

Figure 9. Coronal view of the worst PBA prediction. The coronal plane view of the dose worst PBA sample is shown for (a) PBA and (b) DoTA predictions. From top to bottom, each row corresponds to a 2 millimeter step, where the column 'Difference' displays absolute dose differences between predictions and the 'Target' ground truth MC dose distribution.

Table C1. Gamma pass rate of beamlet dose distributions. Gamma analysis $\Gamma(1\,\text{mm}, 1\%)$ and $\Gamma(2\,\text{mm}, 1\%)$ for DoTA and the pencil beam algorithm (PBA) from matRad (Wieser *et al* 2017) are listed. The reported values include the mean, standard deviation (Std), minimum (Min) and maximum (Max) across all test samples.

Model	Energy (MeV)	Settings	Mean (%)	Std (%)	Min (%)	Max (%)
DoTA (ours)	[70, 220]	$\Gamma(1\,\text{mm}, 1\%)$	96.58	3.83	82.31	100
		$\Gamma(2\,\text{mm}, 1\%)$	98.67	2.04	89.69	100
PBA (matRad)	[70, 220]	$\Gamma(1\,\text{mm}, 1\%)$	92.54	6.07	65.21	99.41
		$\Gamma(2\,\text{mm}, 1\%)$	97.20	4.27	76.49	100

ORCID iDs

Oscar Pastor-Serrano https://orcid.org/0000-0002-2328-1429
Zoltán Perkó https://orcid.org/0000-0002-0975-4226

References

Abadi M *et al* 2015 Tensorflow: large-scale machine learning on heterogeneous distributed systems (https://doi.org/10.5281/zenodo.4724125)
Aerts H *et al* 2014 Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach *Nat. Commun.* **5** 4006
Aerts H *et al* 2015 Data From NSCLC-Radiomics-Genomics (https://doi.org/10.7937/K9/TCIA.2015.L4FRET6Z)
Albertini F, Matter M, Nenoff L, Zhang Y and Lomax A 2020 Online daily adaptive proton therapy *The British Journal of Radiology* **vol 93** 1107
Ba J L, Kiros J R and Hinton G E 2016 Layer normalization arXiv:1607.06450
Bai T, Wang B, Nguyen D and Jiang S 2021 Deep dose plugin: towards real-time monte carlo dose calculation through a deep learning-based denoising algorithm *Mach. Learn.: Sci. Technol.* **2** 25033
Barragán-Montero A M, Nguyen D, Lu W, Lin M-H, Norouzi-Kandalan R, Geets X, Sterpin E and Jiang S 2019 Three-dimensional dose prediction for lung imrt patients with deep neural networks: robust learning from heterogeneous beam configurations *Med. Phys.* **46** 3679–91
Brown T B *et al* 2020 Language models are few-shot learners arXiv:2005.14165 Adv. Neural Inf. Process. Syst.2020-Decem
Chen X, Men K, Li Y, Yi J and Dai J 2019 A feasibility study on an automated method to generate patient-specific dose distributions for radiotherapy using deep learning *Med. Phys.* **46** 56–64
Clark K *et al* 2013 Cancer imaging archive (TCIA): maintaining and operating a public information repository *J. Digit. Imaging* **26** 1045–57
Cohilis M, Sterpin E, Lee J A and Souris K 2020 A noise correction of the γ-index method for Monte Carlo dose distribution comparison *Med. Phys.* **47** 681–92
da Silva J, Ansorge R and Jena R 2015 Sub-second pencil beam dose calculation on GPU for adaptive proton therapy *Physics in Medicine and Biology* vol 60 (Bristol: IOP Publishing) pp 4777–95
D'Ascoli S, Touvron H, Leavitt M, Morcos A, Biroli G and Sagun L 2021 Convit: improving vision transformers with soft convolutional inductive biases arXiv:2103.10697
Devlin J, Chang M W, Lee K and Toutanova K 2019 Bert: pre-training of deep bidirectional transformers for language understanding arXiv:1810.04805 NAACL HLT 2019-2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies—Proceedings of the Conference14171–86
Dong P and Xing L 2020 Deep dosenet: a deep neural network for accurate dosimetric transformation between different spatial resolutions and/or different dose calculation algorithms for precision radiation therapy *Phys. Med. Biol.* **65** 35010
Dosovitskiy A *et al* 2020 An image is worth 16x16 words: transformers for image recognition at scale arXiv:2010.11929
Fan J, Wang J, Chen Z, Hu C, Zhang Z and Hu W 2019 Automatic treatment planning based on three-dimensional dose distribution predicted from deep learning technique *Med. Phys.* **46** 370–81
Fan J, Xing L, Dong P, Wang J, Hu W and Yang Y 2020 Data-driven dose calculation algorithm based on deep u-net *Phys. Med. Biol.* **65** 245035
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
Gajewski J *et al* 2021 Commissioning of gpu-accelerated monte carlo code fred for clinical applications in proton therapy *Front. Phys.* **8** 1
Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y 2014 Generative Adversarial Networks arXiv:1406.2661
Grassberger C, Daartz J, Dowdell S, Ruggieri T, Sharp G and Paganetti H 2014 Quantification of proton dose calculation accuracy in the lung *Int. J. Radiat. Oncol. Biol. Phys.* **89** 424–30
Hendrycks D and Gimpel K 2016 Gaussian error linear units (gelus) arXiv: 1606.08415
Hochreiter S and Schmidhuber J 1997 Long short-term memory *Neural Comput.* **9** 1735–80
Hong L, Goitein M, Bucciolini M, Comiskey R, Gottschalk B, Rosenthal S, Serago C and Urie M 1996 A pencil beam algorithm for proton dose calculations *Phys. Med. Biol.* **41** 1305–30
Hussein M, Heijmen B J M, Verellen D and Nisbet A 2018 Automation in intensity modulated radiotherapy treatment planninga review of recent innovations *Br. J. Radiol.* **91** 20180270
Jagt T, Breedveld S, van de Water S, Heijmen B and Hoogeman M 2017 Near real-time automated dose restoration in IMPT to compensate for daily tissue density variations in prostate cancer *Phys. Med. Biol.* **62** 4254–72
Jagt T, Breedveld S, van Haveren R, Heijmen B and Hoogeman M 2018 An automated planning strategy for near real-time adaptive proton therapy in prostate cancer *Phys. Med. Biol.* **63** 135017
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
Kajikawa T, Kadoya N, Ito K, Takayama Y, Chiba T, Tomori S, Nemoto H, Dobashi S, Takeda K and Jingu K 2019 A convolutional neural network approach for imrt dose distribution prediction in prostate cancer patients *J. Radiat. Res.* **60** 685–93
Kearney V, Chan J W, Haaf S, Descovich M and Solberg T D 2018 Dosenet: a volumetric dose prediction algorithm using 3d fully-convolutional neural networks *Phys. Med. Biol.* **63** 235022
Kontaxis C, Bol G H, Lagendijk J J W and Raaymakers B W 2020 Deepdose: towards a fast dose calculation engine for radiation therapy using deep learning *Phys. Med. Biol.* **65** 75013
Lee H, Kim H, Kwak J, Kim Y S, Lee S W, Cho S and Cho B 2019 Fluence-map generation for prostate intensity-modulated radiotherapy planning using a deep-neural-network *Sci. Rep.* **9** 15671
Low D A, Harms W B, Mutic S and Purdy J A 1998 A technique for the quantitative evaluation of dose distributions *Med. Phys.* **25** 656–61
Ma J, Beltran C, Wan Chan Tseung H S and Herman M G 2014 A gpu-accelerated and monte carlo-based intensity modulated proton therapy optimization system *Med. Phys.* **41** 12
Meyer P, Noblet V, Mazzara C and Lallement A 2018 Survey on deep learning for radiotherapy *Comput. Biol. Med.* **98** 126–46
Neishabouri A, Wahl N, Mairani A, Köthe U and Bangert M 2021 Long short-term memory networks for proton dose calculation in highly heterogeneous tissues *Med. Phys.* **48** 1893–908

Neph R, Lyu Q, Huang Y, Yang Y M and Sheng K 2021 Deepmc: a deep learning method for efficient monte carlo beamlet dose calculation by predictive denoising in magnetic resonance-guided radiotherapy *Phys. Med. Biol.* **66** 35022
Nguyen D, Jia X, Sher D, Lin M-H, Iqbal Z, Liu H and Jiang S 2019 3d radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected u-net deep learning architecture *Phys. Med. Biol.* **64** 65020
Nguyen D, Long T, Jia X, Lu W, Gu X, Iqbal Z and Jiang S 2019 A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning *Sci. Rep.* **9** 1076
Nomura Y, Wang J, Shirato H, Shimizu S and Xing L 2020 Fast spot-scanning proton dose calculation method with uncertainty quantification using a three-dimensional convolutional neural network *Phys. Med. Biol.* **65** 215007
Pastor-Serrano O, Habraken S, Lathouwers D, Hoogeman M, Schaart D and Perkó Z 2021 How should we model and evaluate breathing interplay effects in IMPT? *Physics in Medicine & Biology* vol 66 (Bristol: IOP Publishing) pp 235003–235003
Pastor-Serrano O and Perkó Z 2021 Learning the physics of particle transport via transformers arxiv:2109.03951 [cs.LG]
Peng Z, Shan H, Liu T, Pei X, Wang G and Xu X G 2019 Mcdnet—a denoising convolutional neural network to accelerate monte carlo radiation transport simulations: a proof of principle with patient dose from x-ray ct imaging *IEEE Access* **7** 76680–9
Peng Z, Shan H, Liu T, Pei X, Zhou J, Wang G and Xu X G 2019 Deep learning for accelerating monte carlo radiation transport simulation in intensity-modulated radiation therapy arXiv:1910.07735
Pepin M D, Tryggestad E, Wan Chan Tseung H S, Johnson J E, Herman M G and Beltran C 2018 A monte-carlo-based and gpu-accelerated 4d-dose calculator for a pencil beam scanning proton therapy system *Med. Phys.* **45** 5293–304
Pereira G C, Traughber M and Muzic R F 2014 The role of imaging in radiation therapy planning: past, present, and future *BioMed Res. Int.* **2014** 1–9
Perkó Z, van der Voort S R, van de Water S, Hartman C M H, Hoogeman M and Lathouwers D 2016 Fast and accurate sensitivity analysis of IMPT treatment plans using Polynomial Chaos Expansion *Phys. Med. Biol.* **61** 4646–64
Qin N, Botas P, Giantsoudi D, Schuemann J, Tian Z, Jiang S B, Paganetti H and Jia X 2016 Recent developments and comprehensive evaluations of a gpu-based monte carlo package for proton therapy *Phys. Med. Biol.* **61** 7347–62
Ramachandran P, Bello I, Parmar N, Levskaya A, Vaswani A and Shlens J 2019 Stand-alone self-attention in vision models arXiv:1906.05909
Rojo-Santiago J, Habraken S J M, Lathouwers D, Romero A M, Perkó Z and Hoogeman M S 2021 Accurate assessment of a Dutch practical robustness evaluation protocol in clinical PT with pencil beam scanning for neurological tumors *Radiother. Oncol.* **163** 121–7
Ronneberger O, Fischer P and Brox T 2015 U-net: convolutional Networks for Biomedical Image Segmentation arXiv:1505.04597
Saini J, Maes D, Egan A, Bowen S R, James S 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
Schaffner B, Pedroni E and Lomax A 1999 Dose calculation models for proton treatment planning using a dynamic beam delivery system: an attempt to include density heterogeneity effects in the analytical dose calculation *Phys. Med. Biol.* **44** 27–41
Schreuder A, Bridges D, Rigsby L, Blakey M, Janson M, Hedrick S and Wilkinson J 2019 Validation of the RayStation Monte Carlo dose calculation algorithm using realistic animal tissue phantoms *J. Appl. Clin. Med. Phys.* **20** 160–71
Schreuder A, Bridges D, Rigsby L, Blakey M, Janson M, Hedrick S and Wilkinson J 2019 Validation of the RayStation Monte Carlo dose calculation algorithm using a realistic lung phantom *J. Appl. Clin. Med. Phys.* **20** 127–37
Schuemann J, Giantsoudi D, Grassberger C, Moteabbed M, Min C H and Paganetti H 2015 Assessing the clinical impact of approximations in analytical dose calculations for proton therapy *Int. J. Radiat. Oncol. Biol. Phys.* **92** 1157–64
Souris K, Lee J A and Sterpin E 2016 Fast multipurpose monte carlo simulation for proton therapy using multi- and many-core cpu architectures *Med. Phys.* **43** 1700–12
Srivastava N, Hinton G, Krizhevsky A and Salakhutdinov R 2014 Dropout: a simple way to prevent neural networks from overfitting *J. Mach. Learn. Res.* **15** 1929–58
Taylor P A, Kry S F and Followill D S 2017 Pencil beam algorithms are unsuitable for proton dose calculations in lung *Int. J. Radiat. Oncol. Biol. Phys.* **99** 750–6
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** 1107
Touvron H, Cord M, Douze M, Massa F, Sablayrolles A and Jégou H 2020 Training data-efficient image transformers & distillation through attention arXiv:2012.12877
Tsekas G, Bol G H, Raaymakers B W and Kontaxis C 2021 Deepdose: a robust deep learning-based dose engine for abdominal tumours in a 1.5 t mri radiotherapy system *Phys. Med. Biol.* **66** 65017
van der Voort S, van de Water S, Perkó Z, Heijmen B, Lathouwers D and Hoogeman M 2016 Robustness recipes for minimax robust optimization in intensity modulated proton therapy for oropharyngeal cancer patients *Int. J. Radiat. Oncol.**Biol.*Phys.* **95** 163–70
Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I 2017 Attention is all you need arXiv:1706.03762
Wan Chan Tseung H S, Ma J and Beltran C 2015 A fast gpu-based monte carlo simulation of proton transport with detailed modeling of nonelastic interactions *Med. Phys.* **42** 2967–78
Wang W *et al* 2020 Fluence map prediction using deep learning models—direct plan generation for pancreas stereotactic body radiation therapy *Front. Artif. Intell.* **3** 1–10
Wang Y *et al* 2016 A gpu-accelerated monte carlo dose calculation platform and its application toward validating an mri-guided radiation therapy beam model *Med. Phys.* **43** 4040–52
Wieser H P *et al* 2017 Development of the open-source dose calculation and optimization toolkit matrad *Med. Phys.* **44** 2556–68
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** 15017
Wu Y and He K 2020 Group normalization *Int. J. Comput. Vis.* **128** 742–55
Xing Y, Nguyen D, Lu W, Yang M and Jiang S 2020 Technical note: a feasibility study on deep learning-based radiotherapy dose calculation *Med. Phys.* **47** 753–8
Xing Y, Zhang Y, Nguyen D, Lin M-H, Lu W and Jiang S 2020 Boosting radiotherapy dose calculation accuracy with deep learning *J. Appl. Clin. Med. Phys.* **21** 149–59
Xiong R, Yang Y, He D, Zheng K, Zheng S, Xing C, Zhang H, Lan Y, Wang L and Liu T Y 2020 On layer normalization in the transformer architecture arXiv:2002.04745 Published on ICML20
You Y, Li J, Reddi S, Hseu J, Kumar S, Bhojanapalli S, Song X, Demmel J, Keutzer K and Hsieh C-J 2019 Large batch optimization for deep learning: training bert in 76 minutes arXiv:1904.00962
Zhu J, Liu X and Chen L 2020 A preliminary study of a photon dose calculation algorithm using a convolutional neural network *Phys. Med. Biol.* **65** 20NT02