Sub-second photon dose prediction via transformer neural networks

Pastor-Serrano, Oscar; Dong, Peng; Huang, Charles; Xing, Lei; Perkó, Zoltán

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Sub-second photon dose prediction via transformer neural networks

Oscar Pastor-Serrano[1,2] | **Peng Dong**[2] | **Charles Huang**[3] | **Lei Xing**[2] | **Zoltán Perkó**[1]

[1]Department of Radiation Science & Technology, Delft University of Technology, Delft, Netherlands

[2]Department of Radiation Oncology, Stanford University, Stanford, California, USA

[3]Department of Bioengineering, Stanford University, Stanford, California, USA

Correspondence
Oscar Pastor-Serrano, Delft University of Technology, Department of Radiation Science, and Technology, Delft, Netherlands.
Email: o.pastorserrano@tudelft.nl

Lei Xing and Zoltán Perkó authors share the senior authorship.

Abstract

Background: Fast dose calculation is critical for online and real-time adaptive therapy workflows. While modern physics-based dose algorithms must compromise accuracy to achieve low computation times, deep learning models can potentially perform dose prediction tasks with both high fidelity and speed.

Purpose: We present a deep learning algorithm that, exploiting synergies between transformer and convolutional layers, accurately predicts broad photon beam dose distributions in few milliseconds.

Methods: The proposed improved Dose Transformer Algorithm (iDoTA) maps arbitrary patient geometries and beam information (in the form of a 3D projected shape resulting from a simple ray tracing calculation) to their corresponding 3D dose distribution. Treating the 3D CT input and dose output volumes as a sequence of 2D slices along the direction of the photon beam, iDoTA solves the dose prediction task as sequence modeling. The proposed model combines a Transformer backbone routing long-range information between all elements in the sequence, with a series of 3D convolutions extracting local features of the data. We train iDoTA on a dataset of 1700 beam dose distributions, using 11 clinical volumetric modulated arc therapy (VMAT) plans (from prostate, lung, and head and neck cancer patients with 194–354 beams per plan) to assess its accuracy and speed.

Results: iDoTA predicts individual photon beams in ≈ 50 ms with a high gamma pass rate of $97.72 \pm 1.93\%$ (2 mm, 2%). Furthermore, estimating full VMAT dose distributions in 6–12 s, iDoTA achieves state-of-the-art performance with a $99.51 \pm 0.66\%$ (2 mm, 2%) pass rate and an average relative dose error of $0.75 \pm 0.36\%$.

Conclusions: Offering the millisecond speed prediction per beam angle needed in online and real-time adaptive treatments, iDoTA represents a new state of the art in data-driven photon dose calculation. The proposed model can massively speed-up current photon workflows, reducing calculation times from few minutes to just a few seconds.

KEYWORDS
deep learning, dose calculation, transformer

1 | INTRODUCTION

Modern radiotherapy techniques such as intensity modulated radiation therapy (IMRT) or volumetric modulated arc therapy (VMAT) critically rely on accurate and fast calculations of the radiation dose delivered within the patient by photon beams, typically shaped by multi-leaf collimators (MLC).[1] With modern workflows moving towards online or real-time adaptation, fast dose calculations are critical for quick plan evaluation, re-optimization and finally being able to account for motion due to breathing or anatomical changes.

Commercial treatment planning systems mainly use pencil beam (PB),[2] collapsed cone (CC),[3,4] or Monte Carlo (MC) dose engines. While both PB and CC algorithms are usually faster than MC, the assumptions and approximations they use to solve photon particle transport result in less accurate results. Conversely, MC methods—the gold standard in dose calculation—simulate individual stochastic particle trajectories abiding the physical laws of nuclear interactions and track the deposited dose along these paths. By averaging results from enough particles (typically several millions), MC methods achieve very high accuracy even in the most complex patient geometries, at the cost of high computation times. Current commercial treatment planning systems mainly use improved PB or CC variations yielding close-to-MC accuracy, for example, the anisotropic analytical algorithm (AAA)[5,6] based on the PB convolution[2] in Eclipse (Varian Medical Systems) or the CC convolution algorithm in Pinnacle (Philips).[3] Some recent MC implementations also use the parallelization capabilities of graphics processing units (GPUs) to reduce dose calculation times from several hours to minutes.[7–9] Despite these advances, the need for accurate and fast dose calculation algorithms is still unmet in most clinical workflows, as neither PB nor MC are fast enough for real-time treatment plan correction.

Recently, deep learning models have been applied to several steps of the radiotherapy workflow,[10] mainly as U-net convolutional architectures[11] or Generative Adversarial Networks.[12] Most works aim to aid treatment planning by predicting clinically optimal doses based on historical data. As a result, they are constrained to a specific site, clinical optimum choice, and often fixed beam configurations, limiting their generalization capabilities. These models typically directly predict the full dose distribution using computed tomography (CT) images,[13] organ masks,[14–18] or additional information about the photon beam configuration[19] as input. To further aid treatment planning, few studies additionally provide the beam intensities needed to deliver the predicted dose distribution.[20,21]

Aiming at predicting dose distributions in generic setups, several subsequent studies present dose calculation models that estimate beam or full dose distributions from CTs and additional physics input such as high noise MC[22–24] or PB doses[25,26]; fluence maps, for example, resulting from simple ray tracing calculations;[27,28] energy released per unit mass[29]; or a combination of the previous with additional beam information.[30,31] The reason for their success are the convolutional layers that excel at capturing local features and are heavily optimized for GPU hardware, but are less appropriate for modeling long-range dependencies, for example, changes along the beam direction through the patient.

Although some of the most recent models can quickly predict dose distributions in most cases with good accuracy,[30–32] there is room for improvement with newer architectures that require less input information and can model distant features in the data. Recent transformer architectures[33] are particularly well-suited to process local and distant features, yielding excellent results in a wide range of sequence modeling tasks.[34–36] For smaller datasets, Transformers perform particularly well when combined with convolutional layers.[37] Based on these synergies between convolutions and Transformers, a recent study presented a transformer-based algorithm predicting proton beamlet 3D dose distributions as a sequence of 2D slices in the beam depth, with state-of-the-art performance and speed.[38,39]

In this study, we present a deep learning model that can predict dose distributions in few milliseconds with clinically acceptable accuracy. As in concurrent work,[40] we harness the power of hybrid transformer and 3D convolutional architectures, adapting the previous transformer-based proton dose calculation model[38] to predict the dose of much bigger photon broad beams. As shown in Figure 1, the proposed improved Dose Transformer Algorithm (iDoTA) combines a series of 3D convolutional layers modeling local dose and tissue variations, with a Transformer backbone routing information along the depth of the entire photon beam. The model treats input 3D CT and projected shape volumes (containing beam geometrical information) as a sequence of 2D slices in the direction of the beam, framing dose calculation as sequence modeling to produce a sequence of 2D dose slices forming the 3D dose distribution. After comparing iDoTA to the best-performing data-driven models, we demonstrate its superior speed and accuracy for photon dose calculation tasks, being capable to speed up beam prediction times down to few milliseconds and reducing treatment plan computation times to few seconds.

2 | METHODS AND MATERIALS

In this section, we present the problem setup and architecture of the iDoTA model, used to predict photon beam doses from 3D CT and projected shape inputs. Additionally, we describe the dataset and training procedure used to optimize the model parameters, as well as the
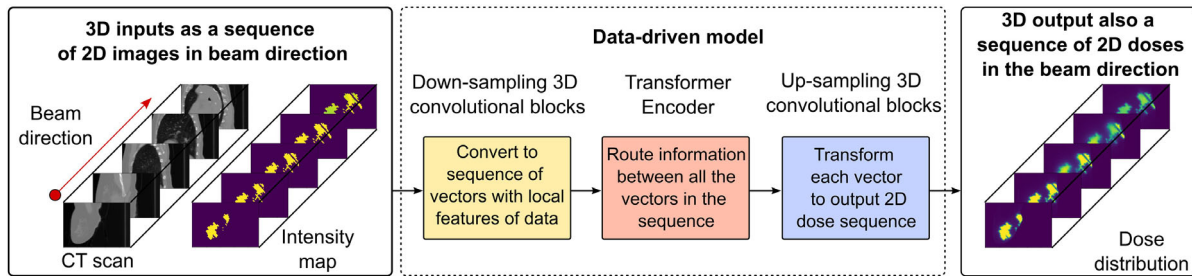
FIGURE 1 **Model overview**. A deep learning data-driven model learns the mapping $y = f_\theta(x, r)$ between input 3D CT x and projected shape r volumes, and the corresponding output 3D dose distributions y. The problem is formulated as a sequence prediction task, where all input and output cubes are treated as a sequence of 2D slices in the beam's eye view. Each 2D slice is mapped into a vector via a series of down-sampling convolutional blocks. A transformer backbone routes information between all elements of the resulting sequence. Finally, a several up-sampling convolutional operations transform each vector into a 2D dose distribution map.
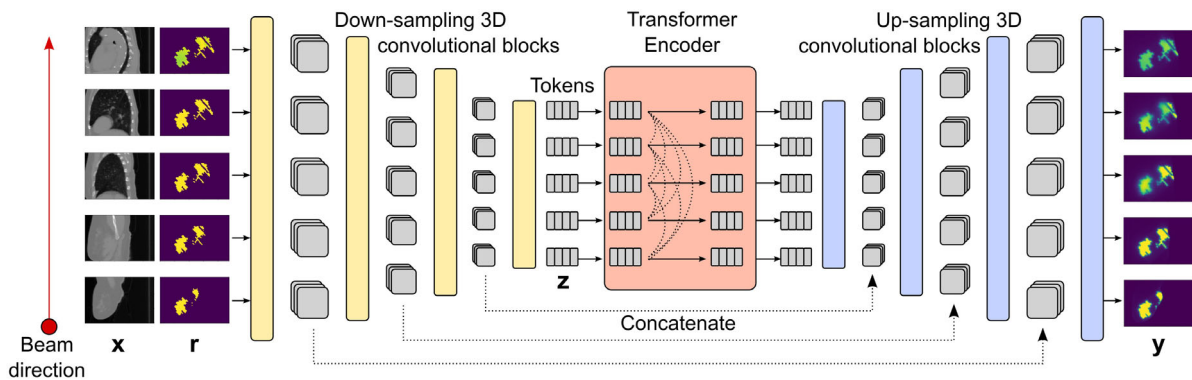

FIGURE 2 **Model architecture**. The proposed model solves the dose prediction task as sequence modeling, mapping two input sequences of 2D CT slices x and projected shapes r with beam shape information into a sequence of 2D dose distributions y. First, a series of down-sampling convolutional blocks merges and compresses the two sequences from the data into a sequence of feature vectors z (referred to as tokens). A Transformer encoder with causal self-attention routes long-range dependencies along the beam direction. Finally, a series of up-sampling convolutional blocks transform the output tokens into a sequence of 2D dose distributions. In each block, the exact same 3D convolution operation is applied to all sequence elements, extracting local features from the preceding and following element in the sequence.

evaluation metrics used to assess iDoTA's performance as a generic photon dose calculation engine.

2.1 | Proposed framework

Photon dose calculation involves estimating the radiation dose delivered in the patient geometry. If the machine parameters do not change, the predicted dose distribution mainly depends on the irradiated geometry and the beam geometrical information such as the MLC aperture shape, the beam angle and the relative position of the isocenter. We assume that all the necessary beam shape information is captured in a 3D projected shape $r \in \mathbb{R}^{D \times H \times W}$ of depth D, height H and width W, containing the result of a simple ray tracing operation propagating the photon beam shape through the patient geometry CT scan $x \in \mathbb{R}^{D \times H \times W}$. The outcome of the dose calculation operation predicted by our model is another grid $y \in \mathbb{R}^{D \times H \times W}$ with the 3D distribution of dose per monitor unit (MU).

As shown in Figure 2, the patient CT x and the 3D projected shape r are inputs to iDoTA, which during training

implicitly learns the mapping $y = f_\theta(x, r)$ via a cascade of neural networks layers with parameters θ. Framing the dose prediction task as modeling a sequence of D elements in the direction of the photon beam, we combine the strengths of both convolutional and transformer architectures into a single model. The input geometry x can be expressed a sequence of D images in the direction of the beam $\{x_i | x_i \in \mathbb{R}^{1 \times H \times W}, \forall i = 1, \ldots, D\}$, while the projected shape 3D input r is similarly viewed as a sequence 2D slices $\{r_i | r_i \in \mathbb{R}^{1 \times H \times W}, \forall i = 1, \ldots, D\}$ containing beam information. Likewise, the final dose volume y is also expressed as the sequence of 2D dose slices $\{y_i | y_i \in \mathbb{R}^{1 \times H \times W}, \forall i = 1, \ldots, D\}$).

2.2 | Model architecture

As seen in Figure 2, the proposed architecture combines a series of convolutional blocks modeling local features with a transformer backbone that processes information along the entire beam depth.

- First, a series of down-sampling convolutional blocks extract local features of the data into a sequence

of vectors $\{\boldsymbol{z}_i | \boldsymbol{z}_i \in \mathbb{R}^N, \forall i = 1, \dots, D\}$—referred to as tokens in the remainder of the paper—of size N. Each block contains a 3D convolutional layer with kernel size equal to 3, modeling local features from the immediately preceding and succeeding elements in the sequence, followed by a layer normalization,[41] a rectified linear unit (ReLU) activation function and a max-pooling operation. All such operations in the block are applied in parallel to every element of the input sequence. Due to the max-pooling operation, the height H and width W of the slices are halved after each block. A total of L blocks result in L resolution levels. After the last block, we apply a final convolution with K filters and flatten the resulting features into tokens of dimension $N = \left(\frac{H}{2}\right)^L \times \left(\frac{W}{2}\right)^L \times K$. As a result, we obtain a sequence of D tokens containing local features about the corresponding input slices, e.g., the third token $\boldsymbol{z}_3$ represents local features from the inputs $\boldsymbol{x}_3, \boldsymbol{r}_3$ and their neighboring slices.

- A Transformer backbone routes information between the extracted features along the depth of the entire volume, with the self-attention mechanism[33] making the information exchange dynamic, that is, each token $\boldsymbol{z}_i$ is independently transformed based on its content and information selectively gathered from other sequence elements. To account for the relative distance between tokens, we add a learnable positional embedding $\boldsymbol{p}_i \in \mathbb{R}^N$ to each token $\boldsymbol{z}_i$, that is, a sequence of vectors $\{\boldsymbol{p}_i | \boldsymbol{p}_i \in \mathbb{R}^N, \forall i = 1, \dots, D\}$ is learned and added to the token sequence before the first operation in the Transformer. We use the pre-Layer Normalization architecture,[42] which consists of a Layer Normalization (LN)[41] operation, followed by a self-attention operation,[33] and two fully-connected layers with Dropout[43] and a Gaussian Error Linear Unit (GeLU) activation.[44]

- Finally, a series of L up-sampling convolutional blocks convert the token sequence into the output dose volume. For each level, we append (along the feature dimension) the sequence previously obtained from the same level down-sampling convolutional block, similar to U-net type architectures. The up-sampling block's architecture is identical to that of its down-sampling counterpart, except for the use of a nearest-neighbor up-sampling interpolation operation instead of the max-pooling.

2.3 | Projected shape and dose calculation

Apart from the values in the CT, the additional 3D projected shape input $\boldsymbol{r}$ encodes beam information such as the MLC aperture shape, the angle or relative distance between the isocenter and the source, including basic material information with a simple correction based on tissue densities. Such projected shape is generated via an algorithm that estimates the dose at each voxel through the percentage depth dose (PDD), corrected by an off-axis factor. The PDD is measured at 100 cm source-to-surface distance (SSD) with a 10 cm × 10 cm field size, adjusting for different SSDs using the Mayneord factor. The depth for determining the percentage dose is the water equivalent distance, calculated via ray tracing for all voxels. The off-axis correction factor is calculated by sampling from a diagonal beam profile for a 40 cm × 40 cm field size at 10 cm depth, projecting it to different depths using the lateral distance of the voxel to the center beam axis and the longitudinal distance from the voxel to the source. This ray tracing calculation estimates the dose using the commissioning data and is optimized for speed over accuracy, taking around 0.1 ms per beam in a GPU. The corresponding ground truth dose distributions (to be predicted by the model) are obtained via the AcurosXB V15.6.05 algorithm in the Varian Eclipse TPS system (with the option of calculating dose to medium). Both the dose and the projected shapes have similar ranges from 0 to ≈ 3, with units cGy/MU.

2.4 | Dataset

iDoTA is trained to predict individual photon beams using a training dataset of 17 clinical patient CTs with disease sites of brain, head neck, lung, abdomen and pelvis. All CTs were recorded using a General Electrics LightSpeed CT scanner with 2.5 mm × 2.5 mm × 2.5 mm resolution. For each patient, 100 different co-planar photon beams were computed, using, for each beam, a random gantry angle and an isocenter location randomly selected within the patient, and an aperture shape that was generated by randomly sampling leaf positions, keeping the couch angle fixed. After calculating the dose per MU and projecting the aperture shape, we cropped 3D CT $\boldsymbol{x} \in \mathbb{R}^{96 \times 96 \times 64}$, projected shape $\boldsymbol{r} \in \mathbb{R}^{96 \times 96 \times 64}$ and dose $\boldsymbol{y} \in \mathbb{R}^{96 \times 96 \times 64}$ blocks covering a volume of approximately $240 \times 240 \times 160$ mm^3, so that the beam always travels in the same direction along the first dimension $D = 96$ with angles between -45° and 45°. All 1700 input CT volumes are normalized to the range [0,1] dividing by using the maximum value of 3071 observed across the entire dataset. Likewise, we normalize both projected shapes and dose distributions using the maximum dose value of 3.075 cGy/MU in the dataset. During training, 10% of the samples are set aside for validation purposes, that is, finding the best model configuration.

We evaluate the best model using an independent test dataset of 584 beam dose distributions corresponding to a prostate and a lung patient unseen during training. Additionally, to assess iDoTA's performance in predicting full dose distributions composed of many photon beams, we obtain 11 additional clinical VMAT treatment

plans with 2 arcs and 99–178 control points per arc, corresponding to one brain, three HN, three lung, and four prostate cancer patients.

2.5 | Training details

We train iDoTA using the mean squared error as a loss function, with mini-batches of 4 samples and the layer-adaptive LAMB optimizer,[45] finding the combination of a low batch size and the LAMB optimizer to be critical for convergence. During training, we augment the dataset size via rotations (in steps of $90°$, perpendicular to the direction of the beam) and random shifts along the beam direction (shifting the entire volume up to 15 positions along the first dimension). Training consists of 10 cycles with 120 epochs/cycle, where the learning rate is set to 10^{-3} at the beginning of each cycle, and halved every 15 epochs.

After hyper-parameter tuning using the validation data, the best-performing model has $H = 4$ transformer heads, $L = 4$ levels with $K = 10$ filters in the last encoder convolution. The four down-sampling operations in the encoder transform the input slices with dimensions $H = 96$ and $W = 64$ into tokens of size $N = H/2^4 \times W/2^4 \times K = 240$. All training and experiments are run in a Nvidia A40® GPU using Tensorflow.[46]

2.6 | Evaluation metrics

For evaluation purposes, we compare iDoTA's predictions to the corresponding ground truth dose distributions in the independent test set of patients unseen during training. The main method to assess dosimetric differences is the gamma analysis,[47] based on the intuition that two neighboring voxels with a similar dose result in equivalent biological effects. Intuitively, a voxel in the predicted dose distribution passes the gamma evaluation $\Gamma(\delta \text{ mm}, \Delta\%)$ if another voxel with a similar value—deviating less than $\Delta\%$ of the maximum dose—is found within a sphere of radius δ mm in the ground truth dose grid. We compute three gamma evaluations $\Gamma(1 \text{ mm}, 1\%)$, $\Gamma(2 \text{ mm}, 2\%)$ and $\Gamma(3 \text{ mm}, 3\%)$, and calculate the gamma passing rate by dividing the number of passed voxels by the total amount of eligible voxels, that is, voxels with values within 10% and 100% of the maximum dose.

As an additional metric to measure explicit voxel dose differences, we compute the average relative error ρ, expressed as a percentage of the maximum dose in the grid. As for the gamma pass rate, the average relative error is calculated only for voxels with values within 10% and 100% of the maximum dose. For model predictions $\boldsymbol{y}$, and corresponding ground truth 3D dose distributions $\hat{\boldsymbol{y}}$ (both with $n_v = D \times H \times W$ voxels), the average

relative error is calculated using the L_1-norm as

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

3 | RESULTS

To assess iDoTA's suitability as a generic photon dose calculation tool and determine its improvements with respect to other data-driven algorithms, we compute the different evaluation metrics on the independent test data. In particular, we compare iDoTA's accuracy and speed to previous approaches when predicting both individual photon beam prediction and full dose distributions from clinical VMAT plans.

3.1 | Individual beams

We compute the $\Gamma(1 \text{ mm}, 1\%)$, $\Gamma(2 \text{ mm}, 2\%)$ and $\Gamma(3 \text{ mm}, 3\%)$ gamma pass rate and the error ρ for the 584 beams in the test dataset. In Table 1 we compare the mean, standard deviation and minimum values to those reported in previous studies achieving state-of-the-art performance, that is, the convolutional architectures for photon dose prediction in standard linear accelerator (Linac)[30] and MR-Linac settings.[31] Additionally, we train and evaluate a purely convolutional version of iDoTA without the transformer encoder, referred to as iDoTA-conv. The overall lower pass rates achieved by iDoTA-conv demonstrate the added benefit of combining transformers and convolutions. In general, iDoTA achieves better pass rates than previous convolutional models, with higher means and smaller standard deviations. Most importantly, the minimum gamma pass rate across all test samples is >20% higher than that of the 3D-U-net based architectures.

iDoTA can better predict photon beams in pelvic anatomies than in lung scans, which is likely caused by the more heterogeneous nature of lung geometries (i.e., the contrasts between bony structures, air, and water-like tissues). Figure 3 further confirms iDoTA's superiority for the pelvic cases over lung, showing $\Gamma(1 \text{ mm}, 1\%)$, $\Gamma(2 \text{ mm}, 2\%)$, and ρ distributions with lower lung pass rates and higher errors. Figure 4 visually compares the target and predicted beam dose distributions for the worst-performing lung and pelvic samples, and an average-performing pelvic beam. The overall errors are low and mostly occur at the beam lateral falloff, which may be caused by the coarse resolution of the input projected shapes. Since the average relative error in test data of 2.18% is similar to the final error in validation data of $2.19 \pm 1.08\%$, and relatively close to the error for training data of $1.54 \pm 0.64\%$, we conclude that the model generalizes well.

TABLE 1 Model accuracy for individual broad beams.

Treatment site	Model	$\Gamma(1,1)$ Mean $\pm$ std	Min	$\Gamma(2,2)$ Mean $\pm$ std	Min	$\Gamma(3,3)$ Mean $\pm$ std	Min
Pelvic	3D-Unet[30]	**89.9 $\pm$ 5.1**	44.5	97.8 $\pm$ 3.0	55.2	99.4 $\pm$ 2.5	62.5
	3D-Unet[31]	87.6 $\pm$ 8.3	47.5	97.9 $\pm$ 2.6	68.2	99.5 $\pm$ 1.0	77.5
	iDoTA - conv	85.8 $\pm$ 8.6	32.5	97.0 $\pm$ 4.6	52.8	99.2 $\pm$ 2.1	76.2
	iDoTA (ours)	89.0 $\pm$ 5.4	**66.9**	**98.1 $\pm$ 1.7**	**87.7**	**99.6 $\pm$ 0.5**	**94.7**
Lung	iDoTA - conv	**84.3 $\pm$ 4.1**	65.5	95.6 $\pm$ 2.0	86.9	98.8 $\pm$ 0.8	94
	iDoTA (ours)	84.1 $\pm$ 4.7	**68.9**	**96.9 $\pm$ 2.0**	**90.1**	**99.2 $\pm$ 0.8**	**94.2**

Gamma pass rates for photon beams are computed using three different criteria in the gamma evaluation. The reported values, which include the mean, standard deviation (std), and minimum across all test samples from pelvic and lung cancer patients, are compared to other state-of-the-art deep learning models as reported in their respective studies. To determine the added benefit of using transformers, a purely convolutional variant of iDoTA—without the transformer encoder, denoted as iDoTA-conv—is trained and evaluated using the same training procedure and dataset.
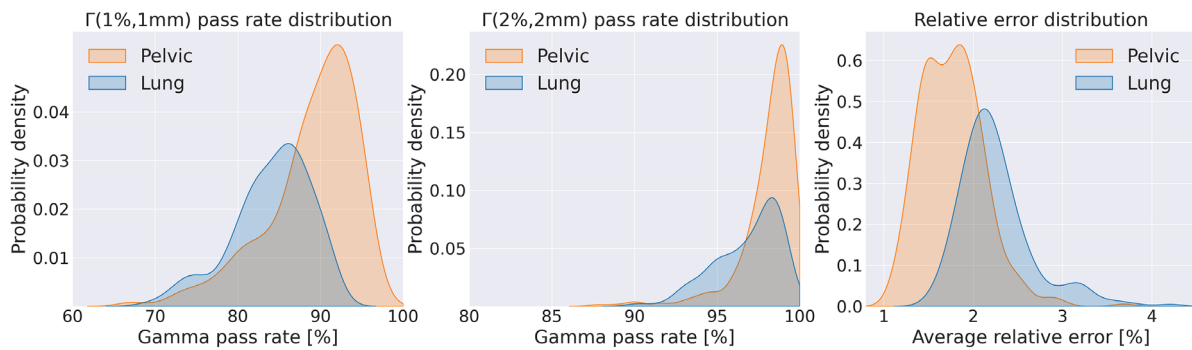

FIGURE 3 **Accuracy metrics distribution**. (Left)$\Gamma(1\,mm, 1\%)$. pass rate, (middle) $\Gamma(2\,mm, 2\%)$ pass rate and (right) average relative error distributions across all beams in the test dataset. The lower errors and higher pass rate values in orange correspond to beams in the pelvic area, while blue distributions are from lung samples.

3.2 | Full dose distributions

For 11 additional patients outside the training dataset with clinical VMAT plans available, we compare the $\Gamma(1\,mm, 1\%)$, $\Gamma(2\,mm, 2\%)$ and $\Gamma(3\,mm, 3\%)$ gamma pass rate to the values reported in previous studies. In particular, iDoTA's accuracy and inference times are compared to those of: convolutional U-net architectures predicting each beam in the plan individually;[30–32] convolutional models de-noising MC dose distributions;[23,24] and a concurrent 3D U-net and transformer model for MR-Linac dose prediction.[40]

Table 2 shows the mean and standard deviation of the gamma pass rates separately for pelvic, lung and HN patients, comparing them to other models. With a $99.51 \pm 0.66\%$ $(2\,mm, 2\%)$ pass rate, an average relative dose error of $0.75 \pm 0.36\%$ across all patients, and higher pass rates in all treatment sites, iDoTA outperforms all previous approaches. Additionally, the average error ρ in HN, lung and pelvic plans is 1.11%, 0.64%, and 0.45%, respectively. For the remaining patient with a brain tumor, a $\Gamma(1\,mm, 1\%)$, $\Gamma(2\,mm, 2\%)$ and $\Gamma(3\,mm, 3\%)$ gamma pass rate of 93.5, 99.7, and 99.9, respectively. As seen in the individual beams, iDoTA is more accurate in pelvic cases and less pre-

cise in HN anatomies, which is also likely due to the bone, water and air (cavities) heterogeneities. Nevertheless, the overall pass rate is still significantly higher than other approaches. Figure 5 shows very similar reference and predicted dose distributions for a prostate and lung VMAT plan, along with the corresponding $\Gamma(2\,mm, 2\%)$ map with mostly all voxels passing the gamma evaluation. To further evaluate the similarity between ground truth and predicted doses, Figure 6 shows dose volume histograms (DVHs) from three test patients. The almost perfectly overlapping DVH lines indicate that iDoTA's predictions are practically identical to the reference data. Finally, Figure 7 visually compares the predicted and ground truth dose distributions in heterogeneous lung and head and heck geometries. The similarity between both distributions and the $\Gamma(2\,mm, 2\%)$ maps with almost all passing voxels further demonstrate iDoTA's accuracy even in the presence of sharp tissue-air and tissue-bone contrasts.

3.3 | Prediction times

Computation speed is critically important in adaptive workflows. In Table 3, we compare iDoTA's total time
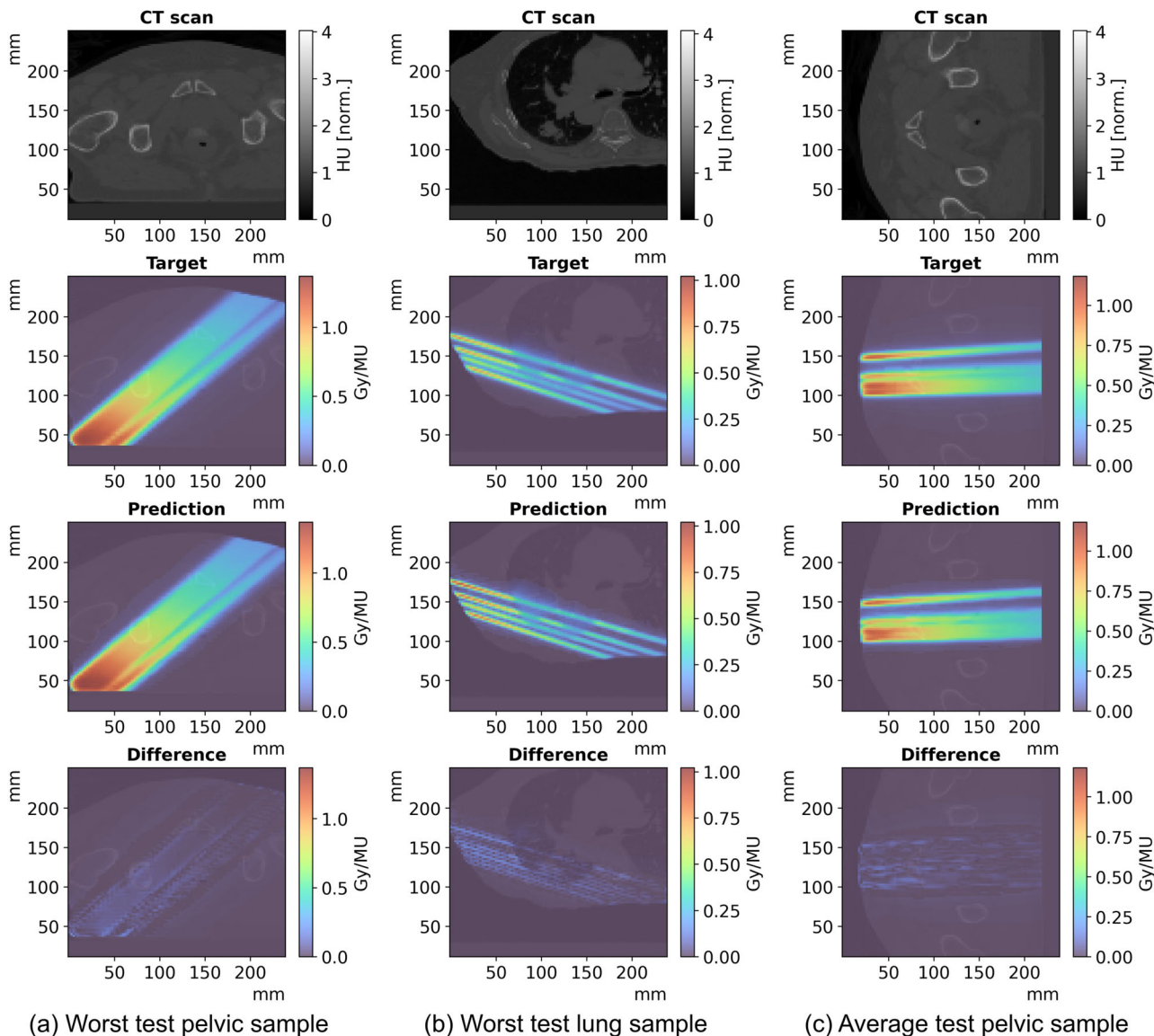
FIGURE 4 **Individual beam test samples**. (a) Worst performing pelvic test sample in the gamma evaluation, with $\Gamma(2\,mm, 2\%)$ gamma pass rate of 87.7%; (b) worst performing prediction in the gamma evaluation across the lung test samples, with $\Gamma(2\,mm, 2\%)$ gamma pass rate of 90.1%, and (c) average performing sample. Given the $96 \times 96 \times 64$ 3D volumes—a sequence of 96 2D slices of dimension 96×64 —all plots show the central slice along the beam direction, that is, slice 32 out of 64. From top to bottom rows, the subplots show the 3D input CT grid, the reference dose distribution, the model's prediction and the dose difference between the predicted and reference beams.

needed to predict individual beams and full plans to the reported values for models in previous studies. All prediction times for all models include the time needed to generate and prepare the inputs, predict the output and (for full dose distributions) accumulate beam doses. For individual beam prediction, iDoTA is significantly faster than any other competitor, being 30–60x faster than the 3D U-net models and 6x faster than the concurrent transformer model TransDose.[40] Likewise, iDoTA predicts full dose distribution from VMAT plans (with 194–354 beams per plan) on average in 8 s, representing a 10–80x speed-up compared to the IMRT (with $\approx$ 10 beams) U-net models. With CPU settings (intel® Core™ i7-8550U 1.8 GHz), iDoTA still remains competitive with previous GPU-based models, predicting beam doses in 1.48 ± 0.13 seconds and full plans in 300 to 600 s, depending on the number of beams.

4 | DISCUSSION

4.1 | Comparison to previous models

Framing photon dose calculation as sequence modeling, iDoTA is able to predict beam doses with high accuracy and speed, achieving an overall $97.72 \pm 1.93\%$

TABLE 2 Model accuracy for full clinical dose distributions.

Treatment site	Model	Voxels	$\Gamma(1,1)$	$\Gamma(2,2)$	$\Gamma(3,3)$
Head & neck	TransDose[40]	Full grid	-	96.7 ± 2.3	-
	Denoising 3D U-net[24]	Full grid	70.9 ± 2.9	89.4 ± 3.7	-
	DoTA (ours)	Full grid	$\mathbf{80.5 \pm 8.6}$	$\mathbf{98.9 \pm 0.9}$	$\mathbf{99.9 \pm 0.1}$
	DoTA (ours)	PTV only	74.6 ± 6.7	98.7 ± 0.4	99.9 ± 0.1
Pelvic	TransDose[40]	Full grid	-	97.9 ± 0.4	-
	3D U-net[30]	Full grid	89.9 ± 3.3	99.5 ± 0.7	99.9 ± 0.3
	3D U-net[31]	Full grid	82.2 ± 9.7	96.1 ± 3.1	99.4 ± 0.6
	3D U-net[32]	Full grid	84.2 ± 2.9	99.0 ± 0.4	99.9 ± 0.1
	Denoising 3D U-net[23]	Full grid	-	95.4 ± 1.6	-
	DoTA (ours)	Full grid	$\mathbf{95.8 \pm 3.1}$	$\mathbf{99.8 \pm 0.2}$	$\mathbf{99.9 \pm 0.0}$
	DoTA (ours)	PTV only	85.0 ± 4.6	99.1 ± 0.7	99.9 ± 0.1
Lung	TransDose[40]	Full grid	-	96.7 ± 1.4	-
	DoTA (ours)	Full grid	$\mathbf{94.3 \pm 1.5}$	$\mathbf{99.8 \pm 0.2}$	$\mathbf{99.8 \pm 0.1}$
	DoTA (ours)	PTV only	84.6 ± 4.6	98.9 ± 1.0	99.9 ± 0.1

For different treatment sites, we display the gamma pass rates of full photon dose distributions, using three different criteria. We include the values from few of the best-performing models as reported in their respective studies, as well as iDoTA's pass rates calculated using only PTV voxels. All pass rates include the average and standard deviation across all available dose distributions.
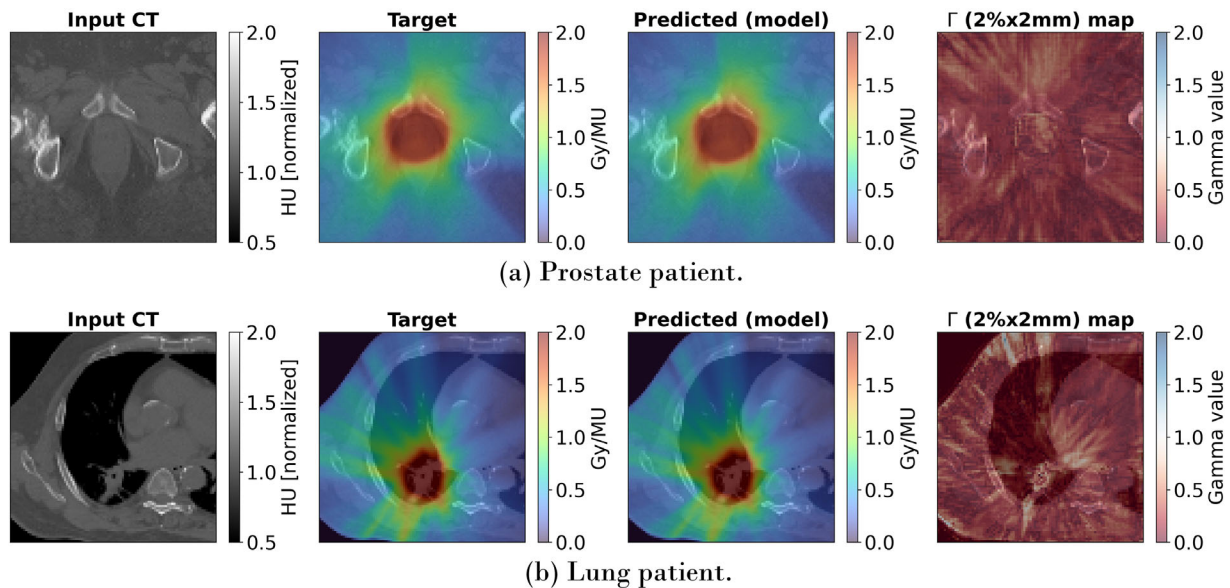

(a) Prostate patient.

(b) Lung patient.

FIGURE 5 **Dose distributions from VMAT plans**. From left to right, the input CT, target and predicted dose distributions and $\Gamma(2 \text{ mm}, 2\%)$ gamma map are shown for two clinical VMAT plans from a (a) prostate and (b) lung cancer patient. To show details of the high dose region, we display crops around the target volume.

$\Gamma(2 \text{ mm}, 2\%)$ pass rate in lung and pelvic geometries. This per-beam prediction precision translates into a very high $\Gamma(2 \text{ mm}, 2\%)$ pass rate of $99.51 \pm 0.66\%$ in dose distributions from clinical VMAT plans, which also outperforms all previous models. Compared to the best-performing convolutional models[30,31] (and even a recently published VMAT model,[32]) iDoTA offers more than 30x faster beam dose prediction even in the most heterogeneous geometries, achieving better gamma

pass rates on average with lower spread, and 20% higher pass rates in the most difficult samples. Furthermore, iDoTA only uses the 3D CT and beam intensity to predict doses, in contrast to the 5 different input volumes containing physics information required by the 3D U-nets, allowing for lower input generation times and faster calculation times overall. iDoTA also convincingly outperforms MC de-noising models,[23,24] with a 5%–10% increase in gamma pass rates and a 20–80x speed-up,
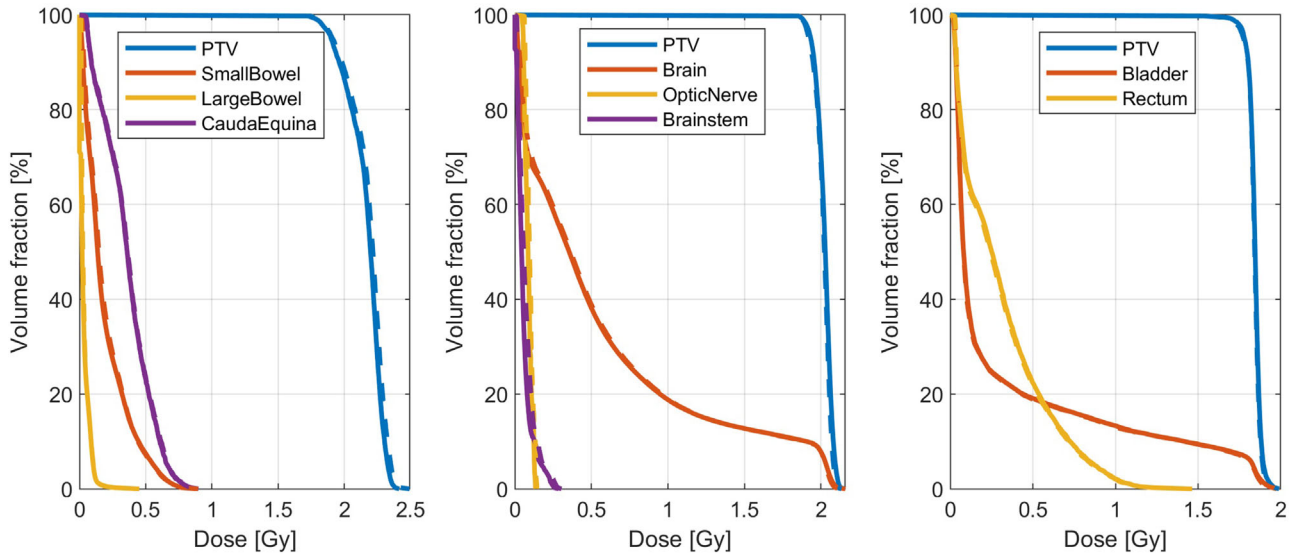
FIGURE 6   Dose volume histograms from 3 VMAT plans. Three dose volume histograms from a (left) pelvic, (mid) brain and (right) prostate test patients are shown, indicating the dose received by a specific fraction of the volume of an organ. All plots include the planning target volume (PTV) and few of the surrounding organs at risk. Solid lines represent iDoTA's predictions, while dotted lines indicate ground truth values. The numeric values included next to each line represent the per-structure $\Gamma(2\ mm, 2\%)$ gamma pass rate.
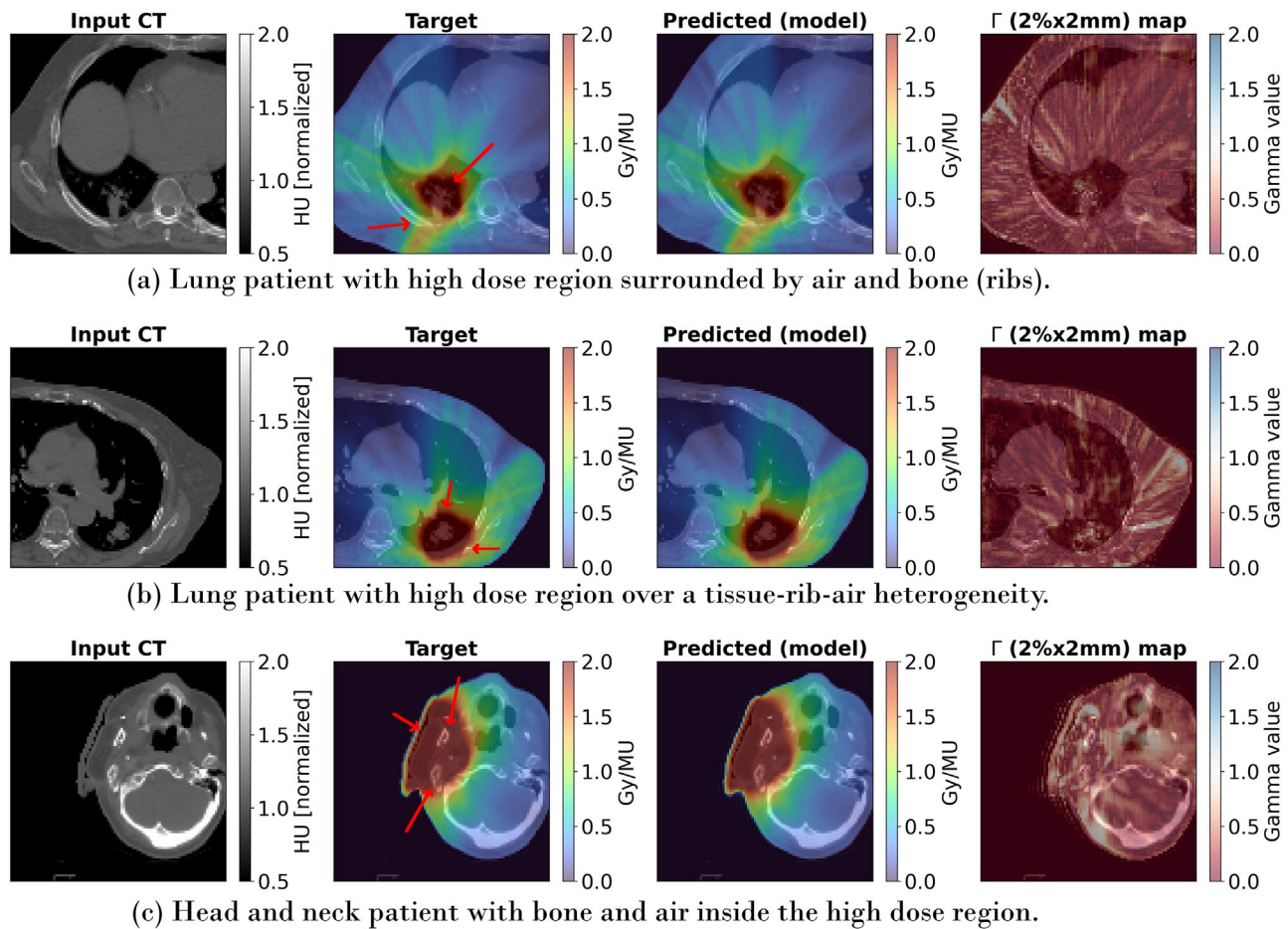

(a) Lung patient with high dose region surrounded by air and bone (ribs).

(b) Lung patient with high dose region over a tissue-rib-air heterogeneity.

(c) Head and neck patient with bone and air inside the high dose region.

FIGURE 7   Dose distributions in heterogeneous anatomies. From left to right, the input CT, target and predicted dose distributions and $\Gamma(2\ mm, 2\%)$ gamma map are shown for three clinical VMAT plans. In all cases, high dose is delivered to heterogeneous anatomies, with (a) part of the high dose region surrounded by air and ribs, (b) part of the high dose directly delivered to an area with a sharp change from lung air to rib bone, and (c) the high dose region covering air and bone heterogeneous tissue. To show details of the high dose region, we display crops around the target volume.

TABLE 3 Average prediction time.

Photon beams	Model	Hardware	Average time (ms)
	TransDose[40]	GPU	310
	3D U-net[30]	GPU	1500
	3D U-net[31]	GPU	3000
	3D U-net[32]	GPU	7000
	DoTA (ours)	CPU	1480
	DoTA (ours)	GPU	**50**
Full plans	Model	Hardware	Average time (s)
	Denoising U-net[23] (<20 beam angles)	GPU	150
	Denoising U-net[24] (<20 beam angles)	GPU	660
	3D U-net[30] (<20 beam angles)	GPU	60
	DoTA (ours, 194–354 beam angles)	CPU	450
	DoTA (ours, 194–354 beam angles)	GPU	**8**

iDoTA's computing speed is compared to the fastest models in literature via the average computing time needed to predict a photon beam or full dose distribution. The reported values include the time needed to generate and process the model inputs. iDoTA's CPU prediction times are also included for comparison, as well as the average number of beams in the evaluated treatment plans.

partially caused by the time needed to generate the high-noise MC dose inputs. In general, iDoTA achieves higher gamma pass rates than all previous convolutional models, also compared to the purely convolutional iDoTA-conv variant trained with identical dataset, training procedure and architecture (except for the transformer encoder). As in previous proton studies[39] and the concurrent TransDose,[40] these findings demonstrate that the addition of the transformer—being able to capture relationships between distant features, as opposed to convolutions—seems to be beneficial for dose prediction tasks.

Moreover, our method outperforms the concurrent TransDose transformer model in both accuracy and speed. Although TransDose is trained to predict photon beams under magnetic fields for MR-Linac applications—which could be a more difficult task to learn—we hypothesize that part of iDoTA's success is due to differences in the model, that is, that the data-demanding transformer architecture in iDoTA routes information only between each of the 96 2D slices, instead of the 5000 voxels that are input to the transformer in TransDose. As a result, iDoTA's transformer has less parameters, which can be favorable with smaller datasets and accelerates inference.

With higher accuracy and lower computing times than any other previously introduced deep learning model, the proposed iDoTA represents a new state of the art in data-driven photon dose calculation. iDoTA can predict full dose distributions in 6–10 s, including CT cropping and rotation time (≈ 25 ms per beam), ray tracing input calculation (≈ 0.1 ms per beam) loading the model and weights (≈ 2 s), inferring the beam dose distribution (≈ 20 ms per beam) and accumulating the doses in the final grid (≈ 5 ms per beam). As a result, iDoTA is an order of magnitude faster than clinically used algorithms or MC approaches adapted to GPU hardware.[7–9] While such MC-GPU implementations are several orders of magnitude faster and almost as accurate as their CPU counterparts, their total calculation times are still in the order of minutes. Furthermore, iDoTA is 20x and 60x faster than the Eclipse Acuros XB and AAA algorithms (Varian Medical Systems) used in $\approx 80\%$ of the clinics, which predict VMAT doses in 2–3 and ≈ 10 minutes, respectively.[48,49] Most importantly, the photon beams can be predicted in parallel in several batches depending on the number of GPUs and their internal memory, practically allowing for further reduction in total calculation times.

4.2 | Limitations

Like all other data-driven algorithms, iDoTA is trained to emulate dose distributions from a specific machine and settings. Deep learning algorithms have limited extrapolation capabilities outside the training domain, which would require a different model each time the machine configuration is changed (or even the CT scanner, unless different CT machines are included in the training dataset). In such cases, fine-tuning iDoTA starting from the provided weights using a smaller dataset can save time without significantly degrading performance.

Ideally, all machine characteristics would be given to the model as separate inputs. Alternatively, to account for geometrical information and machine characteristics, iDoTA requires the additional input projected shape, necessitating ray-tracing pre-calculations. As for the machine parameters, such beam information could be included in the input as separate tokens, e.g., the

aperture shape could be given as 2D binary mask at the beginning of the input sequence.

iDoTA is trained using a certain resolution and grid dimensions, which must be fixed for both training and inference. For dose prediction in finer grid resolutions, iDoTA can be coupled to neural representation models capable of accurate super-resolution.[50] Regarding grid size, predicting dose distributions from treatment plans or beams through anatomies larger than the predetermined voxel grid must be done in steps, obtaining several input volumes and accumulating the outputs along the beam depth. Conversely, all doses can be predicted for the same fixed grid covering the part of the anatomy containing the structures of interests, which neglects the (usually) low doses near patient entrance. As observed in proton dose prediction,[38,39] we expect iDoTA to perform equally well for different grid settings, with calculation times going up for larger grids and finer resolutions, but still within sub-second speed for each beam angle.

Finally, iDoTA is trained and evaluated on a dataset that differs from the ones used in previous models, which can affect the final evaluation metrics. Likewise, the high-end GPU used in our experiments may affect iDoTA's reported prediction times. Nevertheless, we do not expect our GPU to offer significant speed improvements with respect to the one from previous studies, especially if compared to the fastest alternative[40] using a modern GPU similar to ours. iDoTA's intrinsic speed is further confirmed by its competitive prediction times even when using a CPU, as shown in Table 3, which is partly due to using less parameters and a faster input generation. For future work, we recommend a comparison using a unified dataset and computational environment.

4.3 | Applicability

Conditioned only on the beam shape projection and the CT, iDoTA is a versatile algorithm that can drastically reduce computing times in any application involving repeated calculation of dose distributions, for example, checking plan robustness by quickly predicting the dose in each of the many possible error scenarios or anatomical variations of the patient.[51] Given a pre-treatment CT, iDoTA can allow fast quality assurance by comparing the estimated and planned dose distributions, with potential applications in online adaptive workflows. Most critically, iDoTA provides the millisecond speed needed in real-time adaptive treatments, which can be further reduced if pre-computing all beam shape intensity volumes for each angle in the treatment plan. Future work could even include the magnetic field strength as an additional token in the sequence, similar to the energy token in previous transformer-based proton dose prediction models.

5 | CONCLUSION

Combining the convolutional layers extracting local features with a Transformer backbone routing distant information, iDoTA outperforms any previous deep learning model in photon dose calculation. The presented iDoTA model predicts beam dose distributions in few milliseconds with high accuracy. The per-beam prediction speed translates into estimating full VMAT dose distributions in less than 10 seconds on average, instead of the several minutes required by clinical algorithms or previous data-driven models. Given its speed and versatility, iDoTA can accelerate several steps of the radiotherapy workflow: from treatment planning and quality assurance to real-time adaption.

CONFLICT OF INTEREST STATEMENT
The authors have no conflicts of interest to disclose.

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

REFERENCES
1. Hussein M, Heijmen BJM, Verellen D, Nisbet A. Automation in intensity modulated radiotherapy treatment planning—a review of recent innovations. *Br J Radiol*. 2018;91(1092):20180270-20180270.
2. Mohan R, Chui C, Lidofsky L. Differential pencil beam dose computation model for photons. *Med Phys*. 1986;13(1):64-73.

3. Boyer A, Mok Ed. A photon dose distribution model employing convolution calculations. *Med Phys*. 1985;12(2):169-177.

4. Ahnesjö A. Collapsed cone convolution of radiant energy for photon dose calculation in heterogeneous media. *Med Phys*. 1989;16(4):577-592.

5. Ulmer W, Pyyry J, Kaissl W. A 3D photon superposition/convolution algorithm and its foundation on results of Monte Carlo calculations\ast. *Phys Med Biol*. 2005;50(8):1767-1790.

6. Sievinen J, Ulmer W, Kaissl W. AAA Photon Dose Calculation Model in Eclipse™. https://www.yumpu.com/en/document/view/20358138/aaa-photon-dose-calculation-model-in-eclipsetm

7. Jia X, Gu X, Graves YJ, et al. GPU-based fast Monte Carlo simulation for radiotherapy dose calculation. *Phys Med Biol*. 2011;56(22):7017-7031.

8. Jahnke L, Fleckenstein J, Wenz F, Hesser J. GMC: a GPU implementation of a Monte Carlo dose calculation based on Geant4. *Phys Med Biol*. 2012;57(5):1217-1229.

9. Hissoiny S, Raaijmakers AJE, Ozell B, et al. Fast dose calculation in magnetic fields with GPUMCD. *Phys Med Biol*. 2011;56(16):5119-5129.

10. Meyer P, Noblet V, Mazzara C, Lallement A. Survey on deep learning for radiotherapy. *Comput Biol Med*. 2018;98:126-146.

11. Ronneberger O, Fischer P, Brox T. *U-net: Convolutional Networks for Biomedical Image Segmentation*. Springer International Publishing; 2015:234-241.

12. Goodfellow I, Pouget-Abadie J, Mirza M, et al. *Generative Adversarial Nets*. vol. 27. Curran Associates, Inc.; 2014.

13. Kearney V, Chan JW, Haaf S, et al. DoseNet: a volumetric dose prediction algorithm using 3D fully-convolutional neural networks. *Phys Med Biol*. 2018;63(23):235022-235022.

14. Chen X, Men K, Li Y, et al. A feasibility study on an automated method to generate patient-specific dose distributions for radiotherapy using deep learning. *Med Phys*. 2019;46(1):56-64.

15. Fan J, Wang J, Chen Z, et al. Automatic treatment planning based on three-dimensional dose distribution predicted from deep learning technique. *Med Phys*. 2019;46(1):370-381.

16. Nguyen D, Long T, Jia X, et al. A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning. *Sci Rep*. 2019;9(1):1076-1076.

17. Kajikawa T, Kadoya N, Ito K, et al. A convolutional neural network approach for IMRT dose distribution prediction in prostate cancer patients. *J Radiat Res*. 2019;60(5):685-693.

18. Ma M, Buyyounouski MK, Vasudevan V, et al. Dose distribution prediction in isodose feature-preserving voxelization domain using deep convolutional neural network. *Med Phys*. 2019;46(7):2978-2987. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mp.13618

19. Nguyen D, Jia X, Sher D, et al. 3D radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected U-net deep learning architecture. *Phys Med Biol*. 2019;64(6):65020-65020.

20. Lee H, Kim H, Kwak J, et al. Fluence-map generation for prostate intensity-modulated radiotherapy planning using a deep-neural-network. *Sci Rep*. 2019;9(1):15671-15671.

21. Wang W, Sheng Y, Wang C, et al. Fluence map prediction using deep learning models – direct plan generation for pancreas stereotactic body radiation therapy. *Front Artif Intell*. 2020;3(September):1-10.

22. Peng Z, Shan H, Liu T, et al. 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*. 2019;7:76680-76689.

23. Bai T, Wang B, Nguyen D, Jiang S. Deep dose plugin: towards real-time Monte Carlo dose calculation through a deep learning-based denoising algorithm. *Mach Learn: Sci Technol*. 2021;2(2):25033-25033.

24. Neph R, Lyu Q, Huang Y, et al. DeepMC: a deep learning method for efficient Monte Carlo beamlet dose calculation by predictive denoising in magnetic resonance-guided radiotherapy. *Phys Med Biol*. 2021;66(3):35022-35022.

25. Xing Y, Zhang Y, Nguyen D, et al. Boosting radiotherapy dose calculation accuracy with deep learning. *J Appl Clin Med Phys*. 2020;21(8):149-159.

26. Dong P, Xing L. 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*. 2020;65(3):35010-35010.

27. Fan J, Xing L, Dong P, et al. Data-driven dose calculation algorithm based on deep U-Net. *Phys Med Biol*. 2020;65(24):245035-245035.

28. Xing Y, Nguyen D, Lu W, et al. Technical Note: A feasibility study on deep learning-based radiotherapy dose calculation. *Med Phys*. 2020;47(2):753-758.

29. Zhu J, Liu X, Chen L. A preliminary study of a photon dose calculation algorithm using a convolutional neural network. *Phys Med Biol*. 2020;65(20):20NT02-20NT02.

30. Kontaxis C, Bol GH, Lagendijk JJW, Raaymakers BW. DeepDose: Towards a fast dose calculation engine for radiation therapy using deep learning. *Phys Med Biol*. 2020;65(7):75013-75013.

31. Tsekas G, Bol GH, Raaymakers BW, Kontaxis C. DeepDose: a robust deep learning-based dose engine for abdominal tumours in a 1.5 T MRI radiotherapy system. *Phys Med Biol*. 2021;66(6):65017-65017.

32. Tsekas G, Bol GH, Raaymakers BW. Robust deep learning-based forward dose calculations for VMAT on the 1.5T MR-Linac. *Phys Med Biol*. 2022;67:225020.

33. Vaswani A, Shazeer N, Parmar N, et al. *Attention Is All You Need*. vol. 2017. arXiv:1706.03762 [cs.CL].

34. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019-2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*. 2019;1:4171-4186.

35. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst*. 2020-Decem;2020.

36. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale. 2020. arXiv:2010.11929 [cs.CV].

37. D'Ascoli S, Touvron H, Leavitt M, et al. ConViT: improving vision transformers with soft convolutional inductive biases. Proceedings of the 38th International Conference on Machine Learning, PMLR, 2021;139:2286-2296.

38. Pastor-Serrano O, Perkó Z. Learning the physics of particle transport via transformers. *Proc AAAI Conf Artif Intell*. 2022;36(11):12071-12079.

39. Pastor-Serrano O, Z Perkó. Millisecond speed deep learning based proton dose calculation with Monte Carlo accuracy. *Phys Med Biol*. 2022;67(10):105006.

40. Xiao F, Cai J, Zhou X, et al. TransDose: a transformer-based UNet model for fast and accurate dose calculation for MR-LINACs. *Phys Med Biol*. 2022;67(12):125013.

41. Ba JL, Kiros JR, Hinton GE. Layer Normalization. 2016. arXiv:1607.06450 [stat.ML].

42. Xiong R, Yang Y, He D, et al. On layer normalization in the transformer architecture. *37th International Conference on Machine Learning, ICML 2020*, PartF16814:10455-10464, 2020.

43. Srivastava N, Hinton G, Krizhevsky A, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. *J Mach Learn Res*. 2014;15:1929-1958.

44. Hendrycks D, Gimpel K. Gaussian error linear units (GELUs). pp. 1-9, 2016. arXiv:1606.08415 [cs.LG].

45. You Y, Li J, Reddi S, et al. Large batch optimization for deep learning: Training BERT in 76 minutes. 2019. arXiv:1904.00962 [cs.LG].

46. Abadi M, Agarwal A, Barham P, et al. TensorFlow: large-scale machine learning on heterogeneous distributed systems. Technical report. https://www.tensorflow.org/about/bib

47. Low DA, Harms WB, Mutic S, Purdy JA. A technique for the quantitative evaluation of dose distributions. *Med Phys*. 1998;25(5):656-661.

48. Yan C, Combine AG, Bednarz G, et al. Clinical implementation and evaluation of the Acuros dose calculation algorithm. *J Appl Clin Med Phys*. 2017;18(5):195-209.

49. Fogliata A, Nicolini G, Clivio A, et al. Critical appraisal of acuros XB and anisotropic analytic algorithm dose calculation in advanced non-small-cell lung cancer treatments. *Int J Radiat Oncol Biol Phys*. 2012;83(5):1587-1595.

50. Vasudevan V, Shen L, Huang C, et al. Implicit neural representation for radiation therapy dose distribution. *Phys Med Biol*. 2022;67(12):125014.

51. Tilly D, Van De Schoot AJAJ, Grusell E, et al. Dose coverage calculation using a statistical shape model – applied to cervical cancer radiotherapy. *Phys Med Biol*. 2017;62(10):4140-4159.

How to cite this article: Pastor-Serrano O, Dong P, Huang C, Xing L, Perkó Z. Sub-second photon dose prediction via transformer neural networks. *Med Phys*. 2023;1-13. https://doi.org/10.1002/mp.16231