



Published in final edited form as:

Phys Med Biol. ; 66(3): 035022. doi:10.1088/1361-6560/abca01.

DeepMC: A Deep Learning Method for Efficient Monte Carlo Beamlet Dose Calculation by Predictive Denoising in Magnetic Resonance-Guided Radiotherapy

Ryan Neph,

Department of Radiation Oncology, University of California Los Angeles, 200 Medical Plaza, #B265, Los Angeles, California 90095

Qihui Lyu,

Department of Radiation Oncology, University of California Los Angeles, 200 Medical Plaza, #B265, Los Angeles, California 90095

Yangsibo Huang¹,

Department of Radiation Oncology, University of California Los Angeles, 200 Medical Plaza, #B265, Los Angeles, California 90095

You Ming Yang,

Department of Radiation Oncology, University of California Los Angeles, 200 Medical Plaza, #B265, Los Angeles, California 90095

Ke Sheng^{*}

Department of Radiation Oncology, University of California Los Angeles, 200 Medical Plaza, #B265, Los Angeles, California 90095

Abstract

Emerging magnetic resonance (MRI) guided radiotherapy affords significantly improved anatomy visualization and, subsequently, more effective personalized treatment. The new therapy paradigm imposes significant demands on radiation dose calculation quality and speed, creating an unmet need for the acceleration of Monte Carlo (MC) dose calculation. Existing deep learning approaches to denoise the final plan MC dose fail to achieve the accuracy and speed requirements of large-scale beamlet dose calculation in the presence of a strong magnetic field for online adaptive radiotherapy (OART) planning. Our deep learning dose calculation method, DeepMC, addresses these needs by predicting low-noise dose from extremely noisy (but fast) MC-simulated dose and anatomical inputs, thus enabling significant acceleration. DeepMC simultaneously reduces MC sampling noise and predicts corrupted dose buildup at tissue-air material interfaces resulting from MR-field induced electron return effects. Here we demonstrate our model's ability to accelerate dose calculation for daily treatment planning by a factor of 38 over traditional low-noise MC simulation with clinically meaningful accuracy in deliverable dose and treatment delivery parameters. As a post-processing approach, DeepMC provides compounded acceleration

^{*}Corresponding Author: All communications may be addressed to Ke Sheng at ksheng@mednet.ucla.edu or by mail at: 200 Medical Plaza #B265, University of California, c/o Ke Sheng, Los Angeles, California 90095.

¹Author is currently affiliated with: Department of Electrical Engineering, Princeton University, 41 Olden Street, Princeton, New Jersey 08544

of large-scale dose calculation when used alongside established MC acceleration techniques in variance reduction and GPU-based MC simulation.

Introduction

Radiation therapy (RT) is one of the primary modalities for cancer treatment¹. The success of RT hinges on the subtle balance between sufficient tumor doses and normal tissue toxicity. To attain the goal, one must perform tasks that include patient imaging, anatomic structure segmentation, optimization of plan delivery parameters, and validation of accurate radiation delivery. The availability of electron density and 3D anatomy for dose calculation and treatment planning provided by computed tomography (CT) have made it the imaging modality-of-choice in radiation therapy since its inception nearly 50 years ago. CT-guided radiation therapy (CTgRT) has significantly advanced the accuracy of radiotherapy^{2,3}, but its limitations in soft-tissue visualization, real-time monitoring, and harmful ionizing dose have been increasingly recognized. Motivation to overcome these challenges has led to a paradigm shift towards the use of magnetic resonance imaging (MRI), which affords better soft-tissue contrast, real-time tumor tracking, and eliminates the use of ionizing radiation^{4,5}. These benefits are not without their caveats. Besides the lack of electron density information for radiation dose calculation, the strong magnetic field used by MRI perturbs dose deposition, causing significant inaccuracies in fast model-based dose calculation techniques developed for efficient CTgRT.

Image synthesis using generative adversarial networks (GAN) has provided a solution to assigning electron density to the MR images for dose calculation^{6,7}. The second problem to significantly accelerate dose calculation without compromising accuracy for adaptive online radiotherapy (OART)^{8–10} beamlet dose calculation remains unsolved. State of the art methods for calculating dose in the magnetic field-free condition employ deterministic convolution-based algorithms^{11–16}, which are extremely efficient for handling a large amount of beamlet dose calculation¹². These analytical methods reuse pre-calculated energy-specific dose deposition *kernels*, whose shift-invariance is violated by the Lorentz forces exhibited on electrons in a magnetic field. The resultant electron return effects (EREs), commonly observed at the interfaces between anatomies with large density gradients, have detrimental effects on the dose calculation accuracy to the patient^{17,18}. Failing to account for EREs properly can generate disparities between expected and deliverable doses at these tissue-air interfaces in magnetic resonance-guided radiotherapy (MRgRT) by up to 70%^{19,20}. On the other hand, the disparities can be mitigated with accurate EREs modeling in beamlet dose calculation²¹.

An alternative method for calculating radiation dose is Monte Carlo (MC) simulation, which uses a probabilistic sampling of particle physics to simulate the transportation and dose deposition of primary X-rays and their child particles. Different from the analytical methods, MC is based on first principles without assuming shift-invariance that is violated with the magnetic field. Therefore, MC is better suited for MRgRT but substantially slower to simulate the number of particles needed for acceptably low statistical noise²² than the analytical methods. Besides conventional MC acceleration techniques^{22–24}, efforts have

been made to accelerate MC dose calculation using graphics processing units (GPUs)^{25–27} for final plan dose calculation, but additional acceleration is still needed for the orders-of-magnitude larger beamlet dose calculation task for OART to mitigate undesirable EREs. The computational inefficiency of MC creates a bottleneck in time sensitive OART, where accurate beamlets need to be computed to mitigate EREs.

In this work we present a novel deep learning-based MC post-processing framework, termed DeepMC, that employs deep convolutional neural networks (CNNs) as a complement to traditional MC acceleration techniques.

Methods

DeepMC Model Architecture

To address the challenges of efficient MR-guided dose calculation, we've developed a 3D CNN (Figure 1) comprised of a UNet²⁸ with three spatial feature scales, batch normalization²⁹, residual^{30,31} and squeeze-excitation³² blocks, and ReLU activations³³. Skip connections between matched spatial feature scales in the encoder and decoder halves of the UNet carry high resolution features through the network, preserving fine spatial dose detail in the final prediction. Learned stride-two convolution and stride-two transposed convolution are used for changing the spatial feature resolution between each UNet level. High-noise MC-simulated dose and CT number (as Hounsfield units) are passed in as two-channelled 3D input to provide descriptive information for the model to learn to accurately predict dose containing EREs at the interfaces between high- and low-density anatomical media such as soft-tissue and internal airways. Our fully convolutional implementation further enables prediction of dose volumes with varying size³⁴, granting flexible use during prediction of new cases. To guarantee matched feature map dimensions in both the encoder and decoder halves of the network for each spatial scale, the input data batch is zero-padded along each of the three physical dimensions, and later cropped to the original dimensions.

The UNet section of our model is implemented at each spatial scale from a sequence of layer blocks. The blocks are composed of a 3D convolution with $1 \times 1 \times 1$ kernels (for changing the feature map dimensionality), followed by three residual 3D convolution blocks with $3 \times 3 \times 3$ kernels (for spatial feature learning), and finally a squeeze-excitation block with a feature squeeze factor of four (to permit richer exploitation of inter-feature relationships during prediction). Residual convolution blocks carry out batch normalization, followed by activation with the non-linear ReLU operation, and finally convolution. This “pre-activation” ordering in each block promotes better flow of information in deep networks when residual connections are included³¹. Downsampling in the UNet's encoder half uses 3D convolution layers with $3 \times 3 \times 3$ kernels and stride of two for all spatial dimensions. Upsampling, conversely uses 3D transposed convolution layers with $3 \times 3 \times 3$ kernels and stride of two. The number of feature maps produced by each block is updated based on the spatial scales; the number of features is doubled each time the spatial resolution is halved, and the number of features is halved when the spatial resolution is doubled. We begin with 64 features produced in the first block at the full resolution scale. In the UNet's decoder section, the output of each block is concatenated to the output of the corresponding block from the encoder section. The output of the UNet is processed by a 3D convolution layer with

1×1×1 kernels to reduce the feature size to one, followed by a sequence of three residual convolution layers to perform the final dose prediction.

DeepMC is trained using the stochastic gradient descent algorithm with a batch size of 30, a learning rate initialized to 0.03 and decayed along an exponential schedule tracking the number of iterations with a decay factor of 0.993. Training was performed in parallel on four NVIDIA GTX 1080 Ti GPUs for 50 epochs and lasted a total of 22 hours. Loss convergence curves are presented in Figure 2.

In Figure 3, a sample prediction at three consecutive axial slices for one beamlet is shown alongside the high-noise dose input and low-noise ground truth dose, such that both high- and low-dose regions of the beamlet are visible. Without including geometry information (CT Numbers) during prediction, this is a challenging task for such aggressively under-sampled dose inputs, where estimations of EREs in the high-noise dose are otherwise masked by noise corruptions (arrows in Figure 3 indicate these corrupted areas in the high-noise dose). DeepMC, instead generates its dose predictions from both the geometry input (material density) and a high-noise approximation of the beamlet dose.

Our model is implemented in the Python programming language using the TensorFlow v2.10 deep learning library³⁵ and trained via supervised backpropagation using a weighted L2 loss on the voxel wise difference between DeepMC dose and low-noise ground truth dose. Beamlet dose suffers from severe data imbalance between the large number of low-dose voxels and the small number of high-dose voxels, which makes learning unstable. Figure 4 shows the observed voxelized dose density and cumulative density.

To address this imbalance in the data distribution, we defined a voxel wise decaying exponential loss weighting function, shown by equation (1) for ground truth dose volume Y and prediction \hat{Y} , that assigns unitary weight to voxels with dose equal to the maximum ground truth dose for each beamlet, and decays for smaller values.

$$W(Y) = \exp \left[-\alpha \cdot \left(1 - \frac{\frac{1}{2}(\hat{Y} + Y)}{\max(Y)} \right) \right]. \quad (1)$$

The exponential decay factor, α , is a configurable hyperparameter for which we found α equal to 3.0 to be empirically effective in stabilizing the training progress and achieving accurate and generalizable predictions.

Dose Data Generation

For MRgRT dose calculation, both the high-noise dose (model input) and low-noise dose (ground truth) are calculated using stochastic MC simulation methods, granting a physically accurate method of generating virtually unlimited data. Although we can theoretically continue to synthesize data pairs until the model performance on a separate testing dataset converges to the ground truth, there is a practical limitation due to the simulation speed. The Monte Carlo simulation framework, Geant4^{36,37}, used in this study was selected for its accuracy and flexibility in simulating dose for MRgRT, but as a central processing unit

(CPU)-based particle simulation toolkit, it is too slow for direct use in OART. Simulation on a single CPU core, for a few thousand particles to produce highly corrupted high-noise dose is relatively fast, on the order of seconds for a single beamlet. On the other hand, the time to simulate the hundreds of thousands of particles necessary to produce low-noise dose is on the order of minutes. Linear improvements in simulation speed are attainable by enabling Geant4's multi-threaded parallel processing capability to make use of every core of the CPU. For OART and for the purposes of generating data to train our CNN model this is still too slow.

To address these speed concerns, a distributed cluster of 10 computational nodes and a combined total of 216 logical CPU cores was used to achieve asynchronous and scalable Monte Carlo dose calculation. The training dataset consists of 56,000 paired beamlet dose volumes in total; the high-noise dose was simulated using 2,000 particles per beamlet and the low-noise (ground truth) dose for each beamlet was simulated independently for 100,000 particles. A set of 2,400 randomly selected coplanar beam angles and translations (4 mm radius) from the planning target volume (PTV) centroid were defined from each of eight retrospectively collected Head and Neck (H&N) pre-treatment CT scans containing expert-defined anatomic delineations. Since we would like to produce a general model for prediction of low-noise dose, we not only benefit from the simplicity of randomly selecting the beamlet geometries for training, but additionally produce a dataset exhibiting greater geometric diversity. Rather than limiting training to only the most common beamlet orientations, such as those sourced from historical clinical treatment plans, a generality is maintained in the scope of beamlet orientations for which dose can be accurately predicted when creating new plans.

To augment the size of the training dataset, we repeated the high-noise dose calculation five times with identical statistical settings for each beamlet, producing random observations of the same ground truth dose distribution. This relatively inexpensive strategy is made even more efficient by sharing the Geant4 initialization overhead among them. The low-noise simulation is only performed once due to the diminishing variance with an increasing number of histories in MC simulation. We further remove unnecessary training complexity by standardizing the beamlet orientations with simple 2D rotation. This guarantees that every beamlet seen by the model has the same orientation, simplifying the latent understanding that the model must learn.

An independent testing dataset was generated by calculating high- and low-noise dose pairs for 8,043 active beamlets selected to enable creation of an IMRT plan for two separate H&N patients, withheld during model training. For the first patient, seven beams were equally distributed in a 360-degree ring around a single isocenter, placed at the PTV centroid. In the second, six and seven beams were distributed around two disjoint PTVs, respectively.

Experimental Design

To validate our approach to MC dose prediction, we performed three types of evaluation. First, a virtual slab phantom was created to emulate two media interfaces: one dividing high- and low-density media, and another dividing the same low- and high-density media but in reverse order. Two variants of this phantom were used: one with water and lung,

and the other with aluminum and air as the high- and low-density media, respectively. X-ray beamlets were fired into each phantom variant at random angles between -66 and 66 degrees from the interface's normal vector. Beamlet dose data from 80% of the angles of incidence were random selected for model training, and the remaining 20% were held for testing. A prototype 2D version of our DeepMC network (termed DeepMC2D) was trained on slices of the phantom data. Figure 5 shows the phantom geometry used for the first evaluation, which was developed to test the feasibility of CNN-based model's at predicting ERE-induced dose patterns. To demonstrate a proof of concept for DeepMC2D before moving forward to 3D prediction in patient geometries, the normalized mean absolute errors (NMAE) for both the high-noise MC dose and the DeepMC2D predicted dose were calculated with respect to the low-noise ground truth dose.

Next, we trained our proposed model using 56,000 paired high- and low-noise 3D MC dose volumes, generated using our distributed simulation framework. The dose predictions from our 3D DeepMC model (just called DeepMC) were then produced alongside the paired high- and low-noise MC dose for each of the training and testing beamlet configurations. NMAE was calculated between the high-noise dose (model input) and low-noise ground truth dose, and again between the DeepMC-predicted dose and the low-noise ground truth dose. Maximum low-noise ground truth dose for each beamlet served as the normalization reference. Comparing the two NMAE values indicates the overall improvement of dosimetric accuracy achieved by using our prediction model instead of the high-noise MC dose directly.

Finally, an end-to-end analysis was performed to evaluate DeepMC's ability to calculate the beamlet dose for effective treatment planning. First, DeepMC dose was predicted for every active² IMRT beamlet in each of three testing dataset H&N CT scans (shown in Figure 6).

Deliverable treatment plans were created by optimizing plan delivery parameters (X-ray fluence, i.e. the quantity of X-rays to emit per unit area of each beam) according to a previously formulated convex optimization problem³⁸ presented in equation (2).

$$\begin{aligned} & \underset{f}{\text{minimize}} \quad \underbrace{\frac{1}{2} \|W(Af - d_0)\|_2^2}_{\text{Dose Fidelity}} + \underbrace{\lambda(\|D_x f\|_1 + \|D_y f\|_1)}_{\text{Anisotropic Total Variation}} \\ & \text{subject to} \quad f \geq 0. \end{aligned} \quad (2)$$

A is the planning dose matrix with vectorized 3D dose for each beamlet occupying a column, d_0 is the ideal vectorized 3D dose deposition prescribed by the physician and further informed by organ-specific radiation dose tolerances, W is a diagonal matrix encoding the anatomical-structure-specific dose goal importance, f is the optimization variable encoding the per-beamlet X-ray fluence, and λ , D_x , and D_y are the term weighting and first-order finite differencing operators for the x - and y -axes of the beam-specific two-dimensional (2D) fluence maps encoded in f , respectively. The first term encourages the planned dose to match the physician-prescribed dose, while the second term encourages

²Active beamlets are those intersecting with a 1 cm radial expansion of the PTV

the optimal 2D fluence maps to be smooth and, therefore, more efficient to deliver. The convex optimization problem in equation (2) is solved using the Fast Iterative Shrinkage Thresholding Algorithm (FISTA)³⁹, by which f is iteratively updated until convergence to the global minimizer (f^*) of the objective function while obeying the non-negativity constraint. Plans were optimized using three methods for pre-calculating A : low-noise ground truth MC simulation (A_{GT}), high-noise MC simulation (A_{Noisy}), and our model's dose prediction (A_{DeepMC}); these plans will be hereafter referred to as ground truth, noisy, and DeepMC, respectively. The plans were designed such that the planning target volume (PTV) receives a uniform dose equal to the prescription dose. Additionally, the dose to each organ-at-risk (OAR) was reduced as much as possible by adjusting the importance weightings in W while optimizing the ground truth plan, then holding W constant to optimize the other two (noisy, DeepMC) plans, for unbiased comparison. Plan normalization was performed after optimization to set D95 (5th percentile of dose) of the PTV to the prescription dose.

After optimizing each plan, the product Af^* gives a dose estimate (termed *planning* dose), which is recalculated as the *deliverable* A_{GT} dose with higher quality dose calculation. Due to the associated cost of computing A_{GT} with low-noise MC simulation, in a clinical planning timeframe the deliverable dose is calculated only after a plan has first been approved from its planning dose. We investigate both the error in the observable plan dose (planning vs. deliverable) as well as the deliverable plan quality differences generated by each dose approximation during optimization (noisy and DeepMC vs ground truth). We adopt standard plan evaluation metrics, including dose volume histogram (DVH) curves for each PTV and OAR structure and compares various plan quality metrics quantifying first order dose statistics, uniformity, and coverage. Included in this analysis are the D1, D2, D98, D99 (99th, 98th, 2nd, and 1st percentiles of dose), mean dose, maximum dose, homogeneity index (HI)⁴⁰, describing the uniformity of dose within the PTV, and conformity number (CN)⁴¹, describing the compactness of dose to control dose spillage beyond the PTV. Gamma analysis at various gamma criteria (e.g. 1mm/1%, 2mm/2%, etc.) is conducted, and the percentage of total voxels *passing* the gamma test (gamma score less than or equal to 1.0) is reported. Masked gamma passing percentages are also provided, which only consider voxels with dose equal to or greater than a constant threshold, determined as some fraction of the ground truth maximum 3D dose for the treatment plan.

Results

The dose predicted by DeepMC2D for the slab phantom geometry with two different material pairs is given by Figure 7 alongside low-noise (ground truth) dose and high-noise (model input) beamlet dose. DeepMC2D dose (high-noise dose, respectively) from the testing dataset has NMAE 1.253e-3 (2.775e-3). Ignoring voxels with less than 10% of the maximum dose for each beamlet, the NMAE was 3.960e-2 (9.724e-2). Average testing dataset gamma passing scores at a 0.2mm/0.2% gamma criteria for the water/air phantom were 96.36% and 89.38% for DeepMC2D-predicted and high-noise doses, respectively. In the aluminum/air phantom, average gamma passing rates, using the same criteria, were 97.93% and 95.86% for DeepMC2D and high-noise doses, respectively.

In the patient geometry-based treatment planning experiments, 3D dose prediction times were 26 ms per beamlet, amortized over batches of 30 beamlets on a single NVIDIA 1080Ti GPU. High-noise (low-noise, respectively) simulation required 1.76 s (86.35 s) per beamlet on one CPU thread. Utilizing a more clinically realistic computing platform with an Intel i9–7900X CPU featuring 20 logical cores and 1 NVIDIA GTX 1080Ti GPU, high-noise MC simulation and subsequent DeepMC prediction for the IMRT plans of the three testing patients, which modulated 2,111 and 5,932 beamlets, respectively, was achieved in clinically feasible times of 241 s and 658 s compared to 9,114 s and 24,899 s required by low-noise MC simulation.

Figure 8 compares deliverable dose DVHs for the ground truth, DeepMC, and noisy plans. Plans for the first testing patient targeted a single PTV with a prescription dose of 70 Gy. PTV dose for the DeepMC plan shows better agreement with ground truth than the noisy plan, particularly in the high-dose region (lower right curve segment). OAR dose for the DeepMC plan displays smaller error with ground truth than the noisy plan, especially in the spinal cord, left cochlea, both temporomandibular joints (TMJs), and both parotid structures. The plans for the second testing patient targeted two PTVs with prescribed doses of 54 Gy and 59.4 Gy and a boosted internal volume of the second PTV to 70 Gy. As in the first testing patient, deliverable PTV dose for the DeepMC plan agrees more closely with the ground truth, especially in the high dose region of the 70 Gy PTV. For the 54 Gy and 59.4 Gy PTVs, the DeepMC plan shows cold spots, observed as lower curves at the “knee” immediately before the steeply downward-sloping portion of the curve. Agreement of OAR dose for the DeepMC plan with the ground truth plan is better than the noisy plan, particularly in the larynx, spinal cord, right parotid, and left temporomandibular joint. The trend towards PTV overdosing in the noisy plan continues for the third testing patient, with notable OAR sparing again observed in the brainstem when treated with the DeepMC plan.

Figure 9 compares planning and deliverable dose for the noisy and DeepMC plans. Differences are larger between planning and deliverable PTV dose for the noisy plan than the DeepMC plan, indicating improved planning accuracy using DeepMC-predicted vs. high-noise beamlet dose. Better agreement of both planning and deliverable dose for the DeepMC plan to the ground truth plan substantiates this advantage over the noisy plan. The pronounced differences seen in the planning dose (Figure 9B, Figure 9D, and Figure 9F) for the testing patients indicate that during the evaluation of the noisy plan in a clinical setting (via the planning dose), it would be rejected for failing to meet the PTV dose goals prescribed by the physician, while DeepMC plans (Figure 9A, Figure 9C, and Figure 9E) match the ground truth plan in quality and would, therefore, pass clinical evaluation.

Table 1, Table 2, and Table 3 present the plan quality metrics for PTV and OAR structures from the deliverable dose for each of the testing patients’ plans. Dose is listed in Gy. Conformity number (CN) is unitless, with unity being ideal. The homogeneity index is unitless as well, with lower values indicating better dose homogeneity.

The deliverable dose for the DeepMC plan PTV D_{mean} and D_{max} for the first testing patient are closer to the ground truth plan, with 0.3% (0.24 Gy) and 1.8% (1.48 Gy) error, than the noisy plan, with 2.71% (2.01 Gy) and 9.11% (8.51 Gy), respectively. Similarly, most OARs

show deliverable dose that best agrees with the ground truth plan when DeepMC beamlet dose rather than high-noise beamlet dose is used for plan optimization. The plan with lower error compared to ground truth is indicated by green fill for each structure. These results are observed for the other two testing patients as well.

Axial dose color wash slices at the center of the PTV for each testing patient are presented in Figure 10, which compares the deliverable dose of noisy, DeepMC, and ground truth plans. Dose differences for the noisy and DeepMC plans compared to the ground truth plan demonstrate the changes in dose deliverable to the patient when using their respective dose calculation techniques. Differences in the deliverable dose for the noisy plan compared to the ground truth plan are greater in magnitude than for the DeepMC plan, depicted as regions of dark red and blue in the central columns. Gamma index maps for the 1%/1mm gamma criteria are shown by the right-most two columns, with blue voxels “passing” the gamma test, and red voxels “failing” the gamma test. Table 4 provides 3D gamma passing rates for various gamma criteria, evaluated over the entire volume of each testing patient’s plan. Masked passing rates are also included, which were calculated as the fraction of passing to assessed voxels, excluding any voxel with a ground truth dose below 1% of the maximum ground truth dose. For the 2mm/2% criteria, DeepMC predicts planning dose, that is used to create a deliverable plan with equal or greater than 97.31% gamma passing percentage for the three testing patients. This compares to the Noisy plan’s lowest gamma passing percentage at 87.39% for the third testing patient, marking a difference between DeepMC and Noisy plans’ of nearly 10 percentage points. An improvement of approximately 2.5 percentage points for the DeepMC plan is observed for both the first and second testing patients. When the 2mm/2% gamma analysis is masked to only the voxels containing greater than or equal to 1% of the maximum ground truth deliverable dose, the accuracy of each DeepMC plan over their respective Noisy plan broadens, with as much as a 30.89 percentage point improvement, observed for the first testing patient.

Figure 11 reveals the treatment delivery parameters (per-beamlet fluence) resulting from optimization for each plan. For all testing patients, optimization using DeepMC beamlet dose converges upon a set of plan parameters that agree more closely with the ground truth parameters than optimization using high-noise beamlet dose. This suggests that high-noise beamlet dose leads to a very different delivery of X-rays than low-noise beamlet dose, while the DeepMC beamlet dose enables optimization to arrive at the ground truth plan delivery parameters.

In terms of the computational performance, DeepMC using a single logical CPU core and one GPU enables acceleration of end-to-end MC simulation and prediction by a factor of 48.4, approaching the theoretical limit of 50 set by the ratio of primary X-rays simulated for high- (2,000 particles) and low-noise (100,000 particles) dose. For a more realistic local computing environment with 20 logical CPU cores and 1 GPU, the effective acceleration factor enabled by end-to-end DeepMC dose calculation is 37.8, due to the efficiencies in both high- and low-noise MC simulation gained from CPU multithreading. Using only these modest local computational resources, end-to-end DeepMC dose calculation for fixed-beam online-adaptive IMRT planning was enabled in clinically feasible time of under 11 minutes

instead of the prohibitive time of 6.9 hours required of traditional low-noise MC dose calculation.

Discussion

Our initial 2D phantom experiments showed that CNN-based beamlet dose prediction was a feasible means to reducing the computational time of traditional low-noise MC, with its ability to predict ERE-specific dose patterns in the appropriate locations. Extending our DeepMC model into 3D dose prediction and employing it to the task of beamlet dose calculation for treatment planning, demonstrated a definitive improvement over the use of high-noise dose, directly simulated using traditional MC techniques, in a fraction of the time.

Our approach enables meaningful acceleration of MC dose simulation for OART treatment planning purposes by predicting low-noise three-dimensional (3D) dose deposition in the MRgRT setting from extremely noisy (but fast) MC simulation. Our dose prediction method distinguishes itself from traditional MC dose denoising approaches^{42–46} that use edge-preserving low pass filters to reduce dose noise as isolated images. Our method learns the dose deposition patterns from a population of diverse X-ray beam configurations and patient anatomies. As such, our method not only denoises, but reconstructs MR-guided dose deposition from a combination of regional noisy MC-simulated dose and anatomical cues for each newly observed X-ray beam. Also different from more recent works using deep learning for dose denoising^{47,48}, we confront the additional challenges of including 1.5T MRI magnetic field effects, as is the case for Elekta's Unity MR-guided linear accelerator (linac)⁴⁹, where Lorentz forces impose geometry-dependent dose perturbations as EREs, prediction at the finer beamlet-scale (5 mm width) instead of the full-beam (~5-10 cm width), and more severe under-sampling of the initial MC simulated dose due to the need to calculate the large number of beamlets vs. final beams.

Note that even with high-noise dose calculation, MC alone consumes 99% of the total planning time, due to the much greater relative efficiency of modern plan optimization algorithms. Geant4 as the MC engine without additional acceleration techniques such as variance reduction is known to be relatively slow. On the other hand, conventional MC acceleration techniques alone are inadequate, particularly for larger planning problems, such as beam orientation optimization^{50,51} and volumetric modulated arc therapy^{52,53} that involve two to three orders of magnitude more beamlets. DeepMC can be readily combined with these conventional techniques to provide additional acceleration by 38-fold on modest local computational resources. Combining additional distributed computational nodes with our scalable distributed MC simulation framework, the total DeepMC dose calculation time can be further reduced, enabling more complex and beneficial planning paradigms (with greater demands on dose calculation) within the clinically feasible timeframe. Utilizing additional CPU cores for high-noise MC simulation reduces the fraction of time for MC relative to DeepMC-prediction. Conversely, using additional GPUs for DeepMC-prediction increases the acceleration factor and gives end-to-end DeepMC dose calculation a greater speed advantage over low-noise MC simulation. Additional work is necessary to fully characterize

the performance-cost trade-offs and determine the optimal ratio of CPU to GPU resources for realize maximum acceleration; this will be the focus of future work.

In recent years, deep learning has made strides in image processing, classification, and prediction. Its medical applications have met two common roadblocks: the availability of training data and the interpretability of the results. The current task is minimally affected by these roadblocks. Instead of relying on limited, laboriously curated patient images, a diverse training data set of paired of low- and high-noise dose volumes can be almost infinitely produced via MC simulation from a relatively small number of collected patient images.

With DeepMC, we were able to simulate treatment planning using different beamlet calculation methods and quality. Treatment plan parameters determined by optimization with DeepMC-predicted beamlet dose showed better agreement with those produced using ground truth (low-noise) MC simulated beamlet dose with little additional cost. Lower error between planning and deliverable dose for DeepMC compared to noisy plans indicates that DeepMC planning dose is a more reliable estimate of deliverable dose, while high-noise planning dose produces treatment plans that the planning dosimetrist would likely reject before the more costly deliverable dose is calculated. DeepMC plans featured substantially reduced OAR mean and maximum doses and superior PTV conformity and coverage, with better agreement to the ground truth plans, for all testing patients.

The current study is for the head and neck MRgRT problem. The generality of our method for other sites and modalities, including proton and heavy ion therapy planning⁵⁴, scattered X-rays correction for cone-beam CT⁵⁵, etc., will be subject to future investigations.

It is important to note that there is still finite difference between the MC beamlets and the actual machine-produced beams. The beamlets studied in this work are a useful research tool to simulate the behavior of a treatment plan but cannot be directly applied to patients. For future clinical applications, actual measurement data must be adopted for training and validation.

Classical imaging denoising methods for Monte Carlo dose noise reduction such as [43], [42], and [56] have shown a factor of 2-4 reduction in the number of simulated particles for head and neck beamlet calculation, which is an order of magnitude smaller than 38 achieved in the current study, demonstrating the ability of deep learning not only for image denoising, but for information learning and synthesis. However, to further substantiate the statement, DeepMC needs to be directly compared with these classical methods in future studies.

Conclusions

Fast and accurate beamlet dose calculation is essential to the success of online adaptive MRgRT accounting for the actual anatomical configuration and EREs. For effective OART, we use a novel deep learning method that substantially accelerates the accurate but traditionally slow MC dose calculation process. We then performed an end-to-end test to create and evaluate IMRT plans with the novel dose calculation method. Treatment plans optimized with DeepMC-predicted beamlet dose have plan delivery parameters and deliverable dose closely matching the ground truth plan, with substantially lower

computational time. In comparison, the direct utilization of high-noise MC beamlet dose instead produces an unacceptable deviation in the deliverable dose from the planning dose. DeepMC offers relief to the computational costs of MC simulation for a wide scope of applications, enabling new technologies in many new areas.

Acknowledgements

This research is supported by DOE Grants Nos. DE-SC0017057 and DE-SC0017687, NIH Grants Nos. R44CA183390, R43CA183390 and R01CA188300.

References

1. Miller KD, Nogueira L, Mariotto AB, et al. Cancer treatment and survivorship statistics, 2019. *CA Cancer J Clin.* 2019;69(5):363–385. doi:10.3322/caac.21565 [PubMed: 31184787]
2. King RB, McMahon SJ, Hyland WB, et al. An overview of current practice in external beam radiation oncology with consideration to potential benefits and challenges for nanotechnology. *Cancer Nanotechnol.* 2017;8(1). doi:10.1186/s12645-017-0027-z
3. Jaffray DA. Image-guided radiotherapy: from current concept to future perspectives. *Nat Rev Clin Oncol.* 2012;9(12):688–699. doi:10.1038/nrclinonc.2012.194 [PubMed: 23165124]
4. Eccles CL, Adair Smith G, Bower L, et al. Magnetic resonance imaging sequence evaluation of an MR Linac system; early clinical experience. *Tech Innov Patient Support Radiat Oncol.* 2019;12:56–63. doi:10.1016/j.tipsro.2019.11.004 [PubMed: 32095556]
5. Dirix P, Haustermans K, Vandecaveye V. The Value of Magnetic Resonance Imaging for Radiotherapy Planning. *Semin Radiat Oncol.* 2014;24(3):151–159. doi:10.1016/j.semradonc.2014.02.003 [PubMed: 24931085]
6. Maspero M, Savenije MHFF, Dinkla AM, et al. Dose evaluation of fast synthetic-CT generation using a generative adversarial network for general pelvis MR-only radiotherapy. *Phys Med Biol.* 2018;63(18). doi:10.1088/1361-6560/aada6d
7. Qi M, Li Y, Wu A, et al. Multi-sequence MR image-based synthetic CT generation using a generative adversarial network for head and neck MRI-only radiotherapy. *Med Phys.* 2020;0(0):1–15. doi:10.1002/mp.14075
8. Chuter RW, Pollitt A, Whitehurst P, Mackay RI, Van Herk M, McWilliam A. Assessing MR-linac radiotherapy robustness for anatomical changes in head and neck cancer. *Phys Med Biol.* 2018;63(12). doi:10.1088/1361-6560/aac749
9. Wu Q, Chi Y, Chen PY, Krauss DJ, Yan D, Martinez A. Adaptive Replanning Strategies Accounting for Shrinkage in Head and Neck IMRT. *Int J Radiat Oncol Biol Phys.* 2009;75(3):924–932. doi:10.1016/j.ijrobp.2009.04.047 [PubMed: 19801104]
10. Castelli J, Simon A, Louvel G, et al. Impact of head and neck cancer adaptive radiotherapy to spare the parotid glands and decrease the risk of xerostomia. *Radiat Oncol.* 2015;10(1):1–10. doi:10.1186/s13014-014-0318-z [PubMed: 25567003]
11. Sievinen J, Ulmer W, Kaissl W. AAA photon dose calculation model in Eclipse. *Varian Med Syst.* 2005:1–23. http://www.rtsalon.cn/upload/RTsalon_p_3218_2.pdf.
12. Neph R, Ouyang C, Neylon J, Yang YM, Sheng K. Parallel Beamlet Dose Calculation via Beamlet Contexts in a Distributed Multi-GPU Framework. *Med Phys.* June 2019;mp.13651. doi:10.1002/mp.13651
13. Chen Q, Chen M, Lu W. Ultrafast convolution/superposition using tabulated and exponential kernels on GPU. *Med Phys.* 2011;38(3):1150–1161. doi:10.1118/1.3551996 [PubMed: 21520827]
14. Battista JJ. Dose calculations using convolution and superposition principles: The orientation of dose spread kernels in divergent x-ray beams. *Med Phys.* 1993;20(6):1685–1694. doi:10.1118/1.596955 [PubMed: 8309441]
15. Hoban PW, Murray DC, Round WH. Photon beam convolution using polyenergetic energy deposition kernels. *Phys Med Biol.* 1994;39(4):669–685. doi:10.1088/0031-9155/39/4/002 [PubMed: 15552077]

16. Chen Q, Lu W, Chen Y, Chen M, Henderson D, Sterpin E. Validation of GPU based TomoTherapy dose calculation engine. *Med Phys.* 2012;39(4):1877–1886. doi:10.1118/1.3693057 [PubMed: 22482609]
17. Rubinstein AE, Liao Z, Melancon AD, et al. Technical Note: A Monte Carlo study of magnetic-field-induced radiation dose effects in mice. *Med Phys.* 2015;42(9):5510–5516. doi:10.1118/1.4928600 [PubMed: 26328998]
18. Richter S, Pojtinger S, Mönnich D, Dohm OS, Thorwarth D. Influence of a transverse magnetic field on the dose deposited by a 6 MV linear accelerator. *Curr Dir Biomed Eng.* 2017;3(2):281–285. doi:10.1515/cdbme-2017-0058
19. Pfaffenberger A Dose Calculation Algorithms for Radiation Therapy with an MRI-Integrated Radiation Device. 2013. doi:10.11588/heidok.00014551
20. Shortall J, Vasquez Osorio E, Chuter R, et al. Assessing localized dosimetric effects due to unplanned gas cavities during pelvic MR-guided radiotherapy using Monte Carlo simulations. *Med Phys.* 2019;46(12):5807–5815. doi:10.1002/mp.13857 [PubMed: 31600837]
21. Chen X, Prior P, Chen GP, Schultz CJ, Li XA. Technical Note: Dose effects of 1.5 T transverse magnetic field on tissue interfaces in MRI-guided radiotherapy. *Med Phys.* 2016;43(8):4797–4802. doi:10.1118/1.4959534 [PubMed: 27487897]
22. Fippel M Variance Reduction Techniques. In: *Monte Carlo Techniques in Radiation Therapy*. 1st ed. Boca Raton: Taylor & Francis Group; 2013:29–39.
23. Rodriguez M, Sempau J, Brualla L. A combined approach of variance-reduction techniques for the efficient Monte Carlo simulation of linacs. *Phys Med Biol.* 2012;57(10):3013–3024. doi:10.1088/0031-9155/57/10/3013 [PubMed: 22538321]
24. Kawrakow I, Rogers DWO, Walters BRB. Large efficiency improvements in BEAMnrc using directional bremsstrahlung splitting. *Med Phys.* 2004;31(10):2883–2898. doi:10.1118/1.1788912 [PubMed: 15543798]
25. Ahmad SB, Sarfehnia A, Paudel MR, et al. Evaluation of a commercial MRI Linac based Monte Carlo dose calculation algorithm with geant 4. *Med Phys.* 2016;43(2):894–907. doi:10.1118/1.4939808 [PubMed: 26843250]
26. Paudel MR, Kim A, Sarfehnia A, et al. Experimental evaluation of a GPU-based monte carlo dose calculation algorithm in the Monaco treatment planning system. *J Appl Clin Med Phys.* 2016;17(6):230–241. doi:10.1120/jacmp.v17i6.6455 [PubMed: 27929496]
27. Jia X, Gu X, Graves YJ, Folkerts M, Jiang SB. GPU-based fast Monte Carlo simulation for radiotherapy dose calculation. *Phys Med Biol.* 2011;56(22):7017–7031. doi:10.1088/0031-9155/56/22/002 [PubMed: 22016026]
28. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. 2015:1–8. doi:10.1007/978-3-319-24574-4_28
29. Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. 2015. doi:10.1007/s13398-014-0173-7.2
30. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. *Multimed Tools Appl.* December 2015:1–17. doi:10.1007/s11042-017-4440-4
31. He K, Zhang X, Ren S, Sun J. Identity Mappings in Deep Residual Networks. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)*. 2016;9908 LNCS:630–645. doi:10.1007/978-3-319-46493-0_38
32. Hu J, Shen L, Sun G. Squeeze-and-Excitation Networks. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit.* 2018:7132–7141. doi:10.1109/CVPR.2018.00745
33. Glorot X, Bordes A, Bengio Y. Deep Sparse Rectifier Neural Networks. In: *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*. Vol 15. Fort Lauderdale; 2011:315–323.
34. Shelhamer E, Long J, Darrell T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans Pattern Anal Mach Intell.* 2017;39(4):640–651. doi:10.1109/TPAMI.2016.2572683 [PubMed: 27244717]
35. Abadi M, Barham P, Chen J, et al. TensorFlow: A system for large-scale machine learning. In: *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation*. Savannah; 2016:265–283.

36. Allison J, Amako K, Apostolakis J, et al. Recent developments in Geant4. Nucl Instruments Methods Phys Res Sect A Accel Spectrometers, Detect Assoc Equip. 2016;835:186–225. doi:10.1016/j.nima.2016.06.125
37. Agostinelli S, Allison J, Amako K, et al. Geant4—a simulation toolkit. Nucl Instruments Methods Phys Res Sect A Accel Spectrometers, Detect Assoc Equip. 2003;506(3):250–303. doi:10.1016/S0168-9002(03)01368-8
38. Nguyen D, O'Connor D, Yu VY, et al. Dose domain regularization of MLC leaf patterns for highly complex IMRT plans. Med Phys. 2015;42(4):1858–1870. doi:10.1118/1.4915286 [PubMed: 25832076]
39. Beck A, Teboulle M. A Fast Iterative Shrinkage-Thresholding Algorithm. Soc Ind Appl Math J Imaging Sci. 2009;2(1):183–202. doi:10.1137/080716542
40. International Commission on Radiation Units and Measurements. ICRU Report 83 Prescribing, Recording, and Reporting Photon-beam Intensity-modulated Radiation Therapy (IMRT). J ICRU. 2010;10(1).
41. Riet A van't Mak ACA, Moerland MA Elders LH, van der Zee W. A conformation number to quantify the degree of conformality in brachytherapy and external beam irradiation: Application to the prostate. Int J Radiat Oncol. 1997;37(3):731–736. doi:10.1016/S0360-3016(96)00601-3
42. Deasy JO, Wickerhauser MV, Picard M. Accelerating Monte Carlo simulations of radiation therapy dose distributions using wavelet threshold de-noising. Med Phys. 2002;29(10):2366–2373. doi:10.1118/1.1508112 [PubMed: 12408311]
43. Kawrakow I On the de-noising of Monte Carlo calculated dose distributions. Phys Med Biol. 2002;47(17):304. doi:10.1088/0031-9155/47/17/304
44. Fippel M, Nüsslin F. Smoothing Monte Carlo calculated dose distributions by iterative reduction of noise. Phys Med Biol. 2003;48(10):1289–1304. doi:10.1088/0031-9155/48/10/304 [PubMed: 12812447]
45. Miao B, Jeraj R, Bao S, Mackie TR. Adaptive anisotropic diffusion filtering of Monte Carlo dose distributions. Phys Med Biol. 2003;48(17):2767–2781. doi:10.1088/0031-9155/48/17/303 [PubMed: 14516100]
46. Naqa I El, Deasy JO, Vivic M. Locally adaptive denoising of Monte Carlo dose distributions via hybrid median filtering. In: IEEE Nuclear Science Symposium. ; 2003:2703–2706. doi:0-7803-8257-9/04
47. Peng Z, Shan H, Liu T, Pei X, Wang G, Xu XG. 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. doi:10.1109/ACCESS.2019.2921013
48. Fornander H Denoising Monte Carlo Dose Calculations Using a Deep Neural Network. 2019. <http://www.diva-portal.org/smash/get/diva2:1366439/FULLTEXT01.pdf>.
49. Raaymakers BW, Lagendijk JJW, Overweg J, et al. Integrating a 1.5 T MRI scanner with a 6 MV accelerator: Proof of concept. Phys Med Biol. 2009;54(12). doi:10.1088/0031-9155/54/12/N01
50. O'Connor D, Voronenko Y, Nguyen D, Yin W, Sheng K. Fast non-coplanar beam orientation optimization based on group sparsity. October 2017:1–11. <http://arxiv.org/abs/1710.05308>.
51. Lyu Q, Neph R, Yu VY, Ruan D, Boucher S, Sheng K. Many-isocenter optimization for robotic radiotherapy. Phys Med Biol. 2020;65(4). doi:10.1088/1361-6560/ab63b8
52. Lyu Q, Connor DO, Ruan D, Yu V, Nguyen D, Sheng K. VMAT optimization with dynamic collimator rotation. 2018;(May). doi:10.1002/mp.12915
53. Lyu Q, Yu VY, Ruan D, Neph R, O'Connor D, Sheng K. A novel optimization framework for VMAT with dynamic gantry couch rotation. Phys Med Biol. 2018;63(12):125013. doi:10.1088/1361-6560/aac704 [PubMed: 29786614]
54. Kueng R, Frei D, Volken W, et al. Adaptive step size algorithm to increase efficiency of proton macro Monte Carlo dose calculation. Radiat Oncol. 2019;14(1):165. doi:10.1186/s13014-019-1362-5 [PubMed: 31500647]
55. Xu Y, Bai T, Yan H, et al. A practical cone-beam CT scatter correction method with optimized Monte Carlo simulations for image-guided radiation therapy. Phys Med Biol. 2015;60(9):3567–3587. doi:10.1088/0031-9155/60/9/3567 [PubMed: 25860299]

56. El Naqa I, Kawrakow I, Fippel M, et al. A comparison of Monte Carlo dose calculation denoising techniques. *Phys Med Biol*. 2005;50(5):909–922. doi:10.1088/0031-9155/50/5/014 [PubMed: 15798264]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

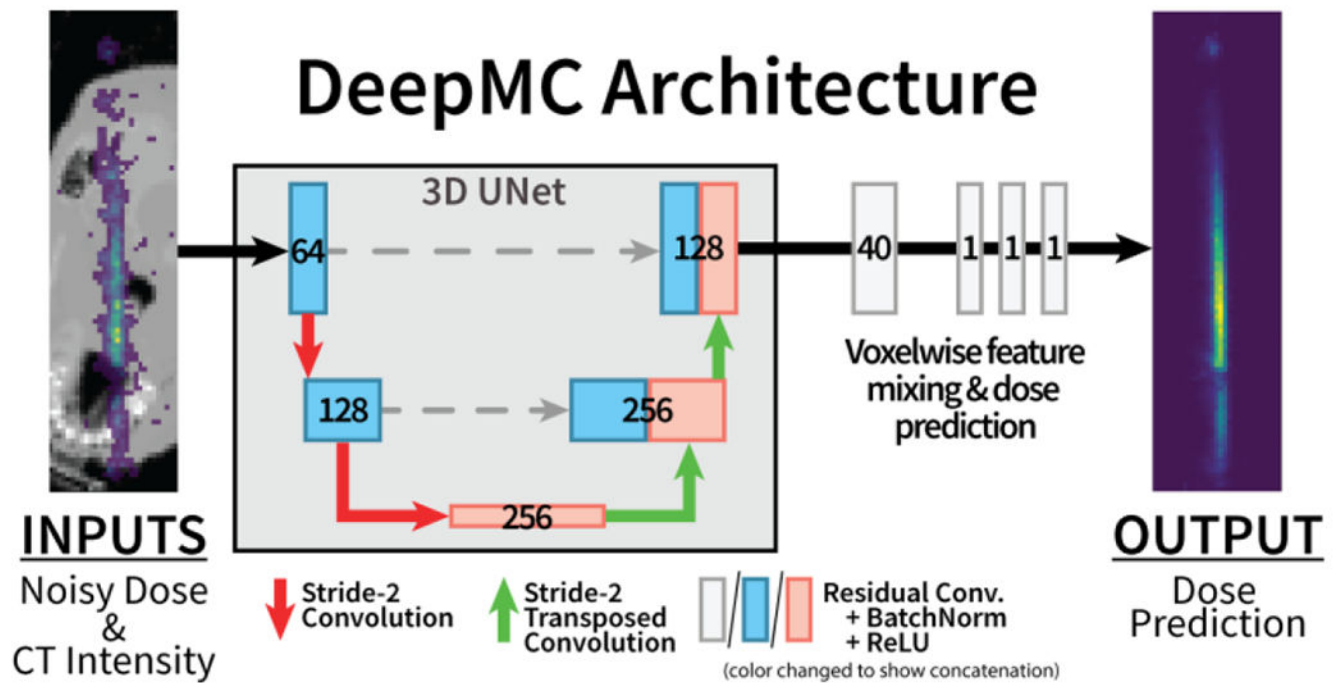


Figure 1.

Fully convolutional model architecture used by our method, DeepMC. Numbers in blocks indicate number of feature channels produced by learned convolutional kernels at each stage. $3 \times 3 \times 3$ kernels are used in UNet layers. $1 \times 1 \times 1$ kernels are used for feature mixing and dose prediction.

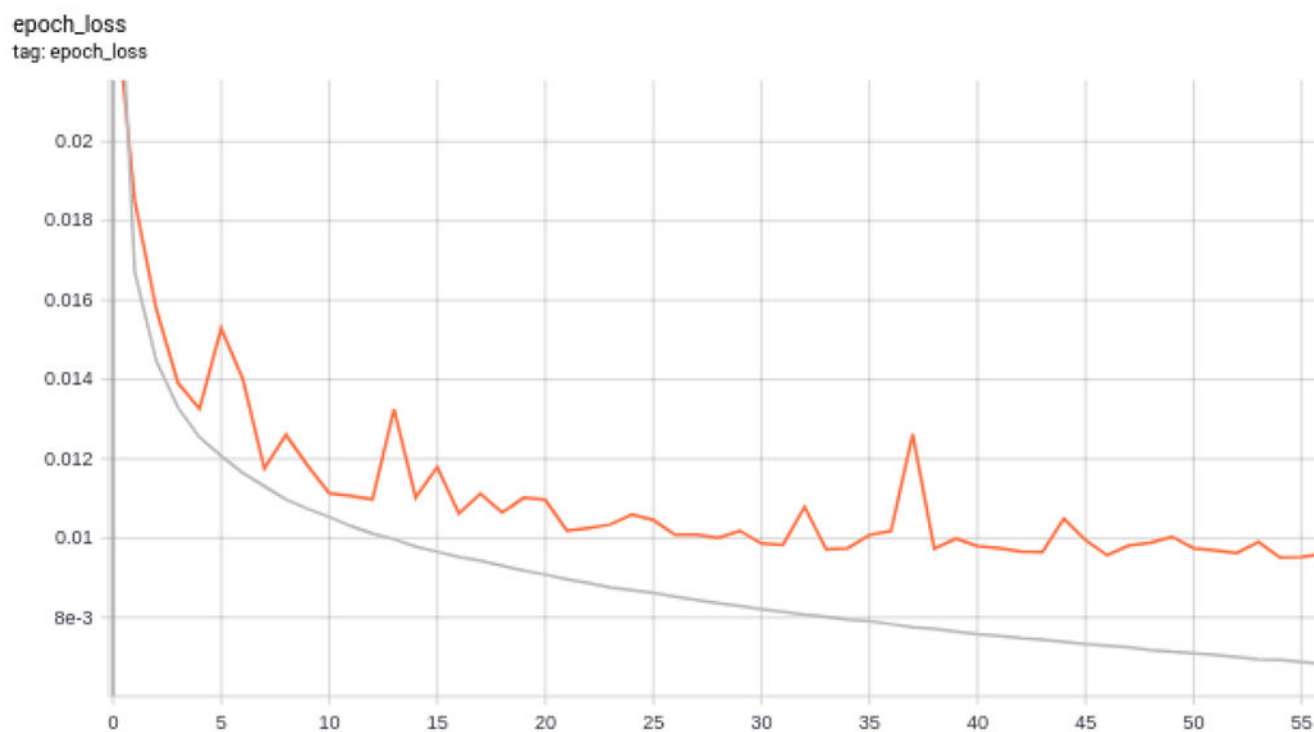


Figure 2. DeepMC training progress for a weighted mean squared error loss function. Per-epoch loss is shown for training dataset (grey) and validation dataset (orange).

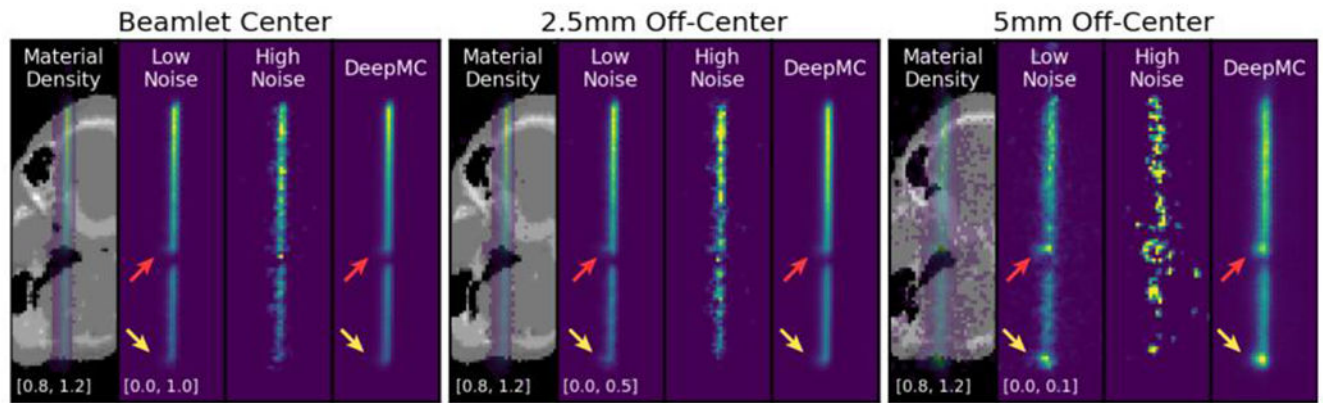


Figure 3.

Dose from low- and high-noise MC simulation and DeepMC prediction for one $5 \times 5 \text{ mm}^2$ X-ray beamlet. Three adjacent slices are shown with their transverse distance from the beamlet's central axis listed in the titles. Color scale limits are displayed in the lower left corner for each slice in g/cm^3 for density and Gy for dose, normalized to the maximum of the low-noise beamlet dose. Arrows show examples of electron return effects asymmetrically perturbing the dose deposition. High-noise simulation fails to accurately estimate dose in these areas while DeepMC dose matches the low-noise ground truth dose.

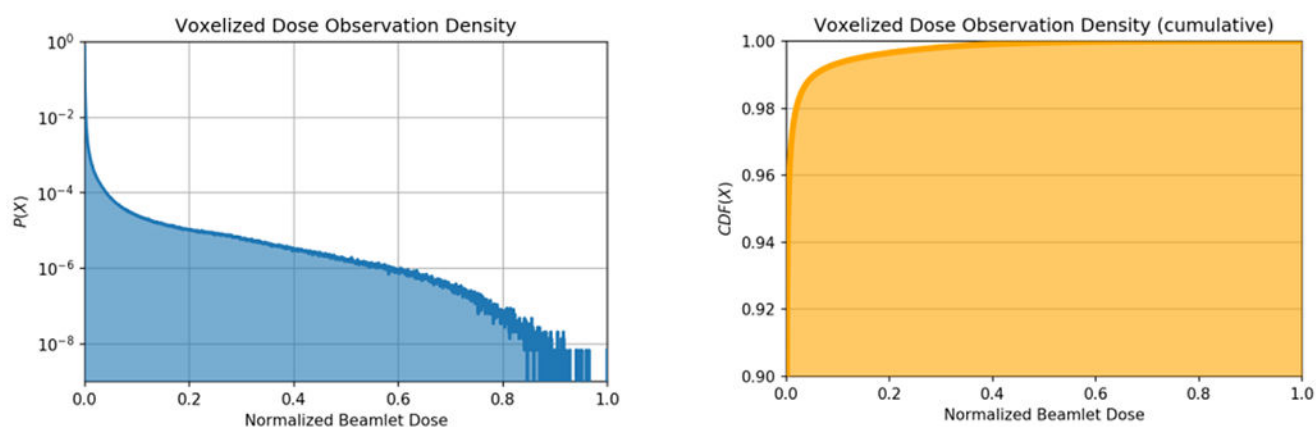
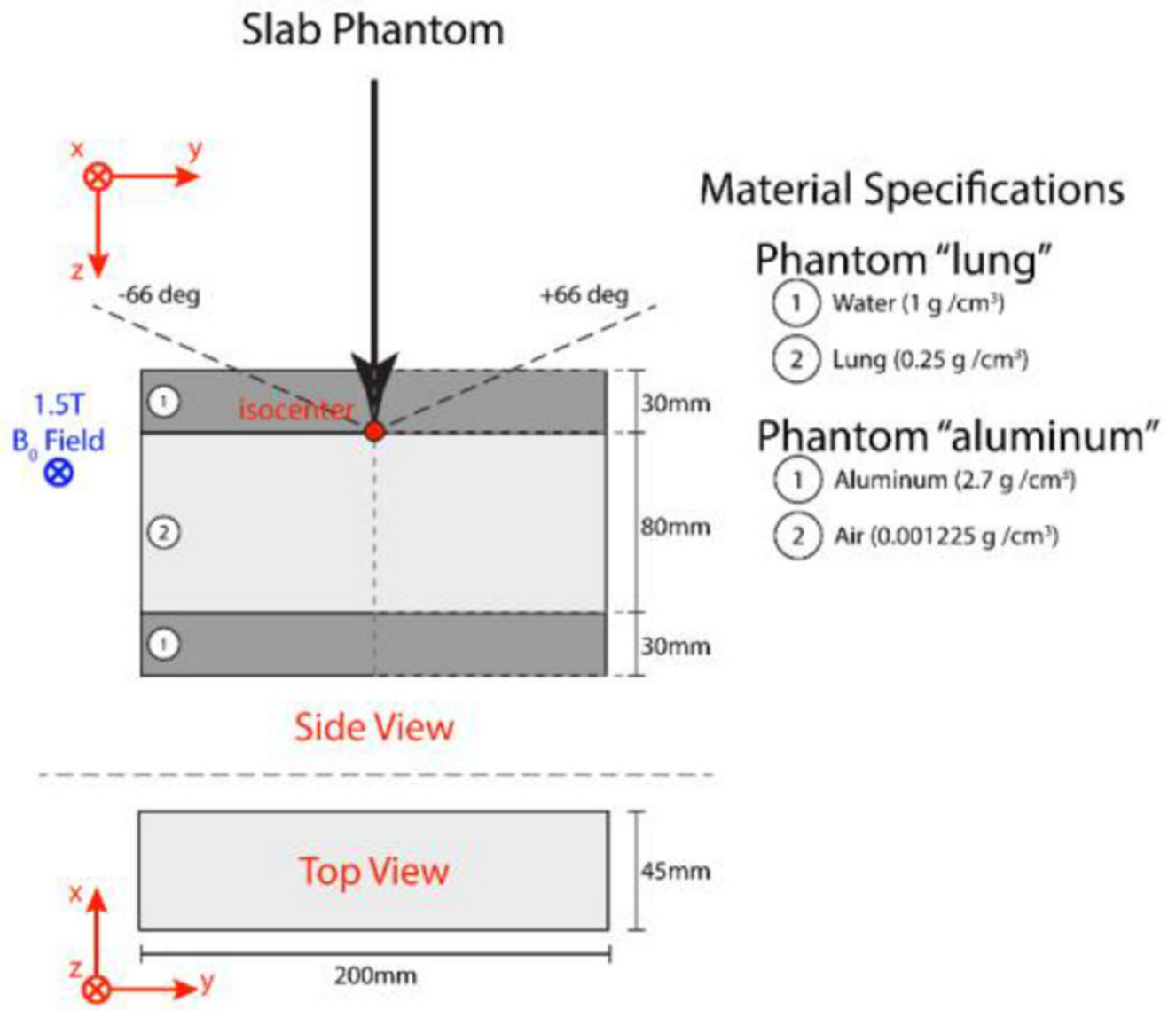


Figure 4. Observed density (left) and cumulative density (right) of voxelized X-ray dose, normalized to per-beamlet maxima. Voxels with dose lower than 10% of per-beamlet maxima account for over 99% of observations. Vertical axes are displayed in log-scale.

**Figure 5.**

Slab phantom geometry specification for studying electron return effects present at high-density-gradient tissue interfaces. Two variants are studied: material 1 is either water or aluminum, and material 2 is lung or air, respectively. The 1.5T magnetic field is directed "into the page". The X-ray beamlet impinges at a variable angle between -66 and 66 degrees, rotating around a vector that is parallel to the magnetic field direction and positioned at the first interface.

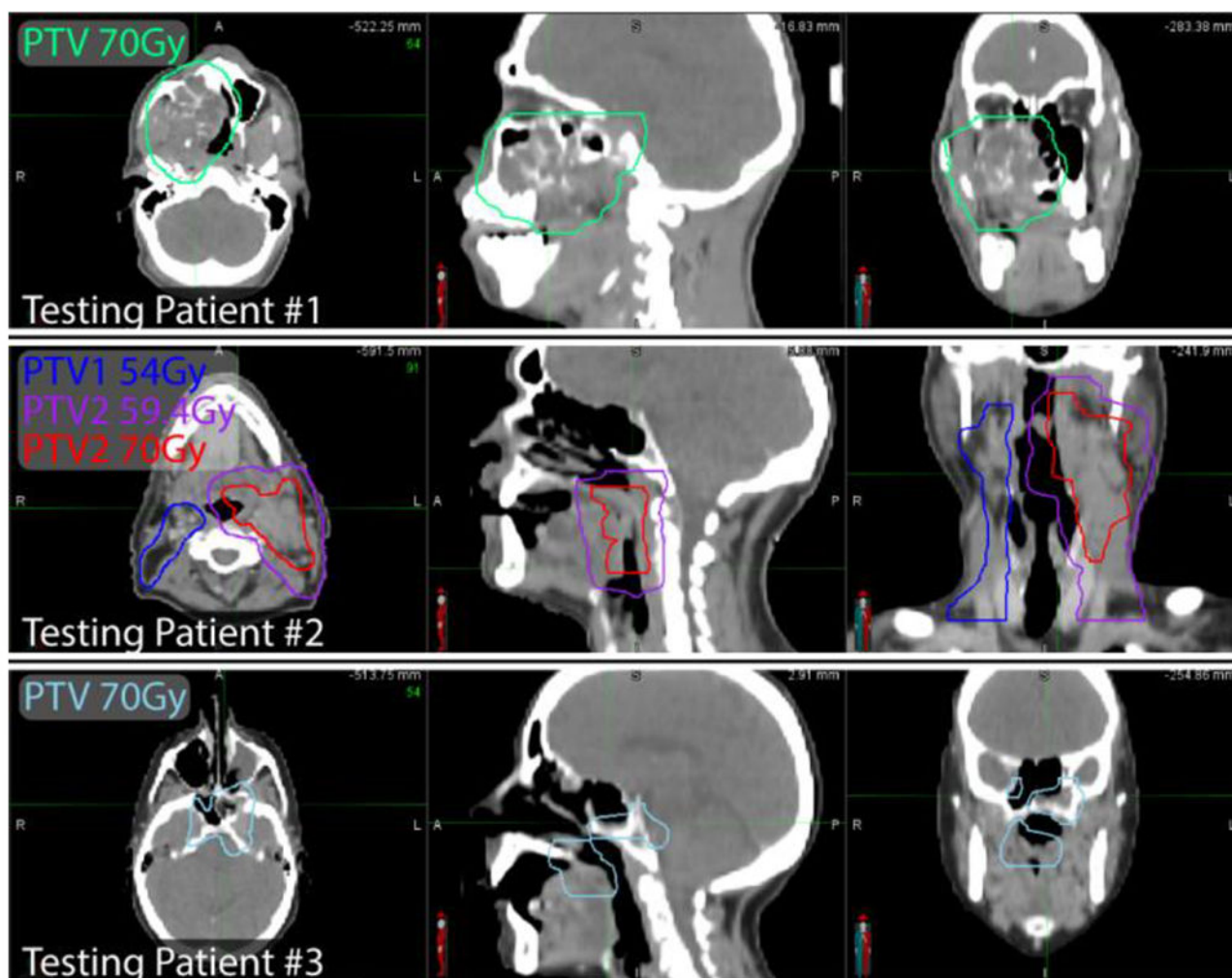


Figure 6.
Orthogonal CT image views for the three testing H&N patients. PTV contours are displayed as colored curves.

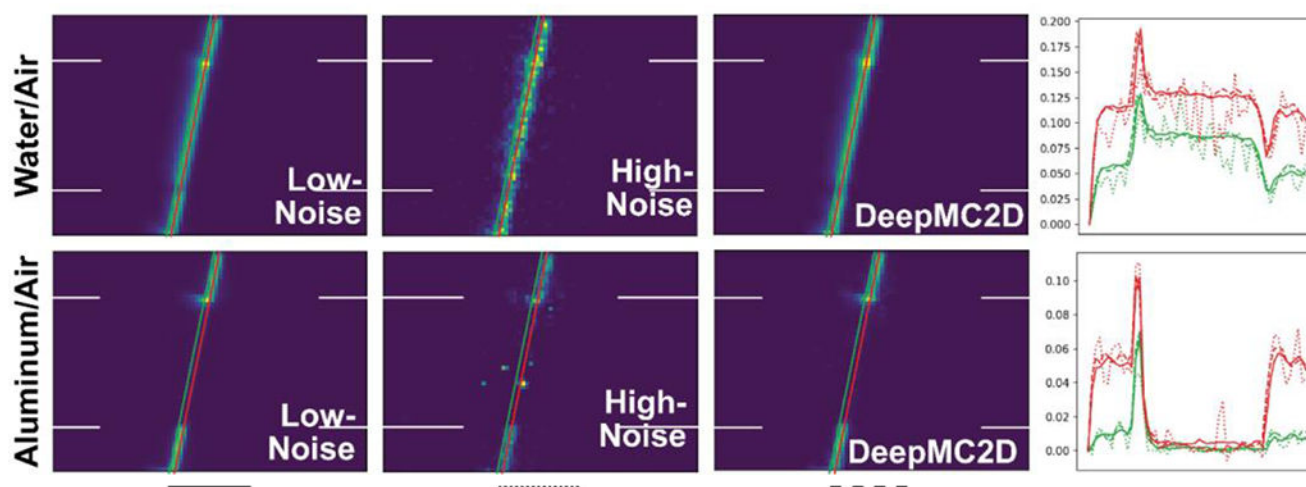
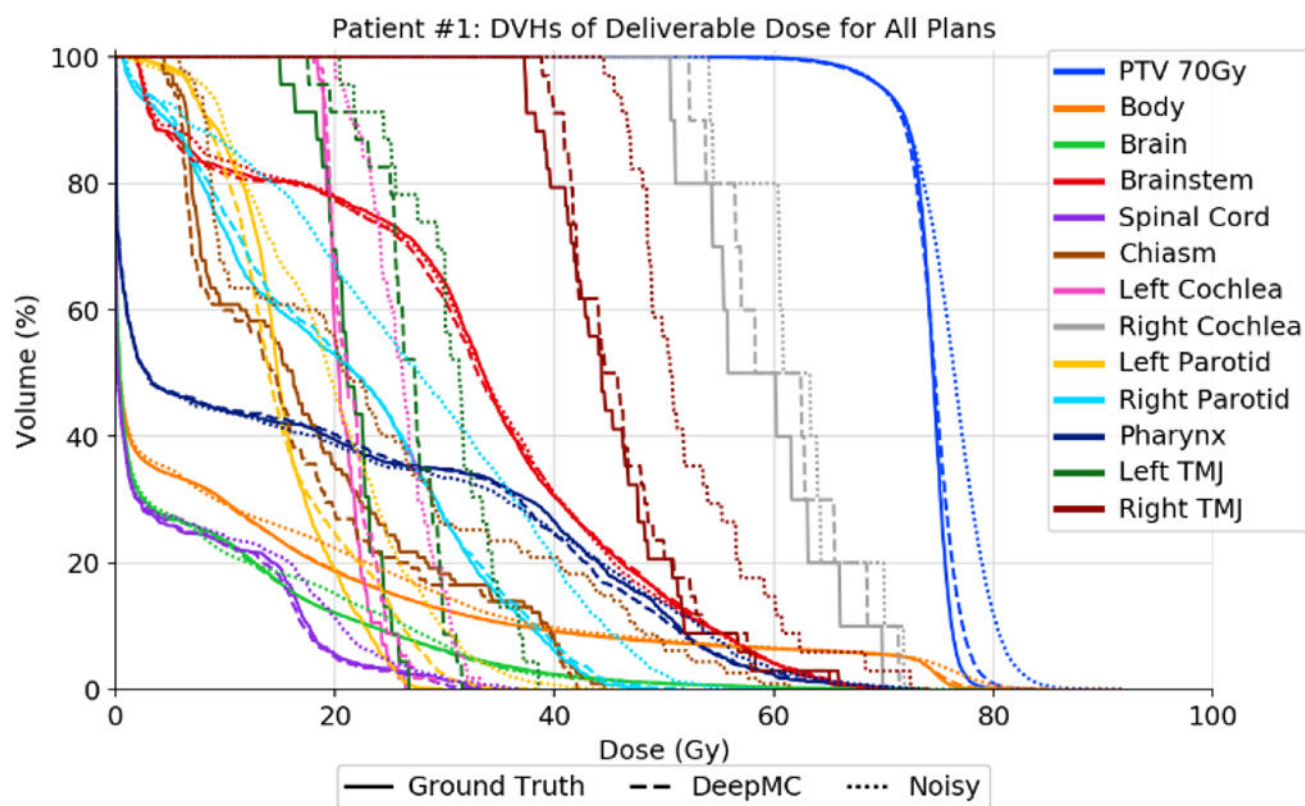


Figure 7.

MC ground truth (low-noise.) dose, model input (high-noise) and DeepMC prediction for an X-ray beamlet in the low dose penumbra region (2.5mm off-axis) of water/air (left) and aluminum/air (right) slab phantoms. White horizontal lines mark the positions of material interfaces. The Right-most column shows dose vs. depth line profiles along the identically colored lines laid over the dose displays in the first three columns. The line styles match with the those provided below each column.



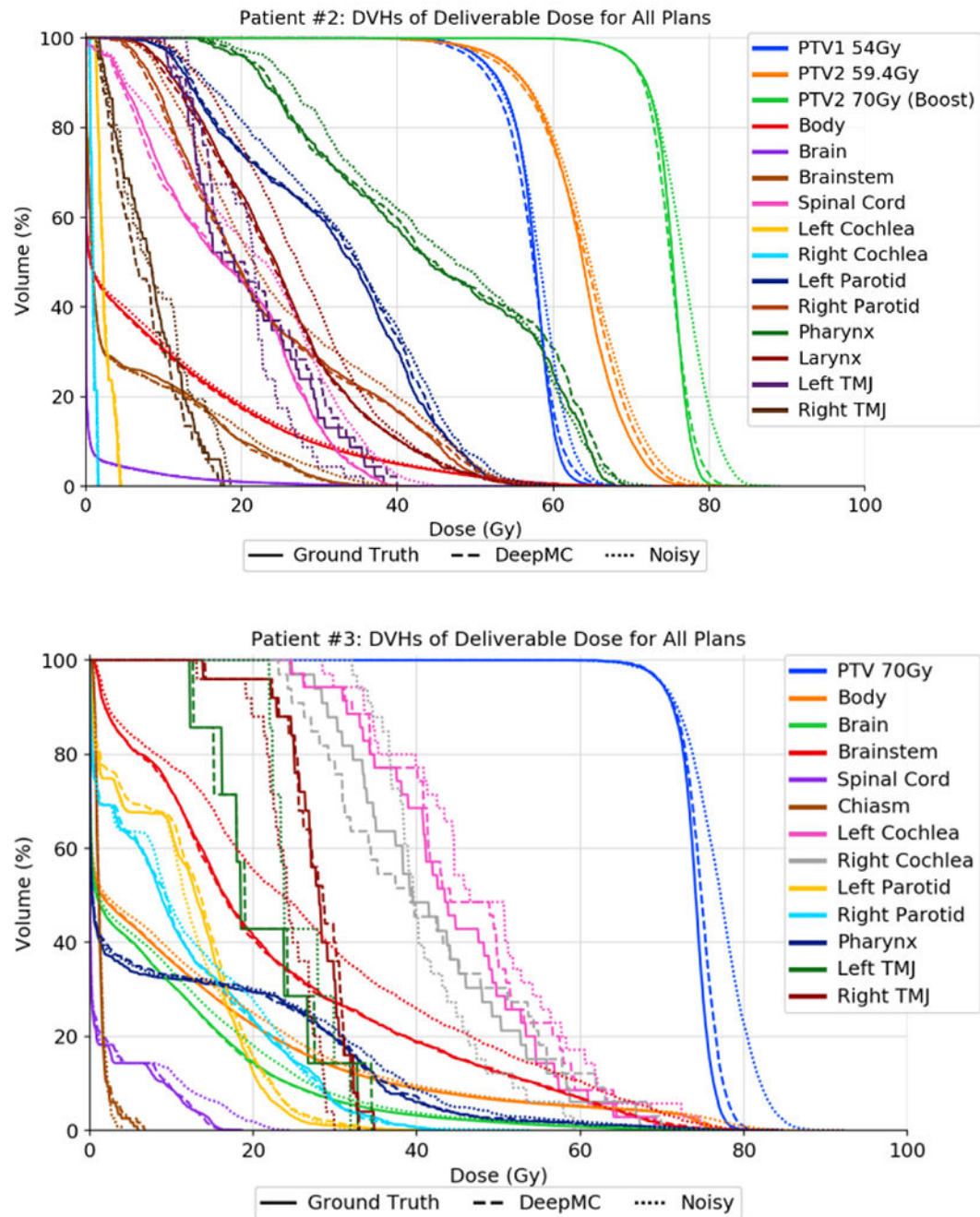
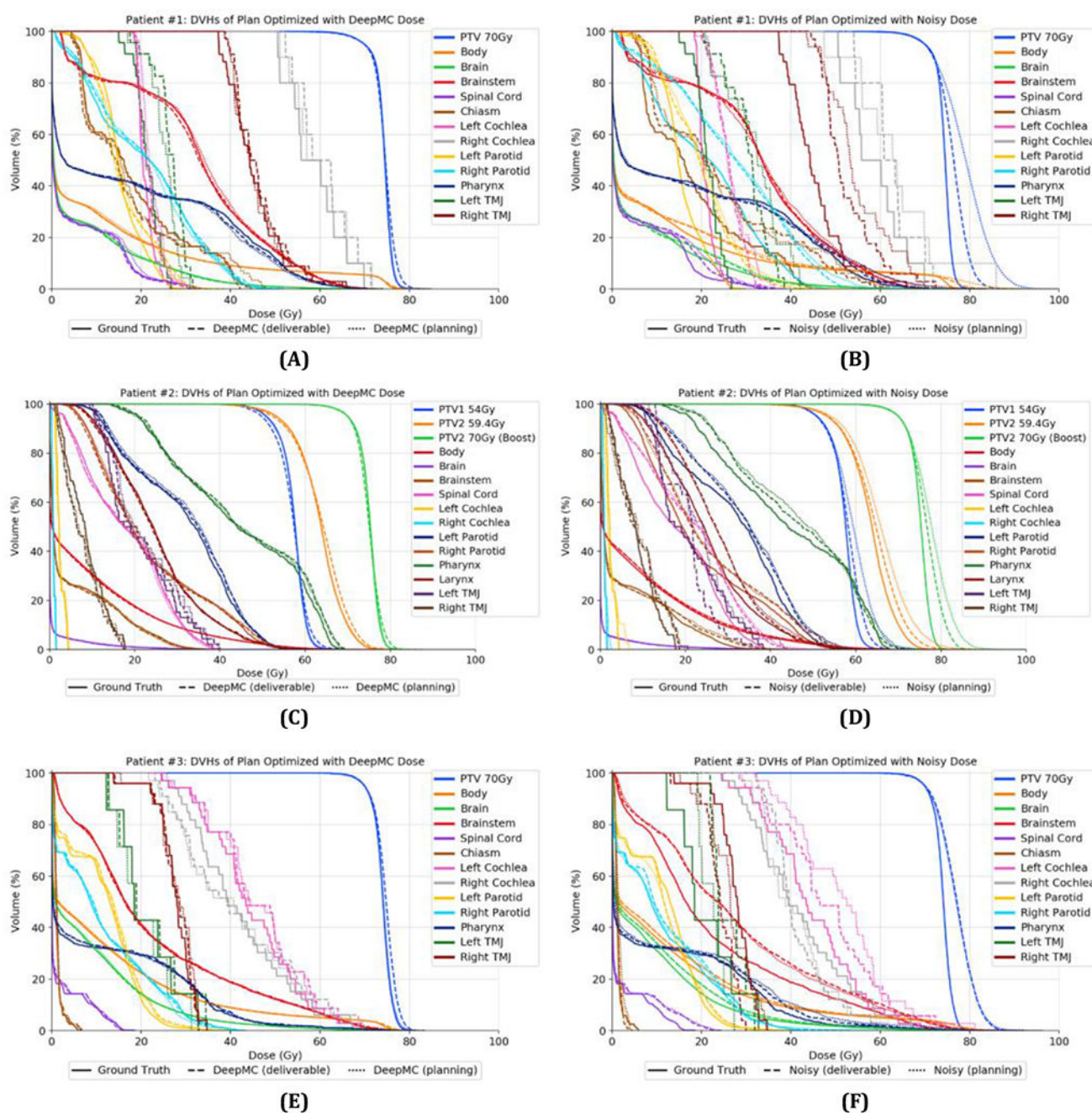


Figure 8.

Dose volume histogram comparing “deliverable” dose for IMRT treatment plans created using low-noise ground truth, DeepMC-predicted, and high-noise beamlet dose for testing patient one (top), two (middle), and three (bottom). Doses for all plans are recalculated after plan optimization using low-noise beamlet dose to reflect the deliverable dose to each patient.

**Figure 9.**

Dose volume histograms for treatment plans created using DeepMC-predicted (A, C, and E) and high-noise MC-simulated (B, D, and F) beamlet dose. Ground truth plans are created using low-noise MC-simulated dose. Dose for “planning” curves is calculated using each plan’s respective beamlet dose after plan optimization. Dose for “deliverable” curves is recalculated using low-noise dose for more accurate, but more computationally expensive plan quality evaluation.

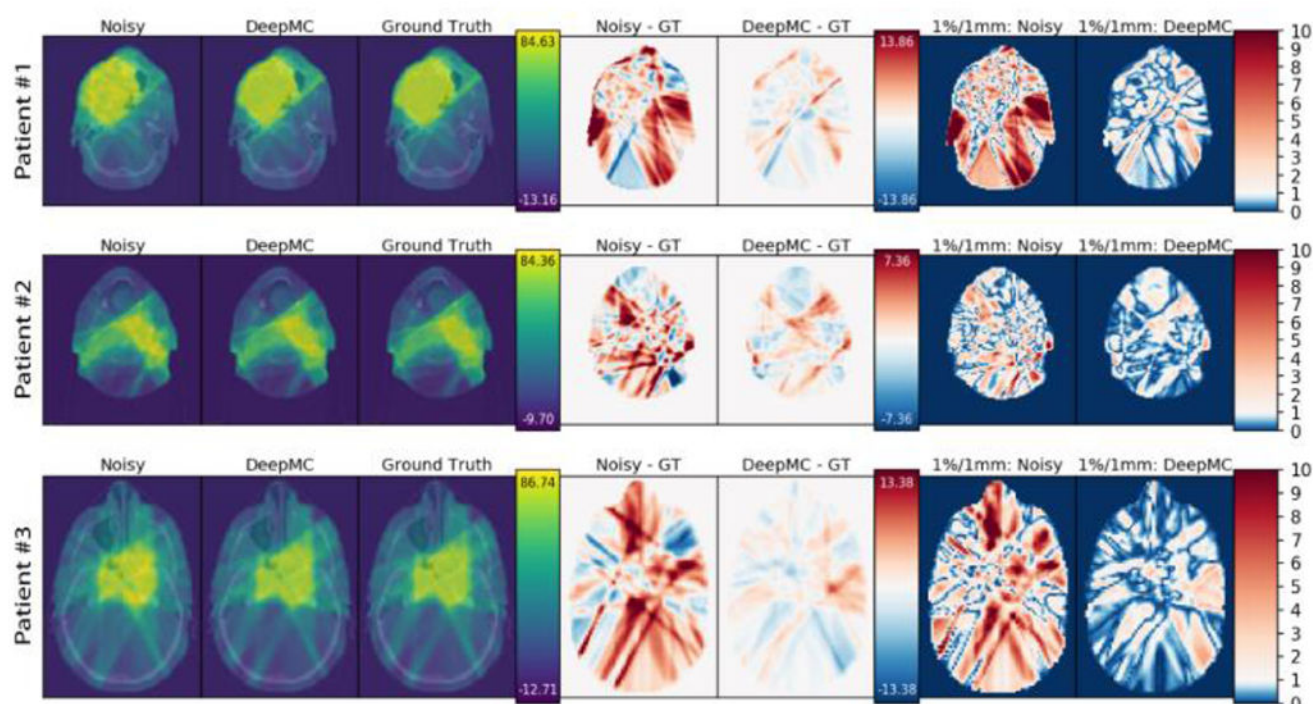
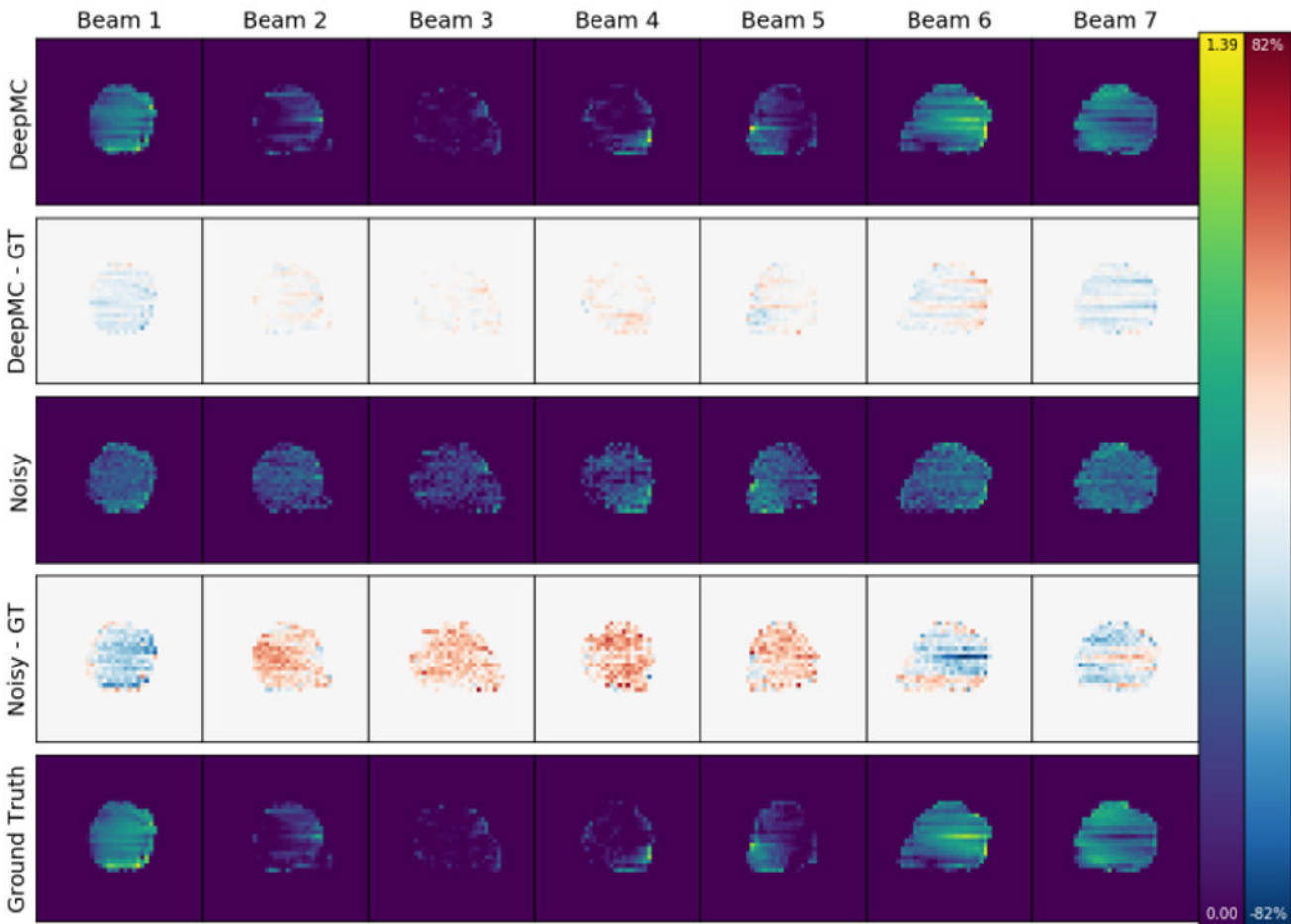


Figure 10.

Deliverable dose color washes for one axial slice through the PTVs of the first (top), second (middle), and third (bottom) testing patient. Deliverable dose for each plan is recalculated using low-noise beamlet dose after plan optimization. The center two columns show differences in deliverable dose attributed to using either high-noise or DeepMC-predicted dose approximations to optimize IMRT beamlet fluence. The last two columns show gamma index maps for each plan compared to the ground truth, evaluated using the 1%/1mm gamma criteria. Colors scales are consistent for each row; scale limits shown on the color bar in absolute dose units of Gy, except for the right-most color bar, which shows “passing” voxels as blue/white (gamma index less than 1.0) and “failing” voxels as red (gamma index greater than 1.0).



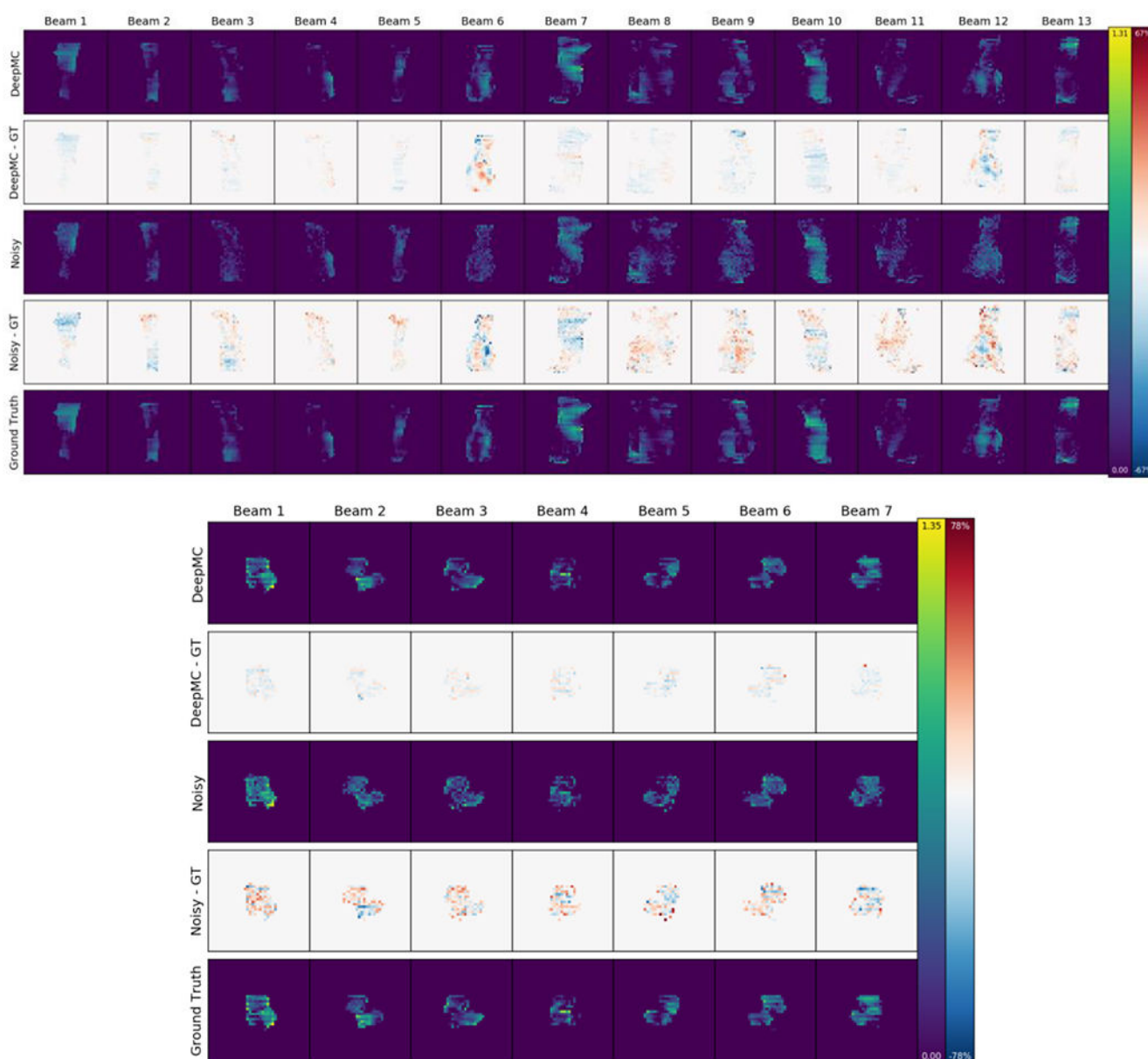


Figure 11.

Plan delivery parameters (X-ray beamlet fluence) for the first (top), second (middle), and third (bottom) testing patient resulting from optimization using DeepMC, high-, and low-noise beamlet dose. All color scales are consistent, and limits are shown in the color bars on the right. All are normalized to the per-beam maximum fluence from the ground truth plan.

Table 1.

Plan quality metrics derived from deliverable dose to the first testing patient for plans optimized using different beamlet dose calculation methods. Absolute dose is presented in units of Gy. Plans with best agreement to ground truth for each metric and structure are shaded green.

Structure	D_{mean}			D_{max}			PTV Dose Metrics			
	GT	Noisy	DeepMC	GT	Noisy	DeepMC	Metric	GT	Noisy	DeepMC
PTV (70Gy)	74.19	76.20	74.43	83.39	91.90	84.87	<i>CN</i>	0.79	0.78	0.79
Body	10.89	11.74	10.92	83.39	91.90	84.87	<i>HI</i>	0.15	0.23	0.18
Brain	6.24	6.53	6.16	73.77	79.40	74.06	<i>D1</i>	78.08	83.98	79.94
Brainstem	32.10	32.18	31.89	70.56	70.43	74.06	<i>D2</i>	77.54	82.97	79.15
Spinal Cord	4.81	5.45	4.69	34.75	36.64	33.01	<i>D98</i>	67.14	67.16	66.89
Chiasm	18.09	23.75	16.79	44.57	61.41	42.06	<i>D99</i>	65.14	65.29	64.70
R TMJ	45.45	52.61	46.37	65.90	72.49	65.13				
L TMJ	21.46	30.56	26.55	26.72	38.53	31.63				
R parotid	20.75	26.98	21.24	48.69	60.00	51.82				
L parotid	15.32	19.51	15.73	29.85	45.88	38.52				
R cochlea	58.80	62.37	60.86	69.93	71.81	71.39				
L cochlea	20.95	26.14	21.66	26.39	33.24	29.26				
Pharynx	18.61	18.56	18.49	70.88	73.21	72.47				

Table 2.

Plan quality metrics derived from deliverable dose to the second testing patient for plans optimized using different beamlet dose calculation methods. Absolute dose is presented in units of Gy. Plans with best agreement to ground truth for each metric and structure are shaded green.

Structure	D_{mean}			D_{max}			PTV Dose Metrics				
	GT	Noisy	DeepMC	GT	Noisy	DeepMC	Metric	GT	Noisy	DeepMC	
PTV #1 (54Gy)	57.17	57.91	56.96	72.77	71.06	72.86	54 Gy	CN	0.19	0.18	0.17
PTV #2 (59.4Gy)	63.34	64.35	63.72	80.80	85.56	85.05		HI	0.26	0.31	0.31
PTV #2 (70Gy)	74.96	76.54	74.90	83.43	89.35	87.60		D2	62.40	64.88	63.76
Body	8.82	9.16	8.75	77.75	78.72	81.86		D98	48.27	48.22	46.87
Brain	0.76	0.76	0.76	56.60	58.62	56.90	59.4 Gy	CN	0.41	0.40	0.39
Brainstem	5.52	5.79	5.35	40.49	42.22	38.54		HI	0.41	0.45	0.44
Spinal Cord	17.96	21.29	17.87	40.14	45.68	40.48		D2	73.63	76.20	74.66
R TMJ	8.44	8.97	7.50	17.70	18.56	17.39		D98	49.37	49.22	48.33
L TMJ	20.89	20.56	22.04	38.30	35.64	39.96					
R parotid	23.09	25.36	22.93	58.95	63.06	58.71	70 Gy	CN	0.78	0.70	0.73
L parotid	31.47	33.31	31.96	59.93	61.87	58.73		HI	0.17	0.23	0.19
R cochlea	0.99	1.05	0.95	1.63	1.72	1.56		D2	79.04	83.57	80.32
L cochlea	2.46	2.49	2.53	4.41	4.39	4.55		D98	67.29	67.41	67.34
Larynx	25.10	27.74	24.80	63.23	62.45	61.45					
Pharynx	44.53	46.19	45.30	67.97	72.53	69.79					

Table 3.

Plan quality metrics derived from deliverable dose to the third testing patient for plans optimized using different beamlet dose calculation methods. Absolute dose is presented in units of Gy. Plans with best agreement to ground truth for each metric and structure are shaded green.

Structure	D_{mean}			D_{max}			PTV Dose Metrics			
	GT	Noisy	DeepMC	GT	Noisy	DeepMC	Metric	GT	Noisy	DeepMC
PTV (70Gy)	73.92	76.99	74.60	81.83	92.36	83.49	<i>CN</i>	0.88	0.84	0.86
Body	12.14	13.05	12.26	81.83	92.36	83.49	<i>HI</i>	0.14	0.24	0.16
Brain	8.19	9.13	8.17	77.69	83.90	79.84	<i>D1</i>	78.28	85.75	79.91
Brainstem	22.55	27.61	22.47	76.16	83.80	77.80	<i>D2</i>	77.75	84.87	79.19
Spinal Cord	1.78	2.32	1.89	18.62	24.38	17.70	<i>D98</i>	68.01	68.03	68.16
Chiasm	1.34	1.31	1.41	5.91	3.93	6.73	<i>D99</i>	66.54	66.14	66.89
R TMJ	27.67	24.24	27.80	34.77	29.92	32.93				
L TMJ	21.14	25.89	21.57	32.75	31.91	34.49				
R parotid	11.30	11.91	11.31	43.19	53.86	43.59				
L parotid	11.41	11.19	12.06	40.11	32.79	39.08				
R cochlea	42.11	41.55	41.55	67.31	59.25	69.86				
L cochlea	44.61	48.36	46.06	69.05	74.65	69.26				
Pharynx	11.14	11.97	11.37	79.65	84.66	79.66				

Table 4.

3D Gamma passing rates (as percent) for varied gamma criteria. A gamma criteria consists of distance-to-agreement and dose difference parameters to quantify the agreement of voxelized dose between two dose fields. All plans have been recalculated using ground truth dose matrices applied to their respective optimized fluence maps (i.e. they are considered “deliverable” dose).

Patient	Comparison	Unmasked			Masked at 1% of GT Dmax		
		2mm/2%	1mm/1%	0.5mm/0.5%	2mm/2%	1mm/1%	0.5mm/0.5%
#1	GT / DeepMC	98.94	97.29	95.67	88.00	69.48	51.28
	GT / Noisy	96.18	94.93	94.00	57.49	43.58	33.29
#2	GT / DeepMC	98.49	95.90	93.08	88.63	69.08	47.86
	GT / Noisy	96.17	93.23	90.97	71.26	49.21	32.18
#3	GT / DeepMC	97.31	91.85	85.09	91.49	74.24	52.86
	GT / Noisy	87.29	80.79	76.06	60.60	40.42	25.77