Title
Accelerating Radiation Dose Calculation with High Performance Computing and Machine Learning for Large-scale Radiotherapy Treatment Planning

Author
Neph, Ryan

Publication Date
2020

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Accelerating Radiation Dose Calculation with High Performance Computing and Machine Learning for Large-scale Radiotherapy Treatment Planning

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Physics and Biology in Medicine

by

Ryan Thomas Neph

2020

ABSTRACT OF THE DISSERTATION

Accelerating Radiation Dose Calculation with High Performance Computing and Machine Learning for Large-scale Radiotherapy Treatment Planning

by

Ryan Thomas Neph

Doctor of Philosophy in Physics and Biology in Medicine

University of California, Los Angeles, 2020

Professor Ke Sheng, Chair

Radiation therapy is powered by modern techniques in precise planning and execution of radiation delivery, which are being rapidly improved to maximize its benefit to cancer patients. In the last decade, radiotherapy experienced the introduction of advanced methods for automatic beam orientation optimization, real-time tumor tracking, daily plan adaptation, and many others, which improve the radiation delivery precision, planning ease and reproducibility, and treatment efficacy. However, such advanced paradigms necessitate the calculation of orders of magnitude more causal dose deposition data, increasing the time requirement of all pre-planning dose calculation. Principles of high-performance computing and machine learning were applied to address the insufficient speeds of widely-used dose calculation algorithms to facilitate translation of these advanced treatment paradigms into clinical practice.

To accelerate CT-guided X-ray therapies, Collapsed-Cone Convolution-Superposition (CCCS), a state-of-the-art analytical dose calculation algorithm, was accelerated through its novel implementation on highly parallelized GPUs. This *context-based* GPU-CCCS approach takes advantage of X-ray dose deposition compactness to parallelize calculation across hundreds of beamlets, reducing hardware-specific overheads, and enabling acceleration by two to three orders of magnitude compared to existing GPU-based beamlet-by-beamlet approaches. Near-linear increases in acceleration are achieved with a distributed, multi-GPU implementation of context-based GPU-CCCS.

Dose calculation for MR-guided treatment is complicated by electron return effects (EREs), exhibited by ionizing electrons in the strong magnetic field of the MRI scanner. EREs necessitate the use of much slower Monte Carlo (MC) dose calculation, limiting the clinical application of advanced treatment paradigms due to time restrictions. An automatically distributed framework for very-large-scale MC dose calculation was developed, granting linear scaling of dose calculation speed with the number of utilized computational cores. It was then harnessed to efficiently generate a large dataset of paired high- and low-noise MC doses in a 1.5 tesla magnetic field, which were used to train a novel deep convolutional neural network (CNN), DeepMC, to predict low-noise dose from faster high-noise MC-simulation. DeepMC enables 38-fold acceleration of MR-guided X-ray beamlet dose calculation, while remaining synergistic with existing MC acceleration techniques to achieve multiplicative speed improvements.

This work redefines the expectation of X-ray dose calculation speed, making it possible to apply new highly-beneficial treatment paradigms to standard clinical practice for the first time.

The dissertation of Ryan Thomas Neph is approved.

Dan Ruan

James Michael Lamb

Xun Jia

You Ming Yang

Ke Sheng, Committee Chair

University of California, Los Angeles

2020

To my parents, my wife Ashley, and my children Graham and Wesley.

TABLE OF CONTENTS

List of Tables x
List of Figures xi
List of Equations xvi
Acknowledgements xvii
Vita xix
1 INTRODUCTION 1
2 OVERVIEW OF RADIATION THERAPY 3
2.1 Radiation Dose Deposition Mechanisms 3
2.1.1 Electromagnetic Radiation 5
2.1.1.1 The Photoelectric Effect 5
2.1.1.2 The Compton Effect 6
2.1.1.3 Pair Production 7
2.1.2 Particulate Radiation 8
2.1.2.1 Electrons 10
2.1.2.2 Protons and Heavy Ions 10
2.1.2.3 Electromagnetic Field Effects 11
2.2 Dose Calculation Algorithms 12
2.2.1 Monte Carlo Dose Calculation 13
2.2.2 Analytical Dose Calculation 14
2.2.3 Linearized Boltzmann Solvers 15
2.3 Radiation Treatment Planning 16
2.3.1 Beamlet Dose Calculation 17
2.3.2 Inverse Treatment Planning 18
2.3.3 Final Dose Calculation 19
2.3.4 Online Adaptive Radiotherapy 20
3 PARALLEL BEAMLET DOSE CALCULATION VIA BEAMLET CONTEXTS IN A DISTRIBUTED MULTI-GPU FRAMEWORK[34] 21
3.1 Introduction 21
3.2 Materials and Methods 24

3.2.1 Nonvoxel-Based Dose Calculation 25
3.2.1.1 Beamlet-based Dose by Intra-beam Parallelization 25
3.2.1.1.1 TERMA Calculation 25
3.2.1.1.2 Beamlet Context Extraction 28
3.2.1.1.3 Nonvoxel-Based Transformation 32
3.2.1.1.4 Dose Ray Convolution 34
3.2.1.1.5 Beamlet Context Dose Extraction 35
3.2.1.2 Distributed Parallelization 35
3.2.2 Measuring Computational Efficiency 36
3.2.3 Measuring Dosimetric Accuracy 37
3.3 Results 40
3.4 Discussion 46
3.4.1 Performance 46
3.4.2 Accuracy 50
3.5 Applications 53
3.5.1 A sparse orthogonal collimator for small animal intensity-modulated radiation therapy[79,80] 54
3.5.1.1 Background 54
3.5.1.2 Methods 55
3.5.1.3 Results and Discussion 58
3.5.2 A novel optimization framework for VMAT with dynamic gantry couch rotation[40] 61
3.5.2.1 Background 61
3.5.2.2 Methods 62
3.5.2.3 Results and Conclusions 64
3.5.3 Single-Arc VMAT optimization for Dual-Layer MLC[44] 65
3.5.3.1 Background 65
3.5.3.2 Methods 65
3.5.3.3 Results and Conclusions 67
3.5.4 Many Isocenter Optimization for Robotic Radiotherapy[43] 69

3.5.4.1 Background 69
3.5.4.2 Methods 70
3.5.4.3 Results and Conclusions 72
3.6 Conclusions 74
4 A HIGH-PERFORMANCE DISTRIBUTED FRAMEWORK FOR LARGE-SCALE MONTE CARLO DOSE CALCULATION 76
4.1 Introduction 76
4.2 Implementation 78
4.2.1 Distributed Computation Model 78
4.2.2 Orchestration and Scalability 83
4.2.3 SimpleDose: A High-Level Treatment Planning Interface 84
4.3 Usage Examples 85
4.4 Performance 88
4.5 Conclusions 89
5 DEEPMC: A DEEP LEARNING METHOD FOR EFFICIENT MONTE CARLO BEAMLET DOSE CALCULATION BY PREDICTIVE DENOISING IN MAGNETIC RESONANCE-GUIDED RADIOTHERAPY 90
5.1 Introduction 90
5.2 Preliminary Investigation in Stacked Slab Phantom Geometries 93
5.3 Methods 96
5.3.1 DeepMC Model Architecture 96
5.3.2 Dose Data Generation 100
5.3.3 Experiment Design 103
5.4 Results 106
5.5 Discussion 112
5.6 Conclusions 114
5.7 Appendices 116
5.7.1 Appendix A. Dose Volume Histograms 116
5.7.2 Appendix B. Plan Quality Metrics 117
5.7.3 Appendix C. 3D Dose Comparison 117

5.7.4 Appendix D. Fluence Map Comparison 119
6 SUMMARY OF WORK 120
7 REFERENCES 123

LIST OF TABLES

Table 3-1. Per-beam Calculation Times (average, in seconds) 41

Table 3-2. Peak Memory Usage For GPU-based CCCS Methods (in Megabytes) 43

Table 3-3. Comparisons between the measured and intended dose distributions for the C-shaped target plan and the mouse phantom whole liver plan. Table reproduced from Woods *et al.* (2019)[80]. 58

Table 3-4. Number of feasible beams, prescription doses and PTV volumes for all patients. Table reproduced from Lyu *et al.* (2020)[43]. 71

Table 5-1. Plan quality metrics derived from deliverable dose to the first testing patient for plans optimized using different beamlet dose calculation methods. 108

Table 5-B1. Plan quality metrics derived from deliverable dose to the second testing patient for plans optimized using different beamlet dose calculation methods. 117

LIST OF FIGURES

Figure 3-1. (a) Intersection map for one beam orientation and target volume definition. Purple-colored cells have no target intersection and are excluded from beamlet-dose calculation, yellow: full intersection, others: partial intersection. (b) Super-sampling ray layout for testing beamlet-target intersection. (c) cross-section of TERMA calculation sub-voxel arrangement for super-sampled averaging (2x and 3x options shown for one voxel). 26

Figure 3-2. (a) Beamlet dose calculation workflow. A single worker node processes beams in parallel across its resident GPU devices. Beamlet processing is further parallelized on a GPU using beamlet contexts. (b) Distributed computing framework. The manager node prepares independent task lists for each worker node to process in parallel and receives the results for delivery to the requestor. 28

Figure 3-3. (a) Visualization of the beamlet context array for a single beam including contextual densities and beamlet-specific dose after calculation. (b) Beamlet context cross sections for various context radii with dose overlaid. (c) Convolution-ray-aligned context array (cross-section) for various kernel rays. Grey area is allocated once and reused for all beams. White subregions are allotted for kernel-ray-specific convolutions geometry. Black cells indicate unused space after packing beamlet contexts into the array. Convolution direction is into page. 31

Figure 3-4. Construction of one beamlet context with implicit kernel tilting. Blue region indicates the volume of non-zero TERMA for a single beamlet. The distance between the blue rings in the transverse view is representative of the context radius setting. The union of red and gold boxes represents the volume in which dose is computed. 32

Figure 3-5. Cross sections of phantom geometries with beam entering from the top; used to assess dosimetric accuracy. Materials and densities are provided. 38

Figure 3-6. Performance for single-node and multi-node scaling strategies. For multi-node measurements, each node was configured with 2 GPUs. The dotted black line indicates theoretical linear scaling in multi-node setups. 41

Figure 3-7. Fractional execution time spent in each sub-procedure on one computational node with 4 threaded post-processing "SparseAgents". Only time spent on the main processing thread is represented. 42

Figure 3-8. Fractional execution time for 1cm context radius with a variable number of background post-processing (dose sparsification) threads. Only time spent on the main processing thread is represented. 42

Figure 3-9. Peak memory usage for various beamlet counts and context radii. 43

Figure 3-10. Single-beamlet depth dose and lateral profiles in the water phantom for increasing beamlet widths. Error is calculated between our context-based GPU-CCCS method and each of CPU-CCCS and Monte Carlo. 44

Figure 3-11. Single-beamlet depth dose and lateral profiles in the stack of slabs phantom for increasing beamlet widths. Error is calculated between our context-based GPU-CCCS method and each of CPU-CCCS and Monte Carlo. 45

Figure 3-12. Central lateral line profile in the water phantom at 10cm depth for various beamlet widths and context radii pairs (top). Maximum errors (%) between non-context-based (infinite radius) and context-based dose profiles are provided (bottom). The dose is normalized to the maximum dose in the volume. Y-axis range is limited to better depict low dose beamlet penumbra region where context-based approximation is active. 46

Figure 3-13. Lateral line profile in the water phantom at various depths for a $5\times5cm^2$ broad beam calculated as the sum of $5\times5mm^2$ beamlets for various context radii at three depths. Dose for "infinite" radius was computed without context-based approximation. 53

Figure 3-14. (Left) Mouse phantom modeled from mouse CT data and 3D-printed with a flexible, tissue-equivalent material and a mid-coronal split for film measurement. Phantom is shown on the previously mentioned rotating couch mount. (Right) 3D-printed block phantom for axial dose measurements. Figure reproduced from Woods *et al.* (2019)[80]. 57

Figure 3-15. (Left) Calculated dose distribution of the C-shaped target plan perpendicular to the gantry rotation axis. (Center) Measured film dose distribution from the center of the solid water phantom for the C target plan delivered with the SOC. Both plans are shown with the same color scale, in units of Gy. (Right) A comparison of the calculated (yellow) and measured (blue) 50% isodose lines, with overlapping regions shown in red. Figure reproduced from Woods *et al.* (2019)[80]. 58

Figure 3-16. (A) Mid-coronal view of the calculated dose for the mouse phantom whole liver plan (units of Gy). (B) The 5 optimal coplanar beam angles selected with the 4π algorithm. (C) Measured film dose from the mouse phantom, treated with the whole liver plan, at the plane shown in A (units of Gy). (D) A comparison of the calculated (yellow) and measured (blue) 60% isodose lines, with overlapping regions shown in red. *Target structure was rotated to account for slight phantom misalignment, which also resulted in the truncated lower left portion of the target. Figure reproduced from Woods *et al.* (2019)[80]. 60

Figure 3-17. (Left) Calculated Audrey test plan with 4 dose levels and an average aperture size of 2.35 mm. (Right) Measured dose distribution of the Audrey plan delivered with

the SOC. Both plans are shown with the same color scale, in units of Gy. Figure reproduced from Woods *et al.* (2019)[80]. 61

Figure 3-18. Demonstration of (A) DLMLC with 10mm leaf width (DLMLC-10mm), (B) SLMLC with 5mm leaf width (SLMLC-5mm), (C) SLMLC with 10mm leaf width (SLMLC-10mm), (D) SLMLC with 10mm leaf width and 5mm leaf step size (SLMLC-10mm-5mm). The grids on (C) and (D) represent the achievable beamlets. Figure reproduced from Lyu *et al.* (2019)[44]. 66

Figure 3-19. DVH for (A) the GBM case, (B) the LNG case, (C) the PRT case, and (D) the REC-SIB case. The solid lines are for the DLMLC plan, and the dotted lines are for SLMLC plans. D95 is normalized to the prescription dose. Figure reproduced from Lyu *et al.* (2019)[44]. 69

Figure 3-20. (A) Demonstration of the robotic arm platform, (B) an isocentric SID-100 beam that covers the entire target, (C) beams of different isocenters are required to efficiently cover the entire target. Figure reproduced from Lyu *et al.* (2020)[43]. 70

Figure 3-21. Final objective value vs the number of beams. The plot with shaded error bar shows a summary of all patients. Each patient plot is titled with the patient number, the number of isocenters for the SID-50 plan, and the number of isocenters for the SID-100 plan. For example, the first patient plot is entitled: '#1: 4(50), 1(100)', showing that the patient #1 has four isocenters for the SID-50 plan, and one isocenter for the SID-100 plan. Figure reproduced from Lyu *et al.* (2020)[43] 73

Figure 4-1. Database document hierarchy. Simulations are organized as children of a sub-beam (beamlet or spot). Sub-beams are children of a beam. Multiple beams can be defined for a single calculation geometry and multiple geometries can be defined for a CT image acquisition. Images contain masks for every structure defined in the RTStruct file. Simulations can produce one or more independent samples of dose for the same configuration as a method for machine learning dataset augmentation. 81

Figure 4-2. Summary of commands available from the SimpleDose interface for treatment planning dose calculation. 85

Figure 4-3. Create-plan command output using the SimpleDose to add dose calculation requests for automatically distributed computation. 86

Figure 4-4. Plan-status command output using the SimpleDose interface to list the simulation progress for all existing *plans*. 87

Figure 4-5. Sparse storage format for calculated planning dose results for a SimpleDose *plan.* The sparse coordinate list (COO) format uses three equal-length arrays to store the row index, column index, and data value for every non-zero element of the represented matrix. 88

Figure 5-1. Slab phantom geometry specification for studying electron return effects present at high-density-gradient tissue interfaces. 94

Figure 5-2. MC dose ground truth (low-noise.), model input (high-noise) and DeepMCv1 prediction for a photon beamlet in the low dose penumbra region (1.25cm off-axis) of water/air (left) and aluminum/air (right) slab phantoms. Lines indicate positions of horizontal material interfaces. Low-noise dose was simulated using 18M particles in ~3 minutes. High-noise dose was simulated using 30K particles in 3 seconds. Prediction took <100ms after high-noise dose simulation. 95

Figure 5-3. 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. 3×3×3 kernels are used in UNet layers. 1×1×1 kernels are used for feature mixing and dose prediction. 98

Figure 5-4. DeepMC training progress. Per-epoch loss is shown for training dataset (grey) and validation dataset (orange). 98

Figure 5-5. Dose from low- and high-noise MC simulation and DeepMC prediction for one 5x5mm^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 normalized units. 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. 99

Figure 5-6. 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. 99

Figure 5-7. 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 (left) and two (right). Doses for all plans are recalculated after plan optimization using low-noise beamlet dose to reflect the deliverable dose to each patient. 107

Figure 5-8. Deliverable dose color washes for axial slices from the first testing patient. Deliverable dose for each plan is recalculated using low-noise beamlet dose after plan optimization. The last two columns show differences in deliverable dose attributed to using either high-noise or DeepMC-predicted dose approximations to optimize IMRT beamlet fluence. Colors scales are consistent for each row; scale limits shown on the color bar in absolute dose units of Gy. 110

Figure 5-9. Planning and deliverable dose color washes for axial slices from the first testing patient. Planning dose is used directly for plan optimization. Deliverable dose for each plan is recalculated using low-noise beamlet dose after plan optimization. The Last two columns show differences in planning and deliverable dose for each plan. Color scales are consistent for each row; scale limits shown on the color bar are in absolute dose units of Gy. 111

Figure 5-10. Plan delivery parameters (X-ray beamlet fluence) for the first 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. 111

Figures 5-A1 through 5-A4. Dose volume histograms for treatment plans created using DeepMC-predicted (top) and high-noise MC-simulated (bottom) 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. 116

Figure 5-C1. Deliverable dose washes for axial slices from the second testing patient. Deliverable dose for each plan is recalculated using low-noise beamlet dose after plan optimization. The last two columns show differences in deliverable dose attributed to using either high-noise or DeepMC-predicted dose approximations to optimize IMRT beamlet fluence. Colors scales are consistent for each row; scale limits shown on the color bar in absolute dose units of Gy. 117

Figure 5-C2. Planning and deliverable dose color washes for axial slices from the second testing patient. Planning dose is used directly for plan optimization. Deliverable dose for each plan is recalculated using low-noise beamlet dose after plan optimization. The Last two columns show differences in planning and deliverable dose for each plan. Color scales are consistent for each row; scale limits shown on the color bar are in absolute dose units of Gy. 118

Figure 5-D1. Plan delivery parameters (X-ray beamlet fluence) for the second testing patient one resulting from optimization using DeepMC, high-, and low-noise ground truth beamlet dose. All color scales are matched consistent, and limits are shown in the color bars on the right. All are normalized to the per-beam ground truth maximum fluence from the ground truth plan. 119

LIST OF EQUATIONS

Equation 2-1..... 6
Equation 2-2..... 7
Equation 2-3.....11
Equation 2-4.....18
Equation 3-1.....26
Equation 3-2.....27
Equation 3-3.....28
Equation 3-4.....55
Equation 5-1.....100
Equation 5-2.....104

ACKNOWLEDGEMENTS

For his incredible support over these years, I'd like to thank my advisor Dr. Ke Sheng, who gave me the stability and encouragement to pursue research in the topics of machine learning and high performance computing that were of great interest to me. I am extremely grateful for the mentorship provided by Ke and his many collaborators, particularly Dr. Dan Ruan and Dr. Youming Yang for lending a critical mind to improve the quality of my work and help steer me in the right direction.

I'd also like to thank all of the UCLA Physics in Biology in Medicine faculty and staff for devoting their efforts to educate my fellow students and me in the principles and practices of Medical Physics, and make our graduate school experience enjoyable and memorable.

I am lucky to have made such great friends as I have in my research lab, and in the PBM program at UCLA. Thank you to Dan, Victoria, Angelia, Kaley, Daniel, Wenbo, Qihui, Elizabeth, Daili, Jiayi, Pav, Ningning, Yang, Nuo, Lingli, Yi, and Alan for your camaraderie in the trenches of scientific research. Thanks also to Geri, John, Kamal, Nastaran, Jason, Ksenia, and Wenbo for the lasting friendship that all began when we arrived for orientation in the fall of 2015.

Special thanks to my parents for raising me and encouraging me to seek out answers to the questions of the world, my sister for walking ahead of me on the path to higher education, my uncle John for exposing me to science and technology at an early age and fostering my process of logical reasoning. I'd especially like to thank my wife Ashley for tolerating the many challenges of raising two children with me in Los Angeles on a single graduate student's income, and allowing me to focus a significant portion of my time on accomplishing

my academic goals. I could not have accomplished all I have without each and every one of you, and I will be forever in your debt.

Section 3.8.1 is a shortened version of:

Woods, K., Nguyen, D., Neph, R., Ruan, D., O'Connor, D., & Sheng, K. (2019). A sparse orthogonal collimator for small animal intensity-modulated radiation therapy part I: Planning system development and commissioning. Medical Physics, 46(12), 5703–5713. https://doi.org/10.1002/mp.13872

and Woods, K., Neph, R., Nguyen, D., & Sheng, K. (2019). A sparse orthogonal collimator for small animal intensity-modulated radiation therapy. Part II: hardware development and commissioning. Medical Physics, 46(12), 5733–5747. https://doi.org/10.1002/mp.13870.

Section 3.8.2 is a shortened version of:

Lyu, Q., Yu, V. Y., Ruan, D., Neph, R., O'Connor, D., & Sheng, K. (2018). A novel optimization framework for VMAT with dynamic gantry couch rotation. Physics in Medicine & Biology, 63(12), 125013. https://doi.org/10.1088/1361-6560/aac704.

Section 3.8.3 is a shortened version of:

Lyu, Q., Neph, R., Yu, V. Y., Ruan, D., & Sheng, K. (2019). Single-arc VMAT optimization for dual-layer MLC. Physics in Medicine & Biology, 64(9), 095028. https://doi.org/10.1088/1361-6560/ab0ddd.

Section 3.8.4 is a shortened version of:

Lyu, Q., Neph, R., Yu, V. Y., Ruan, D., Boucher, S., & Sheng, K. (2020). Many-isocenter optimization for robotic radiotherapy. Physics in Medicine and Biology, 65(4). https://doi.org/10.1088/1361-6560/ab63b8.

Chapter 3 is a version of a published journal article:

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.

Chapter 5 is a version of a manuscript currently under review for publication:

Neph R, Lyu Q, Huang Y, Yang YM, Sheng K. DeepMC: a deep learning method for efficient monte carlo beamlet dose calculation by predictive denoising in magnetic resonance-guided radiotherapy.

VITA

EDUCATION

M.S.	University of California, Los Angeles, Physics and Biology in Medicine	2019
B.S.	Kettering University, Engineering Physics	2015
B.S.	Kettering University, Mechanical Engineering	2015

AWARDS

Norm Baily Award 1st place prize (AAPM Southern California Chapter)	2019
Science Council's Scientific Session - Selected Speaker (AAPM Annual Meeting)	2019
Best Poster (UCLA Physics and Biology in Medicine Research Colloquium)	2017

PEER-REVIEWED PUBLICATIONS

Shang D, Gu W, Landers A, et al. Technical Note: Robust individual thermoluminescence dosimeter tracking using optical fingerprinting. *Med Phys*. 2020;47(1):267-271. doi:10.1002/mp.13895

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

Woods K, **Neph R**, Nguyen D, Sheng K. A sparse orthogonal collimator for small animal intensity-modulated radiation therapy. Part II: hardware development and commissioning. *Med Phys*. 2019;46(12):5733-5747. doi:10.1002/mp.13870

Woods K, Nguyen D, **Neph R**, Ruan D, O'Connor D, Sheng K. A sparse orthogonal collimator for small animal intensity-modulated radiation therapy part I: Planning system development and commissioning. *Med Phys*. 2019;46(12):5703-5713. doi:10.1002/mp.13872

Neph R, Huang Y, Yang Y, Sheng K. DeepMCDose: A Deep Learning Method for Efficient Monte Carlo Beamlet Dose Calculation by Predictive Denoising in MR-Guided Radiotherapy. *Lecture Notes in Computer Science – Workshop on Artificial Intelligence in Radiation Therapy*. 2019:11850:137-145. doi:10.1007/978-3-030-32486-5_17

Gu W, **Neph R**, Ruan D, Zou W, Dong L, Sheng K. Robust Beam Orientation Optimization for Intensity-Modulated Proton Therapy. *Med Phys*. June 2019:mp.13641. doi:10.1002/mp.13641

Lyu Q, **Neph R**, Yu VY, Ruan D, Sheng K. Single-arc VMAT optimization for dual-layer MLC. *Phys Med Biol*. 2019;64(9):095028. doi:10.1088/1361-6560/ab0ddd

Lyu Q, Ruan D, Hoffman JM, **Neph R**, McNitt-Gray M, Sheng K. Iterative reconstruction for low dose CT using Plug-and-Play alternating direction method of multipliers (ADMM) framework. In: Angelini ED, Landman BA, eds. *Medical Imaging 2019: Image Processing*. SPIE; 2019:5. doi:10.1117/12.2512484

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

Landers A, **Neph R**, Scalzo F, Ruan D, Sheng K. Performance Comparison of Knowledge-Based Dose Prediction Techniques Based on Limited Patient Data. *Technol Cancer Res Treat*. 2018;17:153303381881115. doi:10.1177/1533033818811150

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

SELECTED CONFERENCE PRESENTATIONS

Neph R, Huang Y, Yang YM, Sheng K. DeepMCDose: A Deep Learning Method for Efficient Monte Carlo Beamlet Dose Calculation by Predictive Denoising in MR-Guided Radiotherapy. MICCAI Workshop on AI in Radiation Therapy, Shenzhen, China. October 2019

Neph R, Huang Y, Yang YM, Sheng K. Deep Learning MC: Fast CNN-Based Prediction of Monte Carlo Dose for MR-Guided Treatment Planning. AAPM Annual Meeting, San Antonio, TX. July 2019

Neph R, Sheng K. Efficient Multi-GPU Calculation of Local Radiomic Features From 2D and 3D Images. AAPM Annual Meeting, Nashville, TN. June 2018

Neph R, Ouyang C, Neylon J, Sheng K. Distributed Multi-GPU Photon Beamlet Dose Calculation for Efficient Radiation Treatment Planning. AAPM Annual Meeting, Nashville, TN. June 2018

Neph R, Sheng, K. Predicting Risk in NSCLC Patients Using Learned Tumor Sub-Region Appearance From Quantitative Features in CT Images. AAPM Annual Meeting, Denver, CO. June 2017

INVITED TALKS

Neph R, McKenzie E. Deep Learning Theory and Applications. Guest Lecture for PB MED M209: Signal and Image Processing for Biomedicine. December 2019

Neph R, Huang Y, Yang YM, Sheng K. Deep Learning MC: Fast CNN-based Prediction of Monte Carlo Dose for MR-guided Treatment Planning. AAPM-SCC Norm Baily Awards 1st place prize. May 2019

1 INTRODUCTION

This year the X-ray celebrates its 125th birthday in the consciousness of humanity. Since its discovery by Wilhelm Roentgen in late 1985, the X-ray has become one of the most innovative tools in our quest to understand the human anatomy. The years following birthed the invention and steady progression of radiographic medical imaging and radiation therapy that are still very much relied upon in modern medical diagnosis and intervention. The earliest forms of radiation therapy were performed using the novel X-rays for superficial treatment of skin conditions and for hair removal, though it wasn't long before the effects of radiation for the treatment of malignancies began to be understood. By 1935, the practice of Henri Coutard to protract the delivery of radiation into multiple fractions had gained momentum for its healthy tissue-sparing properties.

These innovations paved the way for more than a century of progress, including the invention of the first medical linear accelerator in 1947, computed tomography (CT) imaging in 1972, and inverse radiotherapy treatment planning in the late 1980s, leading to the creation of both intensity modulated radiation therapy and intensity modulated arc therapy in 1995. As the technology surrounding the delivery of therapeutic radiation became more precise, so too became the expectation for computational planning and dose estimation methods.

Modern radiotherapy continues to experience radical paradigm shifts. In recent years, magnetic resonance imaging (MRI) guided radiotherapy has been steadily replacing CT guidance because it offers superior soft tissue contrast compared to CT imaging, doesn't use dose depositing ionizing radiation, enables arbitrary orientation of the imaging plane, and

offers real time target visualization, tracking, and radiation gating. Additionally, alternative therapies have been investigated that take advantage of the radiation dose deposition physics of electrons, protons, and other heavy charged particles to enhance treatment precision. The leap to non-coplanar 4π external beam arrangements with more sophisticated treatment planning algorithms has been investigated extensively with repeated success in improving planning efficiency, treatment outcome, and radiation delivery efficiency.

As the capabilities of radiation therapy have improved, necessarily, the demand for more accurate and faster computational dose modelling to power these state-of-the-art planning and delivery paradigms has also risen. Conventional computational techniques for calculating dose deposition estimates have steadily improved, however, a pointed trade-off between dosimetric accuracy and computational speed has complicated the selection of approach for each treatment paradigm. With the dawn of the new revolution in cloud-based distributed computing and the unprecedented success garnered by deep learning, a class of incredibly flexible machine learning algorithms, it is believed that the next leap forward in the field of radiation therapy will be enabled by these technologies.

Here we investigate the application of both high-performance computing and deep learning to affect this marked change in the highly demanding process of dose calculation. For many new treatment paradigms, significant improvements to the biological outcome can be realized, but the largest hurdle to their wide-spread use remains the difficulty of translation to clinically feasible timeframes. By employing these new computational techniques to the traditional task of radiation dose calculation we aim to bridge the gap between cutting-edge research and clinical practice to provide the best care possible.

2 OVERVIEW OF RADIATION THERAPY

In this chapter a summary of basic principles of radiation dose deposition are first presented, including the physical principles by which radiation dose is deposited in matter and a basic description of the biological effects of radiation in living cells. Next, the most prevalent computational methods for estimating radiation dose are introduced. Finally, an overview of the clinical radiation treatment planning process is described.

2.1 Radiation Dose Deposition Mechanisms

The therapeutic effects of electromagnetic radiation are derived from a class of radiative energy known pragmatically as *ionizing* radiation for its ability to liberate valence electrons from (aka ionize) the matter on which it impinges. The therapeutic benefits of ionization are realized *directly* when an X-ray or γ-ray interacts with the electron cloud of an atom, producing recoil electrons that continues to interact with the atomic electrons of critical molecules within a cellular body, inducing a biological effect. Another consequence of ionizing radiation producing therapeutic benefit is the *indirect* biological effect caused by oxidizing reactions with chemical free radicals that are produced in the cellular environment by recoil electrons (liberated during electromagnetic interactions). Particulate radiation can also induce biological change but does so in a manner that begins differently from that of electromagnetic radiation. Some choices of particulate radiation that are utilized for their therapeutic benefits include electrons, protons, and other heavy ions. Due to differences in the physics underpinning the interactions of each type of particle, the resulting biological effects are also varied in both their magnitudes and distributions within the affected area.

The *direct* and *indirect* effects of radiation both eventually produce biological changes to cellular function. The prevailing explanation for how radiation affects cellular function is by induction of DNA damage in the form of *single*- and *double-stranded breaks* to the polynucleotide strands maintaining the purposeful sequence of amino acid base pairs. Damage dealt directly to the base pairs as well as single strand breaks are significantly more common than double strand breaks and are more easily repaired due to the availability of complementary information from the undamaged strand for reconstruction. By comparison, double strand breaks are both much less common and much more difficult for the cell to properly repair[1]. The repair processes for base damage and single strand breaks are quite successful and are essentially limited to replicating missing information in the base sequence by using the complementary strand as an inverted template. But upon incurring damage to both strands, a variety of function-altering and sometimes lethal (to the cell) chromosome aberrations can be created during the challenging repair processes.

During the application of radiation for medical benefit, it is important that the quantity of radiation delivered be monitored and even targeted so that its effects can be understood and tailored to the needs of the patient. As such, one useful way to quantify the effects of radiation is by the measurement of absorbed radiation *dose* measured in units of energy per unit mass, or Gray (abbreviated as Gy) such that 1 Gy is equal to 1 J/kg. Dose is formally defined as "the expectation value of the energy imparted to matter per unit mass at a point."[2] For a dose of 0 Gy, there are no radiation effect. For reference, many forms of cancer are treated with radiation dose equal to anywhere from 10 Gy to 70 Gy in most cases. It should be noted that for radiographical applications, the goal is to obtain useful image quality and reduce the radiation dose as much as possible, and for therapeutic applications, we are

instead intentionally imparting large radiation dose in a precise fashion to specific parts of a patient's anatomy. In the pursuit of this goal, there are many mechanisms by which dose can be delivered, each with their advantages and disadvantages as determined with the physics by which they are governed. To give the reader a better understanding of these mechanisms, those applied most for therapeutic purposes are presented next.

2.1.1 Electromagnetic Radiation

Photons (electromagnetic radiation) with energies in excess of approximately 25 eV are considered ionizing[2]. Photons used for clinical purposes, however, typically have energies well in excess of 1 keV with energies of approximately 100 keV being preferred for radiological uses, and energies greater than 2 MeV for therapeutic uses. Clinical photons are typically generated using electronic equipment (and are called X-rays) or during the decay of radioactive isotopes in their pursuit of nuclear stability (and are called γ-rays). The interactions that photons have with matter, regardless of the mechanism of their production, can be categorized by one of Rayleigh scattering, photoelectric effect, Compton effect, and pair production, listed in the order of their occurrence probabilities with increasing photonic energy. Interactions that assume the form of Rayleigh (coherent) scattering result in a change to the photon's propagation direction with essentially no loss of energy; they don't contribute dose to the interaction medium and are consequently neglected from consideration in dosimetric analysis and from further description in our presentation.

2.1.1.1 The Photoelectric Effect

Primarily exhibited in high-atomic-number media and at low photon energies, the photoelectric effect occurs when the energy of photon is completely absorbed by an atom

and is subsequently transferred to one of its bound electrons. The excited electron escapes from its atomic orbit with kinetic energy (T) effectively equal to the difference between the original photon energy ($h\nu$) and the binding energy of its orbital shell (E_b), neglecting the minuscule energy transferred to the recoiling atom, as in Equation 2-1.

$$T = h\nu - E_b$$

Equation 2-1

Following electron emission, the atom is left with an orbital vacancy to be filled by an outer-shell electron of higher binding energy, resulting in emission of a characteristic photon of discrete energy equal to the difference in the orbital binding energies, or as one or more Auger electrons as necessary for the enforcement of energy conservation. The possibility for photoelectric interaction to occur with any atomic electron depends on the requirement that the photon energy must exceed the binding energy of that electron's orbital shell. The photoelectric effect is the primary physical process by which image contrast is generated for radiographical applications due to the high dependence of the process's interaction probability on the atomic number of a medium.

2.1.1.2 The Compton Effect

For photons of comparatively higher energy and lower atomic number, the Compton effect takes over for the photoelectric effect as the dominant photonic interaction process. The Compton effect describes the transfer of some of a photon's energy to free electron resulting in both the emission of an electron and energy loss accompanying a direction change for the incident photon. The so-called Compton electron emitted in this interaction has, by conservation of energy, a kinetic energy (T) equal to the difference in the incident ($h\nu$) and scattering ($h\nu'$) photon energies, as in Equation 2-2.

$$T = h\nu - h\nu'$$

Equation 2-2

The full kinematics describing the Compton effect are derived from joint application of the principles of energy and momentum conservation, and while they will not be presented here, provide the analytical tools for characterizing the probabilities of observing each of the possible outcomes of the Compton effect. These *cross-sections* are best characterized by the Klein-Nishina model[3] and are critical for the success of one class of dose estimation methods that employs this probabilistic understanding. The Compton effect is the dominant process on which the therapeutic application of radiation is founded.

2.1.1.3 Pair Production

Photons at even higher energies exhibit lower probabilities of both photoelectric and Compton interactions, giving rise to the dominance of another effect called pair production. Aptly named, pair production explains the emission an electron-positron pair from the complete absorption of a photon in the Coulomb field near an atom's nucleus or under rarer circumstances in the field of an atomic electron where a third product (the excited atomic electron) is also emitted and the process is instead called triplet production). In pair production, some of the energy of a photon is converted into mass in the form of the electron-positron pair and the remainder is given to these particles as kinetic energy. Due to the creation of these product particles, the minimum photon energy necessary for pair production is twice the rest energy of an electron ($2m_e c^2 = 1.022\ MeV$), and the photon energy exceeding this production threshold is randomly divided amongst each of the interaction products. The electron and positron both proceed to interact with the medium as will be described in the next section. However, when the positron's kinetic energy reduces

enough, it finally combines with a nearby free electron in a process called positron annihilation, that converts mass into energy in the form of two opposed photons, each with ≥0.511 MeV of energy, that proceed to interact further with the medium. Though less evident, pair production explains a non-negligible component of the dose deposited from therapeutic applications of ionizing electromagnetic radiation.

2.1.2 Particulate Radiation

The interaction processes of charged particles at therapeutic energies (traditionally much less than 100 MeV) are somewhat simpler than the dose-depositing processes of electromagnetic radiation. Charged particles interact with nearby atoms through the Coulomb force experienced in the electric field of either an atomic electron or an atom's nucleus. In the presentation of these interactions by Attix, they can be broadly classified by the ratio of the "classical impact parameter b vs. the atomic radius a," into one of three groups: "soft collisions ($b \gg a$)", "hard (or 'knock-on' collisions ($b \sim a$)", or "Coulomb-force interactions with the external nuclear field ($b \ll a$)".

Soft collisions occur when the distance between the charged particle and atom is considerably larger than the atomic radius; a very small amount of energy is transferred from the particle to the atom as the particle interacts with the entire atom's electric field. Though it may seem insignificant, the high probability for soft collisions leads to a significant transfer of energy to the medium on the aggregate. Hard collisions result from the interaction of charged particles with an individual orbital electron bound to a nearby atom. The significant transfer of kinetic energy to a single electron results in ejection of the bound electron as a δ-ray (delta ray), and an energy conserving loss to the incident charged particle. The now-vacant space in the atom's electron cloud is resolved by subsequent emission of a

characteristic photon or Auger electrons. For a charged particle passing within a few atomic radii of an atom, the probability of Coulomb-force interactions with the atom's nucleus increases. In more than ~97% of such interactions, the result is an insignificant energy loss and elastic scattering of the charged particle, altering its trajectory; for these no dose is absorbed by the medium. However, for the remaining ~3% of such interactions, the charged particle experiences an inelastic collision, losing energy in the process to the creation of a so-called *bremsstrahlung* X-rays; it is through this process that X-rays are produced for photon-based radiation therapy. The yield of bremsstrahlung X-rays is proportional to the square of the atomic number of the interacting atoms, and inversely proportional to the inverse squared mass of the incident charged particle and is consequently negligible for heavy charged particles such as protons and heavy ions.

Although the interactions of individual charged particles with matter follow a stochastic process, it is useful to describe them as a population using the expectation values of energy loss per unit path length (*stopping power*) and propagation distance (*range*) as a function of the particle energy and properties of the interaction medium. This treatment is called the *continuous slowing down approximation* (CSDA). The stopping power of a particle obeys an inverse square dependency on the velocity of the incident particle, which gives rise to a rapidly increasing rate of energy loss (and dose deposition) for charged particles as they approach the end of their flight paths. This sharp increase in dose deposition rate is called the *Bragg peak,* and it is of great interest for therapeutic applications of radiation for its reduction in entrance dose to healthy tissues on its way to irradiating a precisely targeted region of the patient anatomy.

2.1.2.1 Electrons

Electrons are negatively charged particles with a mass more than 1,800 times smaller than that of protons. Due to their small mass, they experience a significantly larger degree of multiple scattering as they interact with the electric fields of nearby atomic nuclei than do the much heavier protons and heavy ions, causing them to follow very jagged paths through media with high atomic number. Furthermore, due to the inverse square dependence of bremsstrahlung X-ray yield on the incident particle's mass, electrons are much more efficient for use in generating Bremsstrahlung radiation for therapeutic use than other charged particles, forming the operational foundation of modern medical linear accelerators (linacs). Most of the energy lost by electrons in media is deposited as dose through the excitation and ionization of atomic electrons during soft and hard collisions. Emission of secondary radiative products such as δ-rays, Auger electrons, bremsstrahlung, and characteristic X-rays may also contribute non-negligible dose by spreading energy to surrounding media. Some of the electromagnetic radiation produced through such secondary processes escapes the target media entirely in the form of radiative losses. Due to the high degree of scattering experienced by electrons, their Bragg peaks are substantially diffuse, diminishing their beneficial effects in therapeutic application. Another consequence of their tendency for scattering, the range of electrons in water and soft tissue is quite small, limiting their clinical application to more superficial targets.

2.1.2.2 Protons and Heavy Ions

Heavy charged particles are significantly less effective in their ability to produce Bremsstrahlung X-rays, but their high masses preserve the sharp gradient in stopping power that they exhibit at their Bragg peaks. This property makes heavy charged particle therapies

effective at achieving highly precise conformal dose to a targeted volume of tissue. Unlike photons (and to a lesser extent electrons), such heavy charged particles deposit very little *entrance* dose before the Bragg peak, where they still have high energies, and nearly zero *exit* dose after the Bragg peak, due to their pointed loss of energy within the Bragg peak. Despite their advantageous dose deposition properties, heavy charged particles have caveats to clinical use, including the difficulty involved in generating them using a large cyclotrons or synchrotrons, and the necessity for precise patient alignment before treatment resulting from the highly local dose deposition of dose in the Bragg peak.

2.1.2.3 Electromagnetic Field Effects

When any moving charged particle is subjected to an electromagnetic field, the Lorentz force induces a force (**F**) on the charged particle equal to the sum of the electric field (**E**) strength and the vector product of the particles velocity (**v)** with the strength of the magnetic field (**B**), scaled by the charge of the particle (q), as given in Equation 2-3. A bold emphasis denotes a vector quantity.

$$\mathbf{F} = \boldsymbol{q}(\mathbf{E} + \mathbf{v} \times \mathbf{B})$$

Equation 2-3

In radiation therapy, the electric field component of Equation 2-3 is often small enough to be excluded from discussion. However, for charged particles in the strong magnetic field of an MRI scanner, the magnetic field component of the Lorentz force can lead to substantial changes in the deposition of radiation dose.

Two macroscopic effects are observed in the dose from X-rays[4]. First, as charged particles are liberated within homogeneous media (through the photoelectric, Compton, and pair production interactions), their momenta are altered by the Lorentz force, resulting in an

asymmetrical dose penumbra and reduced build-up distance. Second, near the interface between high and low density media, the liberated charged particles exit the high density medium, follow arc-shaped trajectories in the low density medium, due to the vector product component of the Lorentz force, and return back to the high density medium to deposit dose near the interface[5,6]. From Equation 2-3, it is clear that the Lorentz forces produce an effect on dose that is dependent on the strength of the magnetic field, but the mass of the charged particle is also an important factor; the acceleration of the particle is inversely proportional to its mass for a given induced Lorentz force. From this relationship, the small mass of the electron punctuates its importance in the X-ray dose deposition process, inspiring the use of the term electron return effect (ERE) when referring to this class of dosimetric perturbation.

2.2 Dose Calculation Algorithms

Computational dose calculation is a critical component of clinical radiation therapy because it enables the creation of safe and effective radiation delivery *treatment plans* without the need for physical (and often impossible) dose measurement to be performed for each patient. Dose calculation describes the process by which the three-dimensional (3D) voxelized dose within some geometry is algorithmically estimated as a result of a pre-determined delivery of ionizing radiation. Dose calculation is the procedure responsible for providing causal knowledge of how manipulation of the radiation delivery process will affect the delivered dose and is a requirement of every method for treatment planning. There exist many approaches to dose calculation that each strike a balance between the computational speed and the dosimetric accuracy achieved. There is not one method that claims superiority over all others. Rather, some methods are well-suited to some purposes, and other methods

better suited for other purposes, emphasizing the importance of the selection of dose calculation method to the task at hand. In the following sections, the most widely used dose calculation methods along with their advantages and disadvantages for certain applications will be introduced.

2.2.1 Monte Carlo Dose Calculation

The most general technique for dose calculation is to simulate the interactions of individual units of electromagnetic and particulate radiation and virtually "measure" the resulting dose to a voxelized geometry. These simulation-based methods are named after the wider class of Monte Carlo algorithms in which random samples are repeatedly drawn from interesting domain-specific probability distributions to obtain numerical results. In the context of radiation dose calculation, the interesting probability distributions come from the interaction cross sections that are constructed from physical principles and experimental findings. Using a virtualized model of an actual radiation source, particles (or photons) are instantiated one-by-one and tracked through a virtual geometry in discrete "steps", with their *phase space* parameters (positions and momenta) updated after each step according to random samples from a probability density function corresponding to the type of interaction undergone during the step, which is randomly sampled from the interaction cross-section model. Using this stochastic simulation strategy, the dose gradually converges to a solution with increasing certainty.

Since Monte Carlo dose calculation techniques are based on well-established physical principles of particle transport, they are considered the gold standard for dosimetric accuracy, but they require significantly longer computation than deterministic methods to attain reasonable levels of certainty in their solutions. This makes Monte Carlo dose

calculation difficult to apply in the clinical setting where limitations on dose calculation and treatment planning times are often in effect. Many modifications to accelerate the Monte Carlo dose calculation have been proposed such as Macro Monte Carlo (MMC)[7], condensed history[8], variance reduction[9–11], and hardware-enabled parallelization[12–16], but increasing Monte Carlo simulation speed still remains of great interest for clinical application.

2.2.2 Analytical Dose Calculation

Analytical approaches to dose calculation share the property that they provide a deterministic solution to the problem of dose estimation that produces consistent results and algorithmic run-times across independent evaluations for the same inputs. Despite their deterministic nature, analytical approaches often provide some means of adjusting the speed-accuracy trade-off such as by adjusting the spatial dose resolution. By far the most popular form of analytical X-ray dose calculation is achieved by efficient convolution of an analytically defined volume of voxelized TERMA (total energy released in matter) by a set of pre-computed dose deposition *kernels*[17]. Each kernel is calculated such that it describes the spatial distribution of dose deposition resulting from a single electromagnetic interaction in a homogeneous medium (typically water or soft tissue). Kernels are defined for monoenergetic X-ray beams using Monte Carlo simulation and can either be represented numerically, or more efficiently as analytical functions fit to the numerical simulation results. In order to calculate the dose of polyenergetic X-ray beams, the doses resulting from convolution monoenergetic by multiple monoenergetic dose kernels are combined by weighted addition in a process called convolution-superposition (C/S)[18]. Inclusion of more dose kernels in CS intuitively produces more accurate dose but at higher computational cost.

Within the C/S classification, there exist several sub-classes that differ in their choices of model for the dose kernels. Pencil beam convolution (PBC) is a fast C/S method where 2D dose kernels (pencil beams) are convolved with TERMA only along the transverse axes. Varian's Analytical Anisotropic Algorithm (AAA)[19] is a more accurate extension to PBC that further implements transverse kernel scaling via radial exponential functions to account for the lateral inaccuracies of PBC in heterogeneous media. Collapsed-cone convolution-superposition (CCCS)[20] is yet another approach that accelerates 3D C/S by concentrating all the dose spread within radially diverging cones onto each cone's central axis. The CCCS assumption has the effect of reducing the number of kernel parameters and shortening the convolution time at the expense of additional approximation to the resulting dose. With the addition of corrections for kernel tilting and hardening[21], and anatomical heterogeneity by kernel scaling along each cone's axis,[22,23] CCCS is generally regarded as more accurate than both AAA and PBC, but due to its higher computational cost, hasn't seen universal adoption for clinical use[24].

2.2.3 Linearized Boltzmann Solvers

Another class of iterative algorithm for explicitly solving the linearized Boltzmann transport equation (LBTE) – "the governing equation which describes the macroscopic behavior of radiation particles … as they travel through and interact with matter"[25] – has seen advancement over the last decade[26–30]. Unlike both the MC simulation and C/S methods, LBTE solvers employ iterative numerical algorithms that converge upon the "correct" dose based on some pre-defined stopping criteria on the desired accuracy. As in Monte Carlo dose calculation, the accuracy of LBTE solvers is dependent on the calculation time. However, unlike in Monte Carlo, LBTE solvers exhibit error that is systematic, resulting from

discretization of calculation parameters, rather than as statistical noise. LBTE solvers show promise for dose calculation of treatments with high beam counts because the solution time is weakly dependent on the number of radiation sources. Unfortunately, the linearizing assumption of the Boltzmann transport equation employed by LBTE solvers is invalid for particles experiencing the effects of an external magnetic field, disqualifying them from use in the emerging paradigm of MR-guided radiotherapy.

2.3 Radiation Treatment Planning

The goal of radiation therapy is to induce death or disturb the reproductive capacity of cells in a targeted volume, often cancerous in nature, while preserving the normal function of surrounding healthy tissues composing our vital organs. The standard practice for achieving 3D conformal radiotherapy (3DCRT), in which radiation is precisely delivered to a target volume informed by 3D imaging data, for many anatomical sites is external beam radiotherapy (EBRT). EBRT combines multiple beams of radiation generated using external sources and focused on the target volume to simultaneously achieve homogeneous dose within, and minimal dose beyond the target boundaries. An early method for accomplishing 3DCRT prescribed the use of fully conformal beams that each delivered a homogeneous fluence with a cross-sectional beam shape matching the projection of the target volume from advantageous angles of entry (termed the beams-eye-view (BEV)) into patient. To facilitate shaping of each beam, the multi-leaf collimator (MLC) was developed[31]. The MLC provides two opposed banks of radiation attenuating tungsten leaves that can be individually positioned to shape the beam[32].

Following this innovation in beam shaping hardware, an advanced technique called intensity-modulated radiotherapy (IMRT) was developed[33] to increase dose conformality in a target volume. IMRT divides the delivery of radiation for each beam into a set of *fields*, each with homogeneous fluence, such that their sum generates more complex shaping of dose for greater precision of delivery. The individually positioned leaves of the MLC can be mathematically represented as a discrete grid, dividing the total field-of-view (FOV) of a beam into component *beamlets,* for which separate radiation fluence can be delivered; this grid of per-beamlet fluence is referred to as the *fluence map* of the beam and its definition is the goal of inverse treatment planning used in modern radiotherapy.

2.3.1 Beamlet Dose Calculation

Beamlet dose calculation produces causal information linking the choice of individual beamlet fluence to the resulting dose deposited in the patient. To create IMRT treatment plans with complex dose modulating fluence maps, this causal dose information must be calculated for every beamlet involved in the eventual radiation delivery, the number of which can easily reach into the tens or hundreds of thousands[34]. The requirement for such large quantities of dose data severely limits the computational time that can be afforded to the pre-planning beamlet dose calculation procedure in clinical settings. Concurrent with the efficiency considerations of beamlet dose calculation, are the concerns for dosimetric accuracy. Although fast analytical methods exist for many treatment paradigms, such as PBC for X-ray therapy, the sacrificed quality of dose they provide is evident in the outcome of treatment planning[35], resulting in sub-optimal treatment efficacy due to the lack of accurate causal information provided to the treatment planning algorithm. Highly accurate approaches to dose calculation, such as MC simulation and LBTE solvers are attractive in this

regard but fail to even remotely meet the efficiency requirements of the very large scale (VLS) beamlet dose calculation tasks required by advanced treatment modalities. This highlights the difficulty and importance of selecting the optimal balance of accuracy and speed for beamlet dose calculation.

2.3.2 Inverse Treatment Planning

Treatment planning for radiation therapy is enabled by mathematical optimization techniques that iterate through potential plans to accomplish well-defined goals encoded as an *objective function.* The term inverse planning is used to describe this process since our planning goal is an effective and conformal dose distribution, but the machine parameters must instead be defined to indirectly affect the eventual dose delivery. Equation 2-4 shows a simple form of an optimization problem that is used for beamlet-based inverse planning.

$$\begin{aligned} \underset{f}{\text{minimize}} \quad & \|Af - d_0\|_2^2 \\ \text{subject to} \quad & f \geq 0 \end{aligned}$$

Equation 2-4

In this problem formulation, the goal is to select a value for the vector $f \in \mathbb{R}^N$, which encodes a plan's fluence maps, such that the estimate of the deliverable volumetric dose distribution computed as Af agrees with the dose distribution prescribed by the physician and encoded as $d_0 \in \mathbb{R}^M$, with M defining the size of volumetric dose, and N defining the number of beamlets in the plan. For convex inverse planning problems like Equation 2-4, simple iterative gradient-based optimization algorithms such as the well-known gradient descent algorithm, are used for determining the optimal value of f. However, to further improve the biological efficacy of clinical treatment plans, more complex mathematical

terms are often included[36–44], requiring the use of more sophisticated optimization algorithms[45,46] and greater computation time.

The primary goal of treatment planning is to select f such that the voxels contained within the planning target volume (PTV) all receive the physician-prescribed radiation dose, while the remaining voxels, belonging to healthy tissues, called organs at risk (OARs), receive as little dose as possible. As a consequence of the physics of dose deposition, it is generally impossible to achieve perfectly conformal PTV dose while completely sparing OARs, so a balance must be struck based on the radiation tolerances and relative positions of the OARs to the PTV. Although there are many aspects to inverse treatment planning that must be considered, it is important to remember that the outcome of any treatment planning algorithm is fundamentally limited in its ability to produce effective outcomes by the quality of the dosimetric information it has available to it, determined during beamlet dose calculation.

2.3.3 Final Dose Calculation

After inverse treatment planning is used to determine the optimal radiation delivery parameters, the plan must be evaluated for its safety and efficacy before treatment can begin. Plan evaluation uses a *final* dose calculation procedure in which the effects of radiation from all beams are analyzed holistically rather than at the component beamlet-level (as in beamlet dose calculation). This distinction generally relaxes, somewhat, the requirements for speed in favour of increased dosimetric accuracy. Highly accurate algorithms like MC and LBTE solvers, for which the computational speed is independent of or only weakly dependent on the number of beamlets in the plan (when their dose can be immediately combined), are especially useful for final dose calculation. The clinical implementations of end-to-end dose

calculation and treatment planning software almost always include separate strategies for beamlet and final dose calculation to best accomplish the goal of fast and effective radiotherapy, but the details of both strategies are still evolving with the invention of cutting-edge techniques that push the capabilities of dose calculation to new limits.

2.3.4 Online Adaptive Radiotherapy

With the new capabilities made possible by MR-guided radiotherapy, such as enhanced soft tissue image contrast, daily on-board imaging and real-time target tracking, online adaptive radiotherapy (OART) is closer to widespread clnical feasibility than ever before. With OART, daily changes to the shapes and locations of the planning target volume(s) and organs-at-risk are visualized and compensated-for in updated treatment plans. Daily re-planning offers superior dose precision over the current standard of practice, in which a single treatment plan, created from pre-treatment imaging, is re-delivered over the course of many weeks. Unfortunately, treatment planning is traditionally a time-consuming procedure, typically requiring between one and two weeks to complete, depending on the complexity of the treatment. The primary challenge of OART is to condense the entire treatment planning process into a less than 20 minute timeframe, while the patient waits in the treatment position. Although MR-guidance is a key factor to implementing OART, the presence of the strong magnetic field, and resulting EREs affecting the dose, necessitating the use of much slower Monte Carlo dose calculation methods, imposing further strain to the reduction of treatment planning duration.

3 PARALLEL BEAMLET DOSE CALCULATION VIA BEAMLET CONTEXTS IN A DISTRIBUTED MULTI-GPU FRAMEWORK[34]

3.1 Introduction

Modern radiation treatment planning is powered by inverse optimization algorithms that require causal information connecting the plan delivery parameters to the resultant patient dose distribution. This information is encapsulated in a dose influence matrix, consisting of the dose of individual beamlets, which are the smallest deliverable unit whose geometry is typically determined by the multi-leaf collimator width. Monte Carlo (stochastic) simulation is regarded as the gold standard for dosimetric accuracy but remains impractically slow for very large scale (VLS) optimization problems. On the other hand, deterministic approaches provide a faster approximation by convolving reusable dose spread kernels over analytically computed TERMA fields on each unique patient geometry. The popularity of deterministic convolution superposition (C/S) solvers such as collapsed-cone convolution superposition (CCCS)[17,47] and analytical anisotropic algorithm (AAA)[19] have enabled acceptably accurate clinical dose calculation that can typically be performed with an order of magnitude reduction in time required for general purpose (Geant4) and even special purpose (VMC++) Monte Carlo methods in some circumstances[24].

In practice, the dose influence matrix calculation speed is generally acceptable with modern computers for IMRT plans involving only a few pre-determined beams. For arc optimizations and TomoTherapy, two orders of magnitude greater number of beamlets are needed to construct the matrix. Owing to the evidence that non-coplanar beam orientations

and automatic selection of beams and arc trajectories has been shown to produce improved plan quality[40,48–54], there is great research interest in automatic orientation selection from a much larger set of beam candidates by learning-based[55] and dose-driven approaches[38,56–58]. One such method, non-coplanar IMRT with beam orientation optimization[38], selects beams from several hundred candidates, escalating the requirement for dose calculation proportionally. More recently, dynamic collimator rotation[41] and non-coplanar VMAT[40] have been developed for further improved dosimetry and delivery efficiency, pushing the requirement of beamlet dose for optimization to be ~1000 times greater than that of fixed beam IMRT plans to account for the additional degrees of freedom. Dose calculation for the increasing number of beamlets can be a slow process by clinical standards, particularly when higher dosimetric accuracy is desired. There has not yet been an improvement to dose calculation processes using collapsed cone beamlet dose generation that would make these VLS treatment planning methods clinically tractable, despite the clear dosimetric benefits granted for heterogeneous geometries. The purpose of this work is to improve the dose calculation speed for these VLS planning methods so that standard clinical implementation may be achieved.

Since deterministic dose calculation is an embarrassingly parallel computational problem, graphics processing units (GPUs) with a large number of computational cores have found widespread success in accelerating dose calculation for treatment plan optimization and validation purposes. Chen *et al.* employed efficient GPU memory coalescing and analytical dose spread kernels to achieve 1000-3000x speedup over the CPU-CCCS implementation for TomoTherapy dose calculation[59,60]. Neylon *et al.* further optimized memory access speeds during GPU-CCCS convolution by first transforming voxelized TERMA

to a basis aligned with each collapsed-cone direction and subsequently carrying out efficient parallelized line convolutions, demonstrating further acceleration over the CPU method[61]. Tian *et al.* developed a GPU Monte Carlo dose calculator (goMC) based on the OpenCL GPU computing framework to enable widespread adoption of Monte Carlo simulation across all popular GPU hardware architectures[62]. Ziegenhein *et al.* delocalized the dose calculation process with an integrated cloud-based Monte Carlo framework that allows dynamic scaling of computational resources as needed to reduce workstation cost and complexity, improving the expectation for performance scaling with additional hardware on short-lived simulations where gains were previously pervasive[63]. Park *et al.* performed beamlet-based dose convolution with adaptive finite-sized pencil beam kernels to reduce the number of beamlets required to model arbitrary field shapes and accelerate volumetric dose verification for the optimized fluence maps[64]. Cho *et al.* validated the use of a GPU-accelerated convolution-superposition method for kilovoltage dose calculation in small animal irradiation research[65].

The foundation for our approach is the nonvoxel-based (NVB) GPU dose calculation algorithm of Neylon *et al.*[61] which optimizes previous GPU-based CCCS methods[66–68] by employing efficient GPU memory handling practices. The NVB algorithm is an improvement over these algorithms in that it reduces latency in device memory access by successive transformation of the CT density data to align it with each collapsed-cone ray enabling efficient line-convolution. Like the NVB approach of Neylon *et al.*, we treat the convolution operation on a continuous domain with interpolation during dose kernel sampling and dose spread. Unlike the NVBB approach of Lu[69], which treats TERMA and dose calculation in a continuous domain without discretely modeling beamlets, we maintain the standard voxel-based beamlet-superposition (VBS) representation in the output of our algorithm such that

we follow the path of pre-calculating discrete, beamlet-specific dose distributions for use during plan optimization. We make this choice primarily to maintain compatibility with the variety of beamlet-based planning techniques.

Our contributions are two-fold. First, we propose a novel modification to the existing GPU implementation of full beam deterministic dose calculation, enabling efficient low-level parallel computation of beamlet-specific dose using a *beamlet-context transformation*. Second, we implement our beamlet-based GPU dose calculation algorithm in a scalable distributed framework supporting flexible high-level multi-GPU acceleration. Our framework greatly improves the efficiency of the VBS method for use in VLS optimization problems such as dose driven automatic IMRT beam orientation[38] and VMAT trajectory optimization[40,41], 4π[48], and TomoTherapy[70] treatment planning. In this study we introduce the framework for our method, provide some dosimetric characterization for standalone beamlets and their composition as a broad beam, measure its computational performance, and discuss the scalability across networked computational nodes. We also discuss the computational efficiencies enabled by our proposed method and how it could potentially benefit VLS optimization problems but recognize that classification of dosimetric accuracy in clinical treatment planning settings is a more involved matter and leave such investigation to future work.

3.2 Materials and Methods

In this section, we describe our novel beamlet context approach for efficient beamlet-based dose calculation on a GPU. Next, we show that our method may be further parallelized in a scalable manner across a network of multi-GPU compute nodes. Finally, we describe the

experiments designed to quantify our framework's dosimetric accuracy against Monte Carlo and CPU-CCCS reference doses and computational speed in comparison to an existing GPU model-based method.

3.2.1 Nonvoxel-Based Dose Calculation

3.2.1.1 Beamlet-based Dose by Intra-beam Parallelization

Dividing the dose calculation problem into per-beam tasks is a trivial matter. While others have chosen to further separate the problem into per-beamlet tasks, we instead chose to calculate dose for these beamlets simultaneously. Our algorithm processes each beam as a unit, concurrently producing independent dose distributions for each of the beam's active beamlets before writing them to file and continuing with the next beam. This innovation is the key to achieving an efficient and scalable algorithm that minimizes GPU execution and memory management overhead. Details of the low-level parallelization are explained in the subsequent sections.

3.2.1.1.1 TERMA Calculation

Dose calculation for each beam begins by first generating a binary fluence map where active beamlets are assigned a unit fluence. Active beamlets are defined by projecting the target onto the fluence plane at the isocenter (Figure 3-1a) and detecting intersections with the target volume along each of 9 rays configured for each beamlet as shown in Figure 3-1b. If any sample ray intersects any part of the target volume, the beamlet is considered active and its dose will be calculated. This approach is used frequently for beamlet dose calculation and effectively minimizes the computational complexity of the full problem without sacrificing plan quality.

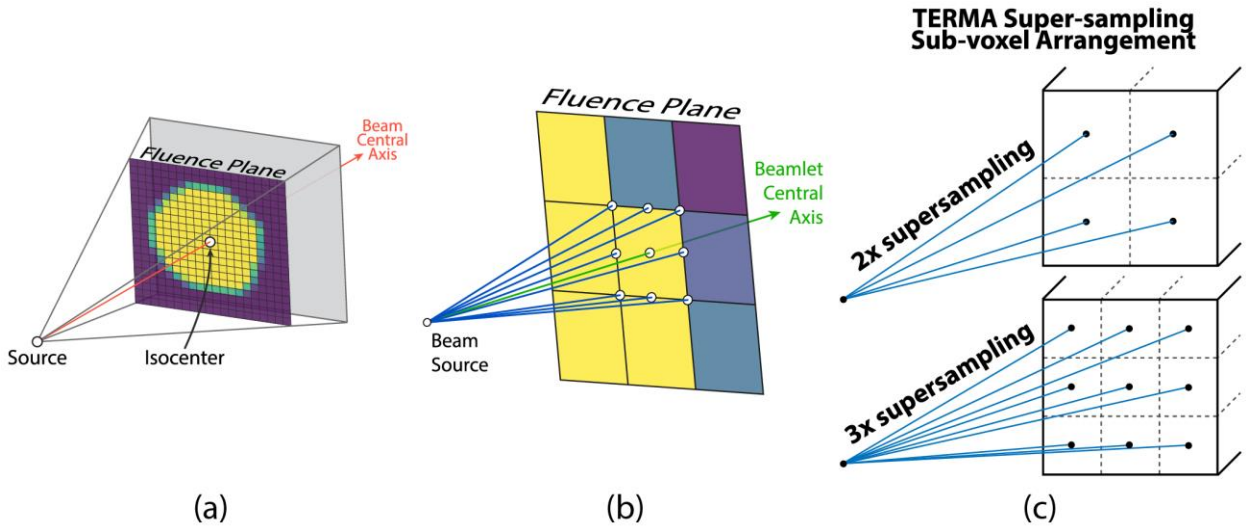


Figure 3-1. (a) Intersection map for one beam orientation and target volume definition. Purple-colored cells have no target intersection and are excluded from beamlet-dose calculation, yellow: full intersection, others: partial intersection. (b) Super-sampling ray layout for testing beamlet-target intersection. (c) cross-section of TERMA calculation sub-voxel arrangement for super-sampled averaging (2x and 3x options shown for one voxel).

The binary fluence map is used in calculating the TERMA by means of a path-length tracing procedure[47] that implements a modified version of Siddon's[71] algorithm better suited to the GPU. For every voxel, i, in the volume, a ray is traced between the source position and the voxel's center, along which, the radiological path length is accumulated, according to Equation 3-1, for a line segment of variable length l_j through each voxel j$\in \mathcal{R}_i$ of density ρ_j .

$$d_i = \sum_{j \in \mathcal{R}_i} l_j \rho_j$$

Equation 3-1

To maximize calculation speed, dose is calculated assuming a constant polyenergetic beam spectrum. Since beam hardening effects change the true beam spectrum within an attenuating medium, an auxiliary value, T_i^*, is calculated for each voxel, i, and used instead of the actual TERMA, T_i, during dose convolution. Equation 3-2 shows the expression for T_i^*,

with respect to T_i at voxel i, including corrections for beam hardening and the inverse square effect of diverging beams.

$$T_i^* = \left(\frac{D_{s,a}^2}{D_{s,i}^2}\right) H_i T_i$$

where

$$T_i = \sum_E \Psi_E \left(\frac{\mu_E}{\rho}\right) e^{-\left(\frac{\mu_E}{\rho}\right) d_i}$$

Equation 3-2

$D_{s,a}$ is the distance from the source to the rotational axis (isocenter), and D_{sv} is the distance from the source to voxel i, in the direction of the beam's central axis. Through the beamlet context extraction process, described in the following section, kernel tilting is implicit and a corrective term ($D_{s,a}^2/D_{s,i}^2$) for the inverse-square effects of a diverging beam is applied directly to each T_i. H_i is a voxel depth- and tissue density-dependent factor based on an effective x-ray attenuation coefficient, which corrects for beam hardening effects. It is interpolated from a table of pre-computed values specific to each beam spectrum and material. For this study, a single fluence-attenuation-table (FAT) was calculated and used for a 6MV bremsstrahlung x-ray spectrum in water. Because the utilized dose kernels are precomputed in homogeneous water, the energy-dependent mass attenuation (μ_E/ρ) coefficients are constant and equal to those of water. Thus, to correct for material inhomogeneity, a standard C/S technique[72] is used, whereby the kernel is instead warped according to the material dependent radiologic pathlength, d_i for each voxel (expressed in Equation 3-1).

Anti-aliasing via uniform super-sampled averaging is employed during TERMA calculation with negligible cost to address aliasing (stair-stepping) otherwise observed

along the beam and beamlet edges. Each voxel is divided into a set of k^3 sub-voxels, for the user-selected integer super-sampling level, k, as depicted by Figure 3-1c. The ray-based path-length and TERMA calculations are performed for each sub-voxel, and the voxel's TERMA is assigned to the average of their values. The entire low-level process is presented for one compute node in Figure 3-2a, including the GPU modules and post-processing tasks. The distributed workflow in Figure 3-2b is explained further in section 3.2.1.2.

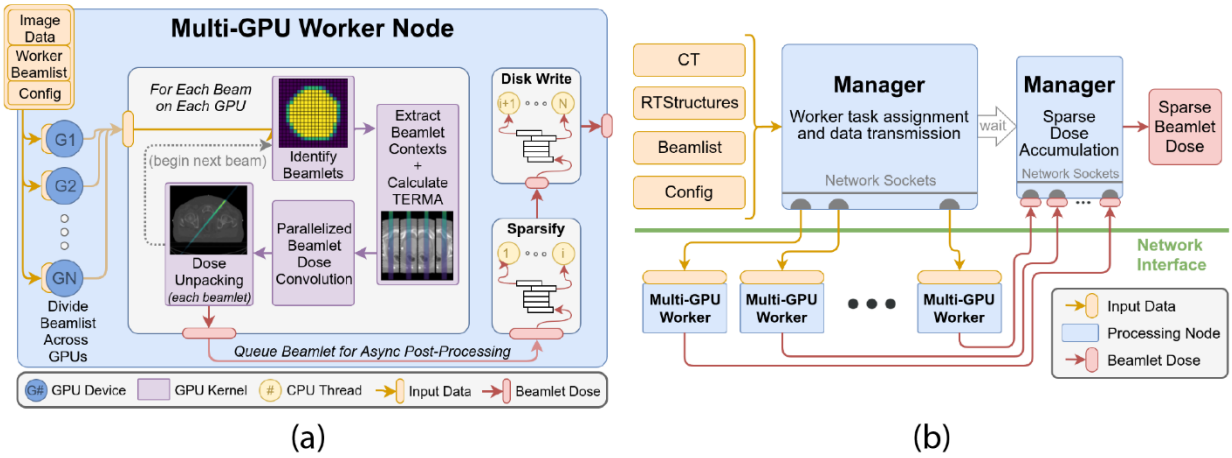


Figure 3-2. (a) Beamlet dose calculation workflow. A single worker node processes beams in parallel across its resident GPU devices. Beamlet processing is further parallelized on a GPU using beamlet contexts. (b) Distributed computing framework. The manager node prepares independent task lists for each worker node to process in parallel and receives the results for delivery to the requestor.

3.2.1.1.2 *Beamlet Context Extraction*

The classical implementation of beamlet-based dose calculation treats each beamlet separately and performs dose calculation for each, one-by-one. While a functional solution, this approach results in a linear scaling of calculation times with the total number of beamlets as in Equation 3-3.

$$\text{Calculation Time} \propto \sum_{b=1}^{B} n_b \quad \text{for} \quad \begin{bmatrix} B: & \#\text{ of beams} \\ n_b: & \#\text{ beamlets in beam } b \end{bmatrix}$$

Equation 3-3

Our method instead reduces the time scaling factor to *B* by calculating individual dose for all beamlets in a beam at once. During dose calculation for one beamlet on the GPU, 3D dose spread is applied for every voxel in parallel. This approach to parallelizing the problem is sub-optimal because the random-access latency of globally stored data during convolution is high and the speed of the algorithm suffers. Trying to directly implement the NVB algorithm with support for beamlet-based dose calculation presents other difficulties such as introducing dose assignment race conditions[66] and inflating the memory footprint beyond feasibility with beamlet specific book-keeping. Attempting to store dose directly as a sparse array on the GPU to overcome memory consumption limits also introduces deleterious race conditions and memory access latencies since constant speed random-access of memory is no longer possible.

To circumvent these problems, we recognize that the dose resulting from common clinical x-ray spectra is spread locally around an interaction point with compact spatial support. For the purposes of radiation beam selection and fluence map optimization, a close approximation of the dose can be obtained by limiting the calculation of dose spread to the immediate neighborhood of each beamlet. This approximation known, as kernel or dose truncation, has been used previously for dose kernel generation[73] and simultaneous Monte Carlo beamlet dose simulation[74] to accelerate dose calculation. Utilizing this approximation, we construct a composite array of independent *beamlet contexts* (hereafter referred to as the *context array;* depicted in Figure 3-3) that each contain only the density and TERMA data necessary for performing the CCCS convolution operations within its beamlet's confined surroundings. The long axis of each context is aligned in parallel to the long axis of its beamlet (called beamlets-eye-view; BEV) such that a minimum distance from any voxel of the

beamlet to the boundaries of the context is enforced by the selected context radius, as described by Figure 3-4. Aligning the contextual data in this way also implicitly corrects for beam divergence effects that would otherwise require costly kernel tilting to be individually applied for every beamlet. Since the contexts are independent and self-containing, their arrangement in the construction of the context array is unremarkable and therefore flexible. To construct each beamlet's context, we directly sample density from the global coordinate system, while mapping each voxel in the context to its corresponding global coordinate and directly calculating TERMA at this location using the method described in section 3.2.1.1.1. We ensure that only TERMA attributable to a context's beamlet is included in the context by projecting onto the fluence plane and testing membership in the fluence element corresponding to that beamlet using the procedure outlined previously.

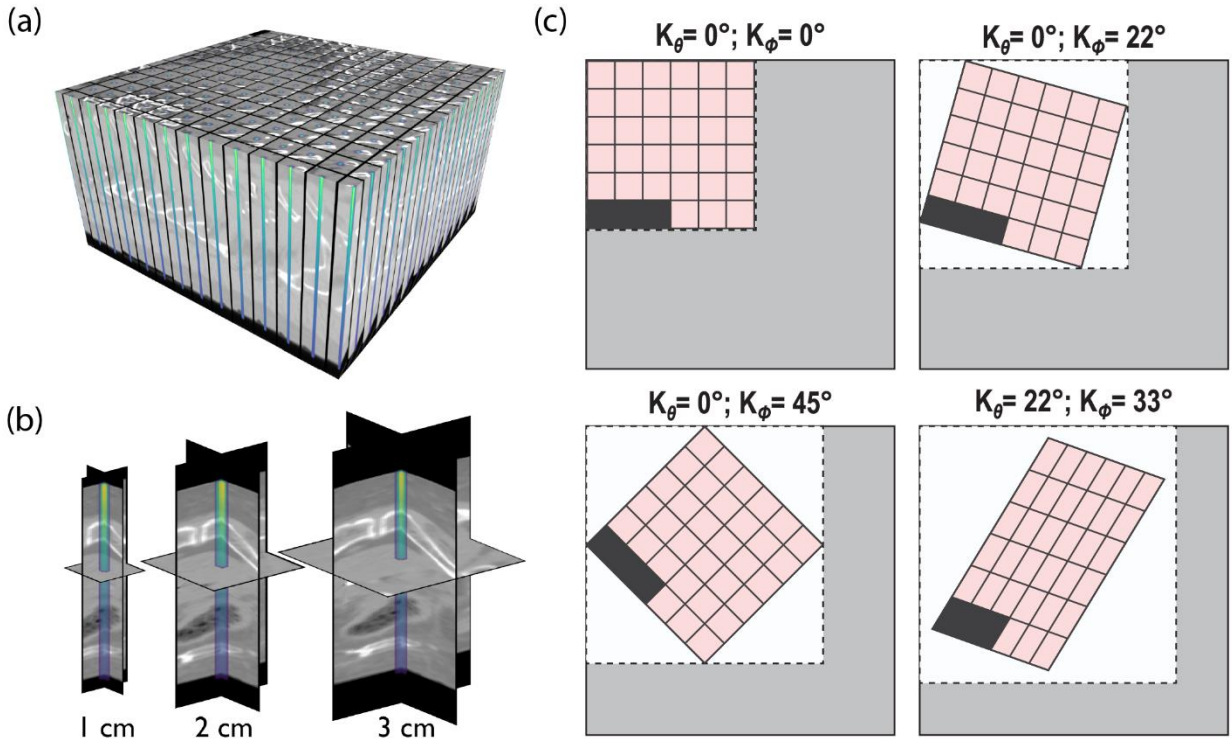


Figure 3-3. (a) Visualization of the beamlet context array for a single beam including contextual densities and beamlet-specific dose after calculation. (b) Beamlet context cross sections for various context radii with dose overlaid. (c) Convolution-ray-aligned context array (cross-section) for various kernel rays. Grey area is allocated once and reused for all beams. White subregions are allotted for kernel-ray-specific convolutions geometry. Black cells indicate unused space after packing beamlet contexts into the array. Convolution direction is into page.

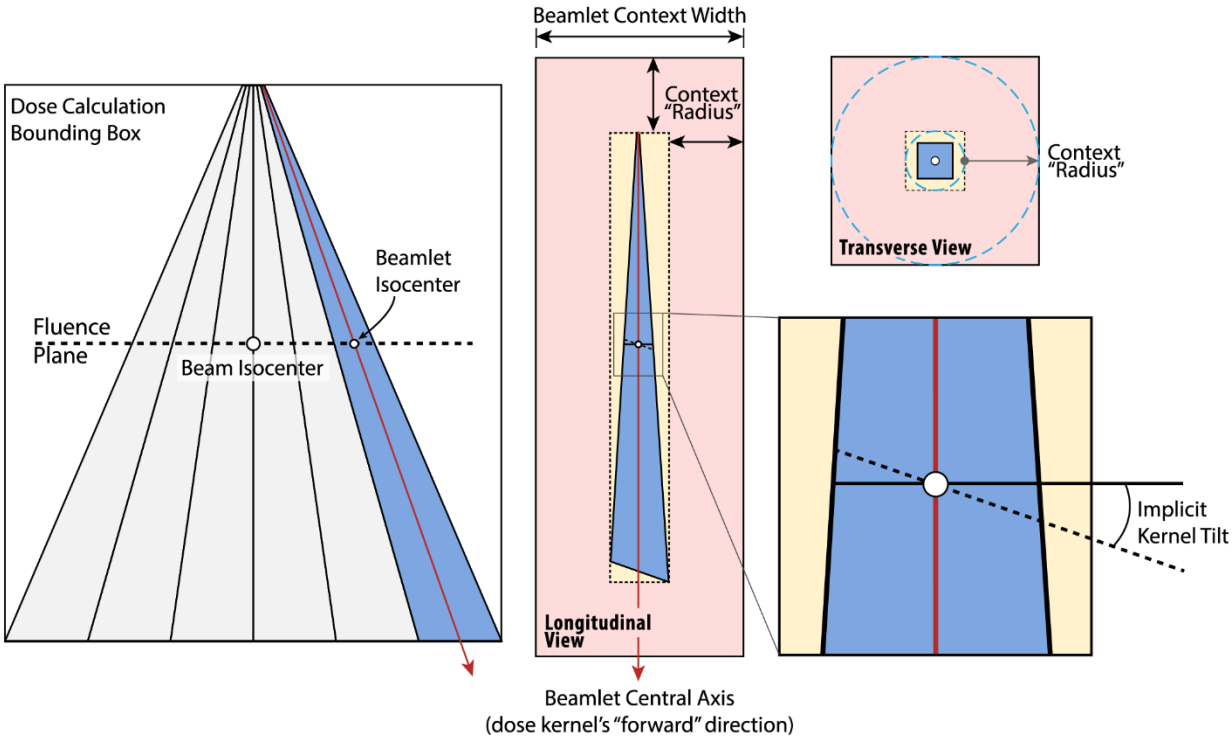


Figure 3-4. Construction of one beamlet context with implicit kernel tilting. Blue region indicates the volume of non-zero TERMA for a single beamlet. The distance between the blue rings in the transverse view is representative of the context radius setting. The union of red and gold boxes represents the volume in which dose is computed.

3.2.1.1.3 *Nonvoxel-Based Transformation*

We combined our novel context transformation with the NVB algorithm[61] to calculate beamlet-specific dose in parallel. To understand the motivation behind the NVB algorithm, we briefly describe the structure of the Monte Carlo point spread kernels; a complete presentation can be found in the literature[17,20,47]. The kernels used in our approach contain coefficients calculated in a polar system of homogeneous water with 24 radial and 48 angular points. By collapsing the full cartesian sampling space into a set of radially divergent cones, the dose distribution may be closely approximated at a much lower computational cost. During convolution, the dose is only calculated for voxels intersecting the central axes of these cones rather than for a radial ray terminating at each of the original sampling points on the dose grid. This is the defining distinction of CCCS over more precise C/S algorithms.

One simple way to structure the CCCS algorithm on the CPU is to iterate over the volume, stopping at each voxel to spread dose to surrounding voxels before moving to the next iterate. In GPU computing, memory read and write latencies usually present the greatest barrier to an efficient implementation. A direct translation of the CPU-based approach to GPU leads to substantial memory latency and introduces race conditions that force expensive synchronization of GPU threads to obtain correct results. Same as the NVB algorithm, we combat these issues by resampling (rotating) the context array along each kernel ray in turn, placing TERMA and density data into a new coordinate system referred to as the rays-eye-view (REV). We use tri-linear interpolation to cast the original density and TERMA data into the context-based REV and back into the global coordinate system. As such, each mapping is affine and performed on a continuous spatial domain. Furthermore, each mapping is invertible to the extent of the data retained after truncation by the selected context radius. After transformation, each row of the contextual data is arranged as a contiguous block of memory permitting its efficient access by coalesced GPU memory transactions during convolution. In doing so, we reduce the memory access overhead and amplify the benefits of GPU parallelization.

As there are many rays along which the dose will be spread during convolution with the dose kernel, the array of Figure 3-3a is reconstructed in the REV specific to each convolution ray prior to dose calculation. The resulting dose is successively transformed into a fixed *common orientation,* which aligns every context central axis in parallel with the data column direction, for accumulation until dose for all rays has been calculated. To efficiently obtain beamlet-specific dose for every one of a beam's active beamlets in parallel, we maintain three arrays in the current REV coordinate system to store the contextual density, TERMA, and

resulting dose. A fourth array is kept in the *common orientation* for accumulation of dose from each REV. GPU memory requirements for the context-based method are primarily determined by the sizes of these four arrays which change based on a number of user-controlled and geometry specific factors such as the number of active beamlets in each beam, and various quality settings (voxel size and context radius among others). To maximize computational efficiency, we pre-determine these memory requirements for every beam before initializing the memory allocations. This allows us to instead allocate a single set of memory for these four arrays with sufficient size to fit all scenarios rather than repeatedly allocating and deallocating smaller memory segments and suffering from the significant overhead that such CUDA API operations impose on overall runtime. The single allocation is represented by a grey box in Figure 3-3c, and convolution along each kernel ray uses its own subset of this memory (white box), dependent upon the rotated geometry of the context array. To provide flexibility of our method to GPU hardware with lower available memory, we further implement optional *beamlet batching* which divides the complete context array for a beam into two or more sub-arrays to process successively. We allow explicit control over the number of batches for all beams when desired, and otherwise, dynamically detect when GPU memory restrictions necessitate batching for each beam on an individual basis.

3.2.1.1.4 Dose Ray Convolution

For each instance of the ray-specific REV-aligned context array, line convolution is carried out over the rows of the REV-aligned context array for every voxel along the kernel ray. By design, the density and TERMA accessed by the voxels in each row are restricted to the values coincident with each ray. This data is cached into shared memory for fast repeated access by neighboring voxels, offering hundreds of times less latency than global memory on

average[75]. GPU thread race conditions are avoided by treating the CCCS operations from the "dose deposition point of view"[47,66], enabling each thread to assign to its own voxel-specific memory address without conflicting with the data write operations of other threads. Each convolution is performed on a nonvoxel basis with linear interpolation using the cumulative kernel (CK) technique developed by Lu[76] and summarized by Neylon[61]. To obtain the full dose distribution for each beamlet, this line convolution procedure is performed along each kernel ray, and the dose from each is transformed into a common orientation and accumulated.

3.2.1.1.5 *Beamlet Context Dose Extraction*

The dense dose distribution attributed to each context's beamlet is stored in the context array in the *common orientation*. The selection of context radius determines the physical spatial extent to which the scattered dose is recorded. To represent this dose distribution in the original coordinate system, a beamlet specific affine transformation is applied to each context and the dose data is converted to a sparse representation in a two-column coordinate list (COO) format. One column contains the linearized volume index of each non-zero element, while the second column contains the corresponding value (dose). A threshold may be configured at this stage to exclude elements of negligible magnitude to further reduce storage size and improve data storage speeds. After conversion to the COO format, the dose data is written in a widely supported and flexible binary format (HDF5) to disk to be recalled and used during treatment planning.

3.2.1.2 Distributed Parallelization

Additional high-level parallelization of the algorithm is achieved by embedding our context-based approach into a distributed multi-GPU framework. We harness the trivial

separability of per-beam processing to build a network of computational workers, each of which may provide one or more GPUs. The division of labor among the worker nodes is simple and flexible with respect to the number available and is based on the computation of each beam as a standalone labor unit.

The beamlet dose calculation task is first executed on a managing node whose job is to prepare the static data (CT and configuration) and assign per-beam processing tasks to the workers. The manager node considers the availability of worker nodes and the number of GPUs provided before transferring the requisite data and task assignments to each. Upon receipt, the worker further assigns per-beam tasks to its resident GPUs which each take responsibility for one beam at a time and run concurrently. Processing on each GPU proceeds as described in section 3.2.1.1. The resulting beamlet dose data is immediately transferred over the network to the managing node for inclusion in the user-facing HDF5 file. The process flow detailing the distributed parallelization structure is described in Figure 3-2b.

3.2.2 Measuring Computational Efficiency

To quantify the performance of our context-based GPU-CCCS method for beamlet dose calculation, we measured and compared its computational efficiency against an existing GPU-CCCS implementation[48] that calculates beamlet dose in sequence. Beamlet doses for two representative 4π plans were calculated with isotropic 2mm voxel sizes and the average calculation times for each beam were recorded. Each plan was composed of the same 1,162 beam specifications distributed spherically around prostate and lung PTV definitions in two distinct CT geometries. The total number of beamlets for each plan was dependent upon the PTV shapes; 434,670, and 302,643 beamlets were calculated in total by each method in the prostate and lung targets respectively. Both our context-based GPU-CCCS and the existing

beamlet-sequential GPU-CCCS implementations shared CCCS quality settings that were set to match one another, such as the number of convolution rays ($N_\theta \times N_\phi$). The additional beamlet context radius parameter of our method was tested at 1, 2, and 3cm to demonstrate the flexibility provided in balancing speed and accuracy. The per-beamlet calculation bounding box margins for the sequential GPU-CCCS method were matched to the context radius to control for the effects of calculation over reduced volumes of different sizes when measuring the performance. N_θ and N_ϕ were set to 8×8 and 16×16 for both methods to quantify computational efficiency in these two common configurations. We also provide results for our method in distributed configurations with 1, 2, and 3 networked worker nodes, each employing two GPUs for a total workforce of 2, 4, and 6 GPUs respectively. To compare peak GPU memory usage for each of the sequential and context-based GPU-CCCS methods, 20 random non-coplanar beam orientations were selected (10 from each of the prostate and lung CT geometries), and doses for rectangular fields composed of various quantities of 5×5mm beamlets were calculated. Isotropic 2mm voxels and 16×16 convolution rays were configured throughout testing, and the context radius of the context-based GPU-CCCS method was additionally varied between 0.3cm and 3cm. For each set of quality parameters, the same 20 beam orientations were processed in a single program execution and the peak GPU memory usage was recorded.

3.2.3 Measuring Dosimetric Accuracy

Accuracy comparisons between the beamlet-sequential GPU-CCCS algorithm[48] and the NVB algorithm[61] of which our method is an extension have already been analyzed and will not be repeated here. Instead, we provide an investigation of the accuracy of our context-based method against Monte Carlo and a reference CPU-CCCS implementation in two

phantom geometries: one homogeneous water, and one heterogeneous stack of slabs, each detailed in Figure 3-5. Monte Carlo dose was obtained using Geant4 for a continuous emission spectrum of a diverging square monoenergetic photon beam, fit to the discrete spectrum used by CCCS. Monoenergetic dose kernels used in our context-based GPU-CCCS and the CPU-CCCS methods were previously synthesized[17] using an EGSnrc code, and the same emission spectrum was used to construct a polyenergetic kernel for dose convolution. Doses for 5mm, 1cm, and 2cm wide beamlets were calculated. A voxel size of 1x1x1mm^3 was selected to compare the dose profiles of all beamlets more accurately. For the context-based GPU-CCCS method, the context radius was fixed at 4cm. Central beamlet-axis percent depth dose (PDD) and lateral beamlet line profile at a depth of 10cm were visualized along with the error of our method from the CPU-CCCS and Monte Carlo results.

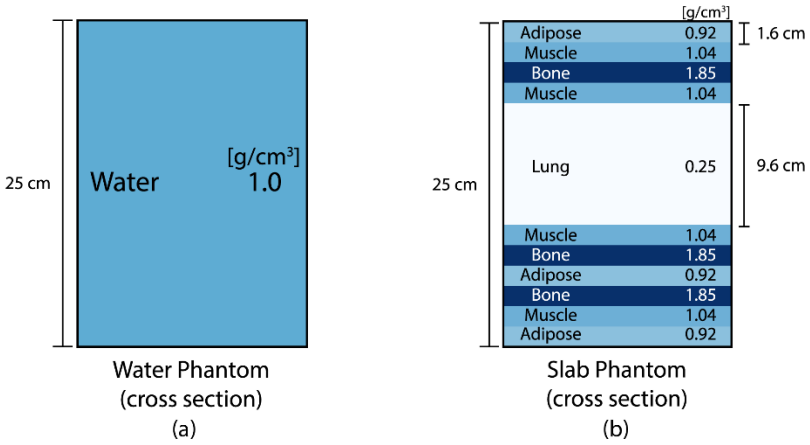


Figure 3-5. Cross sections of phantom geometries with beam entering from the top; used to assess dosimetric accuracy. Materials and densities are provided.

To support the assumption that beamlet dose can be well estimated by calculating only within a limited interaction context, we varied the context radius parameter between 0.1cm and 8cm for each of the three previously tested beamlet widths in the water phantom. From these experiments, the lateral beamlet profile at 10cm depth is included with the maximum

volumetric error for each pairing of beamlet width and context radius compared to dose for the same beamlet without using the context-based approximation by the beamlet-sequential GPU-CCCS method.

We also constructed a $5\times5cm^2$ broad beam by addition of context-based dose for 100 adjacent $5\times5mm^2$ beamlets in the water phantom and compared the broad beam lateral dose profile to that of a broad beam composed of beamlet dose calculated without use of the context-based approximation. Lateral dose profiles were analyzed at depths of 5cm, 10cm, and 15cm and the broad beam error associated with the context-based method was also reported.

Finally, the ability of our approach to scale to increased hardware availability and reducing the overall calculation time was investigated through timing experiments on multiple nodes and code execution profiling. Execution profile data were averaged across three independent application executions with 50 randomly selected 4π beams in each. All experiments were performed using NVIDIA GeForce TITAN X graphics cards from the Maxwell architecture. GPU programming was done using CUDA v9.0. For single-node performance evaluation, one node acted as both the manager and worker, employing an intel Xeon E5-2670 CPU with 8 physical cores and a base clock speed of 2.6GHz. For multi-node evaluation, workers having either an Intel i7-5820K CPU with 6 physical cores and a 3.3GHz clock, or an Intel i7-7700K CPU with 4 physical cores and a 4.2GHz clock were used. All data transfers between host and device memory were facilitated over 16-lane (x16) PCIe 3.0 interfaces.

3.3 Results

In Table 3-1 we present the time required by each algorithm to calculate beamlet dose for one beam averaged over the set of 4π treatment beams[48]. Our framework is implemented such that we measured calculation time for a single GPU as well as in various distributed multi-GPU configurations, emulating simple deployment scenarios. Figure 3-6 shows how the performance of our approach scales in single-node and multi-node configurations as a function of the number of GPUs utilized. The colored dashed lines show the scaling performance in the single-node configuration, where GPUs are simply added to an existing compute node. Colored solid lines indicate performance gains when GPUs on additional worker nodes are introduced instead.

Table 3-1. Per-beam Calculation Times (average, in seconds)

Treatment Site	Prostate		Lung	
$N_\theta \times N_\phi$	8×8	16×16	8×8	16×16
Single Node				
Sequential (1 GPU)	226.4	818.4	47.0	155.2
Context (1 GPU, 0.3cm context)	2.362	2.799	1.686	2.016
(1 GPU, 1cm)	3.066	6.322	2.312	4.958
(1 GPU, 2cm)	5.159	13.731	3.385	9.142
(1 GPU, 3cm)	9.119	29.765	5.343	17.137
Multi-node				
Context (1x 2 GPU, 2cm context)	3.130	11.977	1.964	7.466
(2x 2 GPU, 2cm)	1.643	6.288	1.031	3.919
(3x 2 GPU, 2cm)	1.127	4.312	0.707	2.688

Average per-beam dose calculation times (in seconds) for various hardware configurations, and quality settings

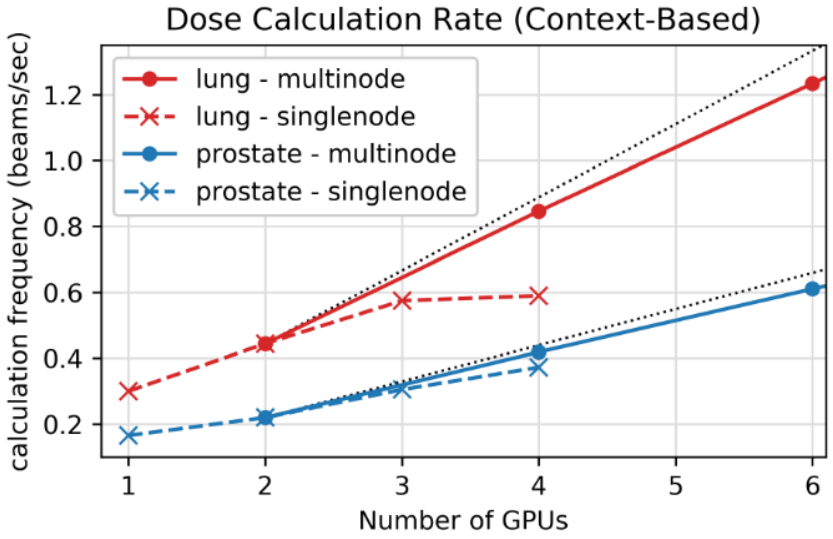


Figure 3-6. Performance for single-node and multi-node scaling strategies. For multi-node measurements, each node was configured with 2 GPUs. The dotted black line indicates theoretical linear scaling in multi-node setups.

A decomposition of our algorithm into the fractions of time spent on each sub-procedure is given in Figure 3-7 for various quality settings on a single node. Additional profiling results for use of various GPU counts on a single node are given in Figure 3-8.

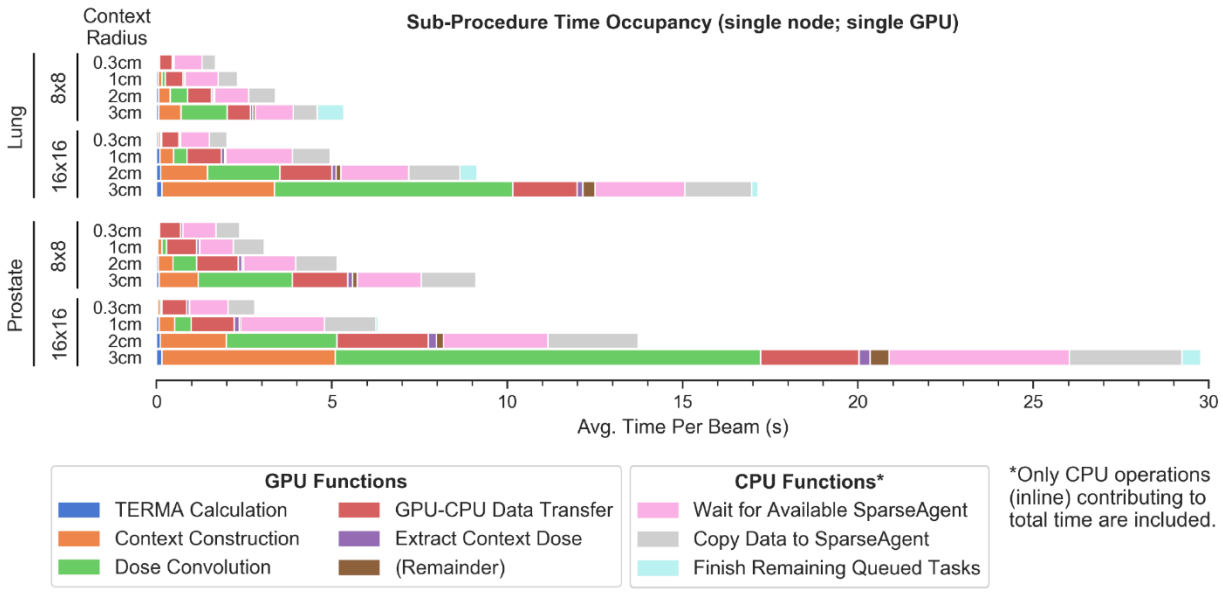


Figure 3-7. Fractional execution time spent in each sub-procedure on one computational node with 4 threaded post-processing "SparseAgents". Only time spent on the main processing thread is represented.

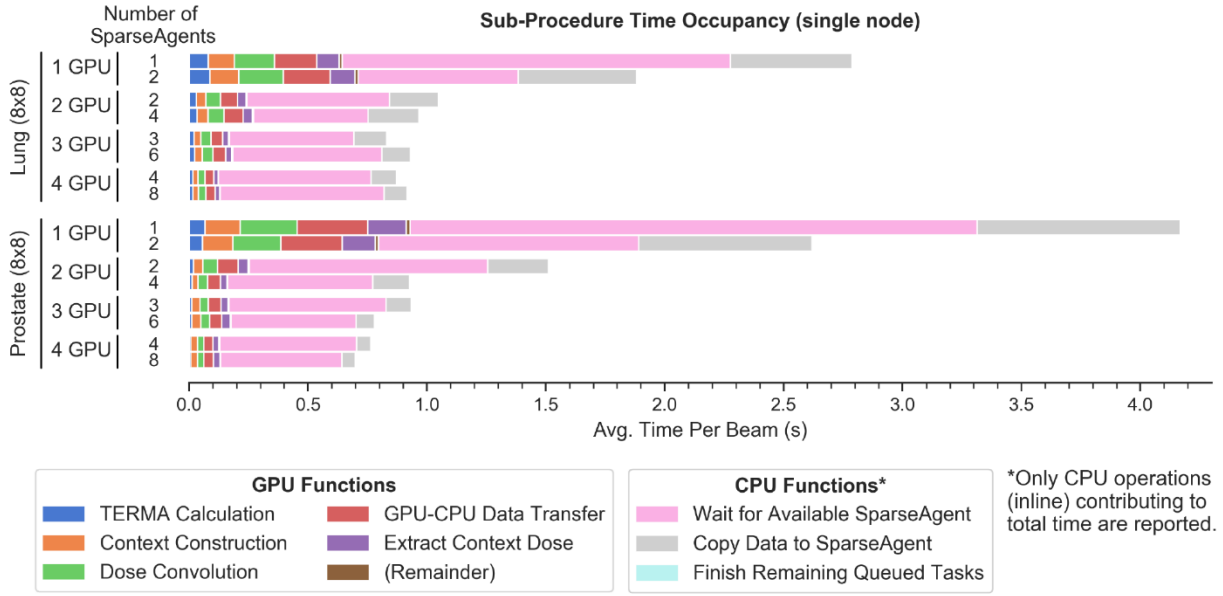


Figure 3-8. Fractional execution time for 1cm context radius with a variable number of background post-processing (dose sparsification) threads. Only time spent on the main processing thread is represented.

The memory usage recorded for both the sequential and context-based GPU-CCCS methods are listed in Table 3-2. Figure 3-9 shows these results in graphical form.

Table 3-2. Peak Memory Usage For GPU-based CCCS Methods (in Megabytes)

	Context-Based GPU-CCCS (by context radius)				Sequential GPU-CCCS
	0.3cm	1cm	2cm	3cm	
50 beamlets	200.07	244.23	528.34	1053.16	959.03
100 beamlets	188.12	382.64	980.64	2136.38	973.84
150 beamlets	255.09	618.59	1774.54	4014.71	985.25
200 beamlets	323.60	864.74	2643.31	5634.15	1001.78

GPU memory usage for sequential and context-based GPU-CCCS methods for 2x2x2mm^3 voxels and 16x16 convolution rays.

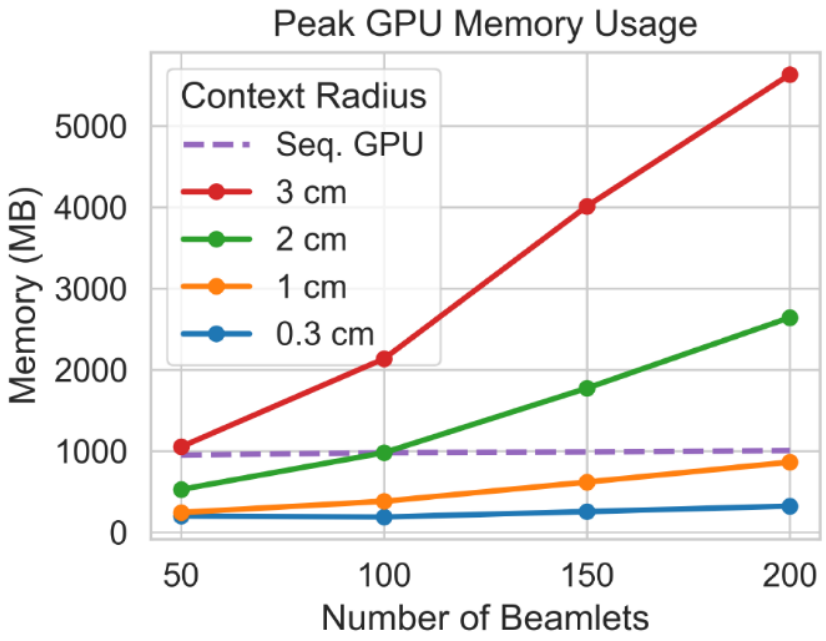


Figure 3-9. Peak memory usage for various beamlet counts and context radii.

Figure 3-10 and Figure 3-11 give the PDD along the beamlet's central axis, the centered lateral line profile at 10cm depth, and the error for each result compared with the CPU-CCCS and Monte Carlo methods. Results of our approach are given in color for each beamlet size in both the water and stacked slab phantoms. Monte Carlo outcomes are presented in grey, and CPU-CCCS dose is given as a dotted curve. Normalized error between our method and each of the Monte Carlo and CPU-CCCS methods are additionally given.

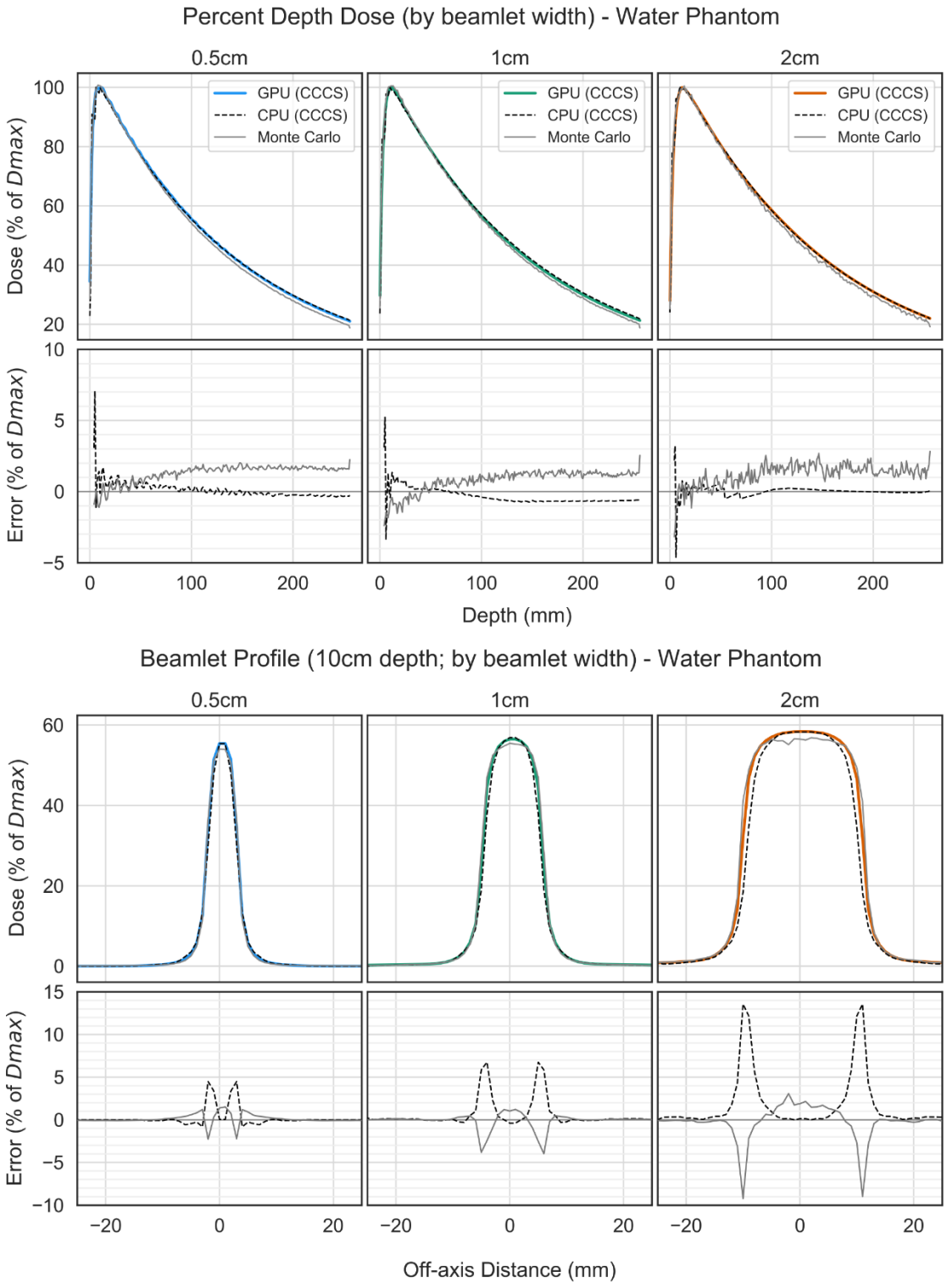


Figure 3-10. Single-beamlet depth dose and lateral profiles in the water phantom for increasing beamlet widths. Error is calculated between our context-based GPU-CCCS method and each of CPU-CCCS and Monte Carlo.

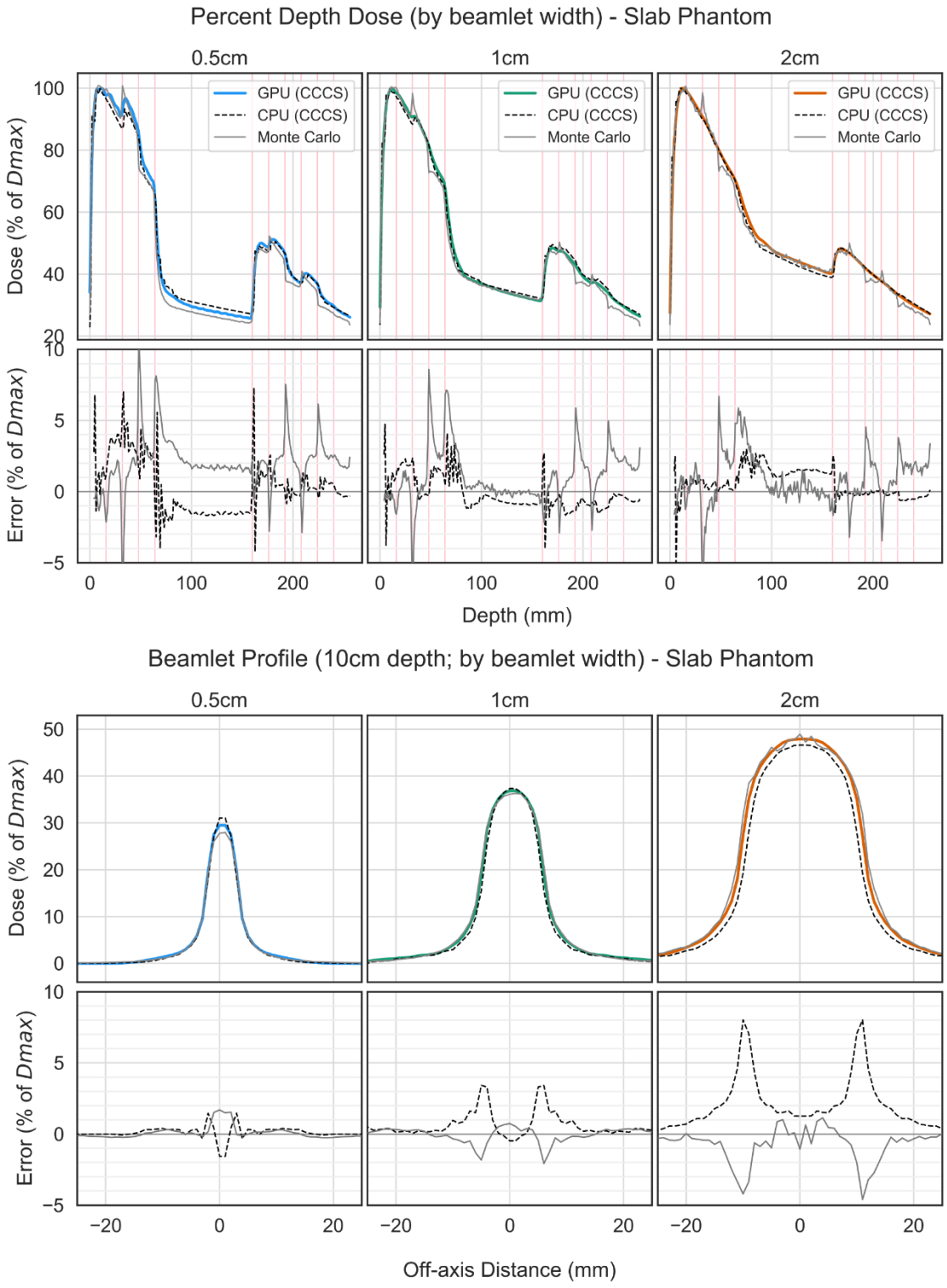


Figure 3-11. Single-beamlet depth dose and lateral profiles in the stack of slabs phantom for increasing beamlet widths. Error is calculated between our context-based GPU-CCCS method and each of CPU-CCCS and Monte Carlo.

The maximum error of the context-based compared to non-context-based beamlet dose resulting from various selections of context radius for the water phantom is given in Figure 3-12 with the resulting lateral line profile at 10cm depth. Absolute errors are expressed as percentages of the maximum reference volume dose calculated without the context-based approximation.

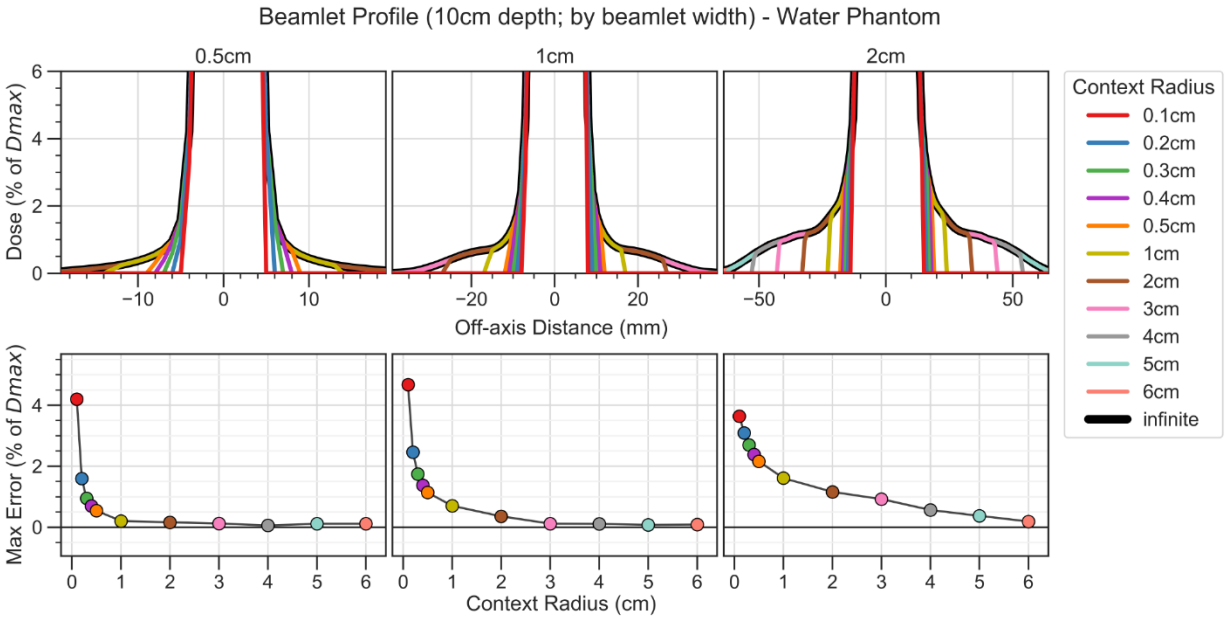


Figure 3-12. Central lateral line profile in the water phantom at 10cm depth for various beamlet widths and context radii pairs (top). Maximum errors (%) between non-context-based (infinite radius) and context-based dose profiles are provided (bottom). The dose is normalized to the maximum dose in the volume. Y-axis range is limited to better depict low dose beamlet penumbra region where context-based approximation is active.

3.4 Discussion

3.4.1 Performance

The performance improvements offered by our context-based method over the beamlet-sequential GPU-CCCS method are evident from Table 3-1. When a 2cm context radius is used on a single GPU for both methods, our approach offers 44-60x speedup and 14-17x speedup for the prostate and lung plans respectively. These results demonstrate a clear efficiency

advantage of our context-based processing. Even for a larger context radius of 3cm, ours demonstrates 25-28x and 9x speedups over the beamlet-sequential GPU-CCCS method configured with its beamlet dose calculation margin matching the context radius; demonstrating pure acceleration without truncation-induced loss of dose fidelity. Analysis of the error reported in Figure 3-12 shows that less than 1% beamlet-specific PDD error could be achieved in the water phantom for the 5x5mm^2 beamlet width by setting the context radius to just 3mm. Targeting this beamlet dose error of less than 1% we additionally timed our approach on the 4π dose calculation task using the reduced 3mm context radius. In this test, our method demonstrated even greater single GPU acceleration rates of 95-292x and 28-77x compared to the beamlet-sequential GPU-CCCS baseline for the prostate and lung plans respectively.

With its simplicity in scaling, our approach was also configured for multi-node calculation. When distributed across three workers employing two GPUs each, for a total of only six GPUs, we measured acceleration factors of 190-200x and 58-66x for the prostate and lung plans respectively, using a 2cm context radius. Applying the results of the single GPU experiments, we expect even greater accelerations for the dosimetrically similar 3mm and 1cm context radii. Figure 3-6 presents complete evidence of our framework's scalability, reaching nearly linear efficiency gains in the number of GPUs used in a multi-node configuration. Network latency didn't contribute significant overhead in our testing. However, we believe multi-node scaling performance can be further enhanced by utilizing dedicated 10 Gigabit inter-node connections to increase the network communication bandwidth over the 1 Gigabit connections used during testing but will leave such confirmation for future work.

Looking more closely at Figure 3-6, we observed that the single node performance gains quickly plateaued beyond use of three GPUs for the lung target. A performance bound isn't observed for the prostate, since we did not have the resources to test the single-node configuration with more than four local GPUs. We hypothesized that as the number of GPUs increased on a node, the beamlet dose of multiple beams is computed much more quickly than it can be transferred to the host, converted to a sparse format, and stored to the hard drive (collectively referred to as *post-processing*). By limiting the number of GPUs on each device, and instead increasing the number of computational nodes, we spread the GPU computational resources across more CPUs and output disks, and better balance the computational speed with the post-processing speeds. We tested this hypothesis for both treatment sites by distributing GPUs over more nodes and found that the multi-node configuration reduces the effects of the bottleneck, overcoming the undesired plateau of performance scaling seen in Figure 3-6.

To confirm our hypothesis moreover, we analyzed our algorithm using standard code-profiling techniques. Figure 3-7 indicates that the only sub-procedures with strongly-dependent runtime contention as the context radius and quality of the dose increase are *Context Construction and Dose Convolution*, both implemented on GPU. This is expected since these functions are dependent on the size of the context array which is directly affected by manipulation of the context radius. Unlike the former two GPU operations, which are executed once for each convolution ray, the *Extract Context Dose* operation is executed once for every beamlet to transform each set of computed beamlet dose data from the context array (in the arbitrary *common orientation*, introduced in section 3.2.1.1.3) to the original

coordinate system. This is a simple transformation and is made efficient by coalesced GPU memory access from the context array.

Undesirable, however, is the observation that copying the dose data from the GPU to the host memory *(GPU-CPU Data Transfer)* follows a weakly scaling trend, indicating that even as the quality of the dose (context radius) is reduced, the total computation time approaches a lower bound, in part determined by the memory transfer bandwidth between the host and GPU device. The other, more dominant factor determining the efficiency bound is the speed of post-processing (dose sparsification and storage). Since these tasks place post-processing requests in a fixed-size queue and are handled by a team of SparseAgents in separate CPU threads, we only see these operations contribute to the total runtime (*CPU Functions*) when the GPU outpaces the CPU. When this occurs, GPU computation is paused to limit the host memory usage while the post-processing queue is sufficiently depleted. The combination of the W*ait for Available SparseAgent* and *Copy Data to SparseAgent* operations indicate the amount of time that the main processing thread must wait while the occupancy of the post-processing queue is reduced. This type of delay is most pronounced when many GPUs are available on each node. This limit is demonstrated in Figure 3-8 where we see that nearly every GPU operation shortens in aggregate as more GPUs are added to a node, while the inline post-processing operations initially shorten as more threaded SparseAgents are provisioned but quickly exhibit diminishing returns; further supporting our hypothesis that the single-node algorithm performance is bounded by the post-processing time consumed on each worker. This time is in turn dominated by CPU core availability, as well as hard drive write and network transfer speeds, that can potentially be alleviated by increasing network

transfer bandwidth and distributing GPU resources over more computational nodes, as suggested.

Using the memory usage results, reported in Table 3-2 and Figure 3-9, we show that the primary factors determining the GPU memory usage of the context-based GPU-CCCS method are the number of beamlets in each beam, and the selected context radius, in addition to the general considerations such as voxel size and patient size (determining the beamlet length) common to all CCCS methods. We observed that for some combinations of beamlet count and context radius, the peak GPU memory usage was similar. Further investigation of these cases confirms our intuition that for each, a context array (Figure 3-3a) of similar size and shape was constructed. As expected, the peak memory usage of the sequential GPU-CCCS algorithm shows no impactful dependence on the number of beamlets in each beam, due to the sequential calculation of beamlet dose inherent to the technique. A weak correlation was observed but is insignificant and likely caused by changes in bookkeeping and the geometry of the calculated beamlets as the field size changes and intersects with different volumes of the CT. As described in section 3.2.1.1.3, we've developed the context-based GPU-CCCS method with optional dynamic beamlet batching to alleviate high memory usage concerns for cases such as that with 200 beamlets and 3cm contexts, showing peak usage of 5.6GB. With this feature, we hope to increase compatibility with budget-friendly GPUs providing less total memory.

3.4.2 Accuracy

Like the NVB algorithm on which we've based the core of our algorithm, calculated dose closely agrees with the CPU-CCCS calculated dose in the water phantom (Figure 3-10), with maximum single-beamlet PDD errors of 2% beyond the high dose gradient region found in

the first few millimeters of the phantoms. Single-beamlet lateral dose profile errors in the water phantom are greatest at the beamlet edges where high dose gradients are again observed. Inspection of the beamlet profile errors in Figure 3-10 indicate that the context-based method consistently displays smaller error in these regions when compared to Monte Carlo dose than when compared to CPU-CCCS dose, likely due to the use of TERMA super-sampling that has been employed in the context-based method to this effect. Errors in profile dose in the primary portion of the beamlets are below 2% on average between context-based and Monte Carlo methods for all beamlets sizes. This small error results from slight depth-dependent difference seen in the beamlet PDDs in the water phantom geometry (Figure 3-10), likely caused by the use of a continuous beam spectrum in Monte Carlo simulation rather than a discrete spectrum as in CCCS.

Single-beamlet dosimetric errors observed in the slab phantom (Figure 3-11) are slightly larger overall than those found in the water phantom. The greatest deviations of the context-based method from Monte Carlo beamlet dose occurs after interfaces between media of substantially different densities (particularly at depths of 32, 64, and 160mm), an effect attributable to the well-known shortcomings of the heterogeneity correction used in the CCCS method that have already been independently investigated[28,77,78]. Closer agreement of our context-based GPU-CCCS method with the reference CPU-CCCS implementation at these interfaces support this explanation. Despite these inherent shortcomings in the CCCS algorithm, the context-based method shows average single-beamlet PDD errors of magnitude less than 1.35% and 2.35% for all beamlet sizes in the water and slab phantoms, respectively.

The dose truncation effects of our context-based approach are evident in Figure 3-12 where its resemblance to cylindrical kernel truncation[73] in the lateral direction is clear. In line with our expectation, the maximum single-beamlet profile error decreases quickly as the context radius is increased. The rate of decrease in the error is smaller as the beamlet size is made larger, because more dose is physically scattered outside of the primary beam, as observed by the longer tails of the 2cm wide beamlet compared to the 0.5cm beamlet. The profile error for the $5x5cm^2$ wide broad beam, composed as a sum of 100 $5x5mm^2$ wide beamlets, presented in Figure 3-13, shows that full beam dose profile errors below 5%, and 10% can be expected for context radii above 2cm and 1cm, respectively. The nature of the context-based method makes it difficult to directly truncate the polyenergetic dose kernel and renormalize its remaining weights to sum to 1, and thus, energy is not strictly conserved in the current implementation. We instead recommend the intuitive use of a small context radius when the beam candidate pool is large, such as in early stage of automatic beam orientation optimization which considers over 1,000 beams. Approximate dose is often sufficient for ruling out trivially unsuitable beam orientations and the context radius can be increased to recompute more accurate beamlet dose once the beam candidate pool has been reduced.

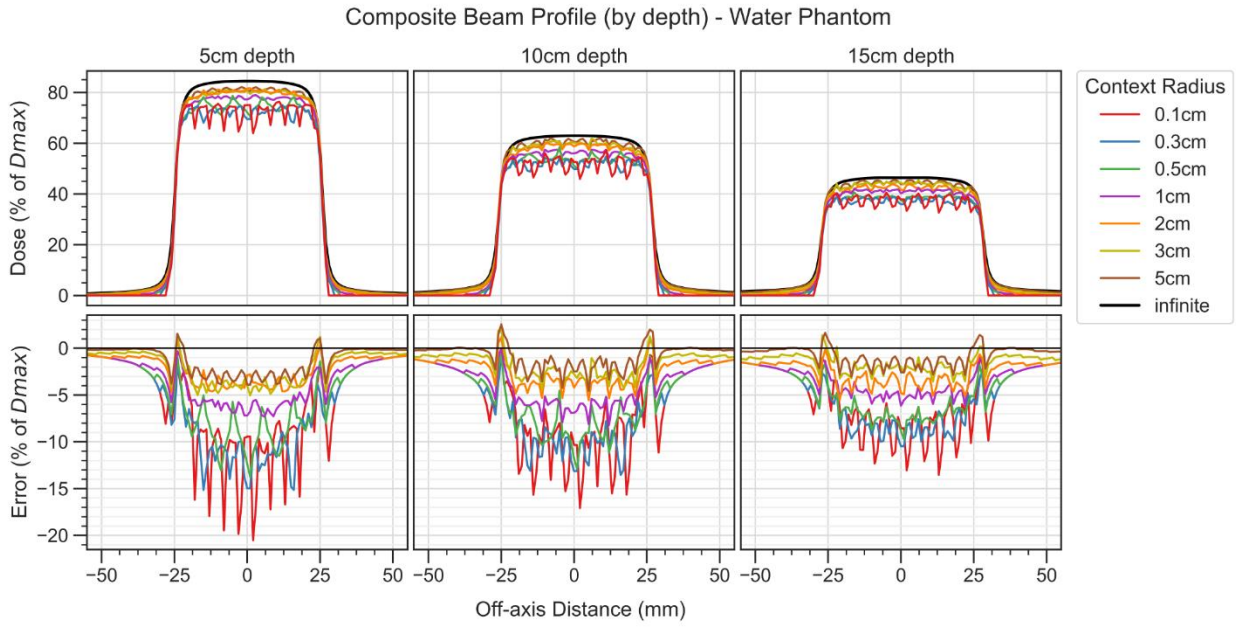


Figure 3-13. Lateral line profile in the water phantom at various depths for a 5×5cm² broad beam calculated as the sum of 5×5mm² beamlets for various context radii at three depths. Dose for "infinite" radius was computed without context-based approximation.

Our current implementation relies on the user-defined context radius which corresponds to a physical path length from the edge of a beamlet. We have also considered dynamically setting the context radius of each beamlet in response to local density patterns. By so doing, beamlets in homogeneous high-density environments would be assigned low radii to match the short radiological path lengths, while those in low density or heterogeneous environments would be assigned higher radii. This adaptive approach would allow more optimal allocation of computational resources to beamlets where distant dose scatter is expected; this work, however, has been left for future investigation.

3.5 Applications

Our accelerated context-based GPU-CCCS dose calculation method has been used to enable effective planning of several new treatment paradigms for which the VLS beamlet dose calculation requirements have traditionally been restricting. Here we discuss a

selection of applications for which our accelerated GPU-CCCS method has supplied the computational impetus.

3.5.1 A sparse orthogonal collimator for small animal intensity-modulated radiation therapy[79,80]

3.5.1.1 Background

Advances in the precision of therapeutic radiation delivery, such as IMRT, have greatly increased the dose conformity and normal tissue sparing, but the clinical benefits from improved dose distributions alone will likely plateau because the limits in radiation tolerance for normal tissues may halt further improvement in the treatment outcome for radioresistant tumors[81]. The next major advancements in radiation therapy are believed to come from developments in knowledge of the underlying radiobiology of healthy and cancerous tissues. In order to better understand which novel treatment techniques will translate to improved clinical outcomes in humans, preclinical validation is required.

To meet this need, preclinical research using small animal models, particularly mouse models, have been used extensively as inexpensive and versatile tools for gaining valuable insight into cancer growth dynamics and the biological effects of radiation[82,83]. However, due to the inconsistencies and inadequacies of irradiation techniques used in several preclinical radiotherapy studies, the use of animal models has shown limited benefit[84–90]. Recent development of dedicated image-guided small animal irradiators, like the X-RAD smART system from Precision X-ray Inc., has greatly expanded the potential for preclinical radiotherapy research[91]. To ensure better translation of preclinical research to clinical application, the modern dose modulation techniques used for IMRT must be emulated in the preclinical setting. For this purpose, we've introduce a new method to delivery small animal

IMRT using the previously described sparse orthogonal collimator (SOC)[92] that can be more easily miniaturized for the small animal scale, and a planning system that supports the use of this simple, yet powerful, dose modulation device.

3.5.1.2 Methods

This novel small animal IMRT method is based on the idea that by adapting the previously formulated direct aperture optimization (DAO) method[36], effective IMRT treatments using only rectangular apertures can be delivered. For treatment planning of SOC-based IMRT we have formulated the Rectangular Aperture Optimization problem (described previously[92]) as

$$\begin{aligned} \underset{\alpha}{\text{minimize}} \quad & \frac{1}{2}\|W(AR\alpha - d_0)\|_2^2 + \lambda\|\alpha\|_1 \\ \text{subject to} \quad & \alpha \geq 0 \end{aligned}$$

Equation 3-4

The optimization variable, $\alpha \in \mathbb{R}^P$, is a vector encoding the optimal fluence for each of the P pre-defined apertures, $R \in \{0,1\}^{NxP}$ is the binary matrix that maps the per-aperture fluence values from α into N per-beamlet fluence values($f \in \mathbb{R}^N$ as presented in Equation 2-4), $A \in \mathbb{R}^{MxN}$ is the standard beamlet dose matrix mapping beamlet fluence to a 3D dose distribution for the plan, and $W \in \mathbb{R}^{MxM}$ is a diagonal weighting matrix defining the relative importance of meeting the ideal dose (d_0) for every voxel or anatomical structure (set of voxels). The L1 regularization term imposing sparsity on α is designed to limit the number of apertures and make the plan's delivery more efficient. A complete description of the optimization procedure is provided by Nguyen *et al*[92].

To ensure accurate planning and radiation delivery, a series of beam commissioning measurements were performed on the X-RAD smART irradiator. Analysis of these

measurements revealed imperfections in the spectral quality and uniformity of the generated X-ray beam. Consequently, in addition to X-ray fluence, the dose delivered by each aperture becomes a function of its size and position in the field of view. The additional dependence required consideration of dose deposition for multiple beam energy spectra, characteristic of each aperture size, greatly inflating the dose calculation requirements for this application. To enable an accurate solution to the problem in Equation 3-4, compatible with the system imperfections, our multi-GPU CCCS dose calculation method was employed to calculate the planning dose data for every beamlet using five different beam energy spectra; introducing a five-fold increase to the standard dose calculation requirement of IMRT. The five energy spectra were defined by estimating the distribution of X-ray energy for five square apertures with side lengths equal to 1, 3, 5, 10, and 20 mm.

Equation 3-4 was solved five times in parallel, once for each energy spectrum, to arrive at five versions of α, each containing a distribution of aperture sizes. Automatic selection of between 5 and 20 beams from a candidate pool of 180 coplanar beams was performed using the 4π beam orientation optimization method, which has been previously published extensively[48,49,99,50,54,93–98]. For each subject, the total number of beamlets requiring dose calculation was as many as 1.5 million, which is over 2 orders of magnitude more than is typically required for a clinical IMRT plan with 7 manually selected beams. For each plan, the apertures were grouped according to their aperture size. Then, each plan's deliverable dose was calculated by summing the dose for all groups, each recalculated using the energy spectrum for the square aperture most closely matched in size to the apertures of the group. From the accurate deliverable dose, the apertures selected in the first optimization were held fixed, and their fluences were re-optimized. This important fine-tuning step enhances the

plan quality from the first optimization, where all apertures assumed a single beam energy spectrum rather than the one most appropriate for their size.

Three dosimetric experiments were conducted in which SOC plans for a C-shaped concave target, a mouse whole liver, and a highly modulated "Audrey" plan, were created using the procedure described in the previous section. Each plan was delivered to pre-calibrated, dose-sensitive Gafchromic EBT3 film using the X-RAD smART system. Radiologically analogous phantoms (Figure 3-14) were 3D-printed using a tissue-equivalent flexible material with integrated slots for placement of the EBT3 film. The measured 2D dose distributions were compared with the calculated deliverable dose distributions obtained from our multi-GPU CCCS dose calculation method on the basis of dose statistics (minimum, maximum, and mean) and using a gamma analysis[100] to produce the percentage of voxels with a sufficiently high gamma index to be considered as *passing* according to the 4%/0.3mm gamma criteria.

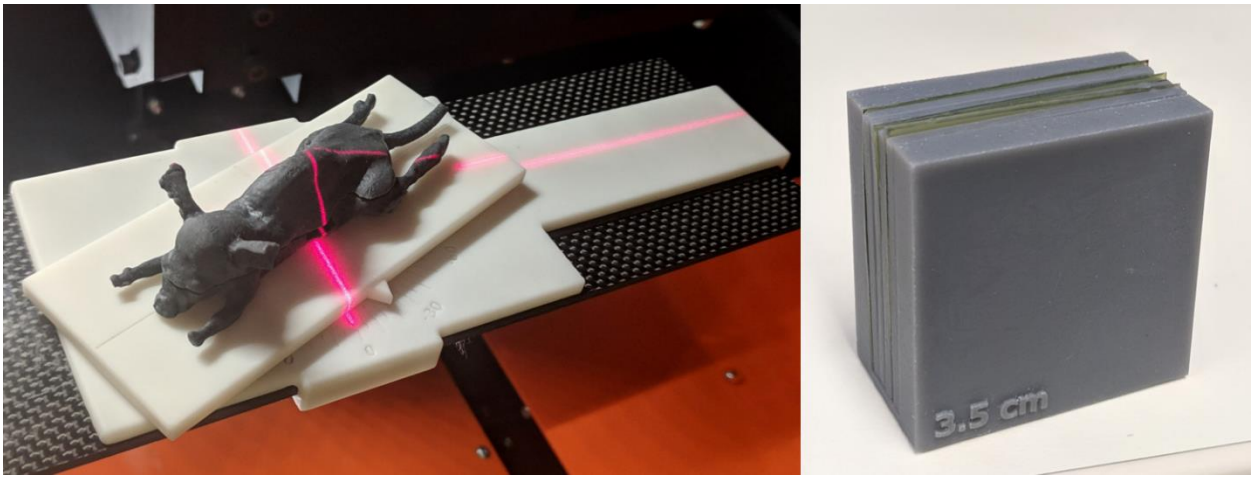


Figure 3-14. (Left) Mouse phantom modeled from mouse CT data and 3D-printed with a flexible, tissue-equivalent material and a mid-coronal split for film measurement. Phantom is shown on the previously mentioned rotating couch mount. (Right) 3D-printed block phantom for axial dose measurements. Figure reproduced from Woods *et al.* (2019)[80].

3.5.1.3 Results and Discussion

Figure 3-15 shows the calculated (left) and measured (center) dose distributions for the C-shaped target plan. The gamma analysis with 4%/0.3 mm criteria revealed a pass rate of ~95% for pixels within the target structure, and 85% for the entire field shown. The maximum and mean absolute pixelwise dose differences were 4.12 and 0.59 Gy, respectively (Table 3-3), and the measured dose to the target had slightly higher maximum (15.8% of the prescription dose), mean (7.0%), and minimum (13.5%) doses. The 50% isodose lines are shown in Figure 3-15 (right), demonstrating excellent overall agreement between the calculated and measured dose distributions.

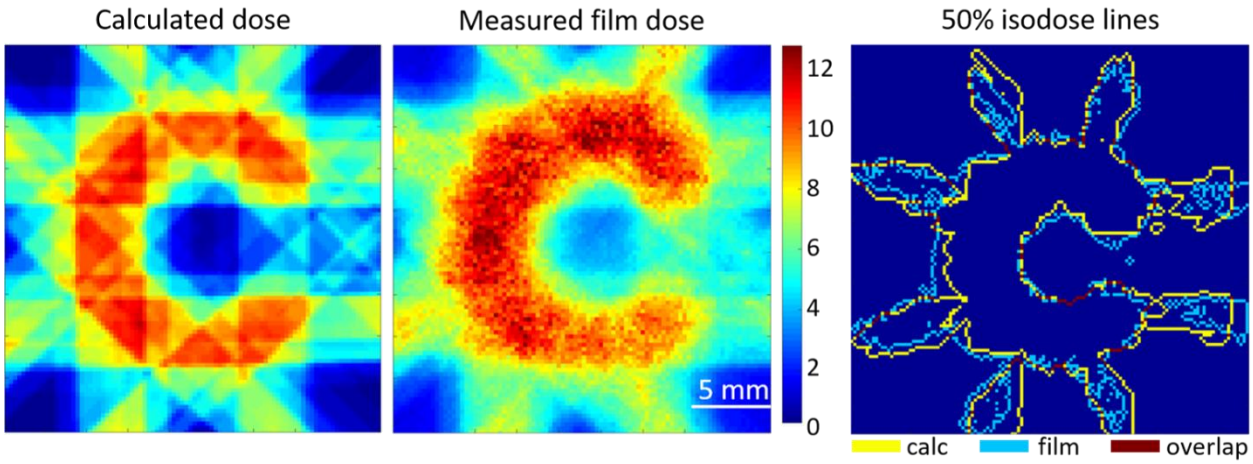


Figure 3-15. (Left) Calculated dose distribution of the C-shaped target plan perpendicular to the gantry rotation axis. (Center) Measured film dose distribution from the center of the solid water phantom for the C target plan delivered with the SOC. Both plans are shown with the same color scale, in units of Gy. (Right) A comparison of the calculated (yellow) and measured (blue) 50% isodose lines, with overlapping regions shown in red. Figure reproduced from Woods *et al.* (2019)[80].

Table 3-3. Comparisons between the measured and intended dose distributions for the C-shaped target plan and the mouse phantom whole liver plan. Table reproduced from Woods *et al.* (2019)[80].

		Dose Statistics (measured – calculated)			Pixelwise Dose Comparison		
Plan	*Structure*	*Max*	*Mean*	*Min*	*Max Diff*	*Mean Diff*	*Gamma Pass Rate*
C Target	C	+1.58 (15.8%)	+0.70 (7.0%)	+1.35 (13.5%)	4.12 (41.2%)	0.59 (5.91%)	94.9%
Mouse Liver	Liver	+1.30 (13.0%)	-1.02 (10.2%)	+1.10 (11.0%)	3.50 (35.0%)	1.19 (11.9%)	98.2%
	Kidneys	+0.43 (4.3%)	+0.24 (2.4%)	+0.11 (1.1%)	0.43 (4.3%)	0.24 (2.4%)	100%

***Max, mean, and min dose differences written as [Gy (% prescription dose)]; Max and Mean Diff are absolute pixelwise dose differences; gamma analysis was performed with 4%/0.3 mm criteria for dose/distance**

The results of the mouse phantom liver test plan are shown in Figure 3-16, with the liver and kidney dose comparisons given in Table 3-3. The maximum measured liver dose was 13.0% higher than the calculated dose, the mean was 10.2% lower, and the minimum was 11.0% higher. As evident in the film dose distribution and isodose comparison shown in Figure 3-16 (C) and (D), the lower left portion of the liver was cut off due to slight phantom misalignment. The affected pixels were omitted from the liver dose analysis. For the unaffected pixels within the liver, the gamma analysis showed a high pass rate of 98.2%. The measured SOC plan was able to significantly spare the dose to the kidneys, with maximum and mean doses of 0.43 and 0.24 Gy, respectively. These are only 4.3% and 2.4% higher than the calculated doses, and therefore all pixel-wise differences were within 5% of the intended dose.

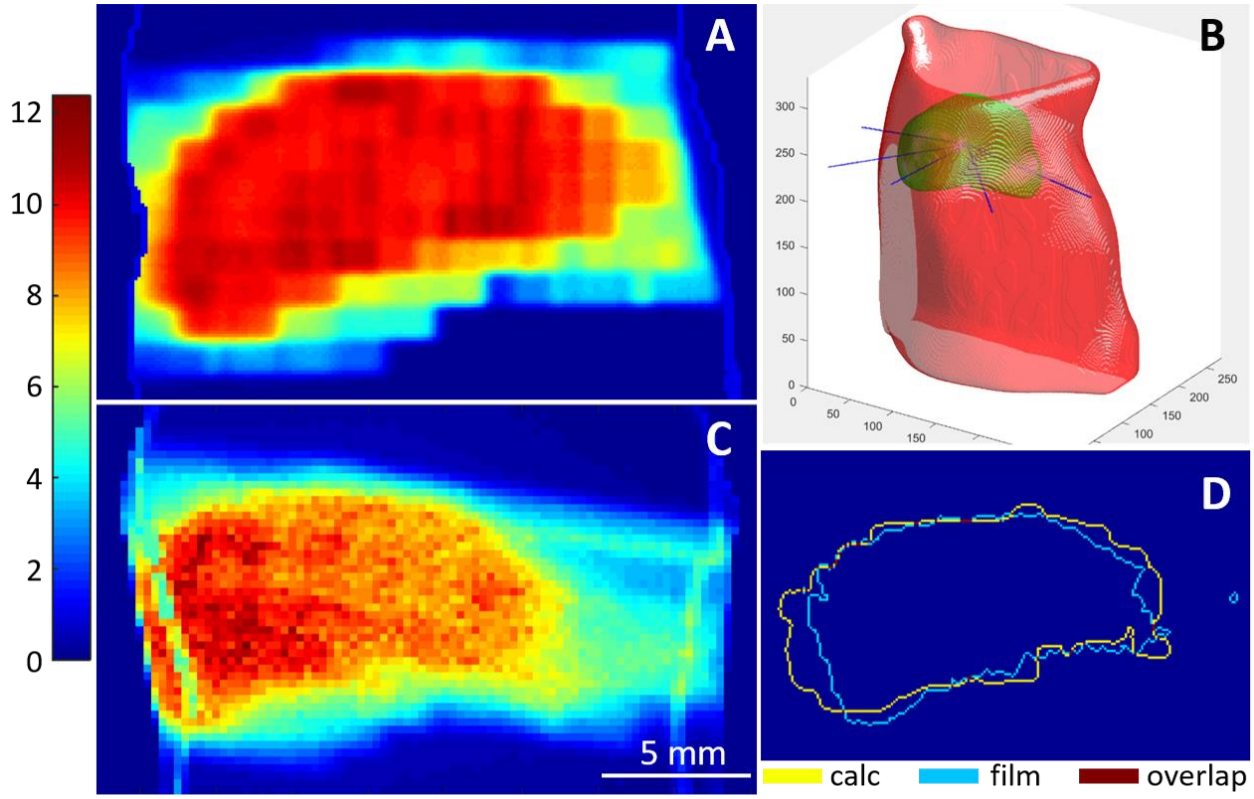


Figure 3-16. (A) Mid-coronal view of the calculated dose for the mouse phantom whole liver plan (units of Gy). (B) The 5 optimal coplanar beam angles selected with the 4π algorithm. (C) Measured film dose from the mouse phantom, treated with the whole liver plan, at the plane shown in A (units of Gy). (D) A comparison of the calculated (yellow) and measured (blue) 60% isodose lines, with overlapping regions shown in red. *Target structure was rotated to account for slight phantom misalignment, which also resulted in the truncated lower left portion of the target. Figure reproduced from Woods *et al.* (2019)[80].

The calculated and measured doses for the 2-dimensional Audrey test plan are shown in Figure 3-17. The maximum and minimum measured film doses were both 1.1 Gy higher than the calculated dose distribution (12.2% of the maximum intended dose), with a mean pixelwise absolute dose difference of 1.6 Gy. Although this plan shows some discrepancies in absolute dose prediction for very small apertures sizes, the sources of which are discussed in Woods *et al.* (2019)[79], the spatial distribution is extremely similar to the calculated plan, validating the overall accuracy of the SOC hardware and control software.

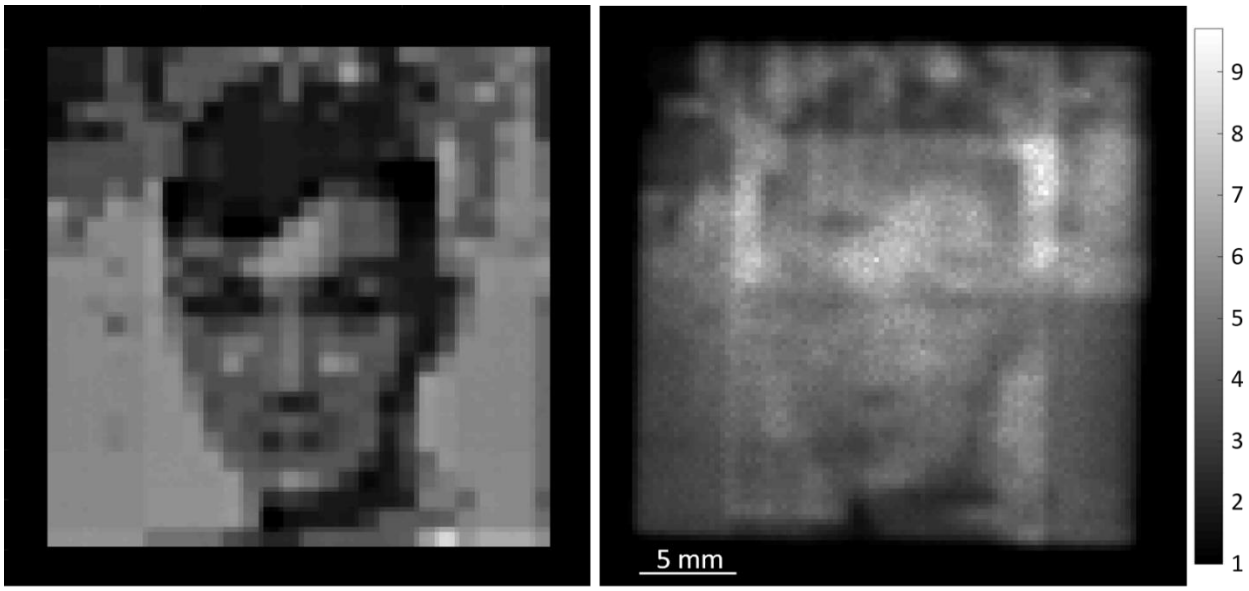


Figure 3-17. (Left) Calculated Audrey test plan with 4 dose levels and an average aperture size of 2.35 mm. (Right) Measured dose distribution of the Audrey plan delivered with the SOC. Both plans are shown with the same color scale, in units of Gy. Figure reproduced from Woods *et al.* (2019)[80].

All three SOC plans showed measured dose deposition that was consistent with the expectation, calculated by our GPU-CCCS method. Furthermore, the entire SOC treatment planning process can currently be completed for one subject in only a few hours, despite the substantial increase in required dose calculation data (approximately 128 times more) compared with traditional IMRT planning.

3.5.2 A novel optimization framework for VMAT with dynamic gantry couch rotation[40]

3.5.2.1 Background

Optimization of plans for Volumetric Modulated Arc Therapy (VMAT) is more challenging than for static beam IMRT due to its large problem size and more complex delivery constraints. Typically, 180 or more beams are included in VMAT delivery compared with fewer than ten beams used in a typical IMRT plan. More importantly, the gantry rotation and

leaf motion are coupled: for efficient VMAT delivery, the MLC leaf movements between adjacent gantry angles cannot exceed the product of the maximal leaf speed and the gantry travel time, which is short to maintain smooth and efficient gantry rotation. The heuristic progressive sampling optimization (PSO) method for VMAT[101] successfully addressed the mechanical constraint problem and kept the computational complexity manageable but fails to provide a globally optimal plans and greatly relies on interactive input from an experienced dosimetrist.

To overcome these limitations, a level-set-based direct aperture optimization for coplanar-arc VMAT was developed[37], which solves the entire arc optimization problem in full angular resolution. This non-progressive sampling approach was shown to generate a single arc coplanar VMAT that outperformed progressive sampling VMAT using two arcs with the same number of control points in each arc. Recent 4π IMRT research, which enables automatic noncoplanar beam orientation optimization (BOO) from a large candidate pool has demonstrated significant dosimetric gains compared to the VMAT plans[48,93,102].

In this study, we propose a novel optimization framework, termed 4πVMAT, that simultaneously solves the complete non-coplanar VMAT trajectory optimization and DAO problems for VMAT, while ensuring deliverability by avoiding couch-gantry-patient collision and enforcing mechanical constraints of MLC leaf motion and gantry rotation.

3.5.2.2 Methods

Our optimization framework takes an alternating approach between solving simpler subproblems for DAO, BOO, and beam trajectory selection (BTS). An objective was formulated to jointly solve the DAO and BOO sub-problems using the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA)[46] for the current solution to the BTS sub-

problem. The BTS sub-problem was modelled after a traveling salesman problem in which the goal is to discover the most efficient path through all the nodes of a graph based on the edge costs connecting them. Here the nodes were defined by the complete set of possible 4π beam from a discretized grid over couch and gantry angles. The edge costs were defined as positive infinity between any two nodes for which the motion connecting them during delivery was impossible. The BTS problem was solved using Dijkstra's algorithm, which is a dynamic programming method that can be used to identify the shortest path between two nodes of a weighted graph. For patient safety and comfort, the edge costs were designed to enforce a constant couch rotation direction within each arc while the gantry was permitted to rotate dynamically.

By alternating between solutions to the DAO&BOO and BTS sub-problems, a deliverable and dosimetrically desirable 4πVMAT plan can be discovered with the properties that the fluence map for each aperture is constant, only a single aperture is selected for each control point, the total number of control points is low, and the apertures are smoothly varying between adjacent control points. Optimization of a 4πVMAT plan considers 2400 candidate beams, for which beamlet dose must be calculated. Using a standard 20×20cm^2 field of view for each beam with 5×5mm^2 beamlets, the total dose calculation requirement is as high as more than 3.8 million individual beamlet doses imposing a significant strain on existing dose calculation approaches which have been developed for the clinically standard purpose of IMRT planning with typically fewer than 10 manually selected beams (up to only 16,000 beamlets) or coplanar VMAT with a substantially lower number of control points in total; typically only 80 per arc, and up to two or three arcs. Our GPU-CCCS beamlet dose calculation method was utilized to efficiently calculate the dose data required by the 4πVMAT

optimization. Plans for 4πVMAT and more traditional, manually selected coplanar arc therapy (termed 2πVMAT) were created for three glioblastoma multiforme, three lung cancer, and three prostate cancer patients. The 4πVMAT and 2πVMAT plans were compared on the basis of their plan quality (PTV and OAR statistics) and delivery efficiency.

3.5.2.3 Results and Conclusions

The 4πVMAT plans demonstrate their flexibility to distribute the dose in any non-colliding direction within the 4π spherical beam space, depending on the benefits to OAR sparing and PTV coverage, whereas 2πVMAT plans are restricted to dosimetrically inferior coplanar delivery that demands planning experience from the human dosimetrist to achieve acceptable but still inferior dosimetry compared with the 4πVMAT plans. The 4πVMAT plans were able to markedly reduce dose to OARs while achieving comparable or better PTV statistics across all patients, especially for the dose limiting organs, such as the brainstem in the GBM #2 and GBM #3, the proximal bronchus in all three LNG patients, the major vessels in LNG #2, and LNG #3, and the seminal vesicle and the rectum in all PRT patients. Optimization of the 4πVMAT plans is still somewhat slow due to the complexity in tuning hyperparameters compounded with the heavy computation costs at each iteration. Planning took between 1 hour for the GBM cases to 9 hours for the lung and prostate cases. Delivery of 4πVMAT plans was between 3 and 5 minutes for all plans which is comparable to the conventional 2 or 3 arc 2πVMAT plans.

Optimization of 4πVMAT plans in its current form is effective at improving plan quality while maintaining efficient delivery, but in its current form is computationally intensive due to the optimization problem size and alternating optimization between DAO&BOO and BTS. Some avenues for further acceleration of the optimization exist such as GPU implementation

of expensive matrix multiplications in each FISTA iteration and progressive up-sampling of the dose data in later iterations of optimization. Nonetheless, the conventionally time-consuming precursor to optimization, beamlet dose calculation, was significantly accelerated by our GPU-CCCS approach to span the gap between the typical dose calculation workload of coplanar arc VMAT and our new dosimetrically superior paradigm of 4πVMAT.

3.5.3 Single-Arc VMAT optimization for Dual-Layer MLC[44]

3.5.3.1 Background

Recently, significant advances have been made in VMAT planning algorithms[37,40,41] but the focus has only been placed on single-layer multi-leaf collimators (SLMLC) thus far. Dual-layer multi-leaf collimators (DLMLC)[103–105] with "stacked and staggered" leaves offer an alternative with two distinct advantages: substantially reduced inter-leaf leakage, and simpler and more cost-effective construction. Two new commercial medical linacs, Halcyon (Varian Medical Systems) and MRIdian (ViewRay) have adopted DLMLC. In this work, we have developed a VMAT optimization method that specifically manages the dosimetric effects of coupling between the two MLC layers by simultaneously solving for both MLC layers and the entire arc in full angular resolution, using an alternating optimization approach that has been investigated in previous studies[37,40,41].

3.5.3.2 Methods

In this study, we focus on optimization of plans for delivery of coplanar VMAT arcs specifically on the Halcyon, without loss of generality. In the Halcyon, the two layers of the DLMLC are stacked and staggered by half of the leaf width to provide more sophisticated modulation than SLMLC with the same leaf width. Figure 3-18a shows the DLMLC with stacked 10 mm wide leaves (DLMLC-10mm) compared to the SLMLC with 5mm leaves in

Figure 3-18b, and the SLMLC with 10mm leaves and either 10mm or 5mm longitudinal leaf positioning step size in Figure 3-18c-d (SLMLC-10mm and SLMLC-10mm-5mm). The DLMLC-10mm affords greater dose modulation capability than the SLMLC-10mm configuration, since one layer of the DLMLC-10mm can be left open to emulate the leaf positioning of the SLMLC-10mm, or both layers can be used for more complex collimation. But the coupling effect of the dual MLC layers expectedly adds complexity to the optimization of identical structures compared to the SLMLC.

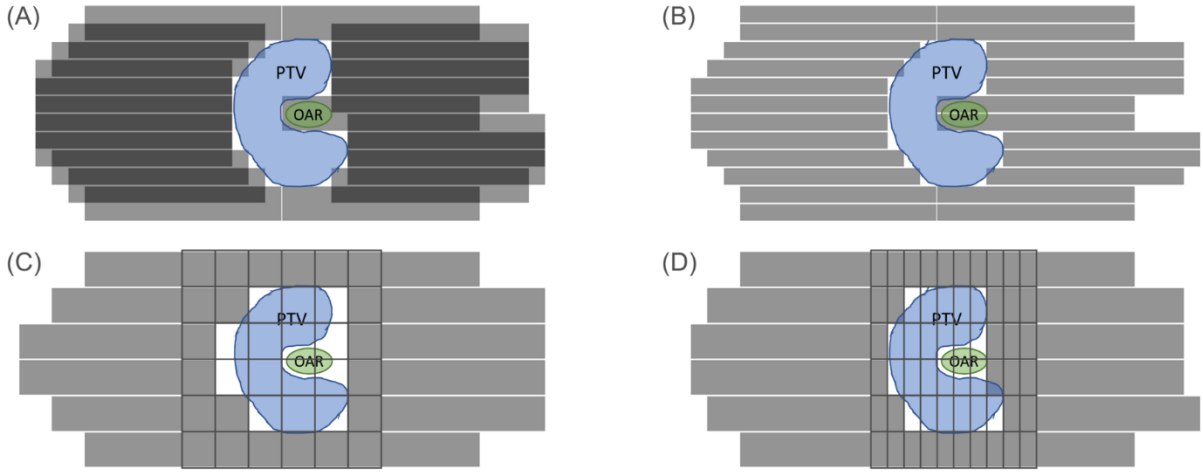


Figure 3-18. Demonstration of (A) DLMLC with 10mm leaf width (DLMLC-10mm), (B) SLMLC with 5mm leaf width (SLMLC-5mm), (C) SLMLC with 10mm leaf width (SLMLC-10mm), (D) SLMLC with 10mm leaf width and 5mm leaf step size (SLMLC-10mm-5mm). The grids on (C) and (D) represent the achievable beamlets. Figure reproduced from Lyu *et al.* (2019)[44].

To solve the DLMLC optimization challenge, a direct aperture optimization problem is formulated[44] to simultaneously optimize dose fidelity, and VMAT deliverability feasibility and efficiency. The optimization problem is solved via an alternating optimization approach by fixing all but one optimization variable in each module, and iteratively updating each variable conditioned on the current states of the rest. To enable efficient planning, our GPU-CCCS beamlet dose calculation method was used. For each of four plans, one glioblastoma

multiforme, one lung cancer, one prostate cancer, and one rectal cancer, 180 beams were selected with 2-degree spacing to form a full VMAT arc. The field of view for each DLMLC-10mm beam was divided into a 40×40 grid of 5×5 mm^2 beamlets, producing a total dose calculation requirement for approximately 290,000 beamlets, representing a 2.25-fold increase in dose calculation over conventional VMAT with only 80 control points-per-arc in general and a SLMLC with 5×5 mm^2. More substantially, however, is the 9-fold increase in beamlet dose requirement for DLMLC-10mm compared to conventional SLMLC-10mm with the same MLC construction cost. To compare plans based on MLC modulation capability alone, all plans were optimized with the same 180 control points.

3.5.3.3 Results and Conclusions

The proposed DLMLC VMAT optimization algorithm optimizes all beams simultaneously and produces 5mm-resolution apertures that are deliverable using DLMLC-10 mm. Figure 3-19 shows the DVHs of the DLMLC-10mm plan and the SLMLC plans for all patients. The DLMLC-10mm and the SLMLC-5mm plans are nearly indistinguishable. With the same leaf width, the DLMLC-10mm plan OAR sparing is superior to both the SLMLC-10mm-5mm and the SLMLC-10mm plans for all patients. Considering the Halcyon's maximum MLC leave movement speed of 50 mm/s, the estimated delivery time is 36 s for the GBM and lung cases, and 90 seconds for the prostate case.

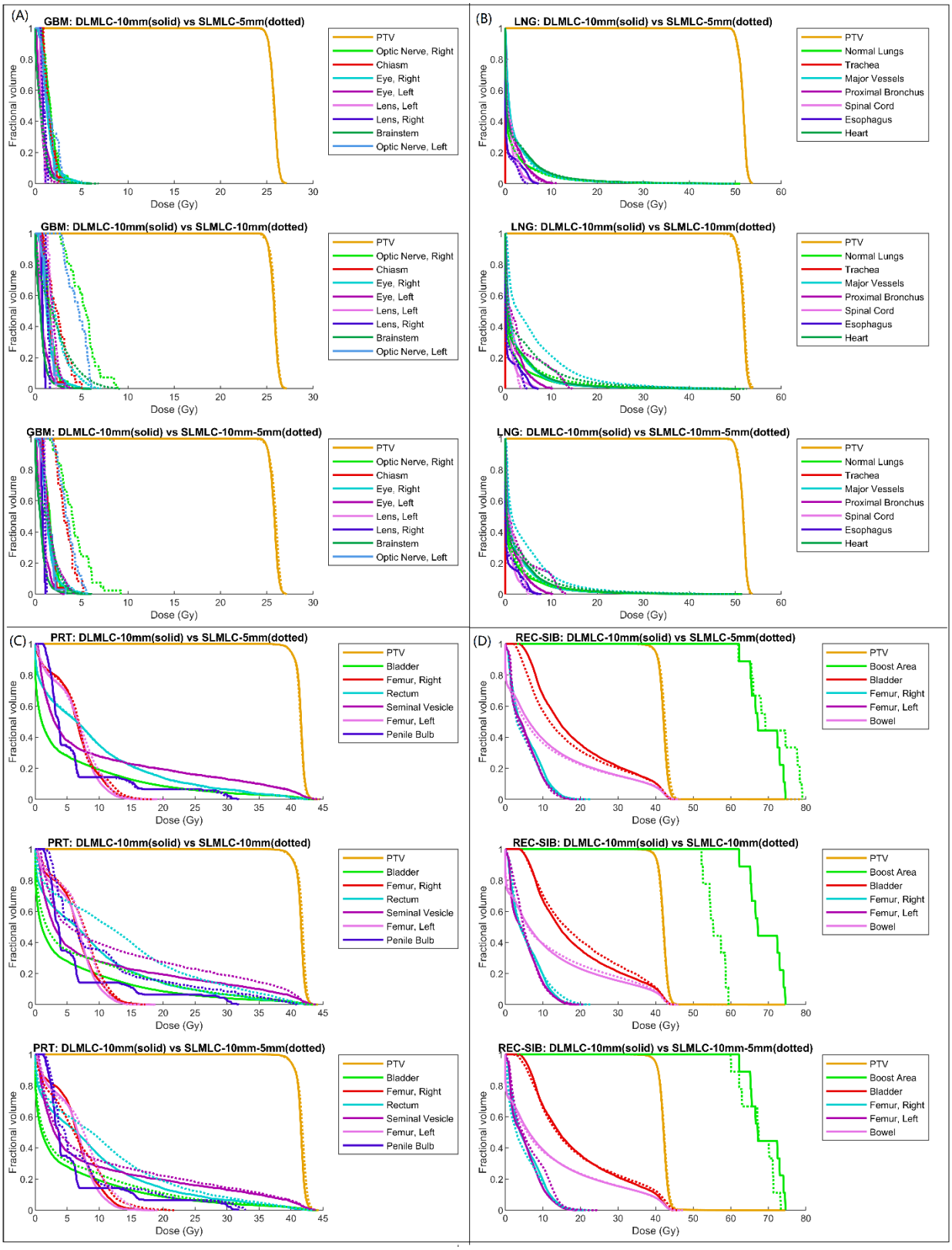

(A)
GBM: DLMLC-10mm(solid) vs SLMLC-5mm(dotted)
PTV
Optic Nerve, Right
Chiasm
Eye, Right
Eye, Left
Lens, Left
Lens, Right
Brainstem
Optic Nerve, Left
GBM: DLMLC-10mm(solid) vs SLMLC-10mm(dotted)
GBM: DLMLC-10mm(solid) vs SLMLC-10mm-5mm(dotted)
(B)
LNG: DLMLC-10mm(solid) vs SLMLC-5mm(dotted)
Normal Lungs
Trachea
Major Vessels
Proximal Bronchus
Spinal Cord
Esophagus
Heart
LNG: DLMLC-10mm(solid) vs SLMLC-10mm(dotted)
LNG: DLMLC-10mm(solid) vs SLMLC-10mm-5mm(dotted)
(C)
PRT: DLMLC-10mm(solid) vs SLMLC-5mm(dotted)
Bladder
Femur, Right
Rectum
Seminal Vesicle
Femur, Left
Penile Bulb
PRT: DLMLC-10mm(solid) vs SLMLC-10mm(dotted)
PRT: DLMLC-10mm(solid) vs SLMLC-10mm-5mm(dotted)
(D)
REC-SIB: DLMLC-10mm(solid) vs SLMLC-5mm(dotted)
Boost Area
Bladder
Femur, Right
Femur, Left
Bowel
REC-SIB: DLMLC-10mm(solid) vs SLMLC-10mm(dotted)
REC-SIB: DLMLC-10mm(solid) vs SLMLC-10mm-5mm(dotted)
Fractional volume
Dose (Gy)

Figure 3-19. DVH for (A) the GBM case, (B) the LNG case, (C) the PRT case, and (D) the REC-SIB case. The solid lines are for the DLMLC plan, and the dotted lines are for SLMLC plans. D95 is normalized to the prescription dose. Figure reproduced from Lyu *et al.* (2019)[44].

In this study we have shown that DLMLC VMAT can be optimized to outperform conventional SLMLC-10mm plans achieve comparable plan quality with SLMLC-5mm VMAT. DLMLC-10mm takes advantage of the faster gantry rotation and leaf speed of new ring gantry linacs such as the Halcyon to enable more efficient plan delivery than coplanar VMAT on a C-arm linac. Compared with the previously reported Halcyon VMAT plans[106,107], which need more arcs to achieve comparable dosimetry to a VMAT plan on traditional TrueBeam C-arm linac, the proposed method clearly elevated the performance of single-arc DLMLC VMAT to be equal to SLMLC with higher resolution leaves. The full angular resolution DLMLC problem can be solved in 5 minutes for the GBM and lung cases, and 20 minutes for the prostate case. With the accelerated dose calculation enabled by our GPU-CCCS algorithm, online adaptive DLMLC planning steps into the realm of clinical feasibility with an approximate total planning time of less than 30 minutes even with relatively low investment in distributed computational hardware.

3.5.4 Many Isocenter Optimization for Robotic Radiotherapy[43]

3.5.4.1 Background

The development of 4π non-coplanar radiotherapy[48,93] has fostered significant improvements to the achievable dosimetric plan quality of IMRT treatment. However, the potential for patient-gantry collision in widely used C-arm linacs limits the optimization and treatment outcome while requiring complex coordinated gantry and couch motion. By replacing the C-arm linac with a robotic arm delivery platform, the number of collision-free beam angles is increased, and the beam-to-beam motion is simplified, but the need for a more

compact X-ray emission head introduces a new conflict between field size and MLC modulation resolution. This study investigates the dosimetry and delivery efficiency of treatment to multiple isocenters to achieve simultaneously high dose modulation resolution and large tumor coverage.

3.5.4.2 Methods

We investigate a novel method of multiple-isocenter non-coplanar IMRT planning with simultaneous fluence map optimization and beam selection. Without loss of generality, we adopt the geometry of the new robotic linac that is currently under development at Celestial Oncology (Figure 3-20).

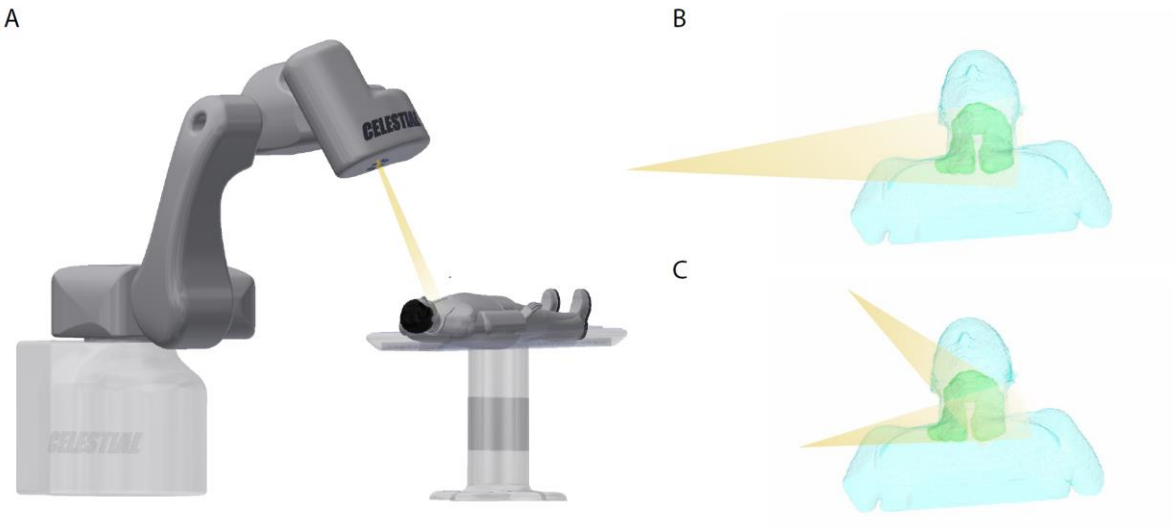


Figure 3-20. (A) Demonstration of the robotic arm platform, (B) an isocentric SID-100 beam that covers the entire target, (C) beams of different isocenters are required to efficiently cover the entire target. Figure reproduced from Lyu *et al.* (2020)[43].

The selection of source-to-isocenter distance (SID) affects both the dose modulation resolution and beam field of view (FOV). To study the effects of SID on plan quality and delivery efficiency, two SIDs are considered: 100 cm and 50 cm. At SID-100, the projected

MLC leaf width at the isocenter is 1 cm, and the FOV is 20×20 cm^2. Likewise, at SID-50 the projected MLC leaf width is 0.5 cm, and the FOV is 10×10 cm^2. Ideally, the projected MLC leaf width should be minimized for maximized dose delivery precision, however for Head and Neck (H&N) cancer with a PTV up to 20 cm in length, the target cannot be fully covered by the reduced FOV of the the SID-50 beam. In order to facilitate use of the SID-50 beams with their enhanced dose modulation resolution, we employed a heuristic isocenter selection approach. The PTV is first tightly bounded by a cuboid aligned to the scanner's coordinate axes, which is then evenly divided into $[N_x, N_y, N_z]$ cells. For each cell, an isocenter position is calculated as the center of mass (CoM) of the enclosed partial PTV. The set of isocenters selected using this method guarantees sufficient coverage of the PTV but does not necessarily guarantee that all isocenters will be used in the final plan; it may be the case that fewer isocenters are necessary for achievement of optimal dosimetry.

An optimization problem was formulated[43] to simultaneously solve for the optimal set of beams and their X-ray fluences from the candidate pool of all considered beams for all isocenters. Plan optimization for 10 H&N patients was performed according to the parameters determining the overall dose calculation and optimization problem sizes in Table 3-4.

Table 3-4. Number of feasible beams, prescription doses and PTV volumes for all patients. Table reproduced from Lyu *et al.* (2020)[43].

Patient	PTV volume (cm^3)	Bounding Box (cm)			Isocenter #		Feasible beam #		Sampled beam #	
		x	y	z	SID-100	SID-50	SID-100	SID-50	SID-100	SID-50
H&N #1	610.5	9.3	14.3	15.8	1	4	776	2785	1162	4648
H&N #2	724.6	10.8	16.0	18.3	1	8	826	1440	1162	2320
H&N #3	947.0	10.0	17.0	22.3	2	12	1579	2117	2324	3480

H&N #4	785.0	10.3	15.5	19.0	1	8	891	1452	1162	2320
H&N #5	686.4	10.5	14.5	18.3	1	8	824	1422	1162	2320
H&N #6	787.0	10.5	16.5	17.5	1	8	842	1443	1162	2320
H&N #7	352.7	8.5	8.5	18.5	1	2	758	1315	1162	2324
H&N #8	555.7	9.8	13.3	14.8	1	4	906	2904	1162	4648
H&N #9	271.3	8.8	6.8	17.0	1	2	781	1385	1162	2324
H&N #10	620.5	10.3	15.0	14.0	1	8	685	1278	1162	2320

To limit the computational costs, for plans with fewer than or equal to four isocenters, 1,162 beams were uniformly sampled in the 4π spherical space around each isocenter with 6° of separation. For other cases with more isocenters, 290 beams were instead uniformly sampled with 12° of separation for each isocenter. The average number feasible beams, after excluding beams with the potential for collision with the patient, was 887 (69% of sampled beams) for SID-100 plans, and 1,754 (60% of sampled) for SID-50 plans. To estimate delivery efficiency of all plans, a traveling salesman problem was formulated[108] and solved, producing an optimized ordering of beams that reduced the amount of linac head travel and MLC leaf motion.

3.5.4.3 Results and Conclusions

Overall, the SID-50 plans resulted in higher dose fidelity values compared with the SID-100 plans using the same number of beams, showing superior dosimetric plan quality to the SID-100 plans. As the number of delivered beams increased, the SID-50 plan quality improved more quickly than the SID-100 plans. When 20 or more beams are used, SID-50 shows clear dosimetric advantage over SID-100 (Figure 3-21) for all cases.

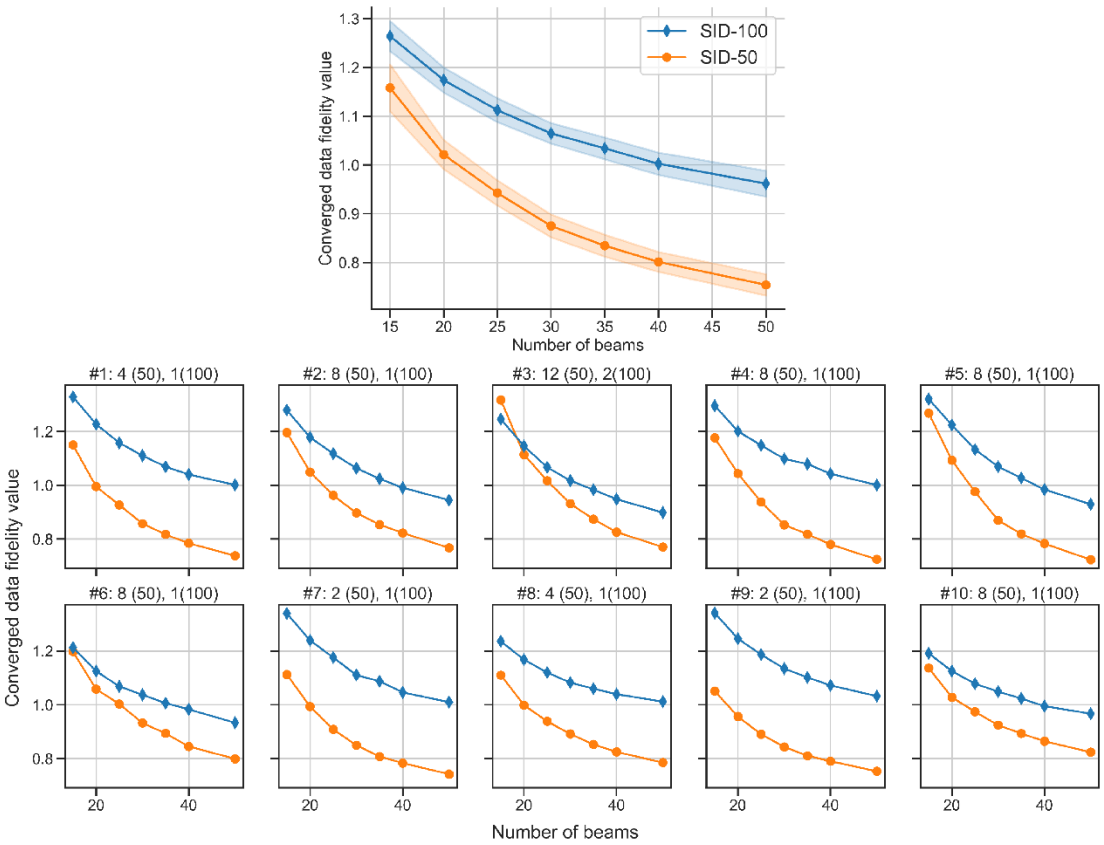


Figure 3-21. Final objective value vs the number of beams. The plot with shaded error bar shows a summary of all patients. Each patient plot is titled with the patient number, the number of isocenters for the SID-50 plan, and the number of isocenters for the SID-100 plan. For example, the first patient plot is entitled: '#1: 4(50), 1(100)', showing that the patient #1 has four isocenters for the SID-50 plan, and one isocenter for the SID-100 plan. Figure reproduced from Lyu *et al.* (2020)[43].

The average delivery time for a 20-beam plan, delivered over 30 fractions, in each case was 12 minutes for SID-50 and 7 minutes for SID-100, with the longer time being attributed to longer MLC travel time, which is the most time consuming aspect among monitor unit (MU) delivery time, linac head travel time, and itself. For a hypo-fractionated delivery over 5 fractions, the higher MU delivery requirement of each fraction tilts the balance of fractional delivery time from an MLC travel time-heavy to an MU delivery-heavy time requirement, which equalizes the average overall delivery for SID-50 (15 minutes) and SID-100 (14 minutes). Due to the inverse square effect of radiation delivery, the SID-50 plans with one-

half the x-ray propagation distance to the target feature a factor of four higher dose rate than the SID-100 plans, allowing for nearly comparable delivery times when MU delivery is the most time consuming component of delivery.

Treatment planning with integrated optimization of multi-isocentric delivery offers dosimetrically comparable outcomes compared to single-isocentric delivery using a C-arm linac and standard SID of 100 cm. The ability of our many-isocenter approach to produce effective plans in applications with reduced FOV, such as for novel compact X-ray emission heads and robotic-arm positioning, expands the possibility for efficient and less restrictive traversal of the entire 4π spherical beam space, affording greater dosimetric gains over the more restrictive C-arm gantry geometry. Using our efficient GPU-CCCS beamlet dose calculation framework, we can efficiently meet the 2-fold greater planning dose data requirements over the traditional C-arm 4π treatment planning problem. With an average planning time of 25 minutes for 15-beam plans, and 22 minutes for 50-beam plans, our approach can be implemented in a clinically feasible timeframe of less than one day in total.

3.6 Conclusions

We developed and implemented a highly efficient GPU-CCCS algorithm for computing beamlet dose with customizable fidelity using an intuitive *context radius* setting for high beam counts in complex static-beam and dynamic arc IMRT planning problems. We've demonstrated that the use of our novel parallel beamlet context-based technique substantially outperforms the naive approach of computing beamlet dose in sequence as is done by existing CCCS algorithms in terms of efficiency, while maintaining similar levels of dosimetric accuracy. Additionally, we embedded this approach in a scalable high-

performance computing architecture that allows the number of independent computing nodes, and the number of GPUs employed by each to be adapted to match the resources and demands of the user. Finally, several applications of novel treatment paradigms, enabled by the efficiency of our GPU-CCCS algorithm for beamlet dose calculation, were presented.

4 A High-Performance Distributed Framework for Large-scale Monte Carlo Dose Calculation

4.1 Introduction

For the purposes of radiation treatment planning in the common clinical setting of CT-guided megavoltage X-ray therapy, Analytical approaches to dose calculation, such as AAA, CCCS, and more recent iterative linearized Boltzmann transport equation solvers strike an effective balance between dosimetric accuracy and calculation speed. For the emerging paradigms of MR-guided X-ray, proton, and heavy ion therapies, however, more complex particle physics results in non-trivial deviations in dose deposition from those for which efficient analytical dose calculation algorithms have historically been developed and validated. For MR-guided X-ray RT, complications to the dose deposition physics arise in the form of electron return effects (EREs) induced by the Lorentz force experienced by electrons traveling through a strong magnetic field, such as the 1.5 tesla field used widely for diagnostic MRI and in modern MRgRT linacs like Elekta's Unity MR-guided linear accelerator[5]. In proton and heavy ion therapies, these complications exist due to the geometry dependent ranges of heavy particles that exhibit highly non-linear dose deposition rates with increasing depth of tissue penetration.

To address the need for accurate dose calculation methods in each of these applications, established Monte Carlo (MC) methods, in which dose depositing particles are individually simulated through a prescribed patient according to a set of physical interaction probabilities, have been thrust to the forefront of clinical practice. While MC-based dose

calculation techniques have proved versatile in handling the complexities involved with these new treatment modalities, they remain substantially slow in comparison to the analytical methods employed in CT-guided X-ray therapy. MC's inefficiency is marked by the necessity for simulating hundreds of thousands of primary particles into complex voxelized geometries derived from a patient's pre-treatment or daily CT image acquisition. Complementary practices for improving treatment effectiveness such as beam orientation optimization and volumetric modulated arc therapy with automatic trajectory optimization create a demand for many orders of magnitude more dose calculation data prior to treatment planning. Another complementary paradigm, online adaptive radiotherapy (OART), has a patient's treatment re-planned from daily image acquisitions to adapt the deliverable dose to changes in patient anatomy and alignment in the radiation delivery equipment for enhanced dosimetric accuracy. In OART, the patient remains in the treatment position while re-planning is performed, thus placing challenging limitations on the amount of time available for dose calculation to be less than 5 minutes. For maximum improvement to treatment efficacy, these methods are ideally combined into a single standard of practice, a goal that remains infeasible in large part due to the insufficient dose calculation speeds of MC.

In order to enable further research of emerging paradigms in radiation therapy and new approaches for the acceleration of MC dose calculation, a stop-gap strategy for performing rapid and large-scale MC dose calculation with existing technologies was necessary. To meet this need, distributed computing concepts were applied to create an integrated framework for the automated and parallelized calculation of dose data with the ability to scale to an arbitrarily large pool of computational resources.

4.2 Implementation

Our platform for rapid MC dose calculation is implemented using an efficient combination of the Python, C++, and CUDA programming languages where most appropriate. Python provides the integration layer tying together the many components of a treatment planning system, such as the data organization, network communication, input/output, and user interface modules. Python is also used for implementing RT-domain-specific algorithms where absolute raw computational efficiency is of little concern, but flexibility and ease of implementation is necessary to maintain the platform's customizability to new applications. Generation of particle source parameters for use by the MC simulation engine is one example of this focus on customization over performance; this task is accomplished by only a few affine transformations and implementation in a compiled language closer to the computational hardware would introduce too much inflexibility with negligible gain in overall performance of the system. For other tasks where computational efficiency is important, such as in raytracing to define each of the dose calculation tasks, C++ and CUDA are selected for implementation; this performant code is then integrated into the high-level Python code when appropriate.

4.2.1 Distributed Computation Model

Enabling the capacity for unfettered parallelization of the expensive MC dose calculation tasks is our model for distributed processing. From a high-level, we organize a set of always-running *compute* nodes (computers) around a central *managing* node. Every compute node runs a persistent process that listens on a TCP/IP network socket and waits for instructions, at which time it begins a locally multi-threaded MC simulation for one beamlet. The contents of the instruction payload that the compute node receives include all necessary input

configuration for executing the beamlet dose simulation task using one of the pre-established MC simulation engines. Based on the usage requirements of the dose calculation framework, these engine choices may include one for traditional X-Ray dose calculation, one for the more specialized considerations of heavy ion dose calculation, and another still for brachytherapy dose calculation. Our framework currently supports the Geant4 MC engine for general purpose dose calculation of nearly every clinically useful particle type, and support for a special purpose proton MC engine is currently in progress. The resulting dose data, runtime log, and anything else requested by the user of the engine are packaged into a response payload that is sent back to the requesting node over the same network socket session. After confirming that the response payload has been successfully delivered, all traces of the previous simulation are cleaned up and the next simulation task in the compute node's local task queue is processed. Safe handling of exceptions and reporting of runtime errors to the requesting node ensure that this persistent process can remain running for months without loss of service.

Compute nodes are only designed to continuously process well-defined MC simulation tasks with no regard for the broader details of treatment planning. The managing node, conversely, handles all other aspects of dose calculation for treatment planning, such as determination of all simulation task configurations, queuing of tasks in prioritized order, scheduling of tasks on remote compute nodes, and efficient organization of the thousands to hundreds of thousands of dose calculation results that are generated to create a single patient plan. Beginning with the request for dose calculation by the end-user, the managing node first ingests the patient's pre-treatment CT images and physician contour definitions (stored in DICOM RTStruct format) along with a list of beam specifications, and a

configuration file describing the machine-specific radiation delivery parameters, such as field of view and beamlet dimensions, as well as the type and energy of radiation to deliver. The remaining actions until the time when dose data is finally exported for treatment planning are all performed automatically and asynchronously from the user's perspective. The sequence of actions performed by the managing server is specific to each treatment modality. For X-Ray RT, the sequence of actions includes the generation of an MC-compatible geometry file that will be supplied to the MC engine at execution, ray tracing to identify active beamlets, and the subsequent creation and enqueueing of new simulation tasks for every active beamlet. The managing node maintains a priority queue of simulation tasks that can be sent to a compute node for simulation when one becomes available. The placement of new simulation tasks into the queue depends on the user-specifiable priority setting, such that new high priority tasks are scheduled for simulation before already-enqueued tasks of lower priority. In the clinical setting, this functionality enables efficient planning amidst a setting in which patients' schedules may change unexpectedly. In the research setting, this enables the user to enqueue many simulation tasks for long term studies such for generating training data for machine learning but still produce urgent results for short term studies in a timely manner.

The managing node organizes all the plan configuration metadata and simulation task statuses in a local document-based database (MongoDB) with the hierarchical organization shown in Figure 4-1. Large data files such as dose calculation results and MC engine geometry inputs are stored on the native filesystem in a protected directory and linked into their corresponding database objects by their relative paths. Unique identifiers (UIDs) are assigned to each database object and are used to define a conflict-free directory hierarchy

on the filesystem. The primary database objects are classified as one of the Image, Structure, Beam, Sub-beam, Simulation, and Sample types. When a hierarchical relationship exists between two types, they are linked to one another by their UIDs. Organization of plan configurations in the database greatly simplifies structured retrieval of results using query filters; for example, dose calculation can be requested for one patient using two different treatment modalities and the results can be easily filtered and exported for direct comparison.

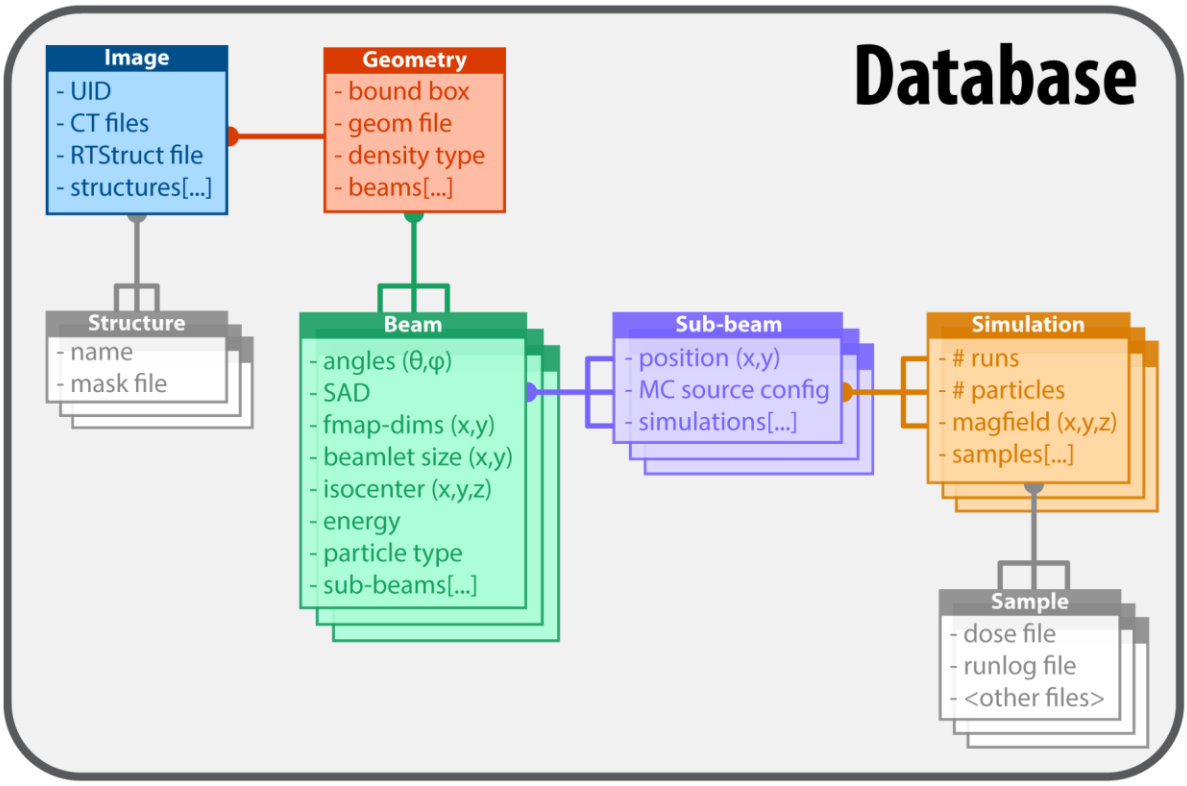


Figure 4-1. Database document hierarchy. Simulations are organized as children of a sub-beam (beamlet or spot). Sub-beams are children of a beam. Multiple beams can be defined for a single calculation geometry and multiple geometries can be defined for a CT image acquisition. Images contain masks for every structure defined in the RTStruct file. Simulations can produce one or more independent samples of dose for the same configuration as a method for machine learning dataset augmentation.

All communication between the managing and compute nodes is performed using an efficient custom-defined protocol that operates over standard TCP/IP networking sockets.

Support for sending arbitrary files and JSON-serializable payloads is available. Every communication follows a standard for a request-response structure with optional payload in either direction. Both the requesting and responding nodes are aware of the communication status such that they can each require confirmation of receipt of their respective payloads if the application requires it. In addition to communication with compute nodes, a client interface was developed, allowing the user to interact with the managing node remotely. A RESTful API on the managing node interprets REST requests sent by the user via the client program to one of many possible actions, including Insertion of Image, Beam, Sub-beam, and Simulation objects into the database, extraction of beamlet dose for treatment planning purposes, and generation of anatomical structure masks. Using the endpoints of the REST API, the client interface can be easily modified to run user-defined scripts for performing various functions. End-to-end treatment planning can be added to the client interface as a user-facing command by chaining together the commands for defining the beams selected for the treatment, defining MC simulation tasks for the active sub-beams in each of the beams, exporting the dose calculation results as a sparse matrix suitable for treatment planning, and finally performing the inverse optimization to acquire the plan delivery parameters. Alternatively, deep learning dose datasets can be automatically generated by creating a script in the client interface that randomly selects beam angles, isocentre coordinates, and active beamlets from a set of patient images, then creates simulation tasks for each to run asynchronously until completion. Furthermore, the use of the REST API model for interaction ensures compatibility with a graphical web-based interface that would simplify the process of performing common tasks for the end-user, especially for clinical use.

4.2.2 Orchestration and Scalability

To make deployment and scaling of our distributed MC simulation framework easier, Docker, a popular software containerization solution, was used. Docker functions similarly to a conventional virtual machine but differs in that instead of encapsulating an entire operating system (OS), which inefficient from both storage and processing perspectives, Docker interfaces with the existing OS at the kernel level, isolating the containerized software within the private network, filesystem, and process namespaces. The details of Docker's implementation make it more lightweight on storage and processing overhead than virtual machines. Additionally, the contents and function of docker containers are defined by images, or snapshots of a virtual filesystem saved by its creator at a specific state – usually between when the targeted software is installed on the computer, and when it is finally executed. Images are composed as layers of changes to a base operating system, so if one base image is used to define multiple images, the base image will only be stored once and reused to reduce the total storage requirements. The primary benefit of Docker's image-based initialization is that a developer can configure the image such that all prerequisites are already installed at their intended versions, and the only thing left to do is run the target application.

We created separate Docker images for the client interface, managing node process, and compute node process from a common base image that already contained the common libraries and python modules used by each process. Containerization of these processes allows fast deployment of the entire framework onto a user's network with only a few simple commands. Furthermore, expansion to a larger pool of compute nodes is achieved simply by starting the compute node container on a new network-accessible computer and informing

the managing node of its network address. For even simpler deployment, Docker's container orchestration tool (docker-compose) was used to make setup of the managing node and any number of compute nodes possible with a single command. Using container orchestration, it is updating all the framework code, across all involved nodes, can be achieved using a single command from the managing node.

4.2.3 SimpleDose: A High-Level Treatment Planning Interface

After our general-purpose MC framework was implemented, a simplified high-level user interface, named SimpleDose, was developed for extremely easy use in treatment planning workflows. In the general-purpose client interface, the user is free to select a managing node with which to issue commands and exchange data – useful when multiple teams on the same network each prefer to maintain their own databases. For simpler deployments, however, the generality of the framework makes user interaction more cumbersome, particularly in overcoming the learning curve associated with using Docker. The SimpleDose interface operates as a wrapper exposing only a subset of the functionality of the general-purpose client interface while completely hiding any details specific to using Docker. Among the features included in this interface were commands for automatically producing the beamlet dose data that were required by the treatment plan optimizer. The only manageable entity exposed by this interface, the *plan* object, consists of a patient CT image sequence, physician-defined anatomical structure contours, the list of beams to be optimized in the plan, the treatment modality, and a set of parameters specific to the radiation delivery device. Through this interface, the end-user can easily add plans with custom priority, delete plans, view the simulation progress for each plan with additional descriptive information from the

plan, and export the simulated beamlet dose, anatomical structure masks, and beam descriptions - all required for treatment planning.

4.3 Usage Examples

Due to its ease of use, the high-level interface, SimpleDose, is the preferred way of interacting with the framework for the treatment planning setting. Figure 4-2 lists the commands provided by the SimpleDose interface.

```
./simpledose usage

usage:  simpledose <target> [target-args]

  Available Targets (<target>):
    target                 description
    -----------            ------------------------------------------------------------
    create-plan            Initialize a new plan for automated Monte Carlo dosecalc
    fmaps                  Generate an fmaps.h5 file describing beams/beamlets
    masks                  Generate a masks.h5 file containing all structure masks
    plandose               Generate a sparse planning dose matrix file (spmat.h5)
    add-simulation         Add simulation tasks to plan with specific particle counts
    delete-plan            Delete a plan and all of its associated data (CAUTION)
    sim-status             List overall simulation tasks status (aggregated)
    plan-status            List simulation tasks status for each plan

    startup                Initialize docker environment, network, and database
    shutdown               Shutdown docker environment and database
    restart                Restart docker environment (if params were changed)
    cleandb                Optimize database by cleaning leftover entries/files
    uninstall              remove all traces of this code (CAUTION)
    bundle-computeserver   Prepare package for starting other computeservers
    start-computeserver    Start compute server on this machine
    (other)                Print this help message

  Examples:
    simpledose create-plan <plan-directory> <plan-name> --nparticles 1000 2000
    simpledose fmaps       <plan-name> <output-file>
    simpledose masks       <plan-name> <output-file>
    simpledose plandose    <plan-name> <output-file> --nparticles 1000
    simpledose delete-plan <plan-name>
    simpledose plan-status
    simpledose sim-status  --rate <update-rate-secs>
```

Figure 4-2. Summary of commands available from the SimpleDose interface for treatment planning dose calculation.

A typical planning workflow begins with the user exporting a CT image sequence and RTStruct file from their clinical radiology database. Next, a *config.json* file is created, indicating the name of the PTV structure, the beam modulation parameters (field of view and collimator dimensions), the desired dose voxel size (spatial resolution), treatment modality and beam energy spectrum, followed by the definition of the *beamlist.txt* file describing the angles and source-to-isocenter distances for every beam that will be considered in plan optimization. Finally, the *create-plan* command is run as demonstrated in Figure 4-3, which begins the automatically distributed MC dose calculation workflow.

```
./simpledose create-plan ../Pt54_replan_day13/ Pt54_replan_day13

plan-directory:  ../Pt54_replan_day13/
plan-name:       Pt54_replan_day13

2020-04-14 16:49:13,388 MainThread (INFO) | client:1559: Setting database reference DOI to "Pt54_replan_day13"
2020-04-14 16:49:14,146 MainThread (INFO) | client:1563: Found 174 ct image files
2020-04-14 16:49:14,146 MainThread (INFO) | client:1564: Found 1 rtstruct image files
2020-04-14 16:49:14,150 MainThread (INFO) | client:1584: Using monte carlo density type "bulk-density"
2020-04-14 16:49:14,150 MainThread (INFO) | client:1585: Using particle type "photon"
2020-04-14 16:49:15,866 MainThread (INFO) | client:1621: Successfully ingested data "/input" with database id "5e95e98b9224b2133cd5aac2" (doi: "Pt54_replan_day13")
2020-04-14 16:49:17,005 MainThread (INFO) | client:1633: Successfully inserted structure "PTV_NNreview" with database id "5e95e98c9224b2133cd5aac3"
2020-04-14 16:49:17,009 MainThread (INFO) | client:1641: retrieved PTV centroid as default isocenter: [-20.053005520343504, -278.0442138621959, -520.6696483336741]
2020-04-14 16:49:17,009 MainThread (INFO) | client:1646: Using custom bbox definition
2020-04-14 16:49:17,009 MainThread (INFO) | client:1666: bounding box: {'start': [-240.0, -364.0, -657.0], 'size': [192, 113, 111], 'spacing': ["2.5", "2.5", "2.5"]}
2020-04-14 16:49:17,018 MainThread (INFO) | client:1674: Successfully inserted geometry with database id "5e95e98d9224b2133cd5aac4"
2020-04-14 16:49:22,501 MainThread (INFO) | client:1702: Successfully inserted beam (1/3) "5e95e98d9224b2133cd5aac5"
2020-04-14 16:49:25,744 MainThread (INFO) | client:1702: Successfully inserted beam (2/3) "5e95e9929224b2133cd5aecb"
2020-04-14 16:49:29,965 MainThread (INFO) | client:1702: Successfully inserted beam (3/3) "5e95e9959224b2133cd5b322"
2020-04-14 16:49:29,966 MainThread (INFO) | client:1703: Successfully inserted 3 beams
2020-04-14 16:50:53,556 MainThread (INFO) | client:1719: Successfully inserted 1029 simulations for beam "5e95e98d9224b2133cd5aac5" (1/3)
2020-04-14 16:52:28,968 MainThread (INFO) | client:1719: Successfully inserted 1110 simulations for beam "5e95e9929224b2133cd5aecb" (2/3)
2020-04-14 16:54:31,100 MainThread (INFO) | client:1719: Successfully inserted 1194 simulations for beam "5e95e9959224b2133cd5b322" (3/3)
2020-04-14 16:54:31,100 MainThread (INFO) | client:1720: Successfully inserted 3333 simulations across 3 beams
```

Figure 4-3. Create-plan command output using the SimpleDose to add dose calculation requests for automatically distributed computation.

A *plan-status* command (Figure 4-4) is available, allowing the user to monitor the progress of dose calculation for each plan. The output is listed in a tabular format with various sorting options available.

```
./simpledose plan-status

Plan Name                                                  Date Created      Particle  Density Type  # Beams  # Particles   Completed Sims        Message
---------------------------------------------------------- ----------------  --------  ------------  -------  ------------  --------------------  -------
phantom                                                    2020-04-05 23:03  photon    bulk-density        1  10k             296 /   296 (100%)
phantom_0.1size                                            2020-04-07 07:48  photon    bulk-density        1  10k            1600 /  1600 (100%)
phantom_1size                                              2020-04-07 07:51  photon    interpolated        1  10k            1600 /  1600 (100%)
phantom_2.5size_shared                                     2020-04-07 07:56  photon    interpolated        1  10k            1600 /  1600 (100%)
phantom_5size                                              2020-04-07 07:56  photon    bulk-density        1  10k            1600 /  1600 (100%)
phantom_5spacing                                           2020-04-07 08:12  photon    bulk-density        1  10k             400 /   400 (100%)
phantom_150MeV                                             2020-04-07 08:18  electron  bulk-density        1  10k            1600 /  1600 (100%)
phantom_250MeV                                             2020-04-07 08:18  electron  bulk-density        1  10k            1600 /  1600 (100%)
phantom_500sad                                             2020-04-07 08:29  electron  bulk-density        1  10k            1600 /  1600 (100%)
phantom_test                                               2020-04-09 02:31  electron  bulk-density        1  10k             263 /   263 (100%)
LNGAS_size0.5_spacing2_energy250_num7_sad50                2020-04-09 22:10  electron  bulk-density        7  10k           11915 / 11915 (100%)
LNGAS_size0.5_spacing2_energy300_num7_sad50                2020-04-09 22:54  electron  bulk-density        7  10k            4819 / 11916 ( 40%)
LNGAS_size0.5_spacing2_energy400_num7_sad50                2020-04-09 23:42  electron  bulk-density        7  10k           11915 / 11915 (100%)
LNGAS_size0.5_spacing2_energy500_num7_sad50                2020-04-10 04:33  electron  bulk-density        7  10k           11915 / 11915 (100%)
LNGAS_size0.5_spacing2_energy300_num11_sad50               2020-04-10 05:31  electron  bulk-density       11  10k           12637 / 18726 ( 67%)
LNGAS_size0.5_spacing2_energy300_num16_sad50               2020-04-10 06:32  electron  bulk-density       16  10k           27235 / 27235 (100%)
```

Figure 4-4. Plan-status command output using the SimpleDose interface to list the simulation progress for all existing *plans*.

For plans that have completed dose calculation, the *plandose* command is provided, which facilitates the export of planning dose data in a sparse format suitable for treatment plan optimization (in a format compatible for use as the matrix A in Equation 2-4). To maintain flexibility with the preferred sparse matrix formats of a variety of computational linear algebra libraries, the coordinate list (COO) format is used for storing the exported planning dose data. The COO format uses three equal-length arrays to store the row index (voxel number), column index (beamlet number), and data value (voxel dose) for every non-zero element of the represented matrix. The specification of this storage format is given in Figure 4-5.

```
GROUP "/" {
   DATASET "data" {
      DATATYPE  H5T_IEEE_F32LE
      DATASPACE  SIMPLE { ( 74226688 ) / ( H5S_UNLIMITED ) }
   }
   DATASET "i" {
      DATATYPE  H5T_STD_U32LE
      DATASPACE  SIMPLE { ( 74226688 ) / ( H5S_UNLIMITED ) }
   }
   DATASET "j" {
      DATATYPE  H5T_STD_U32LE
      DATASPACE  SIMPLE { ( 74226688 ) / ( H5S_UNLIMITED ) }
   }
   DATASET "ncols" {
      DATATYPE  H5T_STD_U32LE
      DATASPACE  SCALAR
   }
   DATASET "nrows" {
      DATATYPE  H5T_STD_U32LE
      DATASPACE  SCALAR
   }
   DATASET "sparse_threshold" {
      DATATYPE  H5T_IEEE_F32LE
      DATASPACE  SCALAR
   }
}
```

Figure 4-5. Sparse storage format for calculated planning dose results for a SimpleDose *plan.* The sparse coordinate list (COO) format uses three equal-length arrays to store the row index, column index, and data value for every non-zero element of the represented matrix.

4.4 Performance

Our framework was installed on a private cluster and tested on its ability to accelerate large-scale MC beamlet dose simulation beyond the capabilities of a single node. Eight compute nodes were instantiated with a combined total of 216 logical CPU cores for use in large-scale MC beamlet dose calculation. IMRT *plans* for 20 electron beam treatment scenarios were submitted for beamlet dose calculation using the *create-plan* command of the SimpleDose interface with a total of 78,545 beamlets with dimensions varying between 1x1 mm^2 and 5x5 mm^2. Each beamlet was simulated for 10,000 primary electron generations with an average time of 1.73 seconds per simulation. Using our framework, the total dose calculation time was 37.8 hours, constituting an average of 2,080 simulations per hour

(equivalently 20.8 million primary electrons per hour). By comparison, for MC dose calculation on one of the eight compute nodes with a total of 32 logical cores, the average rate of completion was approximately 310 simulations per hour (3.1 million primary electrons per hour), constituting a factor of acceleration approximately equal to 6.7, nearly equal to the theoretically expected factor of 6.75, with network transfer overhead suspected as responsible for the small deviation from theory.

4.5 Conclusions

Our distributed MC dose calculation framework provides a stop-gap solution to accelerating MC simulation in support of efficient research of novel treatment planning and dose calculation techniques. The scalability and ease of use of our framework make it a valuable tool for research applications spanning a large range of scales, while requiring little knowledge of dose calculation from the user.

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

5.1 Introduction

Radiation therapy (RT) is one of the primary modalities for cancer treatment[109]. 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[110,111], 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[112,113]. 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[114,115]. The second problem to significantly accelerate dose calculation without compromising accuracy for adaptive online radiotherapy (OART) [116–118] beamlet dose calculation remains unsolved. State of the art methods for calculating dose in the magnetic field-free condition employ deterministic convolution-based algorithms[19,34,59,60,119,120], which are extremely efficient for handling a large amount of beamlet dose calculation[34]. 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[121,122]. Failing to account for EREs properly can generate disparities between expected and deliverable doses at these tissue-air interfaces in MRgRT by up to 70%[123,124]. On the other hand, the disparities can be mitigated with accurate EREs modeling in beamlet dose calculation[125].

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 noise[9] than the analytical methods. Besides conventional MC acceleration techniques[9–11], efforts have been made to accelerate MC dose calculation using graphics processing units (GPUs) [15,16,126] for final plan dose calculation but the speed remains insufficient for the orders-of-magnitude

larger beamlet dose calculation task, which are critical to the OART and mitigation of 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. 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[127–131] with its understanding of the dose deposition patterns learned 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. Our model is tasked with resolving the low-noise 3D dose deposition for each under sampled (high-noise) MC dose simulation. Distinctive from related works[132,133], however, we confront the additional challenges of including 1.5T MRI magnetic field effects, as is the case for Elekta's Unity MR-guided linac[5], 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 (2,000 compared to 100,000 particles for high- and low-noise respectively).

5.2 Preliminary Investigation in Stacked Slab Phantom Geometries

To demonstrate the proof of concept for an earlier, 2D version of our proposed DeepMC model (termed DeepMCv1), we conducted two experiments with limited data and generality. The experimental setup was identical for both: training and testing data sets consisting of 2D high- and low-noise dose image pairs were first synthesized using MC simulation, a model was developed and trained on the training data set, and the dosimetric fidelity to each low-noise image was calculated for the high-noise (model input) and predicted (model output) images. The fidelity was measured using mean squared error (MSE) and the average gamma passing percentage of pixels for various gamma criteria across all samples in the testing data set. To gain more insight into the predictive performance, gamma index maps were visualized at certain intervals for random data samples as the model was trained.

In the first experiment, a simple slab phantom was defined (Figure 5-1). From this basic phantom geometry, two material configurations were selected, yielding two distinct phantoms for study. The first phantom was designed with a water-to-lung, and a subsequent lung-to-water interface. The second was designed to produce more dramatic EREs by replacing water with aluminum and lung with air. For each of the virtual phantoms, a static 1.5T magnetic field was defined perpendicular to the plane of beam angle variation. Dose was simulated for 45 beams spaced at three-degree increments between -66 and 66 degrees. Of these beams, eight equidistant beams were removed to a separate testing data set. Dose from each beam angle contained 18 slices that were separated and independently used for both training and testing. After training the model for two hours on four GPUs, the MSE between high and low-noise dose was compared with the MSE between DeepMCv1-

predicted and low-noise dose. The predicted dose produced a reduction in average MSE by more than one order of magnitude on the full testing data set for both phantoms. Figure 5-2 shows a longitudinal slice of dose for a beamlet in each phantom. The denoising effect of our approach is evident, and its ability to predict the ERE is promising.

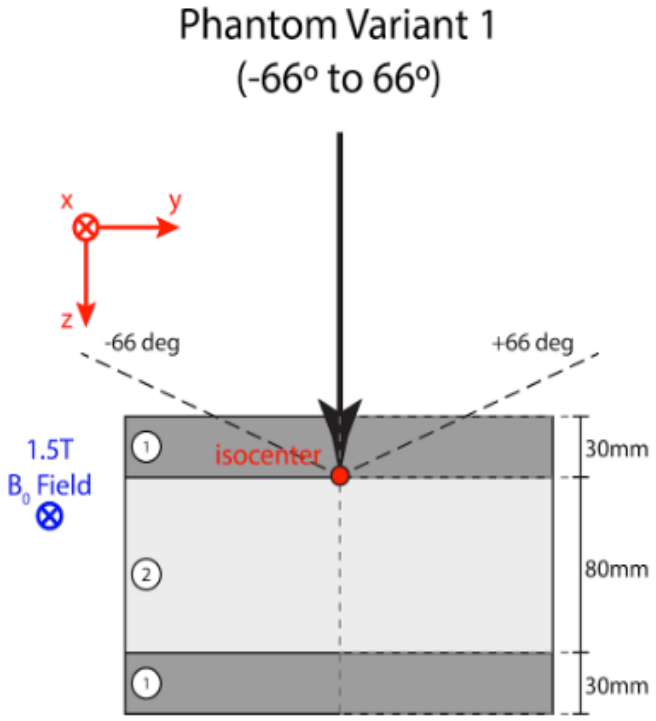


Figure 5-1. Slab phantom geometry specification for studying electron return effects present at high-density-gradient tissue interfaces.

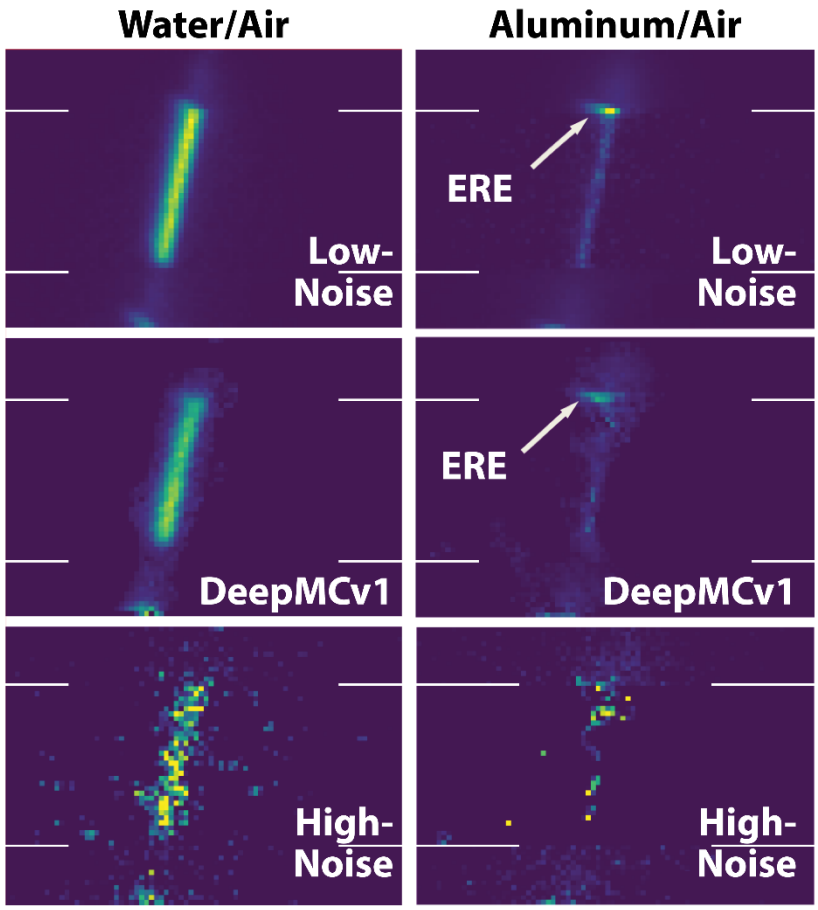


Figure 5-2. MC dose ground truth (low-noise.), model input (high-noise) and DeepMCv1 prediction for a photon beamlet in the low dose penumbra region (1.25cm off-axis) of water/air (left) and aluminum/air (right) slab phantoms. Lines indicate positions of horizontal material interfaces. Low-noise dose was simulated using 18M particles in ~3 minutes. High-noise dose was simulated using 30K particles in 3 seconds. Prediction took <100ms after high-noise dose simulation.

The proof of concept experiments conducted in slab phantoms with varying tissue compositions successfully demonstrated promise for our DeepMC dose prediction model to accurately reduce Monte Carlo dose calculation's statistical noise while also reconstructing dose deposition concentrations at tissue interfaces affected by EREs. To continue this line of research, improvements to the model architecture were made and translation to patient treatment planning applications was investigated as described in the following sections.

5.3 Methods

5.3.1 DeepMC Model Architecture

To address the challenges of efficient MR-guided dose calculation, we've developed a 3D CNN (Figure 5-3) comprised of a UNet[134] with three spatial feature scales, batch normalization[135], residual[136,137] and squeeze-excitation[138] blocks, and ReLU activations[139]. 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-channeled 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 size[140], 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 $1\times1\times1$ kernels (for changing the feature map dimensionality), followed by three residual 3D convolution blocks with $3\times3\times3$ 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 included[137]. 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 $1 \times 1 \times 1$ kernels to reduce the feature size to one, followed by a sequence of three residual convolution layers to perform the final dose prediction.

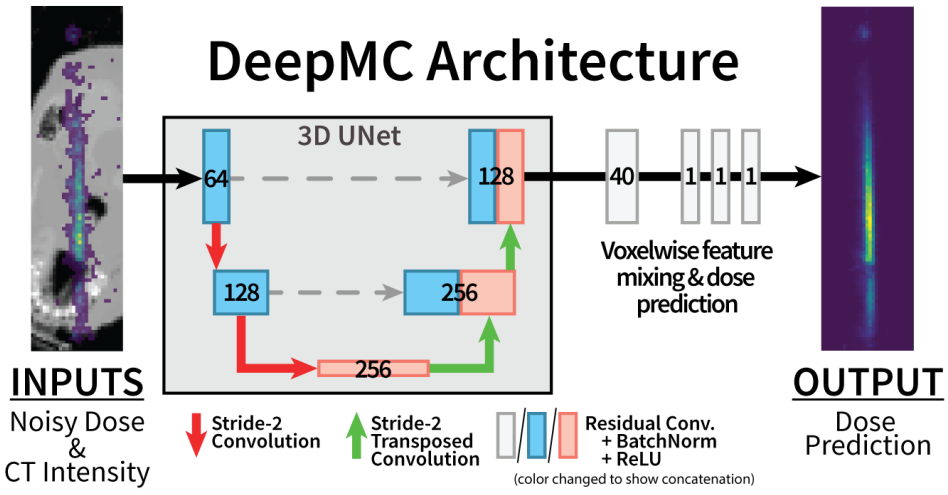


Figure 5-3. Fully convolutional model architecture used by our method, DeepMC. Numbers in blocks indicate number of feature channels produced by learned convolutional kernels at each stage. 3×3×3 kernels are used in UNet layers. 1×1×1 kernels are used for feature mixing and dose prediction.

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 5-4.

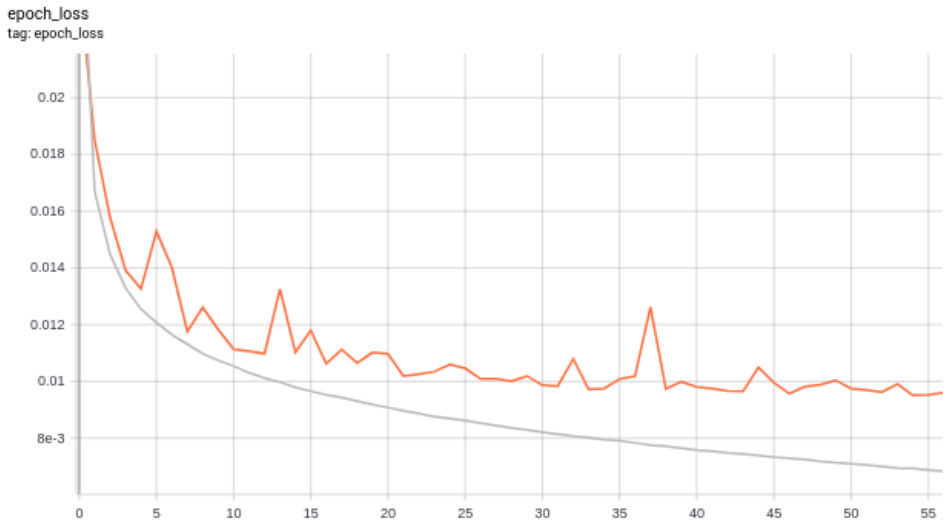


Figure 5-4. DeepMC training progress. Per-epoch loss is shown for training dataset (grey) and validation dataset (orange).

In Figure 5-5, 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), 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 5-5 indicate these corrupted areas in the high-noise dose).

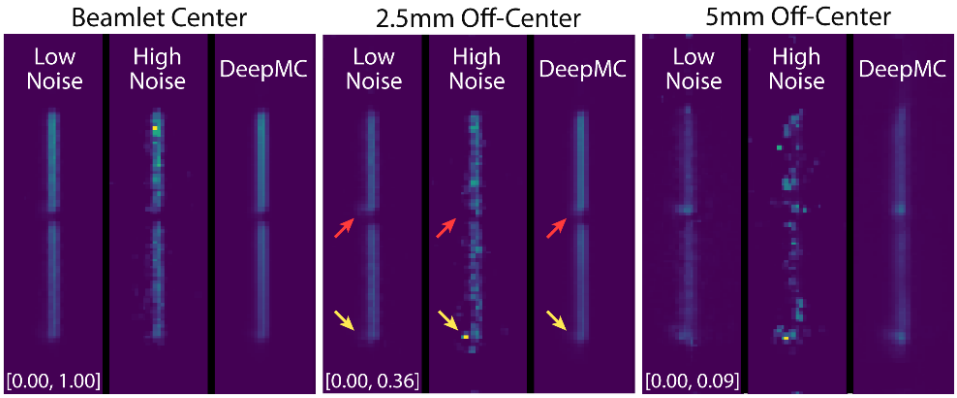


Figure 5-5. Dose from low- and high-noise MC simulation and DeepMC prediction for one 5x5mm² x-ray beamlet. Three adjacent slices are shown with their transverse distance from the beamlet's central axis listed in the titles. Color scale limits are displayed in the lower left corner for each slice in normalized units. 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.

Our model is implemented in the Python programming language using the TensorFlow v2.10 deep learning library[141] 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 5-6 shows the observed voxelized dose density and cumulative density.

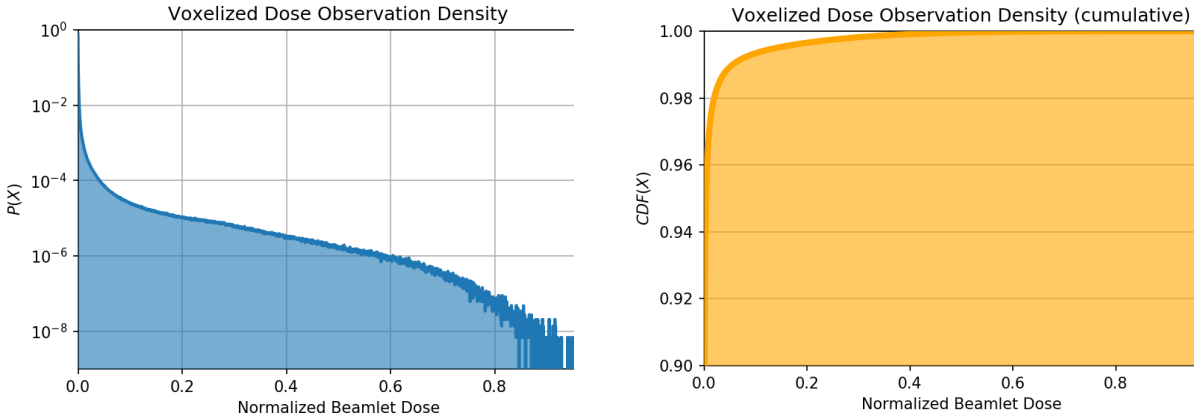


Figure 5-6. Observed density (left) and cumulative density (right) of voxelized x-ray dose, normalized to per-beamlet maxima. Voxels with dose lower than 10% of per-beamlet maxima account for over 99% of observations. Vertical axes are displayed in log-scale.

To address this imbalance in the data distribution, we defined a voxel wise decaying exponential loss weighting function, shown by Equation 5-1 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.

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

Equation 5-1

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.

5.3.2 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 model performance on a separate testing dataset converges to the ground truth, we are burdened by the slow speed of MC simulation using modern software. Our task of calculating x-ray dose for MRgRT poses the added challenge of modeling the complex geometry-dependent physics induced by EREs. While the more traditional setting of CT-guided radiotherapy has benefited remarkably from implementation on graphics processing units (GPUs), similar attempts to realize such hardware-enabled levels of acceleration in the MR-guided setting have been largely unsuccessful. The Monte Carlo simulation framework, Geant4[142,143], 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 clinical 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 clinical OART and for the purposes of generating data to train our CNN model this is still too slow.

To address these speed concerns, we developed an automated and distributed simulation framework that uses a communication model consisting of a single manager node assigning tasks to one or more worker nodes for asynchronous and scalable computation divided at the unit of individual beamlets. We developed a protocol for sending beamlet simulation tasks to any available worker on the network and receiving the simulated dose for retrieval during model training. Docker containerization is used to simplify the setup of new workers. The manager connects with a central database to identify remaining tasks and organize completed task results. Using this framework, we expanded the multi-threaded processing functionality of Geant4 to operate over a scalable pool of network-accessible computational nodes with ease. For our purposes, a cluster of 10 nodes and a combined total of over 160 logical CPU cores was utilized to generate training and testing datasets in condensed timeframe.

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 (≤4mm 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 clinical 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.

Additionally, efficient five-fold data augmentation was achieved with low additional cost by pairing a single low-noise dose volume with each of five independently simulated high-noise dose volumes – justified by the stochastic nature of MC simulation up to a limit in the number of simulated particles[9]. Simulation of the common low-noise dose for each beamlet uses enough particles to approach a deterministic solution with negligible noise corruption, whereas simulating five low-noise dose volumes instead produces five effectively identical dose output at five times the computational cost. Furthermore, for the faster high-noise MC simulations, much of the runtime is spent on initialization of the particle interaction probability tables and simulation geometry. Data augmentation by repeated high-noise simulation of one geometric configuration amortizes the fixed initialization cost across multiple results, minimizing the overall computational cost for training data generation. We have access to the exact geometric parameters of each of our generated beamlet configurations during both training and prediction. This property of the task is exploited to 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 a clinical 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.

5.3.3 Experiment Design

To validate our approach to MC dose prediction, we performed two types of evaluation. First, we trained our proposed model using paired high- and low-noise MC dose volumes, generated using our distributed simulation framework. DeepMC dose predictions were then produced alongside the paired high- and low-noise MC dose for each of the training and testing beamlet configurations. Normalized mean absolute error (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.

A clinically motivated analysis was then performed to evaluate DeepMC's ability to calculate the beamlet dose for effective treatment planning. First, DeepMC dose was predicted for every active IMRT beamlet in each of two testing dataset H&N CT scans. 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 problem[36] presented in Equation 5-2.

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

Equation 5-2

A is the planning dose matrix with vectorized 3D dose for each beamlet occupying a column, d_o 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 the optimal 2D fluence maps to be smooth and, therefore, more efficient to deliver. The convex optimization problem in Equation 5-2 is solved using the Fast Iterative Shrinkage Thresholding Algorithm (FISTA)[46], 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 an efficient dose estimate (termed *planning* dose) with which the clinical dosimetrist evaluates the plan quality for approval prior to radiation delivery. After approval, A_{GT} is then used to calculate the *deliverable* dose for each plan, which is a higher quality indication of the eventual dose delivery. 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 a standard clinical procedure for plan evaluation that analyzes characteristics of 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 (99^{th}, 98^{th}, 2^{nd}, and 1^{st} percentiles of dose), mean dose, maximum dose, homogeneity index (HI)[144], describing the uniformity of dose within the PTV, and conformity number (CN)[145], describing the compactness of dose to control dose spillage beyond the PTV.

5.4 Results

DeepMC 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 is 3.960e-2 (9.724e-2). 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 two 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 noise-free MC simulation.

Figure 5-7 compares deliverable dose DVHs for the ground truth, DeepMC, and noisy plans. Plans for the first testing patient (Figure 5-7 top) targeted a single PTV with a prescription dose of 70 Gray (abbreviated as 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 (Figure 5-7 bottom) 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.

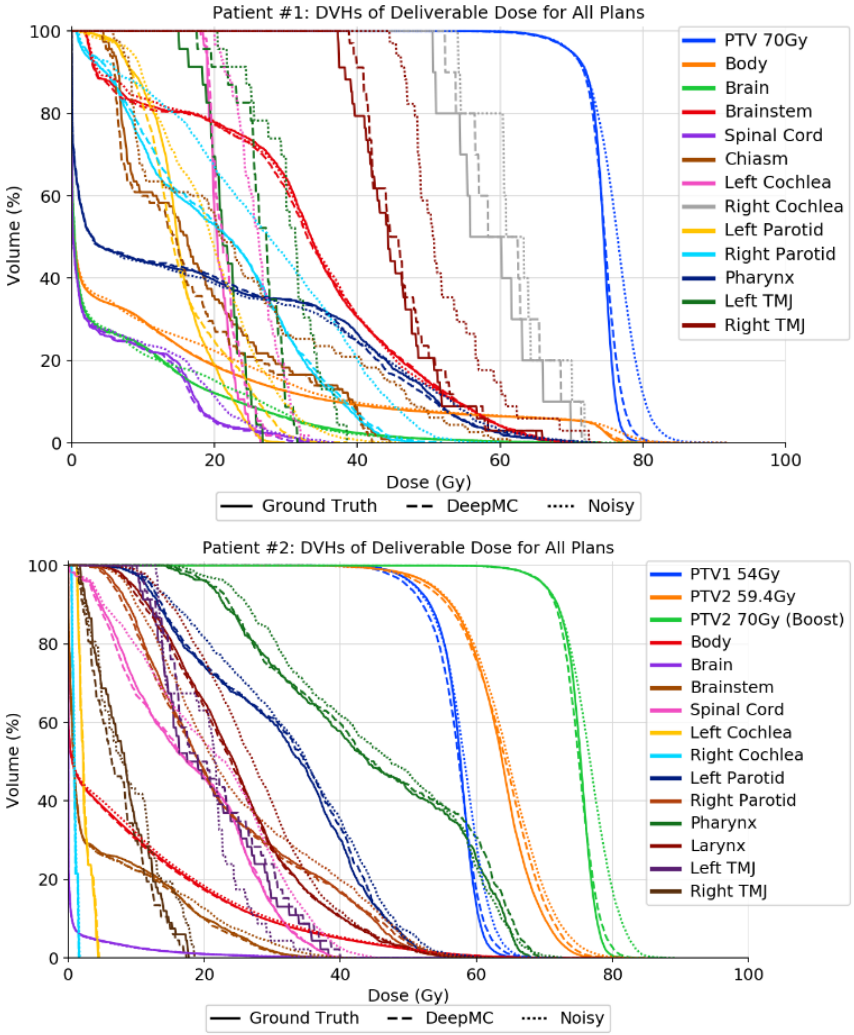


Figure 5-7. Dose volume histogram comparing "deliverable" dose for IMRT treatment plans created using low-noise ground truth, DeepMC-predicted, and high-noise beamlet dose for testing patient one (left) and two (right). Doses for all plans are recalculated after plan optimization using low-noise beamlet dose to reflect the deliverable dose to each patient.

Figures 5-A1 through 5-A4 compare 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 5-A2 and Figure 5-A4) for both 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 5-A1 and Figure 5-A3) match the ground truth plan in quality and would, therefore, pass clinical evaluation.

Table 5-1 presents the clinical plan quality metrics for PTV and OAR structures from the deliverable dose for each of the first testing patient's plans. Table 5-B1 presents the same metrics for the second testing patient. Dose is listed in units of Gy. Conformity number (CN) is unitless, with unity being ideal. The homogeneity index is unitless as well, with lower values indicating better dose homogeneity.

Table 5-1. Plan quality metrics derived from deliverable dose to the first testing patient for plans optimized using different beamlet dose calculation methods.

	D_{mean}			D_{max}		
Structure	**GT**	**Noisy**	**DeepMC**	**GT**	**Noisy**	**DeepMC**
PTV (70Gy)	74.19	76.20	74.43	83.39	91.90	84.87
Body	10.89	11.74	10.92	83.39	91.90	84.87
Brain	6.24	6.53	6.16	73.77	79.40	74.06
Brainstem	32.10	32.18	31.89	70.56	70.43	74.06
Spinal Cord	4.81	5.45	4.69	34.75	36.64	33.01
Chiasm	18.09	23.75	16.79	44.57	61.41	42.06
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

PTV Dose Metrics			
Metric	**GT**	**Noisy**	**DeepMC**
CN	0.79	0.78	0.79
HI	0.15	0.23	0.18
$D1$	78.08	83.98	79.94
$D2$	77.54	82.97	79.15
$D98$	67.14	67.16	66.89
$D99$	65.14	65.29	64.70

Absolute dose is presented in units of Gy. Plans with best agreement to ground truth for each metric and structure are shaded green.

The deliverable dose for the DeepMC plan PTV D_{mean} and D_{max} 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 in Table 5-1.

Axial dose color wash slices at the center of the PTV for the first testing patient are presented in Figure 5-8 and Figure 5-9 (and for the second testing patient in Figure 5-C1 and Figure 5-C2). Figure 5-8 and Figure 5-C1 compare 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.

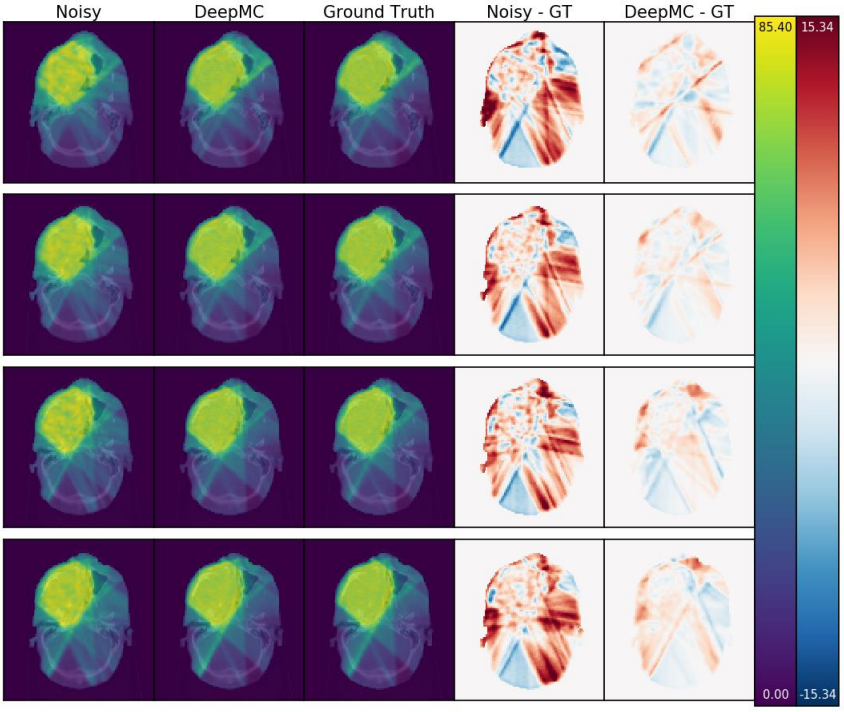


Figure 5-8. Deliverable dose color washes for axial slices from the first testing patient. Deliverable dose for each plan is recalculated using low-noise beamlet dose after plan optimization. The last two columns show differences in deliverable dose attributed to using either high-noise or DeepMC-predicted dose approximations to optimize IMRT beamlet fluence. Colors scales are consistent for each row; scale limits shown on the color bar in absolute dose units of Gy.

Figure 5-9 and Figure 5-C2 compare the planning dose with the deliverable dose for each beamlet dose calculation strategy. Dose difference slices labeled as ΔNoisy and ΔDeepMC show regions where the efficiently-calculated estimate of the plan quality (planning dose) deviates from the true plan quality (deliverable dose), which is considered as the true dose delivered to the patient if using the optimized X-ray fluences in each scenario. ΔNoisy reports a global noise corruption in the planning dose estimate matching the noise signature of high-noise MC dose. ΔDeepMC shows that it's planning dose slightly overestimates the anterior PTV dose and underestimates the posterior PTV, but the magnitude of error in planning dose is substantially lower than that of the noisy plans.

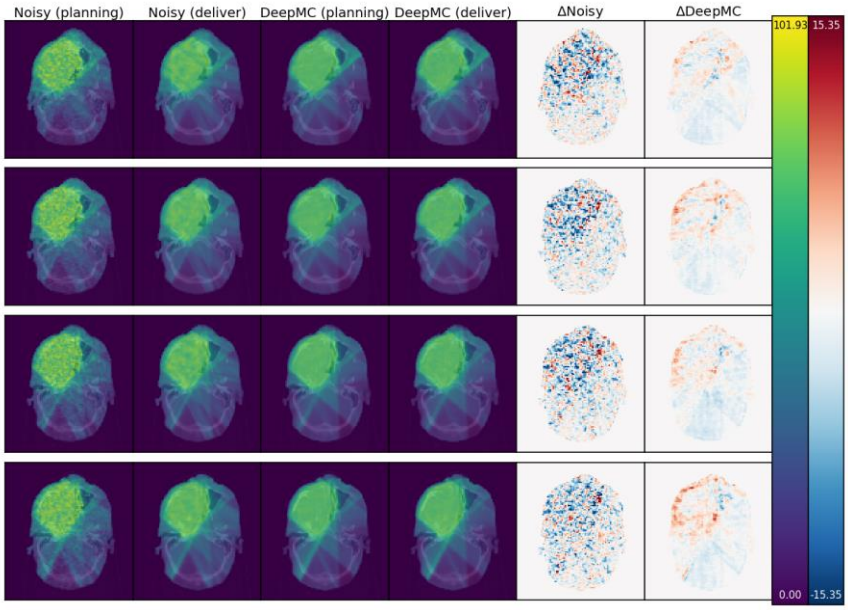


Figure 5-9. Planning and deliverable dose color washes for axial slices from the first testing patient. Planning dose is used directly for plan optimization. Deliverable dose for each plan is recalculated using low-noise beamlet dose after plan optimization. The Last two columns show differences in planning and deliverable dose for each plan. Color scales are consistent for each row; scale limits shown on the color bar are in absolute dose units of Gy.

Figure 5-10 and Figure 5-D1 reveal the treatment delivery parameters (per-beamlet fluence) resulting from optimization for each plan. For both 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.

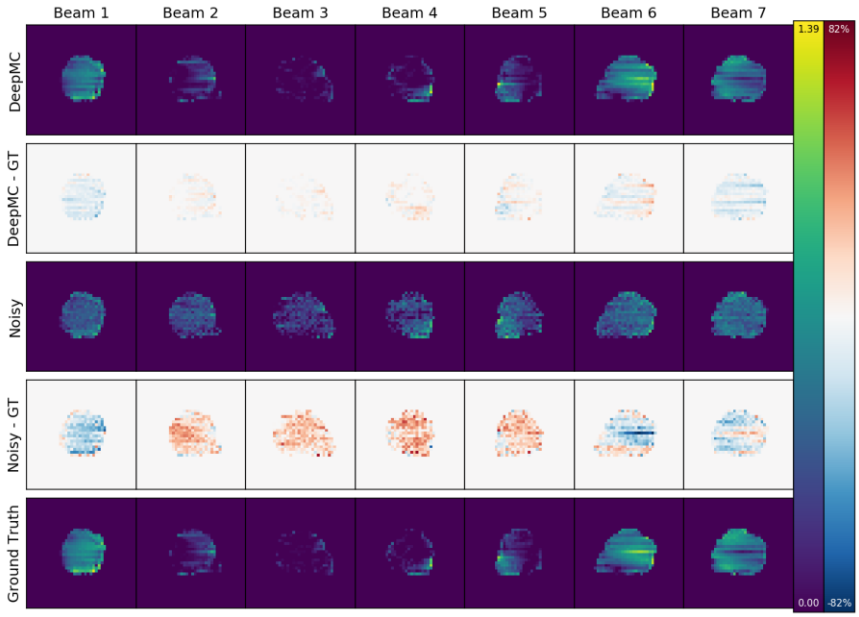


Figure 5-10. Plan delivery parameters (X-ray beamlet fluence) for the first testing patient resulting from optimization using DeepMC, high-, and low-noise beamlet dose. All color scales are consistent, and limits are shown in the color bars on the right. All are normalized to the per-beam maximum fluence from the ground truth plan.

5.5 Discussion

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. 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.

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[38,43] and volumetric modulated arc therapy[40,41] 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.

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. Distinct from MC dose denoising methods that use restrictive imaging filters, the mechanism of DeepMC is established on the physical processes of particle transportation in the presence of the magnetic field and with knowledge of the underlying anatomy. Therefore, DeepMC is more effective than an imaging denoiser because it operates from a mechanistic understanding of the dose deposition process, learned from a diverse population of examples.

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, indicating that DeepMC is more clinically useful for treatment planning than high-noise MC simulation, 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 both testing patients.

Although the current study is for a specific MRgRT problem, our method can be generalized to other areas that benefit from or are enabled by faster MC calculation. Within radiotherapy, the accuracy of proton and heavy ion therapy planning has also been hampered by slow MC dose calculation[146]. For cone-beam CT, estimation of scattered x-rays with MC is effective for reducing imaging artifacts from iterative reconstruction, but is prohibitively slow for online applications[147]. In the wider scope of computer graphics, MC optical ray tracing for photorealistic scene rendering is now used extensively for final rendering of cinematic animation and visual effects[148] but is currently too slow for high quality and noise-free interactive use and real-time applications such as live action environment generation[149] and interactive virtual and augmented reality experiences where more efficient but less realistic rendering approaches are commonplace.

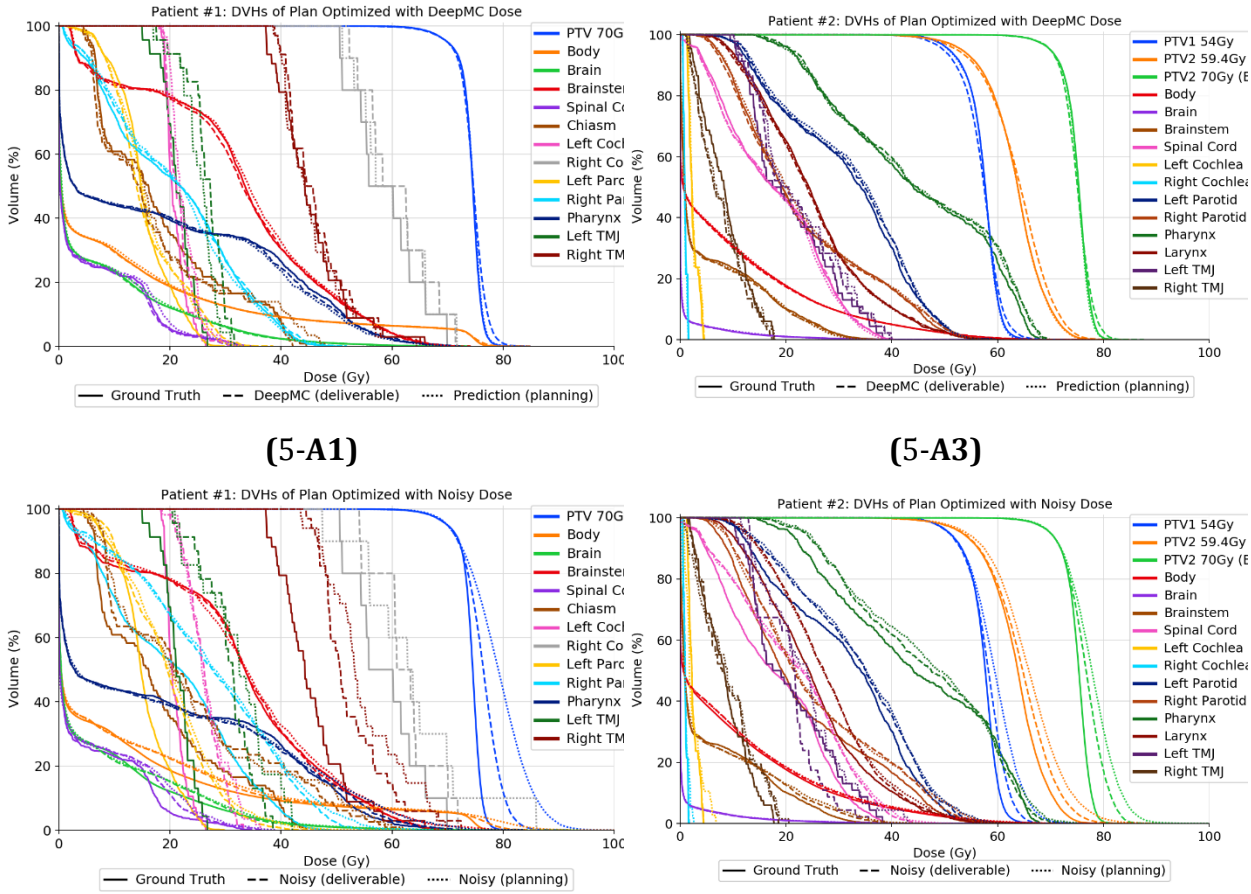
5.6 Conclusions

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 clinical 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.

5.7 Appendices

5.7.1 Appendix A. Dose Volume Histograms



Figures 5-A1 through 5-A4. Dose volume histograms for treatment plans created using DeepMC-predicted (top) and high-noise MC-simulated (bottom) 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.

5.7.2 Appendix B. Plan Quality Metrics

Table 5-B1. Plan quality metrics derived from deliverable dose to the second testing patient for plans optimized using different beamlet dose calculation methods.

	D_{mean}			D_{max}		
Structure	GT	Noisy	DeepMC	GT	Noisy	DeepMC
PTV #1 (54Gy)	57.17	57.91	56.96	72.77	71.06	72.86
PTV #2 (59.4Gy)	63.34	64.35	63.72	80.80	85.56	85.05
PTV #2 (70Gy)	74.96	76.54	74.90	83.43	89.35	87.60
Body	8.82	9.16	8.75	77.75	78.72	81.86
Brain	0.76	0.76	0.76	56.60	58.62	56.90
Brainstem	5.52	5.79	5.35	40.49	42.22	38.54
Spinal Cord	17.96	21.29	17.87	40.14	45.68	40.48
R TMJ	8.44	8.97	7.50	17.70	18.56	17.39
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
L parotid	31.47	33.31	31.96	59.93	61.87	58.73
R cochlea	0.99	1.05	0.95	1.63	1.72	1.56
L cochlea	2.46	2.49	2.53	4.41	4.39	4.55
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

PTV Dose Metrics				
	Metric	GT	Noisy	DeepMC
54 Gy	CN	0.19	0.18	0.17
	HI	0.26	0.31	0.31
	$D2$	62.40	64.88	63.76
	$D98$	48.27	48.22	46.87
59.4 Gy	CN	0.41	0.40	0.39
	HI	0.41	0.45	0.44
	$D2$	73.63	76.20	74.66
	$D98$	49.37	49.22	48.33
70 Gy	CN	0.78	0.70	0.73
	HI	0.17	0.23	0.19
	$D2$	79.04	83.57	80.32
	$D98$	67.29	67.41	67.34

Absolute dose is presented in units of Gy. Plans with best agreement to ground truth for each metric and structure are shaded green.

5.7.3 Appendix C. 3D Dose Comparison

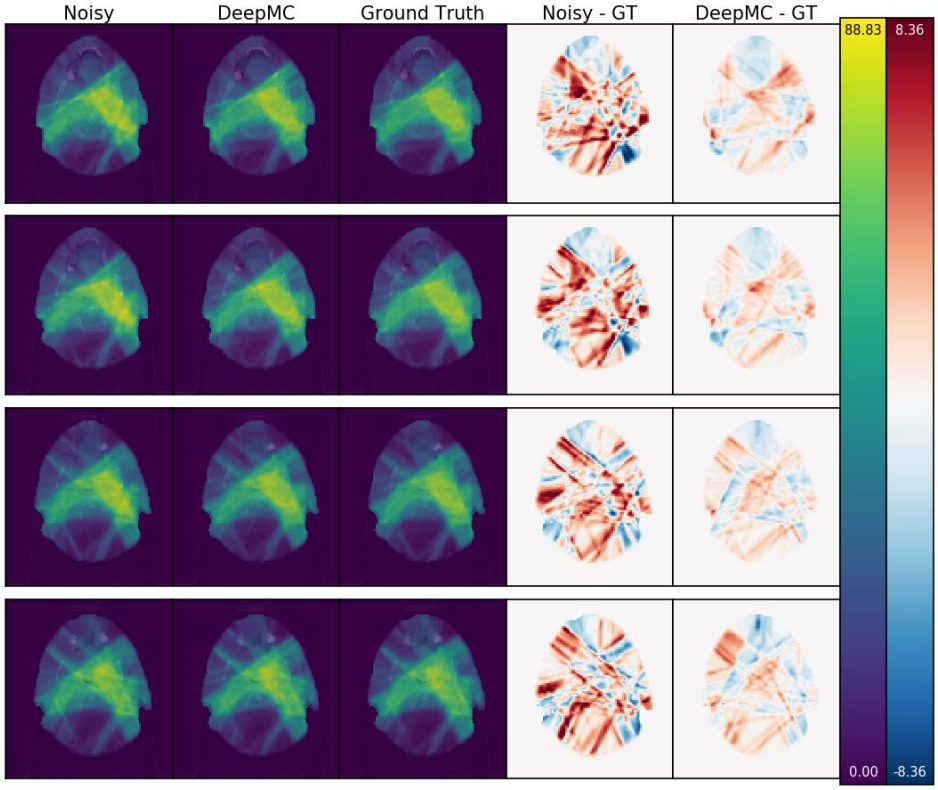


Figure 5-C1. Deliverable dose washes for axial slices from the second testing patient. Deliverable dose for each plan is recalculated using low-noise beamlet dose after plan optimization. The last two columns show differences in deliverable dose attributed to using either high-noise or DeepMC-predicted dose approximations to optimize IMRT beamlet fluence. Colors scales are consistent for each row; scale limits shown on the color bar in absolute dose units of Gy.

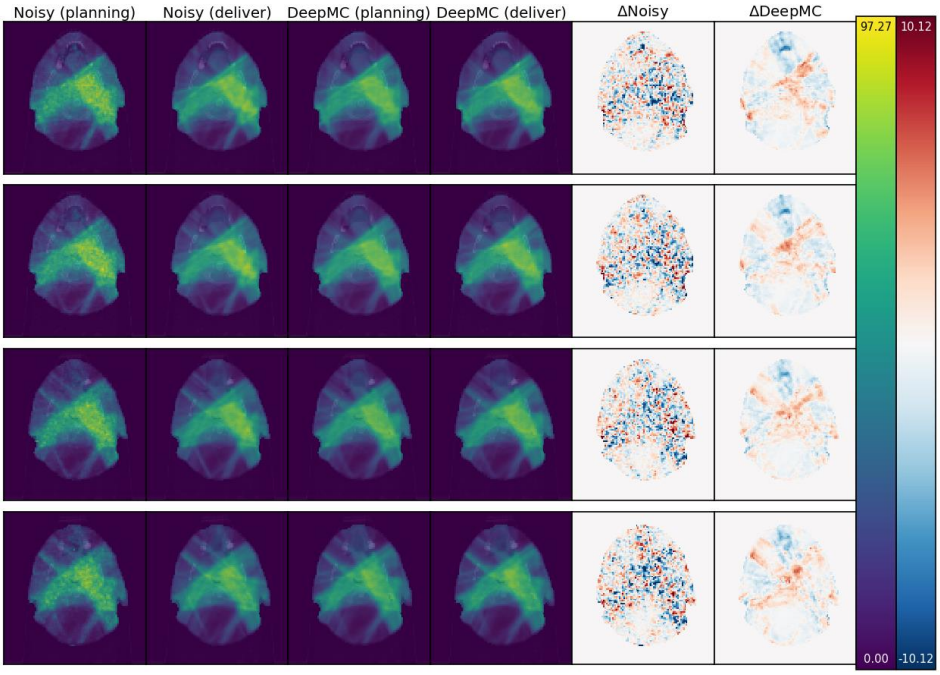


Figure 5-C2. Planning and deliverable dose color washes for axial slices from the second testing patient. Planning dose is used directly for plan optimization. Deliverable dose for each plan is recalculated using low-noise beamlet dose after plan optimization. The Last two columns show differences in planning and deliverable dose for each plan. Color scales are consistent for each row; scale limits shown on the color bar are in absolute dose units of Gy.

5.7.4 Appendix D. Fluence Map Comparison

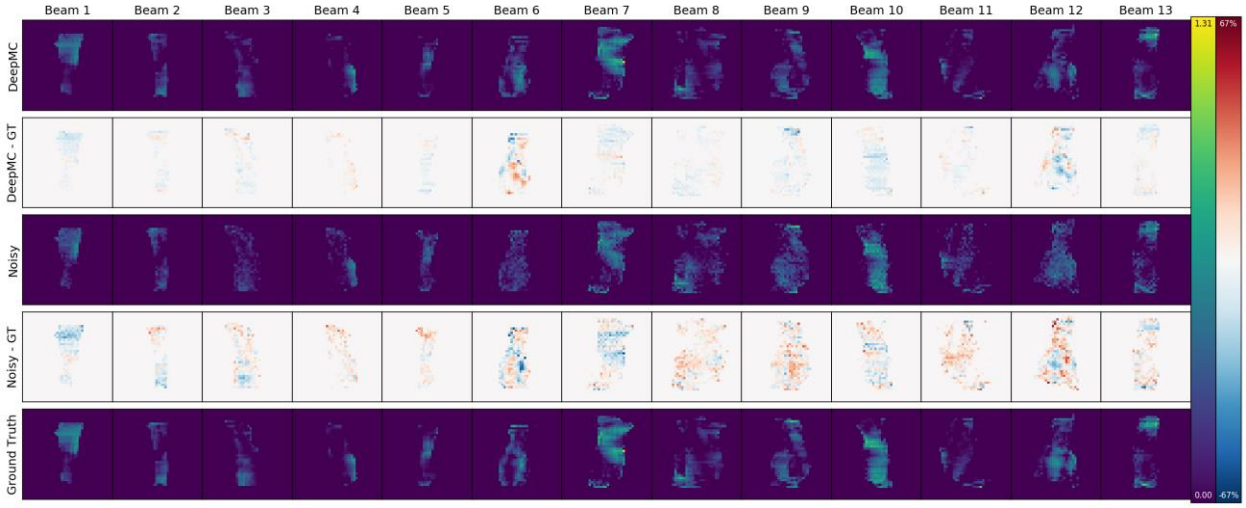


Figure 5-D1. Plan delivery parameters (X-ray beamlet fluence) for the second testing patient one resulting from optimization using DeepMC, high-, and low-noise ground truth beamlet dose. All color scales are matched consistent, and limits are shown in the color bars on the right. All are normalized to the per-beam ground truth maximum fluence from the ground truth plan.

6 SUMMARY OF WORK

Dose calculation speed is a matter concerning the full spectrum of radiation treatment paradigms, especially those at the cutting-edge for which slow dose calculation has barred translation from research to clinical application. This work focuses on identifying applications of clinical radiation dose calculation that, until now, were not suitably fast to accommodate more advanced, and more computationally expensive, planning approaches. To these applications, principles of high performance computing and machine learning were applied in order to provide substantial acceleration to the dose calculation process, affording more time to advanced forms of planning with access to more causal dose information than previously achievable in the same timeframe.

The widely used CCCS algorithm, relied upon for a large fraction CT-guided X-ray treatment planning, was first investigated. A novel approach for the implementation of CCCS dose calculation on GPU hardware took advantage of the compact deposition of dose surrounding narrow X-ray beamlets to parallelize the calculation across hundreds of beamlets at once. This *context-based* GPU-CCCS method realizes higher utilization of GPU hardware, offering two to three orders of magnitude of acceleration compared to existing implementations of CCCS on GPU hardware. Further improvements to acceleration with near-linear scaling in performance were demonstrated via a distributed, multi-GPU framework, offering flexibility of use to a wide variety of applications.

Next, a focus on accelerating large-scale MR-guided X-ray dose calculation resulted in the development of an automatically distributed and scalable framework for MC dose calculation. The framework was implemented using Docker, a software containerization and

orchestration tool, making it simple to install and scale across a new user's network of computational nodes. A low-level user interface was developed, allowing scripted manipulation of a database of simulation task definitions and structured exporting of calculated doses. A high-level interface, SimpleDose, was also developed to conceal Docker-specific details of usage for greater ease. SimpleDose was tested on a cluster of 8 nodes and a combined total of 216 CPU cores, offering a simulation rate of approximately 21 million primary particles per hour, for near theoretical performance scaling,

The distributed MC simulation framework offers a stop-gap solution to very-large-scale dose calculation for accelerating radiotherapy research but remains too slow for the extremely demanding clinical application of MR-guided OART. To simultaneously reduce the MC dose calculation time and benefit from existing MC acceleration techniques, a deep convolutional neural network (CNN) approach was developed to predict low-noise dose from substantially undersampled, noisy MC-simulated dose. The distributed MC framework was utilized to generate a large dataset of paired high- and low-noise doses for model training, and the feasibility of using DeepMC for OART treatment planning was demonstrated for two clinical H&N patients with IMRT dose calculation times under 11 minutes, for a 38-fold acceleration compared to conventional low-noise MC dose calculation. Additionally, DeepMC has potential for application in other treatment modalities such as proton and heavy-ion therapy where MC dose calculation is common-place and cumbersome to the treatment planning process.

This work defines new expectations on the speed of radiation dose calculation time that enable clinical translation of several beneficial new paradigms in radiation therapy. The

focus on scalable techniques for dose calculation make the proposed approaches applicable to the full gamut of treatment scenarios.

7 REFERENCES

1. Lomax ME, Folkes LK, O'Neill P. Biological consequences of radiation-induced DNA damage: Relevance to radiotherapy. *Clin Oncol*. 2013;25(10):578-585. doi:10.1016/j.clon.2013.06.007

2. Attix FH. *Introduction to Radiological Physics and Radiation Dosimetry*. Wiley; 1986. doi:10.1002/9783527617135

3. Klein O, Nishina Y. Über die Streuung von Strahlung durch freie Elektronen nach der neuen relativistischen Quantendynamik von Dirac. *Zeitschrift für Phys*. 1929;52(11-12):853-868. doi:10.1007/BF01366453

4. Raaijmakers AJE, Raaymakers BW, Lagendijk JJW. Magnetic-field-induced dose effects in MR-guided radiotherapy systems: Dependence on the magnetic field strength. *Phys Med Biol*. 2008;53(4):909-923. doi:10.1088/0031-9155/53/4/006

5. 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

6. Raaijmakers AJE, Raaymakers BW, Lagendijk JJW. Integrating a MRI scanner with a 6 MV radiotherapy accelerator: Dose increase at tissue-air interfaces in a lateral magnetic field due to returning electrons. *Phys Med Biol*. 2005;50(7):1363-1376. doi:10.1088/0031-9155/50/7/002

7. Neuenschwander H, Born EJ. A macro Monte Carlo method for electron beam dose calculations. *Phys Med Biol*. 1992;37(1):107-125.

8. Salvat F, Fernández-Varea JM, Sempau J, Mazurier J. Practical aspects of Monte Carlo simulation of charged particle transport: Mixed algorithms and variance reduction techniques. *Radiat Environ Biophys*. 1999;38(1):15-22. doi:10.1007/s004110050133

9. Fippel M. Variance Reduction Techniques. In: *Monte Carlo Techniques in Radiation Therapy*. 1st ed. Boca Raton: Taylor & Francis Group; 2013:29-39.

10. 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

11. 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

12. Jia X, Gu X, Sempau J, Choi D, Majumdar A, Jiang SB. Development of a GPU-based Monte

Carlo dose calculation code for coupled electron–photon transport. *Phys Med Biol*. 2010;55(11):3077-3086. doi:10.1088/0031-9155/55/11/006

13. Jia X, Schümann J, Paganetti H, Jiang SB. GPU-based fast Monte Carlo dose calculation for proton therapy. *Phys Med Biol*. 2012;57(23):7783-7797. doi:10.1088/0031-9155/57/23/7783

14. Hissoiny S, Ozell B, Bouchard H, Despŕs P. GPUMCD: A new GPU-oriented Monte Carlo dose calculation platform. *Med Phys*. 2011;38(2):754-764. doi:10.1118/1.3539725

15. 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

16. 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

17. Mackie TR, Bielajew AF, Rogers DWO, Battista JJ. Generation of photon energy deposition kernels using the Monte Carlo code. *Phys Med Biol*. 1988;33(1):1-20. doi:10.1088/0031-9155/33/1/001

18. Papanikolaou N, Mackie TR, Meger-Wells C, Gehring M, Reckwerdt P. Investigation of the convolution method for polyenergetic spectra. *Med Phys*. 1993;20(5):1327-1336. doi:10.1118/1.597154

19. 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.

20. Ahnesjö A. Collapsed cone convolution of radiant energy for photon dose calculation in heterogeneous media. *Med Phys*. 1989;16(4):577-592. doi:10.1118/1.596360

21. Helen Liu H, Mackie TR, McCullough EC. Correcting kernel tilting and hardening in convolution/superposition dose calculations for clinical divergent and polychromatic photon beams. *Med Phys*. 1997;24(11):1729-1741. doi:10.1118/1.597960

22. Oelfke U, Scholz C. Dose Calculation Algorithms. In: Schlegel WC, Bortfeld T, Grosu A-L, eds. *New Technologies in Radiation Oncology*. 1st ed. Springer; 2006:187-196.

23. Mohan R, Chui C, Lidofsky L. Differential pencil beam dose computation model for photons. *Med Phys*. 1986;13(1):64-73. doi:10.1118/1.595924

24. Hasenbalg F, Neuenschwander H, Mini R, Born EJ. Collapsed cone convolution and analytical anisotropic algorithm dose calculations compared to VMC++ Monte Carlo simulations in clinical cases. *Phys Med Biol*. 2007;52(13):3679-3691. doi:10.1088/0031-9155/52/13/002

25. Failla GA, Wareing T, Archambault Y, Thompson S. Acuros XB Advanced Dose Calculation For the Eclipse Treatment Planning System. *Varian Clin Perspect*. 2015.

26. Kan MWK, Yu PKN, Leung LHT. A review on the use of grid-based Boltzmann equation solvers for dose calculation in external photon beam treatment planning. *Biomed Res Int*. 2013;2013. doi:10.1155/2013/692874

27. Vassiliev ON, Wareing TA, McGhee J, Failla G, Salehpour MR, Mourtada F. Validation of a new grid-based Boltzmann equation solver for dose calculation in radiotherapy with photon beams. *Phys Med Biol*. 2010;55(3):581-598. doi:10.1088/0031-9155/55/3/002

28. Zhen H, Hrycushko B, Lee H, et al. Dosimetric comparison of Acuros XB with collapsed cone convolution/superposition and anisotropic analytic algorithm for stereotactic ablative radiotherapy of thoracic spinal metastases. *J Appl Clin Med Phys*. 2015;16(4):181-192. doi:10.1120/jacmp.v16i4.5493

29. Antonella F, Giorgia N, Alessandro C, Eugenio V, Pietro M, Luca C. Dosimetric validation of the Acuros XB Advanced Dose Calculation algorithm: fundamental characterization in water. *Phys Med Biol*. 2011;56(9):2885. http://stacks.iop.org/0031-9155/56/i=9/a=C02.

30. Wang A, Maslowski A, Wareing T, Star-Lack J, Schmidt TG. A fast, linear Boltzmann transport equation solver for computed tomography dose calculation (Acuros CTD). *Med Phys*. 2019;46(2):925-933. doi:10.1002/mp.13305

31. Takahashi S. Conformation radiotherapy. Rotation techniques as applied to radiography and radiotherapy of cancer. *Acta Radiol Diagn (Stockh)*. 1965:Suppl 242:1+. http://www.ncbi.nlm.nih.gov/pubmed/5879987.

32. Schlegel WC, Grosser KH, Häring P, Rhein B. Beam Delivery in 3D Conformal Radiotherapy Using Multi-Leaf Collimators. In: Schlegel WC, Bortfeld T, Grosu A-L, eds. *New Technologies in Radiation Oncology*. 1st ed. Springer; 2006:257-266.

33. Brahme A, Roos J-E, Lax I. Solution of an integral equation encountered in rotation therapy. *Phys Med Biol*. 1982;27(10):1221-1229. doi:10.1088/0031-9155/27/10/002

34. 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

35. Sini C, Broggi S, Fiorino C, Cattaneo GM, Calandrino R. Accuracy of dose calculation algorithms for static and rotational IMRT of lung cancer: A phantom study. *Phys Medica*. 2015;31(4):382-390. doi:10.1016/j.ejmp.2015.02.013

36. 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

37. Nguyen D, Lyu Q, Ruan D, O'Connor D, Low DA, Sheng K. A comprehensive formulation for volumetric modulated arc therapy planning. *Med Phys*. 2016;43(7):4263-4272. doi:10.1118/1.4953832

38. 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.

39. Gu W, O'Connor D, Nguyen D, et al. Integrated beam orientation and scanning-spot optimization in intensity-modulated proton therapy for brain and unilateral head and neck tumors. *Med Phys*. 2018;45(4):1338-1350. doi:10.1002/mp.12788

40. 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

41. 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

42. Gu W, Ruan D, O'Connor D, et al. Robust optimization for intensity-modulated proton therapy with soft spot sensitivity regularization. *Med Phys*. January 2019. doi:10.1002/mp.13344

43. 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

44. Lyu Q, Neph R, Yu VY, Ruan D, Sheng K. Single-arc VMAT optimization for dual-layer MLC. *Phys Med Biol*. 2019;64(9):095028. doi:10.1088/1361-6560/ab0ddd

45. Chambolle A, Pock T. A first-order primal-dual algorithm for convex problems with applications to imaging. *J Math Imaging Vis*. 2011;40(1):120-145. doi:10.1007/s10851-010-0251-1

46. 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

47. Mackie TR, Scrimger JW, Battista JJ. A convolution method of calculating dose for 15-MV x rays. *Med Phys*. 1984;12(2):188-196. doi:10.1118/1.595774

48. Dong P, Lee P, Ruan D, et al. 4π non-coplanar liver SBRT: A novel delivery technique. *Int J Radiat Oncol Biol Phys*. 2013;85(5):1360-1366. doi:10.1016/j.ijrobp.2012.09.028

49. Yu VY, Landers A, Woods K, et al. A Prospective 4π Radiation Therapy Clinical Study in Recurrent High-Grade Glioma Patients. *Int J Radiat Oncol*. 2018;101(1):144-151.

doi:10.1016/j.ijrobp.2018.01.048

50. Tran A, Zhang J, Woods K, et al. Treatment planning comparison of IMPT, VMAT and 4π radiotherapy for prostate cases. *Radiat Oncol*. 2017;12(1):10. doi:10.1186/s13014-016-0761-0

51. Smyth G, Evans PM, Bamber JC, et al. Non-coplanar trajectories to improve organ at risk sparing in volumetric modulated arc therapy for primary brain tumors. *Radiother Oncol*. 2016;121(1):124-131. doi:10.1016/j.radonc.2016.07.014

52. Panet-Raymond V, Ansbacher W, Zavgorodni S, et al. Coplanar versus noncoplanar intensity-modulated radiation therapy (IMRT) and volumetric-modulated arc therapy (VMAT) treatment planning for fronto-temporal high-grade glioma. *J Appl Clin Med Phys*. 2012;13(4):44-53. doi:10.1120/jacmp.v13i4.3826

53. Hsieh C-H, Liu C-Y, Shueng P-W, et al. Comparison of coplanar and noncoplanar intensity-modulated radiation therapy and helical tomotherapy for hepatocellular carcinoma. *Radiat Oncol*. 2010;5(1):40. doi:10.1186/1748-717X-5-40

54. Woods K, Lee P, Kaprealian T, Yang I, Sheng K. Cochlea-sparing acoustic neuroma treatment with 4π radiation therapy. *Adv Radiat Oncol*. 2018;3(2):100-107. doi:10.1016/j.adro.2018.01.004

55. Amit G, Purdie TG, Levinshtein A, et al. Automatic learning-based selection of beam angles in radiation therapy of lung cancer. In: *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*. IEEE; 2014:230-233. doi:10.1109/ISBI.2014.6867851

56. Bangert M, Unkelbach J. Accelerated iterative beam angle selection in IMRT. *Med Phys*. 2016;43(3):1073-1082. doi:10.1118/1.4940350

57. Li Y, Yao J, Yao D. Automatic beam angle selection in IMRT planning using genetic algorithm. *Phys Med Biol*. 2004;49(10):1915-1932. doi:10.1088/0031-9155/49/10/007

58. Smyth G, Bamber JC, Evans PM, Bedford JL. Trajectory optimization for dynamic couch rotation during volumetric modulated arc radiotherapy. *Phys Med Biol*. 2013;58(22):8163-8177. doi:10.1088/0031-9155/58/22/8163

59. 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

60. 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

61. Neylon J, Sheng K, Yu V, et al. A nonvoxel-based dose convolution/superposition

algorithm optimized for scalable GPU architectures. *Med Phys*. 2014;41(10):101711. doi:10.1118/1.4895822

62. Tian Z, Shi F, Folkerts M, Qin N, Jiang SB, Jia X. A GPU OpenCL based cross-platform Monte Carlo dose calculation engine (goMC). *Phys Med Biol*. 2015;60(19):7419-7435. doi:10.1088/0031-9155/60/19/7419

63. Ziegenhein P, Kozin IN, Kamerling CP. Towards real-time photon Monte Carlo dose calculation in the cloud Towards real-time photon Monte Carlo dose calculation in the cloud. 2017;(Iccr 2016).

64. Park JC, Li JG, Arhjoul L, et al. Adaptive beamlet-based finite-size pencil beam dose calculation for independent verification of IMRT and VMAT. *Med Phys*. 2015;42(4):1836-1850. doi:10.1118/1.4914858

65. Cho N, Tsiamas P, Velarde E, et al. Validation of GPU-accelerated superposition-convolution dose computations for the Small Animal Radiation Research Platform. *Med Phys*. 2018;45(5):2252-2265. doi:10.1002/mp.12862

66. Hissoiny S, Ozell B, Despŕs P. A convolution-superposition dose calculation engine for GPUs. *Med Phys*. 2010;37(3):1029-1037. doi:10.1118/1.3301618

67. Hissoiny S, Ozell B, Després P. Fast convolution-superposition dose calculation on graphics hardware. *Med Phys*. 2009;36(6Part1):1998-2005. doi:10.1118/1.3120286

68. Jacques R, Taylor R, Wong J, McNutt T. Towards real-time radiation therapy: GPU accelerated superposition/convolution. *Comput Methods Programs Biomed*. 2010;98(3):285-292. doi:10.1016/j.cmpb.2009.07.004

69. Lu W. A non-voxel-based broad-beam (NVBB) framework for IMRT treatment planning. *Phys Med Biol*. 2010;55(23):7175-7210. doi:10.1088/0031-9155/55/23/002

70. Piotrowski T, Skośka M, Jodda A, et al. Tomotherapy - A different way of dose delivery in radiotherapy. *Wspolczesna Onkol*. 2012;16(1):16-25. doi:10.5114/wo.2012.27332

71. Siddon RL. Fast calculation of the exact radiological path for a three-dimensional CT array. *Med Phys*. 1984;12(2):252-255. doi:10.1118/1.595715

72. Ahnesjö A, Aspradakis MM, Ahnesj A, Aspradakis MM. Dose calculations for external photon beams in radiotherapy Dose calculations for external photon beams in radiotherapy. *Phys Med Biol*. 1999;44:99-155. doi:10.1088/0031-9155/44/11/201

73. Zhong H, Chetty IJ. Generation of a novel phase-space-based cylindrical dose kernel for IMRT optimization. *Med Phys*. 2012;39(5):2518-2523. doi:10.1118/1.3700403

74. Li Y, Tian Z, Shi F, et al. A new Monte Carlo-based treatment plan optimization

approach for intensity modulated radiation therapy. *Phys Med Biol*. 2015;60(7):2903-2919. doi:10.1088/0031-9155/60/7/2903

75. Kirk DB, Hwu WW. *Programming Massively Parallel Processors*. 2nd ed. Burlington, MA: Elsevier, Inc.; 2010. doi:10.1016/B978-0-12-381472-2.00001-5

76. Lu W, Olivera GH, Chen M, Reckwerdt PJ, Mackie TR. Accurate convolution / superposition for multi-resolution dose calculation using cumulative. *Phys Med Biol*. 2005;655:655-680. doi:10.1088/0031-9155/50/4/007

77. Tillikainen L, Helminen H, Torsti T, et al. A 3D pencil-beam-based superposition algorithm for photon dose calculation in heterogeneous media. *Phys Med Biol*. 2008;53(14):3821-3839. doi:10.1088/0031-9155/53/14/008

78. Arnfield MR, Siantar CH, Siebers J, Garmon P, Cox L, Mohan R. The impact of electron transport on the accuracy of computed dose. *Med Phys*. 2000;27(6):1266-1274. doi:10.1118/1.599004

79. Woods K, Nguyen D, Neph R, Ruan D, O'Connor D, Sheng K. A sparse orthogonal collimator for small animal intensity-modulated radiation therapy part I: Planning system development and commissioning. *Med Phys*. 2019;46(12):5703-5713. doi:10.1002/mp.13872

80. Woods K, Neph R, Nguyen D, Sheng K. A sparse orthogonal collimator for small animal intensity-modulated radiation therapy. Part II: hardware development and commissioning. *Med Phys*. 2019;46(12):5733-5747. doi:10.1002/mp.13870

81. Harrington KJ, Billingham LJ, Brunner TB, et al. Guidelines for preclinical and early phase clinical assessment of novel radiosensitisers. *Br J Cancer*. 2011;105(5):628-639. doi:10.1038/bjc.2011.240

82. Kahn J, Tofilon PJ, Camphausen K. Preclinical models in radiation oncology. *Radiat Oncol*. 2012;7(1):223. doi:10.1186/1748-717X-7-223

83. Rosenthal N, Brown S. The mouse ascending: perspectives for human-disease models. *Nat Cell Biol*. 2007;9(9):993-999. doi:10.1038/ncb437

84. Verhaegen F, Granton P, Tryggestad E. Small animal radiotherapy research platforms. *Phys Med Biol*. 2011;56(12):R55-R83. doi:10.1088/0031-9155/56/12/R01

85. Stuben G, Landuyt W, van der Schueren E, van der Kogel A. Different immobilization procedures during irradiation influence the estimation of alpha/beta ratios in mouse lip mucosa. *Strahlentherapie und Onkol Organ der Dtsch Rontgengesellschaft*. 1993;169(11):678-683. https://www.ncbi.nlm.nih.gov/pubmed/8248845.

86. Kitakabu Y, Yuta S, Keisuke S, Koji O, Mitsuyuki A. Variations of the hypoxic fraction in the SCC VII tumors after single dose and during fractionated radiation therapy:

Assessment without anesthesia or physical restraint of mice. *Int J Radiat Oncol*. 1991;20(4):709-714. doi:10.1016/0360-3016(91)90013-T

87. Nikjoo H, Lindborg L. RBE of low energy electrons and photons. *Phys Med Biol*. 2010;55(10):R65-R109. doi:10.1088/0031-9155/55/10/R01

88. Chow JCL, Leung MKK, Lindsay PE, Jaffray DA. Dosimetric variation due to the photon beam energy in the small-animal irradiation: A Monte Carlo study. *Med Phys*. 2010;37(10):5322-5329. doi:10.1118/1.3488979

89. Verhaegen F, van Hoof S, Granton P V., Trani D. A review of treatment planning for precision image-guided photon beam pre-clinical animal radiation studies. *Z Med Phys*. 2014;24(4):323-334. doi:10.1016/j.zemedi.2014.02.004

90. van Hoof SJ, Granton P V., Verhaegen F. Development and validation of a treatment planning system for small animal radiotherapy: SmART-Plan. *Radiother Oncol*. 2013;109(3):361-366. doi:10.1016/j.radonc.2013.10.003

91. Butterworth KT, Prise KM, Verhaegen F. Small animal image-guided radiotherapy: status, considerations and potential for translational impact. *Br J Radiol*. 2015;88(1045):20140634. doi:10.1259/bjr.20140634

92. Nguyen D, Ruan D, O'Connor D, et al. A novel software and conceptual design of the hardware platform for intensity modulated radiation therapy. *Med Phys*. 2016;43(2):917-929. doi:10.1118/1.4940353

93. Dong P, Lee P, Ruan D, et al. 4π Noncoplanar Stereotactic Body Radiation Therapy for Centrally Located or Larger Lung Tumors. *Int J Radiat Oncol*. 2013;86(3):407-413. doi:10.1016/j.ijrobp.2013.02.002

94. Rwigema J-CM, Nguyen D, Heron DE, et al. 4π Noncoplanar Stereotactic Body Radiation Therapy for Head-and-Neck Cancer: Potential to Improve Tumor Control and Late Toxicity. *Int J Radiat Oncol*. 2015;91(2):401-409. doi:10.1016/j.ijrobp.2014.09.043

95. Nguyen D, Rwigema J-CM, Yu VY, et al. Feasibility of extreme dose escalation for glioblastoma multiforme using 4π radiotherapy. *Radiat Oncol*. 2014;9(1):239. doi:10.1186/s13014-014-0239-x

96. Yu VY, Tran A, Nguyen D, et al. Significant Cord and Esophagus Dose Reduction by 4π Non-Coplanar Spine Stereotactic Body Radiation Therapy and Stereotactic Radiosurgery. *Int J Radiat Oncol*. 2016;96(2):E646. doi:10.1016/j.ijrobp.2016.06.2246

97. Yu VY, Tran A, Nguyen D, et al. The development and verification of a highly accurate collision prediction model for automated noncoplanar plan delivery. *Med Phys*. 2015;42(11):6457-6467. doi:10.1118/1.4932631

98. Woods K, Nguyen D, Tran A, et al. Viability of Noncoplanar VMAT for liver SBRT

compared with coplanar VMAT and beam orientation optimized 4π IMRT. *Adv Radiat Oncol*. 2016;1(1):67-75. doi:10.1016/j.adro.2015.12.004

99. Murzin VL, Woods K, Moiseenko V, et al. 4π plan optimization for cortical-sparing brain radiotherapy. *Radiother Oncol*. 2018;127(1):128-135. doi:10.1016/j.radonc.2018.02.011

100. Low DA, Harms WB, Mutic S, Purdy JA. A technique for the quantitative evaluation of dose distributions. *Med Phys*. 1998;25(5):656-661. doi:10.1118/1.598248

101. Otto K. Volumetric modulated arc therapy: IMRT in a single gantry arc. *Med Phys*. 2007;35(1):310-317. doi:10.1118/1.2818738

102. Sheng K, Shepard DM, Orton CG. Noncoplanar beams improve dosimetry quality for extracranial intensity modulated radiotherapy and should be used more extensively. *Med Phys*. 2015;42(2):531-533. doi:10.1118/1.4895981

103. Liu Y, Shi C, Tynan P, Papanikolaou N. Dosimetric characteristics of dual-layer multileaf collimation for small-field and intensity-modulated radiation therapy applications. *J Appl Clin Med Phys*. 2008;9(2):15-29. doi:10.1120/jacmp.v9i2.2709

104. Liu Y, Shi C, Lin B, Ha CS, Papanikolaou N. Delivery of four-dimensional radiotherapy with TrackBeam for moving target using a dual-layer MLC: dynamic phantoms study. *J Appl Clin Med Phys*. 2009;10(2):21-33. doi:10.1120/jacmp.v10i2.2926

105. Giantsoudi D, Stathakis S, Liu Y, Shi C, Papanikolaou N. Monte Carlo Modeling and Commissioning of a Dual-layer Micro Multileaf Collimator. *Technol Cancer Res Treat*. 2009;8(2):105-114. doi:10.1177/153303460900800203

106. Varian Halcyon Dosimetric Comparison for Multi-Arc VMAT Prostate and Head-and-Neck Cancers. *Med Dosim*. 2019;44(3):301. doi:10.1016/j.meddos.2018.06.004

107. Michiels S, Poels K, Crijns W, et al. Volumetric modulated arc therapy of head-and-neck cancer on a fast-rotating O-ring linac: Plan quality and delivery time comparison with a C-arm linac. *Radiother Oncol*. 2018;128(3):479-484. doi:10.1016/j.radonc.2018.04.021

108. Reinelt G. TSPLIB—A Traveling Salesman Problem Library. *ORSA J Comput*. 1991;3(4):376-384. doi:10.1287/ijoc.3.4.376

109. 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

110. 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

111. 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

112. 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

113. 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

114. 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

115. 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

116. 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

117. 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

118. 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

119. 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

120. 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

121. 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

122. 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

123. Pfaffenberger A. Dose Calculation Algorithms for Radiation Therapy with an MRI-Integrated Radiation Device. 2013. doi:10.11588/heidok.00014551

124. 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

125. 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

126. 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

127. 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

128. 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

129. 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

130. 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

131. Naqa I El, Deasy JO, Vicic 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

132. 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

133. Fornander H. Denoising Monte Carlo Dose Calculations Using a Deep Neural Network. 2019. http://www.diva-portal.org/smash/get/diva2:1366439/FULLTEXT01.pdf.

134. 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

135. Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. 2015. doi:10.1007/s13398-014-0173-7.2

136. 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

137. 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

138. 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

139. 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.

140. 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

141. 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.

142. 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

143. 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

144. 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).

145. 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

146. 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

147. 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

148. Christensen P, Bannister M, Rayner B, et al. RenderMan. *ACM Trans Graph*. 2018;37(3):1-21. doi:10.1145/3182162

149. Industrial Light & Magic. ILM Stagecraft. https://www.ilm.com/hatsrabbits/ilm-stagecraft/. Published 2019. Accessed April 9, 2020.