IRRADIATION PLANNING IN SMALL ANIMAL RADIATION BIOLOGY RESEARCH

PhD thesis

by

Justin MALIMBAN

1

INTRODUCTION

1

1.1. RADIOTHERAPY

According to GLOBOCAN estimates from the International Agency for Research on Cancer (IARC) of the World Health Organization (WHO), at least 20 million people were diagnosed with cancer across 185 countries in 2022, leading to over 9.7 million deaths [1]. By 2050, the incidence of cancer is projected to increase by 77%. Lung cancer is the most prevalent, accounting for 12.4% of all cases, followed by breast cancer at 11.6% and colorectal cancer at 9.6%. The primary treatment options for cancer are surgery, chemotherapy, radiotherapy, or a combination of these approaches. The choice of treatment greatly depends on the type and location of the tumour, its stage and growth rate, and whether metastases are present. More than 50% of cancer patients receive radiotherapy during the course of their treatment, whether for curative or palliative purposes [2]. Available radiotherapy modalities include brachytherapy, where radioactive sources are placed within or around the tumour inside the patient, and external beam radiation therapy (EBRT), which delivers radiation to the cancer site using an accelerator that produces photons, electrons, neutrons, protons or heavy ions.

The goal of radiotherapy is to destroy or shrink the tumour to increase the patient's chances of survival and preserve quality of life. This is accomplished by maximizing the radiation dose to the tumour while minimizing the exposure of surrounding healthy tissues. The majority of patients are treated with conventional megavoltage (MV) X-ray beams, which exhibit an exponentially decreasing energy deposit after a dose build-up close to the skin. This technique results in co-irradiation of tissues proximal and distal to the target volume, which can limit local tumour control due to constraints imposed by normal tissue toxicity. In contrast, proton beams exhibit lower entrance doses and concentrate dose deposition at the end of their range, with little to no dose beyond the Bragg peak as shown in figure 1.1. As a result, dose delivery can be better tailored to the tumour allowing a more conformal dose distribution and reducing unnecessary exposure of surrounding anatomical structures particularly those distal to the target volume. Thus, proton therapy potentially offers substantial improvement in the therapeutic window as compared to conventional radiotherapy techniques by allowing greater escalation of tumour doses and reduction of normal tissue complications.

Because of its physical advantages, there has been a steady increase in the use of proton therapy in treating cancer indications such as brain, head and neck, lung, and breast tumours as well as paediatric malignancies [4, 5]. There is also a rise in the number of proton centres around the world, with 121 currently in operation and 28 more under construction as of June 2024 [6]. As the number of cancer patients treated with proton therapy continues to increase, there is also a demand for high-level evidence support-

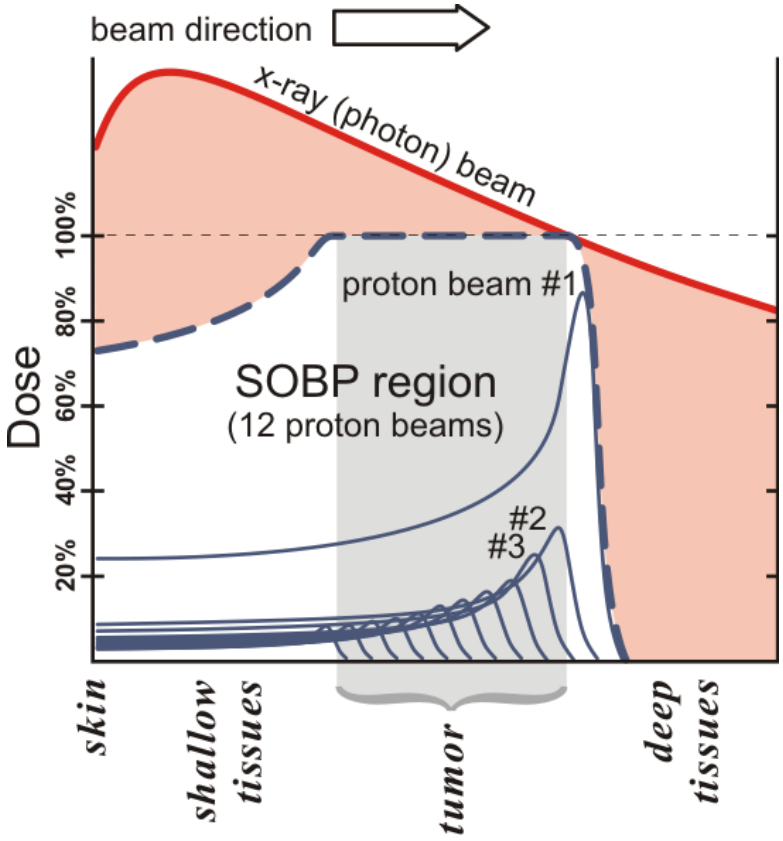


Figure 1.1: Depth dose distribution of X-rays (red) vs protons (blue). Spread-out Bragg peak (SOBP) beam is created by superposition of several pristine Bragg peaks [3].

ing its use. In line with this, preclinical studies are being conducted aiming to address open questions regarding proton radiobiology, long term side effects, regional tissue radiosensitivities, immunomodulatory effects, and new treatment strategies such as very high dose rate irradiations (FLASH) [7, 8] and spatial fractionation [9, 10]. Such research studies both *in vitro* and *in vivo* are critical in determining and demonstrating the clinical potential of proton therapy (or any radiotherapy modality in general) beyond just differences in the dose distribution.

1.2. PRECLINICAL RESEARCH

Preclinical studies play a crucial role in cancer research because they serve as an experimental system for investigating the biological, chemical, and physical aspects of radi-

ation response. Insights gained from preclinical research help to formulate hypotheses about observations in patients, which can then be tested in clinical studies to determine their relevance and potentially improve treatment outcomes.

For example, experiments on *in vitro* and *in vivo* models (findings summarized by Paganetti et al. [11, 12]) have shown that the relative biological effectiveness[1] (RBE) of protons is variable and tends to be progressively higher towards the distal end of the Bragg peak [13, 14]. This is contrary to the simplistic assumption of a constant RBE of 1.1 (relative to ^{60}Co) in routine clinical practice. In reality, RBE depends on the depth, showcasing an increase of as high as 79% in the RBE-weighted dose using experimental RBEs compared to the RBE-weighted dose calculated using the constant clinical RBE [15]. A higher RBE at the distal dose falloff can result in unexpected toxicity to healthy tissues behind the tumour, which raises questions on whether a fixed RBE of 1.1 is an oversimplification given its clinical implications [16]. Considerable variation has also been observed in the experimental RBE as it was found to be influenced by a number of factors such as the energy, dose per fraction, tissue and cell types, and the specific biological endpoint being studied [12, 17]. These uncertainties in the RBE highlight the need for caution in RBE-based treatment planning in the clinic and underscore the necessity for more extensive validation in preclinical models. Furthermore, similar to protons, heavy ions also benefit from the physical characteristics of the beam due to the Bragg peak but exhibit an even higher RBE for cell killing. Additional preclinical research into the differential response of ions to photons and protons will help guide future clinical studies for optimizing heavy charged particle therapy.

Since any radiotherapy treatment inevitably results in some dose delivered to healthy tissues, normal tissue toxicity remains a limiting factor and a key focus of preclinical research. In normal tissue studies, animal models serve as frameworks to observe acute and late effects of radiation damage especially on critical dose-limiting organs. For instance, irradiations of the heart in murine models have demonstrated radiation-induced cardiac injury such as decreased microvascular density [18, 19, 20] and increased fibrosis in the myocardium and pericardium [20, 21]. *In vivo* experiments on heart and lung irradiation with rats showed that pulmonary function loss does not depend only on the dose in the lungs but is also aggravated by co-irradiation of the heart [22]. In parallel, irradiation of the lungs has been shown to contribute to the development of cardiac side effects [23]. These findings have been instrumental in refining management of radiation treatments for thoracic, oesophageal, and breast cancer patients, who are at risk of devel-

[1] Relative biological effectiveness (RBE) is the ratio of the dose of a reference radiation (D_R) to the dose of a test radiation (D_T) required to produce the same biological effect. Commonly, photons (e.g. ^{60}Co) are used as the reference radiation.

oping radiation-induced lung injury (RILI) and radiation-induced heart disease (RIHD). They also suggest that there is an interplay between these two organs and both influence damage in the other. To gain more insights into the individual effects of heart and lung irradiation, to identify organ regions that contribute the most to the development of side effects, and to further clarify the correlation between their responses, animal experiments with more targeted radiation delivery need to be performed (e.g. portion of the heart is irradiated while limiting the dose to the lungs, or vice versa). Such experiments have now become more feasible with the recent technological advancements in preclinical image-guided irradiation platforms. Proton beams, which exhibit greater conformality, can also be used for more localized irradiations, as has been done with stem cell-rich ducts in the parotid glands [24] and hippocampus in the brain [25]. Results from these studies demonstrated regional sensitivities in organs and suggest that sparing these regions may reduce radiation-induced side effects—xerostomia and neurocognitive dysfunction, respectively. These outcomes highlight the need for more comprehensive preclinical studies using more focused dose distributions so that we can gain more understanding and make informed decisions on which tissues or regions to spare and how to optimize dose escalation during radiation treatments.

While normal tissue toxicity is a major concern in radiotherapy, the ultimate goal is to achieve local tumour control. Tumour response studies for proton therapy and proton minibeam radiotherapy have been done with breast [26] and brain tumour-bearing animals [27, 28], respectively. These studies used syngeneic models, where inbred animal cancer cell lines are transplanted into immunocompetent animals. This makes them also suitable for immunological studies. However, it is important to exercise caution as the tumour cells are not of human origin and therefore express different histology, which may present difficulties in interpretation and translation to humans [29]. On the other hand, xenograft models have been employed in preclinical studies of conventional [30] and ultra-high dose rate [31] proton therapy as well as combination treatments with chemotherapy [32] and targeted radionuclide therapy [33]. Xenograft models show greater fidelity to human tumours as they are established from human cancer cell lines or patient tumour biopsies [34]. However, they are implanted in immunodeficient animals and successful engraftment is not always possible for all tumour types [35, 36].

Tumour models can be implanted orthotopically (at the site of origin) or ectopically (at a different location, typically subcutaneous). Orthotopic tumours better reflect clinical conditions as they are engrafted at the site of origin, which offers better resemblance to the native tumour microenvironment, than the more commonly utilized subcutaneous ones [37, 38]. As more sophisticated orthotopic tumour models continue to be-

come available, we now need to rely less on simplistic subcutaneous tumour models and move on to *in vivo* systems that better mimic clinically relevant environments. This provides unique new opportunities for more in-depth studies to answer a variety of radiobiological questions on tumour response, which potentially offer more predictive value for the direction of patient studies. However, a lot of work is yet to be done to establish the validity of these preclinical tumour models.

Ongoing oncological studies are also being conducted on new radiotherapy modalities that are being developed. Preclinical evidence has already indicated that using ultra-high dose rates ($\geq$ 40 Gy/s) in FLASH therapy [39, 40] and spatial dose modulation in minibeam radiotherapy [41, 42] can promote increased healthy tissue sparing without impacting tumour response. This differential response potentially offers reduced side effects and consequently, improved treatment outcomes. Although the protection from radiation-induced toxicity has been demonstrated in *in vivo* models [28, 43, 44, 45], the underlying biological mechanisms mediating the sparing effect in these emerging radiotherapy modalities remain unclear. Contradicting results have also been reported, such as the development of bone necrosis in cats treated with FLASH [46]. Therefore, extensive animal experiments are still needed to fully realize the potential and limitations of these therapies, to identify and optimize the parameters that trigger protective effects, and to support their clinical implementation.

1.3. OVERVIEW OF PRECLINICAL IRRADIATION PLATFORMS

As mentioned in the previous section, there is a need for capabilities in delivering highly controlled dose distributions to small target volumes in order to facilitate next generation radiation biology experiments. Early preclinical experiments were limited in this regard as they were usually performed using cabinet X-ray irradiators, which generally come with a fixed source, non-adjustable collimators, and without on-board imaging system. This type of irradiation platform can only deliver large beam sizes, limiting it to non-conformal, whole-body exposure of animals. Owing to technological developments in the past 15 years, there has now been a transition from cabinet irradiators to more sophisticated platforms with a rotating gantry or rotating couch to allow multiple field delivery, micro-cone beam computed tomography (μ-CBCT) for image-guidance and planning, as well as a dedicated treatment planning system (TPS) for dose calculations. Table 1.1 summarizes the specifications of some notable small animal X-ray irradiation platforms.

Two devices have since been commercialized: the Small Animal Radiation Research Platform (SARRP), developed by Johns Hopkins University in collaboration with Xstrahl

Inc. (Suwanee, GA) [47], and the Small Animal Radiation Therapy (SmART+), developed by Princess Margaret Cancer Center and marketed by Precision X-ray, Inc. (Madison, CT) [48]. Currently, there are over two hundred of these machines in use worldwide. The SmART+ features a C-arm gantry with the X-ray tube mounted opposite an amorphous silicon flat panel detector, which rotates 360° around a stationary specimen during imaging and treatment. On the other hand, acquisition of CT data and arc delivery can be accomplished by gantry and/or animal couch rotation for SARRP. These systems also offer optical imaging such as bioluminescence tomography (BLT) and bioluminescence imaging (BLI) as complementary imaging modalities. Aside from aiding in target localization, these allow longitudinal monitoring of changes in tumour volume, gene expression, and treatment response [49].

More recently, facilities have started to integrate these commercially available X-ray imager/irradiator platforms to experimental proton beamlines to take advantage of their onboard CBCT imaging system and robotic translation stages for image-guided proton irradiation of animals. The use of particle beams is also aimed at improving conformality of the dose distribution to target volumes in animals and for differential response comparison with photons. Several modifications have been implemented to adapt the proton beamline for use with animals. The lead shielding encasing the X-ray system is usually removed to make way for the proton beamline. Collimators with diameters ≤ 5 mm were added to allow creation of very small fields. Energy degraders were also incorporated to tune down the energy of the protons to levels appropriate for irradiating shallower regions (< 40 MeV). Table 1.2 summarizes facilities with SARRP or SmART+ combined with proton beamlines for preclinical research.

Table 1.1: Specifications of dedicated small animal X-ray research platforms (taken and modified from Ghita et al. 2019) [50]. This list does not include irradiators capable of FLASH dose rates.

Device	Vendor/Institute	Beam energy (keV)	Dose rate (Gy/min)	Image-guidance	TPS
Commercial					
SARRP[1]	Xstrahl Inc. (USA)	5-225	1-4	CBCT BLT[6]	Muriplan SmART-XPS
SmART+[2]	Precision X-ray Inc. (USA)	5-225	0.01-4	CBCT BLI[7]	SmART-ATP
Academic					
Micro-CT/RT	Stanford University (GE eXplore RS120) [51]	70-120	2	CBCT	In-house
	The University of Western Ontario (GE eXplore CT/RT 140) [52]	70-140	2	CBCT	In-house
	Heidelberg University (YXLON Y.Fox) [53]	10-160	4.5-6.4	CBCT	In-house
SAIGRT[3] [54]	Technical University of Dresden	10-225	1-4	CBCT	In-house
SACRTD[4] [55]	University of Arkansas	60-225	0.4-3	CBCT	In-house
iSMAART[5] [56, 57, 58]	University of Miami	45-225	2.5-4	CBCT BLT FMT[8]	In-house

[1]Small Animal Radiation Research Platform (SARRP), [2]Small Animal Radiation Therapy (SmART+), [3]Small Animal Image Guided Radiation Therapy (SAIGRT), [4]Small Animal Conformal Radiation Therapy Device (SACRTD), [5]Image guided SMall Animal Arc Radiation Treatment (iSMAART), [6]Bioluminescence tomography (BLT), [7]Bioluminescence imaging (BLI), [8]Fluoresence molecular tomography (FMT)

Table 1.2: List of facilities with small animal X-ray research platforms integrated with a proton beamline. The asterisk (*) indicates facilities that are currently under construction.

Device	Institution	Accelerator	Maximum energy (MeV)	Platform
PPPRP[1] [59]	University of Washington (US)	Scanditronix MC50	50	SARRP
Proton SARRP on rails [60]	University of Pennsylvania (US)	IBA C230	230	SARRP
SIRMIO[2] [61]	Ludwig Maximilian University of Munich (DE)	Depends on facility	Depends on facility	Offline imaging (SARRP)
SAPPHIRE[3] [62]	Technical University of Dresden (DE)	IBA C230	230	SmART+ IB
MiniBEE[4]* [63]	Helmoholtz Zentrum Berlin (DE)	HZB cyclotron (eye tumours)	68	SARRP
University of Manchester*	University of Manchester (UK)	Varian ProBeam	250	SARRP
IMPACT[5]*	Particle Therapy Research Centre Groningen (NL)	AGOR cyclotron	190	SARRP

[1]Precision Preclinical Particle Radiotherapy Platform (PPPRP), [2]Small Animal Proton Irradiator for Research in Molecular Image-guided Radiation-Oncology (SIRMIO), [3]Small Animal Proton and Photon Image-guided Radiation Experiments (SAPPHIRE), [4]MINIbeam BEamline for preclinical Experiments (MiniBEE), [5]Image guided Proton/pArticle infrastructure for preclinical sTudies (IMPACT)

1

1.4. PRECLINICAL IRRADIATION WORKFLOW

The general workflow for image-guided radiation experiments on small animals closely resembles the clinical treatment course in humans as shown in figure 1.2. First, the animal is placed at the irradiation position and 3D volumetric images (μ-CBCT) are acquired. The imaging data is used to identify and delineate volumes of interest such as the target and organs-at-risk (OAR). Based on these organ contours, an irradiation plan is created, in which the beam configuration is decided, and dose objective and constraints to relevant tissues are set. The dose distribution associated with the plan is calculated, and once an optimized plan has been reached, this plan is delivered to the animal. In contrast to the clinic, wherein the treatment planning happens several days prior to the start of treatment, the entire preclinical workflow is ideally performed in a single session, lasting between 20-90 minutes depending on the complexity of the plan [64]. Throughout this process, the animal is kept under anaesthesia, which imposes time constraints since prolonged exposure to anaesthesia may affect the outcome of experiments and the animal's well-being [65, 66].

In the following subsections, current challenges in the irradiation planning process in animals will be discussed. For clarity, it is important to note that the terminologies "contour", "delineation", "segmentation", and "annotation" will be used interchangeably throughout this thesis.

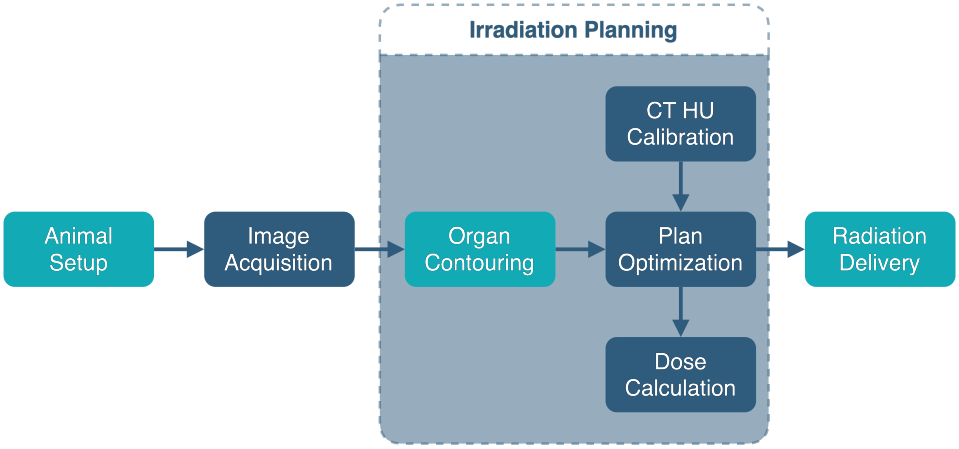


Figure 1.2: Preclinical irradiation workflow

1.4.1. ORGAN CONTOURING

Given that animals are imaged at the treatment position and irradiated shortly thereafter, irradiation planning must be completed in the shortest time possible to minimize exposure to anaesthesia. Similar to the clinic, organ contouring is one of the major bottlenecks that hinders a fast workflow for online dose delivery in animals. In practice, delineation of organs in an animal is mostly performed manually by a biologist. For radiobiological studies wherein a large number of animals may be irradiated, this approach is very time consuming, laborious, and prone to errors due to poor soft tissue contrast in animals especially in the brain and abdomen [67, 68]. Delineating a single organ can take 10 minutes or more [69, 70], potentially extending the process to over an hour per animal, depending on the number of organs to be delineated. Additionally, the quality of the contours, which can have an impact on irradiation quality, can vary from one biologist to the other based on their level of experience. For studies on normal tissue response, wherein around 100 animals are irradiated in a single experiment [71], organ contours of a few representative animals (excluded from the experiment) are usually created beforehand. The average organ contours of these animals are then applied to delineate the entire population. This approach can be deemed acceptable since animals used in *in vivo* experiments are mostly clones of each other, having the same age, sex, weight, food, and activity regimen, and they are also subjected to the same environmental conditions. Therefore, drastic variations in anatomical structures are unlikely compared to humans. However, for studies where morphological variation is critical such as animals implanted with orthotopic tumours, individualized contours specific to the animal are needed to obtain better predictive value of preclinical studies. For such cases, automating the delineation of organs is essential. Not only is the time spent in the contouring process minimized, but the overall human workload is greatly reduced. Auto-contouring tools have also shown to provide more consistent, reliable, and reproducible organ delineations [68, 72]. By reducing the variability in the contours, suboptimal plans that lead to dose errors within the animal population are avoided, and more accurate radiation delivery can be achieved. This can potentially lead to the reduction of the number of animals required to obtain statistically significant results in *in vivo* studies.

Over the years, efforts have been dedicated to developing automated tools for delineating anatomical structures in animals, aiming to both shorten the contouring time and improve consistency. In atlas-based segmentation, a reference model known as an atlas—featuring pre-segmented organ contours from previously acquired images—is employed [73]. The atlas contours are propagated onto the target CT image through deformable registration to facilitate the segmentation. Notable examples of whole-body

atlases of mouse anatomy include the MOBY phantom [74] and the Digimouse atlas [75]. These atlases, based on a single reference animal, are referred to as "classic single atlas" models. However, this approach is reported to produce inferior segmentation accuracy as it exhibits strong bias towards the selected atlas. To mitigate this issue, the "multi-atlas" approach, which incorporates multiple atlases from different animals, has been proposed. Compared to the single atlas, multi-atlas-based segmentation produced more accurate contours and is more robust against posture and anatomic variations [76, 77]. Despite being faster than manual contouring, as evidenced by the multi-atlas-based image segmentation (MABIS) algorithm, which took 12 minutes to delineate six organs in a mouse [78], the process could benefit from further optimization to enhance efficiency.

More recently, unparalleled advances in the field of machine learning have opened opportunities in automatically performing tasks such as medical image segmentation, image registration, and computer-aided detection and diagnosis. The general idea is that the algorithm utilizes relationships and dependencies between input features and output labels inferred from prior training in order to make predictions in new data. Essentially, the algorithm is learning how to create a mapping from a given input to a particular output based on what it has learned from the labelled training data. Demonstrating rapid execution times and improved contour consistency, the focus of auto-segmentation in the clinic has now shifted towards deep learning (DL) with convolutional neural networks [79, 80, 81]. This approach is also gaining traction in mouse organ segmentation as demonstrated by recent DL-based contouring of normal tissues [69, 72, 82], orthotopic lung tumours [83], and skeletal muscle tissue in mice [84]. While thorough validation of DL-based contouring models is being done in the clinic, that is not the case for preclinical models. For example, Schoppe et al. [72] evaluated their model's performance on an independent dataset but of the same type as the training data (i.e. CT scans excluded from the training set but sharing the same properties, such as mouse strain and imaging protocol). However, in reality, especially in facilities where researchers carry out a variety of experiments, it is unlikely that the same type of mouse and imaging protocols will be used all the time. Therefore, the performance of preclinical DL models must also be evaluated on external cases to determine their robustness and generalizability. Comparing 2D and 3D models could also provide valuable insights. While 2D models are typically faster, they rely solely on transverse slices, which can result in loss of craniocaudal information and potentially lead to erroneous results. In contrast, 3D models preserve the Z information, offering a fuller view of the anatomy during training but at the cost of increased computational demands and processing time.

1.4.2. DOSE CALCULATION

Dose computations also impose a bottleneck in the preclinical workflow. Monte Carlo (MC) simulation is considered the gold standard for dose calculations as it provides the most accurate results [85]. Since individual particles (primary and secondary) are transported and radiation interactions are modelled in detail, MC simulations can approximate the dose very well especially in highly heterogeneous regions, but that makes them very computationally expensive and thereby slow. Moreover, the spatial resolution of μ-CBCT images is typically in the order of 100 μm [86], requiring simulation of millions of particles to achieve sufficient precision. This limits its use in routine preclinical practice, where fast dose calculation is crucial while the animal is sedated.

Several methods have already been developed in the clinic to accelerate proton dose calculations. The standard dose engine in most treatment planning systems (TPS) is the pencil beam algorithm (PBA), which can provide reasonably accurate dose distributions in fairly homogeneous media within clinically acceptable timeframes [87, 88]. PBAs are considerably faster than general purpose MC codes, but their performance in the vicinity of highly heterogeneous regions deteriorates particularly when densities of adjacent materials greatly differ such as in the lungs and head and neck [89]. Since MC codes are inherently better at handling heterogeneities, several approaches have been developed in recent years to increase their efficiency by optimizing them to run on massively parallel CPU (central processing unit) [90] and GPU (graphics processing unit) [91, 92, 93] architectures. Alternative proton dose engines, such as those based on deterministic [94] and deep learning methods [95], have also been proposed to speed up dose calculations without sacrificing too much in terms of accuracy (i.e. they can achieve near-MC accuracies).

While computational tools are actively being developed for the clinic, corresponding solutions for preclinical applications remain considerably behind. Commercially available TPS for animals exist such as MuriPlan (Xstrahl Inc., Suwanee, Georgia, USA), SmART-ATP, and SmART-XPS (SmART Scientific Solutions BV, Maastricht, the Netherlands) for photons, while only μ-RayStation (RaySearch Laboratories, Stockholm, Sweden) provides a platform that supports both photons and protons. For kilovoltage (kV) X-ray beams, deep learning techniques have been recently utilized to denoise low statistics MC dose distributions [96, 97]. Since the DL algorithm will predict the equivalent high statistics dose distribution, fewer particles are needed in the MC simulation, thereby saving time. For protons, Vanstalle et al. [98] proposed an analytical model, building upon Bortfeld et al.'s [99] formulation of the Bragg curve. However, their methodology was only validated in a water phantom and not on more complex geometries with a lot of

inhomogeneities like a CT of an animal. Clausen et al. [100] tried to utilize the clinical treatment planning system RayStation to calculate doses for target sizes typical for *in vivo* studies. However, its validity is questionable as the smallest allowable dose grid is 1 × 1 × 1 mm^3, which is too coarse for small animals with anatomical structures smaller than a millimetre. Studies demonstrating fast and accurate proton dose calculations for plan generation in animals are scarce, despite the pressing need for such methods due to constraints introduced by animal sedation. Moreover, fast proton dose engines designed for humans have not been validated yet for use in animals, which could potentially alleviate long computation times.

1.4.3. CT HOUNSFIELD UNIT (HU) CALIBRATION

For spread-out Bragg peak (SOBP) experiments, in which protons stop inside the animal, the accuracy of the dose distribution largely depends on the predicted proton range. Range shifts can lead to underdosage in the target or overdosage in OARs, which degrades treatment quality. CT calibration in irradiation planning is one of the main contributing factors to range uncertainties [101]. When performing dose calculations, the CT data measured in Hounsfield units (HU) must be translated first to quantities required by the dose engine. The basic input for pencil beam algorithms is the stopping power ratio (SPR), while the mass density (ρ) and elemental composition are required for MC-based dose engines. Despite differences in the underlying physics of interaction between X-rays and protons, X-ray CT remains the standard imaging procedure for proton therapy, from which the quantities relevant for dose calculation are indirectly derived.

In most radiotherapy clinics, Hounsfield look-up tables (HLUT) are used for the calibration. These tables are obtained from single-energy CT (SECT) measurements on tissue-equivalent materials of known composition or stoichiometric method to derive values for human tissues [102, 103]. The major issue with SECT calibration is the lack of unique relationship between the CT numbers and the SPR. There is no one-to-one correspondence due to the difference in physics of X-rays and protons. This entails that tissues with identical CT numbers can have different SPRs, potentially resulting in inaccuracies. It has been reported that SECT calibration method can introduce range uncertainties between 3.0% to 3.5% in patients [104, 105, 106]. Literature have suggested that dual-energy CT (DECT) has the potential to extract tissue characteristics more accurately than SECT [107, 108, 109], possibly reducing the range uncertainty margins. By obtaining CT images at two different X-ray spectral distributions, additional information aside from the ρ_e (i.e. Z_{eff}) can be extracted. Z_{eff} is used to determine the mean excitation energy (I) [110], which is a quantity needed for the calculation of the SPR together with ρ_e. DECT

experiments on animal tissue samples have demonstrated a potential reduction in uncertainty to within 1% [108, 109, 111]. Despite these promising results, DECT is not yet standard practice in the clinic.

SECT and DECT methods have also been investigated in the preclinical context for kV X-ray beams. Notably, Schyns et al. [112] tested various energy combinations to determine the optimal parameters for DECT imaging in animals, whereas Vaniqui et al. [113] evaluated the impact of SECT and DECT approaches on the photon dose distribution. However, the merit of SECT and DECT calibration methods for proton dose calculation in animals is yet to be evaluated. Given that the effect of local deviations in the proton range is expected to be aggravated in animals due their small size, DECT could potentially enhance dose calculation accuracy. However, the very small voxel size in μ-CBCT images (~100 μm) necessitates high imaging doses to minimize noise and error in the calibration. According to Schyns et al. [112], a dose of at least 30 cGy per image, totalling 60 cGy per animal, is required to achieve a decent calibration. This prompts the question of whether the potential accuracy gain from DECT justifies the additional imaging dose.

1.5. AIM AND OVERVIEW OF THE THESIS

As discussed in the above, the current preclinical workflow is hindered by technical challenges such as the time-consuming and subjective nature of organ contouring, lack of fast dose calculation platforms, and uncertainties arising from the CT calibration. The overall aim of this thesis is to then establish methods to address these problems, thereby improving the efficiency and accuracy of the preclinical workflow in order to facilitate next generation image-guided radiobiology experiments with proton beams.

In **Chapter 2**, *"Deep learning-based segmentation of the thorax in mouse micro-CT scans"*, different 2D and 3D deep learning algorithms are investigated to develop a reliable auto-contouring platform for organs-at-risk in the mouse thorax. The performance of the trained models is evaluated for an independent dataset (i.e. CT images of the same mice but not included in the training) as well as an external dataset (i.e. CT images of mice of different strain and age taken using a different imaging protocol) to showcase their robustness and generalizability to out-of-distribution data. The interobserver variability was also analyzed to establish a human baseline. Lastly, the time gain in contouring from the deep learning models is presented.

In **Chapter 3**, *"Benchmarking of fast proton dose engines (MCsquare and YODA) for small animal irradiations"*, the suitability of two fast proton dose engines MCsquare and YODA to accelerate irradiation planning in small animals is evaluated. Dose calculations are performed in homogeneous and heterogeneous phantoms as well as in realistic an-

imal geometries (head, thoracic, and abdominal regions). The achievable speed and accuracy from these codes are compared to full Monte Carlo simulations.

In **Chapter 4**, *"A simulation framework for preclinical proton irradiation workflow"*, a simulation framework of the preclinical proton irradiation workflow is developed. This framework can be used to optimize beam properties, imaging protocols, and experimental designs prior to performing actual experiments. A key element of the framework is the μ-CBCT version of the fastCAT CBCT simulator [114], which has been developed to allow generation of realistic CT images of animals without actually using a live one. Beam transport simulations of a dedicated preclinical proton facility have also been performed to create a realistic beam model for treatment planning in an animal. Using this framework, SECT and DECT images needed to perform CT HU-to-SPR calibration has been generated. The accuracy of the calibration methods is evaluated through simulated proton radiographs and SOBP dose distributions in a mouse phantom.

Chapter 5 summarizes the main findings of the work described in this thesis. The impact, future implementation, shortcomings of this work, and remaining challenges in routine preclinical practice are also discussed.

BIBLIOGRAPHY

[1] Freddie Bray et al. “Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries”. In: *CA: A Cancer Journal for Clinicians* 74.3 (2024), pp. 229–263. DOI: 10.3322/caac.21834.

[2] World Health Organization. *New WHO/IAEA publication provides guidance on radiotherapy equipment to fight cancer*. 2021. URL: https://www.who.int/news/item/05-03-2021-new-who-iaea-publication-provides-guidance-on-radiotherapy-equipment-to-fight-cancer#:~:text=More%20than%2050%25%20of%20cancer,%2C%20colorectal%2C%20and%20lung%20cancer (visited on 08/08/2024).

[3] MarkFilipak. *File:Comparison of dose profiles for proton v. x-ray radiotherapy.png*. 2012. URL: https://commons.wikimedia.org/w/index.php?curid=27983203 (visited on 08/26/2024).

[4] Zhe Chen et al. “Proton versus photon radiation therapy: A clinical review”. In: *Frontiers in Oncology* 13.1133909 (2023). DOI: 10.3389/fonc.2023.1133909.

[5] W.P. Levin et al. “Proton beam therapy”. In: *British Journal of Cancer* 93 (2005), pp. 849–854. DOI: 10.1038/sj.bjc.6602754.

[6] PTCOG. *Particle therapy facilities in clinical operation*. 2024. URL: https://www.ptcog.site/index.php/facilities-in-operation-public (visited on 08/26/2024).

[7] Jonathan R. Hughes and Jason L. Parsons. “FLASH Radiotherapy: Current Knowledge and Future Insights Using Proton-Beam Therapy”. In: *International Journal of Molecular Sciences* 21.18 (2020). DOI: 10.3390/ijms21186492.

[8] Michele M. Kim et al. “Development of Ultra-High Dose-Rate (FLASH) Particle Therapy”. In: *IEEE Transactions on Radiation and Plasma Medical Sciences* 6.3 (2022), pp. 252–262. DOI: 10.1109/TRPMS.2021.3091406.

[9] Y. Prezado and G. R. Fois. “Proton-minibeam radiation therapy: A proof of concept”. In: *Medical Physics* 40.3 (2013), p. 031712. DOI: 10.1118/1.4791648.

[10] P. Lansonneur et al. “First proton minibeam radiation therapy treatment plan evaluation”. In: *Scientific Reports* 10.7025 (2020). DOI: 10.1038/s41598-020-63975-9.

[11] Harald Paganetti et al. “Relative biological effectiveness (RBE) values for proton beam therapy”. In: *International Journal of Radiation Oncology, Biology, Physics* 53.2 (2002), pp. 407–421. DOI: 10.1016/S0360-3016(02)02754-2.

[12] Harald Paganetti. “Relative biological effectiveness (RBE) values for proton beam therapy. Variations as a function of biological endpoint, dose, and linear energy transfer”. In: *Physics in Medicine and Biology* 59.22 (2014), R419. DOI: 10.1088/0031-9155/59/22/R419.

[13] Brita Singers Sørensen et al. “Relative biological effectiveness (RBE) and distal edge effects of proton radiation on early damage in vivo”. In: *Acta Oncologica* 56.11 (2017), pp. 1387–1391. DOI: 10.1080/0284186X.2017.1351621.

[14] Richard. A. Britten et al. “Variations in the RBE for Cell Killing Along the Depth-Dose Profile of a Modulated Proton Therapy Beam”. In: *Radiation Research* 179.1 (2012), pp. 21–28. DOI: 10.1667/RR2737.1.

[15] Pankaj Chaudhary et al. “Relative Biological Effectiveness Variation Along Monoenergetic and Modulated Bragg Peaks of a 62-MeV Therapeutic Proton Beam: A Preclinical Assessment”. In: *International Journal of Radiation Oncology, Biology, Physics* 90.1 (2014), pp. 27–35. DOI: 10.1016/j.ijrobp.2014.05.010.

[16] Christopher R. Peeler et al. “Clinical evidence of variable proton biological effectiveness in pediatric patients treated for ependymoma”. In: *Radiotherapy and Oncology* 121.3 (2016), pp. 395–401. DOI: 10.1016/j.radonc.2016.11.001.

[17] K. Ilicic, S.E. Combs, and T.E. Schmid. “New insights in the relative radiobiological effectiveness of proton irradiation”. In: *Radiation Oncology* 13.6 (2018). DOI: 10.1186/s13014-018-0954-9.

[18] Susanne Lauk et al. “Radiation-induced heart disease in rats”. In: *International Journal of Radiation Oncology, Biology, Physics* 11.4 (1985), pp. 801–808. DOI: 10.1016/0360-3016(85)90314-1.

[19] T. K. Yeung et al. “Morphological and Functional Changes in the Rat Heart after X Irradiation: Strain Differences”. In: *Radiation Research* 119.3 (1989), pp. 489–499. DOI: 10.2307/3577520. (Visited on 08/26/2024).

[20] Sarah C. Darby et al. “Radiation-Related Heart Disease: Current Knowledge and Future Prospects”. In: *International Journal of Radiation Oncology, Biology, Physics* 76.3 (2010), pp. 656–665. DOI: 10.1016/j.ijrobp.2009.09.064.

[21] Chang-Lung Lee et al. “Assessing Cardiac Injury in Mice With Dual Energy-MicroCT, 4D-MicroCT, and MicroSPECT Imaging After Partial Heart Irradiation”. In: *International Journal of Radiation Oncology, Biology, Physics* 88.3 (2014), pp. 686–693. DOI: 10.1016/j.ijrobp.2013.11.238.

[22] Peter van Luijk et al. “Radiation Damage to the Heart Enhances Early Radiation-Induced Lung Function Loss”. In: *Cancer Research* 65.15 (2005), pp. 6509–6511. DOI: 10.1158/0008-5472.CAN-05-0786.

[23] Julia Wiedemann, Robert P. Coppes, and Peter van Luijk. “Radiation-induced cardiac side-effects: The lung as target for interacting damage and intervention”. In: *Frontiers in Oncology* 12 (2022). DOI: 10.3389/fonc.2022.931023.

[24] Peter van Luijk et al. “Sparing the region of the salivary gland containing stem cells preserves saliva production after radiotherapy for head and neck cancer”. In: *Science Translational Medicine* 7.305 (2015), 305ra147–305ra147. DOI: 10.1126/scitranslmed.aac4441.

[25] Theresa Suckert et al. “Late Side Effects in Normal Mouse Brain Tissue After Proton Irradiation”. In: *Frontiers in Oncology* 10 (2021). DOI: 10.3389/fonc.2020.598360.

[26] Y. Kwon et al. “uppressive effects of a proton beam on tumor growth and lung metastasis through the inhibition of metastatic gene expression in 4T1 orthotopic breast cancer model”. In: *International Journal of Oncology* 49 (2016), pp. 336–342. DOI: 10.3892/ijo.2016.3520.

[27] Charlotte Lamirault et al. “Spatially Modulated Proton Minibeams Results in the Same Increase of Lifespan as a Uniform Target Dose Coverage in F98-Glioma-Bearing Rats”. In: *Radiation Research* 194.6 (2020), pp. 715–723. DOI: 10.1667/RADE-19-00013.1.

[28] Yolanda Prezado et al. “Tumor Control in RG2 Glioma-Bearing Rats: A Comparison Between Proton Minibeam Therapy and Standard Proton Therapy”. In: *International Journal of Radiation Oncology*Biology*Physics* 104.2 (2019), pp. 266–271. DOI: 10.1016/j.ijrobp.2019.01.080.

[29] K.T. Butterworth. “Evolution of the Supermodel: Progress in Modelling Radiotherapy Response in Mice”. In: *Clinical Oncology* 31.5 (2019), pp. 272–282. DOI: 10.1016/j.clon.2019.02.008.

[30] M Lupu-Plesu et al. “Effects of proton versus photon irradiation on (lymph)angiogenic, inflammatory, proliferative and anti-tumor immune responses in head and neck squamous cell carcinoma”. In: *Oncogenesis* 6.e354 (2017). DOI: 10.1038/oncsis.2017.56.

[31] O. Zlobinskaya et al. “The Effects of Ultra-High Dose Rate Proton Irradiation on Growth Delay in the Treatment of Human Tumor Xenografts in Nude Mice”. In: *Radiation Research* 181.2 (2014), pp. 177–183. DOI: 10.1667/RR13464.1.

[32] Waisse Waissi et al. “Radiosensitizing Pancreatic Cancer with PARP Inhibitor and Gemcitabine: An In Vivo and a Whole-Transcriptome Analysis after Proton or Photon Irradiation”. In: *Cancers* 13.527 (2021). DOI: 10.3390/cancers13030527.

[33] Cristina Müller et al. “Combination of Proton Therapy and Radionuclide Therapy in Mice: Preclinical Pilot Study at the Paul Scherrer Institute”. In: *Pharmaceutics* 11.9 (2019). DOI: 10.3390/pharmaceutics11090450.

[34] Bridget F Koontz, Frank Verhaegen, and Dirk De Ruysscher. “Tumour and normal tissue radiobiology in mouse models: how close are mice to mini-humans?” In: *British Journal of Radiology* 90.1069 (2016), p. 20160441. DOI: 10.1259/bjr.20160441.

[35] Valentina E. Schneeberger et al. “Quantitation of Murine Stroma and Selective Purification of the Human Tumor Component of Patient-Derived Xenografts for Genomic Analysis”. In: *PLOS ONE* 11.9 (2016), pp. 1–16. DOI: 10.1371/journal.pone.0160587.

[36] J. Kahn, P.J. Tofilon, and K Camphausen. “Preclinical models in radiation oncology”. In: *Radiation Oncology* 7.223 (2012). DOI: 10.1186/1748-717X-7-223.

[37] Cekanova M and Rathore K. “Animal models and therapeutic molecular targets of cancer: utility and limitations”. In: *Drug Design, Development and Therapy* 8 (2014), pp. 1911–1922. DOI: 10.2147/DDDT.S49584.

[38] Peter W. Nagle et al. “Patient-derived tumor organoids for prediction of cancer treatment response”. In: *Seminars in Cancer Biology* 53 (2018), pp. 258–264. DOI: 10.1016/j.semcancer.2018.06.005.

[39] Eric S. Diffenderfer et al. “The current status of preclinical proton FLASH radiation and future directions”. In: *Medical Physics* 49.3 (2022), pp. 2039–2054. DOI: 10.1002/mp.15276.

[40] M.C. Vozenin, J.H. Hendry, and C.L. Limoli. “Biological Benefits of Ultra-high Dose Rate FLASH Radiotherapy: Sleeping Beauty Awoken”. In: *Clinical Oncology* 31.7 (2019), pp. 407–415. DOI: 10.1016/j.clon.2019.04.001.

[41] Yolanda Prezado et al. “Spatially fractionated radiation therapy: a critical review on current status of clinical and preclinical studies and knowledge gaps”. In: *Physics in Medicine and Biology* 69.10 (2024), 10TR02. DOI: 10.1088/1361-6560/ad4192.

[42] Yolanda Prezado et al. “Proton minibeam radiation therapy spares normal rat brain: Long-Term Clinical, Radiological and Histopathological Analysis”. In: *Scientific Reports* 7.14403 (2017). DOI: 10.1038/s41598-017-14786-y.

[43] Vincent Favaudon et al. “Ultrahigh dose-rate FLASH irradiation increases the differential response between normal and tumor tissue in mice”. In: *Science Translational Medicine* 6.245 (2014), 245ra93–245ra93. DOI: 10.1126/scitranslmed.3008973.

[44] Karen Levy et al. “Abdominal FLASH irradiation reduces radiation-induced gastrointestinal toxicity for the treatment of ovarian cancer in mice”. In: *Scientific Reports* 10.21600 (2020). DOI: 10.1038/s41598-020-78017-7.

[45] Marie-Catherine Vozenin et al. “The Advantage of FLASH Radiotherapy Confirmed in Mini-pig and Cat-cancer Patients”. In: *Clinical Cancer Research* 25.1 (2019), pp. 35–42. DOI: 10.1158/1078-0432.CCR-17-3375.

[46] Carla Rohrer Bley et al. “Dose- and Volume-Limiting Late Toxicity of FLASH Radiotherapy in Cats with Squamous Cell Carcinoma of the Nasal Planum and in Mini Pigs”. In: *Clinical Cancer Research* 28.17 (2022), pp. 3814–3823. DOI: 10.1158/1078-0432.CCR-22-0262.

[47] John Wong et al. “High-Resolution, Small Animal Radiation Research Platform With X-Ray Tomographic Guidance Capabilities”. In: *International Journal of Radiation Oncology*Biology*Physics* 71.5 (2008), pp. 1591–1599. DOI: 10.1016/j.ijrobp.2008.04.025.

[48] R. Clarkson et al. “Characterization of image quality and image-guidance performance of a preclinical microirradiator”. In: *Medical Physics* 38.2 (2011), pp. 845–856. DOI: 10.1118/1.3533947.

[49] Dan M. Close et al. “In Vivo Bioluminescent Imaging (BLI): Noninvasive Visualization and Interrogation of Biological Processes in Living Animals”. In: *Sensors* 11.1 (2011), pp. 180–206. DOI: 10.3390/s110100180.

[50] Mihaela Ghita et al. “Integrating Small Animal Irradiators with Functional Imaging for Advanced Preclinical Radiotherapy Research”. In: *Cancers* 11.2 (2019). DOI: 10.3390/cancers11020170.

[51] Hu Zhou et al. “Development of a Micro-Computed Tomography–Based Image-Guided Conformal Radiotherapy System for Small Animals”. In: *International Journal of Radiation Oncology*Biology*Physics* 78.1 (2010), pp. 297–305. DOI: 10.1016/j.ijrobp.2009.11.008.

[52] Michael D. Jensen et al. “Implementation and commissioning of an integrated micro-CT/RT system with computerized independent jaw collimation”. In: *Medical Physics* 40.8 (2013), p. 081706. DOI: 10.1118/1.4812422.

[53] Manuela C. Felix et al. “Image-Guided Radiotherapy Using a Modified Industrial Micro-CT for Preclinical Applications”. In: *PLOS ONE* 10.5 (2015), pp. 1–11. DOI: 10.1371/journal.pone.0126246.

[54] Falk Tillner et al. “Precise image-guided irradiation of small animals: a flexible non-profit platform”. In: *Physics in Medicine and Biology* 61.8 (2016), p. 3084. DOI: 10.1088/0031-9155/61/8/3084.

[55] Sunil Sharma et al. “Advanced Small Animal Conformal Radiation Therapy Device”. In: *Technology in Cancer Research and Treatment* 16.1 (2017), pp. 45–56. DOI: 10.1177/1533034615626011.

[56] Hao Sha et al. “An image guided small animal stereotactic radiotherapy system”. In: *Oncotarget* 7.14 (2016), pp. 18825–18836. DOI: 10.18632/oncotarget.7939.

[57] Junwei Shi et al. “Optical molecular imaging-guided radiation therapy part 1: Integrated x-ray and bioluminescence tomography”. In: *Medical Physics* 44.9 (2017), pp. 4786–4794. DOI: 10.1002/mp.12415.

[58] Junwei Shi et al. “Optical molecular imaging-guided radiation therapy part 2: Integrated x-ray and fluorescence molecular tomography”. In: *Medical Physics* 44.9 (2017), pp. 4795–4803. DOI: 10.1002/mp.12414.

[59] E Ford et al. “An image-guided precision proton radiation platform for preclinical in vivo research”. In: *Physics in Medicine and Biology* 62.1 (2016), p. 43. DOI: 10.1088/1361-6560/62/1/43.

[60] Michele M Kim et al. “Design and commissioning of an image-guided small animal radiation platform and quality assurance protocol for integrated proton and x-ray radiobiology research”. In: *Physics in Medicine and Biology* 64.13 (2019), p. 135013. DOI: 10.1088/1361-6560/ab20d9.

[61] Katia Parodi et al. “Towards a novel small animal proton irradiation platform: the SIRMIO project”. In: *Acta Oncologica* 58.10 (2019), pp. 1470–1475. DOI: 10.1080/0284186X.2019.1630752.

[62] Moritz Schneider et al. “SAPPHIRE—establishment of small animal proton and photon image-guided radiation experiments”. In: *Physics in Medicine and Biology* 69.9 (2024), p. 095020. DOI: 10.1088/1361-6560/ad3887.

[63] A. Rousseti et al. “Preclinical proton minibeam radiotherapy facility for small animal irradiation”. English. In: *Proc. 14th International Particle Accelerator Conference* (Venice, Italy). IPAC’23 - 14th International Particle Accelerator Conference 14. JACoW Publishing, Geneva, Switzerland, May 2023, pp. 4522–4524. ISBN: 978-3-95450-231-8. DOI: 10.18429/JACoW-IPAC2023-THPL043.

[64] Marleen Balvert et al. “A framework for inverse planning of beam-on times for 3D small animal radiotherapy using interactive multi-objective optimisation”. In: *Physics in Medicine and Biology* 60.14 (2015), p. 5681. DOI: 10.1088/0031-9155/60/14/5681.

[65] Sara Gargiulo et al. “Mice Anesthesia, Analgesia, and Care, Part I: Anesthetic Considerations in Preclinical Research”. In: *ILAR Journal* 53.1 (2012), E55–E69. DOI: 10.1093/ilar.53.1.55.

[66] M.T. Evans, T. Wigmore, and L.J.S. Kelliher. “The impact of anaesthetic technique upon outcome in oncological surgery”. In: *BJA Education* 19.1 (2019), pp. 14–20. DOI: 10.1016/j.bjae.2018.09.008.

[67] Stefanie Rosenhain et al. “A preclinical micro-computed tomography database including 3D whole body organ segmentations”. In: *Scientific Data* 5.180294 (2018). DOI: 10.1038/sdata.2018.294.

[68] Juan Eugenio Iglesias and Mert R. Sabuncu. “Multi-atlas segmentation of biomedical images: A survey”. In: *Medical Image Analysis* 24.1 (2015), pp. 205–219. DOI: 10.1016/j.media.2015.06.012.

[69] Justin Malimban et al. “Deep learning-based segmentation of the thorax in mouse micro-CT scans”. In: *Scientific Reports* 12.1822 (2022). DOI: 10.1038/s41598-022-05868-7.

[70] Georgios Lappas et al. “Inter-observer variability of organ contouring for preclinical studies with cone beam Computed Tomography imaging”. In: *Physics and Imaging in Radiation Oncology* 21 (2022), pp. 11–17. DOI: 10.1016/j.phro.2022.01.002.

[71] Peter van Luijk et al. “The Impact of Heart Irradiation on Dose–Volume Effects in the Rat Lung”. In: *International Journal of Radiation Oncology*Biology*Physics* 69.2 (2007), pp. 552–559. DOI: 10.1016/j.ijrobp.2007.05.065.

[72] Oliver Schoppe et al. “Deep learning-enabled multi-organ segmentation in whole-body mouse scans”. In: *Nature Communications* 11.5626 (2020). DOI: 10.1038/s41467-020-19449-7.

[73] Dzung L. Pham, Chenyang Xu, and Jerry L. Prince. “Current Methods in Medical Image Segmentation”. In: *Annual Review of Biomedical Engineering* 2 (2000), pp. 315–337. DOI: 10.1146/annurev.bioeng.2.1.315.

[74] William.P. Segars et al. “Development of a 4-D digital mouse phantom for molecular imaging research”. In: *Molecular Imaging and Biology* 6.3 (2004), pp. 149–159. DOI: 10.1016/j.mibio.2004.03.002.

[75] Belma Dogdas et al. “Digimouse: a 3D whole body mouse atlas from CT and cryosection data”. In: *Physics in Medicine and Biology* 52.3 (2007), p. 577. DOI: 10.1088/0031-9155/52/3/003.

[76] Hongkai Wang, David B. Stout, and Arion F. Chatziioannou. “A Deformable Atlas of the Laboratory Mouse”. In: *Molecular Imaging and Biology* 17 (2015), pp. 18–28. DOI: 10.1007/s11307-014-0767-7.

[77] Sophie Lancelot et al. “A Multi-Atlas Based Method for Automated Anatomical Rat Brain MRI Segmentation and Extraction of PET Activity”. In: *PLOS ONE* 9.10 (2014), pp. 1–10. DOI: 10.1371/journal.pone.0109113.

[78] Brent van der Heyden et al. “Automatic multiatlas based organ at risk segmentation in mice”. In: *British Journal of Radiology* 92.1095 (2018), p. 20180364. DOI: 10.1259/bjr.20180364.

[79] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Ed. by Nassir Navab et al. Springer International Publishing, 2015, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28.

[80] Tim Lustberg et al. “Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer”. In: *Radiotherapy and Oncology* 126.2 (2018), pp. 312–317. DOI: 10.1016/j.radonc.2017.11.012.

[81] J. van der Veen et al. “Benefits of deep learning for delineation of organs at risk in head and neck cancer”. In: *Radiotherapy and Oncology* 138 (2019), pp. 68–74. DOI: 10.1016/j.radonc.2019.05.010.

[82] Georgios Lappas et al. “Automatic contouring of normal tissues with deep learning for preclinical radiation studies”. In: *Physics in Medicine and Biology* 67.4 (2022), p. 044001. DOI: 10.1088/1361-6560/ac4da3.

[83] Wouter R. P. H. van de Worp et al. “Deep Learning Based Automated Orthotopic Lung Tumor Segmentation in Whole-Body Mouse CT-Scans”. In: *Cancers* 13.18 (2021). DOI: 10.3390/cancers13184585.

[84] Brent van der Heyden et al. “Automated CT-derived skeletal muscle mass determination in lower hind limbs of mice using a 3D U-Net deep learning network”. In: *Journal of Applied Physiology* 128.1 (2020), pp. 42–49. DOI: 10.1152/japplphysiol.00465.2019.

[85] Frank Verhaegen and Joao Seco, eds. *Monte Carlo Techniques in Radiation Therapy: Applications to Dosimetry, Imaging, and Preclinical Radiotherapy (2nd ed.)* CRC Press, 2021. DOI: 10.1201/9781003212485.

[86] Frank Verhaegen et al. “A review of treatment planning for precision image-guided photon beam pre-clinical animal radiation studies”. In: *Zeitschrift für Medizinische Physik* 24.4 (2014). Schwerpunkt: Medizinphysik in der prä-klinischen Forschung, pp. 323–334. DOI: 10.1016/j.zemedi.2014.02.004.

[87] Linda Hong et al. “A pencil beam algorithm for proton dose calculations”. In: *Physics in Medicine and Biology* 41.8 (1996), p. 1305. DOI: 10.1088/0031-9155/41/8/005.

[88] Barbara Schaffner, Eros Pedroni, and Antony Lomax. “Dose calculation models for proton treatment planning using a dynamic beam delivery system: an attempt to include density heterogeneity effects in the analytical dose calculation”. In: *Physics in Medicine and Biology* 44.1 (1999), p. 27. DOI: 10.1088/0031-9155/44/1/004.

[89] Jatinder Saini et al. “Advanced proton beam dosimetry part I: review and performance evaluation of dose calculation algorithms”. In: *Translational Lung Cancer Research* 7.2 (2018). DOI: 10.21037/tlcr.2018.04.05.

[90] Kevin Souris, John Aldo Lee, and Edmond Sterpin. “Fast multipurpose Monte Carlo simulation for proton therapy using multi- and many-core CPU architectures”. In: *Medical Physics* 43.4 (2016), pp. 1700–1712. DOI: 10.1118/1.4943377.

[91] Xun Jia et al. “GPU-based fast Monte Carlo dose calculation for proton therapy”. In: *Physics in Medicine and Biology* 57.23 (2012), p. 7783. DOI: 10.1088/0031-9155/57/23/7783.

[92] Jan Gajewski et al. “Commissioning of GPU–Accelerated Monte Carlo Code FRED for Clinical Applications in Proton Therapy”. In: *Frontiers in Physics* 8 (2021). DOI: 10.3389/fphy.2020.567300.

[93] Hoyeon Lee et al. “MOQUI: an open-source GPU-based Monte Carlo code for proton dose calculation with efficient data structure”. In: *Physics in Medicine and Biology* 67.17 (2022), p. 174001. DOI: 10.1088/1361-6560/ac8716.

[94] Tiberiu Burlacu, Danny Lathouwers, and Zoltán Perkó. “A Deterministic Adjoint-Based Semi-Analytical Algorithm for Fast Response Change Computations in Proton Therapy”. In: *Journal of Computational and Theoretical Transport* 52.1 (2023), pp. 1–41. DOI: 10.1080/23324309.2023.2166077.

[95] Oscar Pastor-Serrano and Zoltán Perkó. “Millisecond speed deep learning based proton dose calculation with Monte Carlo accuracy”. In: *Physics in Medicine and Biology* 67.10 (2022), p. 105006. DOI: 10.1088/1361-6560/ac692e.

[96] Robert H W van Dijk et al. “A novel multichannel deep learning model for fast denoising of Monte Carlo dose calculations: preclinical applications”. In: *Physics in Medicine and Biology* 67.16 (2022), p. 164001. DOI: 10.1088/1361-6560/ac8390.

[97] Florian Mentzel et al. “Accurate and Fast Deep Learning Dose Prediction for a Preclinical Microbeam Radiation Therapy Study Using Low-Statistics Monte Carlo Simulations”. In: *Cancers* 15.7 (2023). DOI: 10.3390/cancers15072137.

[98] Marie Vanstalle et al. “Analytical dose modeling for preclinical proton irradiation of millimetric targets”. In: *Medical Physics* 45.1 (2018), pp. 470–478. DOI: 10.1002/mp.12696.

[99] Thomas Bortfeld and Wolfgang Schlegel. “An analytical approximation of depth - dose distributions for therapeutic proton beams”. In: *Physics in Medicine and Biology* 41.8 (1996), p. 1331. DOI: 10.1088/0031-9155/41/8/006.

[100] Monika Clausen et al. “Small field proton irradiation for in vivo studies: Potential and limitations when adapting clinical infrastructure”. In: *Zeitschrift für Medizinische Physik* 33.4 (2023), pp. 542–551. DOI: 10.1016/j.zemedi.2022.10.002.

[101] Harald Paganetti. “Range uncertainties in proton therapy and the role of Monte Carlo simulations”. In: *Physics in Medicine and Biology* 57.11 (2012), R99. DOI: 10.1088/0031-9155/57/11/R99.

[102] Uwe Schneider, Eros Pedroni, and Antony Lomax. “The calibration of CT Hounsfield units for radiotherapy treatment planning”. In: *Physics in Medicine and Biology* 41.1 (1996), p. 111. DOI: 10.1088/0031-9155/41/1/009.

[103] Nils Peters et al. “Consensus guide on CT-based prediction of stopping-power ratio using a Hounsfield look-up table for proton therapy”. In: *Radiotherapy and Oncology* 184 (2023), p. 109675. DOI: 10.1016/j.radonc.2023.109675.

[104] Ming Yang et al. “Comprehensive analysis of proton range uncertainties related to patient stopping-power-ratio estimation using the stoichiometric calibration”. In: *Physics in Medicine and Biology* 57.13 (2012), p. 4095. DOI: 10.1088/0031-9155/57/13/4095.

[105] Vicki T. Taasti et al. “Inter-centre variability of CT-based stopping-power prediction in particle therapy: Survey-based evaluation”. In: *Physics and Imaging in Radiation Oncology* 6 (2018), pp. 25–30. DOI: 10.1016/j.phro.2018.04.006.

[106] Nils Peters et al. “Experimental assessment of inter-centre variation in stopping-power and range prediction in particle therapy”. In: *Radiotherapy and Oncology* 163 (2021), pp. 7–13. DOI: 10.1016/j.radonc.2021.07.019.

[107] K B Niepel et al. “Animal tissue-based quantitative comparison of dual-energy CT to SPR conversion methods using high-resolution gel dosimetry”. In: *Physics in Medicine and Biology* 66.7 (2021), p. 075009. DOI: 10.1088/1361-6560/abbd14.

[108] Esther Bär et al. “Experimental validation of two dual-energy CT methods for proton therapy using heterogeneous tissue samples”. In: *Medical Physics* 45.1 (2018), pp. 48–59. DOI: 10.1002/mp.12666.

[109] Christian Möhler et al. “Experimental verification of stopping-power prediction from single- and dual-energy computed tomography in biological tissues”. In: *Physics in Medicine and Biology* 63.2 (2018), p. 025001. DOI: 10.1088/1361-6560/aaa1c9.

[110] M Yang et al. “Theoretical variance analysis of single- and dual-energy computed tomography methods for calculating proton stopping power ratios of biological tissues”. In: *Physics in Medicine and Biology* 55.5 (2010), p. 1343. DOI: 10.1088/0031-9155/55/5/006.

[111] Vicki T Taasti et al. “Validation of proton stopping power ratio estimation based on dual energy CT using fresh tissue samples”. In: *Physics in Medicine and Biology* 63.1 (2018), p. 015012. DOI: 10.1088/1361-6560/aa952f.

[112] Lotte E J R Schyns et al. “Optimizing dual energy cone beam CT protocols for preclinical imaging and radiation research”. In: *British Journal of Radiology* 90.1069 (2016), p. 20160480. DOI: 10.1259/bjr.20160480.

[113] Ana Vaniqui et al. “The impact of dual energy CT imaging on dose calculations for pre-clinical studies”. In: *Radiation Oncology* 12.181 (2017). DOI: 10.1186/s13014-017-0922-9.

[114] Jericho O’Connell and Magdalena Bazalova-Carter. “fastCAT: Fast cone beam CT (CBCT) simulation”. In: *Medical Physics* 48.8 (2021), pp. 4448–4458. DOI: 10.1002/mp.15007.

2

DEEP LEARNING-BASED SEGMENTATION OF THE THORAX IN MOUSE MICRO-CT SCANS

Justin Malimban[1], Danny Lathouwers[2], Haibin Qian[3], Frank Verhaegen[4], Julia Wiedemann[1,5], Sytze Brandenburg[1,7] and Marius Staring[6,7]

[1]Department of Radiation Oncology and Particle Therapy Research Center, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
[2]Department of Radiation Science and Technology, Faculty of Applied Sciences, Delft University of Technology, Delft, The Netherlands
[3]Department of Medical Biology, Amsterdam University Medical Centers (location AMC) and Cancer Center Amsterdam, Amsterdam, The Netherlands
[4]Department of Radiation Oncology (MAASTRO), Research Institute for Oncology & Reproduction, Maastricht University Medical Centre, Maastricht, The Netherlands
[5]Department of Biomedical Sciences of Cells & Systems-Section Molecular Cell Biology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
[6]Department of Radiology, Leiden University Medical Center, Leiden, The Netherlands
[7]These authors contributed equally to this work.

Published in Scientific Reports, 2022, 12:1822
Available at: https://doi.org/10.1038/s41598-022-05868-7

ABSTRACT

For image-guided small animal irradiations, the whole workflow of imaging, organ contouring, irradiation planning, and delivery is typically performed in a single session requiring continuous administration of anaesthetic agents. Automating contouring leads to a faster workflow, which limits exposure to anaesthesia and thereby, reducing its impact on experimental results and on animal wellbeing. Here, we trained the 2D and 3D U-Net architectures of no-new-Net (nnU-Net) for autocontouring of the thorax in mouse micro-CT images. We trained the models only on native CTs and evaluated their performance using an independent testing dataset (i.e., native CTs not included in the training and validation). Unlike previous studies, we also tested the model performance on an external dataset (i.e., contrast-enhanced CTs) to see how well they predict on CTs completely different from what they were trained on. We also assessed the interobserver variability using the generalized conformity index (CI_{gen}) among three observers, providing a stronger human baseline for evaluating automated contours than previous studies. Lastly, we showed the benefit on the contouring time compared to manual contouring. The results show that 3D models of nnU-Net achieve superior segmentation accuracy and are more robust to unseen data than 2D models. For all target organs, the mean surface distance (MSD) and the Hausdorff distance (95p HD) of the best performing model for this task (nnU-Net 3d_fullres) are within 0.16 mm and 0.60 mm, respectively. These values are below the minimum required contouring accuracy of 1 mm for small animal irradiations, and improve significantly upon state-of-the-art 2D U-Net-based AIMOS method. Moreover, the conformity indices of the 3d_fullres model also compare favourably to the interobserver variability for all target organs, whereas the 2D models perform poorly in this regard. Importantly, the 3d_fullres model offers 98% reduction in contouring time.

2.1. INTRODUCTION

Preclinical *in vivo* studies using small animal models serve as an essential experimental system to evaluate potential benefits and radiobiological implications of treatment strategies before clinical implementation. They play an integral role in modelling the disease, disease treatment, and response to treatment under clinically relevant radiation exposure conditions that can potentially translate to improvements in therapeutic outcomes. Over the years, extensive research has been done to develop small animal imaging and irradiation platforms for X-ray therapy [1, 2, 3, 4, 5]. Commercial irradiation units such as the Small Animal Radiation Research Platform (SARRP, Xstrahl Ltd., Camberley, UK) and X-RAD SmART+ (PXI North Branford, CT, USA) are also available, providing image-guided irradiations representative of clinical scenarios [1, 6]. Recently, research groups have also started to adopt these technologies for proton preclinical research by integrating them with a proton beamline to perform image-guided proton irradiations [7, 8, 9].

The preclinical irradiation workflow involves the following stages: (1) animal setup, (2) image acquisition, (3) organ contouring, (4) irradiation planning, and (5) radiation delivery. The process begins with administration of anaesthesia to immobilize the animal and is followed by placement in the irradiation position. Then, 3D volumetric scans of the animal are acquired using micro computed tomography (micro-CT) or other imaging modalities. These images are used to identify the shape and location of target volumes and delineate their boundaries. Then, an irradiation plan is created, and dose distributions are calculated. Once the irradiation objectives and dose constraints are met, the plan is delivered to the animal. For image-guided small animal irradiation, this entire process is preferably carried out consecutively in a single treatment session, which typically lasts for 20–90 min, during which the animal is continuously maintained under anaesthesia [10]. However, prolonged exposure of rodents to anaesthetic agents has been shown to influence physiological parameters which can potentially affect the outcome of experiments [11, 12]. Therefore, a fast irradiation workflow is warranted.

One of the most time-consuming tasks in preclinical image-guided irradiation workflow is organ contouring. Traditionally, the organ contours are created manually by a biologist. This approach is not ideal and can be very tedious since in a single study a large group of animals may be irradiated. As an example, in studies of normal tissue damage well over 100 animals were irradiated in a single session [13]. In this study, organ contours of only five animals were made upon which irradiation plans for the entire population were based. Automating organ contouring not only reduces the overall workload for preclinical irradiations, but also allows plans to be created tailored to individual

animals. This can lead to a better predictive value of preclinical studies and in effect, the number of animals required to meet the objectives of the study may also be reduced. Moreover, individualized contours are essential for animals implanted with orthotopic tumours, which exhibit greater morphological variation compared to the normal tissues.

Over the years, several methods have been developed to automate and speed up the contouring process. One of the most widely used autocontouring techniques for biomedical applications is the atlas method [14]. Several whole-body atlases of mouse anatomy have been constructed, such as the MOBY phantom [15] and the Digimouse atlas [16], which were based on a single reference animal. However, some studies have pointed out that this approach (“classic single atlas") produces inferior segmentation accuracy as it exhibits strong bias towards the selected atlas, and it cannot capture realistic body deformations caused by posture, weight, fat amount, and body length variations [17]. To address this problem, deformable atlases, which can adapt arbitrary poses and adjust organ anatomy based on changes in body weight, length, and fat amount, have been proposed [18, 19]. Another potential solution to compensate for individual variations is the use of multiple atlases constructed from different subjects. The multi-atlas-based image segmentation (MABIS) algorithm developed by van der Heyden et al. [20] was able to complete the contouring process in a relatively short time (~12 mins) and generated accurate segmentations for organs with sharp boundaries, but manual corrections were needed for less sharp ones.

Although atlas-based segmentation methods are generally faster than manual contouring, the effective runtime of the segmentation task is still considerable, and it may be further reduced using deep learning techniques. In particular, convolutional neural networks (CNN) have shown encouraging results in human organ segmentation. Several studies have demonstrated that deep learning-based segmentation yielded more consistent and more accurate results than atlas-based methods for clinical images [21, 22]. It also outperformed the atlas-based methods in terms of speed [23]. CNNs have also found applications in preclinical image segmentation. Van der Heyden et al. [24] used a two-step 3D U-Net model to automatically delineate the skeletal muscle in the lower limb of mice, which was shown to be 150 times faster than manual segmentation. For multi-organ segmentation, Wang et al. [25] developed a 3D two-stage deeply supervised network (TS-DSN) for delineation of major organs in the torso of a mouse with an inference time of less than 2 s. More recently, Schoppe et al. [26] developed a deep learning pipeline based on a 2D U-Net-like network called AIMOS (AI-based Mouse Organ Segmentation), which achieved an inference time of 830 ms. Both models showed superior segmentation accuracy compared to existing studies on atlas-based methods. Moreover,

AIMOS outperformed TS-DSN except for heart segmentation.

In this work, we trained and validated the 2D and 3D U-Net architectures of no-new-Net (nnU-Net) for segmentation of organs in the mouse thorax and compared their performance to the state-of-the-art AIMOS method. We used only native CT scans for the training and validation phase, and we evaluated the trained models' accuracy using an independent testing dataset (i.e. native CTs not included in the training and validation). Unlike previous works, we also tested the trained models against an external dataset (i.e. contrast-enhanced CTs), which does not share the same properties such as the mouse strain and image acquisition parameters as the training data. The external dataset was used to investigate the robustness of the neural networks to datasets that are completely different from what they were trained on. Moreover, we thoroughly compared the accuracy of the automated contours relative to human performance by evaluating the generalized conformity index among three observers. Lastly, we assessed by how much these neural networks can shorten the contouring time compared to manual contouring in order to improve the efficiency of the irradiation workflow. For this segmentation task, we used a publicly available mouse micro-CT dataset [27], and we provide new annotations by two observers for the entire native CT and a subset of the contrast-enhanced CT datasets. These include spinal cord and separate left and right lung segmentations not provided in the original annotations. We make these annotations publicly available at https://doi.org/10.5281/zenodo.5121272.

2.2. METHODS

2.2.1. DATASET

The micro-CT images used in this work were taken from a public database which includes native and contrast-enhanced 3D whole body scans of mice [27]. Supplementary Table S2.1 provides a summary of their properties. The native CT dataset is comprised of 140 images from 20 female BALB/c nu/nu mice, with each animal imaged at seven time points spread over a 72-h period. The entire native CT dataset was utilized, wherein 105 images were allotted to train and validate the models, while the remaining 35 were used as the test set. To create completely independent training and testing datasets, the native CT scans were divided at the animal level: CT images of 15 animals (105 scans) were used for training and validation, and 5 animals (35 scans) were used for testing. In addition, thirty-five scans from the contrast-enhanced CT (CECT) dataset were taken to serve as a second independent test set to further evaluate the trained model's generalizability and robustness. This dataset includes CTs of 10 female BALB/cAnNRj-Foxn1nu mice, which

were also imaged at various time points over a 240-h period. However, only eight animals were considered as two of them did not appear to have contrast enhancement.

In this study, we focused on organs in the thoracic region: heart, spinal cord, right lung and left lung. Both test sets were annotated by three observers. The first observer and two second observers were all trained to follow the same labelling protocol and were supervised by a biologist with more than 5 years of animal contouring experience. All of them used the contouring module of the small animal radiotherapy treatment planning system, SmART-ATP (version 2.0, SmART Scientific Solutions BV, Maastricht, the Netherlands). Delineations by a third observer were taken from the annotations provided together with the CT images [27]. These were resampled to the same voxel resolution of $0.14 \times 0.14 \times 0.14$ mm^3 using nearest neighbor interpolation.

2.2.2. DEEP LEARNING MODELS

The U-Net is one of the most popular architectures for image segmentation. It is a fully convolutional network (FCN) that has a U-shape, with symmetric encoder (contraction) and decoder (expansion) paths. The encoder performs a series of convolution and pooling operations to extract feature representations from the image that the decoder aims to project onto the pixel space through up-sampling in order to restore the original image size. The U-Net was initially proposed for 2D biomedical image segmentation and has been shown to work well even with small training datasets [28]. This is advantageous for preclinical studies where there are restrictions on the number of animals that can be imaged to build the training data. In this work, we investigated the no-new-Net (nnU-Net) deep learning pipeline [29], which offers 2D and 3D U-Net-like architectures, and compared its performance to the 2D U-Net-based AIMOS method [26].

NO-NEW-NET (NNU-NET)

The no-new-Net is an out-of-the-box tool for automated image segmentation and has been widely used for clinical data. It is a self-adapting algorithm that follows certain heuristic rules to decide on the training configuration such as the selection of the batch size, patch size, and network topology depending on the dataset provided by the user [29]. It is a fully automated deep learning pipeline, which offers both 2D and 3D U-Net architectures that closely follow the original U-Net design.

In this work, we trained all the available models in nnU-Net from scratch: 2D U-Net (2d), 3D full resolution U-Net (3d_fullres), 3D low resolution U-Net (3d_lowres), and 3D cascade U-Net (3d_cascade). The network architectures for the 2D and 3D models generated by nnU-Net for this dataset are illustrated in Supplementary Fig. S2.1. nnU-Net generates a 2D U-Net model with a network depth (i.e., number of encoder-decoder

levels) of six. It is configured to accept a patch size of 320 × 224 as input and starts with 32 initial feature channels at the highest layer. The input is downsampled six times in the x and five times in the y direction, resulting in an image size of 5 × 7 at the bottleneck with 480 feature channels. The 2D U-Net model only operates on transverse slices and implements a batch size of 44 during training. nnU-Net also offers three different 3D models with a network depth of five. The 3D full resolution U-Net model runs on the full resolution data and has been shown to be the best performing configuration among all the nnU-Net models in the segmentation challenges where they have participated [29]. It also starts with the same number of initial feature channels but with a patch size of 128 × 96 × 192 and a batch size of 2. Downsampling is performed five times in x and z and four times in y, which reduces the feature maps at the bottleneck to 4 × 6 × 6 with feature channels capped to 320. The 3D low resolution U-Net and 3D cascade U-Net also follow this configuration. However, the 3D cascade U-Net is trained in two stages. The first stage involves training a 3D U-Net on downsampled versions of the training images (3d_lowres). The 3d_lowres model was trained on patches of the dataset at a resolution of 0.19 × 0.19 × 0.19 mm^3, and the resulting segmentations are then upsampled to the original voxel spacing of 0.14 × 0.14 × 0.14 mm^3. These segmentations served as the input for the second stage, and training is performed at full resolution. All five folds of the 3d_lowres model must be completed before the second stage of 3D cascade U-Net can be initiated.

AI-BASED MOUSE ORGAN SEGMENTATION (AIMOS)

Recently, Schoppe et al. [26] developed a fully-automated deep learning pipeline dedicated for organ contouring of mice micro-CT images called AIMOS. It is currently the overall best performing algorithm for mouse segmentation. AIMOS provides pre-processing, network training, and post-processing modules, requiring very little intervention from the user. It offers several 2D U-Net-like architectures that only differ in the number of encoder-decoder stages. For this study, the default architecture, UNet-768, was chosen, which employs six encoder-decoder stages with initial 32 feature channels at the highest layer and 768 feature channels at the bottleneck. The network was trained using all slices with a batch size of 32.

2.2.3. NETWORK TRAINING AND INFERENCE

All neural networks were trained only on native CTs delineated by observer 1. Five-fold cross-validation was performed wherein at each fold, three animals were randomly selected and set aside for validation, while the rest was used for training. The same split configuration was used for all networks. The final predictions were determined through

an ensemble voting by taking the average of the predicted probabilities from the five models resulting from training on the individual folds. All experiments for nnU-Net were carried out using an NVIDIA V100 with 12 GB of GPU memory while AIMOS was trained using an NVIDIA Quadro RTX 6000 with 24 GB of GPU memory. The training of the nnU-Net 3d_fullres model with 1000 epochs took approximately 2 days on our computing system. The inference time for both codes was evaluated on the same system (NVIDIA Quadro RTX 6000) to facilitate comparison. For this, we chose to report the average time the models take to preprocess an image, to make an inference and the total runtime. The runtime was measured starting from data preparation up to exportation of the automated contours.

2.2.4. EVALUATION METRICS AND STATISTICAL ANALYSIS

The quality of the segmentations generated by each model was evaluated in terms of the Dice similarity coefficient (DSC), mean surface distance (MSD) and 95th percentile Hausdorff distance (95p HD) [30, 31]. The DSC measures the degree of overlap between the reference and predicted contours; it increases with overlap and a value of 1 indicates a perfect overlap. MSD and 95p HD give the average and maximum distance measured between closest points on the surface of the contours, respectively. Therefore, smaller values for MSD and 95p HD indicate better correspondence to the ground truth.

To determine whether the difference in the DSC, MSD and 95p HD between the models is significant, a statistical analysis was conducted using a two-tailed Wilcoxon signed rank test with a significance level of $\alpha = 0.05$. The nnU-Net 3d_fullres model was chosen as the base model for comparison because it has been shown to be one of the best performing models in many medical image segmentation tasks [29]. A p-value < 0.05 is considered statistically significant.

2.2.5. INTEROBSERVER VARIABILITY (IOV)

The degree of agreement between observer's delineations was estimated using the generalized conformity index (CI_{gen}). It is defined as the ratio of the sum of the intersecting volumes between all pairs of observers and the sum of union of volumes between the same pairs [32]. The CI_{gen} is the general form of the Jaccard coefficient [30] applicable for comparison of more than two delineated volumes. This reduces to the Jaccard coefficient for the two-observer case. Higher values of CI_{gen} indicate greater similarity between the volumes.

Since the annotations from observer 3 do not include the spinal cord, delineations from only two observers were considered for this organ. There is also no separation of

the left and right lungs for observer 3 so the lungs were combined to form the total lung volume for the other two observers to facilitate comparison. The IOV was then compared to the performance of the models against a consensus segmentation among observers for which we will refer to as the reference contour for the rest of the paper. For the spinal cord, only pixels delineated by both observers were included in the consensus, whereas pixels delineated by 2 out of 3 observers were considered for the heart and lungs.

2.3. RESULTS

2.3.1. NATIVE CT (TEST SET 1)

Figure 2.1 shows the comparison between the automated and manual contours of observer 1 (i.e., observer who annotated the training data) for an example from test set 1. In general, all neural networks showed correct segmentations for the target organs. The boundaries of the predicted contours appear somewhat smoother than the ground truth. Both the AIMOS and nnU-Net 2d models showed cases wherein parts of the left lung were mislabelled as the right lung or vice versa. The mislabelled pixels can be easily corrected manually as they usually occur in clusters and do not exceed 2% of the total organ volume. Moreover, the nnU-Net 2d model misclassified pixels far from the thorax, which are mostly air cavities labelled either as right lung or left lung. For most cases, only a few pixels were mislabelled by this model, but 2 out of 35 samples exhibited numerous incorrectly labelled pixels.

The boxplots of the DSC, MSD and 95p HD for the heart, spinal cord, right lung and left lung when compared against the contours of observer 1 are shown in Supplementary Fig. S2.2, and the mean, median, and standard deviation of each metric are given in Table 2.1. The right and left lungs recorded the highest mean DSC at 0.97 ± 0.01 followed by the heart and spinal cord at 0.95 ± 0.01 and 0.91 ± 0.02, respectively. Using a two-tailed Wilcoxon signed rank test with nnU-Net 3d_fullres as the baseline model for comparison, it was found that the DSC, MSD and 95p HD of nnU-Net 3d_cascade were not significantly different from nnU-Net 3d_fullres for all organs considered. Meanwhile, significant differences ($p < 0.05$) on the DSC and MSD values were observed for the other models. They showed slightly inferior performance on these metrics compared to the nnU-Net 3d_fullres model. Nevertheless, all models achieved a mean MSD less than the in-plane voxel size of 0.14 mm, while the mean 95p HD were all below 0.60 mm for all organs except the right lung segmentation of nnU-Net 2d. Larger surface distances were observed for this model, resulting from false classifications far from the thoracic region (anaesthesia nozzle). Performing connected component analysis reduced the mean MSD and 95p HD to values similar to what was observed for other structures.

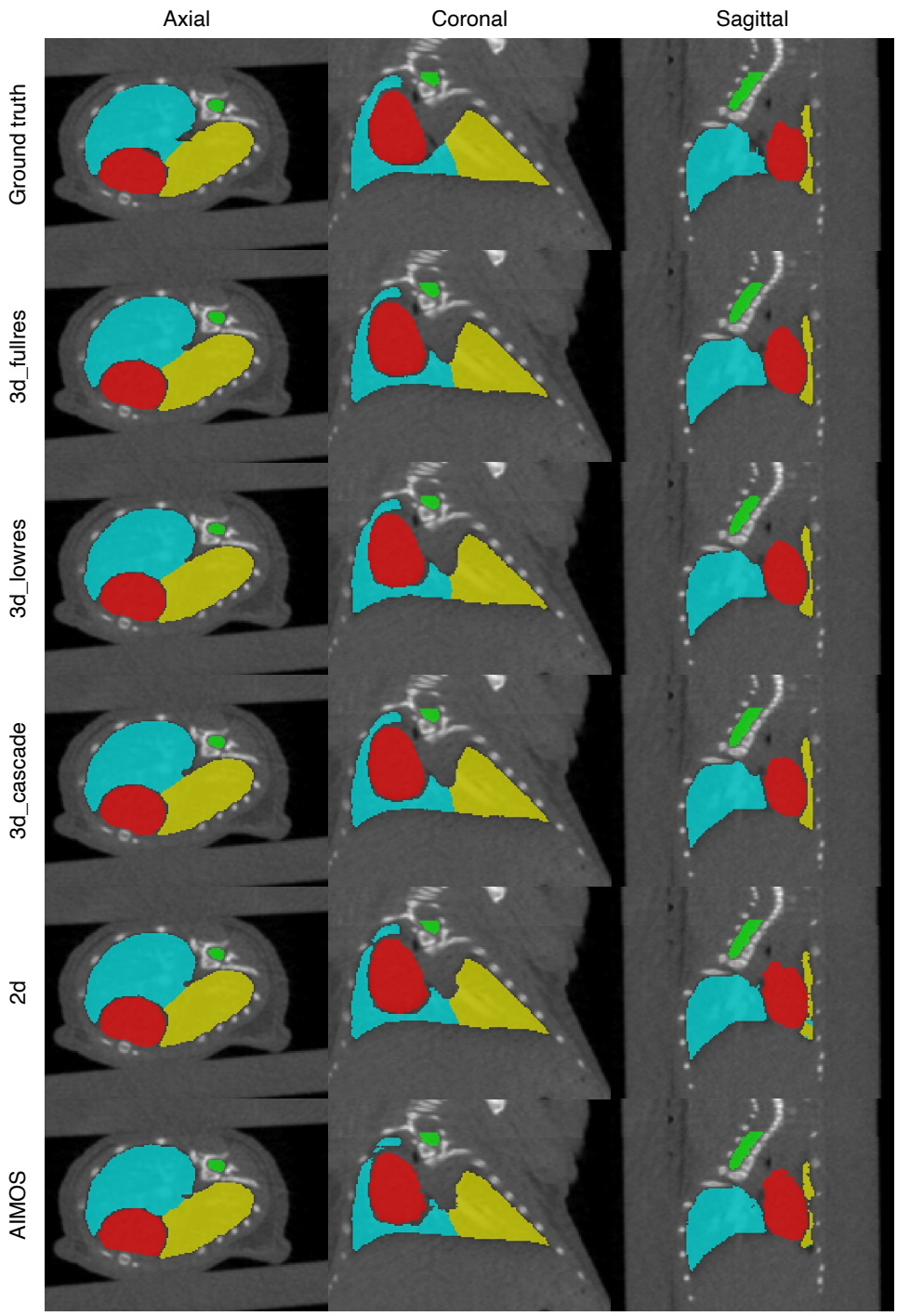


Figure 2.1: An example segmentation in the axial, coronal and sagittal views for test set 1. The first row shows the manual contours of observer 1, while the succeeding rows are the automated contours generated by each model. Contours in red, green, blue and yellow correspond to the heart, spinal cord, right lung and left lung, respectively.

Table 2.1: Comparison of the segmentation accuracy between different models in terms of the DSC, MSD and 95p HD (mean ± SD and median) for test set 1 annotated by observer 1. The asterisk (*) indicates a significant difference with the nnU-Net 3d full resolution model according to the Wilcoxon signed rank test with a significance level of $\alpha = 0.05$. Values in parentheses correspond to re-calculated metrics after connected component analysis.

Organ	Algorithm	Model	DSC		MSD (mm)		95p HD (mm)	
			Mean ± SD	Median	Mean ± SD	Median	Mean ± SD	Median
Heart	nnU-Net	3d_fullres	0.95 ± 0.01	0.950	0.08 ± 0.02	0.088	0.30 ± 0.07	0.28
		3d_lowres	0.95 ± 0.01	0.949*	0.09 ± 0.02	0.090*	0.30 ± 0.07	0.28
		3d_cascade	0.95 ± 0.01	0.950	0.09 ± 0.02	0.089	0.31 ± 0.07	0.28
		2d	0.94 ± 0.01	0.945*	0.10 ± 0.01	0.096*	0.33 ± 0.06	0.31*
	AIMOS	UNet-768	0.94 ± 0.01	0.945*	0.10 ± 0.02	0.094*	0.33 ± 0.07	0.31*
Spinal Cord	nnU-Net	3d_fullres	0.91 ± 0.02	0.913	0.04 ± 0.01	0.034	0.28 ± 0.02	0.28
		3d_lowres	0.90 ± 0.02	0.898*	0.04 ± 0.01	0.041*	0.29 ± 0.03	0.28
		3d_cascade	0.91 ± 0.01	0.912	0.03 ± 0.01	0.034	0.28 ± 0.02	0.28
		2d	0.91 ± 0.02	0.907*	0.04 ± 0.01	0.036	0.28 ± 0.02	0.28
	AIMOS	UNet-768	0.90 ± 0.02	0.909*	0.04 ± 0.01	0.036*	0.28 ± 0.02	0.28
Right Lung	nnU-Net	3d_fullres	0.97 ± 0.01	0.970	0.04 ± 0.01	0.036	0.42 ± 0.00	0.42
		3d_lowres	0.97 ± 0.01	0.967*	0.04 ± 0.01	0.040*	0.42 ± 0.00	0.42
		3d_cascade	0.97 ± 0.01	0.970	0.04 ± 0.01	0.035	0.42 ± 0.00	0.42
		2d	0.97 ± 0.01 (0.97 ± 0.01)	0.966*	0.60 ± 0.84 (0.04 ± 0.01)	0.112*	5.29 ± 19.8 (0.42 ± 0.00)	0.42
	AIMOS	UNet-768	0.96 ± 0.01	0.963*	0.05 ± 0.01	0.044*	0.42 ± 0.00	0.42
Left Lung	nnU-Net	3d_fullres	0.97 ± 0.01	0.966	0.04 ± 0.01	0.035	0.53 ± 0.13	0.56
		3d_lowres	0.96 ± 0.01	0.962*	0.04 ± 0.01	0.042*	0.56 ± 0.00	0.56
		3d_cascade	0.97 ± 0.01	0.966	0.04 ± 0.01	0.035	0.53 ± 0.13	0.56
		2d	0.96 ± 0.01	0.965*	0.04 ± 0.02	0.037*	0.54 ± 0.09	0.56
	AIMOS	UNet-768	0.95 ± 0.01	0.956*	0.05 ± 0.01	0.045*	0.56 ± 0.00	0.56

2.3.2. CONTRAST-ENHANCED CT (TEST SET 2)

Since the overall best performing models for the native CT dataset are the 3d_fullres and 3d_cascade models, and no significant difference was observed between them, the 3d_fullres model was chosen as the representative 3D model to evaluate the contrast-enhanced CT dataset. The results were compared to both nnU-Net 2d and AIMOS. Figure 2.2 shows a visual comparison of the manual contours of observer 1 and the automated contours generated by networks trained on the native CT data. The segmentation performance of the models versus observer 1 in terms of the DSC, MSD and 95p HD are summarized in Table 2.2, and the corresponding boxplots are shown in Supplementary Fig. S2.3. For all structures, the nnU-Net 3d_fullres model produced accurate segmentations albeit with a drop in performance as compared to the native CTs. Except for the spinal cord, the 3d_fullres model achieved a mean DSC > 0.90. The mean MSDs were also smaller than the in-plane voxel size of 0.14 mm except for the heart, while all organs had a mean 95p HD below 0.60 mm.

Consistently, the AIMOS and nnU-Net 2d models exhibited greater variations in the DSC, MSD and 95p HD compared to the 3d_fullres model. Both 2D models failed to generate predictions for the heart and left lung in 1 out of 35 samples. This particular case was excluded in the calculation of the performance metrics presented in Table 2.2. On most of the samples, the 2D models had difficulties in segmenting the heart, right lung and left lung. Several slices were partially or completely unlabelled for these organs and in some cases, half of the volume had no prediction at all. The lower DSC and larger MSD and 95p HD clearly indicate that the 2D models underperformed on this dataset. Notably, the 2D models scored a mean 95p HD > 1 mm for all target organs except for the spinal cord segmentation of AIMOS. This can be attributed to mislabelling pixels in the liver as heart, while pixels associated to air cavities outside the thoracic region such as air pockets in the abdomen were mislabelled as part of the lungs. Segments of the spinal cord were also missing in some cases, but it happened less frequently. Both 2D models also failed to distinguish the left and right lungs as shown in Fig. 2.3, which did not occur for the 3D model.

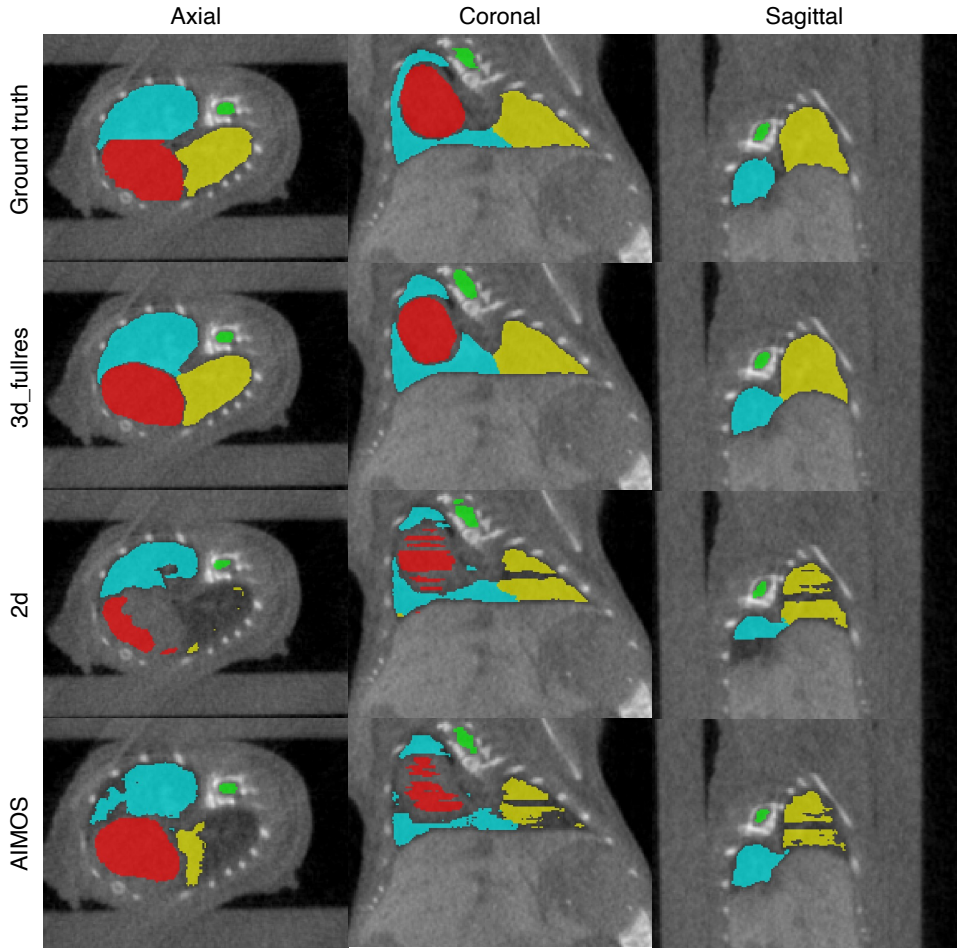


Figure 2.2: An example segmentation in the axial, coronal, and sagittal views for test set 2. The first row shows the manual contours of observer 1, while the succeeding rows are the automated contours generated by each model. Contours in red, green, blue and yellow correspond to the heart, spinal cord, right lung and left lung, respectively.

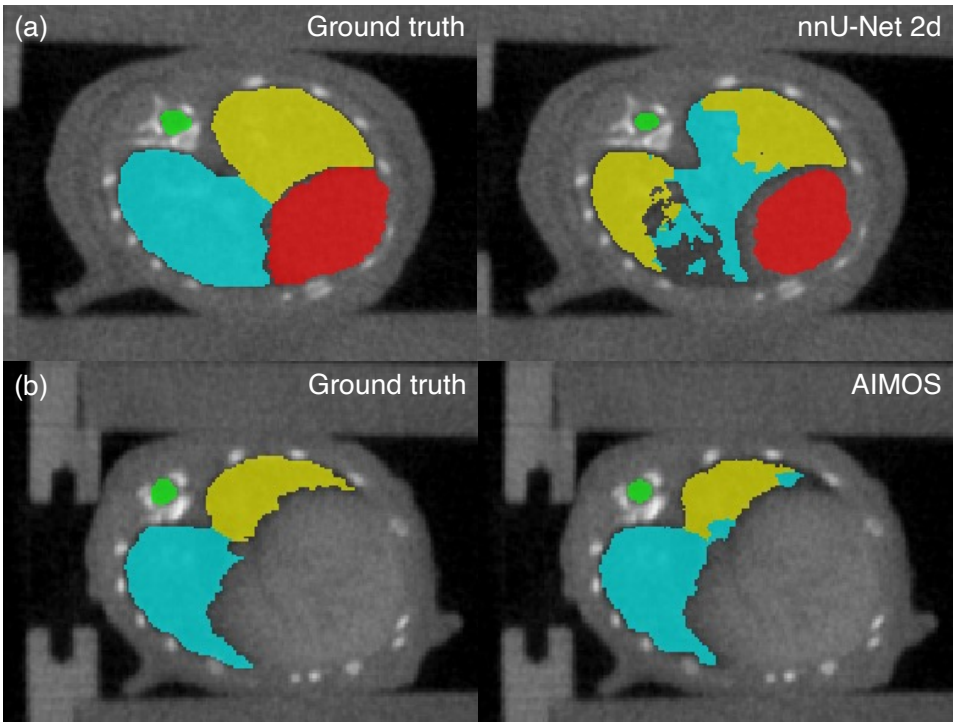


Figure 2.3: Predictions of (a) nnU-Net 2d and (b) AIMOS on contrast-enhanced CTs showing misclassification of the right and left lungs. Corresponding ground truths are given on the left. Contours in red, green, blue and yellow correspond to the heart, spinal cord, right lung and left lung, respectively.

Table 2.2: Comparison of the the segmentation accuracy between different models in terms of the DSC, MSD and 95p HD (mean ± SD and median) for test set 2 annotated by observer 1. The asterisk (*) indicates a significant difference with the nnU-Net 3d full resolution model according to the Wilcoxon signed rank test with a significance level of $\alpha = 0.05$.

Organ	Algorithm	Model	DSC		MSD (mm)		95p HD (mm)	
			Mean ± SD	Median	Mean ± SD	Median	Mean ± SD	Median
Heart	nnU-Net	3d_fullres	0.92 ± 0.02	0.92	0.16 ± 0.04	0.16	0.55 ± 0.18	0.54
		2d	0.81 ± 0.20	0.89*	0.36 ± 0.28	0.28*	1.48 ± 0.93	1.41*
	AIMOS	UNet-768	0.83 ± 0.17	0.89*	0.34 ± 0.24	0.27*	1.50 ± 0.86	1.29*
Spinal Cord	nnU-Net	3d_fullres	0.85 ± 0.03	0.85	0.07 ± 0.04	0.07	0.39 ± 0.11	0.40
		2d	0.76 ± 0.14	0.81*	0.19 ± 0.42	0.10*	1.29 ± 2.24	0.59*
	AIMOS	UNet-768	0.82 ± 0.05	0.83*	0.08 ± 0.02	0.08*	0.45 ± 0.13	0.40*
Right Lung	nnU-Net	3d_fullres	0.96 ± 0.01	0.96	0.06 ± 0.02	0.06	0.45 ± 0.06	0.42
		2d	0.82 ± 0.14	0.86*	0.55 ± 0.39	0.45*	6.14 ± 5.31	3.63*
	AIMOS	UNet-768	0.87 ± 0.16	0.92*	0.21 ± 0.23	0.15*	1.11 ± 1.24	0.60*
Left Lung	nnU-Net	3d_fullres	0.95 ± 0.02	0.95	0.06 ± 0.03	0.06	0.59 ± 0.17	0.56
		2d	0.77 ± 0.18	0.81*	0.80 ± 0.64	0.65*	9.54 ± 9.88	3.52*
	AIMOS	UNet-768	0.77 ± 0.22	0.87*	0.42 ± 0.89	0.17*	2.80 ± 5.29	1.12*

2.3.3. BEST, INTERMEDIATE, AND WORST SEGMENTATIONS

The best, intermediate, and worst segmentation results of the nnU-Net 3d_fullres model for both datasets are shown in Fig. 2.4. The samples were chosen based on the average DSC of the organs. For the native CT dataset, all three contours showed good agreement with the ground truth. For the contrast-enhanced CT dataset, the quality of the best and intermediate results are similar to the native CT. However, the left lung is undersegmented for the worst case. The model had difficulty annotating the lungs because the contrast with soft tissue is not as good as the native CTs. This particular case is actually an extreme outlier and majority of the automated contours for the contrast-enhaced CT dataset did not exhibit such errors. Unlike the native CT dataset, the worst case for the contrast-enhanced CTs required minor manual corrections. Nevertheless, the contours generated by the nnU-Net 3d_fullres model for this case are much better than AIMOS. In fact, AIMOS achieved DSC values of 0.15 (heart), 0.57 (spinal cord), 0.08 (right lung), and 0.001 (left lung), whereas nnU-Net 3d_fullres obtained DSC values of 0.88 (heart), 0.79 (spinal cord), 0.93 (right lung), and 0.82 (left lung).

It can be seen that the proposed model handles organ edges better than humans, particularly for the heart and lungs. The ragged edge details in the ground truth are partly due to the fact that the contours were created in the coronal plane. That choice was made as the organs are more visible and easier to distinguish in that plane. Unfortunately, the software does not allow editing of the contours on planes other than the one initially used to create them.

2.3.4. INTEROBSERVER VARIABILITY (IOV)

Table 2.3 gives the mean CI_{gen} of the heart, total lungs and spinal cord computed between the automated and reference contours for both test sets and Supplementary Fig. S2.4 shows the corresponding boxplots. The performance of nnU-Net 3d_fullres, 2d, and AIMOS was evaluated against the reference established based on the interobserver variability. For the native CT dataset, all models showed comparable results and obtained higher conformity indices than the human observers for all target organs. For the contrast-enhanced CT dataset, only the 3d_fullres model achieved greater conformity than the IOV on all organs, whereas AIMOS indicated better conformity than the human baseline only for the spinal cord delineation. For both datasets, the CI_{gen} is notably lower for the spinal cord, which is due to manual delineation variation in the superior and inferior extent. The small cross-section of a mouse spinal cord also made this structure difficult to delineate, resulting in disagreements along the organ boundary.

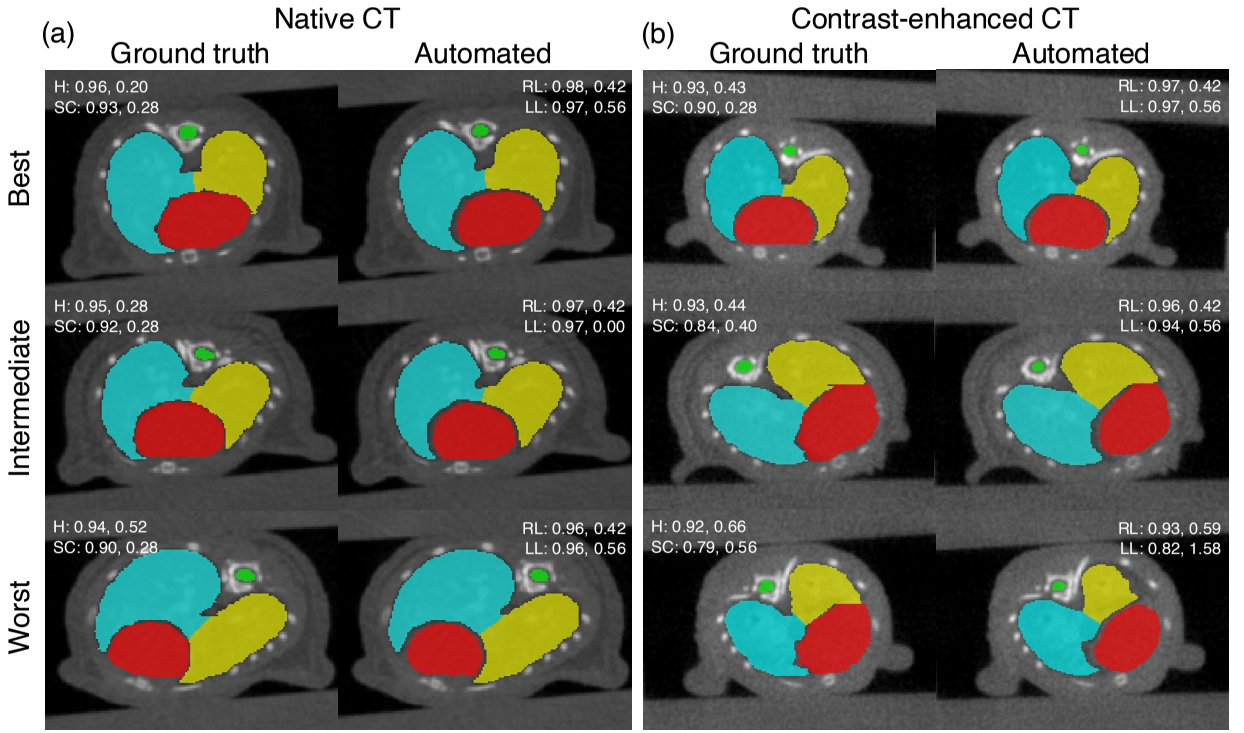


Figure 2.4: Examples of the best (first row), intermediate (second row), and worst (third row) segmentation results for the (a) native CT and (b) contrast-enhanced CT datasets obtained by the nnU-Net 3d_fullres model compared to the ground truth. The DSC scores (first value) and 95p HD in mm (second value) for each organ are also given. H, SC, RL, and LL correspond to the heart (red), spinal cord (green), right lung (blue) and left lung (yellow), respectively.

Table 2.3: Performance of the models against the reference contour in terms of the generalized conformity index (CI_{gen}) (mean ± SD) for the native and contrast-enhanced CT datasets. The model with the best results are shown in bold.

Algorithm	Models	Native CT (test set 1)			Contrast-enhanced CT (test set 2)		
		Heart	Total lungs	Spinal Cord	Heart	Total lungs	Spinal Cord
nnU-Net	3d_fullres	**90% ± 1%**	**95% ± 1%**	65% ± 6%	**87% ± 3%**	**92% ± 2%**	**75% ± 7%**
	2d	89% ± 2%	94% ± 1%	**66% ± 6%**	72% ± 22%	81% ± 16%	65% ± 14%
AIMOS	UNet-768	89% ± 2%	93% ± 1%	64% ± 6%	74% ± 18%	76% ± 18%	73% ± 8%
Interobserver variability		79% ± 3%	86% ± 2%	60% ± 6%	80% ± 3%	81% ± 8%	69% ± 6%

Table 2.4: Comparison of the average preprocessing and inference times, and total runtimes in seconds.

Algorithm	Model	Preprocessing	Inference	Runtime
nnU-Net	3d_fullres	5	27	40
	3d_lowres	14	13	50
	3d_cascade (2nd stage only)	17	28	52
	2d	5	7	21
AIMOS	UNet-768	6	2	20

2.3.5. CONTOURING TIME

Table 2.4 shows the preprocessing and inference time per scan and the total runtime for each model. The variation in the values presented is in the order of a few seconds. As expected, the 2D models had the best inference speeds and total runtimes. Although AIMOS is faster than nnU-Net 2d at inference, both models take similar amount of time to generate the contours. Among the 3D models, the 3d_lowres model had the shortest inference time, and it is faster by a factor of two than the 3d_fullres model. Due to the additional step of downsampling the images in preprocessing, the 3d_lowres model's runtime is longer than the 3d_fullres model. The 3d_cascade model is the slowest as it executes the 3d_lowres model first and uses its prediction as the input for inference at full resolution in the second stage. For comparison, a trained biologist from our institute takes roughly 40 min to create the manual contours per animal whereas all models took less than 1 min.

2.4. DISCUSSION

Typically, image-guided preclinical irradiations require that animals are imaged shortly before irradiation. This entails that the irradiation workflow must be executed in the shortest time possible as animals are continuously exposed to anaesthesia throughout the entire process. One aspect in which time can be effectively reduced is organ contouring. To date, among state-of-the-art methods for autocontouring of mouse organs, deep learning-based algorithms show superior results and outperform atlas-based segmentation techniques [25, 26]. In this study, we further explore deep learning models, in particular 3D U-Net-like neural networks, and compare their performance to the 2D U-Net-based AIMOS, which is the current best performing algorithm for mouse organ segmentation. We trained and validated all the networks for heart, spinal cord, right lung

and left lung segmentation in mice micro-CT images. We used the same micro-CT data as the AIMOS paper. However, we did not train and evaluate the networks separately on both the native and contrast-enhanced CT images. Instead, the training was performed only on the native CT images, while the performance was evaluated on data drawn from the same distribution (i.e., native CT images not used in training and validation) and on out-of-distribution data (i.e., contrast-enhanced CT images). The use of a different strain and age of mouse, imaging with different exposure conditions, and addition of contrast material as in the contrast-enhanced CT dataset represent a large distribution shift from the training data. Evaluation on such dataset gives a worst case estimate of the performance when the models are deployed in routine practice. For most micro-CTs typically taken at preclinical irradiation facilities, the performance is expected to be closer to that of the native CT dataset.

The DSC, MSD and 95p HD were chosen to evaluate the segmentation accuracy of the trained networks. As expected, all neural networks provided accurate segmentations of the target organs when evaluated on micro-CTs drawn from the same distribution as the training data. The DSC and MSD scores of the 2D and 3D models were comparable, and the 95p HDs for all models were well below 1 mm, with two extreme outliers for right lung segmentation of nnU-Net 2d. However, this problem was easily corrected by applying connected component analysis since the pixels are sufficiently far from the region of interest. For preclinical irradiations, a contouring accuracy of about 1 mm is reasonable considering organ movement in the thorax. Moreover, most irradiators use a discrete set of collimators, with differences in size of 1 mm or more [2, 33]. This contouring margin is also large enough to account for the penumbra (20%-80%), which was reported to be around 0.5 mm for x-rays [3] and 0.8 mm for proton beams [7] under standard setup conditions.

Overall, the nnU-Net 3d_fullres and 3d_cascade models showed superior segmentation performance for native CTs. Since no significant difference is observed between them, and since the 3d_fullres model is faster in terms of training and inference, it was deemed the best performing model for this segmentation task. Consistent for all organs, the 3d_fullres model gives a small but significant accuracy benefit compared to AIMOS. For the contrast-enhanced CTs, however, the benefit of the proposed 3D model is very large compared to AIMOS. For instance, AIMOS exhibited unacceptably large Hausdorff distances, resulting mainly from erroneous classifications on other regions of the scan. Such errors, similar to the outliers observed for nnU-Net 2d on native CT scans, can be attributed to the loss of craniocaudal information in 2D networks. Since 2D networks are trained on individual slices, Z position information is not preserved in the training,

which makes them more prone to mislabelling closely resembling pixels far from the region of interest. In effect, 2D networks like AIMOS require more labor-intensive corrections, which render them less useful in practice. Even for the most difficult case in the contrast-enhanced CT dataset, the 3d_fullres model demonstrated a more stable performance than AIMOS. Therefore, for data on which the model has not been trained on, which is common when rolling this out in the field, the proposed model is much more robust and better generalizable. This is an advantage for preclinical facilities where various animal studies are conducted, which typically have different experimental designs. In such facilities, it is difficult to build a training dataset that spans all types of images that the model would face, due to restrictions on the use of animals for imaging experiments.

When evaluated on the contrast-enhanced CT images, the nnU-Net 3d_fullres model trained on the native CTs perform equally well with AIMOS trained on the contrast-enhanced CTs. As reported in the literature, AIMOS achieved a median DSC of 0.92 (heart) and 0.95 (total lungs), and median 95p HD of 0.50 mm (heart) and 0.20 mm (total lungs) [26]. These values are comparable to the median DSC of 0.92 (heart) and 0.96 (total lungs), and median 95p HD of 0.54 mm (heart) and 0.28 mm (total lungs) achieved by our model. However, when AIMOS is not retrained on the contrast-enhanced scans, its performance on those data is considerably worse. This further confirms the superiority of the nnU-Net 3d_fullres model. Aside from employing a 3D neural network, another advantage of nnU-Net is it automatically determines the training configuration such as network depth, batch size, patch size, learning rate, and class sampling strategy tailored to the dataset provided by the user and thus removing the burden of manual tuning. It also employs more extensive data augmentation techniques than AIMOS.

To further establish the usefulness of the models for autocontouring in routine practice, they must maintain good agreement with expert contours. For that, we compared the conformity indices of the nnU-Net 3d_fullres, 2d, and AIMOS models to a consensus segmentation and evaluated their performance against the interobserver variability (IOV) for the heart, total lungs, and spinal cord. Although all three models showed superior results to the IOV for the native CT dataset, only the 3d_fullres model showed higher conformity indices on all target organs for the contrast-enhanced CT dataset. These results indicate the possibility that the 3d_fullres model is better than humans. However, further research is needed to establish this claim. A possible follow up can be a blind scoring study where participants are asked to select their preferred segmentation between manual and automated contours.

Lastly, to determine the impact of integrating these autocontouring tools in the pre-

clinical workflow, we also measured the inference time and total runtimes of the models. Using our computing system, AIMOS achieved the fastest inference time at 2 s, whereas nnU-Net 3d_fullres model took 27 s to generate predictions. The average runtime per animal for the proposed 3D model is 40 s. This runtime includes preprocessing and loading of the micro-CT images, making inference, and exporting the final contours to the desired format. Although we expect this runtime can be substantially shortened with better implementation, it is already a significant improvement from the manual contouring time of about 40 min per animal for this particular segmentation task.

This work has demonstrated that nnU-Net deep learning pipeline can be used and integrated into the preclinical workflow to provide fast and accurate contouring. While more advanced network architectures may be of interest of study, our results with nnU-Net, which employs a generic U-Net architecture, showed that the whole training process is equally important to achieve good performance across datasets. In the future, we intend to assess this method to other treatment sites and imaging modalities. In preclinical studies, MRI scans of the brain and head-and-neck are of particular interest since micro-CTs have poor contrast in these regions.

2.5. CONCLUSIONS

In summary, we reported the segmentation performance, generalizability, and efficiency of nnU-Net and AIMOS for autocontouring of the heart, spinal cord, right lung and left lung in mice micro-CT images. The best performing model for this segmentation task is the nnU-Net 3d_fullres model, which is capable of generating high quality segmentations across diverse datasets while maintaining good levels of agreement with expert contours. It also offers significant improvement in countouring time. Its implementation in routine practice as an autocontouring tool can potentially expedite the preclinical workflow and reduce the overall workload.

DATA AVAILABILITY

Annotations of the heart, spinal cord, right and left lungs used for training and testing are publicly available at https://doi.org/10.5281/zenodo.5121272. The pre-trained nnU-Net 3d_fullres model is accessible at https://doi.org/10.5281/zenodo.5786839.

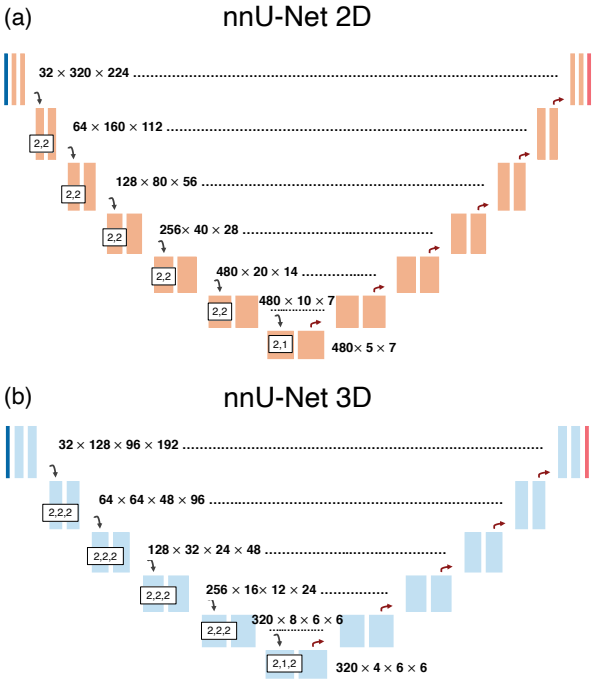
ACKNOWLEDGEMENTS

This work was supported by a grant from the Dutch Cancer Society (KWF research project 12092). The authors would also like to thank Rianne Biemans from Maastricht University for helpful discussions regarding contouring protocols and Mohamed Elmahdy from Leiden University Medical Center for assistance in the analysis.

CONTRIBUTIONS

J.M. annotated the dataset used for training and testing, conducted the experiments, analyzed the results and wrote the manuscript. J.W. annotated test set 1 and H.Q. annotated test set 2 as second observers. D.L, F.V., S.B. and M.S. supervised this study. All authors contributed to reviewing and revising the manuscript.

SUPPLEMENTARY FILE



Dataset	
[0.5, 99.5] percentile	[-664, 277]
Mean	-224.7
Standard deviation	258.9

2D U-Net	
Target spacing	0.14 × 0.14
Median Shape @ Target Spacing	284 × 210
Patch Size	320 × 224
Batch Size	44

3D full resolution U-Net	
Target spacing	0.14 × 0.14 × 0.14
Median Shape @ Target Spacing	284 × 210 × 405
Patch Size	128 × 96 × 192
Batch Size	2

3D low resolution U-Net	
Target spacing	0.19 × 0.19 × 0.19
Median Shape @ Target Spacing	207 × 153 × 295
Patch Size	128 × 96 × 192
Batch Size	2

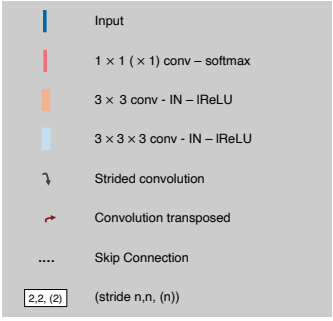


Figure S2.1: (a) 2D and (b) 3D U-Net architectures generated by nnU-Net for this dataset. The orange and blue boxes correspond to a convolution (conv)-instance normalization (IN)-leaky ReLU (lReLU) unit in 2D and 3D, respectively. The numbers beside each encoder level correspond to the channel × image size in x, y, (z) at that stage. Gray arrows indicate strided convolutions while red arrows denote convolution transposed. The final layer uses a 1 × 1 (× 1) convolution and a softmax activation function. The table in the upper right gives the properties of the foreground classes used for preprocessing the images (under Dataset). The target spacing, median shape at target spacing, patch size and batch size for the nnU-Net 2d, 3d_fullres, and 3d_lowres models are also given.

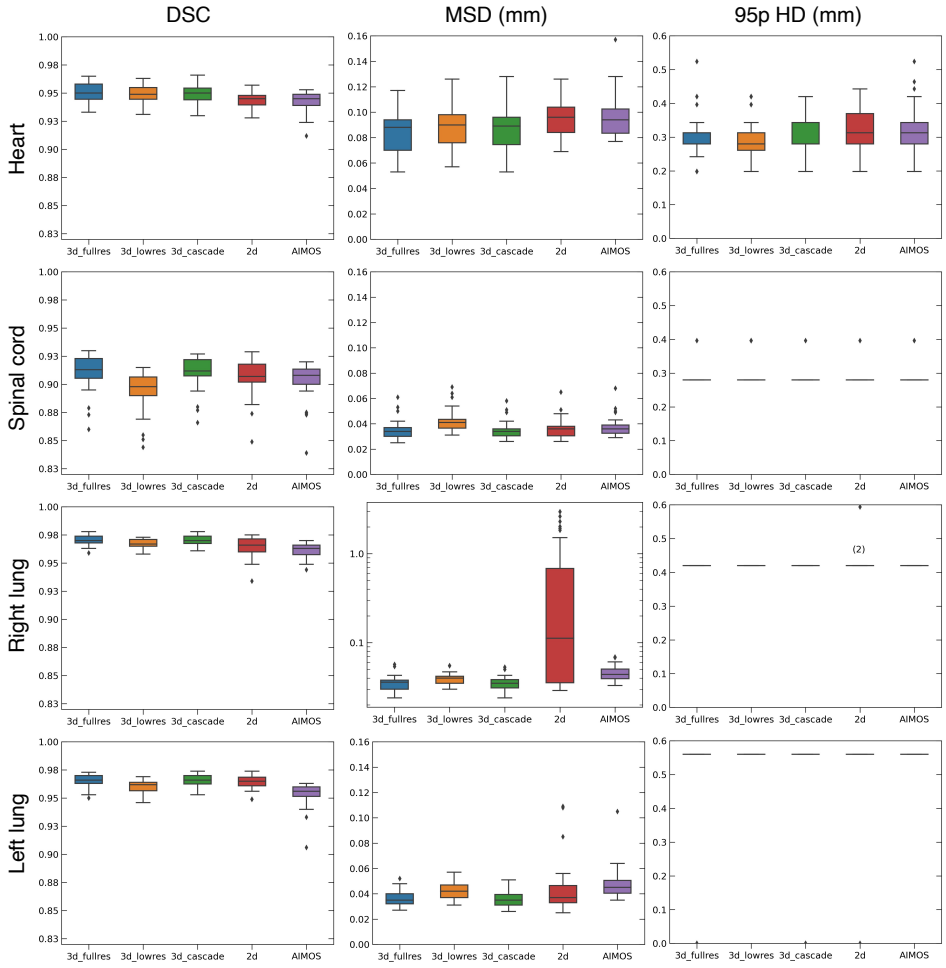


Figure S2.2: Boxplots of the evaluation metrics for test set 1 (native CTs). The first, second, and last columns correspond to the DSC, MSD and 95p HD, respectively. The number above the 95p HD boxplot of the nnU-Net 2d model for the right lung indicates the number of images with a value higher than the maximum y-value in the plot. Tight 95p boxplots for the spinal cord, right and left lungs were observed due to very little variation over the dataset. Legend: box = interquartile range, line = median, whiskers = minimum and maximum, and diamond = outliers.

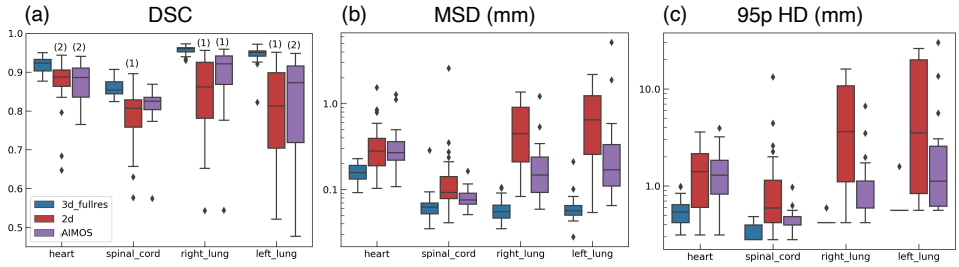


Figure S2.3: Boxplots of the evaluation metrics for test set 2 (contrast-enhanced CTs). The plots on the left, middle, and right correspond to the DSC, MSD and 95p HD, respectively. The numbers above the DSC boxplots indicate the number of images with a value lower than the minimum y-value. Legend: box = interquartile range, line = median, whiskers = minimum and maximum, and diamond = outliers.

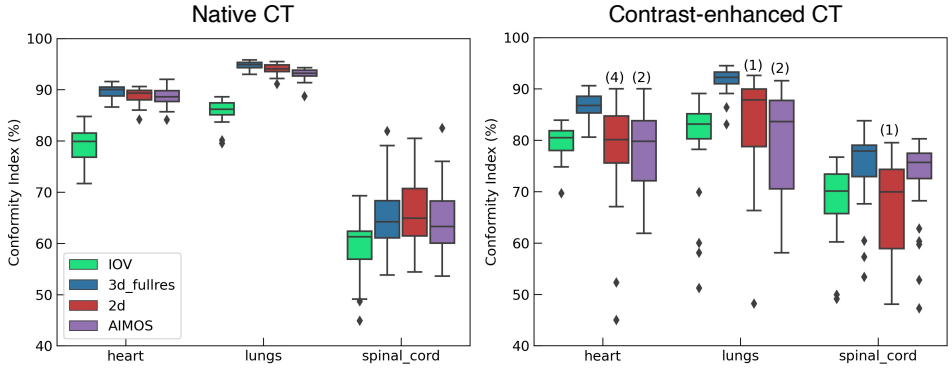


Figure S2.4: Boxplots of the conformity index between the models and the reference contours for test set 1 (left) and test set 2 (right). The conformity index among human observer delineations (green) was the baseline for comparison. The numbers above the boxplots indicate the number of images with a value lower than the minimum y-value. Legend: box = interquartile range, line = median, whiskers = minimum and maximum, and diamond = outliers.

Table S2.1: Properties of the mouse datasets used in this study.

Dataset	**Scanner**	**# Animals/ images**	**Catergory**	**Min and max image size (x,y,z)**	**Voxel spacing (mm^3)**
Native CT	Tomoscope Duo	15/105	train/ validation	280 × 202 × 371 290 × 224 × 421	0.14 × 0.14 × 0.14
		5/35	test set 1	281 × 207 × 531 290 × 220 × 603	
Contrast-enhanced CT	InSyte	10/35	test set 2	308 × 242 × 484	0.14 × 0.14 × 0.14

BIBLIOGRAPHY

[1] John Wong et al. “High-Resolution, Small Animal Radiation Research Platform With X-Ray Tomographic Guidance Capabilities”. In: *International Journal of Radiation Oncology*Biology*Physics* 71.5 (2008), pp. 1591–1599. DOI: 10.1016/j.ijrobp.2008.04.025.

[2] Frank Verhaegen, Patrick Granton, and Erik Tryggestad. “Small animal radiotherapy research platforms”. In: *Physics in Medicine and Biology* 56.12 (2011), R55. DOI: 10.1088/0031-9155/56/12/R01.

[3] Manuela C. Felix et al. “Image-Guided Radiotherapy Using a Modified Industrial Micro-CT for Preclinical Applications”. In: *PLOS ONE* 10.5 (2015), pp. 1–11. DOI: 10.1371/journal.pone.0126246.

[4] Falk Tillner et al. “Precise image-guided irradiation of small animals: a flexible non-profit platform”. In: *Physics in Medicine and Biology* 61.8 (2016), p. 3084. DOI: 10.1088/0031-9155/61/8/3084.

[5] Sunil Sharma et al. “Advanced Small Animal Conformal Radiation Therapy Device”. In: *Technology in Cancer Research and Treatment* 16.1 (2017), pp. 45–56. DOI: 10.1177/1533034615626011.

[6] R. Clarkson et al. “Characterization of image quality and image-guidance performance of a preclinical microirradiator”. In: *Medical Physics* 38.2 (2011), pp. 845–856. DOI: 10.1118/1.3533947.

[7] E Ford et al. “An image-guided precision proton radiation platform for preclinical in vivo research”. In: *Physics in Medicine and Biology* 62.1 (2016), p. 43. DOI: 10.1088/1361-6560/62/1/43.

[8] Michele M Kim et al. “Design and commissioning of an image-guided small animal radiation platform and quality assurance protocol for integrated proton and x-ray radiobiology research”. In: *Physics in Medicine and Biology* 64.13 (2019), p. 135013. DOI: 10.1088/1361-6560/ab20d9.

[9] Katia Parodi et al. “Towards a novel small animal proton irradiation platform: the SIRMIO project”. In: *Acta Oncologica* 58.10 (2019), pp. 1470–1475. DOI: 10.1080/0284186X.2019.1630752.

[10] Stefan Josephus van Hoof. “Radiation planning for image guided preclinical radiotherapy”. English. PhD thesis. Maastricht University, 2019. ISBN: 9789463237123. DOI: 10.26481/dis.20190703sh.

[11] Sara Gargiulo et al. “Mice Anesthesia, Analgesia, and Care, Part I: Anesthetic Considerations in Preclinical Research”. In: *ILAR Journal* 53.1 (2012), E55–E69. DOI: 10.1093/ilar.53.1.55.

[12] M.T. Evans, T. Wigmore, and L.J.S. Kelliher. “The impact of anaesthetic technique upon outcome in oncological surgery”. In: *BJA Education* 19.1 (2019), pp. 14–20. DOI: 10.1016/j.bjae.2018.09.008.

[13] Peter van Luijk et al. “The Impact of Heart Irradiation on Dose–Volume Effects in the Rat Lung”. In: *International Journal of Radiation Oncology*Biology*Physics* 69.2 (2007), pp. 552–559. DOI: 10.1016/j.ijrobp.2007.05.065.

[14] Juan Eugenio Iglesias and Mert R. Sabuncu. “Multi-atlas segmentation of biomedical images: A survey”. In: *Medical Image Analysis* 24.1 (2015), pp. 205–219. DOI: 10.1016/j.media.2015.06.012.

[15] William.P. Segars et al. “Development of a 4-D digital mouse phantom for molecular imaging research”. In: *Molecular Imaging and Biology* 6.3 (2004), pp. 149–159. DOI: 10.1016/j.mibio.2004.03.002.

[16] Belma Dogdas et al. “Digimouse: a 3D whole body mouse atlas from CT and cryosection data”. In: *Physics in Medicine and Biology* 52.3 (2007), p. 577. DOI: 10.1088/0031-9155/52/3/003.

[17] Tobias Gass, Gábor Székely, and Orcun Goksel. “Semi-supervised Segmentation Using Multiple Segmentation Hypotheses from a Single Atlas”. In: *Medical Computer Vision. Recognition Techniques and Applications in Medical Imaging*. Vol. 7766. Springer, 2013, pp. 29–37. DOI: 10.1007/978-3-642-36620-8_4.

[18] A. A. Joshi et al. “Posture matching and elastic registration of a mouse atlas to surface topography range data”. In: *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. 2009, pp. 366–369. DOI: 10.1109/ISBI.2009.5193060.

[19] Hongkai Wang, David B. Stout, and Arion F. Chatziioannou. “A Deformable Atlas of the Laboratory Mouse”. In: *Molecular Imaging and Biology* 17 (2015), pp. 18–28. DOI: 10.1007/s11307-014-0767-7.

[20] Brent van der Heyden et al. “Automatic multiatlas based organ at risk segmentation in mice”. In: *British Journal of Radiology* 92.1095 (2018), p. 20180364. DOI: 10.1259/bjr.20180364.

[21] Sang Hee Ahn et al. “Comparative clinical evaluation of atlas and deep-learning-based auto-segmentation of organ structures in liver cancer”. In: *Radiation Oncology* 14.213 (2019). DOI: 10.1186/s13014-019-1392-z.

[22] Wen Chen et al. “Deep learning vs. atlas-based models for fast auto-segmentation of the masticatory muscles on head and neck CT images”. In: *Radiation Oncology* 15.176 (2020). DOI: 10.1186/s13014-020-01617-0.

[23] Tao Zhang et al. “Comparison between atlas and convolutional neural network based automatic segmentation of multiple organs at risk in non-small cell lung cancer”. In: *Medicine* 99.34 (2020). DOI: 10.1097/MD.0000000000021800.

[24] Brent van der Heyden et al. “Automated CT-derived skeletal muscle mass determination in lower hind limbs of mice using a 3D U-Net deep learning network”. In: *Journal of Applied Physiology* 128.1 (2020), pp. 42–49. DOI: 10.1152/japplphysiol.00465.2019.

[25] Hongkai Wang et al. “Prediction of major torso organs in low-contrast micro-CT images of mice using a two-stage deeply supervised fully convolutional network”. In: *Physics in Medicine and Biology* 64.24 (2019), p. 245014. DOI: 10.1088/1361-6560/ab59a4.

[26] Oliver Schoppe et al. “Deep learning-enabled multi-organ segmentation in whole-body mouse scans”. In: *Nature Communications* 11.5626 (2020). DOI: 10.1038/s41467-020-19449-7.

[27] Stefanie Rosenhain et al. “A preclinical micro-computed tomography database including 3D whole body organ segmentations”. In: *Scientific Data* 5.180294 (2018). DOI: 10.1038/sdata.2018.294.

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Ed. by Nassir Navab et al. Springer International Publishing, 2015, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28.

[29] Fabian Isensee et al. “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation”. In: *Nature Methods* 18 (2021), pp. 203–211. DOI: 10.1038/s41592-020-01008-z.

[30] Abdel Aziz Taha and Allan Hanbury. “Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool”. In: *BMC Medical Imaging* 15.29 (2015). DOI: 10.1186/s12880-015-0068-x.

[31] Mohamed S. Elmahdy et al. “Robust contour propagation using deep learning and image registration for online adaptive proton therapy of prostate cancer”. In: *Medical Physics* 46.8 (2019), pp. 3329–3343. DOI: 10.1002/mp.13620.

[32] Erik Kouwenhoven, Marina Giezen, and Henk Struikmans. “Measuring the similarity of target volume delineations independent of the number of observers”. In: *Physics in Medicine and Biology* 54.9 (2009). DOI: 10.1088/0031-9155/54/9/018.

[33] Mohammad Matinfar et al. “Image-guided small animal radiation research platform: calibration of treatment beam alignment”. In: *Physics in Medicine and Biology* 54.4 (2009). DOI: 10.1088/0031-9155/54/4/005.

3

BENCHMARKING OF FAST PROTON DOSE ENGINES (MCSQUARE AND YODA) FOR SMALL ANIMAL IRRADIATIONS

Justin Malimban[1], Tiberiu Burlacu[2], Felix Ludwig[1], Zoltán Perkó[2], Marius Staring[3,4], Frank Verhaegen[5], Sytze Brandenburg[1] and Danny Lathouwers[2]

[1]Department of Radiation Oncology and Particle Therapy Research Center, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
[2]Department of Radiation Science and Technology, Faculty of Applied Sciences, Delft University of Technology, Delft, The Netherlands
[3]Department of Radiology, Leiden University Medical Center, Leiden, The Netherlands
[4]Department of Radiation Oncology, Leiden University Medical Center, Leiden, The Netherlands
[5]Department of Radiation Oncology (MAASTRO), Research Institute for Oncology & Reproduction, Maastricht University Medical Centre, Maastricht, The Netherlands

Preliminary manuscript, to be submitted

ABSTRACT

Objective: For image-guided proton irradiations, small animals are imaged at the treatment position and irradiated shortly after. This requires that irradiation planning and dose calculations be fast, as animals are maintained under anaesthesia throughout the process. One of the main bottlenecks in this workflow is dose calculation, with Monte Carlo (MC) simulation being the gold standard. However, MC simulations take considerable time to achieve dose distributions at an acceptable uncertainty level. The purpose of this work is to assess the suitability of fast dose engines MCsquare and YODA for proton dose calculation in small animals to enable a more efficient preclinical workflow.

Approach: We benchmarked MCsquare and YODA against the general-purpose MC code TOPAS. Pristine Bragg peaks were calculated in water at energies typically used for proton irradiation of mice. Benchmarking in a heterogeneous phantom was also performed to evaluate the accuracy of the codes in the presence of sharp density changes in the beam path. Dose distributions of proton irradiation plans for head, thorax, and abdominal cases in a mouse were also compared. Lastly, the average runtimes were measured to determine the time gain from these codes.

Main Results: MCsquare and YODA showed excellent agreement with TOPAS in homogeneous water, yielding range shifts < 0.1 mm, gamma pass ratios (3%/0.1mm) > 99.9%, and pencil beam runtimes of 25 s and 0.9 s, respectively. Although YODA is more computationally efficient, its accuracy is sensitive to the split scheme needed to handle lateral heterogeneities. For the heterogeneous phantom, YODA achieved a 95.1% pass rate with a 1+24+24+24+24 split scheme, compared to 99.9% for MCsquare. The accuracy of the split scheme is also dependent on the position of the heterogeneity relative to the beam, and it increasingly becomes worse as the Bragg peak lies nearer to the heterogeneity. For the proton plans, YODA performed well in relatively homogeneous regions like the head and abdomen (lateral field) with pass rates (3%/0.2mm) of 99.8% and 96.8%, respectively. However, it showed reduced performance on more complex regions like the abdomen (anterior field) and thorax, yielding gamma passing rates of 93.2% and 83.3%, respectively. MCsquare was better at handling such cases resulting in gamma pass rates > 97% for all regions. Although YODA is substantially faster, it requires pre-tuning of split schemes for different anatomical regions and beam configurations. This pre-tuning eliminates the need for real-time optimization during irradiation and ensures a smoother implementation in practice. To leverage YODA's speed and MCsquare's accuracy, using YODA for plan optimization and MCsquare for final dose calculations in small animals is recommended.

Significance: Fast dose calculation times for plan generation in animals have been demonstrated, which solves one of the major bottlenecks in the preclinical workflow. This is a crucial step towards realizing efficient image-guided proton irradiation in small animals.

Keywords: Monte Carlo, deterministic dose engine, treatment planning, small animal, preclinical

3.1. INTRODUCTION

There has been a steady increase in the use of proton therapy in managing certain cancer indications. Protons can deliver a much more conformal dose distribution and thus can spare more normal tissues compared to photons. To fully exploit the benefits of proton therapy, radiobiological studies using small animals must be conducted to understand the underlying mechanisms behind normal tissue and tumour response, to further elucidate the differential effects between photons and protons, and to test and develop new irradiation techniques such as ultra-high dose rate (FLASH) [1, 2] and spatially fractionated (minibeam radiation therapy) [3, 4] dose delivery.

In an effort to deliver more well-defined dose distributions for *in vivo* experiments, high resolution micro-cone beam CT (μ-CBCT) scanners are now being integrated into proton beamlines [5, 6, 7, 8] . This progress in radiation delivery opens up possibilities to deliver more complicated dose distributions in small animals such as spread-out Bragg peaks (SOBP). Proton dose distributions can then be shaped to preferentially target organ subvolumes or orthotopic tumour models, while presumably preserving more of the surrounding healthy tissues compared to kilovoltage (kV) X-rays and shoot-through proton beams. Such complex beam configurations, in which the actual stopping point of protons is crucial, require accurate characterization of the dose distribution.

Monte Carlo (MC) simulation is the gold standard for dose calculations [9, 10]. Since individual particles are transported, and the interaction processes and secondary particles can be simulated in detail, it yields the most accurate results, but it is very computationally expensive, which limits its use in daily preclinical practice. The long computation time of MC simulations is far from ideal for image-guided irradiation of animals since the entire irradiation workflow (i.e. imaging, contouring, planning and dose calculation, irradiation) is typically performed in a single session, wherein animals are continuously kept under anaesthesia. As prolonged exposure to anaesthesia may impact the outcome of the experiment and animal well-being, there is an increasingly pressing need for faster proton dose calculation alternatives.

Several different approaches have already been developed to speed up dose computations in the clinic. The majority of clinical treatment planning systems (TPS) employ pencil beam algorithms (PBA) for treatment plan optimization and dose calculations. They are fast and can produce reasonably accurate dose distributions in fairly homogeneous media [11, 12, 13]. However, their performance deteriorates at material interfaces with large density differences. To circumvent this issue, efforts were made to accelerate MC dose calculations, which are inherently better at handling heterogeneities, by capitalizing on the parallelization capabilities of central processing units (CPUs) [14] and

graphics processing units (GPUs) [15, 16, 17]. Nowadays, clinical treatment planning systems like RayStation also feature GPU-based MC dose engines, which have been deemed fast enough for plan optimization, final dose calculation, and robust evaluation in daily clinical practice [18, 19]. Deterministic [20] and deep learning-based [21] proton dose engines that can achieve near MC prediction accuracy within a matter of seconds have also been proposed to overcome bottlenecks for online adaptive proton therapy.

While there are rapid developments of computational tools within the clinical setting, preclinical solutions are somewhat lagging behind. There are commercially available TPS for animals on the market such as MuriPlan (Xstrahl Inc., Suwanee, Georgia, USA), μ-RayStation (RaySearch Laboratories, Stockholm, Sweden), SmART-ATP, and SmART-XPS (SmART Scientific Solutions BV, Maastricht, the Netherlands). SmART-ATP, SmART-XPS, and μ-RayStation feature MC dose engines, while MuriPlan supports both MC and superposition-convolution algorithm for photons. Additionally, μ-RayStation offers both pencil beam algorithm and MC dose engine for protons. For preclinical radiotherapy with kV X-ray beams and microbeams, deep learning has also been explored to accelerate dose calculations in small animals by denoising of low statistics MC dose distributions [22, 23]. Such neural networks can predict dose maps with similar accuracy as those calculated with large number of simulated particles without suffering from long computation times. For proton irradiations, Vanstalle et al. [24] proposed an analytical modelling of the dose in animals based on Bortfeld et al.'s [25] formulation of the Bragg curve, using experimental values for a more accurate prediction of the proton range. However, the validation of their methodology was only performed in a water phantom and not on more complex geometries like CT images of animals. Similarly, Clausen et al. [26] investigated the feasibility of using the GPU-based clinical TPS RayStation for animals by calculating doses on a Gammex phantom with target sizes typical for *in vivo* studies. However, its suitability for preclinical dose calculation purposes remains questionable as the smallest allowed dose grid is $1 \times 1 \times 1$ mm^3. Such voxel resolution is too coarse and does not allow accurate dose calculations in small animals with submillimetre-sized anatomical structures.

In this work, we borrow solutions developed for the clinic and validate their applicability for proton dose calculations in small animals. In particular, we looked into the open-source fast Monte Carlo code MCsquare [14] and the deterministic dose engine YODA [20, 27]. Both codes were benchmarked against a full Monte Carlo simulation in TOPAS [28] for homogeneous and heterogeneous phantoms as well as CT geometries of animals at energies relevant for small animal irradiations. The main objective is to assess the achievable accuracy and potential time gain for these codes.

3.2. MATERIALS AND METHODS

3.2.1. MANY-CORE MONTE CARLO (MCSQUARE)

MCsquare is a fast Monte Carlo algorithm for proton therapy optimized to run on massively parallel CPU architectures [14]. It simulates the transport of primary protons and secondary ions, but does not explicitly track secondary electrons, instead depositing their contribution to the dose locally. Note that a 30 MeV proton beam, which is typical for small animal irradiations, can transfer a maximum energy of 67 keV to secondary electrons. This corresponds to an electron range of 115 μm in tissue. Moreover, neutrons are neglected in the simulation as their contribution to the local dose deposit is less than 0.5% [29, 30].

Energy loss at each step is calculated from user-defined stopping power tables, while multiple Coulomb scattering is calculated using the formalism of Rossi and Greisen [31]. Nuclear interactions are sampled from ICRU 63 differential cross sections [32]. The default physics parameters for the MCsquare "accurate" configuration [14], including transport of secondary protons, deuterons, alphas, and nuclear interactions, were used for all dose calculations in this work.

3.2.2. YET ANOTHER DOSE ALGORITHM (YODA)

YODA is a physics-based, deterministic proton dose engine, which approximates the solution of the linear Boltzmann equation (LBE) [20, 27]. Using the continuous slowing down, energy straggling, and Fokker-Planck approximations, the LBE is reduced to two partial differential equations: (1) one-dimensional Fokker Planck (FP) equation and (2) the Fermi-Eyges (FE) equation. The product of the numerical solution to the FP equation and the analytical solution to the FE equation gives the proton phase space density, from which the dose is derived. Note that nuclear interactions are taken into account in YODA through fluence reduction of the primary protons. The energy deposit by secondary particles resulting from these interactions can be configured to be fully, partially, or not deposited locally. In our case, the fraction of energy released in nuclear interactions that is locally absorbed was calculated based on Berger's [33] fitted data as a function of energy, while the rest is disregarded. Additionally, the dose contribution from secondary electrons is treated as a local deposit.

3.2.3. BEAM SPLITTING IN YODA

Similar to pencil beam algorithms, YODA does not inherently account for lateral density heterogeneities. During the calculation, YODA relies solely on the density informa-

tion along the central axis of the beam, disregarding any other off-axis inhomogeneities present in the beam path. This results in the dose being calculated as though the beam traverses a laterally homogeneous material, leading to discrepancies in the proton range and deposited dose. To better model the effects of density variations on the dose, we follow the approach of Burlacu et al. [27], leveraging the radial symmetry of the Gaussian beam by dividing the original beam into smaller beamlets arranged in $N+1$ concentric rings around the centre of the original beam. The innermost ring contains a single beamlet at the centre, while each of the subsequent rings, positioned at progressively increasing radii, features beamlets with uniform weight and spatial spread within each ring. The radial distance of the rings (i.e. ring radii) as well as the spatial spreads and weights of the beamlets were optimized such that the combined fluence of the beamlets closely approximates the original beam. For the optimization to yield more useful split schemes, the ring radii were bound such that the rings are evenly distributed in $[0,\ f\sigma_s]$, where σ_s is the original spatial spread of the beam. This is to ensure that the rings are not placed too close to each other. Contrary to the implementation of Burlacu et al. [27], where $f = 2$, we varied the factor f in the upper bound between 1.8 and 3.0, depending on the chosen split scheme. Additionally, we noticed that adjusting the constraints on the initial spatial spreads from the original range of $[0.3\sigma_s,\ 0.8\sigma_s]$ to $[0.06\sigma_s,\ 0.6\sigma_s]$ resulted in improved split schemes.

Figure 3.1 illustrates some examples of split schemes implemented in this work for a spot with a beam size of 0.56 mm. It details the actual arrangement of the beamlets within their respective rings, their individual spreads, and shows the fluence difference between the original beam and the summed beamlets. The simplest scheme consists of 25 beamlets, distributed across the rings in a 1+6+6+12 configuration. We also explored increasing the number of beamlets, particularly in the inner rings, to improve the modelling of heterogeneities near the central ray. We also considered adding more rings to increase the spot density and further refine the distribution.

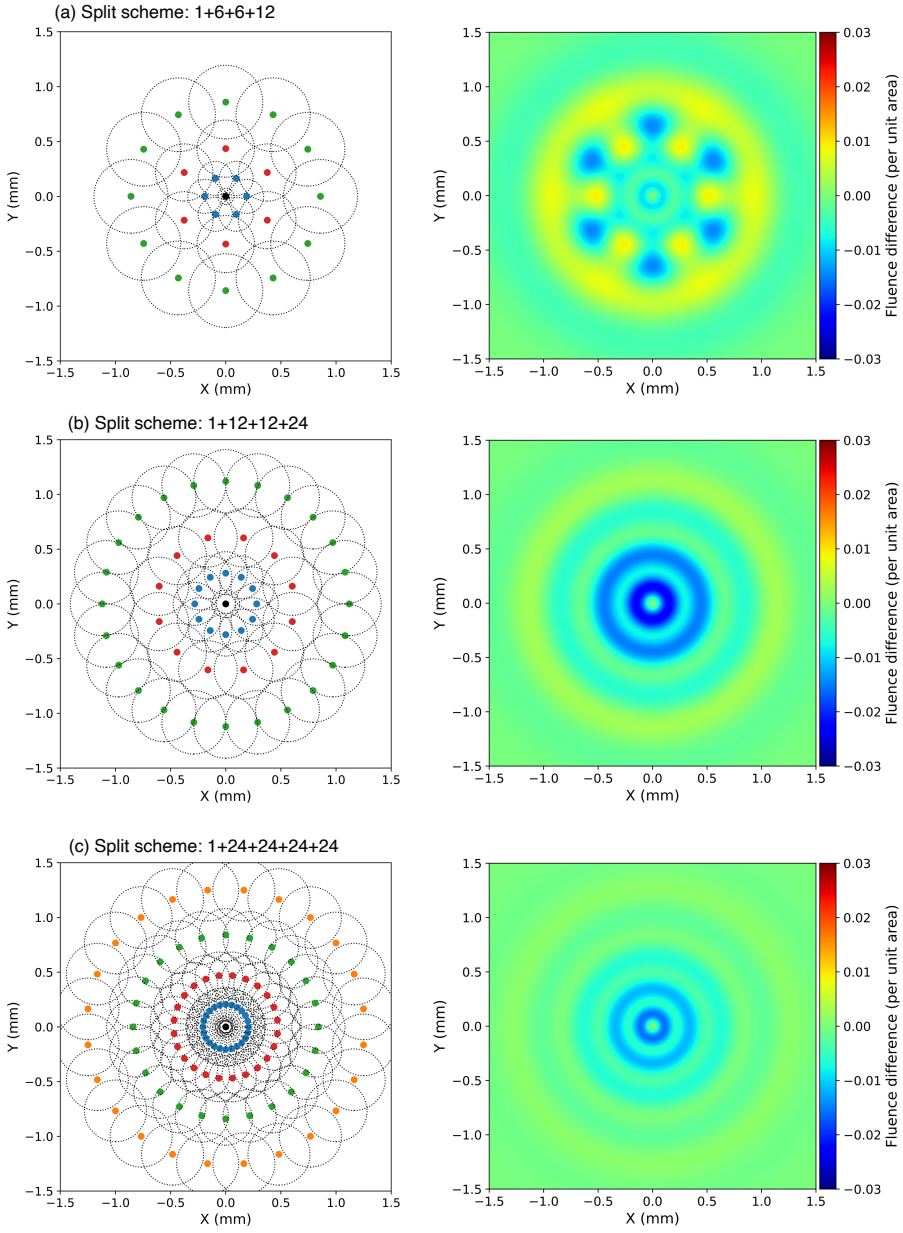


Figure 3.1: Examples of beam splitting schemes implemented in YODA: (a) 1+6+6+12, (b) 1+12+12+24, and (c) 1+24+24+24+24. The left column shows the arrangement of the beamlets in the beam's eye view. Points of the same colour represent beamlets placed in the same ring. The optimized spatial spreads of the beamlets are indicated by the circles. The right column gives the difference between the fluence of the original beam and the fluence of the summed beamlets.

3.2.4. INPUT DATA

The dose calculations were performed in voxelized geometries/DICOM CT images. The same Hounsfield units (HU) to mass density and HU to material calibration tables were used for all three codes. To avoid discrepancies in the predicted range, the standard stopping power tables in MCsquare for each material defined in the calibration were replaced with stopping power tables extracted from TOPAS for the same physics list used in the ground truth simulations. On the other hand, YODA requires separate stopping power tables of the 12 elements normally occurring in human tissues (i.e. H, C, N, O, Na, Mg, P, S, Cl, Ar, K, Ca). These were also extracted from TOPAS for the same physics list and the Bragg additivity rule was subsequently used to calculate the stopping power of each material.

3.2.5. BEAM MODELLING

Any treatment planning system requires a parameterization of the proton phase space. To build the source model, beam transport simulations in the Geant4-based BDSIM toolkit [34] were performed for the IMage guided Proton/pArticle infrastructure for preClinical sTudies (IMPACT) beamline, which is currently being constructed at UMCG-PARTREC. The beam definition for the simulation were taken from emittance measurements close to the cyclotron exit. The phase space distribution of a 66.5 MeV proton beam was scored at the exit foil. The phase space file, which contains information on individual particles (i.e. position in XYZ, momentum in XY, and kinetic energy), was used as the source for the simulation of the final part of the beamline in TOPAS as shown in figure 3.2.

The end of the beamline consisted of the ionization chamber modelled as mylar (10 μm) and aluminium foils (2 μm), polystyrene plates as the range shifter (varying thickness), and a 1-mm diameter 45-mm long brass collimator. The beam parameters were extracted at a distance of 20 mm from the exit of the collimator (i.e. the surface of the phantom) for different range shifter thicknesses in increments of 3 mm and were used to build a beam data library for the succeeding dose computations in this work. This step was needed as YODA does not offer modelling capabilities of other beamline components and the calculation starts at the surface of the object. Also, although the range shifter can be included in the MCsquare simulation, there is no option to model the collimator which comes after.

The spot size, divergence, and energy spread were modelled as a Gaussian in all codes used in this work. However, in reality, the spot size and divergence are not perfectly Gaussian due to the influence of the collimator, and the energy spread becomes

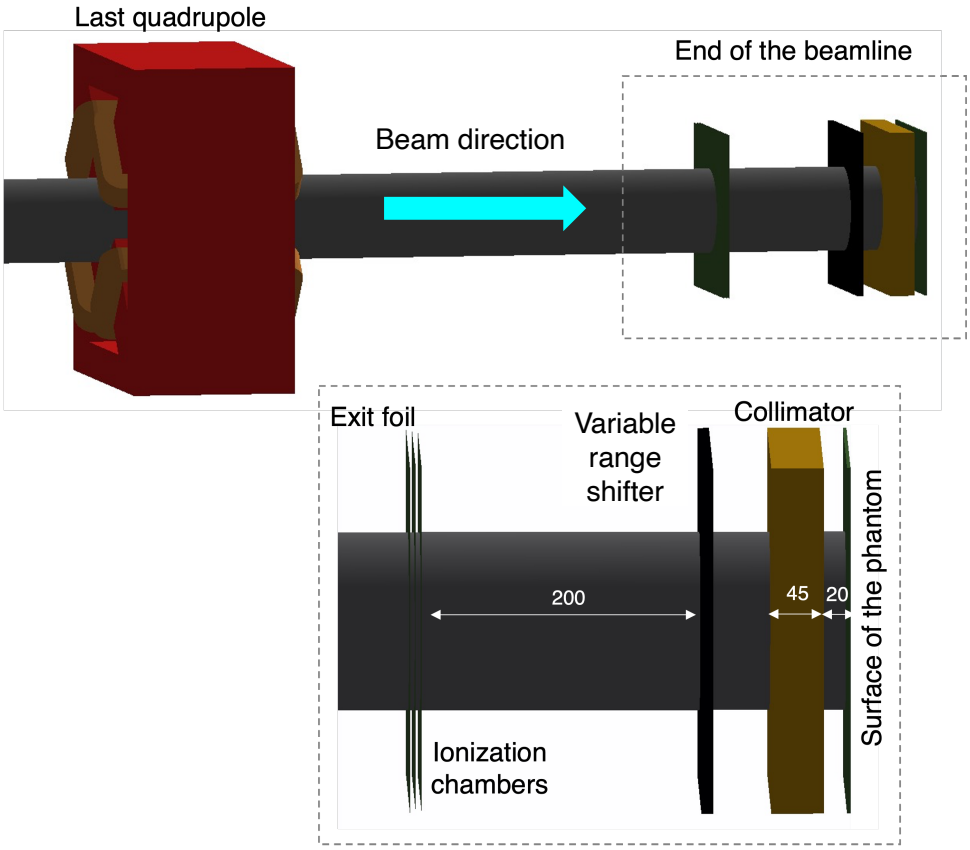


Figure 3.2: Schematic diagram of the final part of the IMPACT beamline simulated in TOPAS. The beam phase space for different range shifter configurations was scored at the surface of the phantom and was used to create the proton beam model for this study.

less Gaussian as the degrader thickness increases. To approximate these quantities, the Gaussian distribution was derived from the phase space file using the full width half maximum (FWHM), with the Gaussian sigma calculated as σ = FWHM/2.35. The values in x- and y- directions were averaged as YODA presently assumes beam symmetry in both directions. Table 3.1 provides the fitted beam parameters for different range shifter thicknesses. Given that the size of the beam hardly changes due to the collimation, the spot size was assumed to be constant at $\sigma_{x/y}$ = 0.56 mm, whereas the divergence and energy spread for intermediate values were interpolated. To maintain consistency across all three simulation codes, the beam definition at the start of the simulations was based on this beam data library.

Table 3.1: Spot size ($\sigma_{x/y}$), divergence ($\sigma_{x'/y'}$), energy (E), and energy spread (σ_E) of the beam for different range shifter (RS) configurations. Values in bold were used to build the beam data library (BDL) in OpenTPS. The spot size ($\sigma_{x/y}$) was assumed to have a fixed value of 0.56 mm in the BDL.

RS (mm)	σ_x **(mm)**	σ_y **(mm)**	$\sigma_{x/y}$ **(mm)**	$\sigma_{x'}$ **(mrad)**	$\sigma_{y'}$ **(mrad)**	$\sigma_{x'/y'}$ **(mrad)**	**E (MeV)**	σ_E **(MeV)**
15	0.55	0.55	0.55	9.64	9.63	**9.64**	**48.7**	**0.54**
18	0.56	0.56	0.56	10.01	9.73	**9.87**	**44.6**	**0.63**
21	0.56	0.56	0.56	10.23	9.94	**10.09**	**40.1**	**0.73**
24	0.56	0.56	0.56	10.39	10.29	**10.34**	**35.3**	**0.87**
27	0.57	0.57	0.57	10.61	10.45	**10.53**	**29.3**	**1.05**
30	0.57	0.56	0.57	11.05	10.86	**10.96**	**23.1**	**1.33**
33	0.57	0.56	0.57	11.69	11.74	**11.72**	**14.6**	**1.84**

3.2.6. BENCHMARK

In order to validate MCsquare and YODA in a preclinical context, where voxel sizes and anatomical structures are extremely small, their results were compared to those obtained using the Geant4-based TOPAS toolkit version 3.9 [28]. The following physics list were implemented: *g4em-standard_opt4*, *g4h-phy_QGSP_BIC_HP*, *g4em-extra*, *g4decay*, *g4h-elastic_HP*, *g4stopping*, *g4ion-binarycascade*. Range cuts of 10 μm for electrons and positrons and 1 mm for gamma rays were imposed. The dose was scored using the *DoseToMedium* scorer. The proton source distribution was sampled according to the spot size, divergence, and energy spread as described in the beam data library. Benchmarking was performed for different phantom geometries of increasing complexity and on different anatomical regions in a digital mouse phantom [35].

VALIDATION IN SIMPLE PHANTOM GEOMETRIES

Validation was initially conducted in a 20 × 20 × 20 mm^3 homogeneous water phantom with voxel size of 0.1 × 0.1 × 0.1 mm^3. To verify the proton range, the integrated depth dose (IDD) curves in water (HU:0) were obtained for a pencil beam with energies typically used in small animal irradiations. Three regions were considered: brain (22.65 MeV), thorax (30.24 MeV), and abdomen (40.33 MeV). The proton energies were determined from a test plan on a mouse phantom delivering beams to the centre of these regions.

To evaluate the impact of heterogeneities on the dose calculation, 3-mm thick lung (ρ = 0.26 g/cm^3) and bone (ρ = 1.92 g/cm^3) slabs were inserted adjacent to each other in a water phantom, wherein the lung/bone interface is parallel to the beam axis as shown

in figure 3.3. A single proton pencil beam was incident on the centre of the phantom, where the boundary between the two materials is located. To assess the robustness of the YODA split schemes, we examined three different scenarios by varying the position of the heterogeneity relative to the beam. In the first scenario, a 40.33 MeV proton beam is incident to the phantom with the heterogeneity positioned way upstream of the Bragg peaks at a depth of 5.5 mm as shown in figure 3.3(a). In the second scenario, using the same beam energy, the heterogeneity was moved to a depth of 10.5 mm, such that the Bragg peak of protons passing through the bone material lies directly behind the heterogeneity (figure 3.3(b)). Lastly, we kept the heterogeneity in the same position as in the first scenario but lowered the beam energy to 32.5 MeV, effectively replicating the second scenario's conditions at a shallower depth in the phantom.

During the optimization of the split schemes, the ring radii, spatial spreads, and weights of the beamlets were fine-tuned to achieve the highest gamma passing rates in the first heterogeneous phantom case. The same optimized parameters were applied to YODA calculations for the other two heterogeneous phantom cases and the full animal irradiation plans.

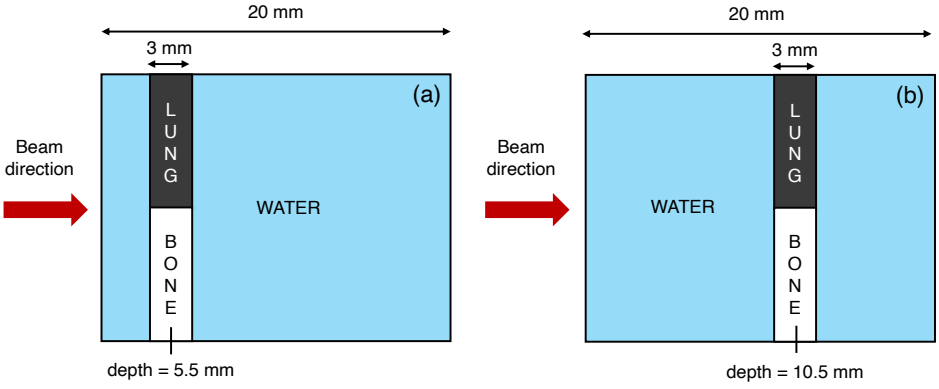


Figure 3.3: Heterogeneous phantom with 3 mm thick lung (ρ = 0.26 g/cm^3) and bone (ρ = 1.92 g/cm^3) slabs placed adjacent to each other. To test the effect of the position of the heterogeneity with respect to the beam, lung and bone slabs were placed at depths: (a) 5.5 mm for cases 1 and 3 and (b) 10.5 mm for case 2.

EVALUATION ON ANIMAL IRRADIATION PLANS

The dose distribution in an animal geometry was investigated by creating proton irradiation plans in OpenTPS [36], which employs MCsquare as its dose engine. For benchmarking purposes, the same plan optimized for MCsquare was used to calculate doses

in TOPAS and YODA. The MOBY digital phantom [35] was used for dose calculations in three anatomical regions: head, thorax, and abdomen. Artificial spherical tumours with diameters 2 mm, 3 mm, and 4 mm were added in the brain, lungs, and pancreas, respectively. These tumours served as the planning target volume (PTV), wherein a prescribed dose of 20 Gy was set. The minimum and maximum doses were constrained to ±3% of the prescribed dose for the fluence optimization. The scoring grid was set to be the same size as the CT grid. For the abdomen, we present two cases wherein the beam is approaching from the: (1) the side (lateral field) and (2) the front (anterior field) of the mouse. The first case is relatively homogeneous, with only soft tissues leading up to the PTV, whereas the second case is more complicated, with air in the beam direction due to the hollow sections of the intestine. It is important to note that the choices made during the plan creation were meant to demonstrate cases with different inhomogeneities and material transitions in the beam path. These choices may not be optimal for an actual irradiation.

It is also important to note that OpenTPS did not allow placement of spots at very shallow depths (i.e. proximal in the target) leading to difficulties in achieving good target coverage in the plans. However, this does not affect the validity of the comparison, as the primary aim of this work is to evaluate the codes. As long as the same plan is consistently applied across all three codes, the comparison remains valid.

3.2.7. DOSIMETRIC EVALUATION

To evaluate the accuracy of MCsquare and YODA for simple phantoms, we compared the integrated depth dose curves and calculated the range shift ($\Delta R_{80} = R_{80,code} - R_{80,TOPAS}$) defined as the difference in the distal range at 80% of the maximum dose. Additionally, to verify the multiple Coulomb scattering implementation, transverse profiles at depths corresponding to the positions of the lung/bone slab as well as at the first and second Bragg peaks within the heterogeneous phantom were compared. We also conducted 3D gamma analysis with TOPAS as the reference. To assess the performance of different YODA split schemes, the gamma passing rates for the homogeneous and heterogeneous phantoms were evaluated using a fixed dose difference (DD) criterion of 3% and varying distance to agreement (DTA) values of 0.1 mm (strictest), 0.2 mm, and 0.3 mm, reflecting achievable spatial resolutions for small animal imaging [37, 38]. A dose cutoff of 10% was also applied.

For the full irradiation plans, the dose difference maps, gamma index maps, and line dose profiles were analyzed. Although no consensus guidelines for gamma evaluation in the preclinical context have been established, a tolerance limit of 3%/0.2mm and 10%

dose cutoff were applied in this work. We adopt the acceptance criteria of the Groningen Proton Therapy Center, where dose distributions with gamma passing rates of 95% or higher are considered acceptable.

3.2.8. RUNTIME EVALUATION

To determine the time gain from using MCsquare and YODA, the runtimes to perform dose calculation of a single spot (E = 40.33 MeV) incident on homogeneous and heterogeneous phantoms were recorded and compared to TOPAS. The dose was calculated on a scoring grid of 0.1 × 0.1 × 0.1 mm^3. For MC-based codes (TOPAS and MCsquare), dose calculations were run to attain a statistical uncertainty of < 1%. This overall statistical uncertainty is the average statistical uncertainty of voxels that have densities greater than 0.2 g/cm^3 and dose greater than half of the maximum dose. For YODA, the runtimes of various beam splitting schemes were also evaluated for the heterogeneous case. All dose computations were performed using the same computational resources with 48 cores and repeated five times using different random starting seeds to obtain the average runtime.

Note that the runtimes reported in this work reflect only the time required for the dose calculation and do not account for the time spent on loading CT images, converting CT HU values to density and elemental compositions, initializing stopping power tables, and exporting the dose volume. The additional computation time from these processes varies based on size of the CT and the range of HU values involved and was therefore excluded in the evaluation.

3.3. RESULTS

3.3.1. PRISTINE BRAGG PEAKS IN HOMOGENEOUS AND HETEROGENEOUS PHANTOMS

Figure 3.4(a) shows the comparison of the integrated depth dose curves for a single spot with energies of 22.65 MeV (brain), 30.24 MeV (thorax), and 40.33 MeV (abdomen), as calculated by TOPAS, MCsquare, and YODA in a water phantom. The proton ranges predicted by MCsquare and YODA closely match those from TOPAS, differing by no more than 0.1 mm—less than the dose grid size—rendering these discrepancies negligible. The lateral profiles for the 40.33 MeV beam at the plateau (Z = 6 mm) and Bragg peak (Z = 14.6 mm) are shown in figures 3.4(b) and (c), respectively. Good agreement can be observed at the plateau, but slightly higher doses were seen for MCsquare and YODA compared to TOPAS at the Bragg peak. Nevertheless, both MCsquare and YODA achieved

high gamma passing rates greater than 99.9% when evaluated with the strictest criteria of 3%/0.1mm (DD/DTA) at all three energies.

Figure 3.5 presents the 2D dose slices for the heterogeneous phantom with lung and bone inserts for the three scenarios described in section 3.2.6. These cases vary based on the positioning of the heterogeneity relative to the proton beam. For YODA, three dose distributions are shown: one without any split scheme, one with the simplest split scheme (1+6+6+12), and one with the most complex scheme (1+24+24+24+24) used in this work. Table 3.2 gives the gamma passing rates for MCsquare and YODA (including all split schemes investigated) using tolerance limits of 3%/0.1 mm, 3%/0.2mm, and 3%/0.3 mm for the dose difference and distance-to-agreement.

Two distinct peaks are expected in the dose distribution: (1) at a shallower depth due to the fraction of the beam passing through the bone insert and (2) further downstream due to protons crossing the less dense lung insert. MCsquare demonstrated excellent agreement with TOPAS, achieving gamma passing rates greater than 99.9% for all three cases using the strictest DD/DTA criteria of 3%/0.1mm. Note that in this section, gamma passing rates are reported using these tolerance limits unless otherwise stated.

The accuracy of the YODA dose distribution is highly dependent on the splitting scheme implemented in the simulation. When a single beam was delivered without enforcing beam splitting, YODA only used the density information of the bone and disregarded the lung insert. As shown in the third column in figure 3.5, this approach resulted in only one peak in the dose distribution at the same depth as the shallower peak as if there is only a bone slab inserted perpendicular to the beam. To better model density heterogeneities, several split schemes were investigated. The simplest scheme employed in this work consists of one central beamlet surrounded by three concentric rings (i.e. 1+6+6+12) (see figure 3.1). Using this scheme, the gamma passing rates for cases 1, 2, and 3 were 83.2%, 82.0%, and 81.0%, respectively. Discrepancies were observed at the Bragg peaks, with the dose being overestimated at the first peak and underestimated at the second peak, as illustrated by the IDD in figure 3.6(a). Selecting a very small spatial spread for the central beamlet and increasing the number of beamlets for the rings adjacent to the interface between the two materials corrected this imbalance and resulted in more accurate dose distributions. For instance, the 1+12+12+24 scheme yielded higher gamma passing rates of 93.9%, 87.4%, and 90.4% for cases 1, 2, and 3, respectively. Further increasing the density of the beamlets in the rings closest to the interface and adding additional rings continued to improve the gamma passing rates, but the effect is minimal. The split scheme 1+24+24+24+24 was found to yield the best results for YODA with 95.1% (case 1), 88.4% (case 2), and 91.4% (case 3) of the voxels passing the test. Overall,

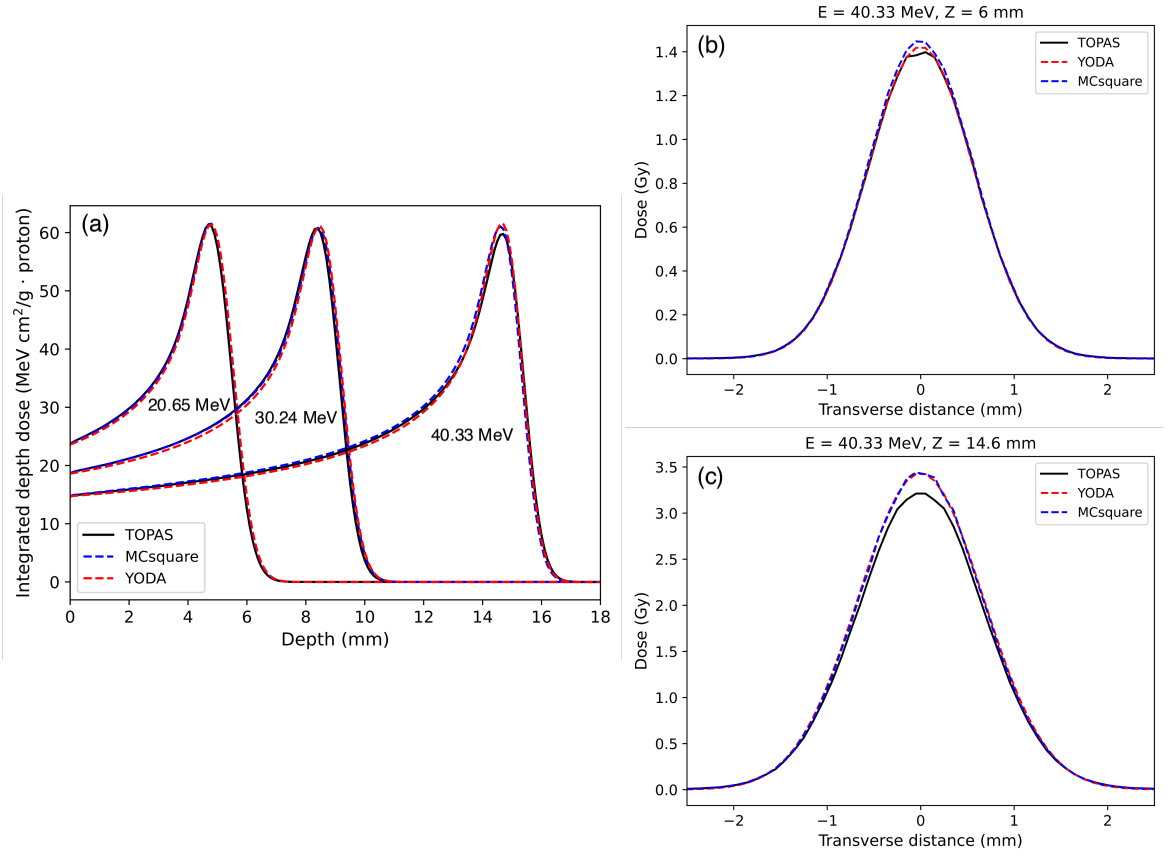


Figure 3.4: (a) Comparison of the integrated depth dose (IDD) curves from TOPAS, MCsquare, and YODA for energies typically used for irradiation of the brain (22.65 MeV), thorax (30.24 MeV), and abdomen (40.33 MeV) of mice. Lateral profiles at the (b) plateau (z = 6 mm) and (c) Bragg peak (z = 14.6) mm are compared only for the highest energy.

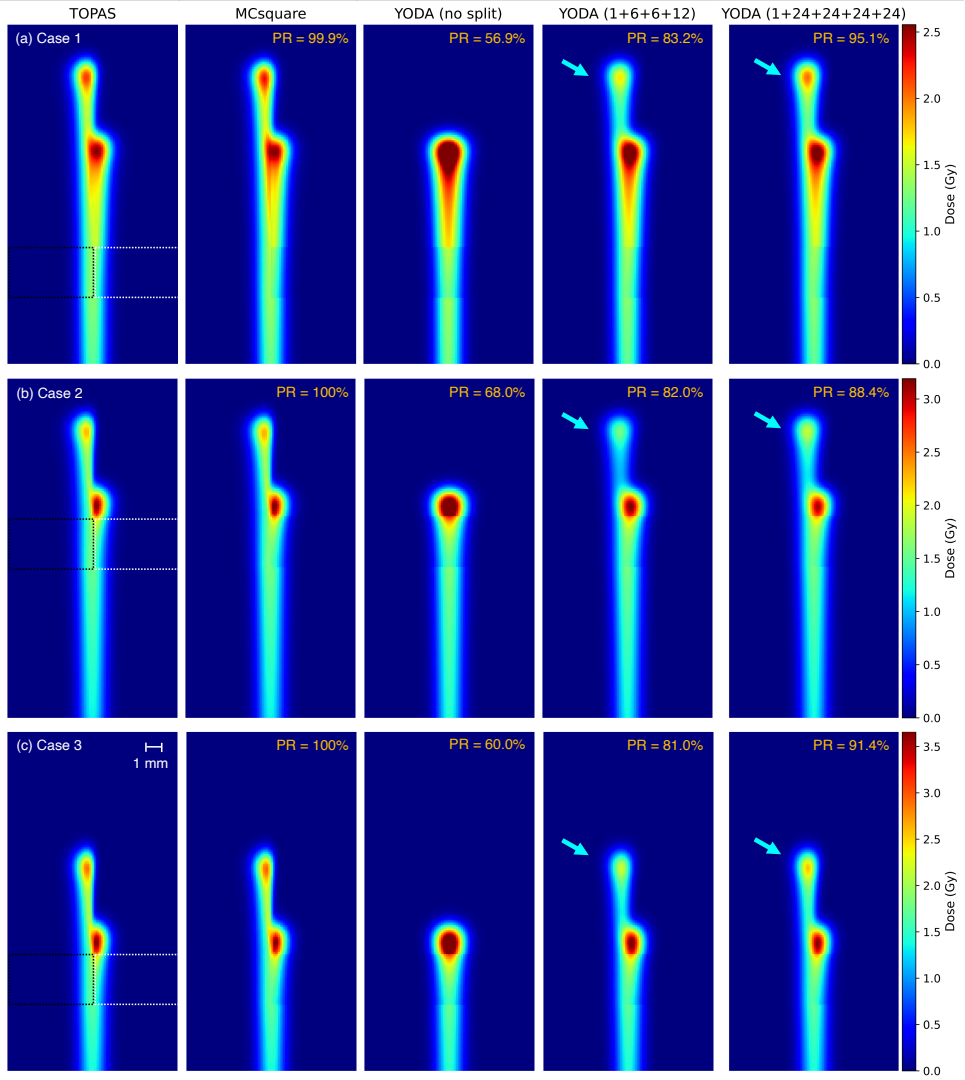


Figure 3.5: Comparison of the TOPAS, MCsquare, and YODA dose distributions in a heterogeneous phantom. (a) Case 1: The heterogeneity is placed at Z = 5.5 mm, with a 40.33 MeV beam incident on the phantom, positioning the Bragg peak well beyond the heterogeneity. (b) Case 2: The same proton beam is used, but the heterogeneity is placed at Z = 10.5 mm, bringing the first Bragg peak closer the heterogeneity. (c) Case 3: The heterogeneity is positioned as in case 1, but with beam energy of 32.5 MeV, effectively replicating case 2 at a shallower depth in the phantom. YODA doses calculated without and with beam splitting are shown. The worst (1+6+6+12) and best (1+24+24+24+24) performing split schemes were chosen for the comparison. The black and white boxes indicate the position of the lung (ρ = 0.26 g/cm^3) and bone (ρ = 1.92 g/cm^3) slabs, respectively. The arrows indicate the area where most of the discrepancies occurred in YODA. PR refers to the gamma passing rate calculated for DD/DTA of 3%/0.1mm. DD: dose difference, DTA: distance-to-agreement.

the lower gamma passing rates observed for cases 2 and 3, where the Bragg peak is close to the heterogeneity, suggest that the accuracy of the split scheme is also influenced by the position of the heterogeneity relative to the beam.

Figure 3.6 and 3.7 compare the IDD and lateral profiles for TOPAS, MCsquare, and YODA, respectively. For case 1, the results are shown for both the simplest (1+6+6+12) and most effective (1+24+24+24+24) split schemes, while for cases 2 and 3, only the best scheme is presented. As shown in figure 3.6(a), the 1+6+6+12 scheme overestimates the protons scattered into the bone insert and underestimates it in the lung insert, resulting in higher doses at the first peak and lower doses at the second peak. This discrepancy is mitigated when using split schemes with a greater number of beamlets near the heterogeneity boundary, allowing more accurate approximation of local density variations in the beam path. Selecting a very small size for the central beamlet can further reduce discrepancies; however, some distortion will still persist due to the beamlet's finite size.

Analysis of the IDDs revealed that the range shifts at both the first and second Bragg peaks for all cases are smaller than the dose grid size (0.1 mm) and can thus be considered negligible. This indicates that the proton range is accurately predicted by both codes even in the presence of low/high density inhomogeneities. Moreover, this result demonstrates that the different implementation of the stopping power calculation in YODA does not compromise the accuracy of proton range predictions. However, further testing across other material and energy combinations is necessary to validate this claim.

Discrepancies were observed on the lateral profiles particularly for YODA at the Bragg peak positions as shown in figure 3.7. While TOPAS and MCsquare showed profiles that are skewed to the left (lung part) or right (bone part) due to the large density change at the interface between the two materials, YODA exhibited a more centrally symmetric Gaussian shape. This discrepancy is linked to the inherent limitation of the Gaussian approximation in the Fermi-Eyges theory [39, 20] used by YODA to model multiple Coulomb scattering. In the calculation, the beam retains a Gaussian distribution as it traverses through the material, failing to capture the non-Gaussian effects that occur when the beam encounters transverse heterogeneities. In principle, choosing a very small spatial spread for the central beamlet helps reduce discrepancies at material interfaces. Our tests also indicate that increasing the density of the spots near the heterogeneity contributed to improved accuracy.

We recognize that the 3%/0.1mm DD/DTA criteria might be overly stringent for evaluation as it is too sensitive to very small shifts and variations between neighboring voxels, which may not have significant dosimetric impact. Using a 3%/0.2mm or 3%/0.3mm DD/DTA gamma index criteria provides a much more practical and realistic assessment

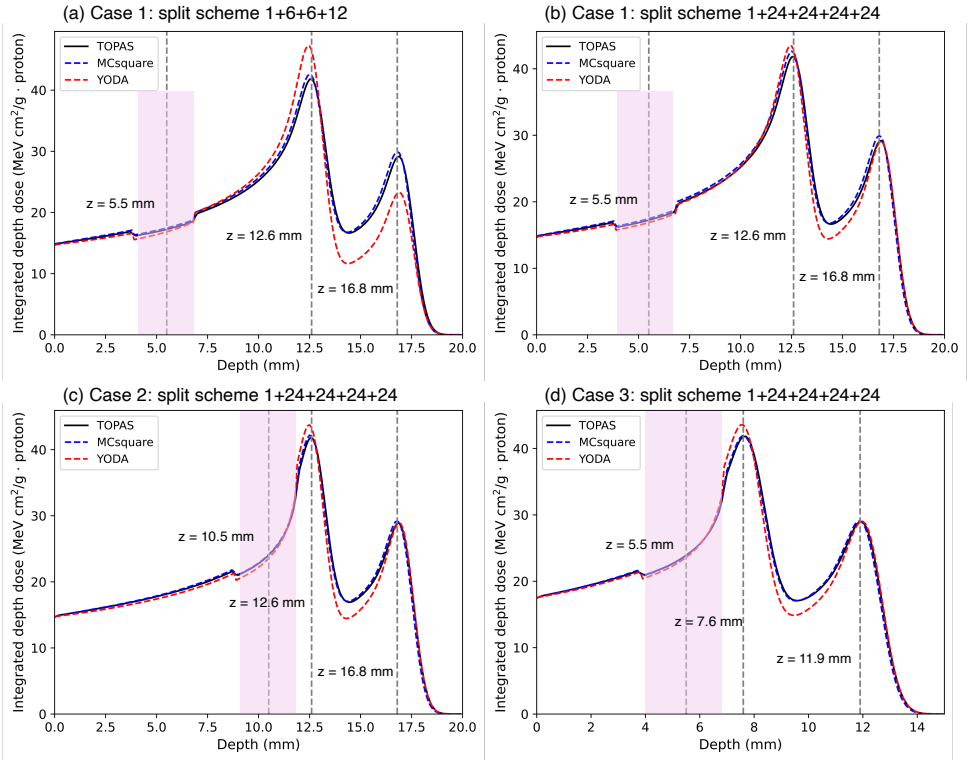


Figure 3.6: Comparison of the integrated depth dose (IDD) curves from TOPAS, MCsquare, and YODA for the heterogeneous phantom: (a) Case 1 with YODA using a 1+6+6+12 split scheme, (b) Case 1 with YODA using a 1+24+24+24+24 split scheme, (c) Case 2 with YODA using a 1+24+24+24+24 split scheme, and (d) Case 3 with YODA using a 1+24+24+24+24 split scheme. The three cases vary based on the position of the heterogeneity relative to the beam as illustrated in figure 3.5. Detailed description of each case can be found in section 3.2.6. Gray dashed lines indicate the depths at which lateral profiles in figure 3.7 were taken. The pink region indicates the position of the lung and bone heterogeneity.

for animal studies. When evaluated using these criteria, even the simpler split schemes for YODA can achieve acceptable gamma passing rates. These simpler split schemes facilitate faster dose calculations as fewer beamlets are involved in the computation.

3.3.2. PROTON IRRADIATION PLANS IN MICE

An overview of the TOPAS, MCsquare, and YODA dose distributions, and the corresponding dose difference and gamma index maps calculated with TOPAS as the reference for the head, thorax, and abdominal cases (lateral and anterior beams) is given in figures

Table 3.2: Gamma pass ratios for MCsquare and YODA evaluated across three different configurations of the heterogeneous phantom as described in section 3.2.6. The results include various split schemes implemented in YODA. DD: dose difference, DTA: distance-to-agreement, #: total number of beamlets for each spot.

YODA beam configuration						
#	Splitting scheme	DD (%)	DTA (mm)	Passing rate (%)		
				Case 1	Case 2	Case 3
1 central + 3 concentric rings						
25	1+6+6+12	3	0.1	83.2	82.0	81.0
		3	0.2	93.4	89.0	90.0
		3	0.3	97.1	93.6	95.1
49	1+12+12+24	3	0.1	93.9	87.4	90.4
		3	0.2	98.0	93.4	96.1
		3	0.3	98.9	95.7	97.7
73	1+24+24+24	3	0.1	94.5	88.0	90.7
		3	0.2	98.6	93.7	96.3
		3	0.3	99.2	95.8	98.2
1 central + 4 concentric rings						
37	1+6+6+12+12	3	0.1	88.3	84.6	85.2
		3	0.2	95.6	91.2	93.3
		3	0.3	97.6	94.7	96.1
73	1+12+12+24+24	3	0.1	94.6	87.9	91.0
		3	0.2	98.5	93.9	96.8
		3	0.3	99.2	96.0	98.2
97	1+24+24+24+24	3	0.1	95.1	88.4	91.4
		3	0.2	98.9	94.1	97.2
		3	0.3	99.4	96.2	98.6
MCsquare						
		3	0.1	99.9	100	100
		3	0.2	100	100	100
		3	0.3	100	100	100

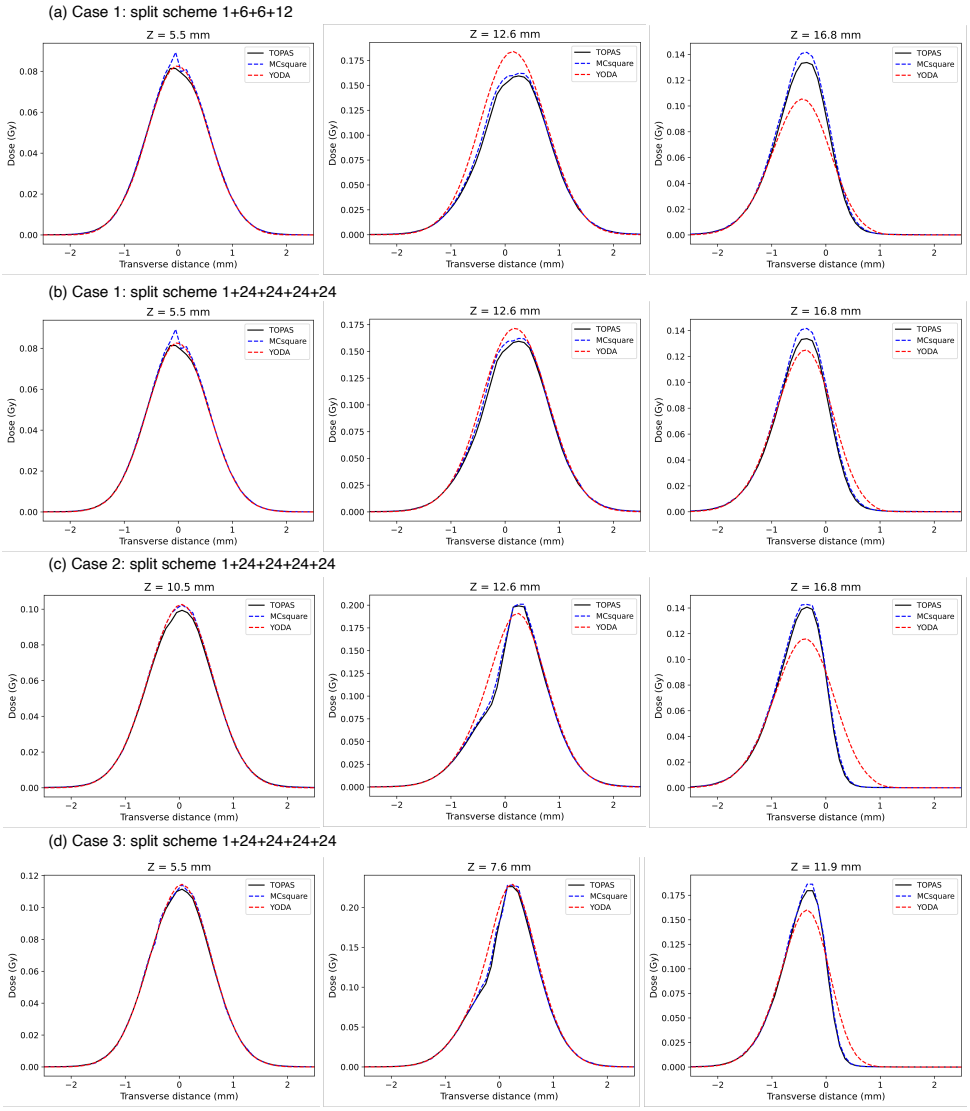


Figure 3.7: Comparison of the lateral profiles from TOPAS, MCsquare, and YODA for the heterogeneous phantom: (a) Case 1 with YODA using a 1+6+6+12 split scheme, (b) Case 1 with YODA using a 1+24+24+24+24 split scheme, (c) Case 2 with YODA using a 1+24+24+24+24 split scheme, and (d) Case 3 with YODA using a 1+24+24+24+24 split scheme. The profiles were extracted at the position of the lung/bone inserts (left), first (middle) and second (right) Bragg peaks as indicated by the grey lines in the IDDs in figure 3.6. The three cases vary based on the position of the heterogeneity relative to the beam as illustrated in figure 3.5. Details for each case are provided in section 3.2.6.

3.8 to 3.11, respectively. Line dose profiles for each case are illustrated in figure 3.12. These dose profiles were extracted at positions indicated in the dose distributions in figures 3.8 to 3.11. The gamma passing rates for both codes considering DD/DTA criteria of 3%/0.2mm are summarized in table 3.3. For YODA, we provide the gamma passing rates for both the 1+6+6+12 and 1+24+24+24+24 schemes. Unless otherwise stated, YODA results presented in this section always refer to the dose distributions calculated using the 1+24+24+24+24 split scheme, which achieved the highest gamma passing rates for the heterogeneous phantom.

Both MCsquare and YODA showed good agreement with TOPAS for the head case, demonstrating high gamma passing rates of 100% and 99.8%, respectively. Voxels that failed the gamma analysis for YODA are mostly located at the interface between bone and soft tissue, where the dose was underestimated as shown in the difference map in figure 3.8.

YODA performed worst on the thorax case with only 83.3% of the voxels passing the gamma evaluation. Dose differences between YODA and TOPAS occurred mostly at the distal portion of the beam after the PTV and air pockets in the lung tissue as shown in figure 3.9. These discrepancies likely stem from the split scheme not being optimized to handle a very heterogeneous region like the thorax, wherein many local density changes exist due to the presence of different types of anatomical structures such as soft tissue, ribs, lungs, and airways. The dose difference map in figure 3.9 shows that the distortion in the dose distribution introduced by the split scheme caused an overdosage or underdosage up to 20% of the prescribed dose at the distal edge of the Bragg peak. The overestimation in the dose is also further illustrated in the representative dose profiles shown in figure 3.12(b). Similarly, underdosage in the intestine (lumen) located after the PTV for the abdomen case with a left lateral field is shown in the dose profiles in figure 3.12(c), for which the gamma pass rate is 96.8%. In contrast, MCsquare performed well and yielded very similar dose distributions with TOPAS for the thorax and abdominal (lateral field) cases, resulting in passing rates greater than 99.95%.

For the abdomen case with an anterior field, both MCsquare and YODA showed overdosage in the intestine located upstream of the PTV as shown in the difference maps in figure 3.11 and line dose profiles in figure 3.12(d). This discrepancy can be attributed to the fact that the contribution of electrons is locally deposited in both codes. However, in low density regions, electrons can have a range larger than the voxel size, resulting in dose deposition away from the voxel of creation. Despite these differences, the gamma index comparison of the MCsquare result with TOPAS is still satisfactory, with 97.1% of the voxels passing the test. On the other hand, the gamma passing rate for YODA is 93.2%

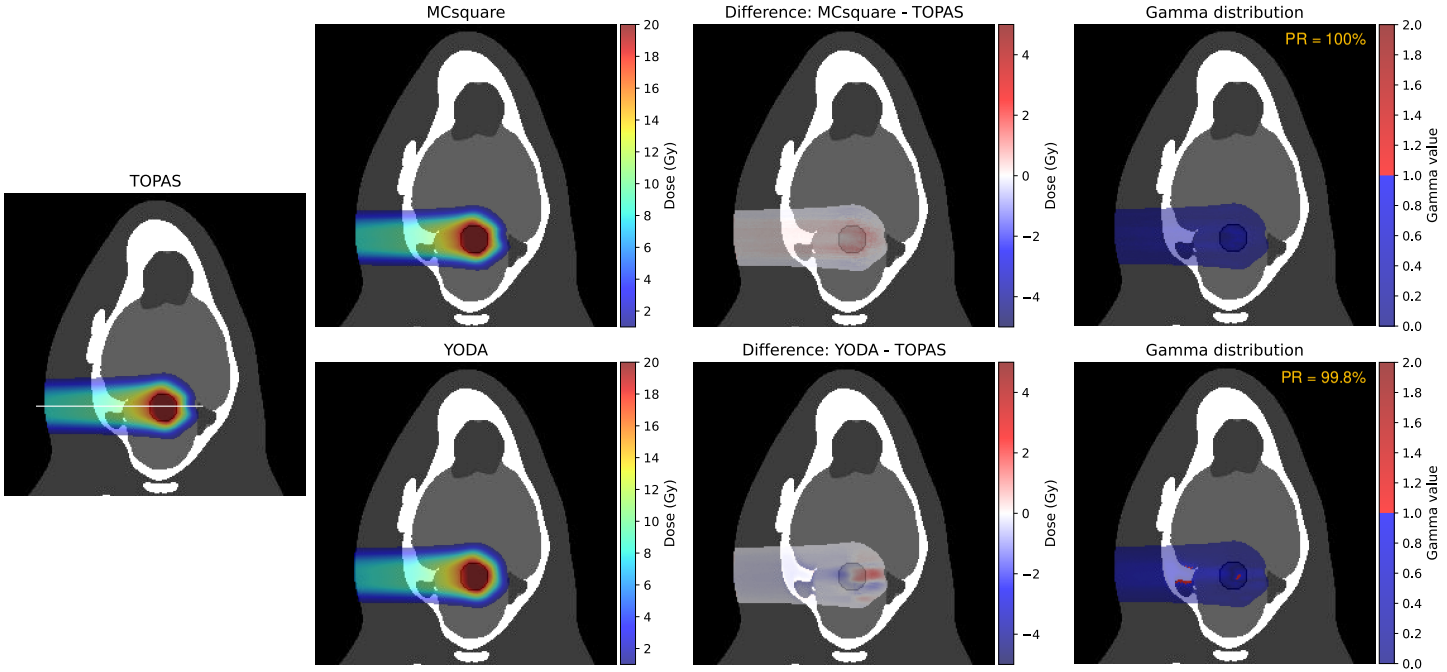


Figure 3.8: Comparison of the dose distributions resulting from TOPAS, MCsquare, and YODA for the head irradiation plan in a mouse. Dose difference (3rd column) and gamma index maps (4th column) using tolerance limits (DD/DTA) of 3%/0.2mm were calculated with TOPAS as the reference dose distribution. The white line indicates where the dose profile in figure 3.12(a) was taken. DD: dose difference, DTA: distance to agreement, PR: passing rate.

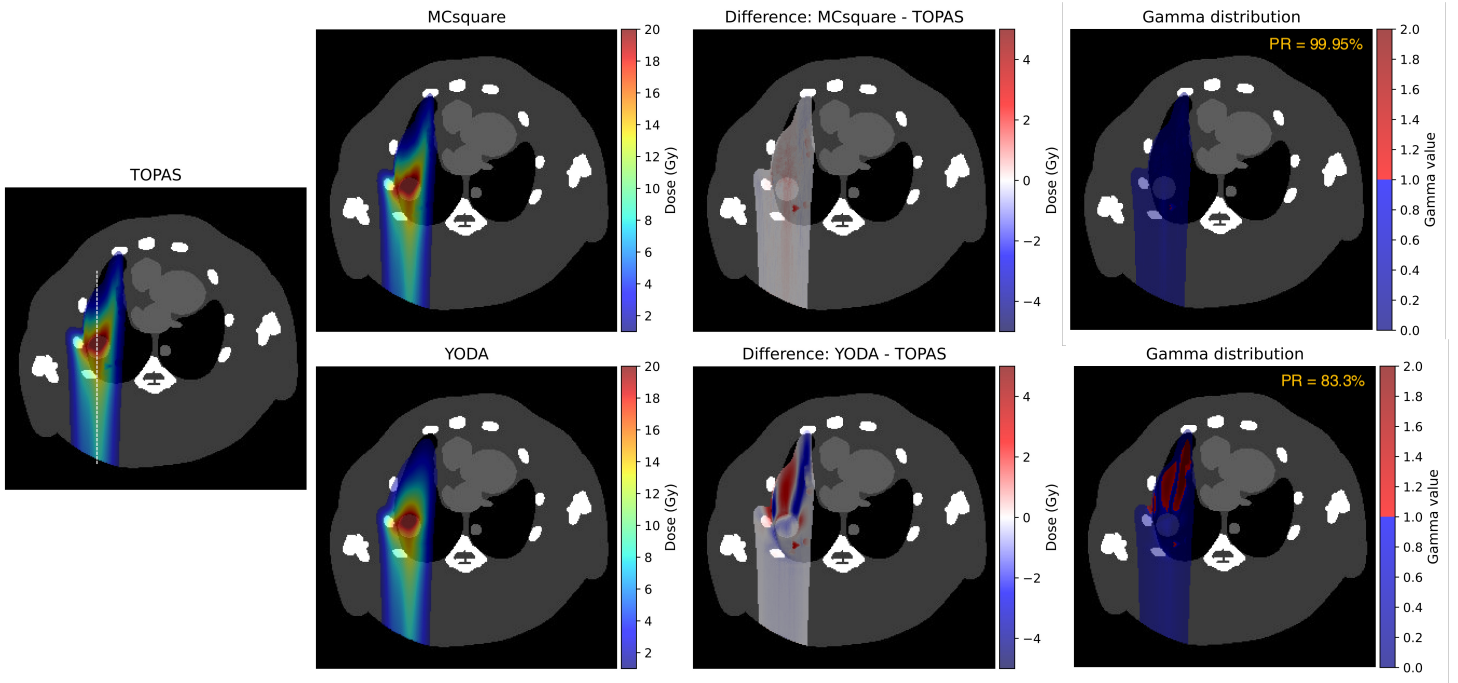


Figure 3.9: Comparison of the dose distributions resulting from TOPAS, MCsquare, and YODA for the thorax irradiation plan in a mouse. Dose difference (3rd column) and gamma index maps (4th column) using tolerance limits (DD/DTA) of 3%/0.2mm were calculated with TOPAS as the reference dose distribution. The white line indicates where the dose profile in figure 3.12(b) was taken. DD: dose difference, DTA: distance to agreement, PR: passing rate.

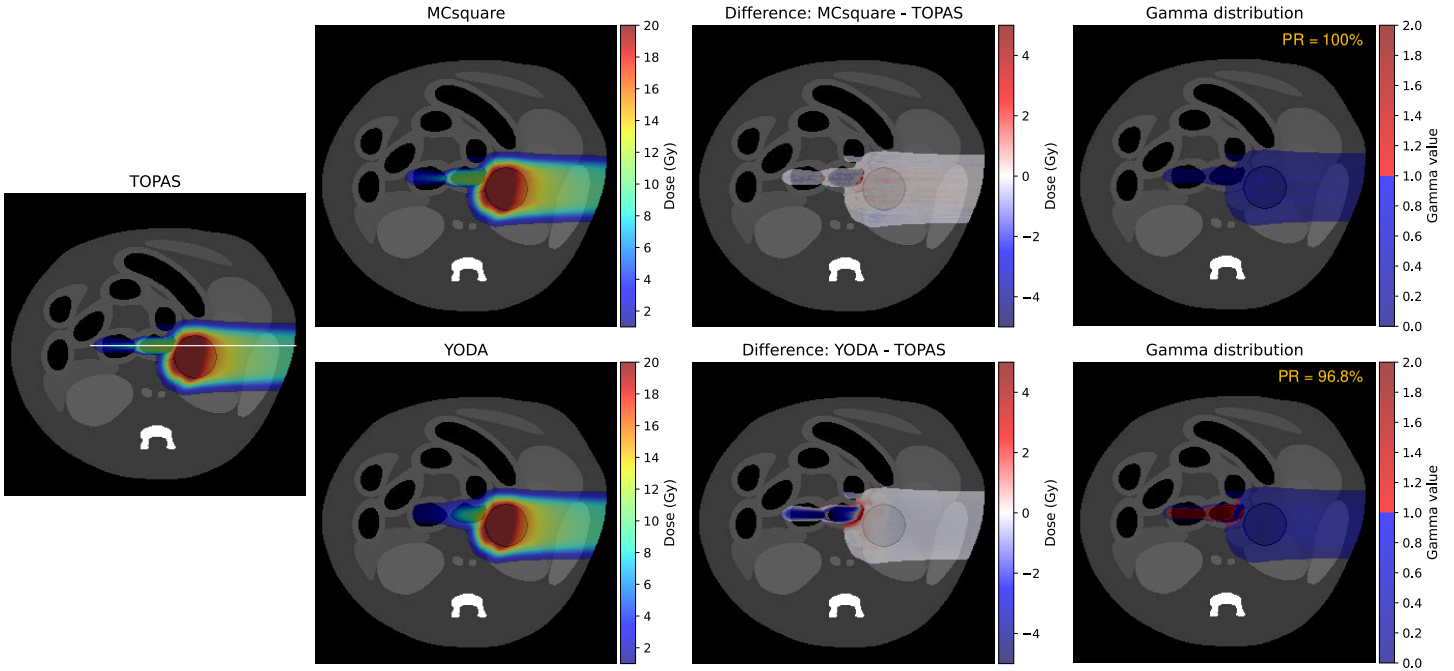


Figure 3.10: Comparison of the dose distributions resulting from TOPAS, MCsquare, and YODA for the abdominal irradiation plan in a mouse using a left lateral field. Dose difference (3rd column) and gamma index maps (4th column) using tolerance limits (DD/DTA) of 3%/0.2mm were calculated with TOPAS as the reference dose distribution. The white line indicates where the dose profile in figure 3.12(c) was taken. DD: dose difference, DTA: distance to agreement, PR: passing rate.

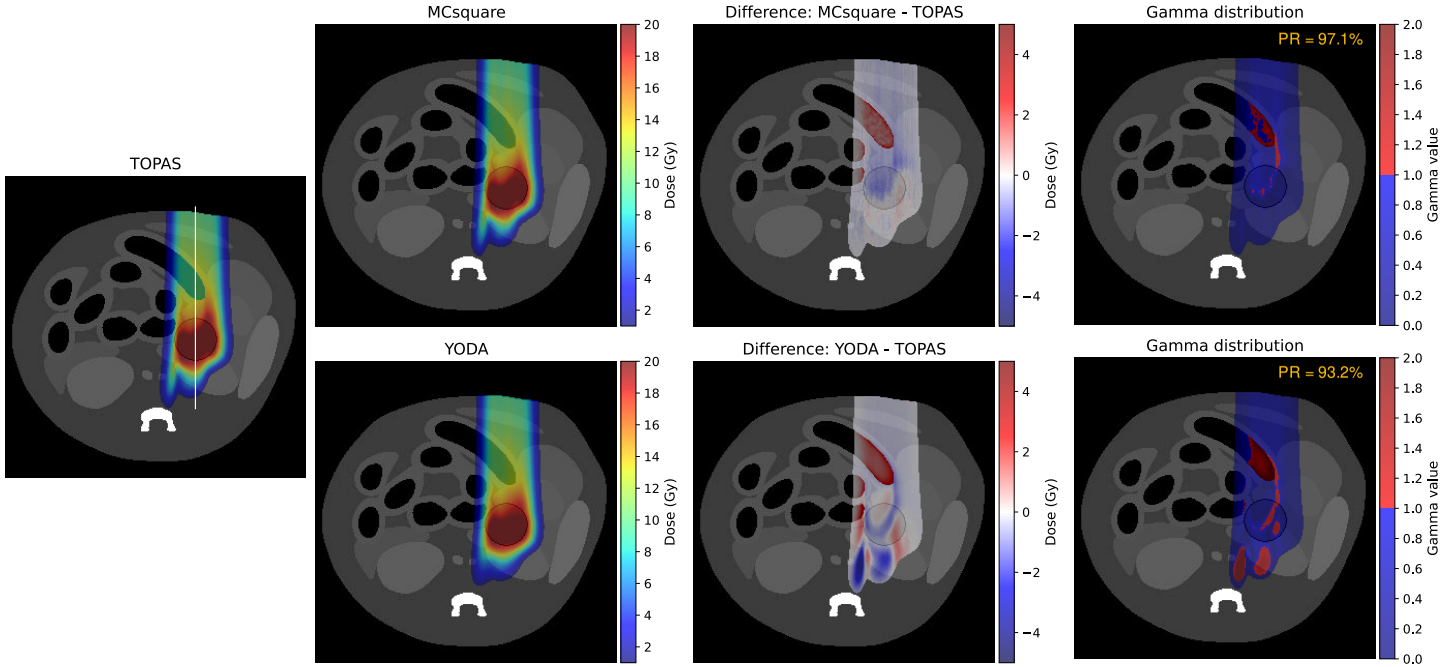


Figure 3.11: Comparison of the dose distributions resulting from TOPAS, MCsquare, and YODA for the abdominal irradiation plan in a mouse using an anterior field. Dose difference (3rd column) and gamma index maps (4th column) using tolerance limits (DD/DTA) of 3%/0.2mm were calculated with TOPAS as the reference dose distribution. The white line indicates where the dose profile in figure 3.12(d) was taken. DD: dose difference, DTA: distance to agreement, PR: passing rate.

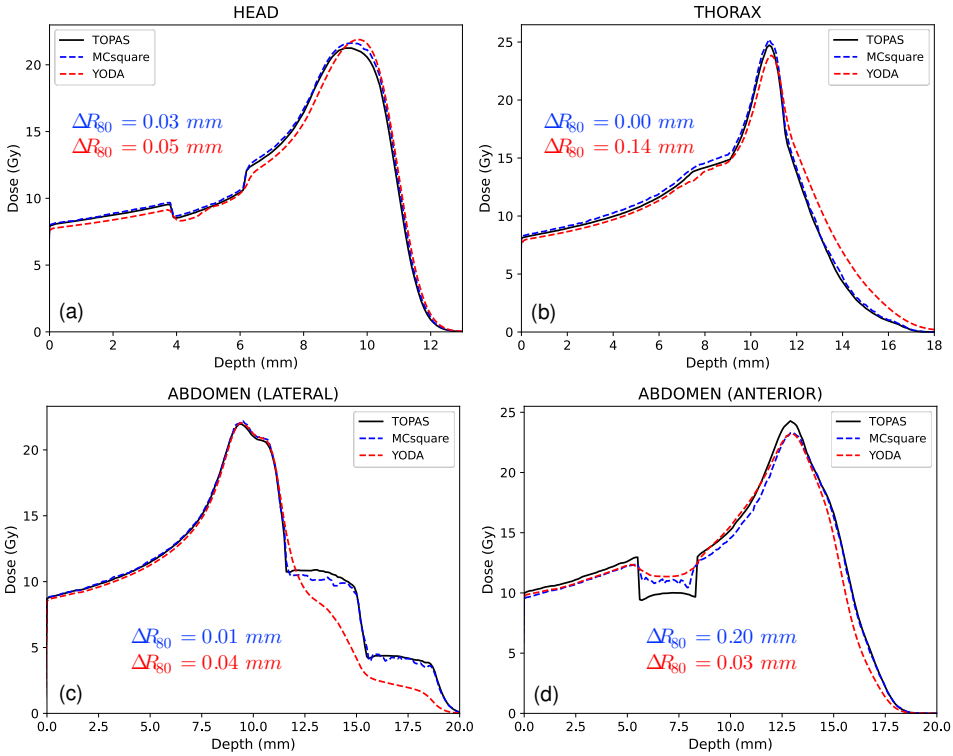


Figure 3.12: Representative line dose profiles in the (a) head, (b) thorax, and (c-d) abdomen (lateral/anterior beam) to quantify range shifts ($\Delta R_{80} = R_{80,code} - R_{80,TOPAS}$). These dose profiles were extracted at positions as indicated by the white lines in figures 3.8 to 3.11.

due to additional discrepancies at the distal end brought about by errors associated with the beam splitting scheme. Although differences at the distal edge were observed for YODA for all cases except the head, range shifts were still within acceptable limits, with less than 0.2 mm variations.

As shown in table 3.3, the use of the more complex 1+24+24+24+24 split scheme in YODA for full animal plans provided only minimal accuracy gains compared to the simpler 1+6+6+12 scheme, in contrast to the substantial improvements observed with the heterogeneous phantom. This suggests that in reality where multiple spots are delivered in close proximity, the added complexity of the 1+24+24+24+24 may be unnecessary. The simpler 1+6+6+12 scheme proves to be more efficient while still maintaining comparable plan quality.

Table 3.3: Gamma passing rates for MCsquare and YODA calculated using 3%/0.2mm DD/DTA criteria for irradiation plans created in the head, thorax, and abdomen of a mouse. YODA results for 1+6+6+12 (simplest) and 1+24+24+24+24 (most accurate) split schemes are presented.

Anatomical region	Code	Passing rate (%)
Head	MCsquare	100
	YODA (1+6+6+12)	99.7
	YODA (1+24+24+24+24)	99.8
Thorax	MCsquare	99.95
	YODA (1+6+6+12)	82.8
	YODA (1+24+24+24+24)	83.3
Abdomen (lateral)	MCsquare	100
	YODA (1+6+6+12)	96.5
	YODA (1+24+24+24+24)	96.8
Abdomen (anterior)	MCsquare	97.1
	YODA (1+6+6+12)	92.0
	YODA (1+24+24+24+24)	93.2

3.3.3. RUNTIME COMPARISON

Table 3.4 summarizes the average runtimes to calculate a pencil beam with a statistical uncertainty of less than 1% for the dose engines investigated in this work. While a full MC simulation with TOPAS took over 10 minutes to calculate a pencil beam dose in a homogeneous medium, MCsquare required 25 seconds, and YODA took less than 1 second to perform the same calculation. TOPAS also exhibited a substantial increase in the computation time in the presence of heterogeneities unlike MCsquare. For YODA, the runtime for heterogeneous cases depends on the split scheme used, with a slight increase in the computation time observed as more elaborate schemes are implemented. The 1+24+24+24+24 split scheme, which has a total of 97 beamlets per spot and produced the best results for the heterogeneous phantom, took less than 2 seconds to complete.

3.4. DISCUSSION

This work assessed the feasibility of using fast proton dose engines, MCsquare and YODA, originally developed for patients, to speed up dose calculations in small animals. The runtime evaluation showed that both codes substantially reduce the computation time compared to a full Monte Carlo (MC) simulation in TOPAS. With 48 cores running in

Table 3.4: Average runtimes of TOPAS, MCsquare, and YODA for calculation of a pencil beam in homogeneous and heterogeneous phantoms. The last two columns indicate the total number of beamlets and the split scheme implemented for the heterogeneous case in YODA.

Code	Runtime (s)			
	Homogeneous	Heterogeneous	#	Scheme
TOPAS	818.8 ± 1.5	1426.0 ± 3.6	-	-
MCsquare	25.2 ± 0.4	25.4 ± 0.3	-	-
YODA	0.9 ± 0.0	0.9 ± 0.1	25	1+6+6+12
		0.9 ± 0.0	37	1+6+6+12+12
		1.3 ± 0.3	49	1+12+12+24
		1.8 ± 0.2	73	1+12+12+24+24
		1.8 ± 0.2	73	1+24+24+24
		1.9 ± 0.4	97	1+24+24+24+24

parallel, MCsquare showed reasonably fast execution time of about 25 seconds per spot. YODA, however, proved even more efficient, completing the same calculation within 2 seconds using the most complicated split scheme implemented in this study, which involves a total of 97 beamlets. The potential for further speed gains through GPU implementation makes YODA even more promising. These reductions in computation time will facilitate rapid plan generation and optimization during animal irradiations, greatly enhancing the overall efficiency of the preclinical workflow.

To evaluate the accuracy, we benchmarked these codes against the general-purpose MC code TOPAS for a variety of configurations. Test cases on homogeneous and heterogeneous phantoms demonstrated that both MCsquare and YODA can accurately predict the proton range for energies commonly used in small animal irradiations, even with large density variations in the beam path.

To create irradiation plans in animal geometries, we used OpenTPS [36], which is an open source proton treatment planning system for humans. During our use of this TPS, we encountered several limitations. A major issue is the energy threshold that restricted the spot placement at very shallow depths. This led to no spots placed at the proximal side of the PTV, which resulted in poor dose coverage as shown in figures 3.8 to 3.11. The dose objectives for plan optimization were also limited to basic criteria such as maximum, minimum, and mean dose in a structure. More advanced optimization functions, such as dose volume histogram (DVH) related objectives, typically found in commercial TPS were not available. These limitations presented challenges in generating good irradiation plans for anatomical regions in animals investigated in this work. However, this

issue is not considered critical for our purpose. Since our goal is to compare the performance of different dose engines, using the same plan for the calculation in each code ensures the comparison remains valid.

Validation on different anatomical regions in a mouse showed that MCsquare can generally predict dose distributions with accuracies comparable to TOPAS. MCsquare yielded gamma passing rates (DD/DTA = 3%/0.2mm) greater than 99.95% for the head, thorax, and abdomen with a lateral field. The worst comparison was seen in the abdomen case with an anterior beam, where differences were observed in low density regions such as the hollow portion of the intestine (consisting of air), for which MCsquare predicts higher doses. The same behaviour was observed for YODA as shown in figure 3.11. This can be explained by the lack of electron transport, forcing their contribution to the dose to be locally absorbed [14]. Such an approximation may not always hold true especially since the voxel size of animal CT images is in the order of 0.1 mm. For example, the highest proton energy in this plan is 41.2 MeV, which can transfer a maximum energy of about 91 keV to secondary electrons. This corresponds to a range of ~190 μm in tissue. Since the lumen of the intestine is modelled as air (ρ = 0.0012 g/cm^3), these secondary electrons would have an even larger range and deposit their dose contribution elsewhere when explicitly simulated, which was not the case in MCsquare and YODA. Despite this, MCsquare still achieved a gamma passing rate of 97.1% for this plan. For all investigated cases, MCsquare met the gamma acceptance criteria and demonstrated sufficiently fast calculation times, confirming its suitability as a dose calculation platform for irradiation planning in animals.

On the other hand, the achievable quality of the dose distribution for YODA is highly dependent on the beam splitting scheme implemented in the calculation. As illustrated by the YODA dose distribution calculated without beam splitting for a single spot incident on a phantom with lung and bone inserts in figure 3.5, the range and multiple Coulomb scattering of the proton beam are determined by the material along the central axis. Therefore, when the width of the beam is relatively large compared to the heterogeneities, as is the case in animals where anatomical structures are of the order of 300 to 400 μm [37], the off-axis density variations across the beam will be neglected, resulting in considerable errors in the dose distribution.

Dividing the beam into smaller beamlets helps mitigate this issue and better account for lateral heterogeneities. Burlacu et al. [27] performed rigorous tests on a heterogeneous phantom with a slab of bone or air moving laterally into the beam. From their examinations using beam energies and voxel sizes typical in clinical settings (i.e. E = 70 to 230 MeV and voxel size = 1 mm), they found that even a simple split scheme of 1+6+6+12

for YODA can achieve high gamma passing rates > 95% using a strict criteria of 1%/1mm (DD/DTA). However, our findings showed that the same split scheme optimized for a proton beam and resolution typical for small animals performed worse, with only 83.2% of the voxels passing the gamma analysis using our strictest criteria of 3%/0.1mm (case 1). The results also progressively deteriorate as the position of the Bragg peak gets closer to the heterogeneity (cases 2 and 3). The worse performance metrics could be due to finer voxel resolution in animals compared to humans, rendering it more sensitive to very small shifts and changes in the dose between neighbouring voxels.

We also found that the split schemes are not universally applicable and has to be optimized on a case-by-case basis. For example, the 1+6+6+12 scheme obtained a passing rate of 93.4% (DD/DTA: 3%/0.2 mm) for the heterogeneous phantom, but it achieved a very high passing rate of 99.7% for the full plan in the head, where more complex structures are present in the beam path. In contrast, the 1+24+24+24+24 scheme, which achieved a 98.9% passing rate for the heterogeneous phantom, showed poorer outcomes for the thorax and abdomen cases. Furthermore, comparisons of the 1+6+6+12 and 1+24+24+24+24 split schemes in table 3.3 revealed only a slight improvement in the gamma passing rate for full plan calculations, whereas substantial improvement was observed for the heterogenous phantom case in table 3.2. These tests on the heterogeneous phantom and on different anatomical sites in a mouse indicate that the effectiveness of the split scheme depends on what local inhomogeneities are present in the beam path and where they are located with respect to the beam. It is also important to point out that the ring radii, spatial spreads, and weights of the beamlets for each split scheme were specifically optimized for the first heterogeneous phantom case. While the optimized parameters worked relatively well on the head and abdomen with a lateral beam, their effectiveness deteriorated on the thorax and abdomen with an anterior field as well as on the other heterogeneous phantom cases. The accuracy of the YODA results could be further improved by optimizing the split scheme parameters for each individual case.

To facilitate the use of YODA in practice, it is essential to develop site-specific protocols or class solutions. Detailed guidelines and recommendations for optimized split schemes tailored to different anatomical regions and for a range of possible beam configurations must be established. This approach will avoid additional time spent for optimization during the irradiations and minimize the need for user intervention. Given that animals used in a single experiment are clones of each other, the optimization only needs to be study-specific rather than animal-specific, simplifying the overall process. With proper tuning, YODA can be effectively used for real-time dose calculations in animals especially in anatomical regions that are relatively homogeneous and without critical

heterogeneities such as the head. While discrepancies were observed in more complex areas like the thorax and abdomen, these were generally not dosimetrically significant for the PTV. However, it is important to note that some degree of dose overestimation or underestimation may occur, particularly in tissues located distally. Given that YODA is considerably faster than MCsquare, accurately predicts proton range (given that beam splitting is properly implemented), and only exhibits dose degradation at the distal end, it may be beneficial to use YODA for plan optimization and reserve MCsquare for the final dose calculations. Overall, YODA proves to be a valuable tool for quickly evaluating dose distributions and aiding in treatment planning considerations for animals, provided that pre-optimized split schemes are established.

While previous works have focused on reducing the contouring time in an effort to implement a faster preclinical workflow [40, 41], this work aims to address another bottleneck in proton irradiation of small animals: the dose calculation. To accelerate dose calculations for kV X-ray beams, van Dijk et al. [22] proposed a deep learning model for denoising low statistics MC dose distributions. The dose calculation then becomes a two-step process, requiring an MC simulation run with fewer initial particle histories ($N = 10^6$) and subsequently, feeding the resulting noisy dose map and the dose uncertainty map to the deep learning model to predict its equivalent dose distribution with less statistical uncertainty. The entire process took 6 s for their test case. For protons in the preclinical context, Vanstalle et al. [24] proposed an analytical method to calculate dose distributions in small animals. While analytical solutions tend to be faster than tracking simulations, the runtime for their proposed algorithm was not evaluated in their work. Benchmarking of their method against a full MC simulation was performed only on a simple water phantom and more rigorous testing on a heterogeneous phantom with a material interface parallel to the beam as well as calculation of 3D dose distributions in realistic animal geometries, as evaluated in our work, was not conducted. Almeida et al. [42] also performed proton dose calculations on a mouse with lung tumour, utilizing a clinical beam and an adaptive aperture to create small fields. They simulated both shoot-through beams and single Bragg peaks delivered at different angles, with the dose calculations performed in TOPAS. While they reported the estimated treatment times, they did not provide information regarding the duration of the dose calculations.

To our knowledge, this is the first study demonstrating the capabilities of MCsquare and YODA for accelerating dose calculations for small animal proton irradiations. Although we performed rigorous validation in phantom and animal geometries, further testing on real micro-CT images of animals and using more complex beam configurations (i.e. multiple fields) should also be conducted. While this work only focused on

CPU-based proton dose engines, another potential solution to be explored in the future is GPU-based MC codes such as the dose engine of the preclinical commercial treatment planning system, μ-RayStation. GPU MC approaches can predict dose distributions faster than CPU-based MC codes [18], which can further boost animal throughput.

3.5. CONCLUSION

This study focused on evaluating the accuracy and time efficiency of the fast CPU-based Monte Carlo code MCsquare and the deterministic dose engine YODA for real-time proton dose calculations in small animals. Validation was first performed in water, wherein excellent agreement with TOPAS was achieved by both codes while significantly reducing the computing time. In all cases investigated, MCsquare demonstrated the ability to perform rapid dose calculations, produce reasonably accurate results, and handle more complex anatomical regions, making it suitable for dose computations in routine preclinical practice. Although YODA is considerably faster than MCsquare, its accuracy relies on the beam splitting scheme needed to properly account for local inhomogeneities. Evaluations on a heterogeneous phantom and CT-based animal geometries revealed that the effectiveness of the split schemes depend on the distribution of local inhomogeneities and their position relative to the beam. This emphasizes the importance of pre-tuning split schemes for different anatomical sites and beam configurations to facilitate YODA's seamless use in daily practice. Despite these considerations, YODA can accurately predict the proton range and only exhibits dose differences towards the distal end. To capitalize on YODA's speed, we suggest using it for plan optimization while relying on MCsquare for the final dose calculation to ensure accuracy.

ACKNOWLEDGEMENTS

This work was partially financially supported by a grant from the Dutch Cancer Society (KWF research project 12092).

BIBLIOGRAPHY

[1] Jonathan R. Hughes and Jason L. Parsons. “FLASH Radiotherapy: Current Knowledge and Future Insights Using Proton-Beam Therapy”. In: *International Journal of Molecular Sciences* 21.18 (2020). DOI: 10.3390/ijms21186492.

[2] Michele M. Kim et al. “Development of Ultra-High Dose-Rate (FLASH) Particle Therapy”. In: *IEEE Transactions on Radiation and Plasma Medical Sciences* 6.3 (2022), pp. 252–262. DOI: 10.1109/TRPMS.2021.3091406.

[3] Y. Prezado and G. R. Fois. “Proton-minibeam radiation therapy: A proof of concept”. In: *Medical Physics* 40.3 (2013), p. 031712. DOI: 10.1118/1.4791648.

[4] Yolanda Prezado et al. “Spatially fractionated radiation therapy: a critical review on current status of clinical and preclinical studies and knowledge gaps”. In: *Physics in Medicine and Biology* 69.10 (2024), 10TR02. DOI: 10.1088/1361-6560/ad4192.

[5] E Ford et al. “An image-guided precision proton radiation platform for preclinical in vivo research”. In: *Physics in Medicine and Biology* 62.1 (2016), p. 43. DOI: 10.1088/1361-6560/62/1/43.

[6] Michele M Kim et al. “Design and commissioning of an image-guided small animal radiation platform and quality assurance protocol for integrated proton and x-ray radiobiology research”. In: *Physics in Medicine and Biology* 64.13 (2019), p. 135013. DOI: 10.1088/1361-6560/ab20d9.

[7] Katia Parodi et al. “Towards a novel small animal proton irradiation platform: the SIRMIO project”. In: *Acta Oncologica* 58.10 (2019), pp. 1470–1475. DOI: 10.1080/0284186X.2019.1630752.

[8] Moritz Schneider et al. “SAPPHIRE—establishment of small animal proton and photon image-guided radiation experiments”. In: *Physics in Medicine and Biology* 69.9 (2024), p. 095020. DOI: 10.1088/1361-6560/ad3887.

[9] Frank Verhaegen and Joao Seco, eds. *Monte Carlo Techniques in Radiation Therapy: Introduction, Source Modelling, and Patient Dose Calculations (2nd ed.)* CRC Press, 2021. DOI: 10.1201/9781003211846.

[10] Frank Verhaegen and Joao Seco, eds. *Monte Carlo Techniques in Radiation Therapy: Applications to Dosimetry, Imaging, and Preclinical Radiotherapy (2nd ed.)* CRC Press, 2021. DOI: 10.1201/9781003212485.

[11] Linda Hong et al. "A pencil beam algorithm for proton dose calculations". In: *Physics in Medicine and Biology* 41.8 (1996), p. 1305. DOI: 10.1088/0031-9155/41/8/005.

[12] Barbara Schaffner, Eros Pedroni, and Antony Lomax. "Dose calculation models for proton treatment planning using a dynamic beam delivery system: an attempt to include density heterogeneity effects in the analytical dose calculation". In: *Physics in Medicine and Biology* 44.1 (1999), p. 27. DOI: 10.1088/0031-9155/44/1/004.

[13] Jatinder Saini et al. "Advanced proton beam dosimetry part I: review and performance evaluation of dose calculation algorithms". In: *Translational Lung Cancer Research* 7.2 (2018). DOI: 10.21037/tlcr.2018.04.05.

[14] Kevin Souris, John Aldo Lee, and Edmond Sterpin. "Fast multipurpose Monte Carlo simulation for proton therapy using multi- and many-core CPU architectures". In: *Medical Physics* 43.4 (2016), pp. 1700–1712. DOI: 10.1118/1.4943377.

[15] Xun Jia et al. "GPU-based fast Monte Carlo dose calculation for proton therapy". In: *Physics in Medicine and Biology* 57.23 (2012), p. 7783. DOI: 10.1088/0031-9155/57/23/7783.

[16] Jan Gajewski et al. "Commissioning of GPU–Accelerated Monte Carlo Code FRED for Clinical Applications in Proton Therapy". In: *Frontiers in Physics* 8 (2021). DOI: 10.3389/fphy.2020.567300.

[17] Hoyeon Lee et al. "MOQUI: an open-source GPU-based Monte Carlo code for proton dose calculation with efficient data structure". In: *Physics in Medicine and Biology* 67.17 (2022), p. 174001. DOI: 10.1088/1361-6560/ac8716.

[18] Francesco Fracchiolla et al. "Clinical validation of a GPU-based Monte Carlo dose engine of a commercial treatment planning system for pencil beam scanning proton therapy". In: *Physica Medica* 88 (2021), pp. 226–234. DOI: 10.1016/j.ejmp.2021.07.012.

[19] Martin Janson et al. "Treatment planning of scanned proton beams in RayStation". In: *Medical Dosimetry* 49.1 (2024), pp. 2–12. DOI: 10.1016/j.meddos.2023.10.009.

[20] Tiberiu Burlacu, Danny Lathouwers, and Zoltán Perkó. "A Deterministic Adjoint-Based Semi-Analytical Algorithm for Fast Response Change Computations in Proton Therapy". In: *Journal of Computational and Theoretical Transport* 52.1 (2023), pp. 1–41. DOI: 10.1080/23324309.2023.2166077.

[21] Oscar Pastor-Serrano and Zoltán Perkó. "Millisecond speed deep learning based proton dose calculation with Monte Carlo accuracy". In: *Physics in Medicine and Biology* 67.10 (2022), p. 105006. DOI: 10.1088/1361-6560/ac692e.

[22] Robert H W van Dijk et al. "A novel multichannel deep learning model for fast denoising of Monte Carlo dose calculations: preclinical applications". In: *Physics in Medicine and Biology* 67.16 (2022), p. 164001. DOI: 10.1088/1361-6560/ac8390.

[23] Florian Mentzel et al. "Accurate and Fast Deep Learning Dose Prediction for a Preclinical Microbeam Radiation Therapy Study Using Low-Statistics Monte Carlo Simulations". In: *Cancers* 15.7 (2023). DOI: 10.3390/cancers15072137.

[24] Marie Vanstalle et al. "Analytical dose modeling for preclinical proton irradiation of millimetric targets". In: *Medical Physics* 45.1 (2018), pp. 470–478. DOI: 10.1002/mp.12696.

[25] Thomas Bortfeld and Wolfgang Schlegel. "An analytical approximation of depth - dose distributions for therapeutic proton beams". In: *Physics in Medicine and Biology* 41.8 (1996), p. 1331. DOI: 10.1088/0031-9155/41/8/006.

[26] Monika Clausen et al. "Small field proton irradiation for in vivo studies: Potential and limitations when adapting clinical infrastructure". In: *Zeitschrift für Medizinische Physik* 33.4 (2023), pp. 542–551. DOI: 10.1016/j.zemedi.2022.10.002.

[27] Tiberiu Burlacu, Danny Lathouwers, and Zoltán Perkó. "Yet anOther Dose Algorithm (YODA) for independent computations of dose and dose changes due to anatomical changes". In: *Physics in Medicine and Biology* 69.16 (2024), p. 165003. DOI: 10.1088/1361-6560/ad6373.

[28] J. Perl et al. "TOPAS: An innovative proton Monte Carlo platform for research and clinical applications". In: *Medical Physics* 39.11 (2012), pp. 6818–6837. DOI: 10.1118/1.4758060.

[29] Matthias Fippel and Martin Soukup. "A Monte Carlo dose calculation algorithm for proton therapy". In: *Medical Physics* 31.8 (2004), pp. 2263–2273. DOI: 10.1118/1.1769631.

[30] Harald Paganetti. “Nuclear interactions in proton therapy: dose and relative biological effect distributions originating from primary and secondary particles”. In: *Physics in Medicine and Biology* 47.5 (2002), p. 747. DOI: 10.1088/0031-9155/47/5/305.

[31] Bruno Rossi and Kenneth Greisen. “Cosmic-Ray Theory”. In: *Reviews of Modern Physics* 13.4 (1941), pp. 240–309. DOI: 10.1103/RevModPhys.13.240.

[32] ICRU. *Nuclear Data for Neutron and Proton Radiotherapy and for Radiation Protection*. Tech. rep. International Commission on Radiation Units and Measurements, 2000. URL: https://www.icru.org/report/nuclear-data-for-neutron-and-proton-radiotherapy-and-for-radiation-protection-report-63/.

[33] Martin Berger. *Penetration of proton beams through water I. Depth-dose distribution, spectra and LET distribution*. Tech. rep. National Institute of Standards and Technology, 1993. URL: https://nvlpubs.nist.gov/nistpubs/Legacy/IR/nistir5226.pdf.

[34] L.J. Nevay et al. “BDSIM: An accelerator tracking code with particle–matter interactions”. In: *Computer Physics Communications* 252 (2020), p. 107200. DOI: 10.1016/j.cpc.2020.107200.

[35] William.P. Segars et al. “Development of a 4-D digital mouse phantom for molecular imaging research”. In: *Molecular Imaging and Biology* 6.3 (2004), pp. 149–159. DOI: 10.1016/j.mibio.2004.03.002.

[36] S. Wuyckens et al. *OpenTPS – Open-source treatment planning system for research in proton therapy*. 2023. arXiv: 2303.00365 [physics.med-ph]. URL: https://arxiv.org/abs/2303.00365.

[37] Chris D Johnstone et al. “Multi-institutional MicroCT image comparison of image-guided small animal irradiators”. In: *Physics in Medicine and Biology* 62.14 (2017), p. 5760. DOI: 10.1088/1361-6560/aa76b4.

[38] Ana Vaniqui et al. “The effect of different image reconstruction techniques on pre-clinical quantitative imaging and dual-energy CT”. In: *British Journal of Radiology* 92.1095 (2018), p. 20180447. DOI: 10.1259/bjr.20180447.

[39] Bernard Gottschalk. *Techniques of Proton Radiotherapy: Transport Theory*. 2012. arXiv: 1204.4470 [physics.med-ph]. URL: https://arxiv.org/abs/1204.4470.

[40] Justin Malimban et al. “Deep learning-based segmentation of the thorax in mouse micro-CT scans”. In: *Scientific Reports* 12.1822 (2022). DOI: 10.1038/s41598-022-05868-7.

[41] Georgios Lappas et al. “Automatic contouring of normal tissues with deep learning for preclinical radiation studies”. In: *Physics in Medicine and Biology* 67.4 (2022), p. 044001. DOI: 10.1088/1361-6560/ac4da3.

[42] Isabel P Almeida et al. “Exploring the feasibility of a clinical proton beam with an adaptive aperture for pre-clinical research”. In: *British Journal of Radiology* 92.1095 (2018), p. 20180446. DOI: 10.1259/bjr.20180446.

4

A SIMULATION FRAMEWORK FOR THE PRECLINICAL PROTON IRRADIATION WORKFLOW

Justin Malimban[1], Felix Ludwig[1], Danny Lathouwers[2], Marius Staring[3,4], Frank Verhaegen[5] and Sytze Brandenburg[1]

[1]Department of Radiation Oncology and Particle Therapy Research Center, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
[2]Department of Radiation Science and Technology, Faculty of Applied Sciences, Delft University of Technology, Delft, The Netherlands
[3]Department of Radiology, Leiden University Medical Center, Leiden, The Netherlands
[4]Department of Radiation Oncology, Leiden University Medical Center, Leiden, The Netherlands
[5]Department of Radiation Oncology (MAASTRO), Research Institute for Oncology & Reproduction, Maastricht University Medical Centre, Maastricht, The Netherlands

Under revision in Physics in Medicine and Biology

ABSTRACT

Objective: The integration of proton beamlines with X-ray imaging/irradiation platforms has opened up possibilities for image-guided Bragg peak irradiations in small animals. Such irradiations allow selective targeting of normal tissue substructures and tumours. However, their small size and location pose challenges in designing experiments. This work presents a simulation framework useful for optimizing beamlines, imaging protocols, and design of animal experiments. The usage of the framework is demonstrated, mainly focusing on the imaging part.

Approach: The fastCAT toolkit was modified with Monte Carlo (MC)-calculated primary and scatter data of a small animal imager for the simulation of micro-CT scans. The simulated CT of a mini-calibration phantom from fastCAT was validated against a full MC TOPAS CT simulation. A realistic beam model of a preclinical proton facility was obtained from beam transport simulations to create irradiation plans in matRad. Simulated CT images of a digital mouse phantom were generated using single-energy CT (SECT) and dual-energy CT (DECT) protocols and their accuracy in proton stopping power ratio (SPR) estimation and their impact on calculated proton dose distributions in a mouse were evaluated.

Main Results: The CT numbers from fastCAT agree within 11 HU with TOPAS except for materials at the center of the phantom. Discrepancies for central inserts are caused by beam hardening issues. The root mean square deviation in the SPR for the best SECT (90kV/Cu) and DECT (50kV/Al–90kV/Al) protocols are 3.7% and 1.0%, respectively. Dose distributions calculated for SECT and DECT datasets revealed range shifts < 0.1 mm, gamma pass rates (3%/0.1mm) greater than 99%, and no substantial dosimetric differences for all structures. The outcomes suggest that SECT is sufficient for proton treatment planning in animals.

Significance: The framework is a useful tool for the development of an optimized experimental configuration without using animals and beam time.

Keywords: micro-CT simulation, single-energy CT, dual-energy CT, stopping power, preclinical proton dose calculations, Monte Carlo

4.1. INTRODUCTION

Over the past decade, there has been a rapid increase in the number of proton therapy facilities around the world [1]. Owing to the characteristic dose fall off after the Bragg peak, proton therapy can deliver a much more conformal dose to the tumour, thereby more effectively sparing surrounding healthy tissues. To fully realize the potential and exploit the benefits of proton therapy, preclinical studies are needed to better understand the biological mechanisms of tumour and normal tissue response and to study differential effects of X-rays and protons.

Several *in vivo* studies have already been conducted to investigate the interplay between radiation induced damage to the heart and lungs [2] as well as inhibition of repair mechanisms by low dose irradiations around the primary high dose areas in the spinal cord [3]. Previous works have also looked into the relative biological effectiveness (RBE) by irradiating mice with different parts of the Bragg curve [4, 5, 6]. Other treatment strategies are also being explored such as using an array of narrow, spatially fractionated beams—proton minibeam [7, 8] and ultra-high dose rates—FLASH [9, 10, 11] to promote normal tissue sparing. Targeted irradiations of critical normal tissue structures such as stem cell-rich ducts in the parotid glands [12] and hippocampus [13] also suggest that sparing these regions may reduce associated radiation-induced side effects (i.e. xerostomia and neurocognitive dysfunction, respectively).

To get more insight into the radiation response of both normal tissue structures and tumours, highly accurate irradiations must be performed, which can be challenging due to motion, location, and size of the targets in small animals. To create opportunities for these experiments, facilities have started to integrate preclinical X-ray CT imaging and irradiation platforms with proton beamlines to provide the image guidance needed to achieve the required accuracy [14, 15, 16, 17]. This opens up possibilities for delivering spread-out Bragg peak (SOBP) irradiations, which allow a much better conformity of the dose distributions to the targets. However, designing such experiments is not straightforward. To achieve highly conformal dose distributions, millimetre-sized pencil beams must be delivered. The beam must also be degraded down to energies much lower than a clinical proton beam (~30 MeV). These require additional components such as collimators and range shifters to be integrated into the beamline. The material chosen and the position of these components with respect to the animal affect the quality of the dose distributions. Furthermore, correct positioning of the Bragg peaks must be ensured and the actual dose distribution in the animal must be accurately determined.

In this work, we present a simulation framework of the preclinical irradiation workflow, which can be used to optimize beam properties, evaluate imaging protocols, and

characterize dose distributions prior to performing experiments. The framework provides a quick way to discover sensitivities and weak points of experimental setups so they can be addressed beforehand. This leads to a more efficient workflow and also enhances experiment quality and capacity.

The framework allows generation of realistic X-ray micro-CT images using the fastCAT CBCT simulator [18]. We have modified the original code to enable creation of CT scans consistent with a small animal imager. This was validated against a full Monte Carlo (MC) X-ray CT simulation in TOPAS [19]. Beam transport simulations in the Geant4-based beam delivery simulation (BDSIM) toolkit [20] were performed to obtain a realistic proton beam model optimized for small animal irradiations. The proton beam model was used to generate the beam data library in matRad [21] for treatment planning.

We focus on the imaging issues to demonstrate an application of this framework. Single-energy CT (SECT) and dual-energy CT (DECT) calibration methods have already been investigated for preclinical studies but only for X-ray irradiations [22, 23]. Here, we extend that to protons by evaluating different published SECT and DECT approaches for estimation of proton stopping power ratios (SPR) and material identification in animal CTs. The uncertainties associated with the calibration methods are assessed by simulating proton radiographs to get range error maps. Their impact on proton dose distributions was also evaluated to establish whether the differences observed between SECT and DECT were large enough to warrant DECT-based proton treatment planning for animals.

4.2. Materials and Methods

4.2.1. Micro-CBCT model for fastCAT

FastCAT is a cone-beam CT (CBCT) simulation toolkit relying on scatter kernels and detector response functions pre-calculated using Monte Carlo (MC) simulations [18]. The X-ray source is modelled using SpekPy, which is a python-based program that allows calculation of polychromatic X-ray spectra for a wide range of X-ray tube specifications [24]. For kilovoltage beams, which are typically used for small animal CT acquisition, fastCAT offers a 450-μm thick Cesium Iodide (CsI) detector. To create CT images, a 3D voxel geometry of the object is imported in the form of a matrix with integers corresponding to the material index. For each of these materials, the linear attenuation coefficients (μ) as a function of energy taken from the NIST XCOM[1] database is used as input in the sim-

[1] NIST XCOM: Photon Cross Sections Database is a web program that can be used to calculate photon cross sections and attenuation coefficients of elements, compounds, or mixtures. Access at: https://physics.nist.gov/PhysRefData/Xcom/html/xcom1.html

Table 4.1: Parameters used in fastCAT based on the X-RAD 225Cx machine.

Parameter	Value
Focal spot	0.4 mm based on IEC 336
Inherent filtration	0.8 mm Be
Anode	Tungsten (W)
Anode angle	20°
Source-to-axis distance (SAD)	303.4 mm
Source-to-detector distance (SDD)	622.9 mm
Number of projections	360

ulation. The CT parameters such as source-to-axis distance (SAD), source-to-detector distance (SDD), and imaging dose must also be defined. Then, projection images are created using raytracing, which are subsequently used for CT reconstruction in TIGRE [25].

In this way, fastCAT is capable of generating realistic CT images in minutes. However, in its original form it is not suitable for micro-CT simulations since its base data (i.e. primary and scatter kernels) were modelled after a clinical CBCT machine, and the phantoms available are in the human scale. Therefore, we performed separate MC simulations in TOPAS to generate base data for micro-CT acquisitions using the X-RAD 225Cx (Precision X-Ray Inc., Madison, CT) as the reference CT model. Table 4.1 gives the CT parameters for this small animal imager.

The following subsections detail the changes we have made in fastCAT and the Monte Carlo validation of the modified version. The physics list and range cuts used in TOPAS can be found in table S4.1 in the supplementary file.

PRIMARY AND SCATTERED X-RAY MODELLING IN TOPAS

CT images in fastCAT are created by first calculating forward projections using raytracing in TIGRE for 18 discrete energies (i.e. 10 to 100 keV in increments of 10 keV, 300 to 900 keV in increments of 200 keV, and 1, 2, 4 and 6 MeV). Each projection is turned into an intensity image by using the primary field and then, the scatter contribution is subsequently added. The final intensity image is obtained by weighting the 18 projection images by the X-ray spectrum and energy deposition efficiency.

We calculated the primary field and scatter contribution for our micro-CT model following the simulation procedure described in the paper of O'Connell and Bazalova-Carter [18] but using the CT parameters given in table 4.1. The primary field was obtained by irradiating the detector without an object. The source was modelled as a cone

beam with a Gaussian focal spot size with σ = 0.4 mm. The *SurfaceTrackCount* scorer in TOPAS was used to count the X-rays incident on a 2D air slab (512 × 512, pixel size = 0.15 mm). A radial profile of the X-ray intensity was taken and the same curve as in the original fastCAT paper was fitted on the resulting profile. Since the shape of the primary field does not change much with energy, the simulation was only performed for 90 keV X-rays and the same profile was used for the other energies.

The phantom-specific scatter kernels were generated for a spherical water phantom with radius of 15 mm for imaging a mouse. These were calculated for all energies required by fastCAT. Similar to the primary field, a radial profile exists for each energy. The scatter contribution is assumed to be independent of the projection angle.

IMAGE NOISE

FastCAT applies Poisson noise to the final intensity image to create a more realistic CT. To generate CTs at different noise levels (or dose levels), fastCAT scales the raw intensity and the noise using the ratio of the user-defined total particle fluence and the reference value (i.e. number of particles for which the base data was calculated for in the MC simulation). Here, the dose in a spherical water phantom was scored using the *DoseToMedium* scorer in TOPAS for a given number of X-rays. This gives a relation that allows to assess image noise as a function of imaging dose.

PHANTOMS

Two types of phantoms were used in this work: a mouse-sized CT calibration phantom (SmART Scientific Solutions BV, Maastricht, the Netherlands) and the mathematical MOBY mouse phantom [26]. The mini-calibration phantom, which contains tissue-equivalent materials of known elemental composition (supplementary table S4.2), was used for the CT HU calibration. It has a diameter of 30 mm and holds 11 cylindrical tissue-equivalent inserts (Gammex Inc., WI, USA) with a diameter of 3.5 mm. Figure 4.1 shows the schematic diagram of the phantom, and table 4.2 gives the corresponding material properties.

The MOBY digital phantom was used to generate simulated CT images of the head and thorax of a mouse as shown in figure 4.2. Artificial spherical tumours were created in the brain and lungs with diameters 2 mm and 3 mm, respectively. The material assignment for each organ is given in table 4.3, and the corresponding elemental composition can be found in supplementary table S4.3.

Using the known material compositions of the phantoms, a ground truth image was created that was used for the evaluation of the CT-based simulations. For example, dose distributions calculated on CT images were validated against the dose computed for an

Table 4.2: Effective atomic number $\left(Z_{eff} = \sqrt[m]{\frac{\sum_i w_i \frac{Z_i}{A_i} Z_i^m}{\sum_i w_i \frac{Z_i}{A_i}}}, m = 3.3\right)$ [27], relative electron density (ρ_e), and stopping power ratio (SPR) relative to water (calculated using the Bethe Bloch equation for 100 MeV protons) of the Gammex materials. Values for the lung insert are not given as it is a highly heterogeneous material.

#	**Material**	Z_{eff}	ρ_e	SPR
1	Air	7.71	0.001	0.001
2	LN-450 Lung	-	-	-
3	AP6 Adipose	6.21	0.928	0.947
4	SR2 Brain	6.09	1.047	1.075
5	BR12 Breast	6.93	0.956	0.972
6	Solid Water	7.74	0.992	1.005
7	LV1 Liver	7.74	1.064	1.078
8	IB3 Inner Bone	10.42	1.086	1.082
9	B200 Bone	10.42	1.103	1.099
10	CB2-30% CaCO3	10.90	1.276	1.270
11	CB2-50% CaCO3	12.54	1.469	1.436
12	SB3 Cortical Bone	13.64	1.695	1.631

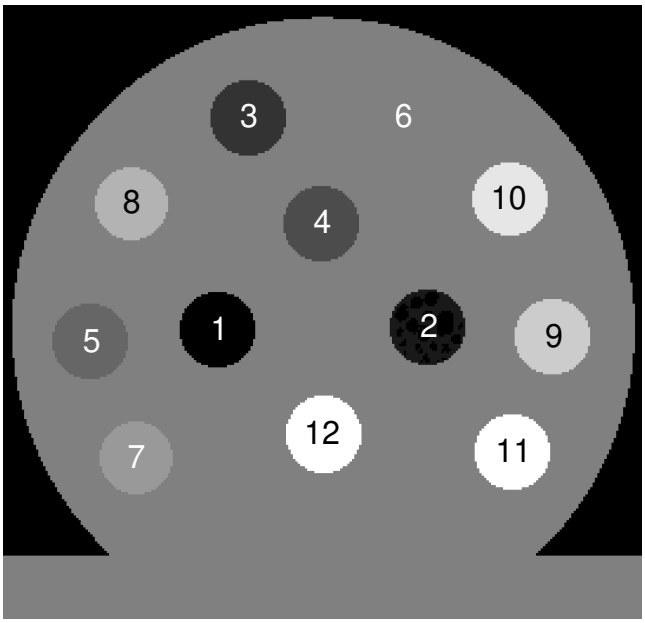


Figure 4.1: Schematic diagram of the mini-calibration phantom. The numbers correspond to materials in table 4.2.

image in which the actual materials are assigned.

VALIDATION OF X-RAY CT IMAGING WITH TOPAS

A full Monte Carlo (MC) CBCT simulation of the mini-calibration phantom, including X-ray scattering, was performed in TOPAS to validate the micro-CT model in fastCAT. The same CT geometry was used with the voxelized model of the mini-calibration phantom as the object and the energy spectrum of 90 kVp X-rays filtered with 2 mm aluminium taken from SpekPy as the source. Projection images were taken at 1° interval over a 360° rotation for 9×10^8 X-rays per projection. To compare fastCAT and TOPAS, the mean Hounsfield Units (HU) of Gammex materials were extracted from a region of interest (ROI) with a diameter of 2.5 mm centred at each rod averaged over 5 mm thickness at the centre of the phantom body.

4.2.2. X-RAY CT-BASED CONVERSION TO PROTON STOPPING POWER

Figure 4.3 shows the workflow developed to perform the HU to SPR calibration from which material and density assignments were derived for the Monte Carlo simulations

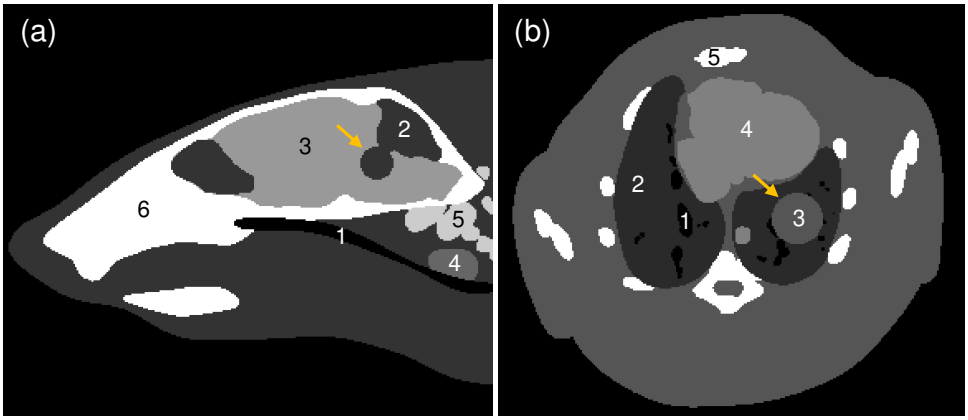


Figure 4.2: (a) Head and (b) thorax of the MOBY phantom. The numbers correspond to the materials assigned to each organ given in table 4.3. The spherical tumours with diameters 2 mm (brain) and 3 mm (lung) artificially added in the phantom are indicated by the yellow arrow.

of proton irradiations. First, simulated CT images of the mini-calibration phantom were generated with fastCAT using imaging protocols typically used for small animal imaging. Scans at tube potentials 50 kVp and 90 kVp filtered with 2 mm aluminium were created as it was determined to be the best preclinical DECT combination in a previous study [22]. We also looked into using stronger filtration for the high kV beam to reduce the overlap between the low and high kV spectra for DECT (supplementary figure S4.1). For this, copper with thickness of 0.32 mm was used, which is a filter also available in the X-RAD 225Cx machine. The 50 kVp + 2mm Al, 90kVp + 2 mm Al, and 90 kVp + 0.32mm Cu protocols will be referred to as 50kV/Al, 90kV/Al, and 90kV/Cu, respectively, for the rest of the paper.

The mean HU values of 10 Gammex materials (i.e. materials 3-12 in table 4.2) given in supplementary table S4.4 were used to fit parameters for SECT/DECT-to-SPR coversion algorithms. The values were obtained for the same ROI as described in the previous section. The LN-450 lung rod was excluded because it spanned a large HU range due to the clearly noticeable air pockets in the CT images.

SINGLE-ENERGY CT (SECT)

Schneider's stoichiometric calibration method was used to obtain the HU-to-SPR conversion for SECT [31]. The fit parameters k_1 and k_2 were derived using least squares fitting on the Gammex tissue substitutes for the three imaging protocols implemented in this study. The *scipy.optimize.least_squares* function in python was used for the mini-

Table 4.3: Material assignment to organs in the head and thorax of the MOBY phantom. Materials 6-8 in the thorax are not shown in the image slice in figure 4.2(b). The elemental composition of each material is given in supplementary table S4.3.

#	Head		Thorax	
	Organs	**Material Assignment**	**Organs**	**Material Assignment**
1	Esophagus	Air[a]	Airsacs	Air[a]
2	Body, cerebellum, olfactory bulb, tumour	Soft tissue[a]	Lung	Lung-inflated[c]
3	Cerebral cortex, brainstem, striatum, rest of the brain	Brain[b]	Body, tumour	Soft tissue[a]
4	Thyroid	Thyroid[c]	Heart	Heart[b]
5	Spinal cord	Vertebral column (C4)[d]	Ribs, spine, bones	Ribs (2nd, 6th)[d]
6	Skull	Cranium[d]	Liver	Liver[b]
7			Stomach	Stomach[c]
8			Gall bladder	Gall bladder[b]

[a]Geant4 material database (https://www.fe.infn.it/u/paterno/Geant4_tutorial/slides_further/Geometry/G4_Nist_Materials.pdf), [b]ICRP Publication 110 [28], [c]Woodard and White [29], [d]White et al. [30]

mization, and the fitting was constrained to positive k-values to avoid physically meaningless parameters. To check our fit procedure, the measured HU values of different phantom materials given in Schneider et al. [31] were also analysed, resulting in the same k-values as reported by them.

Once k-values were determined, the CT numbers of reference human tissues given in Woodard and White [29] and White et al. [30] were calculated. Their corresponding SPR for 100 MeV protons were calculated using the Bethe-Bloch equation given below

$$SPR = \rho_e \frac{\ln\left(\frac{2m_e c^2 \beta^2}{I(1-\beta^2)}\right) - \beta^2}{\ln\left(\frac{2m_e c^2 \beta^2}{I_w(1-\beta^2)}\right) - \beta^2} \tag{4.1}$$

where ρ_e is the relative electron density, m_e is the electron mass, c is the speed of light, $\beta = \frac{v}{c}$ is ratio between the speed of the proton and speed of light ($\beta = 0.428$ for 100 MeV protons), I is the mean excitation energy of the material calculated using the Bragg additivity rule, and I_w is the mean excitation energy of water taken as 78 eV [32]. Then, the calibration curve was created by performing linear fits between the predicted HU and SPR on the lung, soft tissue, and bone regions. The k-values resulting from the stoichiometric fit and the root-mean-square deviation of the predicted CT numbers of Gammex materials for the three scan protocols are given in table 4.4, and an example of the SECT calibration curve is shown in figure 4.4.

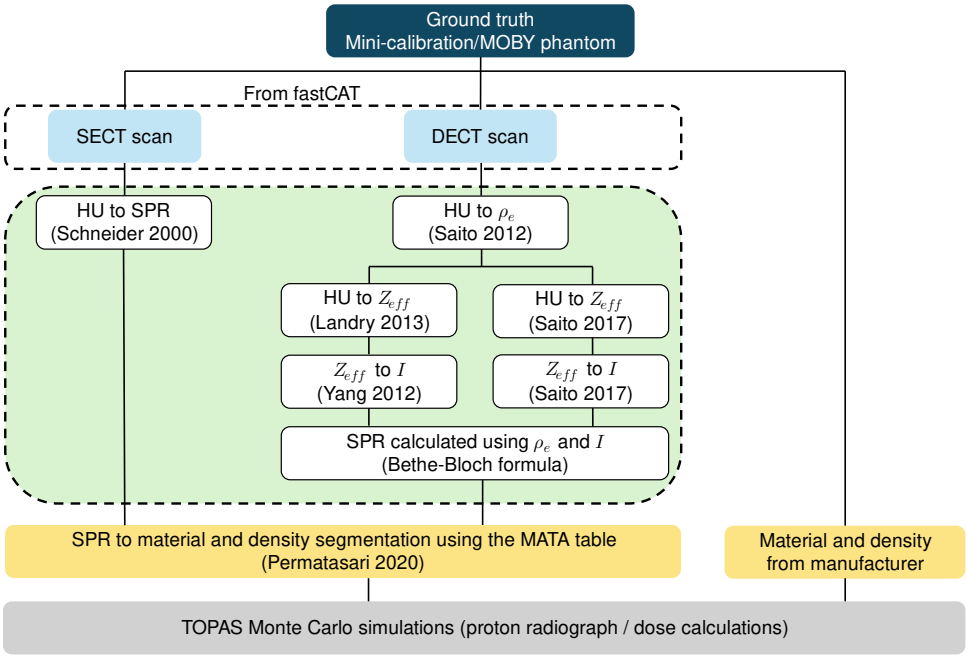


Figure 4.3: Overview of the simulation framework used in this work. Simulated CT scans of preclinical phantoms were generated with fastCAT. The CT scans are then converted to SPR images through SECT and DECT calibration methods. The MATA table is used to convert SPR datasets to material and density maps, which is the input required for Monte Carlo simulations. Comparisons can be made with the ground truth image, which is the phantom with tissue composition used to generate the CT scans.

DUAL-ENERGY CT (DECT)

The 50kV/Al – 90 kV/Al and 50kV/Al – 90 kV/Cu DECT image pairs were used to extract the relative electron density (ρ_e) and effective atomic number (Z_{eff}). The Z_{eff} was subsequently used to derive the mean excitation energy (I), which, together with ρ_e, is a quantity needed for the calculation of the SPR. Saito's [33] approach was used to derive the ρ_e from the weighted subtraction of the low and high kV CT numbers. Two different methods were employed to extract the Z_{eff} from the DECT scans. The first is Landry et al.'s [27] model, which has been implemented in previous small animal DECT studies for X-ray irradiation [22, 23]. It uses the ratio of the attenuation coefficients at low and high X-ray energies $\left(\mu_{low}^{high}\right)$ to obtain the Z_{eff} image. Similar to the stoichiometric method, the fit coefficients were obtained using the *scipy.optimize.least_squares* function, wherein the initial estimates for the μ_{low}^{high} was taken from the NIST XCOM database for water at the effective energies of the X-ray spectra (i.e. 25.5 keV, 30.8 keV, and 49.7 keV

Table 4.4: Fit coefficients (k_1,k_2) derived from stoichiometric calibration and root-mean-square deviation ($RMSD_{HU}$) of the predicted CT numbers obtained for Gammex materials for imaging protocols: 50kV/Al, 90kV/Al, and 90kV/Cu

	50kV/Al	90kV/Al	90kV/Cu
k_1	4.33×10^{-14}	1.28×10^{-19}	3.79×10^{-8}
k_2	2.19×10^{-4}	1.23×10^{-4}	8.05×10^{-5}
$RMSD_{HU}$	76	53	12

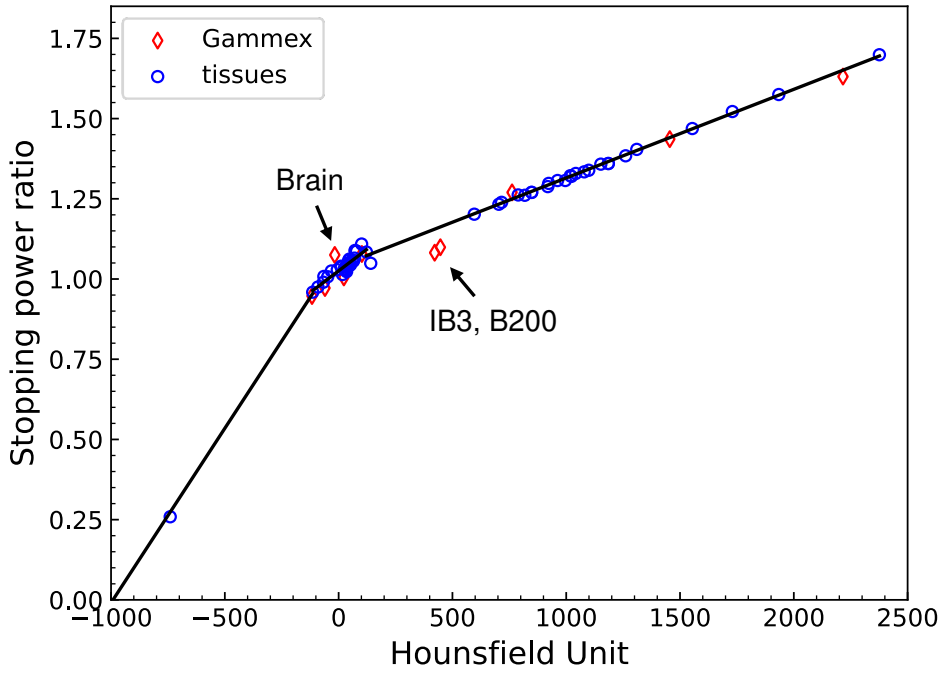


Figure 4.4: Stoichiometric calibration curve for 90kV/Cu. The diamonds correspond to Gammex materials with CT numbers from fastCAT simulation, whereas the circles represent datapoints for reference human tissues with CT numbers calculated from the calibration (i.e. using the k_1 and k_2 parameters).

for 50kV/Al, 90kV/Al, and 90kV/Cu, respectively) and an Z_{eff} of 7.48. This method was combined with Yang et al.'s [34] parameterization of $\ln I$ as a function of Z_{eff}. The second is Saito and Sagara's [35] method, which was found to be the superior DECT model in a recent study [36]. They proposed a simple formulation that relates the Z_{eff} to the

Table 4.5: Fit coefficients obtained from the DECT images of the mini-calibration phantom.

Method	Parameter	50kV/Al – 90 kV/Al	50kV/Al – 90 kV/Cu
Saito (ρ_e)	α	1.68	0.76
	a	0.981	0.976
	b	0.982	1.003
Landry (Z_{eff})	A_{50kV}	1	1
	B_{50kV}	3.923×10^{-3}	1.921×10^{-18}
	C_{50kV}	3.641×10^{-4}	6.508×10^{-4}
	A_{90kV}	1.241	1.402
	B_{90kV}	2.120×10^{-3}	4.586×10^{-21}
	C_{90kV}	2.440×10^{-4}	3.462×10^{-4}
Saito and Sagara (Z_{eff})	γ_L	2.92	2.88

Table 4.6: Fit coefficients for the parameterization of the mean excitation energy (I).

Method	Parameter	Soft tissue	Bone
Yang	a	0.124	0.100
	b	3.377	3.329
Saito and Sagara	c_1	0.316	0.075
	c_0	0.054	0.115

low energy CT numbers and ρ_e. The I-values were calculated using their parameterization of $\ln \frac{I}{I_w}$ as a function of Z_{eff}. For both methods to derive the Z_{eff}, m = 3.3 was used [27]. The fitting methods were validated by using the values from Table I, Table 3, and Table I given in Saito [33], Landry et al. [27], and Saito and Sagara [35], respectively. The largest deviation for ρ_e and Z_{eff} from our calibration are 0.1% and 1.6% from the original values, respectively.

Table 4.5 gives the fit coefficients for each method derived from the DECT scans used in this work. The fit results for Yang et al. [34] and Saito and Sagara's [37] parameterization of I are provided in table 4.6. These coefficients were obtained from fitting of the I and Z_{eff} of reference human tissues [29, 30], which were calculated using Bragg's additivity rule.

SPR TO MATERIAL CONVERSION

To perform Monte Carlo simulations, the elemental composition and mass density of each voxel in the CT are needed instead of the SPR. The SECT- and DECT-based SPR

predictions were converted to composition and density using the MATerial Assignment (MATA) table of Permatasari et al. [38]. The MATA table contains 40 reference human tissues each of which falls in a predefined SPR interval to facilitate material assignment. The density is determined using a linear relationship with the SPR. In this way, calibration-specific methods to generate material and density maps from the SPR dataset are avoided.

SPR ACCURACY

The SPR of Gammex materials predicted from both SECT and DECT calibration were compared to the theoretical values (SPR_t) calculated using the Bethe-Bloch equation for 100 MeV protons (table 4.2). For each insert, the mean SPR (SPR_{mean}) was extracted over a cylindrical ROI (same as previously described) and the deviation was calculated as follows

$$\Delta SPR(\%) = \frac{SPR_{mean} - SPR_t}{SPR_t} \times 100. \tag{4.2}$$

The root mean square deviation ($RMSD$) as given below

$$RMSD(\%) = \sqrt{\frac{\sum_{i=1}^{N} (\Delta SPR_i)^2}{N}} \tag{4.3}$$

was also estimated for each calibration procedure for the N=10 inserts considered in this work.

To get an idea on how the SPRs obtained with the different calibration methods contribute to range errors, proton radiographs of CTs converted to SPR maps using the SECT and DECT methods were calculated and compared to the proton radiograph of the ground truth phantom (i.e. geometry with actual elemental composition assigned). Only the SECT (90kV/Cu) and DECT (50kV/Al – 90kV/Al) protocols that showed the least SPR deviation for Gammex materials were used in the comparison. The MOBY thorax was chosen because it is a highly heterogeneous region with air/soft tissue/bone interfaces. The simulated proton radiographs were obtained in TOPAS by irradiating the CT with a 100 MeV parallel proton beam and scoring the residual kinetic energy of protons (KE_{res}) exiting the volume on a 2D air slab placed directly behind it.

A separate simulation was also performed to obtain the calibration curve relating the residual kinetic energy to the water equivalent thickness (WET). Following the same simulation setup, the (KE_{res}) was scored for water slabs of increasing thickness in increments of 1 mm. The average value in a 1 cm region of interest at the centre of the scorer was used for the WET calibration. The resulting calibration curve is given in supplementary figure S4.2.

The final analysis is expressed in terms of the range error, which is quantified as the pixel-by-pixel difference of the WET map of the calibrated CT from the WET map of the ground truth phantom ($\Delta WET = WET_{CT} - WET_{GT}$).

4.2.3. DOSE CALCULATIONS

To assess the impact of SECT and DECT calibration methods on the dose distribution in animals, irradiation plans were created using the open-source matRad treatment planning system (TPS) [21]. The following sections detail the creation of the matRad beam model for small animal irradiations used in this work (*Beam Modelling*), describe the parameters used to create a pencil beam scanning (PBS) plan for a tumour in the middle of a mouse brain (*Irradiation Plans*), and explain the criteria for evaluating dose distributions (*Dosimetric Evaluation*).

BEAM MODELLING

MatRad requires a database containing the integral depth dose (IDD) curves and lateral size of the beam as a function of depth for a series of proton energies. A realistic model of a 66.5 MeV proton beam for pencil beam irradiations at the IMPACT beamline for small animal radiation biology research at UMCG-PARTREC, which is currently under construction, was created following these steps:

1. The settings of the beam line magnets were optimized with the ion optics code TRANSPORT [39], requiring a dispersion free beam waist in both transverse planes at the centre of the 1 mm diameter, 45 mm length brass collimator used to shape the beam. The initial transverse emittance in the calculations was obtained from measurements close to the cyclotron exit.

2. The magnet settings obtained from the TRANSPORT calculation and the initial emittance were then used as input for Monte Carlo particle tracking simulations with the Geant4-based BDSIM toolkit [20]. The initial phase space file containing the phase space coordinates (i.e. XYZ position, XY momentum components, and kinetic energy) of the individual particles included the transverse - longitudinal correlations introduced by the extraction system of the cyclotron. In the simulation, the interactions of the particles with the vacuum exit window, the foils of the ionization chamber, the air traversed by the particles, and the collimator are taken into account as shown in figure 4.5. The output of the simulation is a phase space file that contains the phase space coordinates of the particles that have passed the collimator. The lateral penumbra (20% – 80%) of the beam at the collimator exit

was 0.33 mm, while its divergence was 16 mrad. This phase space file was used as the source for a subsequent simulation in TOPAS, wherein the 3D total energy deposit in water was scored for 1×10^6 primaries. From this 3D distribution, the quantities needed for the beam model were derived.

3. To generate data for lower energies without performing additional simulations, we pulled back the Bragg curve of the 66.5 MeV proton beam in increments of 0.2 mm. This method gives the correct beam properties provided that the range shifter is right in front of the irradiated object and is water equivalent in terms of multiple scattering.

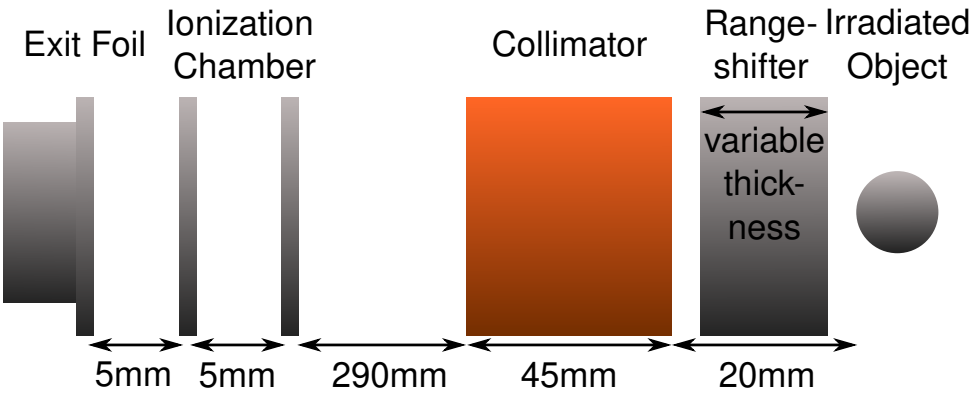


Figure 4.5: Schematic diagram of the final part of the IMPACT beamline.

IRRADIATION PLANS

The irradiation plan was created for the MOBY phantom with a 2 mm diameter spherical tumour in the brain as shown in figure 4.2(a). The planning target volume (PTV) was created by adding a safety margin of 0.2 mm around the tumour to account for setup uncertainties [14, 40]. The PBS plan was made following these steps:

1. First, an initial plan was calculated using the pencil beam algorithm (PBA) of matRad. Since the PBA requires the SPR for the computation of water equivalent thickness, the theoretical SPR (i.e. SPR calculated with the Bethe-Bloch formula using the elemental composition assigned to the organs) was assigned to the MOBY head phantom and used as the input in the calculation. This initial PBA plan evaluation was used to determine the optimal beam arrangement in order to achieve a prescribed dose of 10 Gy in the PTV. A lateral spot spacing of 0.8 mm was implemented and a spread-out Bragg peak (SOBP) was created by superimposing several Bragg

peaks spaced 0.8 mm apart. These parameters were determined from an initial optimization on a target of the same size in water. For the spot fluence optimization, squared dose deviation from the prescribed dose was used for the PTV and no dose constraints were imposed to organs-at-risk such as the brain. The optimization was based on the RBE-weighted dose, which was calculated by applying a constant relative biological effectiveness (RBE) of 1.1 to the physical dose. The doses presented onwards always refer to the RBE-weighted dose.

2. Second, the dose distribution of each pencil beam was recomputed in TOPAS. The phase space file for 66.5 MeV protons obtained in the *Beam Modelling* section was used as the source in the simulation. To match the beam energies in the PBA plan, a range shifter in the form of water slabs was explicitly simulated in TOPAS with thicknesses equal to the difference between the range of the original 66.5 MeV beam and the range in water of the proton beam at each energy layer. The dose was scored in the ground truth MOBY phantom using the *DoseToMedium* scorer. The ground truth is the geometry with material compositions given in table 4.3 assigned (i.e. the same materials used to generate the corresponding CTs in fastCAT).

3. The initial PBA doses were then replaced by the MC-calculated ones. The spot weights were reoptimized in matRad based on the new MC pencil beam doses.

Using the optimized plan, the final dose distribution for the ground truth MOBY head phantom was calculated in TOPAS. The same plan was delivered to the SECT and DECT images to demonstrate how the calibration methods affect the dose distribution. All dose distributions were calculated on a 0.1 mm isotropic dose grid (voxel size of the CT).

DOSIMETRIC EVALUATION

To assess the accuracy of the dose distributions, dose volume histograms (DVHs) were calculated. The criteria for dose coverage is $V_{95} \geq 98\%$ in the PTV. This requires that at least 98% of the PTV receives 95% of the prescribed dose, which is 9.5 Gy. Local differences were also investigated for organs-at-risk (OAR) by looking into the mean (D_{mean}) and maximum (D_{max}) doses received by the brain (i.e. whole brain excluding the PTV) and cranium. The CT-based dose distributions were also compared to the ground truth by performing 3D global gamma analysis with dose difference (DD) / distance to agreement (DTA) set to 3%/0.1 mm. The range shift ($\Delta R = R_{80,CT} - R_{80,GT}$) defined as the difference in the distal range at 80% of the maximum dose was also calculated for 366 line dose profiles over the target area along the beam direction.

Table 4.7: CT numbers of Gammex inserts from TOPAS and fastCAT simulations. The last column lists the difference in Hounsfield units.

Material	HU (TOPAS)	HU (fastCAT)	ΔHU
Adipose	-154	-143	11
Brain	-93	-66	27
Breast	-75	-68	7
Solid Water	40	47	7
Liver	109	115	6
IB3	635	628	7
B200	653	647	6
CB2-30%	1033	1031	2
CB2-50%	1920	1928	8
SB3	2780	2843	64

4.3. RESULTS

4.3.1. SIMULATED CTS FROM FASTCAT

Figure 4.6(a) and (b) shows the simulated CT images of the mini-calibration phantom generated with fastCAT and with a full CBCT simulation in TOPAS, respectively. The comparison of the line profiles is given in figures 4.6(c) and (d). Qualitatively, the contrast is very similar between the two CT images. This observation is supported by the good agreement in the mean HU of the Gammex inserts extracted from fastCAT and TOPAS as shown in table 4.7. The difference in mean HU values between fastCAT and TOPAS were within 11 HU except for the brain (27 HU) and SB3 cortical bone (63 HU). The TOPAS simulations resulted in lower HU values for these materials. For SB3 cortical bone, the cupping artifact is more pronounced in TOPAS leading to a much larger difference. Notably, these two inserts are located in the middle of the phantom. To check whether there is a position dependence of the extracted HU values, an additional simulation was performed wherein the brain and SB3 cortical bone were moved to the outer ring in both the fastCAT and TOPAS simulations (i.e. brain and SB3 cortical bone were switched with the liver and CB2-50%, respectively). The HU difference reduced to 15 HU and 24 HU for the brain and SB3 cortical bone, respectively. In contrast, the deviation for the liver and CB2-50% increased to 15 HU and 26 HU, respectively.

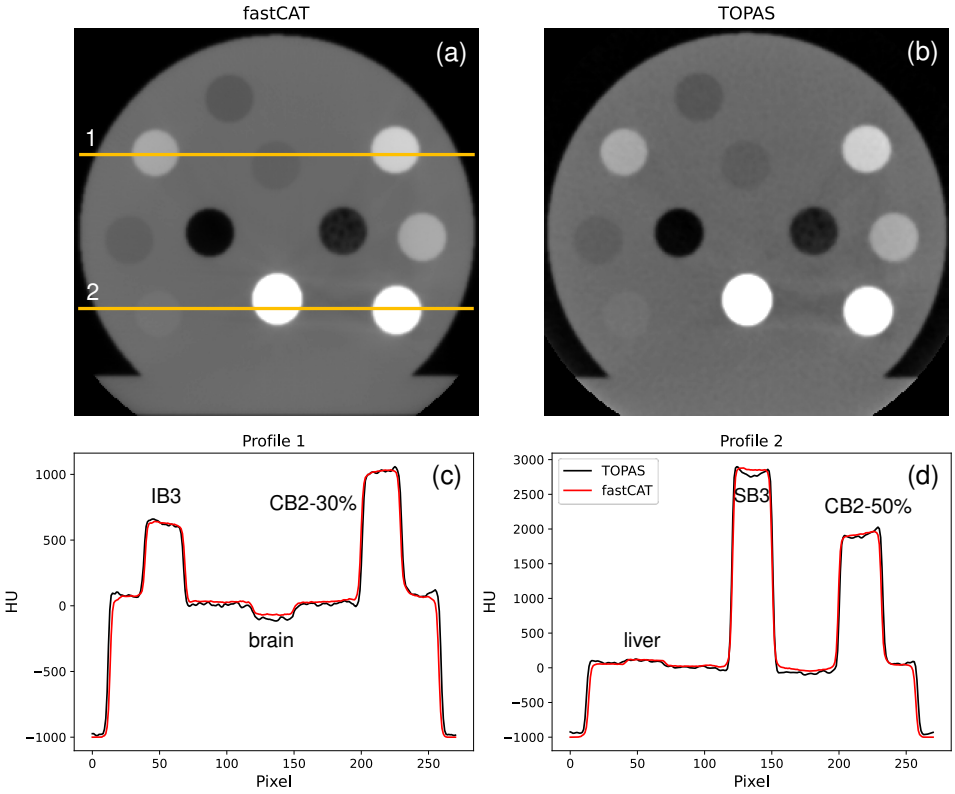


Figure 4.6: Simulated CT images of the mini-calibration phantom from (a) fastCAT and (b) TOPAS averaged over ten slices for 90kV/Al. The line profiles indicated by the yellow lines going through IB3/brain/CB2-30% and liver/SB3/CB2-50% inserts are plotted in (c) and (d), respectively.

4.3.2. SPR ACCURACY OF GAMMEX MATERIALS

Figure 4.7 shows the deviation of the SECT- and DECT-predicted SPR of tissue-equivalent materials from the theoretical value (i.e. SPR calculated using the Bethe-Bloch equation). The corresponding root mean square deviation (RMSD) calculated for each protocol are given in table 4.8. The brain, IB3 inner bone, and B200 bone inserts exhibited the largest deviation (>5%) for SECT. The datapoints for these three materials can be seen to be farther away from the calibration curve (figure 4.4) leading to divergent results. Overall, the CT scan taken with the 90kV/Cu yielded smaller SPR deviations and the lowest RMSD among the SECT imaging protocols implemented in this work.

On the other hand, between the two DECT configurations, the 50kV/Al – 90kV/Cu image pair performed worse than 50kV/Al – 90kV/Al combination as illustrated in fig-

ure 4.7 and table 4.8 despite having better separation in the low and high energy spectra. The 50kV/Al – 90kV/Al DECT yielded prediction errors within ±2%, while 50kV/Al – 90kV/Cu DECT had larger deviations up to -5.2% for IB3 inner bone with Saito and Sagara's method. Compared to the best SECT case, the 50kV/Al – 90kV/Al DECT also exhibited better results, which is consistent with findings using clinical imaging protocols [41, 42, 43]. The two DECT approaches (Landry, Saito and Sagara) at 50kV/Al – 90kV/Al performed similarly with RMSD of 0.9% and 1.0%, respectively. The 90kV/Cu SECT and 50kV/Al – 90kV/Al DECT imaging protocols have then been used to generate SECT and DECT scans of the animal phantom for dose calculations.

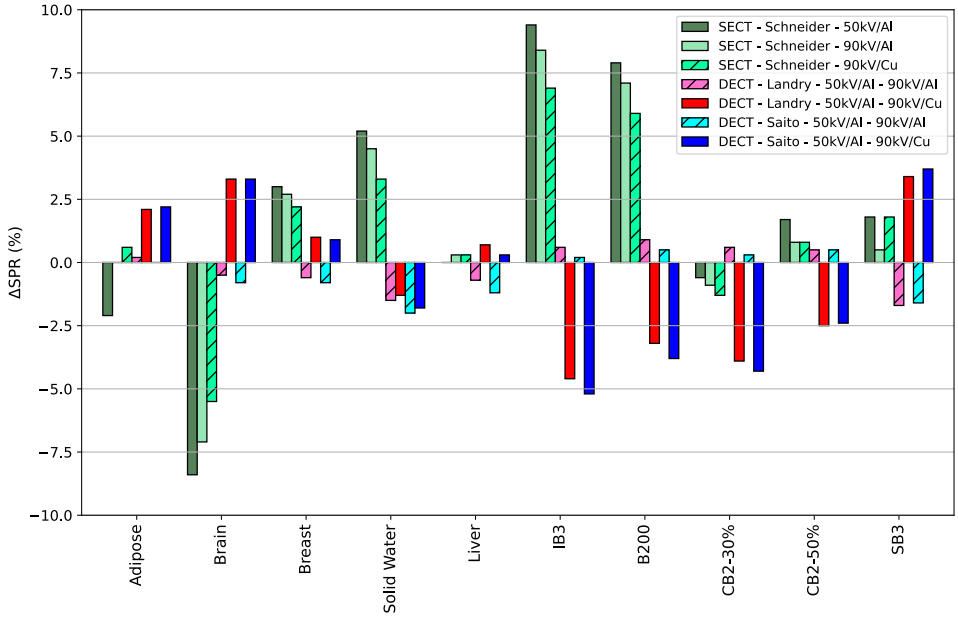


Figure 4.7: Deviation of the SECT- and DECT-based SPR predictions from theoretical values for Gammex materials. Three imaging protocols were evaluated for SECT: 50kV/Al, 90kV/Al, and 90kV/Cu, whereas two combinations were implemented for DECT: 50kV/Al – 90kV/Al and 50kV/Al – 90kV/Cu.

4.3.3. EVALUATION OF RANGE ERRORS IN THE MOBY THORAX

Figure 4.8 shows the distribution of the range differences for the MOBY thorax between the ground truth and the SECT and DECT calibration methods. The range errors are within ±1 mm with SECT having larger errors and larger width of the distribution than DECT. The mean ± standard deviations of the range errors are 0.41 ± 0.16 mm, -0.23 ±

Table 4.8: Root mean square deviation (RMSD) of SPR values of Gammex materials

	RMSD (%)		RMSD (%)	
Protocol	**SECT**	**Protocol**	**DECT Landry**	**DECT Saito**
50kV/Al	5.2	50kV/Al – 90kV/Al	0.9	1.0
90kV/Al	4.5	50kV/Al – 90kV/Cu	2.9	3.1
90kV/Cu	3.7			

0.12 mm, and -0.24 ± 0.12 mm for SECT, DECT-Landry, and DECT-Saito, respectively. SECT calibration systematically gives positive range errors, which indicates that WET values are overestimated as a direct consequence of higher SPR predictions than the ground truth. On the contrary, DECT approaches lean towards negative range errors due to overall underestimation of the SPR.

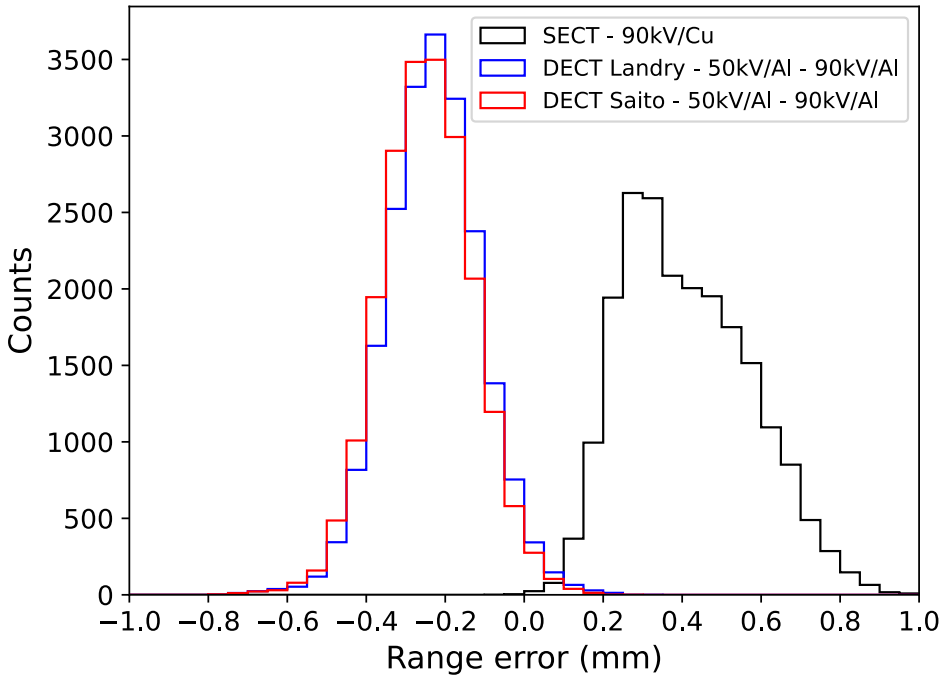


Figure 4.8: Distribution of the range errors ($\Delta WET = WET_{CT} - WET_{GT}$) for the MOBY thorax resulting from SECT and DECT calibration.

4.3.4. IMPACT OF CT CALIBRATION METHODS TO THE DOSE DISTRIBUTION

Figure 4.9(a) presents the proton dose distribution for the treatment plan optimized on the ground truth MOBY phantom. To gauge the uncertainty introduced by the calibration methods to the dose calculation, the same plan was delivered to all other cases. The gamma index maps calculated for each calibration method are shown in figure 4.9(b). Gamma values greater than 1 appear mostly at the distal edge of the PTV and along transitions between soft tissue and bone. Nevertheless, gamma analyses of the SECT-, DECT-Landry-, and DECT-Saito-based dose distributions using 3%/0.1 mm criteria give passing rates of 99.8%, 99.5% and 99.3%, respectively, which show that there are no major dosimetric differences with the ground truth.

Figure 4.10(a) shows a representative dose profile for each case. Both SECT and DECT methods show an underestimation of the dose in regions with high density gradient (tissue/bone interface) as a consequence of CT image blurring at material transitions. This difference is also depicted by γ values > 1 for brain or soft tissue voxels close to the cranium as shown in figure 4.9(b). Analysis of 366 dose profiles revealed that SECT systematically induces negative range shifts with a median of -0.04 mm, while DECT-Landry and DECT-Saito result in positive range shifts with median values of 0.05 mm and 0.06 mm, respectively. These observations are consistent with the results in section 4.3.3 indicating that SECT leads to overestimation of the WET, whereas DECT underestimates it. As shown in figure 4.10(b), the range shifts are below the CT voxel size of 0.1 mm and can be considered negligible.

The comparison of the dose volume histograms (DVHs) is displayed in figure 4.11, and the dose metrics are given in table 4.9. From the DVH analysis, it can be deduced that adequate PTV coverage (i.e. $V_{95} \geq 98\%$) was not achieved for any of the dose distributions based on the calibrated CTs. Only 95.9%, 96.3%, and 96.4% of the PTV received at least 9.5 Gy for dose distributions computed with SECT, DECT-Landry, and DECT-Saito images, respectively. Voxels that did not meet the dose coverage criteria are mostly towards the edge of the PTV as shown in figure 4.9(b). Meanwhile, all calibration methods showed no considerable change in the mean and maximum doses to the surrounding OAR.

4.4. DISCUSSION

This work presents a framework to simulate preclinical proton irradiations. Using this platform, one can design and optimize preclinical setups and gain insights about how it affects the dose distribution in an animal prior to experiments. A major part of this framework is the fastCAT CBCT simulator. It was modified based on the CT geometry of a small animal CT scanner to allow generation of realistic micro-CBCT images. The va-

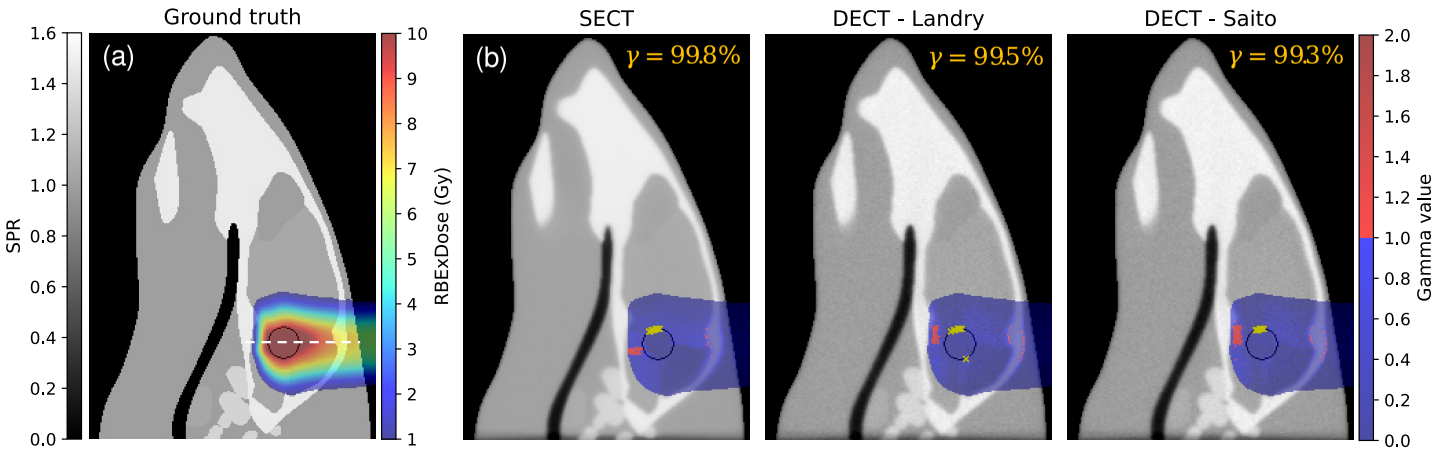


Figure 4.9: (a) Dose distribution for the treatment plan optimized on the ground truth MOBY head phantom. (b) Gamma index maps (from left to right) for the same plan delivered to SECT, DECT-Landry, and DECT-Saito calibrated images. Tolerance limits (DD/DTA) of 3%/0.1mm were used for the gamma evaluation. The images onto which the dose distribution and gamma index maps were overlayed are the theoretical SPR (GT) and the SPR distribution resulting from the calibration (SECT and DECT). The black circle illustrates the PTV, and the white line indicates where dose profiles in figure 4.10(a) were extracted. The yellow 'x' marks show the region in the PTV that failed the dose coverage criteria (i.e. $D_{PTV} < 9.5$ Gy). DD: dose difference, DTA: distance-to-agreement.

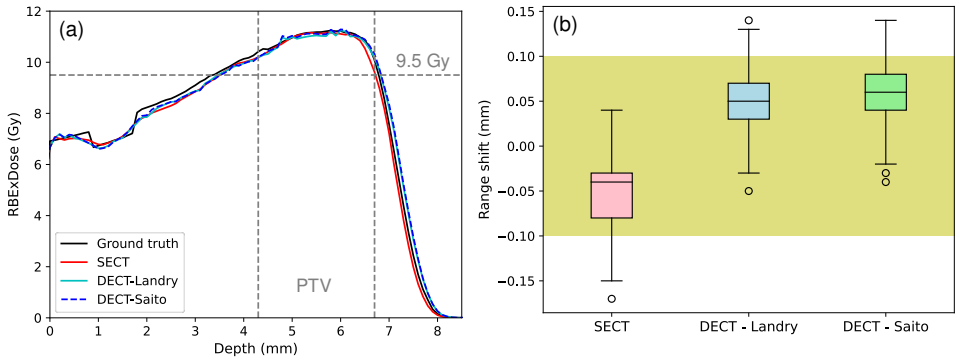


Figure 4.10: (a) Dose profiles in the beam direction taken at the centre of the PTV indicated by the white line in figure 4.9(a). (b) Range shift ($\Delta R = R_{80,CT} - R_{80,GT}$) distribution for SECT and DECT calibration methods. Yellow shade indicates the region within one voxel size of the CT (±0.1 mm).

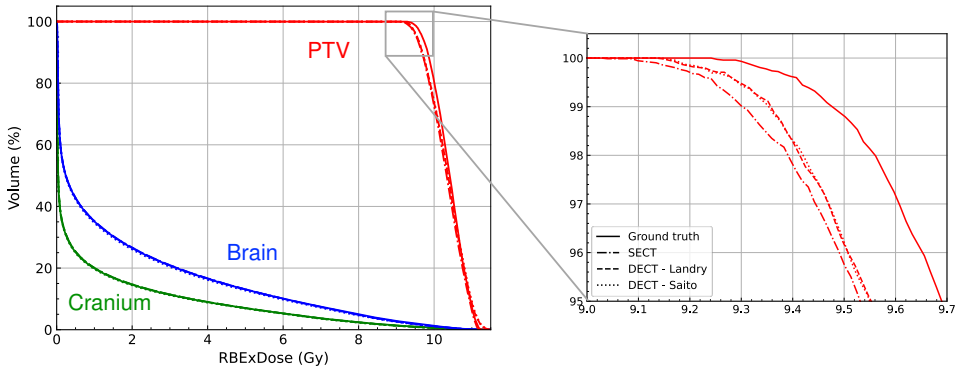


Figure 4.11: Dose volume histograms calculated for the PTV (prescribed dose: 10 Gy), OAR-brain, and cranium. The solid, dash-dotted, dashed, and dotted lines correspond to the ground truth, SECT, DECT-Landry, DECT-Saito, respectively.

lidity of the fastCAT micro-CBCT scans was demonstrated by the overall good agreement of the CT numbers with full Monte Carlo (MC) simulations. The large discrepancies observed for the brain and SB3 cortical bone inserts are likely caused by beam hardening not being properly handled in fastCAT. These two inserts are located at the center of the phantom, where the beam has the hardest spectrum (i.e. beam with a higher mean energy due to low energy X-rays being preferentially absorbed). The attenuation coefficient then becomes lower, which results to lower HU values as demonstrated by TOPAS. The cupping artifact in the SB3 cortical bone is also much more pronounced in TOPAS than

Table 4.9: Mean (D_{mean}) and maximum (D_{max}) doses, and percent volume that received 95% of the prescribed dose (V_{95}). The last one is only for the PTV.

Structure Metric	Ground truth	SECT	DECT-Landry	DECT-Saito
PTV				
D_{mean} (Gy)	10.4	10.3	10.4	10.4
D_{max} (Gy)	11.3	11.3	11.5	11.5
V_{95} (%)	98.8	95.9	96.3	96.4
Brain				
D_{mean} (Gy)	1.6	1.5	1.6	1.6
D_{max} (Gy)	10.7	10.5	10.7	10.7
Cranium				
D_{mean} (Gy)	0.2	0.2	0.2	0.2
D_{max} (Gy)	7.5	7.8	7.8	7.8

in fastCAT, which further supports this claim.

In comparison to Vaniqui et al.'s [23] measured HU values for the same mini-calibration phantom at 50kV/Al and 90kV/Al, fastCAT CT numbers are generally higher. The difference can have a number of explanations. The reconstruction protocol and the region of interest from which the HU values were extracted may have been different. The thickness of the CsI detector and the pixel pitch are not the same. Furthermore, the simulations to create the micro-CBCT fastCAT model were rather simplistic. Collimators and other CT components were not included in the modelling, which could generate additional scatter. Also, the angular distribution of the X-ray intensity was assumed to be uniform while in reality, it is not. A better agreement with experimental values could be obtained by replacing the detector response functions in fastCAT and performing more detailed modelling of the micro-CT scanner. Nevertheless, the general trend of the CT numbers of Gammex materials from fastCAT remains consistent with published values.

The potential use of fastCAT simulated CT scans in the evaluation of SECT- and DECT-based stopping power ratio (SPR) estimation for proton treatment planning in animals was also demonstrated in this study. Previous works have already investigated the feasibility of using DECT, and its impact on preclinical X-ray dose distributions [22, 23]. Here, we extended it by exploring another energy combination with better spectral separation (50kV/Al – 90kV/Cu), using different DECT approaches to predict proton stopping power ratios, and comparing the results to the SECT stoichiometric calibration. The best performing SECT and DECT protocols were also evaluated in terms of range errors (from

MC-simulated proton radiographs) and proton dose calculation accuracy in animal CTs.

The SECT stoichiometric calibration has been reported to yield uncertainties up to 3.5% [44]. For the SECT imaging protocols (i.e. 50kV/Al, 90kV/Al and 90kV/Cu) investigated in this work, the SPR deviation of Gammex materials from theoretical values is in a similar range except for the brain, IB3, and B200 inserts, where deviations >5% were obtained. These tissue surrogates are seen to lie away from the calibration curve as shown in figure 4.4. This suggests that these materials have a poor ability in reproducing CT numbers of biological tissues. Although not as pronounced as in our case, the same three materials also slightly deviate from the HU-to-ρ_e calibration curve by Goma et al. [45] for human size phantoms. Since the SPR is directly proportional to the ρ_e, we expect a similar trend. Overall, the SECT SPR prediction improves as the mean energy of the X-ray spectra increases, with 90kV/Cu yielding the lowest RMSD. It is important to note that the Cu filter was not explicitly modelled in the simulations and at such thickness, it will produce more scattered X-rays, which might be detrimental to the image quality and may affect the accuracy of the calibration.

For DECT, two different acquisition settings were investigated. First, the 50kV/Al – 90kV/Al combination was selected as it was reported by Schyns et al. [22] to be the optimal energy combination for preclinical DECT showing the lowest deviation in the estimated ρ_e and Z_{eff} of tissue equivalent materials. However, their work only compared DECT results for different energy combinations filtered by the same material. To improve spectral separation, we implemented a stronger filtration material for the high kV beam (i.e. 90 kVp filtered with 0.32 mm Cu) in SpekPy for the second DECT image pair (50kV/Al – 90kV/Cu). The idea is that by reducing the overlap between the X-ray spectra, the attenuation values (or CT numbers) from the high and low kV datasets become more independent from each other, which should improve material discrimination [46]. However, contrary to expectation, the 50kV/Al – 90kV/Cu performed poorly compared to 50kV/Al – 90kV/Al. To understand these findings, further studies including experimental measurements are needed.

The 50kV/Al – 90kV/Al DECT pair achieved better results for Gammex materials compared to 90kV/Cu SECT with deviations within ±2%. At this setting, the DECT conversion algorithms of Landry et al. [27] and Saito and Sagara [35] were comparable although the latter offers a much simpler implementation. Note that several DECT studies using clinical imaging protocols have shown higher accuracy in SPR estimation than what we have achieved [42, 47, 48, 49]. These studies were usually performed with 80kV – 140kV pair where tin (Sn) filtration was applied for the high tube voltage. This combination exhibits a large X-ray spectra separation, with the high kV image being dominated by

Compton effect. Yang et al. [50] and Li et al. [51] have demonstrated that increasing the energy separation leads to a reduction in the uncertainties associated with SPR estimation. Although it is desirable to use higher energies for DECT, there are limitations for pre-clinical imaging. For instance, the tube voltage of some micro-CT machines only go up until 100 kV. Moreover, higher energies will also result to poorer contrast in the image.

To get an idea on how the SPR uncertainties associated to the CT calibration methods translate into errors in the proton range, we obtained WET maps through proton radiograph simulations on SECT and DECT images. The mean shift and variation in the proton range for DECT is smaller in magnitude than SECT, which reflects its superiority in terms of tissue characterization. However, it is important to note that performing dual energy CT imaging results in additional dose to the animal and the imaging dose required to obtain acceptable DECT calibration is high (30 cGy per CT image) [22].

To assess whether DECT indeed offers a potential gain over SECT in treatment planning for animals, proton dose calculations in a mouse brain were performed on SECT and DECT images. The treatment plans were made in matRad to which we have added a beam data library based on a realistic beam model for small animal irradiations. The proton beam model was obtained from beam transport simulations of a preclinical proton beamline in BDSIM. The accuracy of the dose distributions was evaluated by comparing it with the one calculated based on the ground truth phantom. It is worth highlighting that this is another advantage of the framework. Since the actual tissue compositions used to generate the CTs in fastCAT are known, the same materials can be assigned to the MOBY geometry to create a ground truth image for dose calculations, which is usually lacking when experimental CTs of animals are used.

Results from proton dose calculations on SECT- and DECT-calibrated CT images showed very small differences with the ground truth. A slight shift in the dose profiles can be observed in figure 4.10(a), wherein the SECT profile is more upstream and DECT ones are deeper than the ground truth. These observations are in line with the proton radiograph results in which SECT was shown to overestimate the SPR, while DECT tend to underestimate it. However, SECT- and DECT-calibrated scans did not really lead to considerable range errors as the average shift is smaller than the dose grid (0.1 mm). There was also negligible effect on the global dose distribution as shown by the high gamma pass ratios (>99% for both SECT and DECT) and on the local mean doses to anatomical structures (both target and OARs). Overall, the results suggest that SECT is sufficient for CT characterization of tissues in a micro-CT for animal proton treatment planning and that the additional dose required for DECT imaging is not warranted.

To our knowledge, this work presented the first dosimetric evaluation of SECT and

DECT calibration methods in the context of proton irradiation of small animals. It should however be emphasized that the merit of implementing DECT imaging in proton preclinical practice cannot be evaluated just with a single case study. The investigation was limited to the head, which can be considered relatively homogeneous with only the brain and the skull along the beam direction. The framework can also be used to assess other treatment regions particularly those with more complex geometry and increased heterogeneity to test the robustness of the calibration methods.

4.5. CONCLUSION

In this study, we proposed a framework that enables *in silico* modelling of preclinical proton irradiations. The framework will be helpful for the development and optimization of irradiation setups, assessment of the quality of small animal irradiations, and quantification of associated uncertainties in preclinical proton dose delivery. As an example, we have demonstrated how the framework can be used to assess the impact of SECT and DECT calibration methods on proton dose distributions in small animals. Calculations on a mouse brain revealed that treatment planning based on DECT offered no added benefit to the accuracy of the dose.

ACKNOWLEDGEMENTS

This work was partially financially supported by a grant from the Dutch Cancer Society (KWF research project 12092).

SUPPLEMENTARY FILE

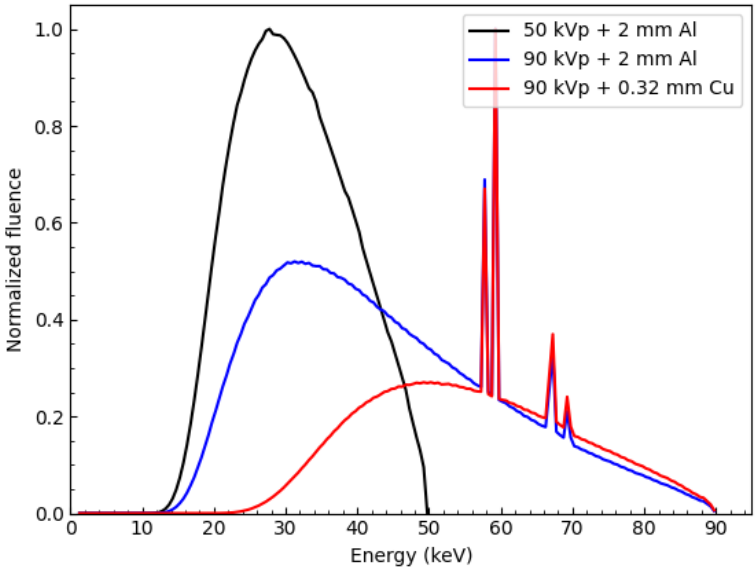


Figure S4.1: X-ray energy spectra extracted from SpekPy for 50 and 90 kVp filtered with 2 mm aluminium (Al) and 90 kVp filtered with 0.32 mm copper (Cu).

Table S4.1: Physics list and range cuts used for TOPAS simulations

CT simulation	
Physics list	g4em-standard_opt4
Range cut	electron/positron: 0.1 mm (phantom), 1 mm (everywhere else) gamma: 1 mm
Proton radiograph / Dose calculation	
Physics list	g4em-standard_opt4, g4h-phy_QGSP_BIC_HP, g4h-elastic_HP, g4ion-binarycascade, g4stopping, g4em-extra, g4decay, g4radioactivedecay
Range cut	electron/positron: 0.1 mm (phantom), 1 mm (everywhere else) gamma: 1 mm

Table S4.2: Elemental composition of materials assigned to the mini calibration phantom obtained from the manufacturer

#	Material	ρ	H	C	N	O	Z > 8
1	Air	0.001	0.0	0.0	75.5	23.2	Ar(1.3)
2	LN-450 Lung	0.428	8.47	59.56	1.97	18.11	Mg(11.21), Si(0.58), Cl(0.10)
3	AP6 Adipose	0.95	9.06	72.29	2.25	16.27	Cl(0.13)
4	SR2 Brain	1.05	10.83	72.54	1.69	14.86	Cl(0.08)
5	BR12 Breast	0.98	8.59	70.10	2.33	17.90	Cl(0.13), Ca(0.95)
6	Solid Water	1.02	8.00	67.29	2.39	19.87	Cl(0.14), Ca(2.31)
7	LV1 Liver	1.10	8.06	67.01	2.47	20.01	Cl(0.14), Ca(2.31)
8	IB3 Inner Bone	1.13	6.67	55.65	1.96	23.52	P(3.23), Cl(0.11), Ca(8.86)
9	B200 Bone	1.15	6.65	55.51	1.98	23.64	P(3.24), Cl(0.11), Ca(8.87)
10	CB2-30% $CaCO_3$	1.33	6.68	53.47	2.12	25.61	Cl(0.11), Ca(12.01)
11	CB2-50% $CaCO_3$	1.56	4.77	41.61	1.52	32.00	Cl(0.08), Ca(20.02)
12	SB3 Cortical Bone	1.82	3.41	31.41	1.84	36.50	Cl(0.04), Ca(26.80)

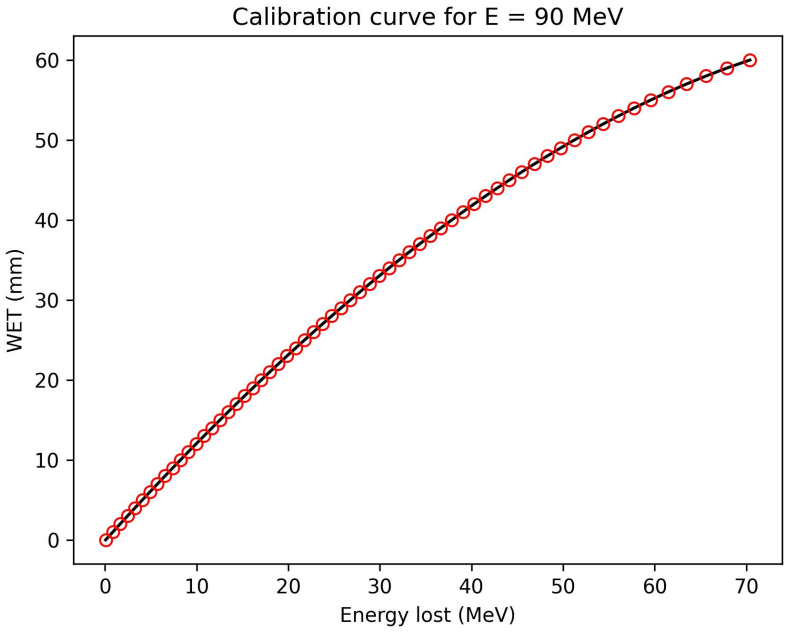


Figure S4.2: Calibration curve to convert the residual kinetic energy (MeV) of protons to water equivalent thickness (WET). The calibration was obtained for 100 MeV protons in steps of 1-mm water thickness. A 3rd degree polynomial (WET [mm] = $2.96 \times 10^{-5} KE_{res}^3 - 1.16 \times 10^{-2} KE_{res}^2 + 1.17 \times 10^{-1} KE_{res} + 7.39 \times 10^{1}$) was fitted to the datapoints from the simulation.

Table S4.3: Elemental composition of materials assigned to the MOBY phantom

#	Material	ρ	H	C	N	O	Z > 8
Head							
1	Air[a]	0.001	0.0	0.0	75.5	23.2	Ar(1.3)
2	Soft tissue[a]	1.00	10.1	11.1	2.6	76.2	
3	Brain[b]	1.05	10.7	14.3	2.3	71.3	
4	Thyroid[c]	1.05	10.4	11.9	2.4	74.5	Na(0.2), P(0.1), S(0.1), Cl(0.2), K(0.1), I(0.1)
5	Male vertebral column (C4)[d]	1.42	6.3	26.1	3.9	43.6	Na(0.1), Mg(0.1), P(6.1), S(0.3), Cl(0.1), K(0.1), Ca(13.3)
6	Cranium[d]	1.61	5.0	21.2	4.0	43.5	Na(0.1), Mg(0.2), P(8.1), S(0.3), Ca(17.6)
Thorax							
1	Air[a]	0.001	0.0	0.0	75.5	23.2	Ar(1.3)
2	Lung (inflated)[c]	0.26	10.3	10.5	3.1	74.9	Na(0.2), P(0.2), S(0.3), Cl(0.3), K(0.2)
3	Soft tissue[a]	1.00	10.1	11.1	2.6	76.2	
4	Heart[b]	1.05	10.4	13.8	2.9	71.9	Na(0.1), P(0.2), S(0.2), Cl(0.2), K(0.3)
5	Liver[b]	1.05	10.2	13.0	3.1	72.5	Na(0.2), P(0.2), S(0.3), Cl(0.2), K(0.3)
6	Stomach[c]	1.05	10.4	13.9	2.9	72.1	Na(0.1), P(0.1), S(0.2), Cl(0.1), K(0.2)
7	Gall bladder[b]	1.03	10.4	23.1	2.8	62.7	Na(0.1), P(0.2), S(0.3), Cl(0.2), K(0.2)
8	Male ribs (2^{nd}, 6^{th})[d]	1.41	6.4	26.3	3.9	43.6	Na(0.1), Mg(0.1), P(6.0), S(0.3), Cl(0.1), K(0.1), Ca(13.1)

[a]Geant4 material database (https://www.fe.infn.it/u/paterno/Geant4_tutorial/slides_further/Geometry/G4_Nist_Materials.pdf), [b]ICRP 110 [28], [c]Woodard and White [29], [d]White et al. [30]

Table S4.4: Mean HU ± standard deviation of Gammex inserts

Material	**50kV/Al**	**90kV/Al**	**90kV/Cu**
Adipose	-194 ± 7	-143 ± 7	-117 ± 7
Brain	-139 ± 8	-66 ± 8	-17 ± 7
Breast	-91 ± 9	-68 ± 8	-60 ± 7
Solid Water	72 ± 8	47 ± 7	23 ± 6
Liver	134 ± 10	115 ± 9	102 ± 7
IB3	935 ± 18	628 ± 13	422 ± 7
B200	954 ± 21	647 ± 15	447 ± 8
CB2-30%	1459 ± 22	1031 ± 15	762 ± 8
CB2-50%	2771 ± 33	1928 ± 23	1455 ± 11
SB3	4114 ± 23	2843 ± 16	2216 ± 12

BIBLIOGRAPHY

[1] PTCOG. *Particle therapy facilities in clinical operation*. 2024. URL: https://www.ptcog.site/index.php/facilities-in-operation-public (visited on 08/26/2024).

[2] Ghazaleh Ghobadi et al. "Physiological Interaction of Heart and Lung in Thoracic Irradiation". In: *International Journal of Radiation Oncology*Biology*Physics* 84.5 (2012), e639–e646. DOI: 10.1016/j.ijrobp.2012.07.2362.

[3] Hendrik P. Bijl et al. "Influence of adjacent low-dose fields on tolerance to high doses of protons in rat cervical spinal cord". In: *International Journal of Radiation Oncology*Biology*Physics* 64.4 (2006), pp. 1204–1210. DOI: 10.1016/j.ijrobp.2005.06.046.

[4] Brita Singers Sørensen et al. "Relative biological effectiveness (RBE) and distal edge effects of proton radiation on early damage in vivo". In: *Acta Oncologica* 56.11 (2017), pp. 1387–1391. DOI: 10.1080/0284186X.2017.1351621.

[5] Michelle E. Howard et al. "Dosimetric Assessment of a High Precision System for Mouse Proton Irradiation to Assess Spinal Cord Toxicity". In: *Radiation Research* 195.6 (2021), pp. 541–548. DOI: 10.1667/RADE-20-00153.1.

[6] Janet M. Denbeigh et al. "Characterizing Proton-Induced Biological Effects in a Mouse Spinal Cord Model: A Comparison of Bragg Peak and Entrance Beam Response in Single and Fractionated Exposures". In: *International Journal of Radiation Oncology*Biology*Physics* 119.3 (2024), pp. 924–935. DOI: 10.1016/j.ijrobp.2023.12.031.

[7] Yolanda Prezado et al. "Proton minibeam radiation therapy spares normal rat brain: Long-Term Clinical, Radiological and Histopathological Analysis". In: *Scientific Reports* 7.14403 (2017). DOI: 10.1038/s41598-017-14786-y.

[8] Yolanda Prezado et al. "Tumor Control in RG2 Glioma-Bearing Rats: A Comparison Between Proton Minibeam Therapy and Standard Proton Therapy". In: *International Journal of Radiation Oncology*Biology*Physics* 104.2 (2019), pp. 266–271. DOI: 10.1016/j.ijrobp.2019.01.080.

[9] Elke Beyreuther et al. “Feasibility of proton FLASH effect tested by zebrafish embryo irradiation”. In: *Radiotherapy and Oncology* 139 (2019), pp. 46–50. DOI: 10.1016/j.radonc.2019.06.024.

[10] Michele M. Kim et al. “Comparison of FLASH Proton Entrance and the Spread-Out Bragg Peak Dose Regions in the Sparing of Mouse Intestinal Crypts and in a Pancreatic Tumor Model”. In: *Cancers* 13.16 (2021). DOI: 10.3390/cancers13164244.

[11] Brita Singers Sørensen et al. “Pencil beam scanning proton FLASH maintains tumor control while normal tissue damage is reduced in a mouse model”. In: *Radiotherapy and Oncology* 175 (2022), pp. 178–184. DOI: 10.1016/j.radonc.2022.05.014.

[12] Peter van Luijk et al. “Sparing the region of the salivary gland containing stem cells preserves saliva production after radiotherapy for head and neck cancer”. In: *Science Translational Medicine* 7.305 (2015), 305ra147–305ra147. DOI: 10.1126/scitranslmed.aac4441.

[13] Theresa Suckert et al. “Late Side Effects in Normal Mouse Brain Tissue After Proton Irradiation”. In: *Frontiers in Oncology* 10 (2021). DOI: 10.3389/fonc.2020.598360.

[14] E Ford et al. “An image-guided precision proton radiation platform for preclinical in vivo research”. In: *Physics in Medicine and Biology* 62.1 (2016), p. 43. DOI: 10.1088/1361-6560/62/1/43.

[15] Michele M Kim et al. “Design and commissioning of an image-guided small animal radiation platform and quality assurance protocol for integrated proton and x-ray radiobiology research”. In: *Physics in Medicine and Biology* 64.13 (2019), p. 135013. DOI: 10.1088/1361-6560/ab20d9.

[16] Katia Parodi et al. “Towards a novel small animal proton irradiation platform: the SIRMIO project”. In: *Acta Oncologica* 58.10 (2019), pp. 1470–1475. DOI: 10.1080/0284186X.2019.1630752.

[17] Moritz Schneider et al. “SAPPHIRE—establishment of small animal proton and photon image-guided radiation experiments”. In: *Physics in Medicine and Biology* 69.9 (2024), p. 095020. DOI: 10.1088/1361-6560/ad3887.

[18] Jericho O’Connell and Magdalena Bazalova-Carter. “fastCAT: Fast cone beam CT (CBCT) simulation”. In: *Medical Physics* 48.8 (2021), pp. 4448–4458. DOI: 10.1002/mp.15007.

[19] J. Perl et al. “TOPAS: An innovative proton Monte Carlo platform for research and clinical applications”. In: *Medical Physics* 39.11 (2012), pp. 6818–6837. DOI: 10.1118/1.4758060.

[20] L.J. Nevay et al. “BDSIM: An accelerator tracking code with particle–matter interactions”. In: *Computer Physics Communications* 252 (2020), p. 107200. DOI: 10.1016/j.cpc.2020.107200.

[21] Hans-Peter Wieser et al. “Development of the open-source dose calculation and optimization toolkit matRad”. In: *Medical Physics* 44.6 (2017), pp. 2556–2568. DOI: 10.1002/mp.12251.

[22] Lotte E J R Schyns et al. “Optimizing dual energy cone beam CT protocols for preclinical imaging and radiation research”. In: *British Journal of Radiology* 90.1069 (2016), p. 20160480. DOI: 10.1259/bjr.20160480.

[23] Ana Vaniqui et al. “The impact of dual energy CT imaging on dose calculations for pre-clinical studies”. In: *Radiation Oncology* 12.181 (2017). DOI: 10.1186/s13014-017-0922-9.

[24] Gavin Poludniowski et al. “Technical Note: SpekPy v2.0—a software toolkit for modeling x-ray tube spectra”. In: *Medical Physics* 48.7 (2021), pp. 3630–3637. DOI: 10.1002/mp.14945.

[25] Ander Biguri et al. “TIGRE: a MATLAB-GPU toolbox for CBCT image reconstruction”. In: *Biomedical Physics and Engineering Express* 2.5 (2016), p. 055010. DOI: 10.1088/2057-1976/2/5/055010.

[26] William.P. Segars et al. “Development of a 4-D digital mouse phantom for molecular imaging research”. In: *Molecular Imaging and Biology* 6.3 (2004), pp. 149–159. DOI: 10.1016/j.mibio.2004.03.002.

[27] Guillaume Landry et al. “Deriving effective atomic numbers from DECT based on a parameterization of the ratio of high and low linear attenuation coefficients”. In: *Physics in Medicine and Biology* 58.19 (2013), p. 6851. DOI: 10.1088/0031-9155/58/19/6851.

[28] Hans-Georg Menzel and Paul DeLuca Christopher Clement. “Realistic reference phantoms: An ICRP/ICRU joint effort”. In: *Annals of the ICRP* 39.2 (2009), pp. 3–5. DOI: 10.1016/j.icrp.2009.09.001.

[29] H. Q. Woodard and D. R. White. “The composition of body tissues”. In: *British Journal of Radiology* 59.708 (1986), pp. 1209–1218. DOI: 10.1259/0007-1285-59-708-1209.

[30] D. R. White, H. Q. Woodard, and S. M. Hammond. "Average soft-tissue and bone models for use in radiation dosimetry". In: *British Journal of Radiology* 60.717 (1987), pp. 907–913. DOI: 10.1259/0007-1285-60-717-907.

[31] Wilfried Schneider, Thomas Bortfeld, and Wolfgang Schlegel. "Correlation between CT numbers and tissue parameters needed for Monte Carlo simulations of clinical dose distributions". In: *Physics in Medicine and Biology* 45.2 (2000), p. 459. DOI: 10.1088/0031-9155/45/2/314.

[32] Esther Bär et al. "Optimized I-values for use with the Bragg additivity rule and their impact on proton stopping power and range uncertainty". In: *Physics in Medicine and Biology* 63.16 (2018), p. 165007. DOI: 10.1088/1361-6560/aad312.

[33] Masatoshi Saito. "Potential of dual-energy subtraction for converting CT numbers to electron density based on a single linear relationship". In: *Medical Physics* 39.4 (2012), pp. 2021–2030. DOI: 10.1118/1.3694111.

[34] M Yang et al. "Theoretical variance analysis of single- and dual-energy computed tomography methods for calculating proton stopping power ratios of biological tissues". In: *Physics in Medicine and Biology* 55.5 (2010), p. 1343. DOI: 10.1088/0031-9155/55/5/006.

[35] Masatoshi Saito and Shota Sagara. "A simple formulation for deriving effective atomic numbers via electron density calibration from dual-energy CT data in the human body". In: *Medical Physics* 44.6 (2017), pp. 2293–2303. DOI: 10.1002/mp.12176.

[36] K B Niepel et al. "Animal tissue-based quantitative comparison of dual-energy CT to SPR conversion methods using high-resolution gel dosimetry". In: *Physics in Medicine and Biology* 66.7 (2021), p. 075009. DOI: 10.1088/1361-6560/abbd14.

[37] Masatoshi Saito and Shota Sagara. "Simplified derivation of stopping power ratio in the human body from dual-energy CT data". In: *Medical Physics* 44.8 (2017), pp. 4179–4187. DOI: 10.1002/mp.12386.

[38] Felicia Fibiani Permatasari et al. "Material assignment for proton range prediction in Monte Carlo patient simulations using stopping-power datasets". In: *Physics in Medicine and Biology* 65.18 (2020), p. 185004. DOI: 10.1088/1361-6560/ab9702.

[39] Urs Rohrer. *PSI graphic transport framework*. 2007. URL: https://aea.web.psi.ch/Urs_Rohrer/MyWeb/trans.htm (visited on 08/31/2024).

[40] Frank Verhaegen, Patrick Granton, and Erik Tryggestad. “Small animal radiotherapy research platforms”. In: *Physics in Medicine and Biology* 56.12 (2011), R55. DOI: 10.1088/0031-9155/56/12/R01.

[41] Nace Hudobivnik et al. “Comparison of proton therapy treatment planning for head tumors with a pencil beam algorithm on dual and single energy CT images”. In: *Medical Physics* 43.1 (2016), pp. 495–504. DOI: 10.1118/1.4939106.

[42] Esther Bär et al. “The potential of dual-energy CT to reduce proton beam range uncertainties”. In: *Medical Physics* 44.6 (2017), pp. 2332–2344. DOI: 10.1002/mp.12215.

[43] Esther Bär et al. “Experimental validation of two dual-energy CT methods for proton therapy using heterogeneous tissue samples”. In: *Medical Physics* 45.1 (2018), pp. 48–59. DOI: 10.1002/mp.12666.

[44] Ming Yang et al. “Comprehensive analysis of proton range uncertainties related to patient stopping-power-ratio estimation using the stoichiometric calibration”. In: *Physics in Medicine and Biology* 57.13 (2012), p. 4095. DOI: 10.1088/0031-9155/57/13/4095.

[45] Carles Gomà, Isabel P Almeida, and Frank Verhaegen. “Revisiting the single-energy CT calibration for proton therapy treatment planning: a critical look at the stoichiometric method”. In: *Physics in Medicine and Biology* 63.23 (2018), p. 235011. DOI: 10.1088/1361-6560/aaede5.

[46] Bernhard Krauss et al. “The Importance of Spectral Separation: An Assessment of Dual-Energy Spectral Separation for Quantitative Ability and Dose Efficiency”. In: *Investigative Radiology* 50.2 (2015). DOI: 10.1097/RLI.0000000000000109.

[47] Nora Hünemohr et al. “Experimental verification of ion stopping power prediction from dual energy CT data in tissue surrogates”. In: *Physics in Medicine and Biology* 59.1 (2013), p. 83. DOI: 10.1088/0031-9155/59/1/83.

[48] Vicki T. Taasti et al. “A robust empirical parametrization of proton stopping power using dual energy CT”. In: *Medical Physics* 43.10 (2016), pp. 5547–5560. DOI: 10.1118/1.4962934.

[49] Guillaume Landry et al. “Technical Note: Relative proton stopping power estimation from virtual monoenergetic images reconstructed from dual-layer computed tomography”. In: *Medical Physics* 46.4 (2019), pp. 1821–1828. DOI: 10.1002/mp.13404.

[50] M Yang et al. “Does kV–MV dual-energy computed tomography have an advantage in determining proton stopping power ratios in patients?” In: *Physics in Medicine and Biology* 56.14 (2011), p. 4499. DOI: 10.1088/0031-9155/56/14/017.

[51] B Li et al. “Comprehensive analysis of proton range uncertainties related to stopping-power-ratio estimation using dual-energy CT imaging”. In: *Physics in Medicine and Biology* 62.17 (2017), p. 7056. DOI: 10.1088/1361-6560/aa7dc9.

5

GENERAL DISCUSSION AND OUTLOOK

5.1. INTRODUCTION

The global incidence of cancer continues to climb, with over half of all patients receiving radiotherapy as a form of treatment [1, 2]. To boost its effectiveness and reduce normal tissue toxicity, new treatment approaches in the field of radiotherapy are continuously being explored including the use of particle beams [3, 4], radioimmunotherapy [5, 6], radiation and drug combination treatments [7, 8], ultra-high dose rate FLASH therapy [9, 10, 11], and spatial fractionation [12, 13, 14]. In conjunction with remaining knowledge gaps in radiobiology, these developments further underscore the need for comprehensive preclinical studies:

- to characterize normal tissue, tumour, and immunological responses
- to identify early and long-term side effects
- to assess cellular, tissue, and organ-specific radiosensitivities
- to further elucidate DNA damage and repair mechanisms and how they coordinate differently compared to conventional photon beams
- to identify appropriate targets for combination treatments with radiation and to gain insights into their added value
- to optimize parameters in order to maximize therapeutic outcomes and support translation to the clinic.

Recent technological advancements in small animal irradiation platforms—such as inclusion of high-resolution image-guidance, dynamic translation and collimation systems, flexibility and increased conformality in dose delivery, and integration with particle beamlines [15, 16, 17, 18, 19, 20]—along with the increasing availability of more clinically relevant animal models [20, 21] have created new opportunities to conduct preclinical experiments with a much higher level of precision and relevance to the clinic. The insights obtained from these preclinical studies are invaluable for refining and gaining confidence in radiotherapy treatment schemes and can play a pivotal role in shaping the design of patient studies in the clinic.

However, small animal experiments present a unique challenge. Unlike in the clinic, where imaging, contouring, and planning can be conducted over a period of several days, the preclinical workflow typically requires these steps to be carried out consecutively while the animal is in the irradiation position under sedation. To mitigate effects of the anaesthesia and ensure the animal's wellbeing, a fast workflow is thus essential. Therefore, this thesis explores the use of deep learning (DL) for contouring (**Chapter 2**)

and fast proton dose engines for dose calculations (**Chapter 3**) to expedite irradiation plan generation for small animals.

The small size of animals, coupled with the even tinier size of targets, also imposes stringent constraints on the dose accuracy, particularly when using particle beams. To put this into perspective, van de Worp et al. [22] delineated lung tumours in 60 mice and found that the size ranged from 1.1 mm^3 to 60 mm^3. This entails that range uncertainties from particle beams should be in the submillimetre level to avoid the risk of underdosing the target or overdosing surrounding healthy tissues. Therefore, the simulation framework of the preclinical workflow (**Chapter 4**), developed as a part of this work, has been used to evaluate range uncertainties associated with CT Hounsfield unit (HU) calibration methods, providing valuable insights into their impact on proton dose distributions in small animals.

The findings and implications of these works as well as remaining issues and potential avenues for future research are discussed in the following sections.

5.2. DEEP LEARNING FOR AUTO-CONTOURING IN SMALL ANIMALS

While deep learning-based auto-contouring tools are being actively developed for clinical use [23, 24, 25, 26], their application specifically for small animals has only emerged in recent years [27, 28, 22, 29, 30, 31, 32, 33]. Noteworthy work by Schoppe et al. [28] is the AIMOS (AI-based Mouse Organ Segmentation) pipeline based on a 2D U-Net, which demonstrated that deep learning is also a powerful tool in delineating organs-at-risk (OAR) in mice. At the time of its publication, AIMOS established itself as the state of the art in the preclinical world. However, there are a few caveats to their work. For instance, the segmentations of the μ-CBCT images [34] they used in the training exhibit inaccuracies in organ contours as illustrated in figure 5.1. Since deep learning models learn to delineate structures from the training data, the quality of DL-based contours depends on the accuracy of the reference contours. Errors in the training data will be propagated, potentially leading to suboptimal predicted contours. Although their results show high values in similarity measures, this only means that the AIMOS-based contours are comparable to the reference contours, not necessarily that they are of high quality. Moreover, they tested their models exclusively on an independent dataset of the same type as the training data. Therefore, caution is warranted when using their trained models, as they might not perform as well on new data with a different distribution, such as CT images of mice with different strain or age or CT images taken using a different imaging protocol.

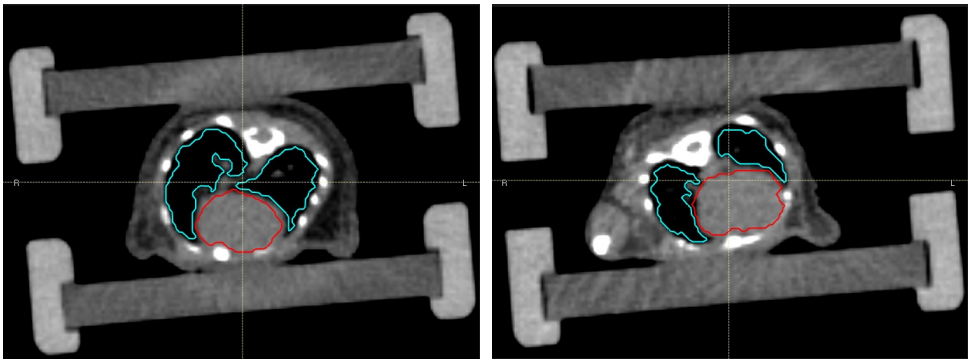


Figure 5.1: Examples of the training data used in developing AIMOS [28, 34]

QUALITY OF MANUAL SEGMENTATIONS

The DL contouring work for mice presented in **Chapter 2** improves upon Schoppe et al.'s [28] work by implementing stricter guidelines for creating the manual segmentations of the Rosenhain dataset [34]—the same dataset used by Schoppe et al.—to train the deep learning algorithm. Given that these manual segmentations serve as the ground truth from which the DL models learn, ensuring high-quality segmentations in the training gives a certain level of confidence on the reliability of the trained model and on the quality of the predicted contours. As demonstrated by Lappas et al. [35], clear contouring protocols also lead to smaller interobserver variability for organs-at-risk in the thorax and head of rodents. Adhering to clear contouring guidelines also reduces inconsistencies when assessing normal tissue volumes, which is important when radiation-induced side effects are the relevant endpoints in the experiment. While consensus guidelines for delineating organs-at-risk in humans are well-established [36, 37], the standardization of contouring practices is far less common for small animals. Although some institutions have likely developed their own contouring protocols [35], establishing an internationally agreed-upon guideline would be highly beneficial. Standardization would enhance the robustness of deep learning models for contouring and potentially facilitate their use across multiple institutions. However, we recognize that achieving this is not straightforward and may require concurrent standardization of other practices such as imaging protocols.

2D VS 3D NEURAL NETWORKS

The deep learning work was further extended by training not only 2D but also 3D neural networks, all of which are U-Net-based architectures [38]. Utilizing the nnU-Net pipeline

[39], all available configurations (2d, 3d_fullres, 3d_lowres, and 3d_cascade) were trained and their performance was compared to the 2D U-Net-based AIMOS [28]. While 2D networks are faster—in our case, the 2D models predicted contours in half the time of 3D networks—they suffer from the loss of craniocaudal information. In contrast, 3D networks leverage volumetric information, capturing spatial relationships along the longitudinal direction, which may be missed when processing individual CT slices in 2D. For example, the nnU-Net 2d model, in some cases, misclassified voxels in the anaesthesia nozzle (located far from the lungs) as lung tissue due to their similarly low HU values. Moreover, both nnU-Net 2d and AIMOS occasionally misclassified the left as the right lung and vice versa. These errors did not occur with the 3D models. Despite this, 2D models should not be disregarded outright. For the test set comprising of native CT scans, which share the same properties as the training data but were not used in the training, the predicted contours from the 2D networks are quite comparable to those from the 3D models. Errors due to isolated voxels delineated away from the organ can be easily removed through connected component analysis, although this method is more effective for single, whole organs and less suitable for complex structures with numerous interconnected parts like the skeleton. The outliers also constituted only a small fraction of the total volume, which can be easily corrected manually. This suggests that while 2D networks may require constant monitoring, which we believe should anyway be the standard practice for any DL-based contouring software, they remain a viable option. They typically require less computational power and GPU memory, leading to lower cost and, more importantly, faster training and inference times. This is particularly advantageous for time-sensitive applications like animal contouring. In this work, contouring time for OARs in the thorax of a mouse was reduced substantially from 40 minutes with a human expert to around 20 seconds using the nnU-Net 2d and AIMOS DL models. The nnU-Net 3d_fullres, which produced the most accurate segmentations, completed the same task in 40 seconds. The significant reduction in contouring time not only streamlines the preclinical workflow for online irradiations but also aligns with the 3R principle of *Refinement* for animal experiments, which seeks to reduce animal burden and improve their welfare.

EXTERNAL VALIDATION

In radiobiological studies, tens or even hundreds of animals are often used in a single study. These animals, being of the same strain, age, and sex, and subjected to identical diet, activity, and environmental conditions, exhibit minimal anatomical variations compared to humans. Thus, a deep learning model trained on such a homogeneous

dataset, provided that high quality contours were used in the training, is expected to perform well within this specific domain. However, variations in animal models used in experiments are likely in a multi-user facility. Given the large number of animals to be imaged, the expertise required, and the time needed for manual delineations, it is impractical to train a separate model for each experiment. Therefore, auto-contouring models that are robust across a range of cases are preferable.

To determine how well the 2D and 3D models generalize to new data, despite being trained exclusively on one type of image (native CT), they were tested against the contrast-enhanced CT (CECT) dataset of Ronsehain et al. [34], which involved a different mouse (varying in strain and age) and were taken using different image acquisition parameters. The external validation revealed the limitations of the 2D models, which often struggled with contour predictions, resulting in multiple slices being partially or fully missed. In contrast, the best 3D model (nnU-Net 3d_fullres) significantly outperformed nnU-Net 2d and AIMOS, demonstrating greater robustness to out-of-distribution data. Despite a slight decline in performance with CECT compared to native CTs, the predictions of nnU-Net 3d_fullres on the CECT dataset did not exhibit substantial qualitative differences to the manual contours.

The results demonstrated that although the 2D models are faster to deploy and worked reasonably well within the training domain, they lack the robustness of 3D models, which makes them less reliable to use in reality especially at a multi-user environment. Additionally, the need for manual corrections can further prolong the contouring process. On the other hand, 3D models offer better generalization to unseen data but require more time and resources, which could also be a hindrance in practice. This raises the question of which approach is more suitable for preclinical applications where time constraints are a significant factor. The choice largely depends on the situation and expectations of the users. If computational resources are limited and the training data to which the model was trained on is representative of the intended target, 2D models would probably be enough. However, if the workflow can accommodate the extra time and resources required, 3D models may be the better option. Alternatively, one can train the model on a dataset encompassing various cases, which can make more generic and robust models regardless of the approach.

AVAILABILITY OF ANIMAL CTS

While it is desirable to create a diverse training dataset to improve the robustness of deep learning models, this task is particularly challenging in the preclinical community. Ideally, one should curate a dataset with a broad distribution, which could include images

of animals with different strains, ages, and sexes as well as with variations in weight, posture, and orientation or images acquired with different imaging protocols. However, strict regulations governing the use of animals in experiments make it difficult to image a large number of animals solely for the purpose of building such a comprehensive dataset. It is also essential to emphasize that creating high-quality manual segmentations for these images would require significant amounts of time, effort, and expertise of a biologist. Currently, most DL-based auto-contouring studies on animals do not share their dataset [22, 29, 27] except those based on publicly available data [28, 30]. In this thesis, although the training data was also sourced from a public database, the manual segmentations were recreated, following stricter contouring guidelines to ensure quality. These segmentations have been made publicly available, and other researchers are strongly encouraged to do the same. This kind of practice will make it easier to create large datasets with a wide variety of cases on which deep learning models can be trained. This, in turn, could improve their robustness and generalizability across different scenarios. A common dataset also promotes proper and fair comparison of DL auto-contouring models across multiple institutions by eliminating variation in the training data, ensuring that differences lie in the deep learning implementation rather than the data used. Sharing datasets also lessens the need for individual institutions to repeatedly invest resources towards creating manual annotations from scratch, which is particularly beneficial for smaller institutions with limited manpower and expertise. However, shared datasets would only be useful for training and validating deep learning models if participating institutions adhere to a consensus guideline for contouring, avoiding variations in interpretation of what should and should not be included in the organ segmentations. To prevent bias in the training data, information on the CT imaging system, acquisition protocols, reconstruction and calibration methods must also be provided.

LIMITATIONS AND FUTURE PERSPECTIVES

The public mouse dataset [34] utilized in this work contained two types of data: native CT and CECT. Our approach was to only use the native CT images in the training and the CECT data was mainly intended for external validation to assess how well the models handle images that are significantly different from the training set. Results showed that the 2D models severely underperformed, while the best 3D model, although producing acceptable contours, demonstrated reduced performance on the CECT data as expected. In the future, improving the performance on the CECT images is possible either by training a separate model only on the CECT dataset or merging both datasets to create a single, larger, and more diverse training set. However, training a separate deep

learning model for each dataset would result in having multiple auto-contouring models in the workflow, which could hinder seamless integration and operational efficiency. On the other hand, when creating diverse datasets, it is essential to ensure balanced representation of all image types in the training set to prevent sample bias and achieve more reliable predictions.

The deep learning work in this thesis was focused exclusively on OARs in the thorax of a mouse. Since the training data consisted solely of CT images from healthy mice, it would be valuable to investigate whether the DL models also perform well when applied to mice with lung tumours. Moreover, the auto-contouring efforts can be easily extended to other body sites although a drop in performance in regions like the abdomen is expected due to poorer soft tissue contrast. This lack of contrast also poses challenges for human experts tasked with creating the manual segmentations. Administering contrast agents could be a potential solution to enhance organ visibility. Additionally, the deep learning techniques can be adapted to complementary imaging modalities typically used for animal experiments like dual-energy CT, magnetic resonance imaging (MRI), positron emission tomography (PET), and bioluminescence imaging (BLI) to improve tumour delineation and tracking of disease progression in animal models.

Since atlas-based segmentation still takes over 10 minutes to delineate organs in an animal [40, 41], the results in **Chapter 2** of this thesis clearly demonstrate that deep learning, once properly trained, has the potential to increase efficiency and facilitate online irradiation of animals. Our implementation achieves contouring times of under 1 minute, which could be further accelerated by using more powerful graphics processing units (GPUs) or by optimizing the architecture and network parameters. However, additional work is still needed to fully implement DL auto-contouring in-house, integrate it seamlessly into existing workflows, and make it accessible in the animal treatment planning system. Quality assurance procedures also need to be established to detect erroneous predictions even after deployment in practice.

Furthermore, we recognize that our evaluation of the DL-based contours is somewhat premature, only relying on geometric comparisons with manual contours. While these measures reveal differences in the shape of the contours, they do not provide an insight on how these differences impact the resulting dose distribution or their correlation. To verify the significance of the observed geometric differences, extensive dosimetric validation must be carried out. Additionally, as shown in a recent study on animal auto-contouring [29], expert evaluation of the contours by rating them on a scale of 0 to 5—where 0 indicates the predicted contour is unacceptable and must be redone from scratch, and 5 signifies no manual adjustments are needed for the predicted contour—

could offer further insights into the acceptability of DL-based contours. A blind evaluation can also be conducted, wherein manual and automated contours are presented for assessment without prior knowledge of their origin. These steps could further promote confidence in the community to adopt DL auto-contouring tools in preclinical workflows.

5.3. PROTON DOSE CALCULATIONS IN SMALL ANIMALS

To ensure the effective delivery of radiation beams with small animal irradiation platforms, it is essential to have tools for creating irradiation plans and calculating dose distributions in animals. Given that planning occurs shortly before irradiation, these tools must be efficient and capable of meeting the fast-paced demands of preclinical practice. Unfortunately, general purpose Monte Carlo (MC) codes, despite offering the most accurate dose distributions, are far too slow for time-sensitive applications. In MC simulations, individual particles along with any secondary particles they generate (if included in the simulation) are tracked step-by-step within the geometry. To achieve precise dose distributions, a large number of particles must be simulated to ensure statistical convergence, which results in significant computational burden. Although considered as the gold standard for dose calculations [42, 43], the long computation time has rendered MC simulations impractical for clinical applications [44]. Therefore, commercial treatment planning systems in the clinic often rely on analytical pencil beam algorithms (PBA) for their speed [45, 46]. However, these algorithms fall short in accurately modelling the dose distribution in highly heterogeneous regions. More recently, advancements in computer technology and increasing interest in rapid, patient-specific approaches, such as online adaptive radiotherapy [47, 48] and robust optimization [49, 50], have spurred the development of fast proton dose engines [51, 52, 53, 54, 55, 56, 57, 58]. These implementations often employ simplified physics, deterministic techniques, or deep learning approaches, along with the computational power of multiple central processing units (CPUs) or graphics processing units (GPUs), to accelerate dose calculations.

Given that lengthy dose computation times impose a bottleneck in the preclinical workflow, **Chapter 3** explores the feasibility of leveraging these fast dose calculation platforms to meet the demand for efficient and timely dose calculations in small animals. Specifically, the fast Monte Carlo code MCsquare [54] and the deterministic algorithm YODA [57, 59] were investigated. While these proton dose calculation platforms have been previously validated for human use, which typically involve treatment energies above 70 MeV and CT voxel sizes $\geq$ 1 mm, most small animal irradiations are conducted with energies below 40 MeV and require much finer voxel sizes around 0.1 mm.

Therefore, alongside evaluating the efficiency gains provided by these codes, a key contribution of this work is extending their validation to these smaller scales for potential preclinical applications.

ACCURACY OF THE PREDICTED PROTON RANGE

In **Chapter 3**, both MCsquare and YODA were benchmarked against the general-purpose MC code TOPAS [60], evaluating their performance on simple phantoms and more complex mouse phantom geometries [61]. In all simulations, we incorporated a voxel size of 0.1 mm, which is standard for small animal imaging systems. Validation on homogeneous (0 HU) and heterogeneous (i.e. water phantom with lung and bone inserts) phantoms at energies pertinent to small animal irradiations demonstrated that both MCsquare and YODA accurately reproduced the proton range observed in TOPAS to within 0.1 mm, even in the presence of density inhomogeneities in the beam path. The good agreement in the proton range was achieved by ensuring that the CT HU-to-density conversion, material conversion, and stopping power tables were consistently matched across all three codes. For MCsquare, stopping power tables of materials were directly extracted from TOPAS, though the energy binning was modified to meet the specific requirements of MCsquare. The YODA stopping power tables were also sourced from TOPAS, but it employs a slightly different method by requiring elemental stopping powers instead and calculating the stopping power of compounds through the Bragg additivity rule. This methodological difference did not influence the proton range in YODA as TOPAS/Geant4 calculates the stopping power tables in the same way internally by default. However, it is important to note that such an approach may not always yield accurate results. With Bragg additivity, the elemental mean excitation energies (I) are used in the calculation of the stopping power of a compound, which neglects molecular binding effects. For instance, while Bragg additivity estimates the I-value of water at 75 eV [62], reported values for water range from 67.2 eV to 82.4 eV, with the current recommendation being 78 eV for proton treatment planning [63]. Previous works have already shown that these variations in the I-value used in stopping power calculations result in range differences [64, 65, 66]. As uncertainties in the I-value influence all three codes, an important next step is to perform actual measurements to verify our calculations.

DEALING WITH LATERAL HETEROGENEITIES

MCsquare, as a Monte Carlo code that simulates the propagation of individual particles, is highly effective at modelling dose distributions in heterogeneous regions. On the other hand, YODA, which relies on the analytical Fermi-Eyges theory to model multiple Coulomb scattering (MCS), inherits the intrinsic limitations suffered by pencil beam algorithms when dealing with lateral heterogeneities. The simulations on a heterogeneous phantom (figure 3.5) containing adjacent lung and bone slabs, with the material interface positioned directly at the centre of the beam, illustrate the limitation of YODA's MCS implementation. Although the expected outcome is to observe two distinct Bragg peaks—one associated to protons traversing the lung, which extends deeper, and another associated to protons passing through the bone, which stops earlier—YODA produces only a single Bragg peak as it considers only one of the two materials and disregards any other off-axis heterogeneities in the beam path. To accurately describe changes in the range and lateral beam spreading caused by heterogeneities, spot decomposition into narrower beamlets was implemented, a technique commonly employed in pencil beam algorithms. However, it is important to note that this approach leads to increased calculation times. For example, as shown in table 3.4, a 1+6+6+12 split scheme with a total of 25 beamlets took 0.9 s, whereas a 1+24+24+24+24 split scheme with a total of 97 beamlets took 1.9 s.

Unlike other beam splitting methods that position sub-spots on a grid with their origins offset laterally from the central axis of the original beam [67, 68, 69], the beam splitting approach used in this work takes advantage of the radial symmetry of the Gaussian beam by placing sub-spots on concentric circles around the original beam's central axis. Moreover, contrary to the approach of Yang [70] and the commercial TPS RayStation, where splitting is limited to a number of pre-defined schemes, this method is customizable, allowing for flexible adjustments in the number of concentric rings and the distribution of spots within each ring. Several configurations were tested in **Chapter 3**, ranging from the simplest 1+6+6+12 scheme to the more intricate 1+24+24+24+24 scheme, incorporating denser spot distributions and additional rings.

Single spot simulations across three heterogeneous phantom configurations, which varied in the position of the heterogeneity relative to the beam, revealed that increasing the density of the spots near the material interface results in more accurate dose distributions. The dosimetric evaluation for the heterogeneous phantom cases 1, 2, and 3 resulted in 3%/0.1mm gamma pass rates of 83.2%, 82.0%, and 81.0%, respectively, for the 1+6+6+12 scheme. These results are inferior to the 95.1%, 88.4%, and 91.4% pass rates observed for the 1+24+24+24+24 scheme for the same cases. However, when applied to

full plan calculations, the 1+24+24+24+24 scheme offered only a marginal improvement in the gamma pass rates compared to the 1+6+6+12 scheme. When multiple spots are delivered in close proximity, the discrepancies introduced by the split schemes tend to average out, reducing their impact on the overall dose distribution. This suggests that the increased complexity of the 1+24+24+24+24 scheme does not provide a substantial accuracy benefit over simpler schemes and only prolongs the calculation time. For full plan calculations, which involves tens to hundreds of spots, simpler schemes may be preferable as they are more efficient, while still achieving comparable dose distributions.

It is also important to point out that we initially optimized the parameters of the split schemes (i.e. ring radii, spatial spreads, and weights of the beamlets) to achieve the highest gamma passing rates for case 1, which featured lung and bone inserts located far upstream from the Bragg peaks. These optimized parameters were then applied to the other two heterogeneous phantom cases, where the heterogeneities were closer to the Bragg peaks, as well as to full animal plan calculations. This choice was deliberately made to assess the robustness of the optimization across varying conditions and heterogeneity distributions. The results indicate that while the optimized split schemes for case 1 worked well in relatively homogeneous regions, such as the head and abdomen with a lateral field, they were less effective in regions with more complex heterogeneities, like the thorax and the abdomen with an anterior field. Additionally, the optimized schemes showed reduced performance in cases where the Bragg peak was positioned closer to the heterogeneity.

To improve results, fine-tuning the split schemes for YODA for each specific experiment may be necessary. However, given the time constraints in animal irradiations, real-time or on-the-fly optimization is not practical. Instead, follow up optimization studies involving different animals (e.g. mice, rats, rabbits), anatomical regions, voxel resolutions, and beam configurations (e.g. single or multiple fields, co-planar and non-coplanar beams) are required to establish standardized solutions. By doing so, a comprehensive lookup table for split schemes can be created, thereby eliminating the need for optimization during irradiation planning and minimizing user intervention.

IRRADIATION PLANNING IN SMALL ANIMALS

The open-source treatment planning system OpenTPS [71], originally developed for proton therapy planning in humans, was adapted for use in animals in **Chapter 3**. At the time of writing, some limitations were encountered that hindered the creation of optimized plans for mice. Notably, the software's restriction on spot placement at shallow depths posed significant challenges, given the small size of animals and the shallow po-

sitioning of the target volumes. This highlights the importance of exercising caution in applying tools intended for human use to animals. Moreover, the dose optimization criteria in OpenTPS were limited to mean, minimum, and maximum doses in structures, lacking DVH related objectives typically available in commercial TPS. These constraints contributed to suboptimal planning outcomes in animals. Although the plans we created did not achieve optimal dose coverage, the comparison of the dose distributions calculated from MCsquare, YODA, and TOPAS remains valid since the same plan was consistently used for each code.

In the future, a dedicated irradiation planning system for small animals, such as μ-RayStation, would be more suitable. While μ-RayStation is designed for preclinical use, it has not yet been validated for protons, so thorough validation must be performed.

BEAM MODELLING OF THE IMPACT BEAMLINE

To enable the use of OpenTPS for animal irradiation planning, a realistic beam model of the preclinical IMPACT beamline at our institute was built from beam transport simulations of a 66.5 MeV proton beam (**Chapter 4**). To make the beam suitable for animal irradiations, a range shifter and collimator system was incorporated at the end of the beamline. The collimator was placed downstream of the range shifter to cut-off scattered particles, with its exit positioned 2 cm from the animal container to minimize the contribution of air scatter. Although OpenTPS supports modelling of the range shifter, it does not offer functionality for simulating the collimator. Consequently, to create the beam data library (BDL), we had to perform separate simulations for each range shifter thickness. In each configuration, the beam phase space was recorded at the entrance of the animal container, from which beam parameters such as the spot size, divergence, energy, and energy spread were derived.

It is important to note that the spot size, divergence and energy spread are modelled as a Gaussian in OpenTPS. However, as shown in the beam characterization in figure 5.2, these parameters do not closely follow a Gaussian distribution. The spot profiles exhibit sharp penumbra and a flat top due to the collimation, while the divergence shows large angle tails. The low energy tail in the energy distribution also becomes more prominent with increasing range shifter thickness. Despite these differences, using a Gaussian approximation of the beam parameters in the BDL is acceptable for this study. Since the aim is to compare the codes, ensuring a consistent beam definition at the start of each simulation across all codes preserves the validity of the comparison. However, if OpenTPS is to be used for irradiation planning in preclinical practice, experimental measurements must be conducted to verify the validity of these approximations and ensure

that the planned dose accurately reflects the delivered dose.

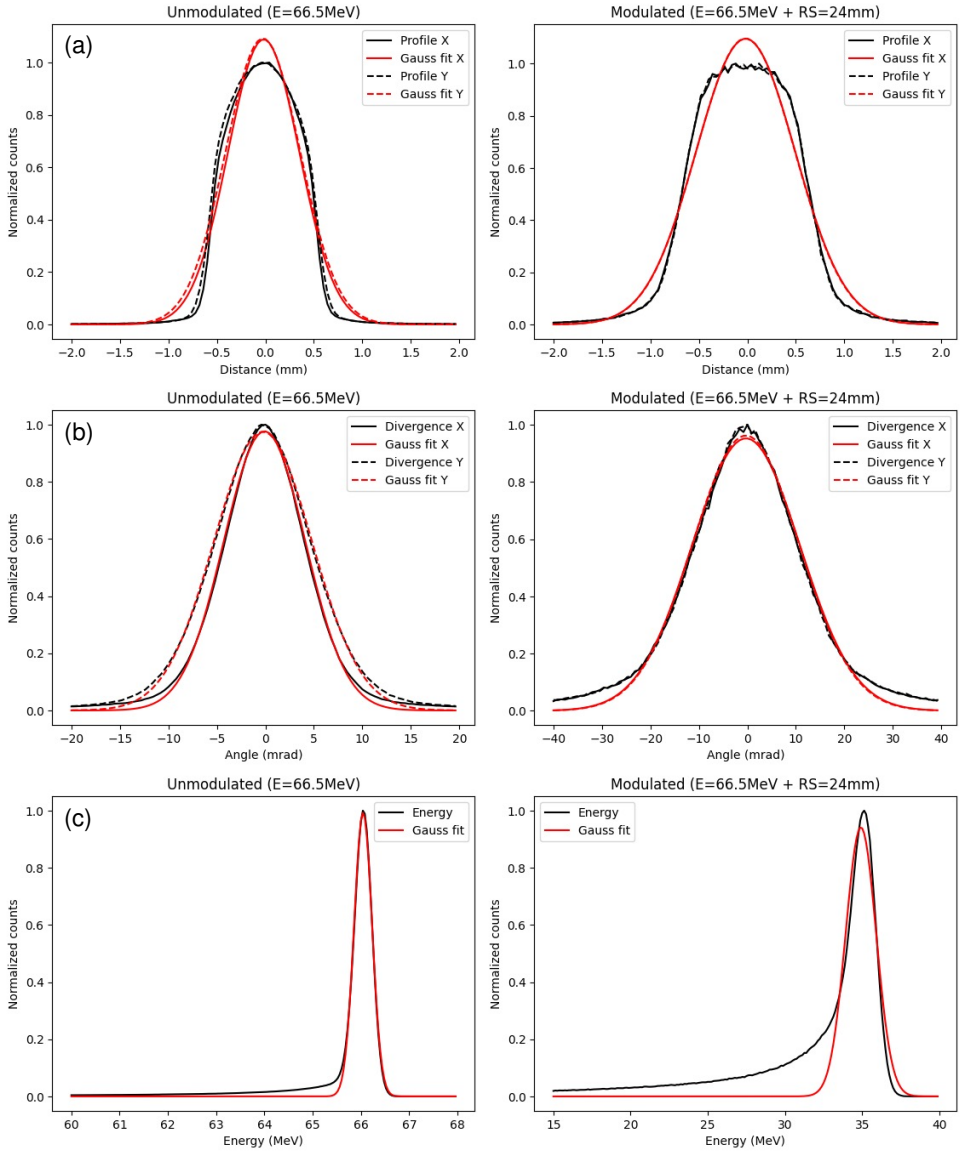


Figure 5.2: (a) Spot profile, (b) divergence, and (c) energy distribution for the unmodulated (i.e. 66.5 MeV proton beam) (first column) and modulated beam (i.e. 66.5 MeV proton beam with 24 mm range shifter) (second column) at the entrance of the animal container.

Moreover, unlike in the clinic, where treatment machines are generally standardized,

preclinical beamlines can vary widely depending on the facility, experimental setup, and specific requirements of each study. Therefore, it should be an important consideration for developers and vendors of animal irradiation planning systems to support the use of variable collimators, range shifters, ridge filters, and other beamline components to accommodate these setup variations. Another key consideration is how to incorporate into the planning system the fact that the beam and beamline components are typically fixed, with spot scanning achieved by moving the animal, in particular for very small fields.

DOSE EVALUATION IN ANIMALS

To assess the accuracy of the proton dose calculations, two types of evaluation were performed: (1) range shift and (2) 3D gamma analysis. The range shift analysis reveals how well the codes predict the proton range, which is crucial given the small targets and tight margins used in preclinical studies. While a 1-mm range shift might appear minor in patients, it can lead to substantial underdosing of the target and overdosing of adjacent healthy tissues in small animals. On the other hand, the gamma analysis is one of the most commonly used metrics for dose comparison, providing insights into how closely two dose distributions match and pinpointing where discrepancies lie. While acceptance criteria for gamma analysis are well established in the clinic, there is lack of consensus on the appropriate values for dose difference (DD), distance-to-agreement (DTA), and passing criteria when applied to animal studies. In **Chapter 3**, various tolerance limits were applied by fixing the DD at 3%, in line with our institute's dose homogeneity criteria, and varying the DTA at 0.1,0.2, and 0.3 mm, based on the spatial resolutions achievable with small animal imaging systems [72, 73]. A DD/DTA criteria of 3%/0.2mm were ultimately adopted for evaluating full plans, with gamma passing rates greater than 95% considered acceptable. A DTA of 0.1 mm was deemed too strict for the evaluation as it is highly sensitive to variations between adjacent voxels that may not have significant dosimetric implications. For context, van Dijk et al. [74] used a DD/DTA criteria of 3%/0.3mm, based on clinical guidelines of 3%/3mm, whereas Vanstalle et al. [75] applied a dose difference range of +7% to -5%, based on ICRU 62 recommendations on the acceptable dose heterogeneity for X-ray radiotherapy [76], and DTA of 0.1 mm, reflecting the targeting accuracy recommended for mice [77]. In order to facilitate more meaningful comparison, interpretation, and evaluation of the quality of dose distributions, the preclinical community should establish consensus guidelines for gamma analysis specific to animal studies. These guidelines would support more informed decision-making and ensure consistent reporting of outcomes in preclinical research.

Although not utilized in this study, dose evaluation using dose volume histogram parameters should also be considered. While the gamma analysis helps assess the voxel-wise similarity between two dose distributions, it does not adequately reflect how local dose differences impact relevant anatomical regions. In some cases, even if the gamma analysis indicates a failure, large dose differences may be permissible in areas outside critical OARs, provided that the dose remains within prescribed limits to prevent complications. Unfortunately, due to the suboptimal quality of the plans generated in this study, interpreting DVH parameters proved challenging, leading to their exclusion from the evaluation.

MCSQUARE VS YODA

While the work presented in **Chapter 2** aimed at reducing the contouring time through deep learning automation, **Chapter 3** examines the suitability of fast proton dose engines (MCsquare and YODA) in speeding up dose calculations for animals, targeting another stage that hampers the efficiency of the preclinical workflow. The results on both simple phantom geometries and SOBP plans involving various anatomical sites in a mouse clearly demonstrate the superior performance of MCsquare compared to YODA in predicting dose distributions. This is not surprising, as MCsquare uses a Monte Carlo method, whereas YODA employs a semi-analytical approach. MCsquare consistently met the gamma acceptance criteria, achieving pass ratios (3%/0.2mm) above 97% in all cases. Conversely, while YODA proved effective in relatively homogeneous regions like the head (99.8%) and abdomen with lateral field (96.8%), it struggled with more heterogenous sites like the thorax (83.3%) and abdomen with anterior field (93.2%). For these worst-case scenarios, the dose discrepancies in YODA were observed predominantly at the distal end of the dose distribution, which generally had minimal impact on the dose distribution in the target volume. The observed dose degradation is attributed to less optimal beam splitting schemes used in these more complex cases and refining these schemes could enhance YODA's performance to a level comparable to MCsquare. Additionally, it must be pointed out that both MCsquare and YODA exhibited higher doses in regions with very low-density materials due to the lack of detailed electron transport. Therefore, caution is warranted when such low-density regions play a critical role in the optimization and assessment of the dose.

Although MCsquare demonstrated greater versatility in dose computations, our runtime comparisons for a single spot showed that it is considerably slower than YODA. Even when employing a split scheme that divides the spot into 97 sub-spots, YODA completed the calculations within 2 s, whereas MCsquare took 25 s. Future integration of

GPU support could also further accelerate YODA's performance. This highlights YODA's speed advantage, making it the more efficient option, particularly for applications that require numerous rapid calculations, such as plan optimization, where the 3D dose distribution for each individual spot in the plan must be computed. To benefit from both MCsquare's accuracy and YODA's speed, we recommend using YODA for plan optimization and then combining it with MCsquare for the final dose calculations to ensure high quality dose distributions.

Moving forward, GPU-based MC codes [53, 56] and deep learning-based dose calculators [55, 74] could also be explored to further expedite dose computations for preclinical applications. However, rather than focusing solely on seeking faster dose engines, we argue that automating the entire planning process is equally important. For example, the clinical TPS RayStation now offers deep learning planning [78], where a 3D dose distribution is initially predicted by analysing the target and OARs and based on this prediction, a plan is automatically generated without the need for time-consuming iterative plan optimization. Such automation would greatly reduce the time between imaging and radiation delivery, further improving the overall efficiency.

5.4. PROTON RANGE UNCERTAINTIES IN SMALL ANIMALS

Proton beams can deliver more conformal dose distributions but suffer from range uncertainties. Aside from anatomical changes and setup errors, proton range accuracy is also heavily influenced by the CT calibration in treatment planning. The most commonly used method is the single energy CT (SECT) stoichiometric calibration [79, 80], which has been reported to introduce range uncertainties of up to 3.5% in current clinical practice [81, 82, 83]. Dual energy CT (DECT), however, has shown promise in further reducing these uncertainties [84, 85, 86, 87]. While previous studies have evaluated SECT and DECT methods for tissue characterization in kilovoltage X-ray irradiation planning in animals [88, 89], there have been no reported investigations into their application for protons in the preclinical context. Given the tighter margins imposed by the small size of animals and targets within, the work described in **Chapter 4** focused on determining the accuracy benefits of DECT-based proton irradiation planning compared to SECT in animal studies.

SIMULATION FRAMEWORK

To conduct SECT and DECT evaluations without the use of live animals, a simulation framework for the preclinical proton irradiation workflow was developed. This includes a μ-CBCT model, adapted from a preclinical X-ray imager, which was integrated into

the fastCAT CBCT toolkit initially developed for simulating CBCT scans of humans [90]. The CT numbers from the simulated μ-CBCTs of a preclinical phantom generated by the model closely matched those obtained from full MC simulation in TOPAS, validating its accuracy and applicability.

When comparing these CT numbers to those reported in the literature measured for the same phantom [89], the fastCAT-generated values followed the same trend but were slightly higher. This is not surprising, given that the simulations we performed to create the μ-CBCT model were relatively simplistic and did not incorporate collimators, filters, and other CT components, which could introduce additional scatter. Moreover, the detector response functions, reconstruction protocol, and region of interest for HU value extraction were not identical. Future work could benefit from more detailed simulations to allow a fair and proper comparison with experimental values. Nevertheless, these variations are not critical to the objectives of this study, which focus on evaluating CT calibration methods. As long as the same "CT system" is consistently used, the conclusions drawn from the assessment of the calibration methods remain valid.

It is also important to note that the use of simulated CT images offers the advantage of having a ground truth image for comparison. Because the elemental composition of the tissues assigned to the geometry used to generate the simulated CT scans are known, these same materials can be used to create the ground truth image. This is in contrast with experimental CT scans of animals, where the actual tissue composition is unknown. The lack of ground truth image in experimental studies complicates the comparison of the dose distribution resulting from SECT and DECT calibrations, as there is no definitive reference for the evaluation.

A model of the dedicated preclinical IMPACT beamline was also created in the Geant4-based BDSIM toolkit [91] for Monte Carlo beam transport simulations. This model is useful for developing and optimizing beam properties and experimental setups tailored to the specific needs of the study. It provides a platform for identifying potential issues and weaknesses in experimental setups prior to performing actual experiments, allowing for pre-emptive adjustments. This helps conserve beam time and resources, thereby enhancing the overall efficiency of the workflow.

For the work in **Chapter 4**, this beamline model was used to generate the beam data library for the matRad treatment planning system [92] for dose calculations. For simplicity, a single simulation at full energy (E = 66.5 MeV) was performed to characterize the beam and the pull-back method was used to derive the beam data for lower energies. It should be noted that this method is valid only when the range shifter is positioned directly in front of the imaged object and is water equivalent in terms of multiple scat-

tering. The same beamline model from this framework was used in **Chapter 3** to create the beam data library for OpenTPS, but with the range shifter placed upstream of the collimator. For this setup, the pull back method is not applicable, requiring separate simulations for each energy to build the proton beam model in the irradiation planning system. Future work should include experimental validation of the proton beam models in matRad and OpenTPS developed for preclinical proton irradiations.

SECT vs DECT

Using the simulated μ-CBCTs of a mini-calibration phantom generated by fastCAT, the accuracies of SECT and DECT stopping power ratio (SPR) predictions for tissue equivalent materials were compared. To improve upon previous work, various imaging protocols were tested, including stronger filtration, to further refine DECT performance. Two DECT algorithms were also evaluated: Landry et al.'s approach [93], which has been previously applied in small animal DECT studies for X-ray irradiation [88, 89], and Saito and Sagara's method [94], which was recently shown to yield superior DECT results in the clinical context [84]. Consistent with findings in the literature, our results demonstrated that DECT can improve tissue characterization, yielding lower root mean square deviations (RMSD) in the SPR compared to SECT for Gammex materials. For the best imaging protocols, the RMSD for SECT was 3.7%, while DECT reduced it to 1% across the ten materials evaluated in this study, with both DECT methods showing comparable performance. Contrary to expectation, the DECT imaging protocol that utilized stronger filtration (i.e. greater energy separation) exhibited poorer performance. It is left to future work to experimentally confirm these findings.

Although this evaluation clearly highlights DECT's advantage in predicting SPR, it does not provide insights on its impact on range accuracy. As an initial step, proton radiographs of a mouse phantom [61] were simulated in TOPAS using the SECT and DECT calibrated CTs, and these were compared to the proton radiograph of the ground truth image. Range errors were quantified by taking the voxel-wise differences in water equivalent thickness (WET) between the calibrated CTs and the ground truth. In line with earlier findings, DECT demonstrated smaller mean shifts and less variation in WET values (i.e. proton range) compared to SECT. However, proton dose calculations showed that both SECT and DECT methods achieved high gamma passing rates (> 99%) and minimal range shifts (< 0.1 mm). These outcomes show that DECT did not offer substantial accuracy benefit for proton dose calculations in small animals. SECT is adequate for proton irradiation planning, and the additional dose burden from DECT imaging may not be justified. However, it should be emphasized that this is a single case study and addi-

tional work, including evaluations on different anatomical regions, is necessary to fully evaluate the merit of DECT-based proton irradiation planning in small animals.

5.5. CONCLUDING REMARKS

In conclusion, this thesis explored the application of deep learning for auto-contouring and fast proton dose engines for dose calculations, with the primary goal of streamlining irradiation planning for small animal studies. These tools greatly benefit the preclinical community by enhancing workflow efficiency, increasing experiment capacity, and reducing the overall workload of physicists and biologists. The reduction in planning time not only boosts animal throughput but also contributes positively to animal welfare. Additionally, the research on CT HU calibration methods has provided valuable insights into the benefits of SECT and DECT calibration for proton irradiation planning in preclinical settings. Ultimately, the work described in this thesis brings us one step closer to achieving more accurate and efficient image-guided irradiations of small animals for radiobiological studies.

BIBLIOGRAPHY

[1] Freddie Bray et al. “Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries”. In: *CA: A Cancer Journal for Clinicians* 74.3 (2024), pp. 229–263. DOI: 10.3322/caac.21834.

[2] World Health Organization. *New WHO/IAEA publication provides guidance on radiotherapy equipment to fight cancer*. 2021. URL: https://www.who.int/news/item/05-03-2021-new-who-iaea-publication-provides-guidance-on-radiotherapy-equipment-to-fight-cancer#:~:text=More%20than%2050%25%20of%20cancer,%2C%20colorectal%2C%20and%20lung%20cancer (visited on 08/08/2024).

[3] Marco Durante, Roberto Orecchia, and Jay Loeffler. “Charged-particle therapy in cancer: clinical uses and future perspectives”. In: *Nature Reviews Clinical Oncology* 14 (2017), pp. 483–495. DOI: 10.1038/nrclinonc.2017.30.

[4] Marco Durante and Jay Flanz. “Charged particle beams to cure cancer: Strengths and challenges”. In: *Seminars in Oncology* 46.3 (2019), pp. 219–225. DOI: 10.1053/j.seminoncol.2019.07.007.

[5] Jean-Pierre Pouget et al. “Clinical radioimmunotherapy—the role of radiobiology”. In: *Nature Reviews Clinical Oncology* 8 (2011), pp. 720–734. DOI: 10.1038/nrclinonc.2011.160.

[6] Aurélie Rondon, Jacques Rouanet, and Françoise Degoul. “Radioimmunotherapy in Oncology: Overview of the Last Decade Clinical Trials”. In: *Cancers* 13.21 (2021). DOI: 10.3390/cancers13215570.

[7] Ricky Sharma et al. “Clinical development of new drug–radiotherapy combinations”. In: *Nature Reviews Clinical Oncology* 13 (2016), pp. 627–642. DOI: 10.1038/nrclinonc.2016.79.

[8] Saif S. Ahmad et al. “Clinical Development of Novel Drug–Radiotherapy Combinations”. In: *Clinical Cancer Research* 25.5 (2019), pp. 1455–1461. DOI: 10.1158/1078-0432.CCR-18-2466.

[9] Jonathan R. Hughes and Jason L. Parsons. “FLASH Radiotherapy: Current Knowledge and Future Insights Using Proton-Beam Therapy”. In: *International Journal of Molecular Sciences* 21.18 (2020). DOI: 10.3390/ijms21186492.

[10] Michele M. Kim et al. “Development of Ultra-High Dose-Rate (FLASH) Particle Therapy”. In: *IEEE Transactions on Radiation and Plasma Medical Sciences* 6.3 (2022), pp. 252–262. DOI: 10.1109/TRPMS.2021.3091406.

[11] Eric S. Diffenderfer et al. “The current status of preclinical proton FLASH radiation and future directions”. In: *Medical Physics* 49.3 (2022), pp. 2039–2054. DOI: 10.1002/mp.15276.

[12] Y. Prezado and G. R. Fois. “Proton-minibeam radiation therapy: A proof of concept”. In: *Medical Physics* 40.3 (2013), p. 031712. DOI: 10.1118/1.4791648.

[13] P. Lansonneur et al. “First proton minibeam radiation therapy treatment plan evaluation”. In: *Scientific Reports* 10.7025 (2020). DOI: 10.1038/s41598-020-63975-9.

[14] Yolanda Prezado et al. “Spatially fractionated radiation therapy: a critical review on current status of clinical and preclinical studies and knowledge gaps”. In: *Physics in Medicine and Biology* 69.10 (2024), 10TR02. DOI: 10.1088/1361-6560/ad4192.

[15] John Wong et al. “High-Resolution, Small Animal Radiation Research Platform With X-Ray Tomographic Guidance Capabilities”. In: *International Journal of Radiation Oncology*Biology*Physics* 71.5 (2008), pp. 1591–1599. DOI: 10.1016/j.ijrobp.2008.04.025.

[16] R. Clarkson et al. “Characterization of image quality and image-guidance performance of a preclinical microirradiator”. In: *Medical Physics* 38.2 (2011), pp. 845–856. DOI: 10.1118/1.3533947.

[17] E Ford et al. “An image-guided precision proton radiation platform for preclinical in vivo research”. In: *Physics in Medicine and Biology* 62.1 (2016), p. 43. DOI: 10.1088/1361-6560/62/1/43.

[18] Michele M Kim et al. “Design and commissioning of an image-guided small animal radiation platform and quality assurance protocol for integrated proton and x-ray radiobiology research”. In: *Physics in Medicine and Biology* 64.13 (2019), p. 135013. DOI: 10.1088/1361-6560/ab20d9.

[19] Katia Parodi et al. “Towards a novel small animal proton irradiation platform: the SIRMIO project”. In: *Acta Oncologica* 58.10 (2019), pp. 1470–1475. DOI: 10.1080/0284186X.2019.1630752.

[20] Frank Verhaegen et al. “Roadmap for precision preclinical x-ray radiation studies”. In: *Physics in Medicine and Biology* 68.6 (2023), 06RM01. DOI: 10.1088/1361-6560/acaf45.

[21] Theresa Suckert et al. “Models for Translational Proton Radiobiology—From Bench to Bedside and Back”. In: *Cancers* 13.16 (2021). DOI: 10.3390/cancers13164216.

[22] Wouter R. P. H. van de Worp et al. “Deep Learning Based Automated Orthotopic Lung Tumor Segmentation in Whole-Body Mouse CT-Scans”. In: *Cancers* 13.18 (2021). DOI: 10.3390/cancers13184585.

[23] Lisanne V. van Dijk et al. “Improving automatic delineation for head and neck organs at risk by Deep Learning Contouring”. In: *Radiotherapy and Oncology* 142 (2020), pp. 115–123. DOI: 10.1016/j.radonc.2019.09.022.

[24] Tim Lustberg et al. “Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer”. In: *Radiotherapy and Oncology* 126.2 (2018), pp. 312–317. DOI: 10.1016/j.radonc.2017.11.012.

[25] Paul J. Doolan et al. “A clinical evaluation of the performance of five commercial artificial intelligence contouring systems for radiotherapy”. In: *Frontiers in Oncology* 13 (2023). DOI: 10.3389/fonc.2023.1213068.

[26] Charlotte L. Brouwer et al. “Machine learning applications in radiation oncology: Current use and needs to support clinical implementation”. In: *Physics and Imaging in Radiation Oncology* 16 (2020), pp. 144–148. DOI: 10.1016/j.phro.2020.11.002.

[27] Brent van der Heyden et al. “Automated CT-derived skeletal muscle mass determination in lower hind limbs of mice using a 3D U-Net deep learning network”. In: *Journal of Applied Physiology* 128.1 (2020), pp. 42–49. DOI: 10.1152/japplphysiol.00465.2019.

[28] Oliver Schoppe et al. “Deep learning-enabled multi-organ segmentation in whole-body mouse scans”. In: *Nature Communications* 11.5626 (2020). DOI: 10.1038/s41467-020-19449-7.

[29] Georgios Lappas et al. “Automatic contouring of normal tissues with deep learning for preclinical radiation studies”. In: *Physics in Medicine and Biology* 67.4 (2022), p. 044001. DOI: 10.1088/1361-6560/ac4da3.

[30] Justin Malimban et al. “Deep learning-based segmentation of the thorax in mouse micro-CT scans”. In: *Scientific Reports* 12.1822 (2022). DOI: 10.1038/s41598-022-05868-7.

[31] Francesco Sforazzini et al. “Deep Learning–based Automatic Lung Segmentation on Multiresolution CT Scans from Healthy and Fibrotic Lungs in Mice”. In: *Radiology: Artificial Intelligence* 4.2 (2022), e210095. DOI: 10.1148/ryai.210095.

[32] Elena Vincenzi et al. “A fully automated deep learning pipeline for micro-CT-imaging-based densitometry of lung fibrosis murine models”. In: *Respiratory Research* 23.308 (2022). DOI: 10.1186/s12931-022-02236-x.

[33] Martina Buccardi et al. “A fully automated micro-CT deep learning approach for precision preclinical investigation of lung fibrosis progression and response to therapy.” In: *Respiratory Research* 24.126 (2023). DOI: 10.1186/s12931-023-02432-3.

[34] Stefanie Rosenhain et al. “A preclinical micro-computed tomography database including 3D whole body organ segmentations”. In: *Scientific Data* 5.180294 (2018). DOI: 10.1038/sdata.2018.294.

[35] Georgios Lappas et al. “Inter-observer variability of organ contouring for preclinical studies with cone beam Computed Tomography imaging”. In: *Physics and Imaging in Radiation Oncology* 21 (2022), pp. 11–17. DOI: 10.1016/j.phro.2022.01.002.

[36] Charlotte L. Brouwer et al. “CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG Oncology and TROG consensus guidelines”. In: *Radiotherapy and Oncology* 117.1 (2015), pp. 83–90. DOI: 10.1016/j.radonc.2015.07.041.

[37] Romaana Mir et al. “Organ at risk delineation for radiation therapy clinical trials: Global Harmonization Group consensus guidelines”. In: *Radiotherapy and Oncology* 150 (2020), pp. 30–39. DOI: 10.1016/j.radonc.2020.05.038.

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Ed. by Nassir Navab et al. Springer International Publishing, 2015, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28.

[39] Fabian Isensee et al. “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation”. In: *Nature Methods* 18 (2021), pp. 203–211. DOI: 10.1038/s41592-020-01008-z.

[40] Brent van der Heyden et al. “Automatic multiatlas based organ at risk segmentation in mice”. In: *British Journal of Radiology* 92.1095 (2018), p. 20180364. DOI: 10.1259/bjr.20180364.

[41] Theresa Suckert et al. “High-precision image-guided proton irradiation of mouse brain sub-volumes”. In: *Radiotherapy and Oncology* 146 (2020), pp. 205–212. DOI: 10.1016/j.radonc.2020.02.023.

[42] Frank Verhaegen and Joao Seco, eds. *Monte Carlo Techniques in Radiation Therapy: Applications to Dosimetry, Imaging, and Preclinical Radiotherapy (2nd ed.)* CRC Press, 2021. DOI: 10.1201/9781003212485.

[43] Frank Verhaegen and Joao Seco, eds. *Monte Carlo Techniques in Radiation Therapy: Introduction, Source Modelling, and Patient Dose Calculations (2nd ed.)* CRC Press, 2021. DOI: 10.1201/9781003211846.

[44] Jatinder Saini et al. “Advanced proton beam dosimetry part I: review and performance evaluation of dose calculation algorithms”. In: *Translational Lung Cancer Research* 7.2 (2018). DOI: 10.21037/tlcr.2018.04.05.

[45] Linda Hong et al. “A pencil beam algorithm for proton dose calculations”. In: *Physics in Medicine and Biology* 41.8 (1996), p. 1305. DOI: 10.1088/0031-9155/41/8/005.

[46] Barbara Schaffner, Eros Pedroni, and Antony Lomax. “Dose calculation models for proton treatment planning using a dynamic beam delivery system: an attempt to include density heterogeneity effects in the analytical dose calculation”. In: *Physics in Medicine and Biology* 44.1 (1999), p. 27. DOI: 10.1088/0031-9155/44/1/004.

[47] Francesca Albertini et al. “Online daily adaptive proton therapy”. In: *British Journal of Radiology* 93.1107 (2019), p. 20190594. DOI: 10.1259/bjr.20190594.

[48] Lena Nenoff et al. “Experimental validation of daily adaptive proton therapy”. In: *Physics in Medicine and Biology* 66.20 (2021), p. 205010. DOI: 10.1088/1361-6560/ac2b84.

[49] Jan Unkelbach et al. “Robust radiotherapy planning”. In: *Physics in Medicine and Biology* 63.22 (2018), 22TR02. DOI: 10.1088/1361-6560/aae659.

[50] Chenbin Liu et al. “Robust Optimization for Intensity Modulated Proton Therapy to Redistribute High Linear Energy Transfer from Nearby Critical Organs to Tumors in Head and Neck Cancer”. In: *International Journal of Radiation Oncology*Biology*Physics* 107.1 (2020), pp. 181–193. DOI: 10.1016/j.ijrobp.2020.01.013.

[51] Matthias Fippel and Martin Soukup. “A Monte Carlo dose calculation algorithm for proton therapy”. In: *Medical Physics* 31.8 (2004), pp. 2263–2273. DOI: 10.1118/1.1769631.

[52] Pablo Yepes et al. “Monte Carlo fast dose calculator for proton radiotherapy: application to a voxelized geometry representing a patient with prostate cancer”. In: *Physics in Medicine and Biology* 54.1 (2008). DOI: 10.1088/0031-9155/54/1/N03.

[53] Xun Jia et al. “GPU-based fast Monte Carlo dose calculation for proton therapy”. In: *Physics in Medicine and Biology* 57.23 (2012), p. 7783. DOI: 10.1088/0031-9155/57/23/7783.

[54] Kevin Souris, John Aldo Lee, and Edmond Sterpin. “Fast multipurpose Monte Carlo simulation for proton therapy using multi- and many-core CPU architectures”. In: *Medical Physics* 43.4 (2016), pp. 1700–1712. DOI: 10.1118/1.4943377.

[55] Oscar Pastor-Serrano and Zoltán Perkó. “Millisecond speed deep learning based proton dose calculation with Monte Carlo accuracy”. In: *Physics in Medicine and Biology* 67.10 (2022), p. 105006. DOI: 10.1088/1361-6560/ac692e.

[56] Hoyeon Lee et al. “MOQUI: an open-source GPU-based Monte Carlo code for proton dose calculation with efficient data structure”. In: *Physics in Medicine and Biology* 67.17 (2022), p. 174001. DOI: 10.1088/1361-6560/ac8716.

[57] Tiberiu Burlacu, Danny Lathouwers, and Zoltán Perkó. “A Deterministic Adjoint-Based Semi-Analytical Algorithm for Fast Response Change Computations in Proton Therapy”. In: *Journal of Computational and Theoretical Transport* 52.1 (2023), pp. 1–41. DOI: 10.1080/23324309.2023.2166077.

[58] Ahmad Neishabouri et al. “Long short-term memory networks for proton dose calculation in highly heterogeneous tissues”. In: *Medical Physics* 48.4 (2021), pp. 1893–1908. DOI: 10.1002/mp.14658.

[59] Tiberiu Burlacu, Danny Lathouwers, and Zoltán Perkó. “Yet anOther Dose Algorithm (YODA) for independent computations of dose and dose changes due to anatomical changes”. In: *Physics in Medicine and Biology* 69.16 (2024), p. 165003. DOI: 10.1088/1361-6560/ad6373.

[60] J. Perl et al. “TOPAS: An innovative proton Monte Carlo platform for research and clinical applications”. In: *Medical Physics* 39.11 (2012), pp. 6818–6837. DOI: 10.1118/1.4758060.

[61] William.P. Segars et al. “Development of a 4-D digital mouse phantom for molecular imaging research”. In: *Molecular Imaging and Biology* 6.3 (2004), pp. 149–159. DOI: 10.1016/j.mibio.2004.03.002.

[62] M.J. Berger and S.M. Seltzer. *Stopping powers and ranges of electrons and positrons*. Tech. rep. International Commission on Radiation Units and Measurements, 1982. URL: https://nvlpubs.nist.gov/nistpubs/Legacy/IR/nbsir82-2550A.pdf.

[63] ICRU. *Key Data For Ionizing-Radiation Dosimetry: Measurement Standards And Applications*. Tech. rep. International Commission on Radiation Units and Measurements, 2014. URL: https://www.icru.org/report/icru-report-90-key-data-for-ionizing-radiation-dosimetry-measurement-standards-and-applications/.

[64] Pedro Andreo. “On the clinical spatial resolution achievable with protons and heavier charged particle radiotherapy beams”. In: *Physics in Medicine and Biology* 54.11 (2009). DOI: 10.1088/0031-9155/54/11/N01.

[65] Abigail Besemer, Harald Paganetti, and Bryan Bednarz. “The clinical impact of uncertainties in the mean excitation energy of human tissues during proton therapy”. In: *Physics in Medicine and Biology* 58.4 (2013), p. 887. DOI: 10.1088/0031-9155/58/4/887.

[66] Esther Bär et al. “Optimized I-values for use with the Bragg additivity rule and their impact on proton stopping power and range uncertainty”. In: *Physics in Medicine and Biology* 63.16 (2018), p. 165007. DOI: 10.1088/1361-6560/aad312.

[67] Nobuyuki Kanematsu et al. “Dynamic splitting of Gaussian pencil beams in heterogeneity-correction algorithms for radiotherapy with heavy charged particles”. In: *Physics in Medicine and Biology* 54.7 (2009), p. 2015. DOI: 10.1088/0031-9155/54/7/010.

[68] Hermann Fuchs et al. “A pencil beam algorithm for helium ion beam therapy”. In: *Medical Physics* 39.11 (2012), pp. 6726–6737. DOI: 10.1118/1.4757578.

[69] Stewart Mein et al. “Fast robust dose calculation on GPU for high-precision ^{1}H, ^{4}He, ^{12}C and ^{16}O ion therapy: the FRoG platform”. In: *Scientific Reports* 8.14829 (2018). DOI: 10.1038/s41598-018-33194-4.

[70] Jinhe Yang et al. “An improved beam splitting method for intensity modulated proton therapy”. In: *Physics in Medicine and Biology* 65.18 (2020), p. 185015. DOI: 10.1088/1361-6560/ab9b55.

[71] S. Wuyckens et al. *OpenTPS – Open-source treatment planning system for research in proton therapy*. 2023. arXiv: 2303.00365 [physics.med-ph]. URL: https://arxiv.org/abs/2303.00365.

[72] Chris D Johnstone et al. “Multi-institutional MicroCT image comparison of image-guided small animal irradiators”. In: *Physics in Medicine and Biology* 62.14 (2017), p. 5760. DOI: 10.1088/1361-6560/aa76b4.

[73] Ana Vaniqui et al. “The effect of different image reconstruction techniques on pre-clinical quantitative imaging and dual-energy CT”. In: *British Journal of Radiology* 92.1095 (2018), p. 20180447. DOI: 10.1259/bjr.20180447.

[74] Robert H W van Dijk et al. “A novel multichannel deep learning model for fast denoising of Monte Carlo dose calculations: preclinical applications”. In: *Physics in Medicine and Biology* 67.16 (2022), p. 164001. DOI: 10.1088/1361-6560/ac8390.

[75] Marie Vanstalle et al. “Analytical dose modeling for preclinical proton irradiation of millimetric targets”. In: *Medical Physics* 45.1 (2018), pp. 470–478. DOI: 10.1002/mp.12696.

[76] ICRU. *Prescribing, Recording and Reporting Photon Beam Therapy (Supplement to ICRU 50)*. Tech. rep. International Commission on Radiation Units and Measurements, 1999. URL: https://www.icru.org/report/prescribing-recording-and-reporting-photon-beam-therapy-report-62/.

[77] Frank Verhaegen, Patrick Granton, and Erik Tryggestad. “Small animal radiotherapy research platforms”. In: *Physics in Medicine and Biology* 56.12 (2011), R55. DOI: 10.1088/0031-9155/56/12/R01.

[78] Martin Janson et al. “Treatment planning of scanned proton beams in RayStation”. In: *Medical Dosimetry* 49.1 (2024), pp. 2–12. DOI: 10.1016/j.meddos.2023.10.009.

[79] Uwe Schneider, Eros Pedroni, and Antony Lomax. "The calibration of CT Hounsfield units for radiotherapy treatment planning". In: *Physics in Medicine and Biology* 41.1 (1996), p. 111. DOI: 10.1088/0031-9155/41/1/009.

[80] Wilfried Schneider, Thomas Bortfeld, and Wolfgang Schlegel. "Correlation between CT numbers and tissue parameters needed for Monte Carlo simulations of clinical dose distributions". In: *Physics in Medicine and Biology* 45.2 (2000), p. 459. DOI: 10.1088/0031-9155/45/2/314.

[81] Ming Yang et al. "Comprehensive analysis of proton range uncertainties related to patient stopping-power-ratio estimation using the stoichiometric calibration". In: *Physics in Medicine and Biology* 57.13 (2012), p. 4095. DOI: 10.1088/0031-9155/57/13/4095.

[82] Vicki T. Taasti et al. "Inter-centre variability of CT-based stopping-power prediction in particle therapy: Survey-based evaluation". In: *Physics and Imaging in Radiation Oncology* 6 (2018), pp. 25–30. DOI: 10.1016/j.phro.2018.04.006.

[83] Nils Peters et al. "Experimental assessment of inter-centre variation in stopping-power and range prediction in particle therapy". In: *Radiotherapy and Oncology* 163 (2021), pp. 7–13. DOI: 10.1016/j.radonc.2021.07.019.

[84] K B Niepel et al. "Animal tissue-based quantitative comparison of dual-energy CT to SPR conversion methods using high-resolution gel dosimetry". In: *Physics in Medicine and Biology* 66.7 (2021), p. 075009. DOI: 10.1088/1361-6560/abbd14.

[85] Esther Bär et al. "Experimental validation of two dual-energy CT methods for proton therapy using heterogeneous tissue samples". In: *Medical Physics* 45.1 (2018), pp. 48–59. DOI: 10.1002/mp.12666.

[86] Christian Möhler et al. "Experimental verification of stopping-power prediction from single- and dual-energy computed tomography in biological tissues". In: *Physics in Medicine and Biology* 63.2 (2018), p. 025001. DOI: 10.1088/1361-6560/aaa1c9.

[87] Isabel P Almeida et al. "Monte Carlo proton dose calculations using a radiotherapy specific dual-energy CT scanner for tissue segmentation and range assessment". In: *Physics in Medicine and Biology* 63.11 (2018), p. 115008. DOI: 10.1088/1361-6560/aabb60.

[88] Lotte E J R Schyns et al. "Optimizing dual energy cone beam CT protocols for preclinical imaging and radiation research". In: *British Journal of Radiology* 90.1069 (2016), p. 20160480. DOI: 10.1259/bjr.20160480.

[89] Ana Vaniqui et al. “The impact of dual energy CT imaging on dose calculations for pre-clinical studies”. In: *Radiation Oncology* 12.181 (2017). DOI: 10.1186/s13014-017-0922-9.

[90] Jericho O’Connell and Magdalena Bazalova-Carter. “fastCAT: Fast cone beam CT (CBCT) simulation”. In: *Medical Physics* 48.8 (2021), pp. 4448–4458. DOI: 10.1002/mp.15007.

[91] L.J. Nevay et al. “BDSIM: An accelerator tracking code with particle–matter interactions”. In: *Computer Physics Communications* 252 (2020), p. 107200. DOI: 10.1016/j.cpc.2020.107200.

[92] Hans-Peter Wieser et al. “Development of the open-source dose calculation and optimization toolkit matRad”. In: *Medical Physics* 44.6 (2017), pp. 2556–2568. DOI: 10.1002/mp.12251.

[93] Guillaume Landry et al. “Deriving effective atomic numbers from DECT based on a parameterization of the ratio of high and low linear attenuation coefficients”. In: *Physics in Medicine and Biology* 58.19 (2013), p. 6851. DOI: 10.1088/0031-9155/58/19/6851.

[94] Masatoshi Saito and Shota Sagara. “A simple formulation for deriving effective atomic numbers via electron density calibration from dual-energy CT data in the human body”. In: *Medical Physics* 44.6 (2017), pp. 2293–2303. DOI: 10.1002/mp.12176.