dortmund university

Microbeams - quick and dirty

A milestone towards treatment plan optimization for microbeam radiation therapy: development of a fast and portable machine learning-based dose prediction method

Florian Mentzel

01.06.2023
Dortmund

A document submitted in partial fulfillment of the requirements for the degree of

Doctor rerum naturalium (Dr. rer. nat.)

at Department of Physics, TU Dortmund University

Dieses Dokument ist eine Dissertation in der Fakultät Physik an der Technischen Universität Dortmund.
Termin und Ort der mündlichen Prüfung: 01.06.2023, Dortmund

1. Gutachter: Prof. Dr. Kevin Kröninger
2. Gutachter: Prof. Dr. Armin Lühr
Vertretung der wissenschaftlichen Mitarbeitenden: Dr. Doris Reiter
Vorsitz der Prüfungskommission: Prof. Dr. Heinz Hövel

Abstract

Microbeam radiation therapy (MRT) is a promising yet preclinical radiotherapy treatment for several tumour diagnosis such as gliosarcoma and radioresistant melanoma for which even modern clinical treatments such as intensity-modulated radiation therapy (IMRT) and volumetric modulated arc therapy (VMAT) yield poor outcome perspectives. The dose prediction during MRT treatment planning, as for most other novel radiotherapies, is mostly performed with very time-consuming Monte Carlo (MC) simulations. This slows down preclinical research processes and renders treatment plan optimization infeasible.
In this thesis, several milestones for the introduction of a fast machine learning (ML) dose calculation method for MRT are presented. First, a 3D U-Net-based ML dose engine is developed using MC training data obtained with Geant4 simulations of a synchrotron broadbeam incident on different bone slab models and a simplified human head phantom as a proof of concept. The developed model is shown to produce dose predictions within less than 100 ms which is substantially faster than the used MC simulations with up to 20 hours and also the currently fastest approximative MRT dose prediction approach, called *HybridDC*, with approximately 30 minutes. The model is also shown to be superior to a dose prediction approach using generative adversarial networks (GANs) and also a novel transformer-based ML model called Dose Transformer (DoTA), with which it is compared for application in proton minibeam radiation therapy (pMBRT) in a subsequent study. Secondly, the developed ML model and the MC simulations for data generation are extended to account for the spatially fractionated nature of MRT. For this, a novel MC scoring method is developed which is able produce separate dose estimations for the high-dose *peak* regions where the microbeams traverse the phantoms and the low-dose *valley* regions in-between those beams. Finally, the developed ML model and the MC scoring method are deployed in a first application of an ML dose prediction method in a preclinical MRT study in collaboration with the University of Wollongong, Australia, conducted at the Imaging and Medical Beamline (IMBL) at the Australian Synchrotron which aimed at treating rats after implanting gliosarcoma cells. It is shown that the ML model can be trained to provide unbiased dose estimations in complex target phantoms even when trained on high-noise MC data, in important finding for the acceleration of future developments of ML models as such datasets can be produced significantly faster. The ML predictions in the rat phantoms deviate at most 10% from the MC simulations, rendering the proposed model a suitable candidate for fast dose predictions during treatment plan optimization in the future.

Kurzfassung

Microbeam radiation therapy (MRT) ist eine vielversprechende vorklinische Strahlentherapie für einige Tumordiagnosen, wie beispielsweise Gliosarcome und radioresistente Melanome, für die auch moderne Therapiemethoden wie intensity-modulated radiation therapy (IMRT) und volumetric modulated arc therapy (VMAT) schlechte Therapieaussichten haben. Die Dosisvorhersage während der Behandlungsplanung für MRT, ebenso wie für viele andere neue Strahlentherapien, wird meistens mit sehr zeitaufwändigen Monte Carlo (MC) Simulationen durchgeführt. Dies zieht die Forschungsschritte in vorklinischen Studien in die Länge und verhindert vor allem die Optimierung von Behandlungsplänen. In dieser Arbeit werden mehrere Meilensteine für die Einführung einer schnellen MRT-Dosisberechnungsmethode auf der Basis von ML präsentiert. Zuerst wird ein machine learning (ML)-Dosisberechnungsmodell auf der Grundlage eines 3D U-Nets entwickelt. Dazu werden zunächst MC Trainingsdaten mithilfe von Geant4 Simulationen erzeugt, die die Dosisverteilung in verschiedenen Knochenscheibenphantomen und einem vereinfachten Kopfphantom nach Bestrahlung mit einem sogenannten Synchrotron broadbeam vorhersagen. Das entwickelte Modell erzeugt Dosisvorhersagen innerhalb von weniger als 100 ms, was signifikant schneller als die Laufzeit der verwendeten MC Simulationen (bis zu 20 Stunden) und ebenfalls die zur Zeit schnellsten MRT Dosisberechnungsmethode mithilfe von Approximationen, der sogenannten HybridDC Methode (ca. 30 Minuten). Darüber hinaus wird gezeigt, dass das vorgestellte Modell sowohl bessere Vorhersageergebnisse als ein alternativer ML-Ansatz auf Basis von generative adversarial networks (GANs), als auch ein neues Transformer-basiertes ML-Modell namens Dose Transformer (DoTA) erreicht. Der Vergleich mit dem DoTA-Modell erfolgt in einer Studie zur Dosisvorhersage einer anderen neuen Strahlentherapiemethode, der proton minibeam radiation therapy (pMBRT). Anschließend wird das entwickelte ML-Modell und die MC Simulationen weiterentwickelt, um der räumlich fraktionierten Natur von MRT gerechnet zu werden. Dazu wird eine neue MC Scoringmethode entwickelt, welche separate Dosisverteilungen für den Peakbereich, in dem die Microbeams die Phantome durchqueren und eine hohe Dosis deponieren, und für den Valleybereich mit deutlich geringeren Dosisdepositionen dazwischen erstellt. Abschließend werden das entwickelte ML-Modell und die neue MC Scoringmethode in einer ersten Anwendung von ML-Dosisvorhersagemethoden in einer vorklinischen MRT-Studie einer Forschungsgruppe der University of Wollongong angewendet, in der mit Gliosarcomen implantierte Ratten an der Imaging and Medical Beamline (IMBL) am Australian Synchrotron bestrahlt wurden. Es wird gezeigt, dass das ML-Modell nach dem Training Dosisvorhersagen ohne Bias erzeugen kann, obwohl es mithilfe von MC Simulationen mit einer hohen statistischen Unsicherheit trainiert wird. Dies ist eine wichtige Erkenntnis für die beschleunigte Entwicklung zukünftiger ML-Modelle, da solche Daten deutlich schneller erzeugt werden können. Die produzierten Dosisvorhersagen weichen zumeist höchstens 10% von den MC Simulationen ab, daher wird das entwickelte Modell als geeigneter Kandidat für zukünftige schnelle Dosisvorhersagen für die Planungsoptimierung von MRT-Bestrahlungen eingeordnet

Publications, conference contributions and thesis supervisions

During working on my PhD project, I contributed to ten publications, six of those as main author. I held six presentations at international conferences in addition to several presentations at several smaller, national workshops and conferences. In total, I was directly involved in the scientific supervision of 18 B. Sc. and M. Sc. thesis and involved to a lower degree, especially in the design of the research projects, of additionally 20 research projects resulting in B. Sc. and M. Sc. theses.

List of Publications

F. Mentzel et al., *Accurate and fast deep learning dose prediction for a preclinical microbeam radiation therapy study using low-statistics Monte Carlo simulations.* Submitted for publication to Cancers.

E. Derugin, ..., F. Mentzel et al., *Deep TL: Progress of a machine learning-aided personal dose monitoring.* Accepted for publication in J. Radiol. Prot.

F. Mentzel et al., *Small beams, fast predictions - A comparison of machine learning dose prediction models for proton minibeam therapy.* Medical Physics 49(12), 7791-7801

M. Rempe, F. Mentzel, et al., *k-strip: A novel segmentation algorithm in k-space for the application of skull stripping.* Submitted for publication to Magnetic Resonance in Medicine

F. Mentzel et al., *A step towards treatment planning for microbeam radiation therapy: fast peak and valley dose predictions with 3D U-Nets.* Accepted for publication in IFMBE Proceedings on the World Congress on Medical Physics and Biomedical Engineering 2022

F. Mentzel et al., *Fast and accurate dose predictions for novel radiotherapy treatments in heterogeneous phantoms using conditional 3D U-Net generative adversarial networks.* Medical Physics 49 (5), 3389–3404.

S. Zwiehoff, ..., F. Mentzel et al., *Enhancement of Proton Therapy Efficiency by Noble Metal Nanoparticles Is Driven by the Number and Chemical Activity of Surface Atoms.* Small Structures (2021)

F. Mentzel et al., *No more glowing in the dark: How deep learning improves exposure date estimation in thermoluminescence dosimetry.* J. Radiol. Prot. 41 S506 (2021)

F. Mentzel et al., *Extending information relevant for personal dose monitoring obtained from glow curves of thermoluminescence dosimeters using artificial neural networks*, Radiat. Meas. 136 (2020) 106375

M. Hötting, ..., F. Mentzel et al., *Study of radiation-induced frequency shifts in quarz crystal oscillators.* 2020 IEEE International Conference on Wireless for Space and Extreme Environments (WiSEE), Vicenza, Italy, 2020, pp. 24-28

List of conference contibutions

F. Mentzel et al.,*Accurate and fast deep learning dose prediction for a preclinical microbeam radiation therapy study using low-statistics Monte Carlo simulations.* MMND-ITRO 2022, Noosa, Australia. Presented by Prof. Susanna Guatelli.

F. Mentzel et al., *A step towards treatment planning for microbeam radiation therapy: fast peak and valley dose predictions with 3D U-Nets.* IUPESM 2022, Singapore

F. Mentzel et al., *Dose prediction for radiotherapy using conditional 3D-Unet generative adversarial networks.* MMND-ITRO 2022, Bowral, Australia (hybrid, digital participation)

F. Mentzel, *Towards fast dose calculations for novel radiotherapy treatments with generative adversarial networks.* AUM2021: ANSTO User Meeting 2021, (digital)

F. Mentzel, *Machine learning for personal dose monitoring - Insights into irradiation scenarios with a novel TL dosimeter.* IRPA15: 15th International Congress of the International Radiation Protection Association, 18.01.-5.02.2021, Seoul, South Korea (hybrid, digital participation)

F. Mentzel, *From HEP to hospital: utilizing ATLAS detector and analysis technology for medical physics projects.* MMND-ITRO 2020, Wollongong, Australia

F. Mentzel, *Extending information relevant for personal dose monitoring gained from glow curves of passive solid state dosemeters using artificial neural networks.* 19th International Conference on Solid State Dosimetry, 2019, Hiroshima, Japan

Contents

List of Abbreviations

BDA beam-defining aperture
CPU central processing unit
CT Computer Tomography
DoTA Dose Transformer
ESRF European Synchrotron Radiation Facility
FWHM full width at half maximum
GAN generative adversarial networks
GPU graphics processing unit
HU Hounsfield units
IMBL Imaging and Medical Beamline
IMRT intensity-modulated radiation therapy
MAE mean-absolute error
MC Monte Carlo
ML machine learning
MRT microbeam radiation therapy
MSC multi-slit collimator
MSE mean-squared error
pMBRT proton minibeam radiation therapy
PSF phase space file
PVDR peak-valley dose ratio
ReLU rectified linear unit
VMAT volumetric modulated arc therapy
WGAN Wasserstein GAN

List of Figures

List of Tables

1 Introduction

Killing cancer cells is easy.
It can be difficult, however,
to keep the patient around them intact.

F. Mentzel, 2022

Since the first investigations of the potential of x-rays for tumour treatment [1] soon after their discovery, many innovations and developments in medical imaging and radiotherapy treatment methods have improved the clinical outcome of a wide variety of cancer diagnoses. Especially the introduction of highly conformal treatment techniques, which allow precise dose application to tumour targets while sparing healthy tissue, such as intensity-modulated radiation therapy (IMRT) [2], volumetric modulated arc therapy (VMAT) [3], and more recently proton therapy [4], significantly improved the long-term survival rates for many diagnoses [5, 6].

With increasing conformality and accuracy of available treatments, the fast and accurate estimation of the radiation dose delivered to the tumour target and the surrounding healthy tissue has equally gained in importance for a successful therapy. For this, dedicated fast dose estimation methods, often based on approximative analytical computations (e.g. [7, 8]), have been developed. Those methods, however, are highly specialized and often proprietary software. Therefore, they are not suitable for research on novel and pre-clinical treatments which try to tackle shortcomings of existent treatments. Instead, dose estimation for those treatments is commonly performed using time-intensive Monte Carlo (MC) particle tracking simulations (e.g. [9, 10]) with software tools such as *Geant4* [11], a multi-purpose MC toolkit developed at CERN. While MC simulations allow for excellent agreement with experimental dosimetry even for complex irradiation scenarios [12], the required computation time delays research and hinders potential optimizations of treatment plans. This holds true especially for *spatially fractionated therapies* such as microbeam radiation therapy (MRT) [13], which is a proposed treatment for e.g. gliosarcoma [10], for which the clinical outcome has barely improved over the last decades [14].

MRT utilizes arrays of coplanar sub-mm synchrotron radiation beamlets depositing high *peak doses* along the path of those narrow beams and relatively low *valley doses* in-between them. For current pre-clinical MRT studies at the Imaging and Medical Beamline (IMBL) [15] at the Australian Synchrotron, dose estimation are performed with MC simulations based on Geant4 [10]. The high spatial resolution required close to the resulting steep dose gradients lead to MC simulation times of up to several hours for accurate dose predictions [16]. While at the European Synchrotron Radiation Facility (ESRF) [17], another facility researching MRT, dose estimations are also performed with MC simulations [18] based on the *PENELOPE*[19] computer code system and MC methods are still under development [20], a faster approximate dose estimation method, the *HybridDC* model, has been developed [21, 22, 23]. It is currently being extended to also be available at the IMBL [24].

While the HybridDC model is an important development for faster MRT planning, the execution time of approximately half an hour per irradiation field is too long to allow for sophisticated plan optimization techniques [25] or the application of adaptive treatment techniques [26]. This incentivises the development of even faster methods for MRT dose prediction. While there are, in fact, already very fast MRT dose estimation methods available in the literature [21, 27], those were found to not suffice in accuracy especially at tissue interfaces to be used in plan optimization.

Throughout this thesis, an MRT dose estimation model is developed, which is both very

fast and at the same time sufficiently accurate to accelerate treatment plan generation and also potentially make optimizations of those viable. A class of algorithms capable of approximating a wide range of complex data [28] with very short execution times stem from the field of machine learning (ML). While ML-based dose estimation methods are published with increasing frequency for clinically available treatments such as IMRT and VMAT in recent years [29, 30, 31], those studies are usually facilitated by large databases of past treatments or the possibility to quickly generate new datasets using specialized fast algorithms. Moreover, this is also not possible for other new and pre-clinical treatments other than MRT, resulting in the requirement to create the datasets from scratch using the time-consuming MC simulations. In the course of this thesis, existent approaches to mitigate this obstacle are discussed and new ones are developed, which make the development of fast ML-based dose estimation methods more accessible to research in new and pre-clinical radiotherapy treatments. Parts of the presented work have already been published [16, 32, 33, 34].
The thesis is structured as follows: In a first step, the underlying concepts of radiotherapy, MRT, dose calculation methods and ML, are presented in Chapter (2). Subsequently, the development of the foundations of the proposed model in a proof-of-concept study predicting a synchrotron broadbeam instead of microbeams is presented in Chapter (3). The transferability of the obtained ML model by retraining it on a dose deposition dataset for proton minibeam radiation therapy (pMBRT) and comparing the results to other state-of-the-art ML models in Chapter (4). In the subsequent Chapter (5), the spatial fractionation of MRT doses into peak and valley doses is included into the ML model. To obtain the required peak and valley dose data sets, a specialized MRT MC dose scoring method is introduced. The applicability of the developed MRT-ML model in pre-clinical research is demonstrated in a retrospective treatment planning for a rodent radiation study in Chapter (6). Finally, an outlook on potential future paths of research is given in Chapter (7), before summarizing and concluding the thesis in Chapter (8).

2 Getting up to speed: microbeam radiation therapy, dose prediction and machine learning

This chapter introduces some theoretical concepts which are required for the developments in this thesis. First, the generation of microbeams and the physical background of their interaction with tissue is discussed in Section (2.1). Afterwards, methods of Monte Carlo (MC) particle tracking and machine learning are explained with focus on microbeam radiation therapy (MRT) dose estimation in Section (2.2) and Section (2.3), respectively.

2.1 Microbeam radiation therapy (MRT)

Already at the beginning of the 20th century it was found that adverse effects of x-ray therapies, especially skin lesions, can be reduced by treating a target with a grid of smaller beams instead of a homogeneous field [35]. The enhanced survival of healthy tissue when being exposed to narrow radiation beams in contrast to wider fields was later described as *dose-volume effect* [36] and subsequently led to the development of x-ray microbeam therapy [37, 13]. While the observation of enhanced healthy cell survival following spatially fractionated irradiation is well documented (e.g. [38, 39, 40]), the biological cause of the success of MRT remains disputed; Common explanation approaches involve blood vessel disruption together with associated repair processes [41, 42], the infiltration of tumour tissue by immune cells [43], or effects of tumour reoxygenation [44], but also on the so-called *FLASH effect* [45], which describes enhanced radioresistance of healthy tissue when being exposed to very high dose rates (e.g. [46]).

2.1.1 Generation of synchrotron microbeams

To maintain the narrow microbeam blades while traversing a patient, it is important to use an x-ray source with a very low divergence. At the same time, a high photon flux helps minimizing the positioning uncertainty due to potential movements of a patient and contributes to the possible contribution of the aforementioned FLASH effect. Resulting from those requirements, synchrotron beam lines such as the Imaging and Medical Beamline (IMBL) are suitable for MRT research, although compact microbeam sources are subject to ongoing research [47].

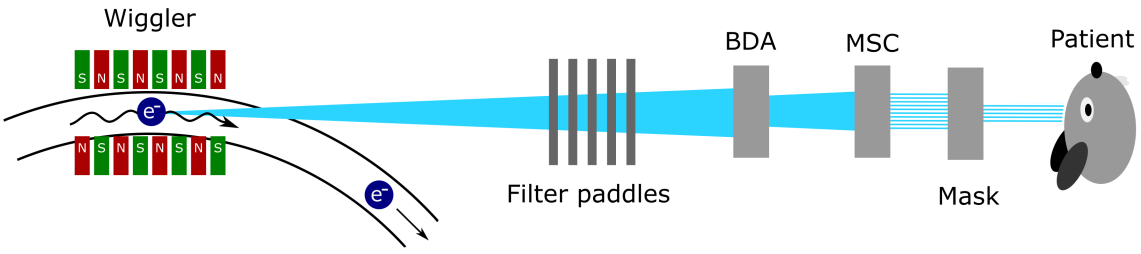


Figure 2.1: Schematic of relevant parts involved in the generation of microbeams at the IMBL (not to scale). The blue electron passing the wiggler emits synchrotron radiation shown in bright blue, passing first the filter paddles, then the beam-defining aperture (BDA) and subsequently the multi-slit collimator (MSC), before reaching a patient. Inspired by [12].

Figure (2.1) schematically shows the generation of microbeams at the IMBL. The electrons within the synchrotron ring have a kinetic energy of 3 GeV [15]. The alternating magnetic

field causes the electrons to emit synchrotron x-ray radiation (shown in blue) in their direction of flight [48]. The emitted radiation can be passed through up to five filter paddles which are described in detail in [15], adjusting the photon flux and energy spectrum depending on the materials placed in the beam.

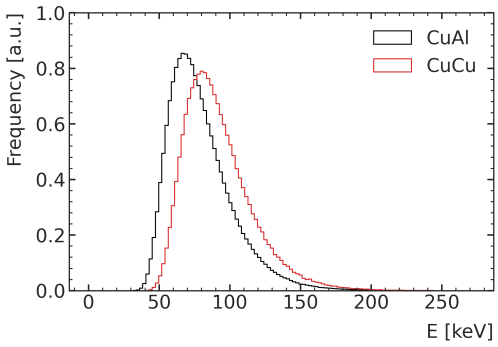


Figure 2.2: Synchrotron beam energy spectrum after filtering with two copper blades (red) and one copper blade together with an aluminium blade (black). Produced using simulation data from [12].

The energy spectra used in the studies presented in this thesis - *CuAl* and *CuCu*, named after the main constituents of the paddles being copper and aluminium - are obtained from a previously published simulation [12] and are shown in Figure (2.2). It can be seen that using the CuCu paddle set, the mean energy is higher ($\overline{E}_{\mathrm{CuCu}} = 91.9\,\mathrm{keV}$) than with the CuAl paddle set ($\overline{E}_{\mathrm{CuCu}} = 79.5\,\mathrm{keV}$). In addition, the CuCu spectrum is slightly wider when compared by full width at half maximum (FWHM): $\mathrm{FWHM}_{\mathrm{CuCu}} = 46.1\,\mathrm{keV}$ vs. $\mathrm{FWHM}_{\mathrm{CuAl}} = 41.8\,\mathrm{keV}$.
The beam is subsequently shaped using a BDA made from tungsten. For this thesis, a BDA with a nominal height of 1.053 mm and a nominal width of 30.0 mm is used. The resulting beam is referred to as *broadbeam* in the further course of this thesis. From the broadbeam, an array of microbeams is obtained using a MSC which is also made from tungsten. The individual slits have a nominal width of 50 μm and are spaced 400 μm apart.

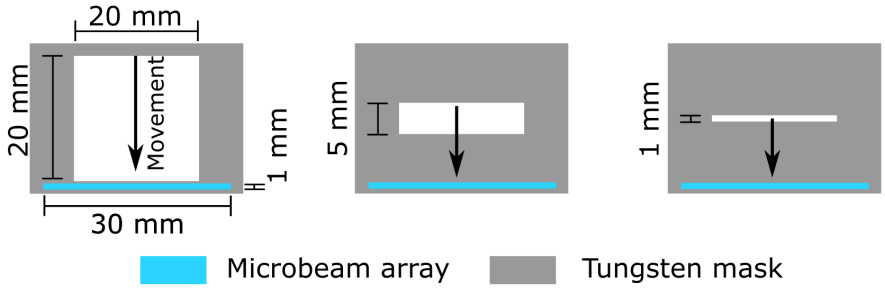


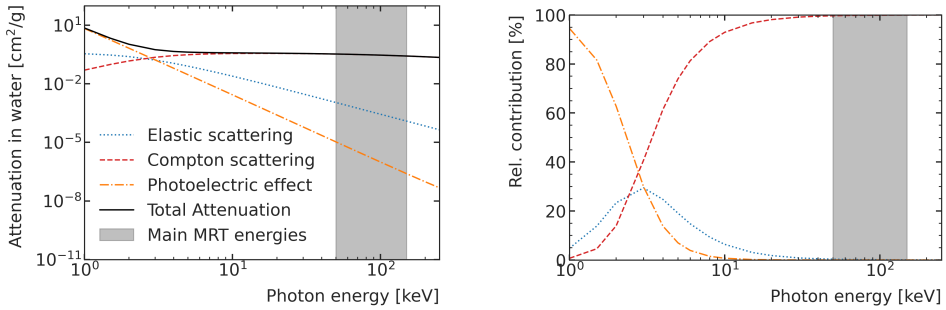
Figure 2.3: Schematic showing how the microbeam field height is adjusted using masks with cutouts of different sizes.

The delivery of fields of different heights is schematically shown in Figure (2.3). The microbeam array is shaped using a mask which is placed in front of the treatment target. With the beam from the synchrotron having a fixed position and direction, the treatment target is moved together with the mask in front of the beamline. Thereby, the synchrotron beam

is *painted* over a larger area than its initial size of approximately 1 mm beam height. This method is exemplarily shown for three rectangular fields of different height (20 mm, 5 mm and 1 mm), all being 20 mm wide.

2.1.2 Interaction of microbeams with tissue

The most important effects resulting in energy transfer from the microbeams to the target material are the same effects [49] as for conventional x-ray therapy: the photoelectric effect, the Compton effect and pair production. The importance of each effect is dependent both on the material being traversed by the x-rays and the energy of the beam. A comparison of their importance in water is shown in Figure (2.4). For better visibility of the energies important for MRT, the range from 50 keV to $50 - 150$ keV is highlighted.

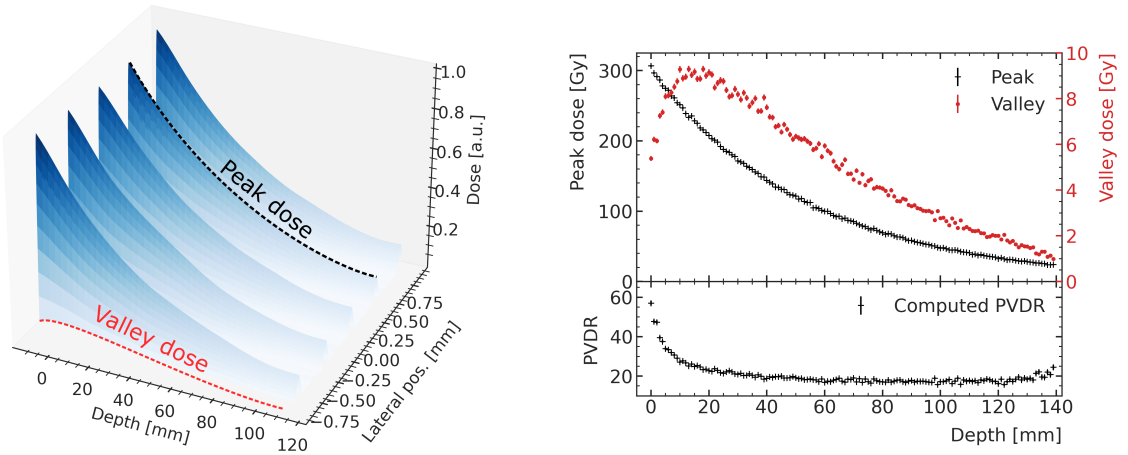


(a) (b)

Figure 2.4: (a): Composition of the total attenuation of x-ray radiation in water in dependence of the energy according to [49]. (b): Relative contributions of different relevant effects. The most relevant energies for MRT are marked grey in both plots.

For the relevant energies, the Compton effect is the most important interaction for the initial photons entering a phantom, as shown in Figure (2.4b). For scattered photons, the elastic scattering and the photoelectric become more dominant in the range of a few keV of remaining x-ray energy. The energies in MRT are not high enough for pair production which requires a photon energy of at least 1.024 MeV. All the aforementioned effects cause the generation of secondary electrons which subsequently deposit energy in the surrounding tissue. The primary dose deposition in the path of the microbeams is referred to as *peak dose*. As the electrons do not only move in the direction of the photons, they are scattered into the areas between the microbeams resulting in the deposition of a *valley dose*.
A visualization of the dose profile following the irradiation of a homogeneous water phantom with microbeams is shown in Figure (2.5a). While the near-exponential decay of the peak dose with the depth of the phantom is visible, the development of the valley doses occurs at a significantly lower absolute value. A comparison of the development of the depth profiles is possible in the one-dimensional plot showing depth-dose curves for the central microbeam and one of its adjacent valley (towards positive lateral position) shown in Figure (2.5b). While the peak dose decreases from the entrance of the phantom towards the distal end, the valley dose peaks in approximately 20 mm depth before subsequently decreasing. The initial increase in valley dose results from the build-up and scattering of secondary electrons. At the end of the phantom, a steeper decrease of the valley dose compared to the development of the slope before occurs, which is due to missing backscatter. The different axes for peak dose and valley dose should be noted, as the peak dose as a maximum value of around 300 Gy while

the valley dose only peaks at approximately 9 Gy. The peak-valley dose ratio (PVDR), shown in the lower subplot, is a commonly used measure to describe the relation of the peak and valley dose at a certain position within a target phantom (e.g. [50]). Because the valley dose results from photon and electron scattering, the PVDR depends strongly on the microbeam energy and the size and material of the target phantom.



(a) (b)

Figure 2.5: (a): 2D dose profile for a 20x20 mm^2 MRT field in a water phantom with exemplary peak (black) and valley dose profile (red). (b): Peak and valley depth dose curves with the resulting PVDR for the central peak and one adjacent valley (towards positive lat. position).

2.2 Current dose estimation methods for microbeam radiation therapy

As for all radiotherapies, the dose deposition following an irradiation has to be estimated accurately to be able to deliver as much dose as prescribed to the target volume while not overdosing the healthy tissue around it. In the following, central aspects of the MRT dose estimations with MC simulations at the IMBL and the HybridDC model are presented.

2.2.1 Monte Carlo simulations with phase space files

The advantage of multi-purpose MC simulations with e.g. Geant4 is that all included physical interaction processes are modelled explicitly using probability distributions for each interaction of a particle traversing materials in the simulation world and their results [11]. Implemented processes differ in computation speed and experimentally validated accuracy. They are bundled in so-called *physics lists* allowing a user to trade-off between the computation time and modelling accuracy as required by individual simulations. The implemented physical processes are continuously extended and validated by groups from a variety of research fields (e.g. [51]). Resulting from this, MC codes are often referred to as the *gold standard* in dosimetry simulations (e.g. [52, 53]), especially when utilizing the most accurate physics lists.
An experimentally well-validated Geant4 simulation, closely described in [12], has been developed in the last years to facilitate accurate MRT dose predictions [12]. The electromagnetic interactions of the microbeams with matter are modelled using a custom physics list comprising the most accurate set of electromagnetic interactions implemented (*EM Option 4*,

details on involved processes: [54]). To account for the high degree of polarization of synchrotron radiation [55], the *Livermore Polarized Physics* are included in the simulation [56]. The MRT simulation covers the generation of the synchrotron radiation from the accelerated electrons over the different beam shape and energy spectrum adjustments described in subsection (2.1.1). After the MSC, the simulated particles can be stored instead of being tracked further downstream towards a patient phantom. Each stored particle is characterized by its type, their energy, position, momentum and polarization in a so-called phase space file (PSF). Such PSFs have been created for all commonly used beamline setting with respect to the wiggler magnet field strength, inserted filter paddles, etc., and are used in subsequent simulations of patient treatments. This approach allows a significant acceleration of the whole simulation procedure as the microbeam generation process can be recycled and does not have to be re-executed. Based on those PSFs, all simulations within this thesis are performed with Geant4 version 10.6p01., using the aforementioned physics list. In addition to the used processes, so-called *production cuts* and *step limits* are implemented in the physics list. Their meaning and used settings are explained in the following:
(1) Production cuts: Geant4 tracks all particles, once created, until they absorbed or have lost all kinetic energy. Not all particles are created, however, in the first place. Only particles travelling a minimal distance within the material of their creation, or having a minimal energy, are generated and subsequently tracked in the simulation. This is a design feature to accelerate simulations by making it possible to neglect computationally very expensive low-energy interactions which potentially do not contribute to the accuracy of the result, depending on the required spatial resolution. Following the very high spatial resolution required for MRT simulations, the production cuts are set to 1 µm or 250 eV, respectively.
(2) Step limits: By default, particles are stopped after travelling the distance which Geant4 has sampled for the next occurring physical interaction and additionally at material boundaries in the simulation geometry. While most interactions of tracked particles are modelled as discrete processes in Geant4, a part of the energy loss and scattering is simulated as continuous process along the simulated trajectory. This includes especially the energy of suppressed secondary particles below the production cuts. The dose depositions in this thesis are usually scored using a 3D voxel grid. Energy depositions within a voxel is counted towards its total reported dose. To allow for an accurate dose deposition profile in phantom targets of interest, the step limits within phantoms are set to be 1/5th of the spatial resolution of the respective voxel grids, leading to Geant4 stopping the current simulation step after a maximum of the given distance. The step limit is only enforced for charged particles.

2.2.2 HybridDC: photon Monte Carlo and kernel-based electrons

HybridDC [22] is the currently fastest MRT dose estimation method used in pre-clinical research and takes approximately 30 min for one field according to the provided reference above, which is a significant improvement compared to up to several hours using a accurately modelled Geant4 simulation [16].
The HybridDC algorithm separates the dose simulation into a simplified MC simulation and the subsequent application of a *dose kernel* [57] to achieve the resulting microbeam dose distribution. The approach is schematically shown in Figure (2.6) and explained below following the descriptions in [22]: The simplified MC simulation only tracks photons and suppresses the generation of any secondary particles such as electrons. The energy depositions are instantly deposited locally in a 3D voxel grid at each computed interaction point of the photons. The depositions are further separated into two contributions: the first interaction of a photon within the phantom is scored as *primary dose*. All subsequent interactions of that photon are scored as *scatter dose*. Using a map of the initial photon fluences of the microbeams, a

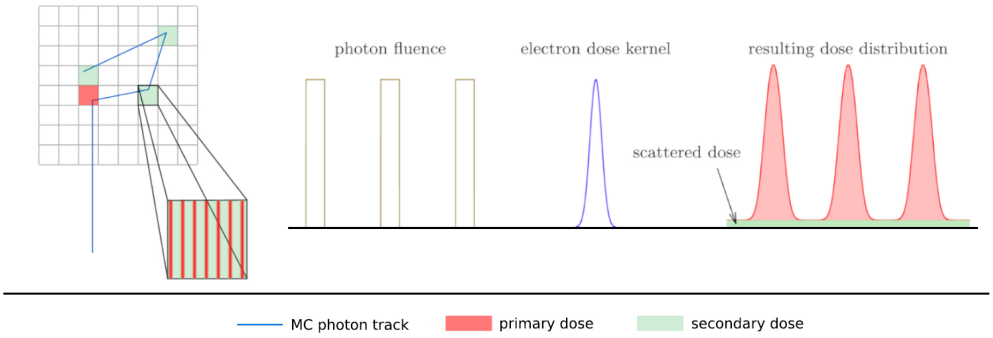


Figure 2.6: Schematic of the HybridDC algorithm showing the simplified MC simulation left and the required subsequent photon fluence and electron dose kernel to achieve the final MRT dose on the right. Adapted from [22].

primary photon profile can be located within the voxels using the primary dose. The final primary dose is then calculated by applying an electron dose kernel specifically developed for MRT. The total dose is obtained by adding the scatter dose to the output of the dose convolution algorithm.

2.3 Accelerating dose estimation with machine learning

Most machine learning (ML)-based dose estimation methods belong to the ML sub-field of so-called *supervised learning*, which comprises scenarios in which a neural network or other algorithms are trained with data samples that contain both an input to the algorithms and the desired output. In the dose estimation case, the input usually is derived from a Computer Tomography (CT) scan of the target phantom, describing the geometry and contained materials. Additionally, the input can be extended to also comprise more information required for the prediction such as the beam shape or its energy spectrum. The output of such an ML model is the dose distribution in the given input geometry. Those dose distributions need to be simulated before the training of the ML model. For MRT, this can be done with either MC simulations or the HybridDC algorithm. The HybridDC model comes with several approximations such as the assumption of perfect coplanarity of the microbeams, i.e., no residual divergence, and the neglection of the impact of the polarization of the synchrotron photons on the resulting dose distributions (see [22]). An ML model trained on HybridDC data therefore would reproduce those assumptions. For this reason, it was decided to train the proposed MRT ML model directly on high-quality MC simulations. By doing so, all physical effects modelled in the MC simulation can be embedded into the prediction of the ML model, while at the same time potentially achieving very fast prediction times.

2.3.1 Blueprint of the proposed model: the U-Net structure

There are several ML algorithms potentially suitable for dose predictions in the literature. In this thesis, the three-dimensional extension of the so-called *U-Net* [58, 59], a convolutional neural network originally developed for 3D segmentation tasks, is chosen as the base for the MRT ML model. Adaptions and variations of the 3D U-Net were shown to perform well in dose estimation tasks in recent years [60, 61, 62]. They are especially suitable for new treatments as they are known for achieving good results also in environments with small data sets (e.g. [63]). Figure (2.7) shows a schematic overview of the 3D U-Net model as originally proposed [59].

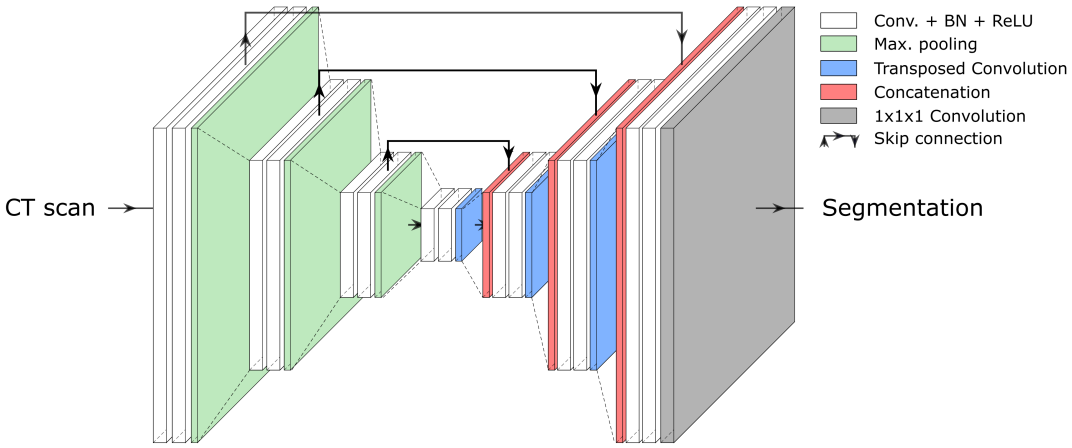


Figure 2.7: Schematic of the 3D U-Net. Modelled after the description in [59].

The 3D U-Net is a type of convolutional network [64] which performs multi-scale feature extraction utilizing a *compression* or *analysis* and a subsequent *decompression* or *synthesis* path, connected by skip connections to allow information from each resolution level to bypass the respective lower levels, facilitating convergence of the model [65]. The respective paths contain blocks of layers, each comprising two convolutional layers with subsequent *batch normalization* [66] and the use of the rectified linear unit (ReLU) [67] function for activation. After each of those blocks, the data is compressed in the downwards path using a *max pooling* operation [68] and decompressed in the upwards path using a *transposed convolution* or *deconvolution* operation [69]. The final segmentation is produced by applying a 1x1x1 convolution matrix to the previous output containing multiple channels to project them onto a single channel per segmentation target, i.e. one for tumour, multiple for organ segmentation.

2.3.2 Training methods for dose estimation neural networks

In this thesis, two different methods to train neural networks for dose predictions are used: training as *regression model* and as *generative adversarial networks*. While the differences are explained in the following sections, they also have very basic common feature with respect to the handling of the data.
To train a neural network and subsequently estimate its performance on future, unseen data (*generalization*), the available data set needs to be divided into three sub-sets: the *training*, *validation*, and *test* data. The training data is used in the process of updating the weights of the network as described in the next sections. The performance on the validation data is used as comparative measure between different neural networks defined by external parameters, the so-called *hyperparameters*. The resulting best network model is subsequently applied to the test data, which is completely left out of the process until this step to allow for an an unbiased performance estimation. To improve network convergence, all values in data samples are normalized to the range between 0 and 1 using the respective maximum and minimum value of the training data set. In the case that validation or test samples exhibit larger or smaller values, this also results in values outside this range.

Training regression models

A neural network can be trained as a regression model by using a *loss function* to compare the output of the network for a given input with the respective associated *target*. Commonly

used loss functions include for example the mean-squared error (MSE) or mean-absolute error (MAE). In the case of radiotherapy dose estimations, the given input can be a phantom CT or a derived density matrix, the output is the dose distribution.
The loss is usually computed on a *batch* of training samples before being used to update the trainable weights of the network using an optimization algorithm such as in this case *Adam* [70]. One iteration through all training samples is referred to as *epoch*.

Training generative adversarial networks

When it comes to ML algorithms which are trained to mimic MC simulations, an alternative training approach has gained high interest: generative adversarial networks (GANs) [71]. Instead of deterministically producing an output for a given input, GANs allow for statistical variations in their predictions. An early notable application of GANs as a fast alternative to MC simulations was the generation of detector simulations samples for the ATLAS detector at CERN [72]. Since then, they have also been applied successfully in the field of radiotherapy research [73, 29, 74].
GANs are built using two neural networks: a *generator* network and a *discriminator* or *critic*, where the naming in the literature depends on the type of GAN. The generator network is passed a vector of random numbers and produces a so-called *fake* data sample. The critic network subsequently receives a set of fake samples produced by the generator and *true* samples from a dataset which should be mimicked by the generator. In the case of Wasserstein GANs (WGANs) [75], which are used in this work, the critic network is trained to tell apart those two groups of samples using a so-called *Wasserstein distance* [76]. This measure is then provided to the generator to adjust its weights resulting in the predictions being more similar to the real data set as judged by the critic network. Upon successful training, this results over time in the generator producing samples which are ideally indistinguishable from the original dataset.
As for the regression training, the networks are updated batch-wise. It is common to train the critic network on several batches before continuing to train the generator network, ensuring a high quality of the critic feedback to the generator. By adding *conditional information* to each data sample, it is possible to train the generator to create samples matching real samples in dependence of this [77]. In the case of the dose estimation, the conditional information is the phantom geometry and beam characteristics such as the energy spectrum, allowing the generator to create a dose estimation for the presented scenario while allowing for statistical variations by means of the random noise. Compared to the more conventional training of neural networks as regression models, this method is more flexible which was found to potentially help learning to predict doses in complex phantom geometries [29].

3 A proof of concept: synchrotron broadbeam dose prediction

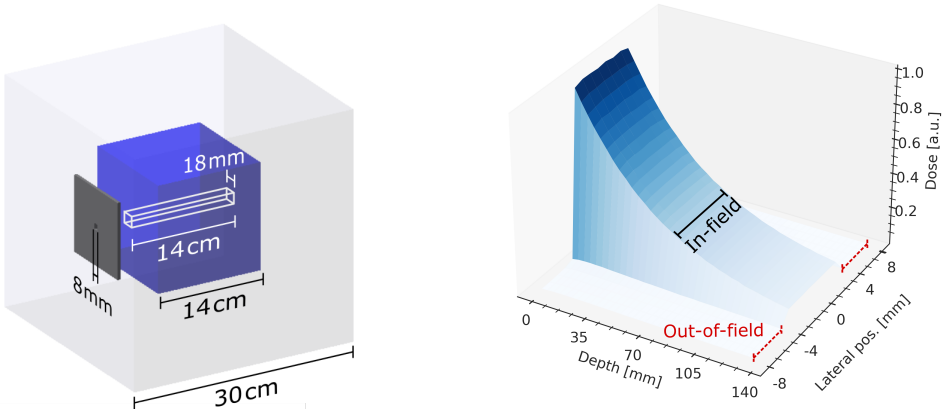
In this section, the development of a first machine learning (ML) model, for dose predictions at the Imaging and Medical Beamline (IMBL) is presented. Due to the potential benefits of better generalization in complex scenarios, the model is designed as generative adversarial networks (GAN) as described in the previous section. High resolution microbeam simulations result in very large files containing a very large number of voxels. This poses a problem for the development of ML models because predicting larger volumes requires a model with more parameters leading to an increase in both training time and required memory to contain the model during the training. For example, in the case of this thesis the used graphics processing units (GPUs) do not exceed 11 GB of memory. In early studies for this thesis, the process of directly predicting high-resolution microbeam simulations was found to be infeasible. Instead, this first proof-of-concept ML model is developed using simulations with a synchrotron broadbeam instead of microbeams. The broadbeam dose distribution is relatively similar to the sum of the primary and secondary dose used by the HybridDC algorithm. This opens the perspective of potentially transferring a developed model to primary and secondary dose prediction, accelerating the first slow Monte Carlo (MC) step involved in the HybridDC algorithm.
Major parts of the results shown in this section have been already published prior to the submission of this thesis [16].

3.1 Model development

In the following, the simulation setup being used during the development of the ML model is presented before describing the adaption of the 3D U-Net for synchrotron broadbeam prediction as part of a GAN.

3.1.1 Data from a digital phantom: a Geant4 bone slab model

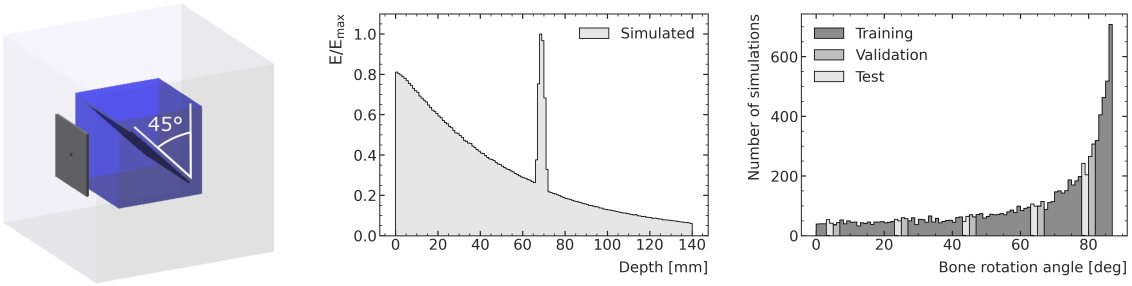
A simple digital bone slab phantom is built to allow the generation of well-defined data samples allowing for systematic investigation of strengths and weaknesses of the proposed ML model. The basic simulation setup before the insertion of a bone slab is shown in Figure (3.1a). The synchrotron broadbeam, based on the phase space file (PSF) for the *AlAl-filtration* ([12], energy spectrum: Figure (2.2), enters the simulation world, shown in light grey, from the left. The field is cropped to 8x8 mm^2 using a tungsten mask (*G4_W*, [78]) before entering a 14x14x14 cm^3 water cube (*G4_WATER*, [78]), placed 4 cm behind the mask. The energy depositions are scored using a grid of 1x1x1 mm^3 over a volume of 18x18x140 mm^3. This size allows scoring the full depth of the phantom and also includes some of the out-of-field region. A resulting visualisation of the resulting dose to water in the central plane is shown in Figure (3.1b), also highlighting the in-field and out-of-field region. In the training of the ML models, the energy deposition is used instead of the dose.
To produce training samples for the ML model, a bone slab (*G4_BONE_COMPACT_ICRU*, [78]) of 2.5 mm thickness is placed into the centre of the water phantom.
The adapted simulation setup containing a bone slab at 45 ° rotation is shown in Figure (3.2a). The energy deposition-depth curve resulting from a simulation with the bone slab inserted is shown in Figure (3.2b). The energy deposition peak caused by the bone insert is clearly visible in the depth profile.
For each sample, the bone slab is rotated with an angle between $\alpha \in [0, 87]^\circ$ relative to the plane perpendicular to the beam. The upper limit is defined by the angle at which the beam would be incident on the top of the slab. As the traversed bone material increases with



(a) (b)

Figure 3.1: (a): Schematic of simulation setup with the water phantom (blue), scoring volume within (white) and tungsten mask (grey) in the simulation world (light grey). Reproduced from [16]. (b): Dose distribution at the central plane of the beam showing the in-field region (black) and the out-of-field region (red).

the rotation angle and is expected to pose a more difficult energy deposition distribution, the sampling of angles is not uniform but instead follows a $1/\cos(\alpha)$ distribution, resulting in an increasing number of samples for larger rotation angles. The distributions of simulation samples split into training, validation and test samples are shown in Figure (3.2c). In contrast to many ML studies which remove samples randomly from the training data for validation and testing, it is performed in a systematic way in this case. Five sets of 2-degree intervals for each validation ($\alpha \in \{[5,7) \cup [25,27) \cup [45,47) \cup [65,67) \cup [80,82)\}^\circ$) and testing ($\alpha \in \{[3,5) \cup [23,25) \cup [43,45) \cup [63,65) \cup [78,80)\}^\circ$) are removed from the training dataset. The systematic allocation of validation and test data allows for a dedicated investigation of the interpolation capability of a developed model in those regions as they are clearly separated from the training data samples in the rotation angle parameter space. During ML training, the density and energy deposition matrices are subject to random flipping around the central plane, leading to a mirrored distribution towards negative bone slab rotations.



(a) (b) (c)

Figure 3.2: (a): Extended simulation setup with included bone slab (dark blue) at 45° (white) rotation. (b): Energy deposition-depth curve at the centre of the beam, normalised to its maximum. (c): Distribution of the simulated training (dark grey), validation (medium grey) and test (light grey) dataset with respect to the bone slab rotation angle. Reproduced from [16].

3.1.2 Design of a 3D U-Net GAN for dose prediction

Due to extremely long training times on the order of days for each model, a manual search for an optimal combination of parameters for the generator and critic network is conducted which only includes a relatively small set of options.

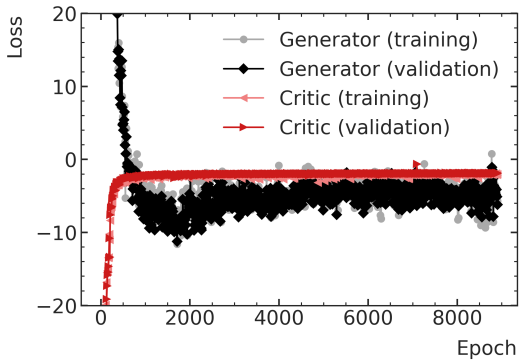


Figure 3.3: Typical loss curves of a dose prediction GAN training showing both the generator (black, grey) and critic loss (red, light red) for training and validation data, respectively.

A common technique to monitor the progress of an ML model during training is the inspection of *loss curves* which show the progression of the loss value over the epochs. In the case of GANs, however, this method is not useful. Figure (3.3) shows an exemplary loss curve of a generator and critic network during GAN training. After an initialisation phase, the losses do not show definitive trends, rather moving around a relatively constant value for each of the networks. This is an expected behaviour of the losses during GAN training: as both networks learn in parallel, the losses are expected to stay relatively stable over the course of the training. While the critic learns to tell apart generated and original samples increasingly well, the generator simultaneously learns to make this task harder for the critic by generating even more similar data samples. While the resulting quality of produced data samples cannot be judged by the loss curves for this reason, they can be used to inspect the stability of the ongoing training: an increase or the observation of oscillations in the loss curves hints at instability of the training process.
In the search for an optimal set of generator and critic comprising the GAN being developed in this section, a different measure is defined: the δ (delta) index, which is modelled after the *global gamma index* [79]. To compute the delta index, first a quantity δ is calculated:

$$\delta = \frac{E_{\text{gen}} - E_{\text{sim}}}{E_{\text{sim}}^{\text{max}}} \,. \tag{3.1}$$

In Eq. (3.1), E_{gen} is the energy deposition predicted by the generator for a given voxel, while E_{sim} is the respective value from the corresponding Geant4 simulation. E_{sim}^{max} is the maximum energy deposition amongst all voxels in the data sample. From this, a δ index *passing rate* can be calculated subsequently. It is defined as the ratio of in-field voxels in a data sample that exhibit a δ value of less than 1$ or 3% for the 1% δ index and the 3% δ index, respectively. To search for an optimal ML model, 1% passing rate on the validation dataset is used as criterion for model selection. The determined best architectures for the generator and critic are described in the following. All ML models are implemented using the Keras[80] interface to Tensorflow 2.2[81].

The generator model

While the generator is modelled after the presented 3D U-Net, there are several changes to the network which were found to achieve better dose predictions in the given scenario. The best overall generator network is shown in Figure 3.4 and will be explained in the following.

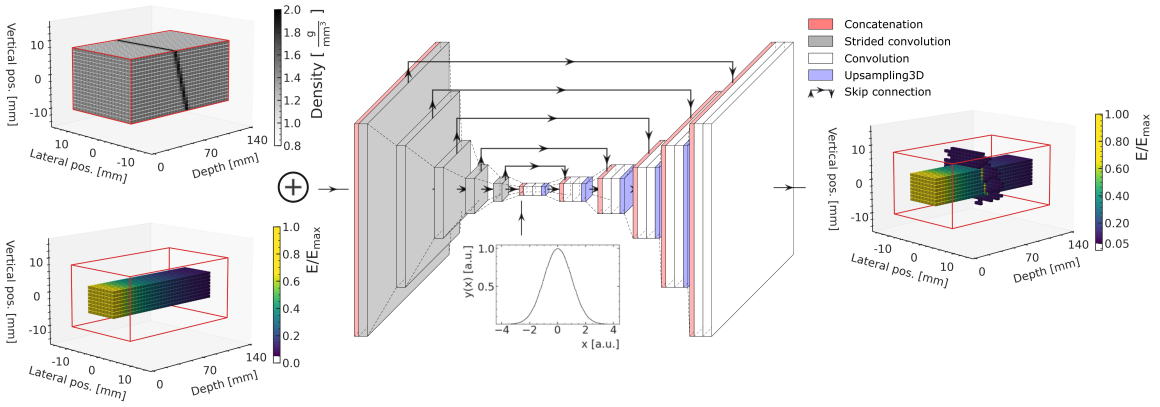


Figure 3.4: Schematic of the best generator 3D U-Net structure with input and output data. Reproduced from [16]

In addition to the density matrix also used in the Geant4 simulation, the generator is also provided with the Geant4 energy deposition simulation in a water-only phantom. This was found to improve the predictions at the edges of the field by simplifying learning the steep gradients occurring there. A motivation and possible explanation for the success of the inclusion of the water-only simulation is the simplification of the learning task for the network: Instead of learning to predict the entire prediction of the energy deposition from scratch, the network only needs to learn the impact that a bone slab insertion has on the provided water-only distribution. This impact is schematically shown in Figure (3.5).
Moreover, the used additional input describes both the shape of the field and the characteristics of the simulated beam, as both are embedded in the water-only Geant4 simulation result. There, this method potentially allows for the inclusion of different beam shapes and energies in later development stages.

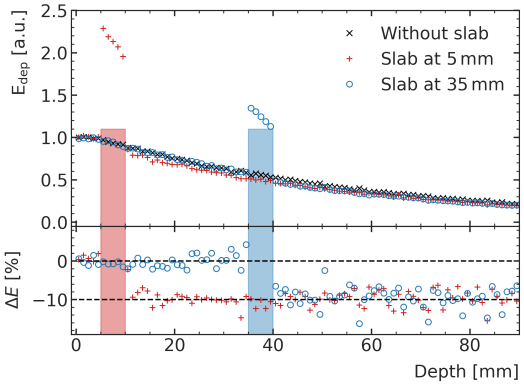
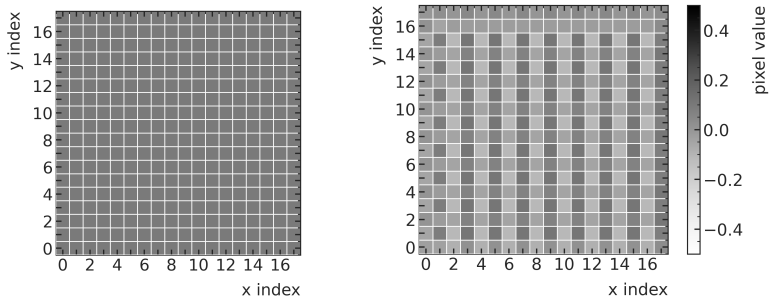


Figure 3.5: Visualisation of the impact a bone insert at different depths has on the energy deposition from a synchrotron broad beam.

On the compression path, the blocks of two convolutions with a subsequent pooling operation are replaced with a single *strided convolution* [82], which are often used in U-Net dose

prediction models (e.g. [83]) and allow a more sophisticated data compression. To allow for more interaction between the different regions in the prediction volume, additional U-Net stages (six instead of originally four) are included in the model. A random number vector (length: 100), which is required for the operation as a GAN, is inserted in the *bottleneck layer* in the centre of the phantom. Different insertion methods such as the concatenation to the input data led to less accurate predictions. Instead of using an increasing number of filters for each stage of the U-Net, all convolutions in the generator network are performed using 64 filters. A smaller number of filters was found to result in less accurate predictions. The originally proposed progression doubling the filters on every layer [59] was not feasible due to the memory required for this. 128 filters were the maximum number fitting into memory but did not notably increase the prediction performance while leading to very long training durations, resulting in a decision against them. On the decompression path, the originally proposed upconvolution operations are replaced with 3D upsampling operations, which increase the size by a factor of two by duplicating each voxel in each spatial dimension. The replacement of the upconvolutions was chosen after finding *checkerboard patterns* as a result of the application of that operation. An example for their occurrence is shown in Figure (3.6). Figure (3.6a) shows an image of constant pixel value. Figure (3.6b) shows the output following the application of an upconvolution operation. The effect has also been reported by other sources (e.g. [84]). It was found during the model optimisation that effect is reduced with increasing training duration, but this mitigation technique led to extremely long training times rendering the inclusion of upconvolutions infeasible.



(a) (b)

Figure 3.6: Schematic of the resulting checkerboard pattern (b) when applying an upsampling operation to an image with constant pixel values (a).

The originally proposed ReLU activation function is replaced with the novel *swish* function [85] due to faster model convergence. As optimizer, the Adam optimizer [70], using a learning rate of $\alpha = 2 \cdot 10^{-5}$ and a batch size of 32, which is limited due to the required memory, is used. A dropout regularization [86] of 15% applied to each convolutional layer was found to provide the best generalization results.

The critic model

The critic is built as 3D fully convolutional network [87] using strided convolutions with 128 filters each. It is is schematically shown in Figure (3.7). The input is a concatenation of the input of the generator, the density matrix and the water-only energy deposition, together with the respective simulation or generator prediction for the given density matrix. For all layers except the last, the swish function is used as activation. The final layer comprises a linear output function, producing a single number output for each sample representing a rating of their belonging to either the fake or real data set.

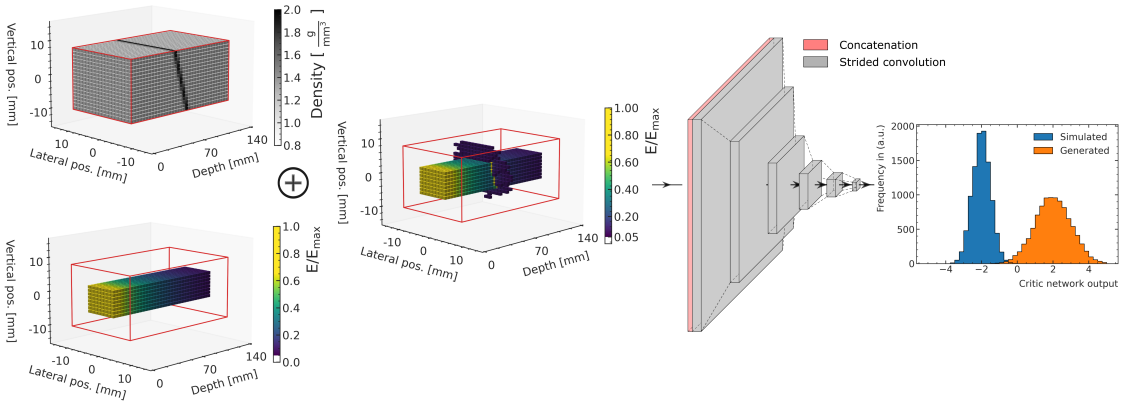


Figure 3.7: Schematic of the critic network with input and output data. Reproduced from [16]

As described in [71], it is required to restrict the weight update process for the critic during the training of a GAN. The best results were obtained using a *gradient penalty* term [88], while weight clipping [71] and the use of spectral normalization [89] resulted in less accurate predictions. It is found that training the critic five times for each weight update of the generator yielded a good compromise of good results and acceptable training duration. With regard to the optimizer and dropout, the same configurations as for the generator are used.

Prediction accuracy, speed and generalization of the best model

The development of the average passing rates over the training epochs up to the best model are shown in Figure (3.8). The best 1% validation passing rate ($(96.9 \pm 0.4)\%$) is obtained by training the described model for 1800 epochs. No further improvement was achieved with longer training times. The passing rates of the training and validation data are very similar, indicating good generalization and no overfitting.

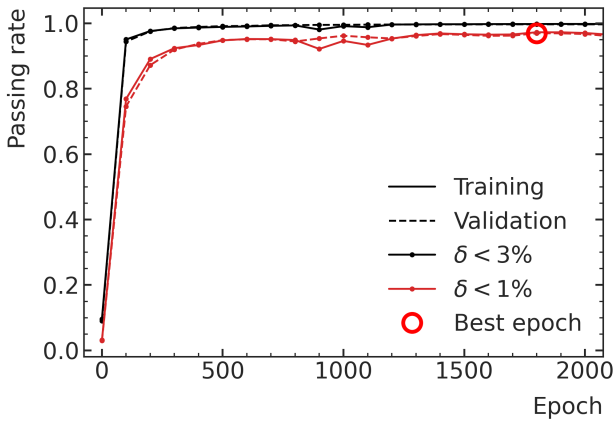


Figure 3.8: Passing rates as a function of trained epochs. The best 1% validation passing rate determining the chosen model is marked with a red circle. Reproduced from [16].

To assess the generalization, the obtained model is applied to the test data. A comparison of the average 1% and 3% passing rates on all three datasets are shown in Table (3.1). The

nominal passing rates is slightly higher for the training data set but this is found not to be statistically significant with respect to the uncertainties of the mean values.

Table 3.1: Average δ index passing rates for all three datasets as reported in [16].

	Passing rate [%]		
	Training	Validation	Test
$\delta < 1\%$	97.2 ± 0.5	96.9 ± 0.4	96.9 ± 0.7
$\delta < 3\%$	99.8 ± 0.2	99.7 ± 0.2	99.6 ± 0.2

To allow for a more systematic inspection of the passing rates, they are shown as a function of the bone slab rotation angle in Figure (3.9). The data samples within each validation and test segment as shown in Figure (3.2c) are grouped. The training data samples are grouped according to the angular separation shown there as well, although the central intervals are separated into two values each, resulting in the mean angular values in Figure (3.9).

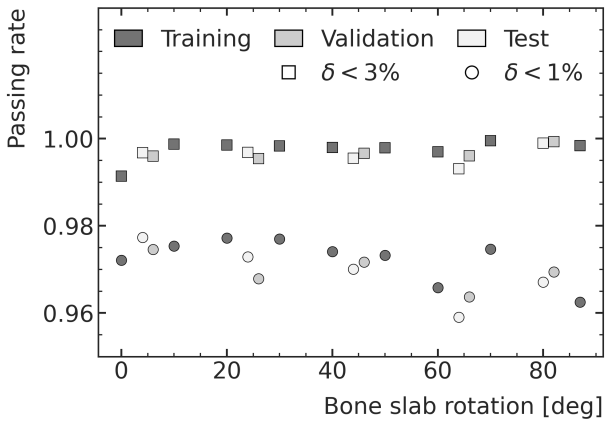


Figure 3.9: Grouped δ passing rates with respect to the bone slab rotation angle for the different datasets. Reproduced from [16].

For all three data sets and both passing rates, the obtained results are found to be generally in good agreement, confirming good generalization. With regard to the dependence on the bone slab rotation, it can be seen that the training passing rates are slightly lower for low rotation angles i.e. the bone being placed perpendicularly to the beam. This may result from the especially large gradients in this case. Larger rotation angles from 10° on achieve higher passing rates both for 1% and 3% deviation from the maximum energy deposition. Starting from 40° on, a downwards trend in the 1% passing rate can be observed in the training data. While the 3% passing rates of the validation and test data follow closely the results obtained on the training dataset, the 1% passing rates exhibit more deviation. Nevertheless, observed deviations are found to be small and mainly visible due to the chosen field of view in the figure.
Figure 3.10 shows 2D visualisations of the density profiles at the centre of the beam of two exemplary test data samples ((a): $\alpha = 4°$, (b): $\alpha = 80°$) together with the obtained passing rates. Especially the material interface between water and bone results in higher rates of the occurrence of larger deviations.

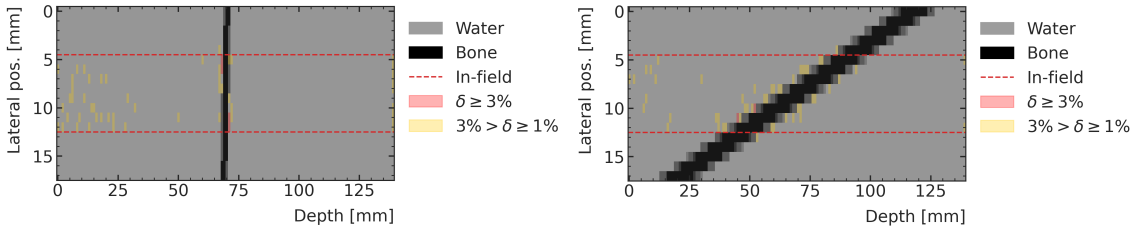


Figure 3.10: 2D slice of the density matrix of two exemplary test samples with a bone slab rotation of $\alpha = 4°$ (a) and $\alpha = 80°$ (b). Voxels with a deviation of more than 1% or 3% of the maximum of the energy deposition are marked with yellow and red, respectively. Dashed lines show the extent of the in-field region. Reproduced from [16].

Figure 3.11 shows the Geant4-simulated and ML-generated energy deposition depth profile together with the statistical uncertainty of the MC simulation at the centre of the beam (lateral pos. 0 mm, Fig. (3.10)) and the edge of the field of view (lateral pos. 17 mm, Fig. (3.10)). Shown examples include the two examples from Figure (3.10) and additionally a test data sample with a bone slab rotation angle of $\alpha = 64°$.

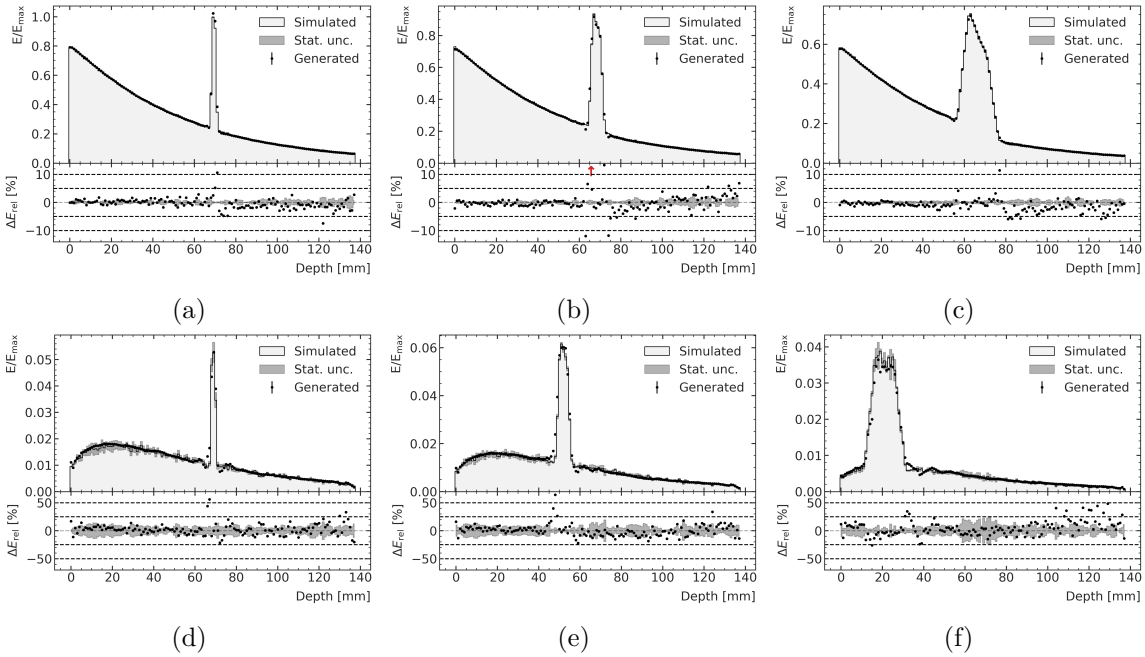


Figure 3.11: Comparison of Geant4-simulated (grey) and ML-generated (black) rnergy deposition-depth curves of exemplary test data samples. (a-c): Centre of the field. (d-f): Edge of the scoring volume. The bone slab rotation angles shown are $\alpha = [4, 64, 80]°$ (a+d, b+e, c+f). Below the curves, the relative deviations are shown, including outliers as red arrows. Adapted from [16].

Most individual relative deviations are below 5% in the field and below 25% percent in the out-of-field region. The latter is driven by the statistical uncertainties of the MC simulation which can be seen in the bottom row of Figure (3.11), showing that the deviations are often less than one standard deviation. Around the bone slab inserts, larger deviations both in-

field and out-of-field can be observed, whereas the energy deposition prediction within the bone region is mostly accurate within a few percent. Towards the distal end of the phantom, some ML predictions tend to systematically deviate from the MC simulations, as can be seen e.g. in Figure (3.11b). Nevertheless, the deviations remain within 10% of the energy deposition. From the entrance of the beam into the phantom up to the bone slab insert, the more frequent occurrence of *yellow entries* in Figure (3.10a) and Figure (3.10b), showing deviations larger than 1% of the maximum, results from the overall larger absolute value of the energy depositions in that area. The same effect suppresses the occurrence of any of those deviations in the out-of-field region by design, which led to their exclusion from the performance measure determining the best model. As shown for example in Figure (3.11d), the maximum energy deposition at the edge of the field of view does not exceed 5% of the maximum energy deposition at the centre of the field at all.
From the described findings, the model is found to be suitable for preliminary predictions of energy depositions. To be a candidate for the use as fast dose estimation engine in a treatment plan optimization scenario, the execution times play a very important role. The respective duration of a single energy deposition generation using either Geant4 on a central processing unit (CPU) (Intel Xeon E5-2630 v4 @ 2.2 GHz), or the generator model from the GAN using the same CPU or a GPU (*Nvidia GeForce GTX 1080i*) are shown in Table 3.2. The presented times do not include the preprocessing of the data set, which is done prior to the execution in all cases.

Table 3.2: Energy deposition prediction duration with the GAN generator compared to Geant4 MC simulations as reported in [16].

Model	Time per prediction [s]	Rel. speed
Geant4 (1 CPU)	$9.5 \cdot 10^5$	1
GAN (1 CPU)	0.6	$\sim 1.6 \cdot 10^6$
GAN (1 GPU)	0.1	$\sim 9.5 \cdot 10^6$

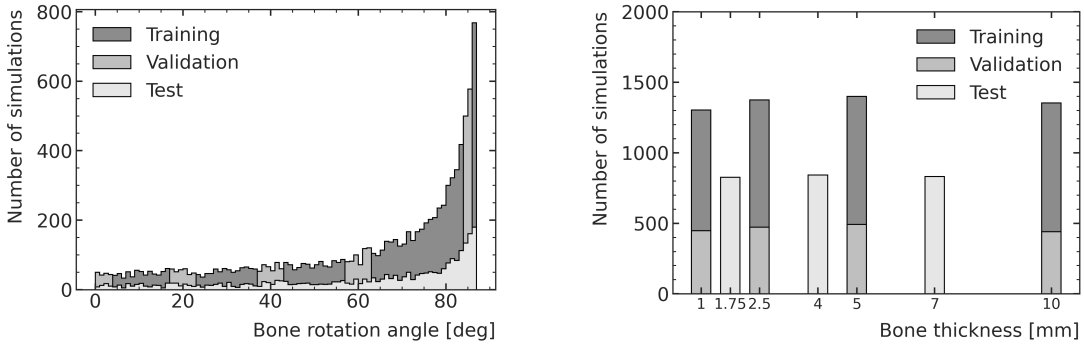
The Geant4 MC simulations used to generate the datasets within this study requires approximately 264 computing hours. In reality, this process is highly parallelized and therefore significantly shortened. Nevertheless, even considering additional methods to reduce the MC simulation time (e.g. by decreasing the very small step limit or increasing the production cuts used), the GAN generator takes only a fraction of the required computing time even on a CPU (0.6 s) and especially when using a GPU for prediction (0.1 s). Even compared to the MC step of the HybridDC model (30 min), the proposed ML model still achieves a significant speed-up. Batch processing additionally allows the prediction of up to several samples in parallel without notable increase in computation time, resulting in an even shorter effective prediction time.

3.2 Performance studies with more complex phantom models

The development of the model was done using an extremely simple target phantom model, which is relatively far away from realistic treatment scenarios. In this section, the transferability of the model to more complex phantoms is investigated. First, the previous bone slab model is extended. Afterwards, a simplified paediatric head phantom to test the model in a more realistic scenario.

3.2.1 Extended bone slab model

In the development of the microbeam radiation therapy (MRT) GAN model, the used bone slab model was restricted to only one slab thickness. In this second performance assessment, a new dataset is generated using a second bone slab model which also includes a variation in bone slab thickness. The rotation angles remain the same from 0° to 87° including flipping of the samples resulting in negative rotation angles. Considering the bone slab thicknesses, discrete values $d \in \{1, 1.75, 2.5, 4, 5, 7, 10\}$mm are used.
Figure 3.12 shows the separation of simulated samples into training, validation, and test data for this second study. Training and validation samples include only bone slab thicknesses of $d \in \{1, 2.5, 5, 10\}$mm. The remaining slab thicknesses $d \in \{1.75, 4, 7\}$mm are reserved for the test dataset. Considering the rotation angle, again several groups of rotation angles are excluded from the training data to serve as validation data The angles excluded are: $\alpha \in \{[0, 4) \cup [17, 23) \cup [37, 43) \cup [57, 63) \cup [84, 86)\}°$. The last validation interval is chosen shorter due to the significant changes in the geometry occurring at those rotation angles. The test samples are simulated with all bone slab rotation angles as they already differ in bone slab thickness. This choice of test data allows for the investigation of a more complex task for the network: as there are test data samples for which neither the rotation angle, nor the slab thickness are included in the training, requiring an interpolation in two dimensions of the parameter space.



(a) (b)

Figure 3.12: Stacked histograms of training (dark grey), validation (medium grey) and test dataset (light grey) with regard to their bone slab rotation (a) and thickness (b). Reproduced from [16]

Prediction accuracy and generalization assessment

The training with this second dataset is a bit less stable which can be noticed in the less monotonous development of the passing rates over the training epochs, shown in Figure (3.13). The best 1% validation passing rate is after training for 3,300 epochs which is substantially longer than for the simple bone slab model, translating to a wall time of several days of training.

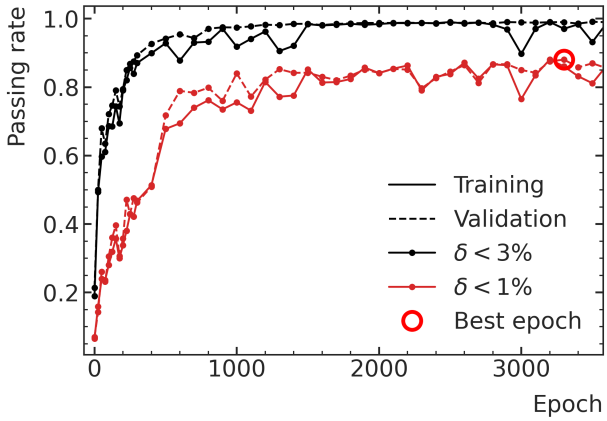


Figure 3.13: Passing rates against the training epochs using the dataset generated from the extended bone slab model. Reproduced from [16].

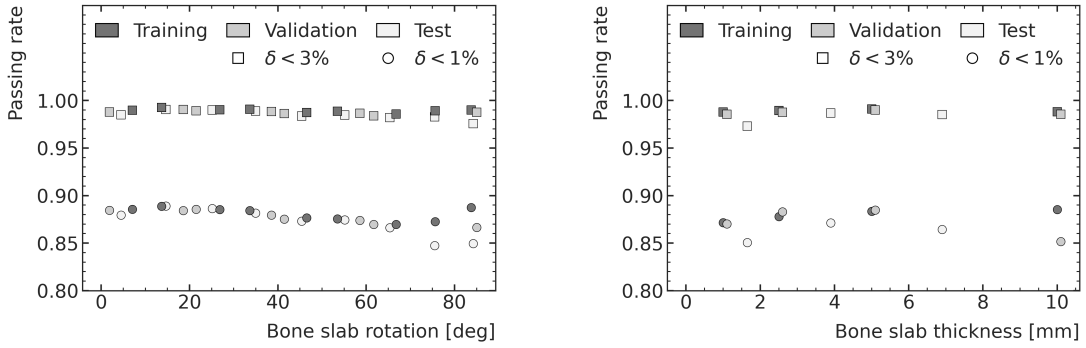
The obtained average δ passing rates using the final model to predict all samples in the training, validation and test dataset, are shown in Table (3.3). Within their reported statistical uncertainties, the passing rates are in agreement. This finding indicates successful generalisation of the model on the slab rotations and thicknesses not part of the training data set, thereby confirming the potential usability of the model also in more complex scenarios.

Table 3.3: Averaged δ index passing rates for the training, validation and test data sets as reported in [16].

	Passing rate [%]		
	Training	Validation	Test
$\delta < 1\%$	88.1 ± 0.7	87.7 ± 0.2	87.2 ± 1.4
$\delta < 3\%$	98.9 ± 0.2	98.8 ± 0.2	98.5 ± 0.4

All reported passing rates are found to be on a very high level. However, the reported uncertainties of the test dataset is higher than for the other datasets, especially for the 1% passing rate. The reason for this can be inspected more closely in Figure (3.14) showing the passing rates in dependence of the bone slab rotation angle and its thickness. The reported passing rates are averaged over the quantity not shown, respectively, in both cases.
With respect to the bone slab rotations (Figure (3.14a)), the average passing rates are very similar for the three datasets. A notable difference to the training with only a fixed bone slab thickness is that with the more complex phantom the decrease in performance for low slab rotations (bone slab perpendicular to beam) is not visible anymore.

For higher rotation angles, the result on the test dataset shows a lower passing rate as compared to the training and validation data. Especially the increase in passing rates for the training dataset while the performance on the test dataset drops is a strong indicator for overtraining in this part of the parameter space.
Moreover, Figure (3.14b) seems to indicate overfitting regarding the bone slab thicknesses as well because of lower scores for all test bone slab thicknesses compared to the respective training and validation datasets. This, however, needs to be analysed together with Figure (3.14a). Combining the two figures' results in the observation, that the lower results throughout all slab thicknesses stems from the high rotation angles. Following from this it can be concluded that while the interpolation between bone slab thicknesses seems to be successful for smaller slab rotation angles, clearly shown by the agreement of the datasets in Figure (3.14a), the model struggles with to generalisation for large rotation angles.



(a) (b)

Figure 3.14: Passing rates in dependence of the bone slab rotation angle (a) and its thickness (b). Reproduced from [16].

Exemplary energy deposition depth profiles, at the centre of the field and out of field, are shown in Figure 3.15, similar to the first simple bone slab model investigation. The shown data samples stem from the test dataset. More specifically, they all are samples with both a bone slab rotation angle and thickness excluded from the ML training. The performance of the model on those samples is of special interest to evaluate the potential of the proposed model.
Generally, all shown energy depositions are predicted relatively accurately, rarely exceeding 10% deviation in-field and 25% out-of-field. The previously observed weakness of the model considering high rotation angles is confirmed in the depth curve for $\alpha = 80°$ in Figure (3.15c).Right before and behind the bone slab insert the model systematically first over- and subsequently under-predicts the energy deposition. This is most likely to be associated with the model not generalising on those rotation angles and rather producing an output matching the learnt distribution for a different, in the shown case larger, rotation angle which is part of the training dataset. One reason making the prediction for large rotation angles especially difficult for the model is visible in Figure (3.15f): the bone slab does not reach the edge of the prediction field anymore due to the combination of its high rotation angle and the small thickness.
All in all, the model is found to perform well especially in the more realistic cases including less steep bone slab angles. The time required per prediction is not changed compared to the simple bone slab model as no hyperparameters of the model, including its structure, were not changed.

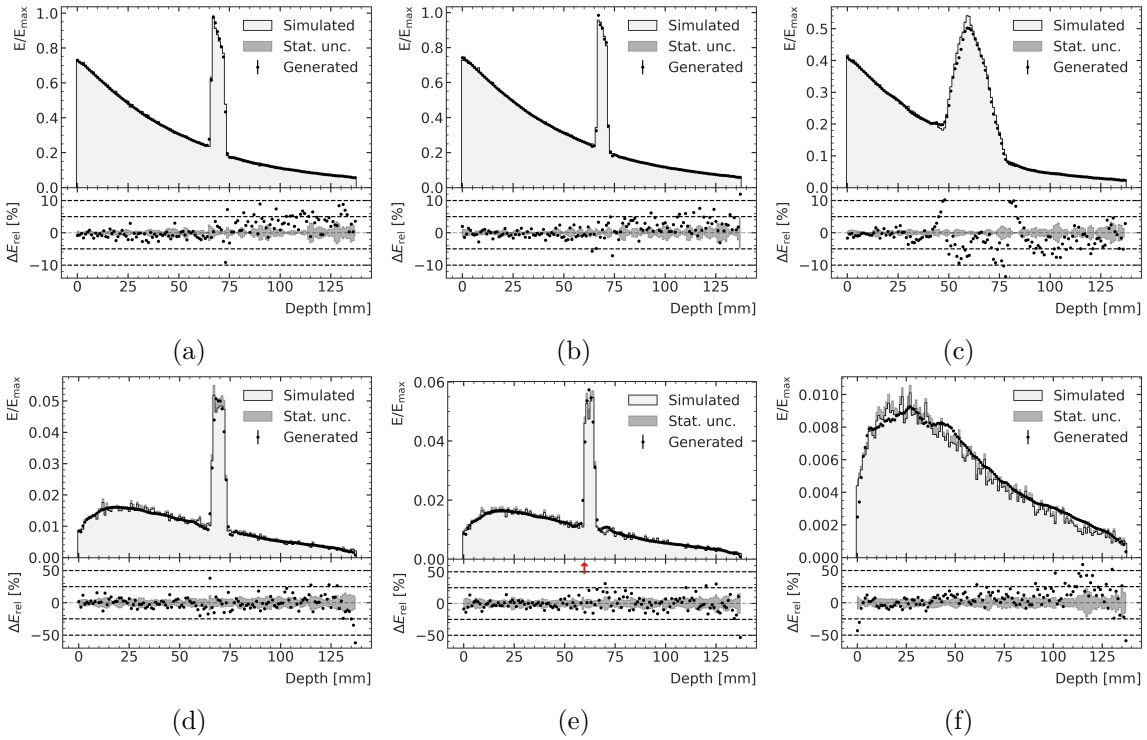


Figure 3.15: Comparison of Geant4-simulated (grey) and ML-generated (black) energy deposition-depth curves of exemplary test data samples. (a-c): Centre of the field. (d-f): Edge of the scoring volume. The bone slab rotation angles and thicknesses are $\alpha = [0, 40, 85]°$ $d = [7, 4, 1.75]$ mm, respectively. Below the curves the relative deviations are shown, including outliers as red arrows. Adapted from [16].

3.2.2 Simplified paediatric head

After two rather technical phantoms, the performance of the model for application on a more realistic scenario is investigated, posed by a simplified paediatric head phantom, built from nested spheroids. A schematic overview is shown in Figure (3.16).

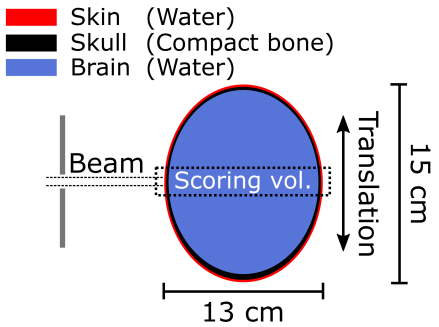
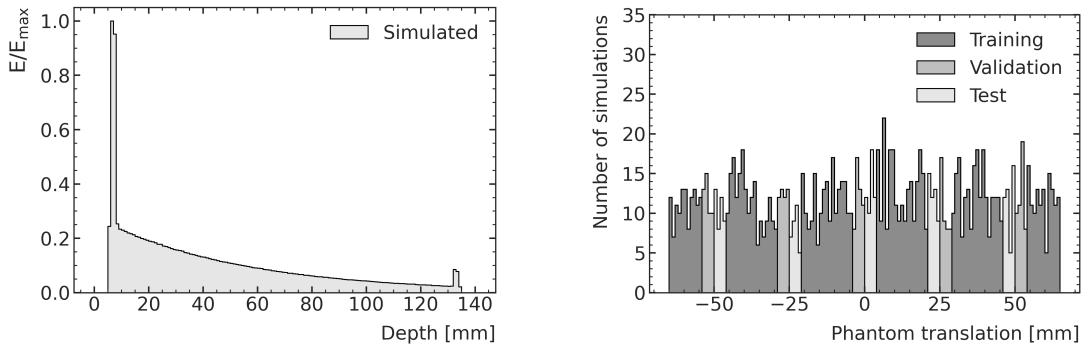


Figure 3.16: Schematic overview (not to scale) of paediatric head phantom used for MC dataset creation. The beam enters from the left, is shaped by the tungsten mask (dark grey), before entering the head phantom consisting up of skin (red), skull (black) and the brain (blue). The scoring volume is marked with a dotted line. Reproduced from [33].

The outer spheroid representing the skin as outer perimeter of the phantom has a length of 15 cm. The width and height are both 13 cm. Its assigned material in the simulation is water (*G4_WATER*, [78]). Inside of this spheroid, a bone spheroid (*G4_BONE_COMPACT_ICRU*, [78]) of 14.8 cm length and 12.8 cm width and height is placed. Because in the Geant4 framework, nested geometries automatically result in the inner objects overwriting the materials in place from outer objects. In this case, this results in the first, large water spheroid being reduced to a 1 mm shell around the inner bone sphere. Subsequently, another spheroid is inserted into the second one, resulting in total in a skin (water) layer on a skull shell around the inner one, presenting the brain and being simulated using water (*G4_WATER*, [78]). This inner sphere is 14.25 cm long and 12.6 cm both wide and high. Instead of being placed centred in the bone spheroid, the inner one is shifted 1.25 cm to the front. This results in a thinner bone shell at the forehead compared to the back of the head.

The phantom is placed perpendicular to the incident beam resulting in a radiation from the side of the head. In contrast to the previous simulations in which the whole scoring volume was covered in either water or bone material, in this scenario the air filling the simulation world (*G4_AIR*, [78]) fills up different amounts of the scoring volume, depending on the position of the phantom in front of the beam. An exemplary energy deposition-depth profile in the case of the head being centred in front of the beam is shown in Figure (3.17a). The energy deposition in air (*G4_AIR*, [78]) before entering the phantom is negligible. Subsequently, the thin skin layer is visible before the energy deposition peaks in the first bone layer. Next, the near-exponential decrease of the energy deposition inside the approximated brain volume can be observed before a smaller peak appears where the beam hits the distal part of the skull, followed again by the visible thin skin layer before energy deposition drops to approximately zero in air after the phantom again.



(a) (b)

Figure 3.17: (b) Exemplary energy deposition for a phantom centred on the beam. (c) Distribution of data sets in dependence of the phantom translation in front of the beam. Reproduced from [16].

To create a dataset for the training of the ML model, the phantom head is centred on the beam first and then translated horizontally in front of the beam to create different samples. The translations are sampled from a flat distribution between -65 mm to 65 mm, so that the in-field region is always aiming at brain volume and does not primarily aim on the skull of the forehead or back of the head. Similar to the previous sections, the resulting dataset is split systematically into training, validation and test data. The distribution into those three is shown in Figure (3.17b). For validation, the translations t in the ranges $t \in \{[-54, -50) \cup [-29, -25) \cup [-4, 0) \cup [25, 29) \cup [50, 54)\}$ mm are removed from the

dataset, for testing $t \in \{[-50, -46) \cup [-25, -21) \cup [0, 4) \cup [21, 25) \cup [46, 50)\}$ mm).

Prediction accuracy and generalization assessment

The best model training on data generated with the simplified paediatric head phantom is achieved after training for 6,900 epochs. The development of the passing rates over the training epochs is shown in Figure (3.18). Compared to training with the data of the extended bone slab model, the stability of the training further decreases, shown by larger jumps between subsequent passing rate values. Nevertheless, the model overall converges although the required training time is relatively long with nearly one week of wall-time.

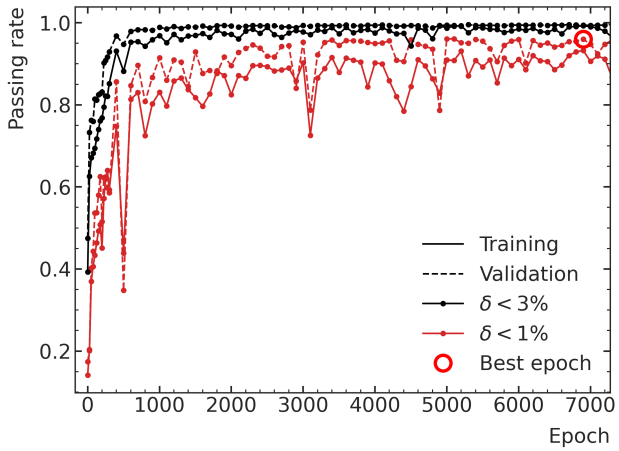


Figure 3.18: Passing rates against the training epochs using the dataset generated from the simplified paediatric head phantom. Reproduced from [16]

The averaged passing rates when applying the final model on all datasets is shown in Table 3.4. For calculation, voxels containing only air are not considered. In addition to the overall passing rates, δ_{brain} is also shown which is calculated only considering voxels inside the brain. This results in a more realistic performance estimation because the normalization factor in the δ index is the maximum energy deposition which occurs in the bone and is significantly higher than the maximum energy deposition in the brain.

Table 3.4: Average δ and δ_{brain} passing rates as reported in [16] of prediction using the ML model being trained on data from the simplified head phantom.

	Passing rate [%]		
	Training	Validation	Test
$\delta < 1\%$	93 ± 7	96.0 ± 1.7	96.3 ± 0.9
$\delta < 3\%$	99.3 ± 1.5	99.4 ± 0.5	99.6 ± 0.1
$\delta_{\text{brain}} < 1\%$	87 ± 10	90 ± 6	90.4 ± 2.4
$\delta_{\text{brain}} < 3\%$	98.9 ± 2.1	99.2 ± 0.4	99.5 ± 0.2

In the values summarized in the tables, the performances on the training samples appears lower than on both the validation and test dataset. This is a counter-intuitive finding and requires more investigation.

A reason for this result can be found in Figure (3.19) showing all reported passing rates for the three datasets as a function of the phantom translation. The performance of the model degrades for larger phantom translations and samples with large deviations are over-proportionally included in the training dataset over the validation and test dataset. For this reason, the absolute numbers as reported in Table (3.4) cannot be used alone as measure of model generalisation. Considering the more central part of the parameter space, from around $-40\,\text{mm}$ to $+40\,\text{mm}$, the predictions on training, validation and test dataset are very similar, indicating a good generalisation for that region.

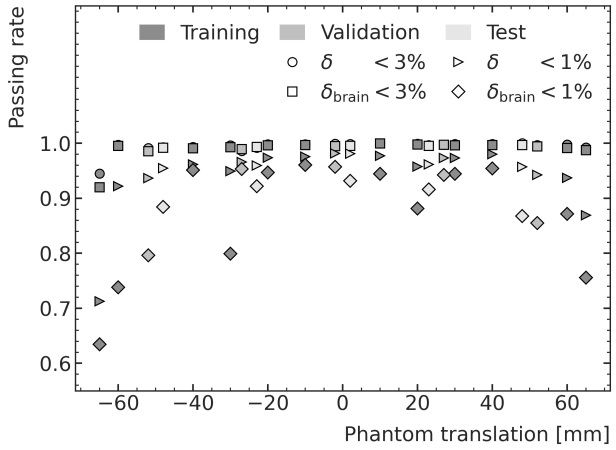


Figure 3.19: Passing rates in dependence of the phantom translation. The results are separated by dataset (dark grey: training, grey: validation, light grey: test), by the δ criterion (circle and square: 3%, triangle and diamond: 1%), and the volume taken into account being the whole phantom or only the brain. Reproduced from [16].

The decrease in performance stems from two main sources: (1) the further the phantom is translated, the less voxels overall are considered due to more voxels being covered with air only. At the same time a larger fraction of voxels contain bone, and a smaller fraction contains water material, resulting in more energy depositions with large absolute values. This difference between samples with the phantom centred on the beam and translated to the maximum can be seen in Figure (3.20a) and Figure (3.20b), respectively, showing the density matrix in the prediction volume at the vertical centre of the beam. Deviations considering those larger values contributes more significantly to the chosen δ measure. In Figure (3.20b), the occurrence of a relatively large number of voxels with $\delta > 3\%$ can be seen in bone voxels. (2) the prediction involving more bone material is generally more complex as more bone being traversed results in a energy deposition landscape including many voxels with large gradients.
Figure 3.21 shows the predictions of the ML model at the centre of the field and the edge of the prediction volume in comparison to the respective Geant4-simulated energy depositions. For the in-field region, all voxels inside of the brain are predicted accurately within 10% with most voxels deviating less than 5%, even for the worst-case example of $t = -65\,\text{mm}$, shown in Figure (3.21c). Larger deviations occur around the proximal and distal skull passing of the beam, leading to some values with larger deviations. Overall, the deviations increase with depth in the phantom. This is partly attributed to the increasing statistical uncertainty of the MC-simulated samples which can be seen by the enlarging of the shown error bands. Some systematic deviations are visible, such as the over-estimation of the energy deposition

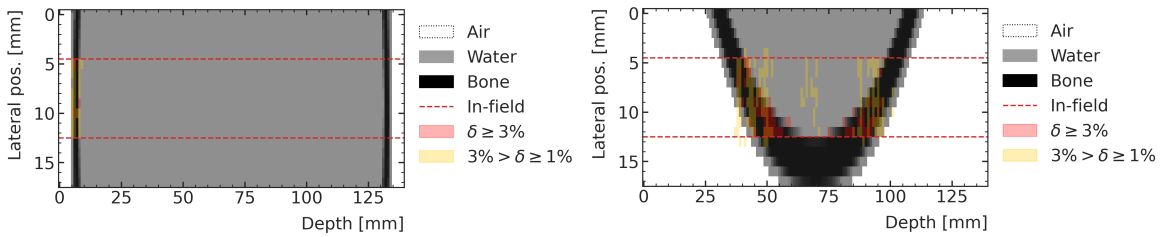


Figure 3.20: Comparison of Geant4-simulated (grey) and ML-generated (black) rnergy deposition-depth curves of exemplary data samples. (a-c): Centre of the field. (d-f): Edge of the scoring volume. The phantom translations are (a+d) $t = 2\,\text{mm}$ (test data), (b+e) $t = 48\,\text{mm}$ (test data) and (c+f) $t = -65\,\text{mm}$ (training data, worst overall example). Below the curves the relative deviations are shown, including outliers as red arrows. Adapted from [16].

in Figure (3.21a).

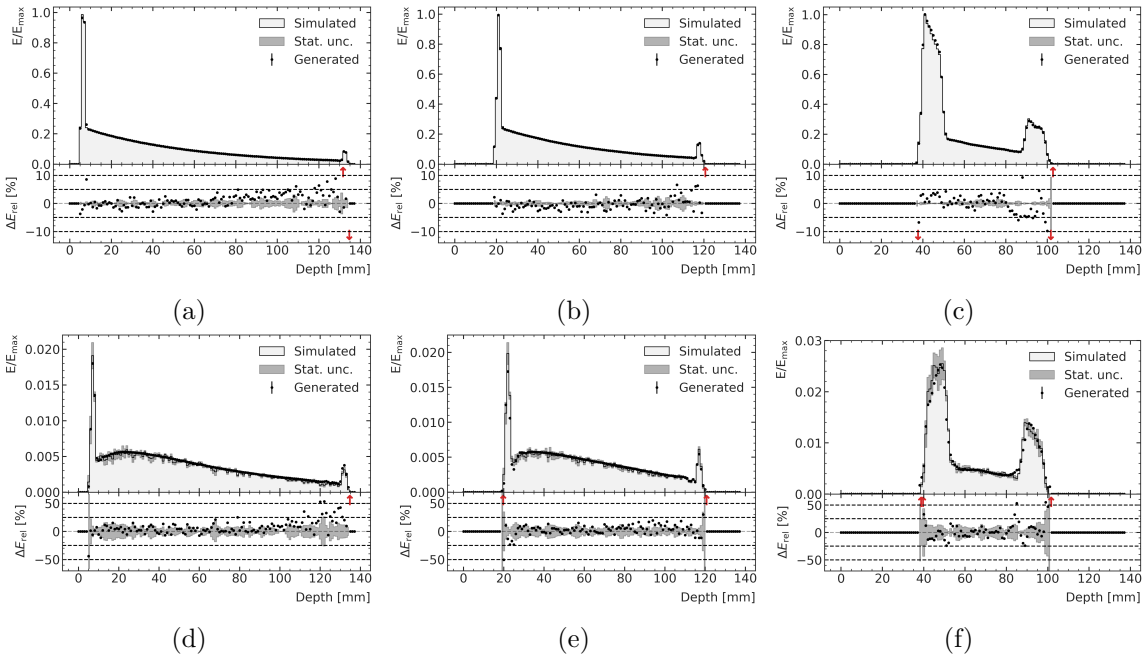


Figure 3.21: Comparisons of normalized simulated and generated energy depositions inside the phantom along the beam using the test data for the in-field (a-c) and out-of-field (d-f) region of the beam with phantom translations of $t = [2, 48, -65]\,\text{mm}$ (a+d, b+e, c+f). The lower part of the plots shows the relative energy deviation ΔE_{rel} in percent. Outliers are indicated with a red arrow. Adapted from [16].

3.2.3 Generalisation test with a CT-based skull model

In the following, a significantly more complex but interesting generalisation test is performed. The ML model trained on the simple head phantom data set is used to predict the energy deposition of a synchrotron broadbeam in a more realistic, Computer Tomography (CT)-based phantom. The simulation geometry is based on the ICRP110 phantom [90] which is

available as a Geant4 example [91].
The existent example reads a realistic CT-based head phantom and assigns materials to it. The voxels are of size $(dx, dy, dz) = (2.137, 2.137, 8.0)$ mm. $dz = 8$ mm is the slice distance of the used CT scan in vertical direction. Figure (3.22) shows the phantom being integrated into the previously used synchrotron broadbeam simulation. The energy depositions are still scored using the 140x18x18 voxel grid with a resolution of 1 mm.

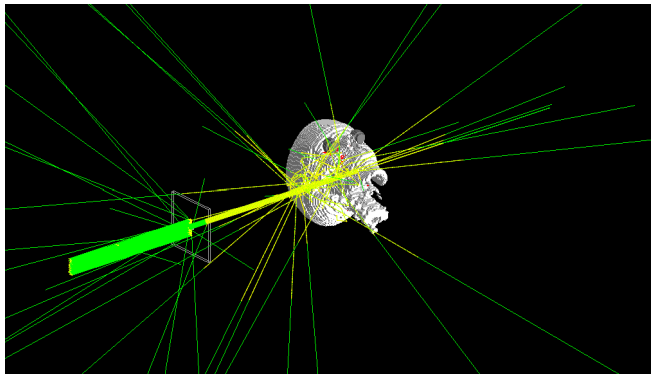


Figure 3.22: Screenshot from Geant4, showing the CT-based voxelized skull simulation. The tungsten mask is shown with wireframes, the green lines are the broad beam photons, and secondary electrons are shown in red. Yellow marks interaction points of particles with the simulation world. The realistic skull is shown in light grey, the brain inside of the skull is not visible.

To obtain a density from the Geant4 simulation as input for the ML model, both energy deposition E and dose D are scored. Subsequently, the density per scored voxel can be calculated as

$$\rho = m/V = m/E \cdot E/V = 1/D \cdot E/V = E/D/V \tag{3.2}$$

where V is the voxel volume of $1\,\mathrm{cm}^3$. An example of a derived density matrix is shown in Figure (3.23). While the overall features are similar to the simple paediatric head phantom constructed for the training of the ML model, there are several differences. First, the skin layer is significantly thicker. Also, the skull is constructed with a layer structure using a denser inner and outer layer of compact bone together with a less dense material in between, which represents the *cranium*. Where the double-layer structure cannot be seen, the cranium layer is too thin for the voxel resolution.

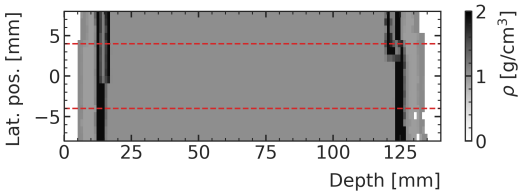
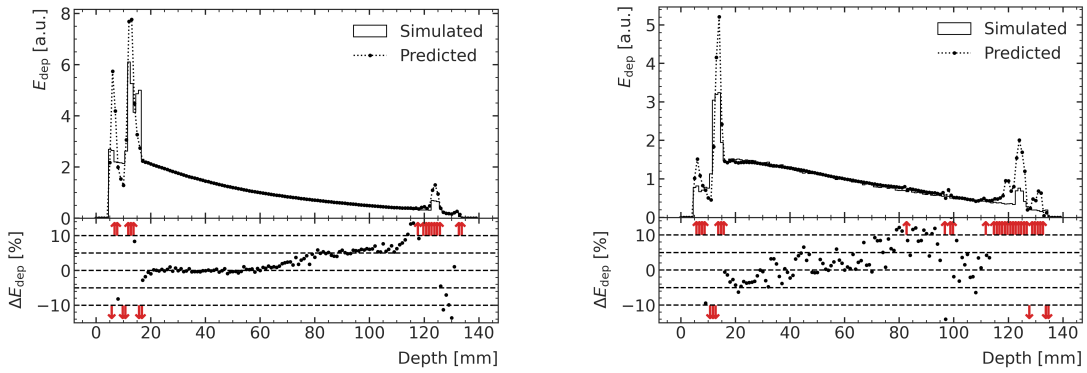


Figure 3.23: Exemplary density matrix in the prediction volume slice obtained from the ICRP110-based phantom. The red dashed lines indicate the in-field region.

The resulting ML prediction is compared to the Geant4-simulated energy deposition in Figure (3.24). The ML model predicts a peak of energy deposition right after the entrance into the phantom. This is likely due to the simple paediatric phantom exhibiting only the thin

layer of skin over the bone, which was not varied during the training, resulting in the model having *memorized* the energy deposition peak there. At the depth of the actual skull, the ML-model successfully predicts an increased energy deposition. However, the prediction is too high and too narrow, resulting from the training data only containing one layer of skull being made from compact bone instead of considering the less dense cranium in between compact bone layers as done in this more realistic phantom. In the region of the brain, in fact, the ML-model predicts nearly all voxels with an accuracy of 10%. Even in the out-of-field prediction, shown in Figure (3.24b), it can be seen that the predictions in the brain region rarely exceed 10%.



(a) (b)

Figure 3.24: Comparison of Geant4-simulated and ML-generated energy deposition-depth curves at the centre of the beam (a) and the edge of the prediction volume (b). Dotted lines connecting the predicted energy depositions are included as visual aid. The lower plots show the relative deviations, red arrows indicate deviations outside the shown range.

For different positions of the phantom in front of the beam, the deviations are comparable to the shown example. While the generalisation of the model is unsuccessful in terms of accurate energy predictions in all regions, the accuracy inside the brain is found to be a sign for a successful generalisation. It can be seen as plausible that in future studies the energy depositions in CT-based geometries might be predicted successfully without the acquisition of those for training data generation, by instead using a more sophisticated phantom including more physiological variations to train an ML model.

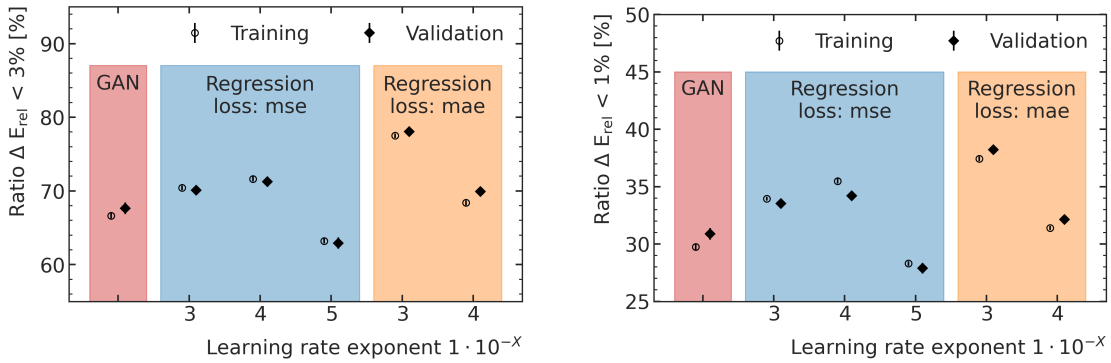
3.3 Is the GAN approach worth it? Comparison to regression models

Although the developed ML model has been found to be able of accurate predictions for a variety of phantoms, a large drawback are the very long training times of the GAN-based generator on the order of days to weeks on the used hardware in this thesis. In this section, neural network models identically built to the generator used in the developed model are trained as regression models instead. The target phantom used is the simplified paediatric head phantom developed in Section (3.2.2). Leaving out the critic network from training speeds up the model training process especially as it is trained five times for every weight update of the generator network.
The network is trained using the Adam optimizer with a batch size of 32. As loss function, the mean-squared error (MSE) and the mean-absolute error (MAE) are compared in their resulting performance, using learning rates of $1 \cdot 10^{-2}$, $1 \cdot 10^{-3}$, $1 \cdot 10^{-4}$, $1 \cdot 10^{-5}$. The ratio of voxels

Table 3.5: Mean absolute error and passing rates for in-field voxels based on different criteria for the validation data set.

Model	Dataset	MAE $[1 \cdot 10^{-3}]$	$\Delta E_{rel} < 1\%$ [%]	$\Delta E_{rel} < 3\%$ [%]
GAN	Training	2.4 ± 0.1	29.7 ± 0.3	64.3 ± 0.4
	Validation	2.0 ± 0.1	30.9 ± 0.5	66.2 ± 0.8
	Test	1.87 ± 0.02	30.0 ± 0.4	66.6 ± 0.5
Regression	Training	1.43 ± 0.04	37.4 ± 0.2	77.5 ± 0.3
	Validation	1.5 ± 0.1	38.2 ± 0.4	78.1 ± 0.7
	Test	1.35 ± 0.03	39.0 ± 0.3	79.5 ± 0.4

exhibiting a relative deviation of less than 1% and 3%, respectively, is used instead of the δ index for model comparison. Figure (3.25) shows the respective training and validation rates of voxels for the GAN and regression trainings with different configurations. The training with a learning rate of $1 \cdot 10^{-2}$ did not converge successfully. The training using a learning rate of $1 \cdot 10^{-5}$ the MSE loss was not finalized after it became obvious that it would result in worse results than the higher learning rates.The overall best validation results are achieved by the model which was trained as regression with an MAE loss and a learning rate of $1 \cdot 10^{-5}$. In fact, most regression configurations result in better validation results than the GAN training. The final performance of the best regression model and the GAN model are evaluated on the



(a) (b)

Figure 3.25: Rates of voxels with a relative deviation of less than 1% (a) and 3% (b) respectively for the GAN training and regression training using the MSE and MAE loss function together with different learning rates.

test data set. The summary of the performance measures is shown in Table (3.5). For the same reason as in the previous section, the reported results are better on the validation and test data. As computed by all used performance measures, however, the regression-trained 3D U-Net proves to allow for more accurate predictions than the GAN-trained model.
Figure (3.26) allows for a closer inspection of the accuracy of the predictions in dependence of the phantom translation. As seen in the previous section for the GAN model, also the performance of the regression-trained model decreases towards larger absolute translations.

The performance degradation is more significant for the GAN model.
The generalization of the model is evaluated to be satisfactory as no systematic difference can be found between the performance of training, validation, or test data as they all fall on a similar trend curve.

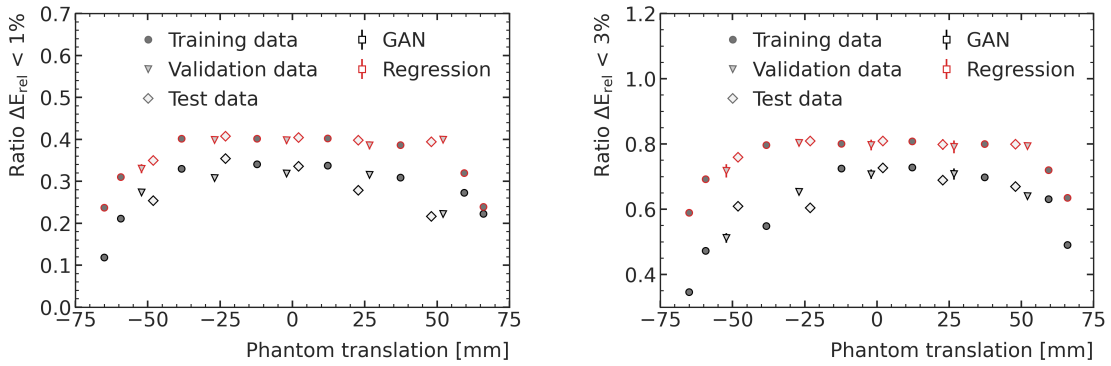


Figure 3.26: Performance evaluation using the 1% (a) and 3% (b) passing rates in dependence of the phantom translation on the training, validation, and test data sets.

When looking at the 2D or 1D depth energy deposition curves, the difference between the two models is clearly visible as well. Figure (3.27) shows a 2D slice of the energy deposition predictions from the GAN-based and the regression-based model and compares it to the Geant4 simulation result. The energy predictions outside of the head phantom do not contribute to

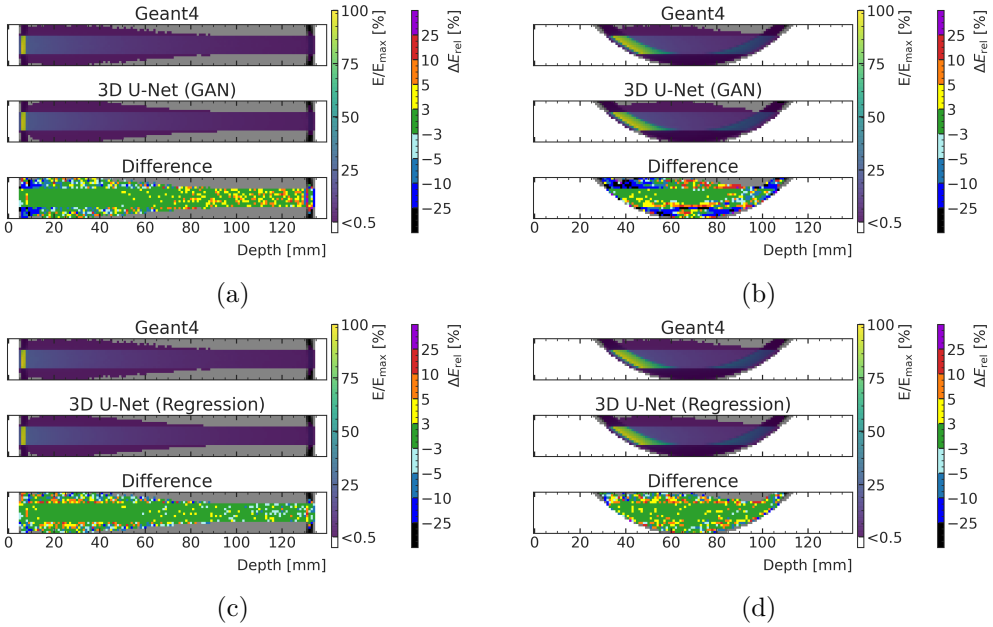


Figure 3.27: 2D energy deposition slices from the GAN-trained (a, b) and the regression-trained (c, d) networks at the vertical centre of the field of view for phantom translations of 2 mm (a, c) and -65 mm (b, d), compared to the Geant4-based simulation.

the performance measure as they can be assumed to be zero. Within the field, most deviations are below 5%, which can be seen in more detail in the 1D depth energy deposition

curves in Figure (3.28). Outside of the beam area, the deviations are larger and up to around 25%, although at a low absolute dose level. In the case of the predictions in the centre of the phantom, the GAN-based model underestimates the out-of-field dose at the entrance of the phantom while overestimating it from approximately the middle of the phantom. The regression-based model exhibits less biased predictions in the first half of the phantom and then tends to underestimate the dose in the distal parts of the phantom. In the case of the prediction at the extremes of the phantom translation, the GAN-based model exhibits strong underestimation in the area next to the beam, which is not visible for the regression-based model. The overall less biased prediction can clearly be seen using the regression model.
This can be confirmed looking at depth energy deposition curves in Figure (3.28) and comparing them to the ones shown in Figure (3.21f) for the GAN model: both the spread of the deviations are lower and the predictions show less of a trend throughout the phantom (overestimating for the GAN, underestimating for the regression network).

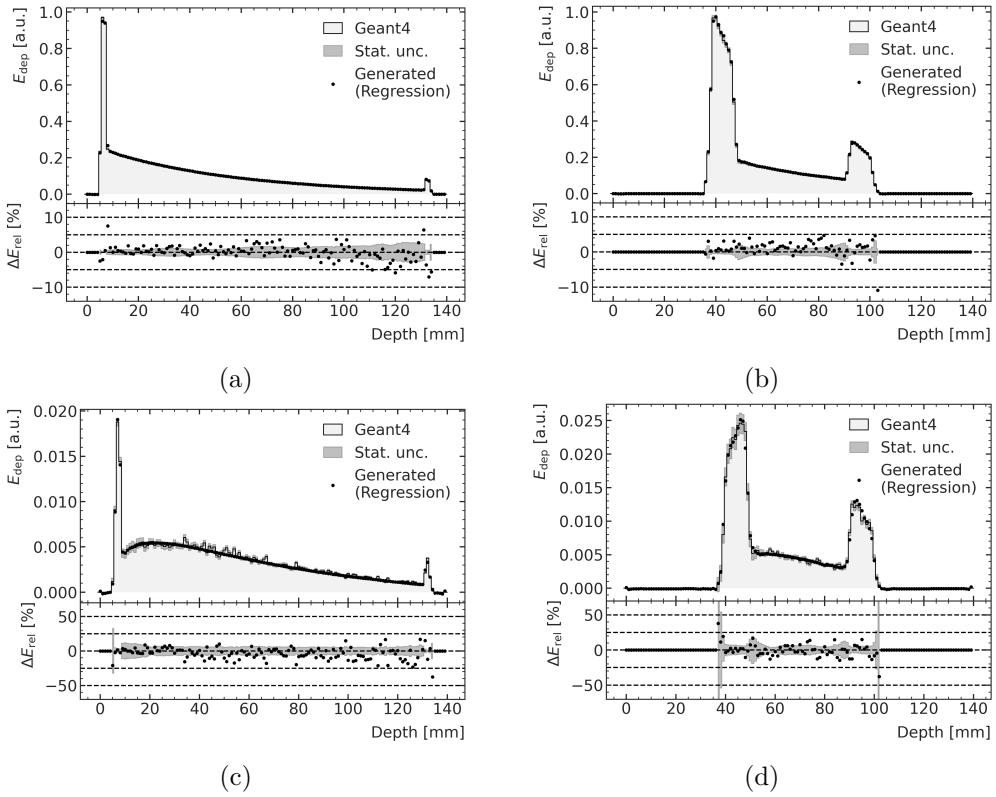


Figure 3.28: Depth energy deposition curves from the regression-trained (c, d) networks at the central line of voxels (a, b, e, f) and at the edge of the field of view (c, d, g, h) for phantom translations of 2 mm (a, c, e, g) and -65 mm (b, d, f, h), compared to the Geant4-based simulation.

3.4 Summary and conclusion of the proof-of-concept study

The presented 3D U-Net-based ML models are all capable of predicting energy deposition following an irradiation with a synchrotron broadbeam at the IMBL with an accuracy of about 10% within 100 ms when being executed on a GPU. Batch processing allows the parallel prediction of up to 32 samples, reducing the effective time required per prediction.
The model was developed using a simple bone slab model and its ability to be applied also to a more complex bone slab model and a simple paediatric head phantom was shown. In

an outlook application study, the fully trained model was found to even extrapolate to some extent to a more complex CT-based phantom, predicting the energy depositions inside the brain region within 10% although the phantom differs significantly from the training phantom. The observed accuracy does not suffice for the use as sole treatment planning dose calculation engine. Combined with its prediction speed, however, the models are found to be feasible candidates for preliminary predictions during treatment planning, enabling for example plan optimization or support online-adaptive treatment at some stage. Nevertheless, significant differences were found between the presented GAN approach and the regression training. The regression-trained energy prediction ML model is both significantly faster in training, allowing for a more in-depth search for optimal parameter configurations, and achieved more accurate predictions on the simple paediatric head model. One aspect of the GAN approach which as originally thought to be an advantage is its ability to adapt to statistical variations in the datasets and reproduce them. Instead, adapting to the noise in the training samples, i.e. statistical uncertainty of the MC simulation, is not a desired behaviour of the ML model in the case of energy deposition predictions based on an input geometry. In the presented studies, the use of the GAN-trained generator results in more noisy predictions, especially in the out-of-field regions, while no benefit for the quality of the predictions in-field can be observed.

While this study did not reveal any benefits of the GAN model, future studies involving training on significantly more complex phantom geometries might result in the flexibility of the GAN allowing it to perform better than a regression-trained model. Nevertheless, at this stage it seems improbable that a GAN approach is the best choice for an energy deposition prediction task.

4 Transfer & comparison study: predicting proton minibeams

microbeam radiation therapy (MRT) is not the only novel preclinical treatment method which currently performs dose estimations using time-consuming Geant4 simulations. After the regression model was found superior to the generative adversarial networks (GAN)-based generator in the previous section, these approaches are compared again in this section, testing their transferability to a different treatment: proton minibeam radiation therapy (pMBRT). In addition to comparing the two methods, the proposed model is also compared to the Dose Transformer (DoTA), a machine learning (ML) model based on the so-called *self-attention mechanism* [92] which was recently developed for predicting doses following proton pencil beam irradiations [93, 94]. It will be briefly introduced later in this section.
Most parts of the results shown in this section have been already published prior to the submission of this thesis [33].

4.1 Proton minibeam radiation therapy (pMBRT)

Utilizing a grid of sub-millimetre proton beams, pMBRT is classified as spatially fractionated therapy like MRT. since its proposal in 2013 [95], pMBRT has been reported in multiple studies to increase healthy tissue sparing while at the same time not decreasing the tumour control [96, 97, 98, 99].
Protons are charged particles and their energy deposition in tissue differs significantly from the x-rays utilized in MRT. The energy loss of charged particles with matter is described by the *Bethe-Bloch equation* [100, 101]. A central aspect described by the equation is that a decrease of particle energy leads to an increase of energy loss per track length, in turn further decreasing the particle energy. This process of accelerated stopping of protons in matter leads to the formation of the so-called *Bragg peak*, a phenomenon which, in fact, was discovered well before the formulation of the Bethe-Block equation [102]. The dose deposition behaviour following proton beam irradiation of a water phantom is shown exemplarily for a single proton minibeam with an energy of 100 MeV in Figure (4.1). The entry dose, as summarized over the width and height of the first layer of voxels, is used as normalization for the depth-dose curve and also the 2D visualisation. Doses depositions smaller than 1% of the entry dose are not shown.

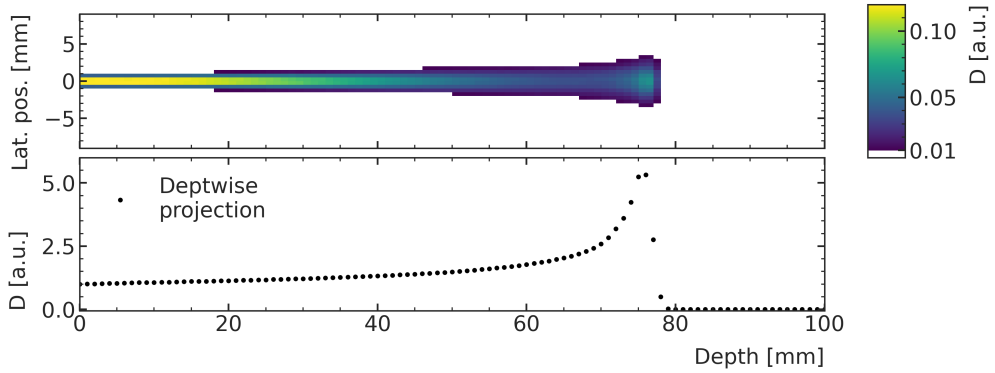


Figure 4.1: 2D dose deposition profile of a proton minibeam in water (upper plot) together with a depth-dose curve summarized over the full width and height of the phantom (lower plot).

The beam is simulated divergence-free with a diameter of 0.8 mm, similar to an early proof-of-concept-study for pMBRT [95]. The interaction of the protons with matter is simulated

using the Geant4 physics list *QGSP_BIC_HP* which is recommended for proton therapy applications [103]. The dose depositions are scored using a voxel grid with a resolution of 1x1x1 mm^3, which is adapted from the synchrotron broadbeam simulations in the previous studies.
The upper plot in Figure (4.1) shows the resulting 2D dose deposition profile. The entrance of the beam from the left can clearly be seen together with the abrupt stop with the Bragg peak at around 75 mm depth. The scattering of the protons in the water leads to a visible beam broadening. The lower part of Figure (4.1) shows the depth-dose profile summarized over the full width and height of the phantom, clearly showing the increase in dose deposition approaching the Bragg peak, which exhibits a dose of over five times the entry dose.
The voxel-wise depth-dose curve along the centre of the beam leading from the entry into the phantom up to the Bragg peak depends strongly on the chosen voxel size. Figure (4.2) shows the depth-dose curves in a single line of voxels at the centre of the beam for the used scoring resolution of 1x1x1 mm^3 and also for two more scoring resolutions of 0.5x0.5x0.5 mm^3 (2x resolution, grey) and 0.25x0.25x0.25 mm^3 (4x resolution, red). All curves are normalized to the maximum of the Bragg peak of the 1x1x1 mm^3 scoring. Using a finer spatial resolution, the Bragg peak is less pronounced because of the beam broadening due to scattering. While this is important for the intuitive understanding of depth curves shown later in this section, it has no further impacts on this study.

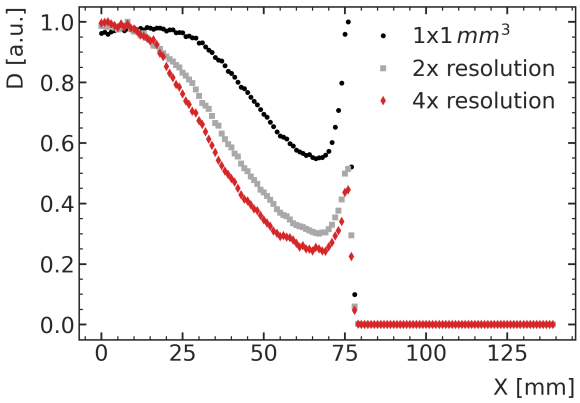


Figure 4.2: Dependence of the 1D voxel-wise depth curve on the voxel size.

4.2 Simulation dataset

The dataset being used for this study is produced using the simple paediatric head phantom introduced in Section (3.2.2). The phantom is centred in front of a single incident proton minibeam. The proton beam energy is varied between 20 MeV and 100 MeV in steps of 2 MeV. The smaller beam in comparison to the synchrotron broadbeam allows for a larger translation until reaching the edge of the brain model. Therefore, the phantom is translated continuously between $\Delta = -70$ mm and $\Delta = +70$ mm. As for the previous studies, the energy deposition is used for training the ML model instead of the dose. Exemplary energy deposition simulations for different beam energies and phantom translations are shown in Figure (4.3). Energy depositions are not shown if they are smaller than 1% of the voxel-wise maximum among all shown simulations.
For high energies and large translation values such as 100 MeV at -70 mm, the beam completely penetrates the phantom. This poses an additional difficulty for the model as the lack of the Bragg peak in this region of the parameter space differs significantly from other sam-

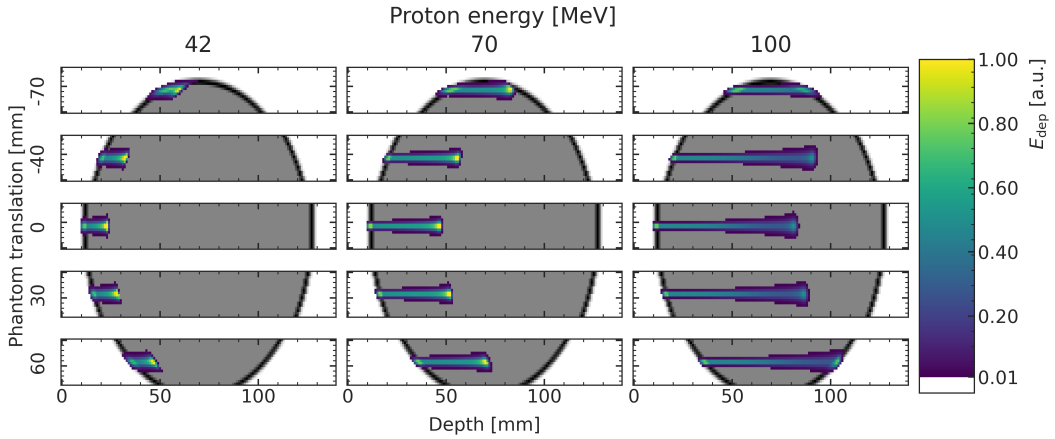


Figure 4.3: 2D profile of simulated energy depositions in phantoms at different translations using different beam energies. Adapted from [33].

ples. On the other side this scenario is seen as not very relevant clinically as the energy of protons entering the phantom are usually chosen so that the Bragg peak is located in the target region. Nevertheless, those samples are included in the dataset. The created simulation dataset comprises 2911 in total. From these, 720 samples are excluded for validation and 741 for subsequent testing, respectively. The split, like in the previous studies, is conducted in a systematic way to allow for better inspection of dependencies of the observed performances after training. The parameter space and its distribution into the different datasets is shown in Figure (4.4). The separation is chosen in a way that there are samples for which both the proton energy and also the phantom translation are excluded both from the training.

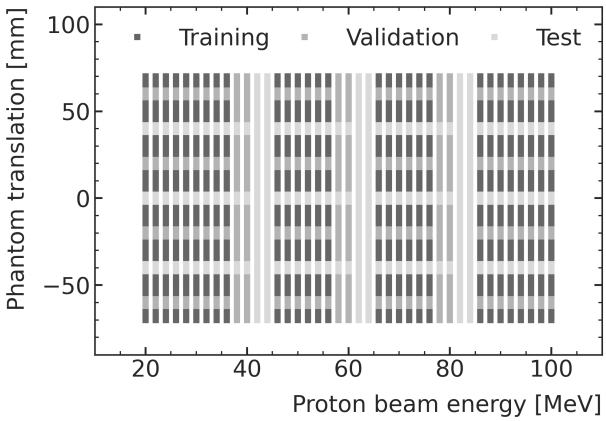


Figure 4.4: Separation of the simulated data samples into training (dark grey), validation (medium grey) and test data (light grey). Reproduced from [33].

4.3 ML models for proton minibeam prediction

This subsection first discusses adaptions of the 3D U-Net based model for application to proton minibeams before briefly introducing the DoTA model.

4.3.1 Adaption of the 3D U-Net-based synchrotron broadbeam model

In the previous section, two inputs were introduced as input for the generator: the density matrix of the predicted volume and the energy deposition in a water-only phantom. This was found to enhance the prediction quality in the case of the synchrotron radiation. A direct transfer of this model for proton minibeam prediction is shown in Figure (4.5).

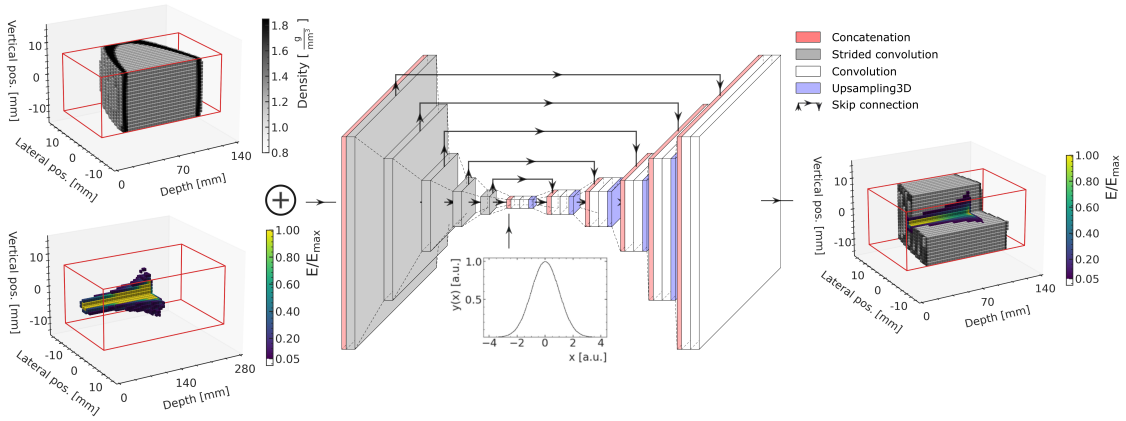


Figure 4.5: Adaption of the 3D U-Net-based generator model for proton minibeam prediction. Reproduced from [33].

In the case of proton therapy this method becomes more difficult to use. While the synchrotron x-ray beam can have a constant energy distribution throughout the irradiation, this is not the case for proton irradiations. Due to the Bragg peak leading to a confined beam range, its energy is constantly varied to deliver doses to different depths of the target. A continuous change in proton beam energy would result in infinite required water-only simulations being pre-computed and stored. The only option is to simplify the approach to only discrete energy steps. In that case, the inserted water-only simulation does not match the required energy for the prediction exactly. In addition to this deviation potentially rather degrading than improving the performance of the model, it is generally questionable whether the energy deposition profile form a water-only simulation helps the model at all because the model would have to shift the Bragg peak contained in the simulation result depending on the traversed material. The position change of the Bragg peak following the passing through two bone slabs of different depth is schematically shown in Figure (4.6).
Accurately learning how to modify the pre-computed energy deposition is suspected to be a at least as difficult for the ML model as predicting the energy deposition profile in the first place, therefore not being useful as additional input. To investigate those ideas, two different additional inputs encoding the proton beam energy are compared: the *water-only condition* and the *scalar energy condition*. For the water-only condition, all available proton energies in the dataset are simulated in a water-only phantom and stored for usage in the prediction, exactly as proposed for synchrotron beams in the previous section. For the scalar energy condition, a matrix of the same size (140x18x18 voxels), filled with the scalar value of the proton energy, normalized by the maximum energy of 100 MeV, is passed to the network instead.

4.3.2 DoTA: a transformer-based model

Transformer models are based on the so-called *attention* mechanism which was introduced as a technique for sequence translation [92]. A type of sequence translation that transformer

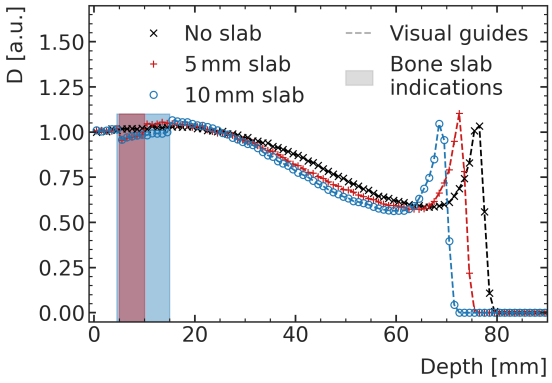


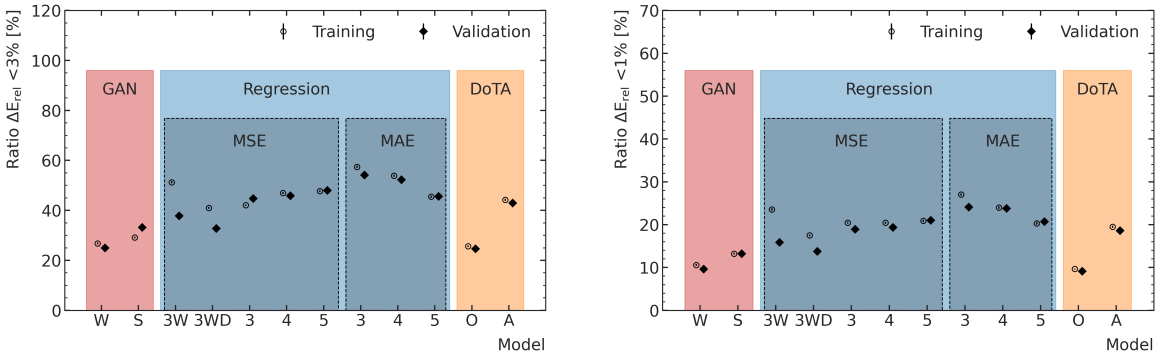
Figure 4.6: Visualisation of the different ranges of proton beams in water only and when traversing bone slabs with different thicknesses.

models have gained much public attention is the generation and translation of natural language content [104, 105]. The great success of transformer models not only in that domain but a whole range of applications (e.g. [106, 107, 108]) has also drawn the attention of researchers from the field of medical physics to them. The DoTA model presents an attempt to interpret dose prediction as translation task: Instead of translating e.g. a German text to English, a sequence of Computer Tomography (CT) slices is translated to a dose distribution [94]. The implementation of the DoTA model is derived from a publicly available Github repository [109] which is linked to the original publication.

4.4 Hyperparameter optimization

In the search for the best ML model, several hyperparameter configurations are explored. The models are trained until the ratio of voxels, which exhibit a deviation of the predicted from the Monte Carlo (MC) simulated energy deposition of less than 1%, does not increase anymore for 100 epochs. Voxels with less than 1% of the maximum energy deposition are not used for performance evaluation. Figure (4.7) shows the aforementioned 1% ratio averaged over the training and validation dataset, respectively. In addition, also the average ratio of voxels with less than 3% deviation between ML prediction and MC simulation is shown. The ML model configurations are described in the following, the x-axis shows model tags for better identification. Both the 3D U-Net GAN and regression models are trained with a batch size of 32, equal to the previous study.
The GAN model trained with the scalar condition (GAN-S) achieves a higher score than with the water-only condition (GAN-W), confirming the initial assumption that the water-only condition is not well-suited for proton beam predictions. The same is found for the regression model. The models *3W* and *3WD* are both trained using the water-only condition. The 3W model exhibits a significantly higher training than validation score, which indicates a lack of generalisation. As a countermeasure, the dropout used for regularisation in the model is increased in the model shown as 3WD. While the gap between training and validation scores is successful reduced, a degradation of prediction accuracy is also observed. Comparing the validation scores to the models being trained with the scalar condition (all regression models without *W* in the model indication), the scalar conditioned models perform generally better and more importantly show no or only small indications for overfitting as training and validation scores are very similar.
Out of the different regression models with the scalar condition, the use of the mean-absolute

error (MAE) loss together with a learning rate of $1 \cdot 10^{-3}$ performs best. Training with an even larger learning rate of $1 \cdot 10^{-2}$ did not converge. An interesting difference between the MAE and mean-squared error (MSE) regression is the observation that the models trained with MAE loss increase in performance when using larger learning rates while this tendency is reversed for the MSE loss.
Considering DoTA, an initial training using the original model, as found in the online repository [109], (DoTA-O) is compared with an adapted version (DoTA-A). The adapted version, using the MAE loss and the Adam optimizer instead of the MSE loss and the so-called LAMB optimizer [110], performs better yet does not exceed the performance of the best 3D U-Net regression model.



(a) (b)

Figure 4.7: Overview of relevant results from the hyperparameter optimization. Shown are the ratio of voxels exhibiting a deviation of ML-predicted energy deposition of less than 1% (a) and 3% (b) from the MC simulation both for the training and validation dataset. The x-axis indicates the used model with tags which are explained in the text. Reproduced from [33]

4.5 Performance assessment: prediction accuracy and generalization

After determining the respective best GAN, regression, and transformer model, each is applied to the test dataset. The DoTA model's execution speed is found to be significantly lower than previously reported [94]. Using the given experimental setup and hardware available in this study, the execution time is determined as (1.05 ± 0.04) s per prediction compared to (0.13 ± 0.07) s taken by the 3D U-Net models.
An overview on the comparison between training and test performance with regard to the average MAE and the ratio of voxels exhibiting a maximum of 1% and 3% deviation between ML prediction and MC simulation respectively, is shown in Table (4.1). It should be remembered that voxels with less than 1% of the maximum energy deposition are not used for performance evaluation. The shown metrics are not computed over the whole parameter space. In the previous section it was shown that the performance of the models degrades increasingly for larger phantom translations. The same is true for the proton beam prediction models. For this reason, a direct comparison between the scores obtained from averaging over all samples in the datasets is not representative for the generalisation, as samples with a large translation are contained also in this study in the training dataset. Instead, Table (4.1) shows scores averaged over all samples with a phantom translation of at most ± 50 mm.
Comparing the scores obtained from averaging over the aforementioned parameter range, the performance of each model is similar on the training and test datasets, indicating good

generalisation. The models, however, show significantly different levels of accuracy: the regression model is found to exhibit the best agreement between prediction and MC simulation with $(61.0 \pm 0.5)\%$ of the samples deviating less than 3% from MC.

Table 4.1: Performance comparison on the training and test datasets of the respective best GAN, regression and transformer model as reported in [33].

Model	Dataset	MAE $[1 \cdot 10^{-4}]$	$\Delta E_{rel} < 1\%$ [%]	$\Delta E_{rel} < 3\%$ [%]
3D U-Net	Training	11.64 ± 0.19	10.50 ± 0.19	30.6 ± 0.5
(GAN)	Test	12.98 ± 0.23	11.00 ± 0.18	33.1 ± 0.4
3D U-Net	Training	4.38 ± 0.01	25.87 ± 0.28	61.2 ± 0.5
(Regression)	Test	4.74 ± 0.03	25.62 ± 0.29	61.0 ± 0.5
DoTA	Training	5.27 ± 0.02	21.99 ± 0.28	48.6 ± 0.5
	Test	6.25 ± 0.06	20.45 ± 0.34	46.1 ± 0.6

Figure (4.8) and Figure (4.11) show the test dataset performance of the three final models in dependence of the phantom translation and the beam energy, respectively averaged over the other parameter space variable. The regression model is found to perform best over the whole parameter range, followed by the DoTA model which is only in a few regions outperformed by the GAN model, ranking last in the comparison. Like in the previous section, a decrease in performance of all three models towards larger phantom translations can be seen in Figure (4.8), leading to the adjusted reporting for comparing training and test performance in this section.

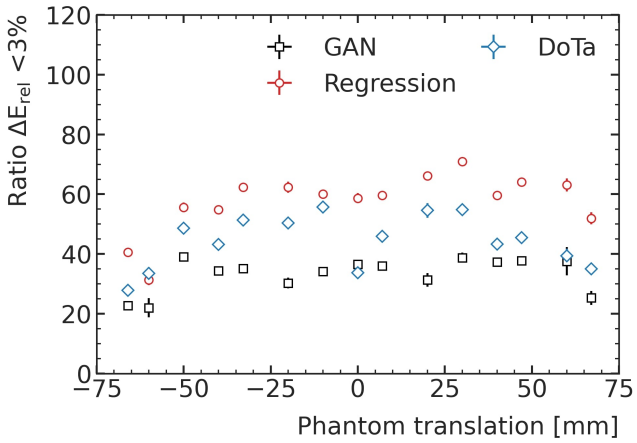


Figure 4.8: Ratio of voxels with less than 3% deviation between ML prediction and MC simulation of test data samples in dependence of the phantom translation, averaged over the proton beam energy. Reproduced from [33].

The decrease is asymmetrical which can be attributed to the asymmetry of the simple head model exhibiting a thicker skull layer at the back of the head, being targeted when using a negative translation. The increased amount of bone material in the path of the beam results in an increased prediction difficulty as the location of the Bragg peak changes quickly with more bone being traversed. In addition, the beam is incident on the skull under a steeper angle for

large translation values. This results in an asymmetric Bragg peak inside the brain, which in turn makes an accurate prediction even more challenging for the ML models. Examples of asymmetric Bragg peaks can be seen in an overview of 2D slices of ML predictions together with the corresponding MC simulations in Figure (4.9).

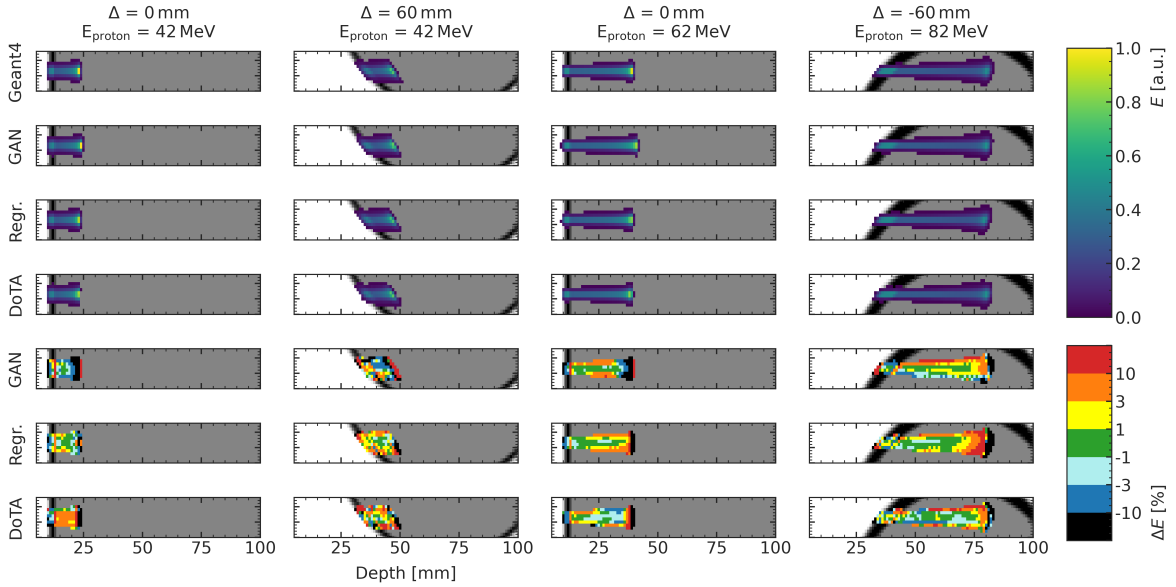


Figure 4.9: Exemplary comparisons between the MC simulation (top row) and respective ML predictions by the three models under investigation (indicated by the y-axis label), shown as 2D slice at the centre of the field and the voxel-wise relative deviation between ML and MC for test data samples with a phantom translations and proton energies of 0 mm and 42 MeV, 60 mm and 42 MeV, 0 mm and 62 MeV, and -60 mm and 82 MeV (from left to right). Voxels with less than 1% of the maximum energy deposition are not shown. Reproduced from [33].

Generally, the scores seem to be low with only two out of three voxels match the set accuracy level even for the best model. However, it has to be noted that the comparison is made in a relatively strict way to get a good differentiation between the models. As it can be seen in Figure (4.9), the majority of voxels with a large deviation do not occur in the centre of the beam but either further away from the beam centre where very low doses occur or around the Bragg peak. Two 1D energy deposition-depth curves of examples shown in Figure (4.9), highlighting the steep gradients involved and the effect of minor range deviations on the energy deposition, are shown in Figure (4.10): the combinations 0 mm and 62 MeV (1), and -60 mm and 82 MeV (2). In Figure (4.9), the GAN model exhibits a significant number of large-deviation voxels for combination (1) towards the end of the proton range, resulting in a low ratio of voxels with high prediction accuracy. Figure (4.10a) shows that this, in fact, is a result of a slight overestimation of the proton range for the given energy of about 1 mm. Similarly, combination (2), presenting actually one of the worst prediction cases of the regression model, shows large areas of deviating voxels in Figure (4.9). In Figure (4.10a), it can be clearly seen that this is a result from a range under-estimation.
While the accuracy around the Bragg peak is very important of course, already very small range deviations result in very large deviations between MC simulation and ML prediction. Many publications allow for a spatial deviation in the reported performance measure as well, increasing the reported score (see the gamma coefficient being used e.g. in [94]). This was found not to be instructive for the performance comparison of the models presented in this

section. However, it should be kept in mind when comparing the reported results to the literature.

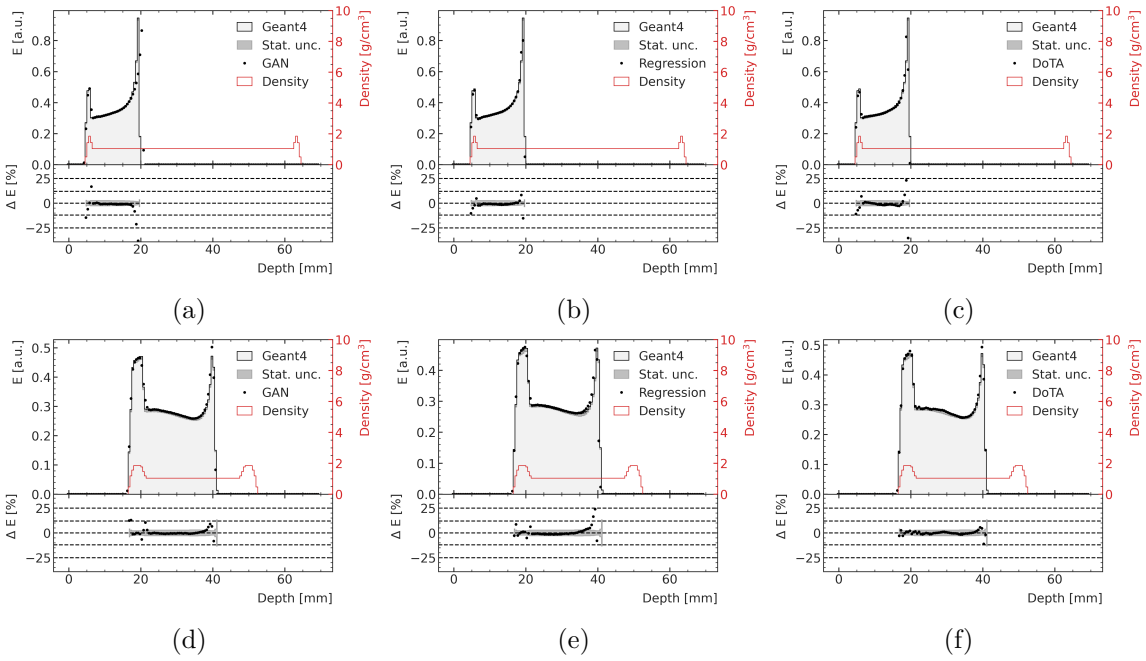


Figure 4.10: Exemplary 1D depth-wise comparisons of ML-predicted and MC-simulated energy depositions, shown for phantom translations and beam energies of 0 mm and 62 MeV (a-c), and -60 mm and 82 MeV (a-c), respectively. For easier visual localisation of the head phantom, the voxel-wise density is shown in red.

In Figure (4.11), the computed scores are averaged over all phantom translations, including the larger translations leading to lower scores, and instead being shown as a function of the proton beam energy. All models show an initial increase in performance peaking around approximately $50 - 60$ MeV, followed by a decrease towards the high energy region. Low energies are most difficult to accurately predict for all models. This is a result from the very short range of low-energy protons in the phantom resulting in very high gradients. This negatively impacts the performance two-fold: (1) fewer voxels exhibiting an energy deposition of at least 1% of the maximum dose results in a larger impact of individual voxels in the overall score. Inaccuracies with regard to e.g. the exact range cannot be compensated by many accurately predicted voxels in the plateau region leading up to the Bragg peak, like it would be possible for higher energies. (2) the involved gradients for low energies are steeper because the beam does not scatter as much as in the case of high energies before reaching the Bragg peak. The difference in the maximum energy deposition can be seen in exemplary 1D depth-wise energy deposition comparisons between the ML models and the MC simulation in Figure (4.10).
The regression model again achieves the best prediction results over the whole parameter range. It does, however, exhibit *dents* in the performance around energies being exclusively used for validation and testing, i.e. around 40 MeV, 60 MeV, and 80 MeV. This observation hints towards a lower degree of generalisation achieved by the regression model in those regions. This behaviour is not visible in the results obtained using the DoTA model. The structured separation into the three datasets allows for a closer inspection of the *local* generalisation in certain regions of the parameter space. This is done by computing the mean relative deviation between ML prediction and MC simulation as a function of the phantom

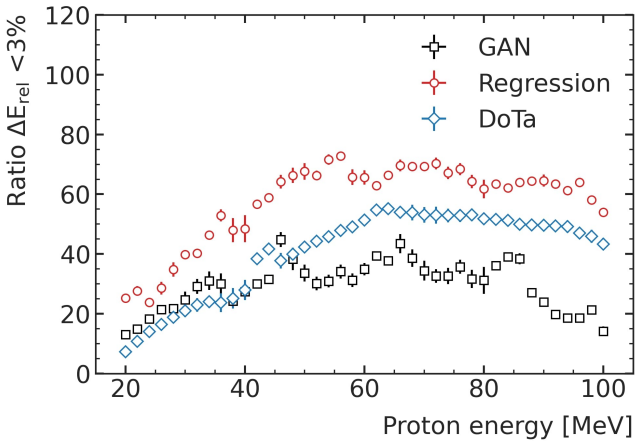


Figure 4.11: Ratio of voxels with less than 3% deviation between ML prediction and MC simulation of test data samples in dependence of the proton beam energy, averaged over the phantom translation. Reproduced from [33].

translation and the proton energy. The result for all three models is shown in Figure (4.12). The shade of grey in the background of the samples is included to make the visual inspection of the performances among the datasets easier. Training samples are shown with white background, validation samples in light grey and test samples in dark grey.

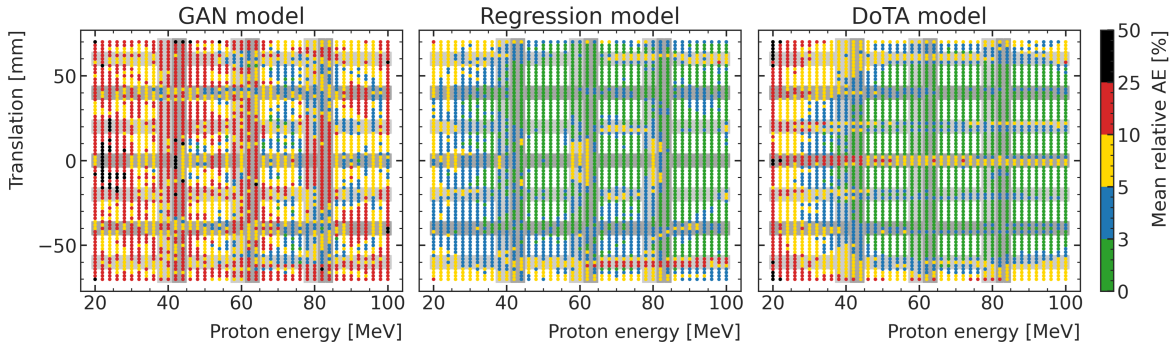


Figure 4.12: Mean relative absolute error (AE) obtained from predicting all datasets. The shade of grey in the background of shown points indicate their respective dataset as described in Figure (4.4), i.e. training (white background), validation (light grey background) and test (dark grey background). Reproduced from [33].

The GAN model is once more found to generally perform significantly worse than the other two models. The trends form the previous performance assessment plots can also be found here: predicting lower proton energies and larger phantom translations is overall more difficult for both the regression and DoTA model. Nevertheless, both models predict the energy deposition distributions with an average deviation of less than 10% (green, blue or yellow) for most parts of the parameter space.
The energy-wise *dents* in the performance of the regression model, visible in Figure (4.11), can be found using the mean deviation measure in this plot as well. Although significantly less prominent around 40 MeV, the regression model is found to systematically perform lower in the regions of validation and test data around small phantom translations (yellow regions). This indicates that the regression model in fact struggles more than the DoTA model with generalising well with regard to the proton beam energy, directly translating to deviations

in the predicted proton range. In contrast, the DoTA model is found to exhibit lower performances for the left-out phantom translations along the whole proton energy range. The DoTA model seems to be well-capable of interpolating between proton energies but struggles more with the interpolation between different geometries. As a consequence, the parameter space would have to be more densely populated with regard to the phantom geometries to improve the DoTA performance whereas the 3D U-Net regression model would benefit most from additional samples along the proton energy parameter space direction.
For the application in preclinical contexts this could be interpreted as an advantage of the 3D U-Net regression model. Using MC simulations, it is easy to provide the model with wide range of beam energies. The construction of different phantoms, however, is limited by a usually rather small available number of e.g. CT samples from preclinical patients. This limitation of the DoTA model coincides with the general observation that transformer-based models rely on very large datasets to perform well [111]. Therefore, models like DoTA are expected to provide more benefit for application cases where the creation of such large datasets is possible or a dataset is even already available. In conclusion, the 3D U-Net regression model is found to be the strongest candidate for developing ML models for new treatments thanks to its transferability and high accuracy in scenarios of small training datasets. It is therefore used in the further course of this thesis.

5 The next step: development of a microbeam dose prediction model

This section covers the development of the developed model from synchrotron broadbeam prediction to the actual prediction of microbeam fields. A possible approach would be training two machine learning (ML) models for predicting the primary and secondary dose following the HybridDC Monte Carlo (MC) scoring method and subsequently applying the electron dose kernel to obtain the final field. While this has the advantage of integrating well with existing techniques, it comes at the disadvantage of relying strongly on the HybridDC model and its approximations. Instead, a stand-alone method for microbeam dose prediction with no further dependencies except the underlying MC simulation is developed in this section.
Parts of the results shown in this section have been already published prior to the submission of this thesis [32].

5.1 Implications of MRT dose prediction by microbeam superposition

A naive and straightforward way to extend the developed model towards the prediction of a full microbeam radiation therapy (MRT) field is by superposition of individual microbeam predictions. A schematic of this approach is shown in Figure (5.1). First, the ML broadbeam model (left) is retrained on a new dataset which exhibits a smaller voxel size and includes only the dose deposition from a small part of a single microbeam (centre). Using this model, the microbeam array can be constructed by superposition (right).

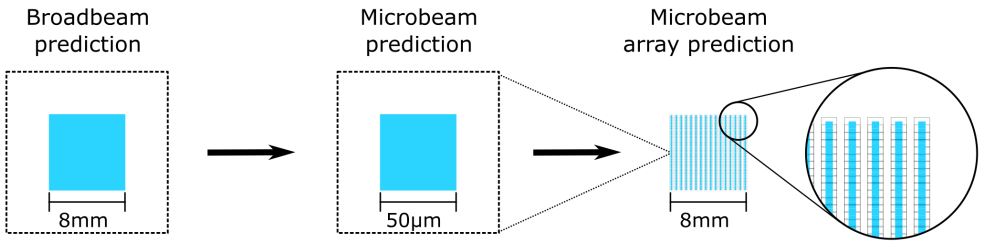
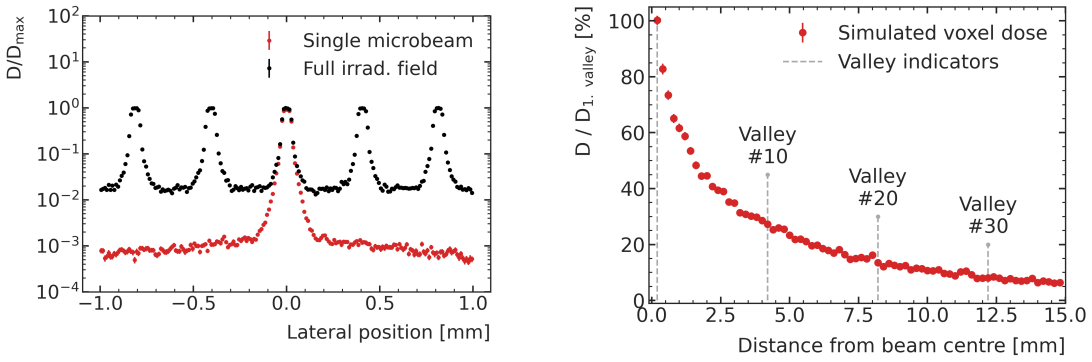


Figure 5.1: Schematic of the procedure of predicting microbeam arrays using a superposition of individual microbeams.

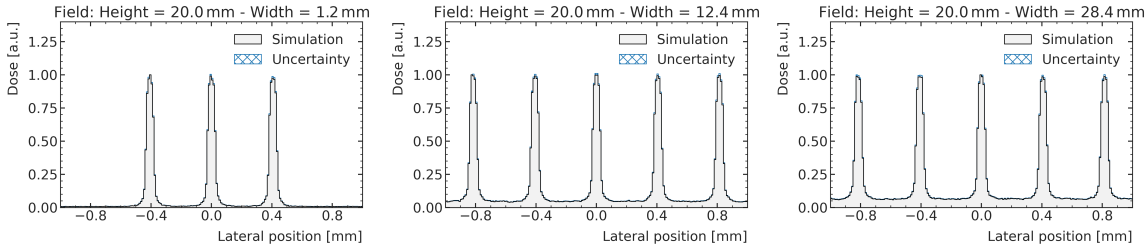
While this method comes with the disadvantage of a substantial increase in computing time because of the repetitive execution of the ML model, it easily allows producing dose distributions for variable field shapes resulting in high flexibility. A difficulty in implementing this method is the required size of the prediction field. Figure (5.2a) shows the lateral dose profile which results from the irradiation of a 14x14x14 cm^3 cubic water phantom with a single microbeam (red) and a 20 mm wide MRT field comprising an array of microbeams (black), scored with a spatial resolution of 5 μm in 7 cm depth.
Moving away from the position of the single microbeam, the dose contribution decreases steeply, already being reduced by a factor of nearly 1000 when in the region of the first valley after 200 μm for a micrometre spacing of 400 μm. When comparing the valley dose of the MRT field containing multiple microbeams, it is around one order of magnitude larger than the dose that is contributed by the single microbeam. This means that not only the next neighbouring microbeams contribute to their respectively adjacent valley, but many neighbouring microbeams have to be considered. Figure (5.2b) shows the dose that the



(a) (b)

Figure 5.2: (a): Comparison of a dose profile from a single microbeam (red) and a 20 mm wide microbeam array field (black). Reproduced from [32]. (b): Dose contribution of a single microbeam as a function of distance from the beam centre, normalised by the dose it contributes to its adjacent valley. Indicators show the 10th, 20th and 30th valley away from the microbeam. Adapted from [32].

shown single microbeam contributes in even larger distances to the beam centre. The doses shown are normalised by the contribution of the beam in the adjacent valley right next to it. Even in the 10th valley away from the single microbeam, it still deposits as much as 30% of the dose it contributes to the valley right next to it, 30 valleys away the contribution still comprises 10% of the next-neighbouring valley. The resulting effect on the valley doses in a field being constructed from a different number of microbeams is shown in Figure (5.3) shows the dose profiles following the irradiation of a with microbeam fields comprising a different number of microbeams.

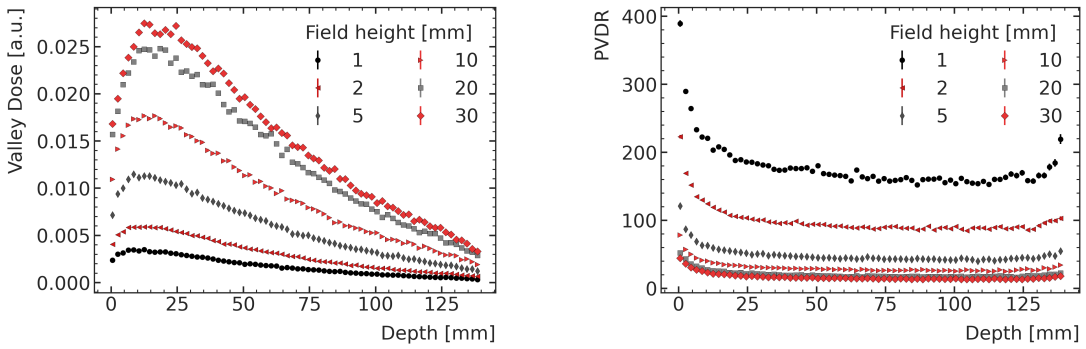


(a) (b) (c)

Figure 5.3: Dose profiles at the centre of a cubic water phantom following irradiation with an MRT field of 1.2 mm (3 microbeams, a), 12.4 mm (31 microbeams, b) and 28.4 mm (71 microbeams, c) width and 20 mm height.

Figure (5.3a) shows the dose profile from an MRT field comprising 3 microbeams, Figure (5.3b) from 31 microbeams, and Figure (5.3c) from 71 microbeams. The peak doses stay nearly identical between the three fields. However, the valley doses between the peaks can be seen to increase. Especially the increase in valley dose between Figure (5.3b) and Figure (5.3c) is notable. The distance of added microbeams from the shown part of the dose profile is upwards from 6 mm to each side. Given a microbeam width of 50 μm, this means that even at a distance of 1200 times its width, a microbeam interferes with the valley doses in other parts of the field.

While this effect was now discussed with respect to a fixed field height and a variable width by including a varying number of microbeams in the field, different field heights also impact the valley doses between the peaks. Those, in turn, still stay mostly constant independent of the field size. Figure (5.4a) shows the valley dose in the first valley next to the central microbeam for a fixed-width field of 20 mm, in this case in dependence of the field height between 1 mm and 30 mm. The valley doses are normalized to the peak dose.



(a) (b)

Figure 5.4: Valley dose (a) and peak-valley dose ratio (PVDR) (b) in dependence of the field height.

Up to the largest field size, the valley doses keep increasing. However, the increase is not linear, which can be seen for example by comparing the 2 mm-field peaking at about 0.6% peak dose (0.006 on the shown axis) with the 20 mm-field peaking at about 2.5% peak dose (0.025 on the shown axis), which amounts to roughly four times the valley dose for a field of ten times the size. In addition to the increase in peak valley dose, a larger field size also results in the valley peak occurring slightly deeper in the phantom. For the smallest field with 1 mm height, the valley dose peaks approximately 7 mm from the surface, going up to a peak at approximately 17 mm from the surface for a field of 30 mm height. The respective depth-wise development of the PVDR for the central valley is shown in Figure (5.4b). Figure (5.5) combines the results regarding the field width and height by showing the PVDRs of the central valley at the centre of the phantom, allowing an easier comparison of the effects resulting from different field sizes. It can be clearly seen that the PVDR decreases significantly for increased field heights and widths. The change in overall PVDR, together with the shift of the peak of the valley dose in the phantom, might be an important aspect to consider when making treatment planning decisions in the future.

Returning to the implications for the mircobeam superposition approach for MRT dose prediction, the long-ranging impact on valley doses poses a very difficult problem. It has been shown that even as far as 15 mm away from a 50 μm wide beam (a 30 mm field size results in 15 mm distance from the centre), its impact is significant and cannot be neglected in a superposition approach. This would result, however, in the need for a similar-sized prediction volume. The previously developed model was capable of predicting a volume of 140x18x18 voxels. Assuming this could be doubled towards the lateral and vertical field size, it would result in 140x36x36 predictable voxels, allowing for a voxel size just below 1 mm. Figure (5.6) shows the effect that too large voxel sizes have on the accurate determination of the peak and valley doses.

Already a voxel size in the same order of magnitude as the nominal microbeam width (50 μm) results to an significant averaging effect, reducing the average dose within the voxel to about

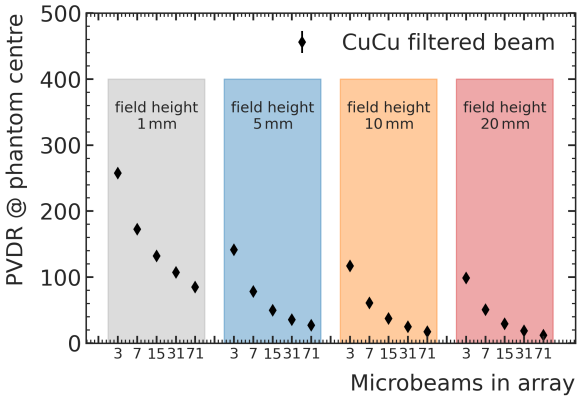


Figure 5.5: PVDR at the centre of a 14x14x14 cm^3 water phantom following the irradiation with different sized MRT fields. The number of microbeams in the field array are a measure of the field width (400 μm microbeam pitch).

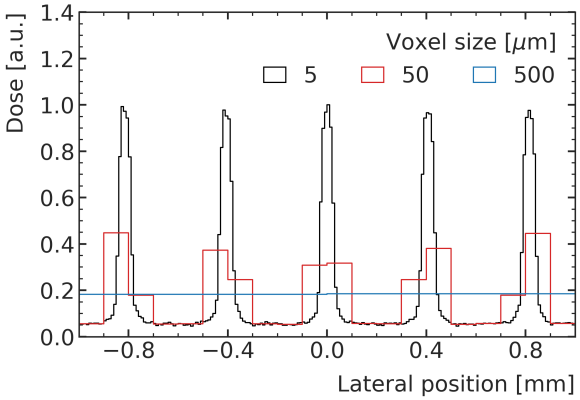


Figure 5.6: Comparison of dose per voxel for different voxel sizes.

30% of its original value. It can be clearly seen that a superposition approach, as discussed up until here, is not feasible.

5.2 Saving the day for superposition: macro voxels for microbeams

In fact, the micro-scale resolution of the individual microbeams is not always of interest during MRT treatment planning. While the exact knowledge about the dose gradient between peak and valley region may be of concern for certain applications, usually the treatment is prescribed as valley dose or peak dose to the target volume and the distribution of valley doses and peak doses, not those in the transition regions between them, is important for planning decisions. If a microscopic description of the dose distribution is not required, a more macroscopic approach is viable: the *macro voxel method*, which is presented in the following before discussing how it can be utilized to achieve a viable method of MRT dose estimation using the superposition of individual microbeams.

5.2.1 The macro voxel method

Using the macro voxel method, the average peak and valley doses in a certain region are saved instead of scoring the dose depositions on a micrometre scale. Those macroscopic descriptions of microscopic quantities can be used both for treatment planning and, in the scope of this

thesis, dose prediction. Figure (5.7) shows the dose distribution in a water phantom resulting from microbeams entering from the top, together with a white grid indicating exemplary macro voxels with 0.5 mm edge length.

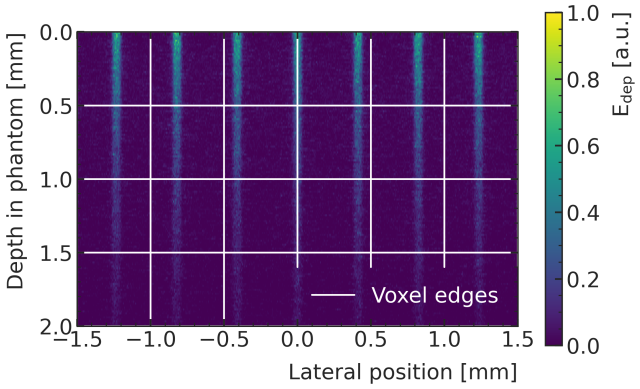


Figure 5.7: Visualization of microbeam energy deposition pattern together with macro voxel edges. Reproduced from [34].

Figure (5.8) schematically describes how the macro voxel method differs from the microscopic scoring approach. It shows energy deposition interactions as seen from the direction of the incident beam. Instead of scoring all energy depositions using a high-resolution voxel grid, only energy depositions in *peak regions* and *valley regions* are considered. For this, all energy depositions which occur at most 5 μm away from a theoretical peak position are saved as *peak dose* and all energy depositions which occur at most 50 μm away from a theoretical valley position are saved as *valley dose* for each of the macro voxels. All energy depositions between those regions are discarded. This method, in contrast to the microscopic scoring approach, can be used with nearly arbitrary macro voxel sizes, simply resulting in averaging the peak and valley doses over a larger region.

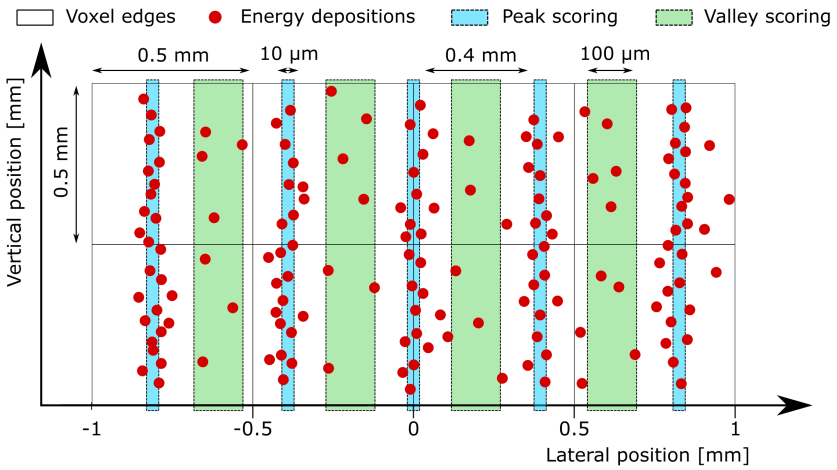


Figure 5.8: Macro voxel scoring schematic. Adapted from [34].

To illustrate in more detail how the method works, Figure (5.9) shows the application of the macro voxel scoring method on an exemplary lateral microbeam dose profile.
Within the scoring regions, the dose depositions are recorded and then averaged per macro voxel. Only energy depositions inside the macro voxels are counted. In the case of a macro

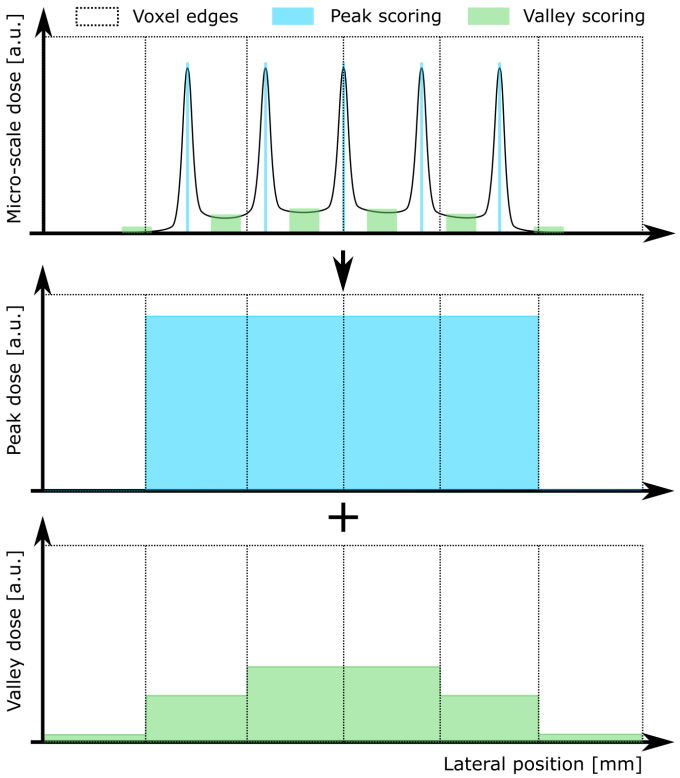


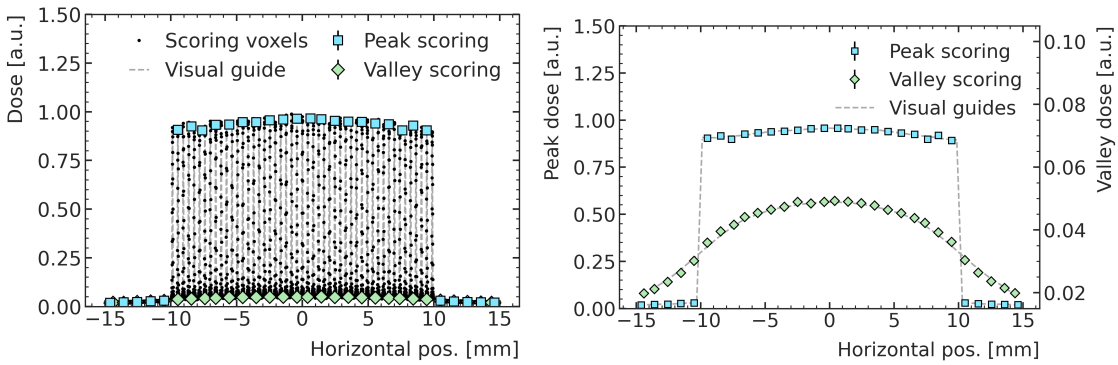
Figure 5.9: High resolution dose profile resulting from microbeams in a water phantom with respective peak and valley scoring regions marked in blue and green, respectively, together with the macro-voxel peak doses and valley doses. Respective doses are not to scale.

voxel boundary being positioned in a peak or valley scoring region, such as the central peak dose or the two valley doses in the most distant voxels from the centre, the energy depositions are split into the neighbouring voxels according to the respective volumetric fraction of the scoring region being located inside them.

The peak and valley doses in Figure (5.9) are not to scale, especially the shown valley dose is exaggerated. The curvature in the valley dose, however, generally is a usual feature in the MRT dose profile. This is caused by the proximity of more beams at the centre of the field resulting in more accumulated valley dose in the central region as compared to those at the edges of the field which are more distant from most peaks.

Figure (5.10) shows the macro voxel method applied to a more realistic MRT field. It shows the microscopic dose scoring on a 5 μm voxel grid in black together with the resulting peak and valley doses for a 0.5 mm macro voxel grid. Figure (5.10b) allows for a better visual inspection of the shape of the peak and valley doses by showing the two dose profiles on separate axes. In both Figure (5.10) and Figure (5.10b), grey lines are included to help guiding the eye. As indicated already in Figure (5.9)), the peak dose is limited to the extent of the microbeam field itself while the valley dose contains relevant entries also in the out-of-field region.

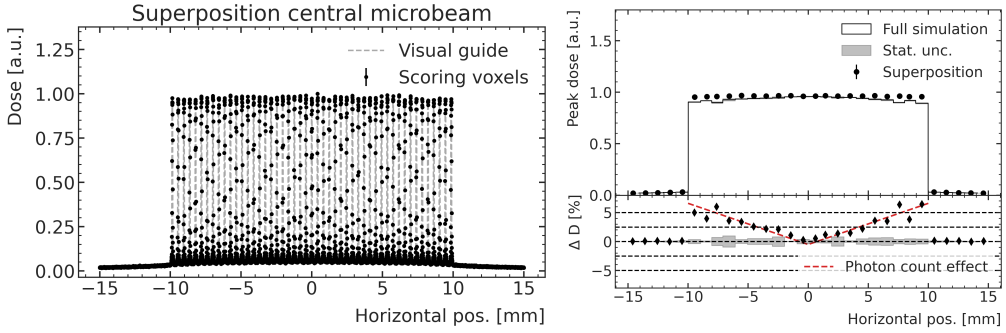
The peak dose profile also shows a roll-off effect to the sides. The reason for this is not the dose aggregation in peak areas in the centre of the field as it is for the valley doses. This can be seen when comparing the simulation result using the original phase space file (PSF) with a modified PSF in which all microbeams are replaced with the central one. In the case of the



(a) (b)

Figure 5.10: (a): Exemplary lateral microbeam dose profile using a 5 μm voxel grid (black, grey lines support visual identification of microbeams) and the respective peak and valley macro voxel fields in blue and green. (b): Peak and valley dose profiles on separate scales. Reproduced from [32].

aggregation of doses being the cause of the roll-off effect, it should still be visible. As can be seen in Figure (5.11a), this is not the case. When replacing the individual beams all with the central one, the peak dose roll-off disappears and the resulting dose profile instead exhibits a flat peak dose.



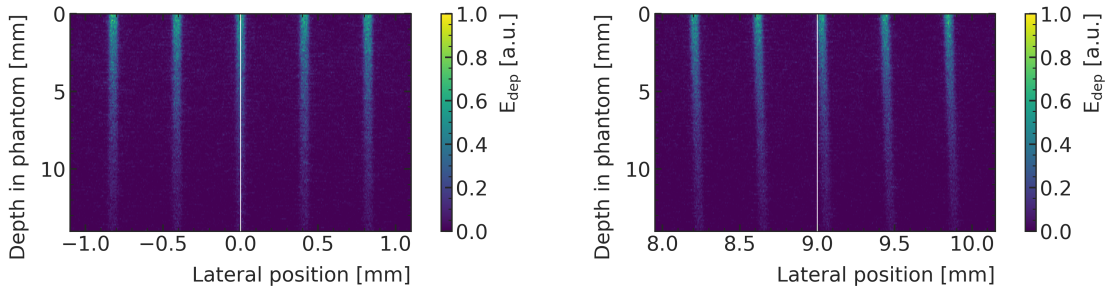
(a) (b)

Figure 5.11: (a): Comparison of the dose profile at the centre of the phantom using multiple versions of the central beam instead. (b): Peak dose comparison between the original PSF (full simulation, histogram) and a modified PSF (black dots) in which all microbeams are replaced with the central one. Adapted from [32].

Instead, the simulated microbeams further away from the centre in fact contain fewer photons and therefore lead to less dose deposition in the phantom. The decrease in photons is attributed to the original synchrotron x-ray intensity profile and is further increased from the small but existent divergence of the beam, being incident on a multi-slit collimator with plan parallel slits, resulting in more photons being cut from beamlets further from the centre. Figure (5.11b) shows a comparison between the peak dose profiles from the original and modified simulation method described before. The lower part of the plot shows the relative deviation between them. In red, the expected effect resulting from the decrease in photons further away from the centre of the field is highlighted. The agreement between the deviation and the photon count effect is found to be good enough to determine this to be the main cause of the roll-off effect.

5.2.2 Inclusion of microbeam divergence

One aspect of microbeams which is already included in the previously shown plots but has not been discussed is the inclusion of the small but existent remaining divergence of the beamlets. This is exemplarily shown in Figure (5.12). In the centre of the MRT field, the beamlets are close to perfectly orthogonal to the entrance surface into the phantom (top of plot). A vertical white line is inserted for better visual inspection. Figure (5.12b), in contrast, shows the microbeams at the edge of the MRT field. They are visibly not perpendicular to the entrance surface of the phantom, a result of the beam divergence. The deviation from the direction of the microbeams at the centre of the phantom can be clearly seen by comparing the microbeams the the white vertical line.

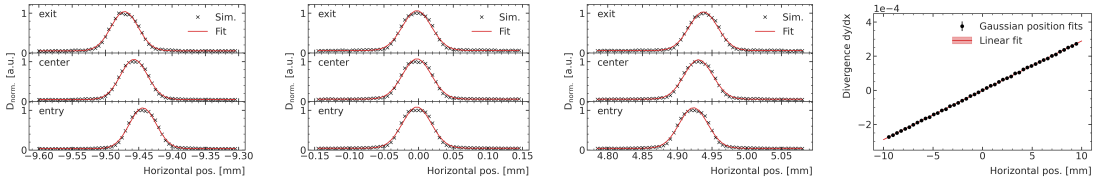


(a) (b)

Figure 5.12: Visualisation of the energy deposition caused by the microbeams in a water phantom in the central part of the field (a) and the edge of the field showing some beam divergence (b). White vertical lines are inserted to facilitate visual inspection.

The divergence is not large, and the aspect ratio of the shown plots should be noted, with only 2 mm on the x-axis and 14 cm on the y-axis. Nevertheless, using the discussed scoring method with pre-determined peak and valley locations, even a micrometer-scale deviation from those positions would lead to a significant difference in scored energy. This is especially true for the very thin peak energy regions. To counteract the effect of the divergence, the scoring regions themselves must *follow* it. This is achieved by assigning the scoring regions in dependence of the lateral position in the MRT field and the depth in a phantom.
To find the pre-computed peak and valley positions under consideration of the divergence, in a first step the lateral dose profile is simulated in a simple water phantom with a 5 μm resolution at several depths. A best estimate of the position of the maximum is computed by fitting a Gaussian function to each peak. The result for three depths, one at the entry, one at the centre and one at the exit of a simple water phantom can be seen in Figure (5.13a) (left edge of MRT field), Figure (5.13b) (centre of MRT field), and Figure (5.13c) (right edge of MRT field). Using the Gaussian functions to estimate the position of peak maxima, the peaks are found to be on average 411 μm apart at the centre of the water phantom in contrast to the nominal 400 μm spacing at the beam source. As all future phantoms will be centred at the same location, this is the effective peak-to-peak pitch for all following studies.
The microbeams are assumed not to be deflected, therefore the scoring regions are still assumed to be straight. For each lateral position, an inclination of the scoring volume against the orthogonal direction with respect to the phantom surface is computed by fitting a straight line to the peak locations previously found using the Gaussian fits. Figure (5.13d) shows the found divergence dy/dx. This can act as calibration curve to pre-compute the expected positions of peak and valley scoring regions throughout the phantom and for variable MRT

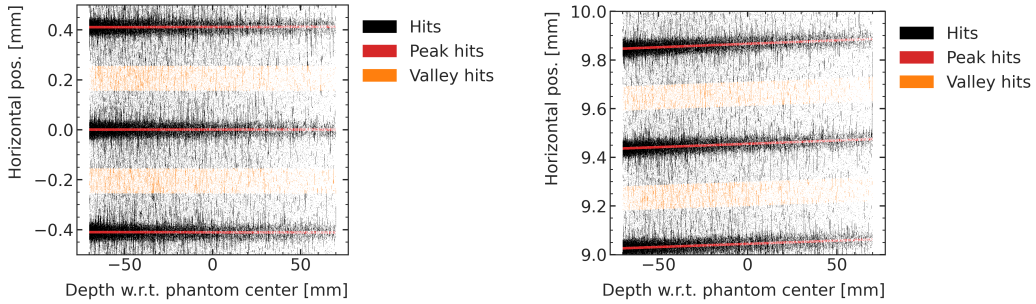
field sizes. For the computation of the correct peak and valley locations it is important to note that in this case, the centre of the used water phantom is used as reference origin, not e.g. the centre of the beam source.



(a) (b) (c) (d)

Figure 5.13: Microbeam locations and Gaussian fits at different positions ((a): -9.453 mm, (b): 0 mm, (c): 4.932 mm) and depths in the phantom.

Figure (5.14) exemplarily shows the divergent scoring regions at the centre (Figure (5.14a)) and the edge of the field (Figure (5.14b)) with energy depositions (black) counted towards the peak dose shown in red and counted towards the valley dose as orange. It can be seen that the divergent scoring regions follow the beam divergence. The plots also show the centre of the used phantom as position 0 as it acts as the origin for the calculations.



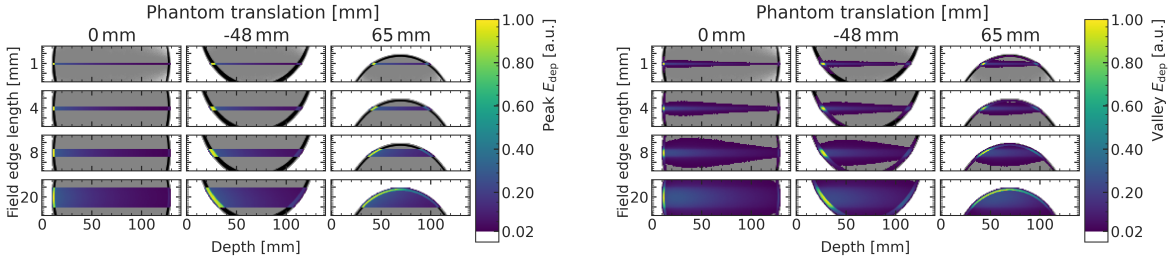
(a) (b)

Figure 5.14: Hit locations in the centre (a) and further out (b) together with indications for peak and valley areas.

While this method is used throughout this thesis, it also comes at the cost of having to be repeated for all MRT fields which are used in the future due to different divergences of different beam configurations.

5.2.3 Agreement between macro and micro voxel scoring

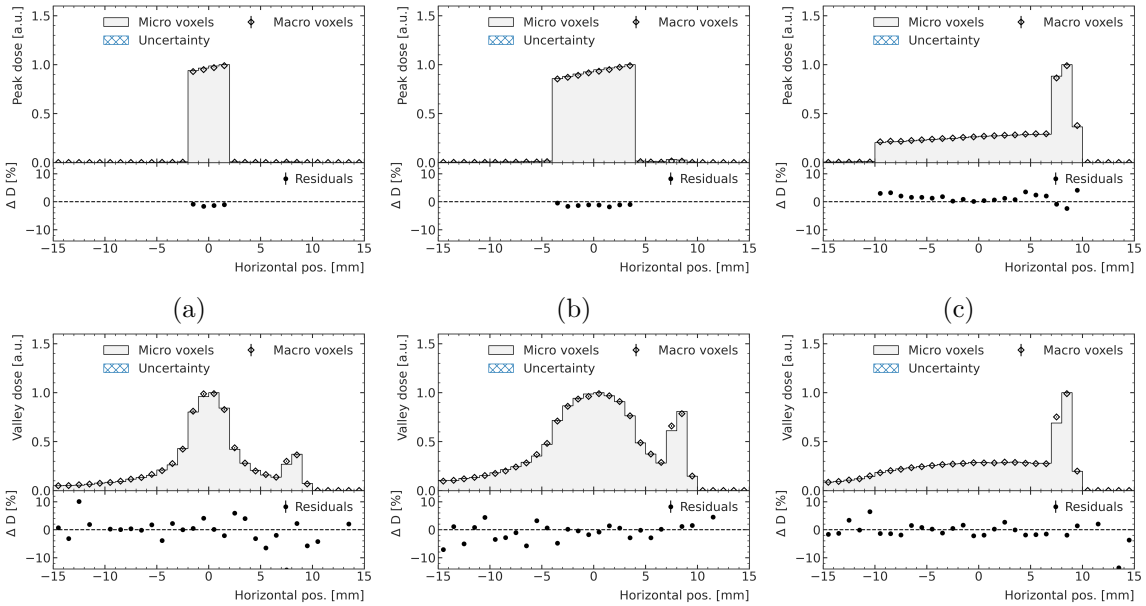
This subsection investigates the agreement of peak and valley doses obtained using the previously developed macro voxel method under influence of beam divergence with the respective doses obtained using micrometre-scale scoring voxels. For this, MRT fields of different sizes incident on the simple head phantom used in the previous sections, are simulated. Figure (5.15) shows resulting peak and valley energy depositions using the macro voxel method and field sizes of 1x1, 4x4, 8x8, and 20x20 mm^2. As before, the sharp field edges of the peak doses can be observed in contrast to the wider spread of the valley doses. Energy depositions lower than 2% of the respective maximum are not shown.
In addition to the macro voxels, the energy depositions are also scored using 2D arrays of 5 μm in different depths of the phantom. Using those, the peaks are located again using Gaussian



(a) (b)

Figure 5.15: Exemplary peak (a) and valley (b) doses in a simple head phantom for different MRT field sizes.

fits. Subsequently, the peak doses are determined by averaging over the two voxels next to the found peak location (therefore also 10 μm) and over 20 voxels (therefore also 100 μm) around the valley locations which are assumed to be right between two peaks. Figure (5.16) shows exemplary lateral dose profiles scored near the entrance into the head phantom for the 4x4, 8x8, and 20x20 mm^2 MRT fields. All valley doses (Figure (5.16c, Figure (5.16d, Figure (5.16e) and also the peak dose profile of the largest field (Figure (5.16c) exhibit an increase in dose where the profile reaches the skull.



(a) (b) (c)

(d) (e) (f)

Figure 5.16: Agreement of peak doses (a-c) and valley doses (d-f) scored using macro voxels and micro voxels, respectively, for a 4x4 (a,d), 8x8 (b, e) and 20x20 mm^2 (c,f) MRT field.

Deviations are mostly within 1-3% with few larger deviations. Those are attributed to slight misalignments between the Gaussian fit to detect the peak maximum and the pre-computation method used before. Overall, the results are found to be in acceptable agreement to proceed with this scoring method. Using this method, MRT fields of large sizes can be composed.

5.2.4 Microbeam superposition using the macro voxel method

This subsection demonstrated how the previously developed macro voxel method can be used to use a superposition of individual microbeam dose depositions to build the entire MRT field dose profile. In a first step, it is required in a first step to acquire the macro voxel scoring for a single microbeam. Figure (5.17) schematically shows how the macro voxel method is applied to a single, centred microbeam.

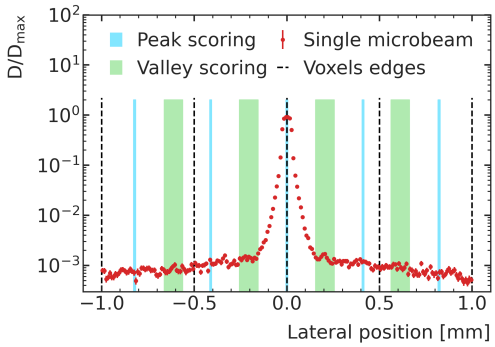
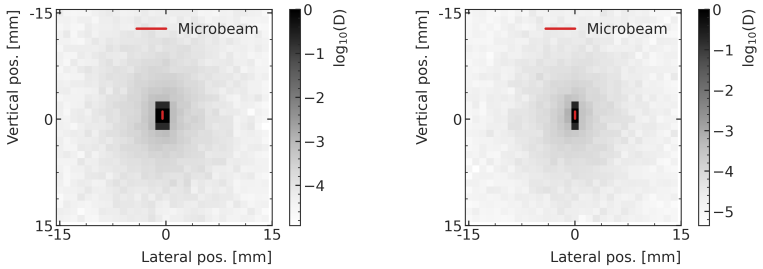


Figure 5.17: Schematic of the macro voxel method applied to a single microbeam. Reproduced from [32].

To allow for the long-ranging effect especially of the valley doses, a scoring field of view of 30 mm is used for a single microbeam of 0.5 mm height. Depending on the exact location of the microbeam with respect to the macro voxel edges, the peak dose maximum covers either one or two macro voxels. Figure (5.18) shows exemplary 2D peak dose predictions for a centred and a non-centred single microbeam. The dose is shown on a log scale due to the rapid decrease of peak doses to the sides. Valley doses are not shown at this point but look very similar to the shown profiles when being shown on a linear scale.



(a) (b)

Figure 5.18: Resulting 2D peak dose profile at the centre of a water phantom for a single, centred microbeam (a) and a single, non-centred microbeam (b). Reproduced from [32].

By predicting the peak and valley doses for single microbeams at their respective location, the total peak and valley dose of an MRT field can be superimposed. The concept is schematically shown in Figure (5.19).
Exemplary superimposed peak and valley dose predictions, 30 mm around the centre of the field, together with the resulting superimposed PVDRs, are shown in Figure (5.20) for MRT field sizes of 4x4, 12x4 and 20x20 mm^2.
While the peak doses are observed to minimally increase with increasing field size, they do

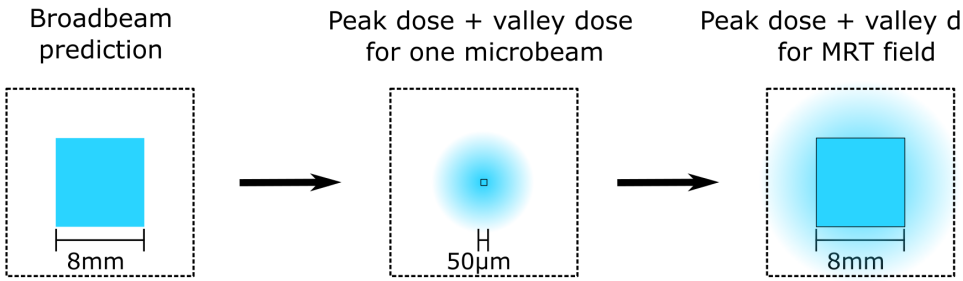
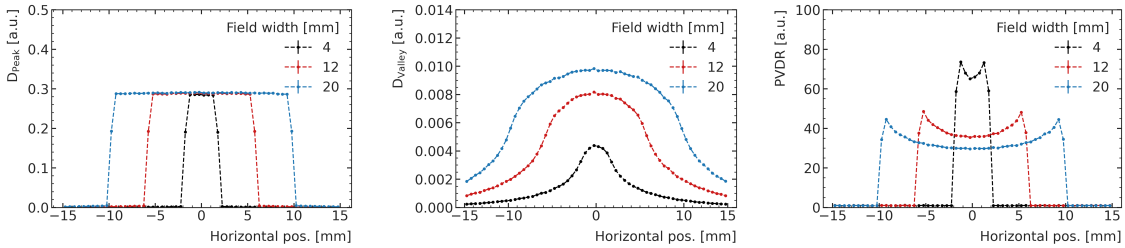


Figure 5.19: Schematic of the microbeam superposition method using the macro voxel method.



(a) (b) (c)

Figure 5.20: Peak (a) and valley (b) doses with the respective PVDR (c) for a field height of 4x4 (black), 12x12 (red) and 20x20 mm^2 (blue).

not exhibit the previously discussed roll-off effect. This is because for simplicity, only the central microbeam is predicted and shifted to the locations of the other microbeams at this stage. As a result, the decrease due to the photon count effect and also the beam divergence are not used. The divergence is found not to significantly contribute to the dose distributions so that not considering it makes a negligible impact. The different contributions of peaks closer to the edge of the original PSF could be considered in a post-processing step in the future.

The superposition with this technique only works when two requirements are met: (1) the predicted microbeams have to be located in the respective peak positions of the MRT field (e.g. 0, ±411 μm, ±822 μm, ...) and (2) the macro voxel edges of all microbeam predictions need to coincide. Although this sounds trivial at first, it is not trivial to achieve. Figure (5.21) shows the scenario of two individual microbeams close to each other incident on an exemplary phantom (grey) being superimposed. Because the macro voxel edges have to coincide, the second microbeam b) cannot be centred in the prediction volume. The microbeam would have to be predicted at position 411 μm of the first macro voxel to the right from the centre line to match both the macro voxel edges and the peak pitch.

Moving further to the side, the next peak would have to be predicted at position 822 μm as shown in Figure (5.22).

This position already falls in the second macro voxel to the side. To account for long-ranging dose effects especially in the valleys, however, it is desired to maintain the approximately 15 mm prediction field to each direction from the microbeam. To achieve this, given a fixed prediction volume, it is more suitable to neglect the voxel furthest to the left and instead shift the prediction volume one voxel to the right as shown in Figure (5.22) b). This results

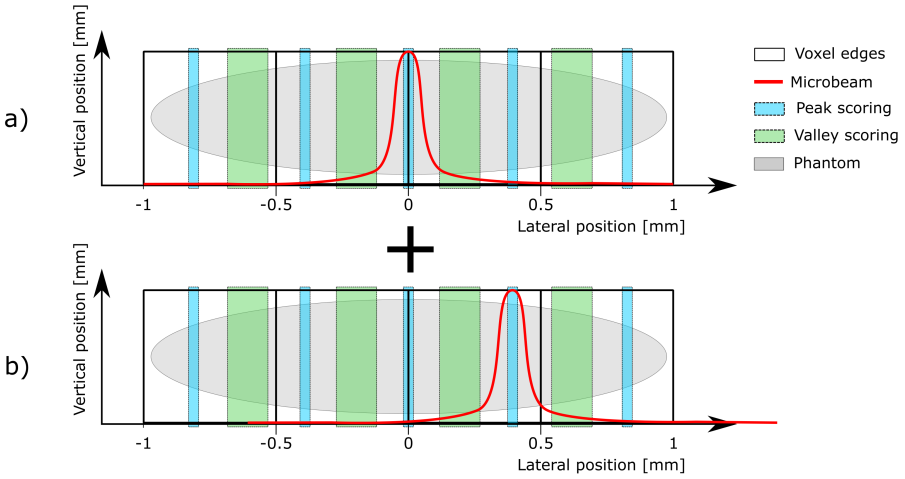


Figure 5.21: Schematic of the superposition of two microbeams a) and b).

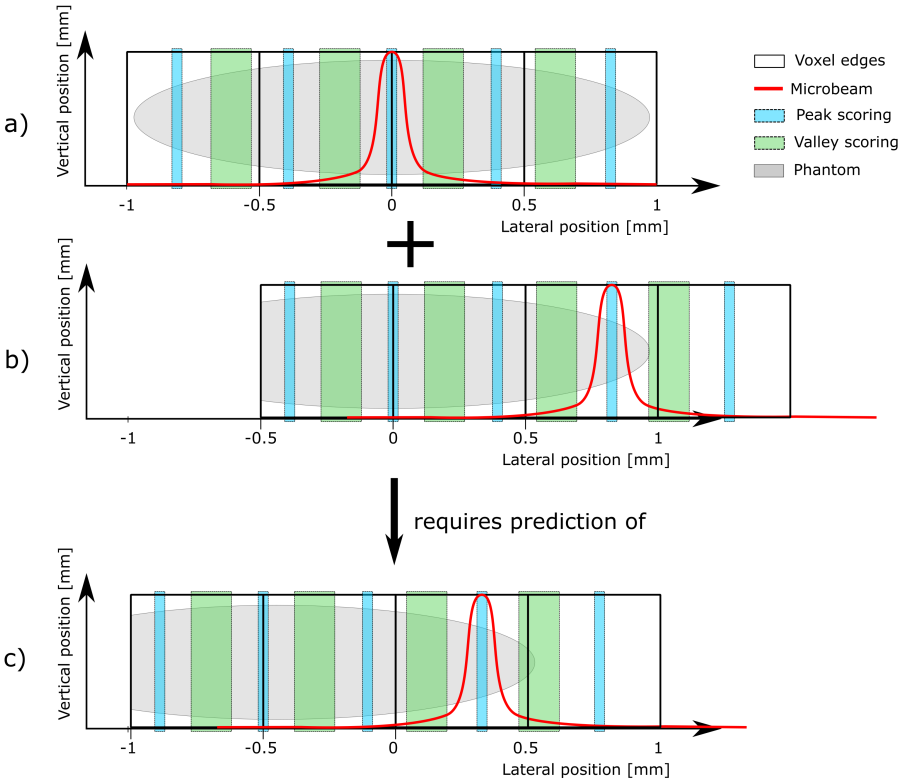


Figure 5.22: Schematic of the superposition of two more distant microbeams a) and b) together with the required prediction c) to generate the distribution b).

in a prediction volume as shown in Figure (5.22) c): the microbeam would be predicted at 322 μm first with the target phantom (grey) being shifted 500 μm to the left. The resulting dose deposition matrix is shifted by 500 μm to one side to achieve the total translation of 822 μm, before being superimposed with the result of prediction a). Following this example,

larger MRT fields can be constructed using the dose fields from single microbeams at a location

$$k = (N \cdot 411\,\mu\text{m}) \quad \text{modulo} \quad 500\,\mu\text{m} \tag{5.1}$$

and applying the respective shift in steps of 500 μm to the target phantom.

5.3 Predicting individual microbeams with machine learning

This section investigates the suitability of the previously developed microbeam superposition approach utilizing the macro voxel method for ML prediction.

5.3.1 Simulation setup and datasets

The target phantom in this study is a simple 3x3x3 cm^3 water block. This size is used because of the current focus of MRT research at the Imaging and Medical Beamline (IMBL) on preclinical rodent studies. This study, therefore, presents a first investigation into the suitability of the macro voxel method for dose predictions on the spatial scale of future rat head phantoms which are of similar size. Figure (5.23) shows a schematic of the simulation setup. A single microbeam enters the water phantom from the left. To account for the long-ranging effect on the valley doses, the scoring volume is chosen to be 3x3x3 cm^3 as well in accordance with the finding of the previous section. To limit the complexity of this first study on the direct prediction of microbeams, a macro voxel size of 1 mm^3 is chosen, resulting in 30x30x30 (27,000 voxels compared to 216,000 with 60x60x60 voxels) simulated and predicted voxels per data sample. To generate different data samples, both the microbeam and the phantom are translated relative to each other, and the scoring volume as discussed in the previous section. The microbeam position is set to discrete values in steps of 500 μm while the phantom is translated continuously. In addition, several microbeam locations derived from Equation (5.1) are simulated: 55 μm (N=5), 144 μm (N=4), 233 μm (N=3), 288 μm (N=8), 322 μm (N=2), 377 μm (N=7), 411 μm (N=1), 466 μm (N=6).

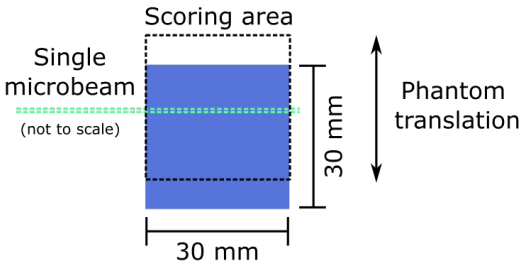
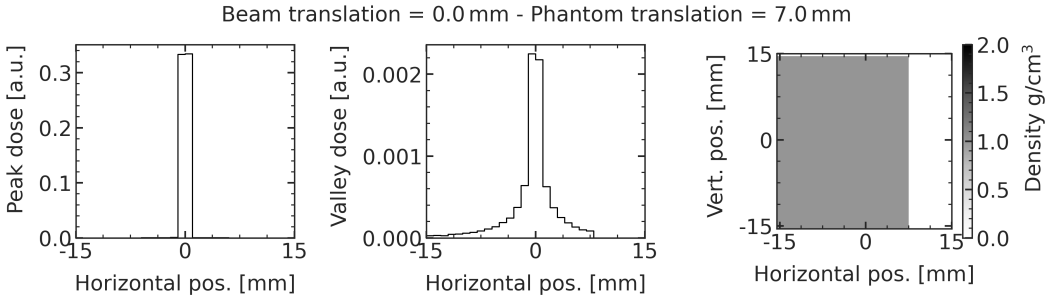


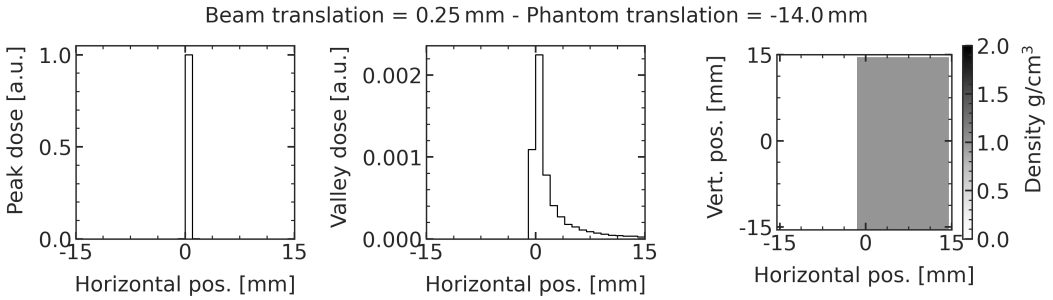
Figure 5.23: Simulation setup for creating a dataset of dose depositions from single microbeams using the macro voxel method showing the incident beam (green), a 3x3x3 cm^3 water phantom (blue) and the scoring volume (black). Reproduced from [32].

Two exemplary data samples resulting from this are shown in Figure (5.24). Figure (5.24a) shows the lateral peak and valley dose profile at the centre of the water cube, obtained using the macro voxel method for a centred beam and a phantom translation of 7 mm to the side. This is indicated by the grey box in the right side plot. On the linear scale, the peak dose can be seen to be invisibly small already in the voxels next to the two central ones. The valley dose profile decreases less quickly in comparison but steeply nevertheless. As soon as the phantom ends, the dose drops to zero. Figure (5.24b) shows the resulting lateral dose

profiles at the centre of the cube for a microbeam translation of 0.25 mm and a phantom translation of -14 mm. As the peak is located in one macro voxel and not on the boundary as in the previous example, the peak dose profile now only covers one voxel before dropping to close to zero.



(a)



(b)

Figure 5.24: Lateral peak and valley dose profile for two exemplary data samples comprising a microbeam and phantom translation of (0 mm | 7 mm) (a) and (0.25 mm | -14 mm), respectively. Reproduced from [32].

The parameter space of the simulated data samples, split into training (grey), validation (black) and test data (red), is shown in Figure (5.25). Samples which exhibit a phantom translation of $[-11, -9]$ mm, $[-4.5, -2.5]$ mm, $[2.5, 4.5]$ mm, and $[9, 11]$ mm are excluded from training and used as validation data set. All samples simulated using the microbeam translation of multiples of 411 μm are used as test data because those resemble most closely the realistic case in application. Such samples, however, are not created for all phantom translations.

5.3.2 Adaption of the ML model

In Section (3), the water-only energy deposition was passed to the ML model to allow the model to be conditioned on the beam position and facilitate learning by adding information about the energy deposition in water. Similar to this method not being suitable for the prediction of proton minibeams, which was found in Section (4), this is not a viable way in this study. First, due to many different microbeam locations, there would be again a need for many water-only simulations. Additionally, the whole simulation only contains a water phantom. Therefore, this type of additional information would make the ML prediction obsolete. Instead, the position of the simulated microbeam with respect to the centre of the prediction volume is encoded using a 3D *distance matrix* of the same shape as the prediction volume, 30x30x30. The layer shapes of the U-Net are modified accordingly to allow for this input and output matrix shape instead of the previously used 140x16x16 voxels (35,840 in

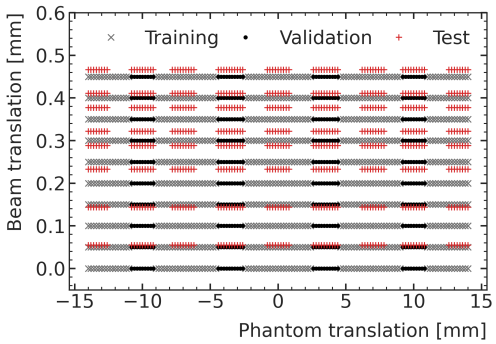
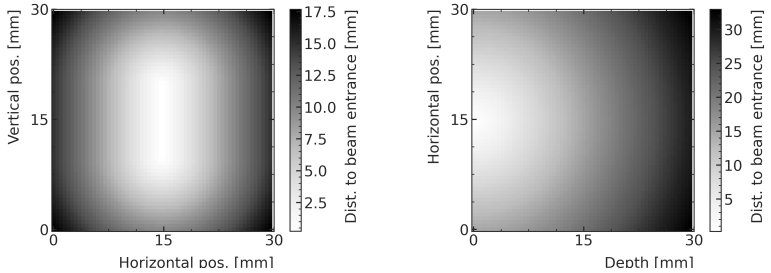


Figure 5.25: Distribution of data samples on the training (grey), validation (black) and test (red) datasets in dependence of the beam and the phantom translation. Reproduced from [32].

total). Two 2D slices of an exemplary distance matrix are shown in Figure (5.26). Each voxel contains the minimum distance of its centre to the projected area of the microbeam entering the phantom.



(a) (b)

Figure 5.26: Slices of a distance matrix in the yz (a) and xy (b) plane.

Figure (5.27) shows a zoom-in on how the matrix changes when the microbeam is moved to the side. The greyscale coding is not to scale and only serves as visualization of the shift of the matrix values together with the microbeam location. In initial tests it was found that the simultaneous prediction of peak and valley doses are not successful. Instead, two ML models are trained, one for the prediction of the peak doses, one for the valley doses. Figure (5.28) shows a schematic of the resulting ML model adapted from the previous one. The two matrices being input into the model are concatenated to a single 30x30x30x2 input matrix before being passed to the network. The output are a 30x30x30 dose matrix for either the valley or the peak regions.

5.3.3 Search for optimal ML models

In a first step, both ML models (for the peak and the valley dose prediction) are trained using the same configuration that was found to work best in Section (4): each convolutional layer exhibits 64 convolutional filters, training is performed using the Adam optimizer with the mean-absolute error (MAE) loss function, a batch size of 32, and a learning rate is $1 \cdot 10^{-3}$. For the valley dose prediction model, no significant improvement is found by varying the batch size, learning rate or number of filters. The peak dose prediction model requires more

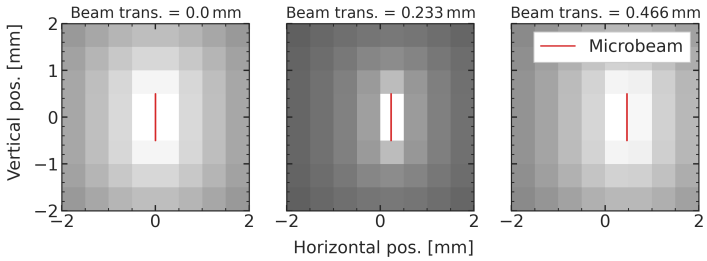


Figure 5.27: Change in microbeam position (red) and resulting change in the distance matrix (not to scale).

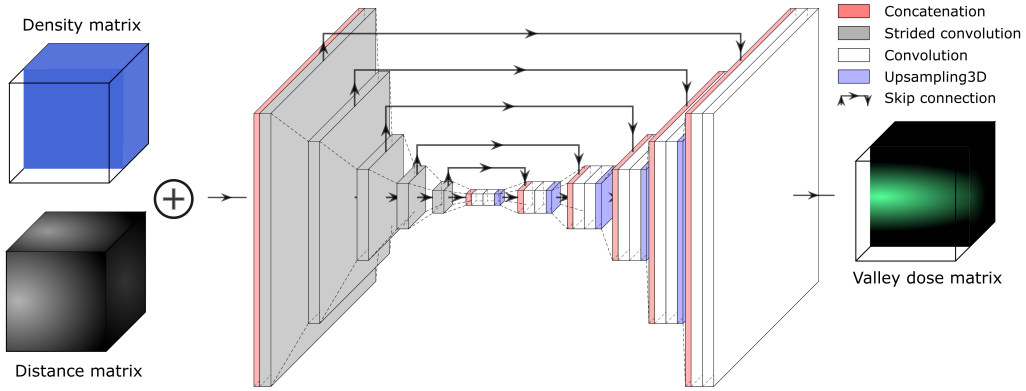


Figure 5.28: Schematic of the adapted ML model taking two input matrices (density and distance) and creating one output matrix (in this case valley dose). All matrices are shown for visualization only and are not to scale with respect to their colour coding.

attention. Due to the large gradients, the ML model does not learn to predict the very low doses left and right of the central voxel. This effect is schematically shown in Figure (5.29). While this is not a large problem due to the doses being very small in comparison to the central peak, it may add up to significant deviations especially for large MRT fields.
For this reason, an alternative loss function for the peak dose prediction model is proposed. One function which transforms data from several orders of magnitude to a more similar scale is the logarithm. A simple approach would be using the logarithm of dose values inside the MAE loss function in the following form:

$$MAE_{\log} = \frac{1}{N}\sum_{i}^{N} |\log_{10}(y_i) - \log_{10}(\hat{y}_i)| \tag{5.2}$$

N is the number of training samples, y_i the peak dose prediction for training sample i and $\hat{y}_i$ the MC simulation result for sample i. Figure (5.30) shows an exemplary peak dose prediction using a model trained using that loss function. While the lateral profile, shown on a log-scale in Figure(5.30a) looks promising, the dose actually deviates significantly which can be seen both in the lower part of Figure(5.30a) showing deviations of more than 10% in the centre of the field and also in the depth dose curve in Figure(5.30b) showing very large discrepancies between the ML prediction and the MC simulation. The log-loss function shown

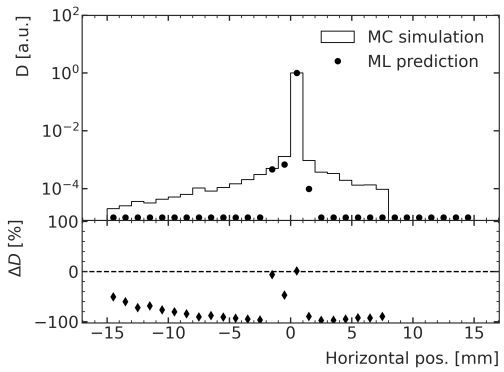
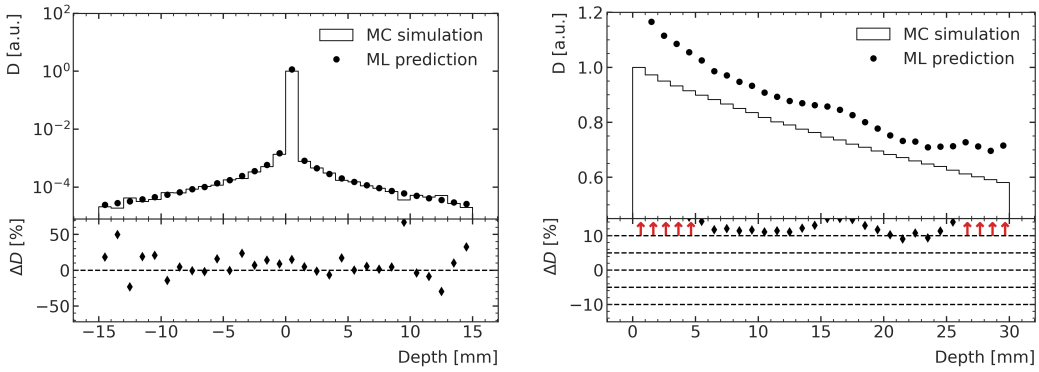


Figure 5.29: Comparison of the lateral valley dose profile (a) and depth valley dose curve (b) in the centre of the phantom as predicted using the trained ML model with the MC simulation for a data sample with a phantom translation 7 mm and a microbeam translation 322 μm. Reproduced from [32].

in Equation (5.2) is found not to be suitable for training dose prediction ML models because it does not capture the importance of accurate dose predictions on a linear scale.



(a) (b)

Figure 5.30: Peak dose prediction of an ML model trained with the $MAE_{\log}$ loss function showing the lateral dose profile (a) and the depth dose curve (b) in comparison to the respective MC simulation.

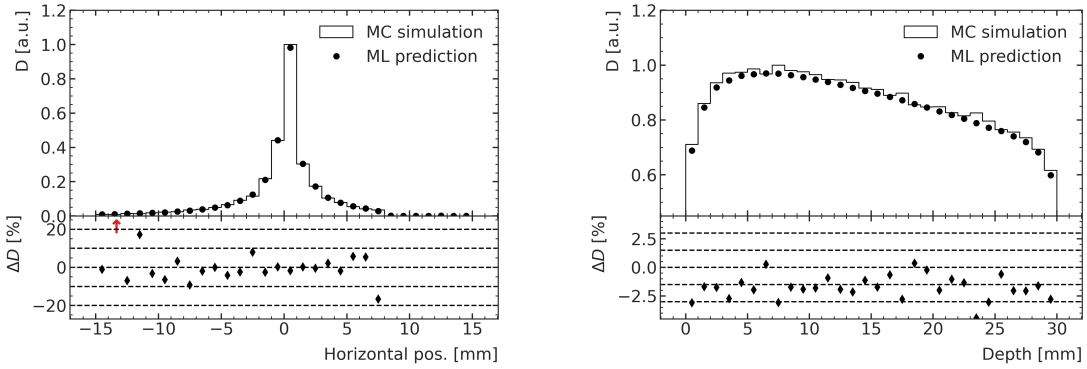
Following this, a mixed loss is proposed: it used the log-loss in Equation (5.2) and adds a linear loss to it. The weight between the two contributing terms can be adjusted using a parameter α:

$$MAE_{\text{mixed}} = \frac{1}{N}\sum_{i}^{N}\left[\frac{1}{\alpha+1}\cdot|\log_{10}(y_i)-\log_{10}(\hat{y}_i)|+\frac{\alpha}{\alpha+1}|y_i-\hat{y}_i|\right] \tag{5.3}$$

with N is the number of training samples, y_i the peak dose prediction for training sample i, $\hat{y}_i$ the MC simulation result for sample i. The weighting parameter α is optimized during the search for an optimal model. A value of $\alpha = 10$ is found to be optimal. Other model settings are not found to significantly impact the result. Therefore, they are kept at 64 convolutional filters, using the Adam optimizer with a batch size of 32, and a learning rate is $1 \cdot 10^{-3}$.

5.3.4 Results

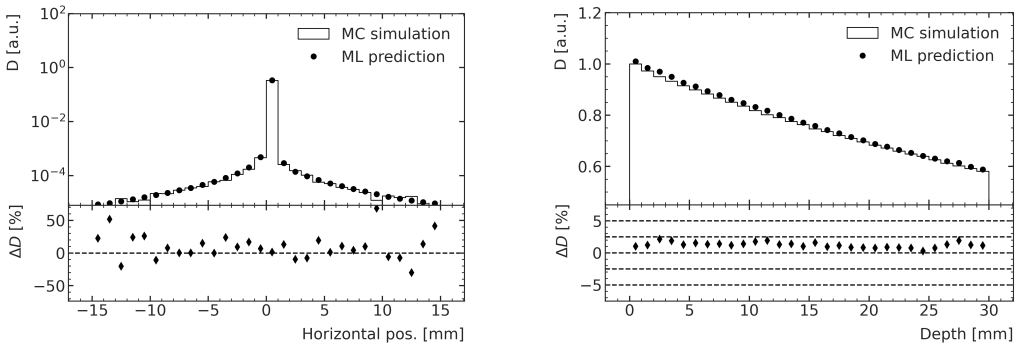
Figure (5.31) shows the lateral valley and peak dose profile together with the depth dose curve at the centre of the phantom, predicted by the ML models for an exemplary test data sample (phantom translation 7 mm and beam translation 322 μm). Other predictions perform very similar to the shown ones.
While for the valley dose prediction, as shown in Figure (5.31), the agreement is found to be within a few percent, a systematic underestimation by roughly 1.5% of the valley dose prediction can be seen in the depth dose curve in Figure (5.31b). Such a behaviour is seen for many beam configurations both over- and underestimating the depth dose by a few percent. While a deviation of only 1.5% is generally a very good agreement of the ML prediction with the MC simulation, it poses a big problem for the superposition use case. Because the deviations do not cancel each other out upon superposition, they are potentially added up resulting in significant deviations of the final superimposed dose prediction.



(a) (b)

Figure 5.31: Lateral valley dose profile (a) and depth valley dose curve (b) in the centre of the phantom as predicted using the trained ML model for a test data sample with a phantom translation 7 mm and a microbeam translation 322 μm. Reproduced from [32].

An exemplary peak dose prediction is shown in Figure (5.32) for a test data sample with a microbeam translation of 0.244 mm and no phantom translation (0 mm). As for the valley doses, the agreements are generally found to be adequate with deviations of at most 2.5% in the centre of the field as shown in the depth-dose curve in Figure (5.32b). Several predictions however, show a small but systematic deviation in the peak dose prediction, similar to the findings for the valley dose prediction model. As in the case of the valley dose prediction model, the deviations in the relevant centre of the field are relatively small but potentially add up when creating large, superimposed fields.
These findings severely limit the applicability of the ML approach for MRT dose prediction by superposition of individual microbeams. In addition, the prediction speed is significantly reduced by the need to predict many individual fields for the superposition. Assuming a final field size of 20x20 mm^2, the proposed model, using a single microbeam of 0.5 mm height, requires 2000 individual predictions. With the observed prediction time of approximately 0.1 seconds, this accumulates to 3.3 minutes per MRT field. Using batch predictions, this could further be reduced to a minimum of about 6.25 seconds (32 simultaneous predictions on one graphics processing unit (GPU)). At that stage, the model would most likely be slowed down by the capability to provide the phantom material matrices quickly enough as for future



(a) (b)

Figure 5.32: Exemplary peek dose predictions using the mixed loss functions for a beam translation of 0.244 mm and a phantom translation of 0 mm shown as lateral profile at the centre of the phantom (a) and the depth-dose curve in the line of the maximum dose deposition (b).

application scenarios, they would have to be produced from Computer Tomography (CT) images which is more time consuming than using pre-processed matrices like in this study. In addition, while prediction times on the order of 10 seconds are significantly faster than existing methods, it is relatively slow for the initial purpose of the ML model to be used in treatment optimization scenarios, potentially requiring many dose prediction evaluations.

The discussed findings are expected to get more severe when including more complex phantoms of actual patients in the future. For this reason and with regard to the previously discussed aspects, a microbeam superposition approach is not further explored in the course of this thesis. This does not mean, however, that the macro voxel approach is not found to be a valuable tool for MRT dose prediction, as will be shown in the next section.

5.4 ML peak and valley dose prediction with macro voxels and rat phantoms

After finding that the superposition of microbeams is found not to be a feasible method to predict the dose distribution following the irradiation with an MRT field, a more direct approach is investigated in this section. In the previous section, separate peak and valley dose distributions were scored in the MC simulation using the macro voxel method. During its introduction, rather as a by-product, it was shown that the method can also be used to score the entire MRT field by including the divergence of the microbeams into the method. This section investigates the prediction of the entire peak and valley dose distributions for one incident MRT field instead of pursuing a superposition approach. This concept is schematically shown in Figure (5.33). By predicting the peak and valley distribution, this approach is very similar to the originally investigated broad beam prediction model in Section (3).

To be able to predict the peak and valley doses for different MRT fields, an ML model then needs to be trained on different field sizes or even shapes. This presents a significant extension of the previously developed broad beam dose prediction model. This section investigates the capability of the developed ML model to predict the peak and valley doses in a simplified rodent skull phantom following the irradiation with a quadratic MRT field of variable size. A rodent skull phantom is chosen as base for the creation of the simplified model because in the current state of preclinical research at the IMBL, rodent brain tumours are the main target.

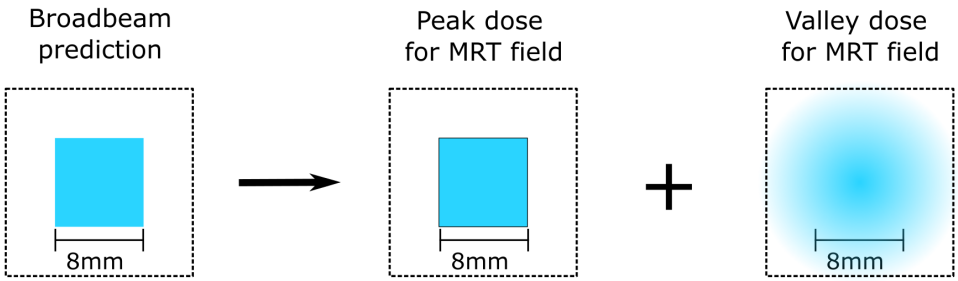
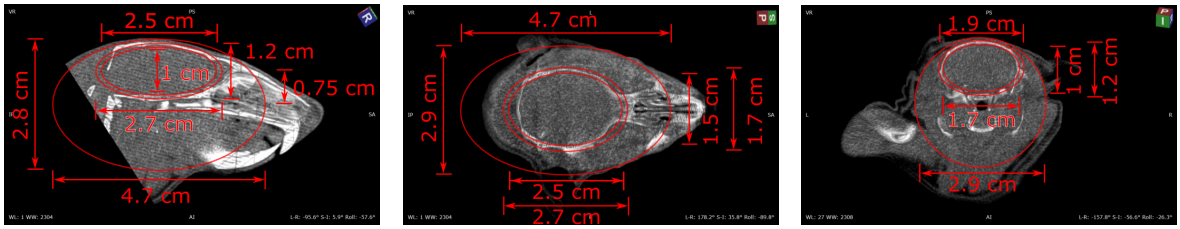


Figure 5.33: Schematic of MRT dose prediction using the macro voxel method to separate the dose into peak and valley dose distribution fields.

5.4.1 Development of a simplified rat head phantom

A rodent skull phantom, similar to the simple paediatric head phantom introduced in Section (3), is developed to serve as simulation model in this study. To adapt the model a bit more closely to a realistic treatment scenario, a CT image of a rat skull is used to derive some measures. Three slice views of the used CT image together with taken measures and insertions of ellipsoids being used for the simplified model construction are shown in Figure (5.34c).



(a) (b) (c)

Figure 5.34: CT scan from side (a) top (b) and back (c) with measures and inserted ellipsoidal digital model (red).

Figure (5.35) shows the resulting simplified rodent skull phantom together with an exemplary MRT field incident from the left on the top of the head of the phantom, in agreement with the current treatment protocols at the IMBL. The dose in the phantom is scored in a 30x30x30 voxel grid, each 1 mm^3 in size, resulting in a scoring volume of 30x30x30 mm^3.

5.4.2 Dataset

In this study, different simulation samples are created by varying three different parameters: the horizontal translation of the phantom Δy, the vertical translation of the phantom Δz and the beam size D_{beam}. The size of the simple rodent head phantom is not modified in this study. To visualize the resulting variation in the data, Figure (5.36) shows schematics of three exemplary data samples.
Data samples are generated by performing random sampling in the ranges $\Delta y \in [-9, 9]$ (rounded to 3 digits), $\Delta z \in [-6, 6]$ (rounded to 3 digits) and $D_{\mathrm{beam}} \in [3, 19]$ (rounded to integers). Due to long simulations times, a rather small dataset comprising 344 samples is creates spanning the following parameter space shown in Figure (5.37).
The data samples are split randomly into training (60%, 206 samples), validation (20%, 69

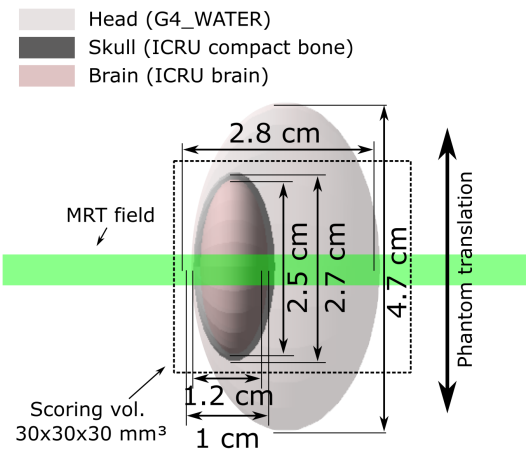
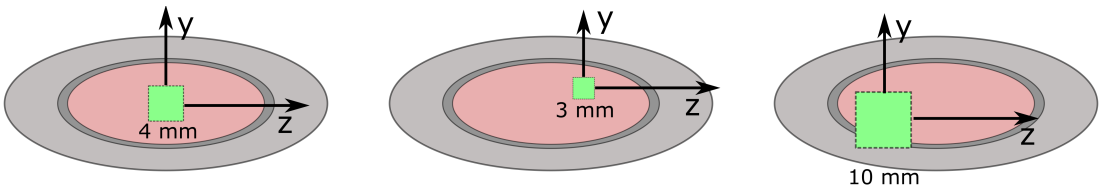
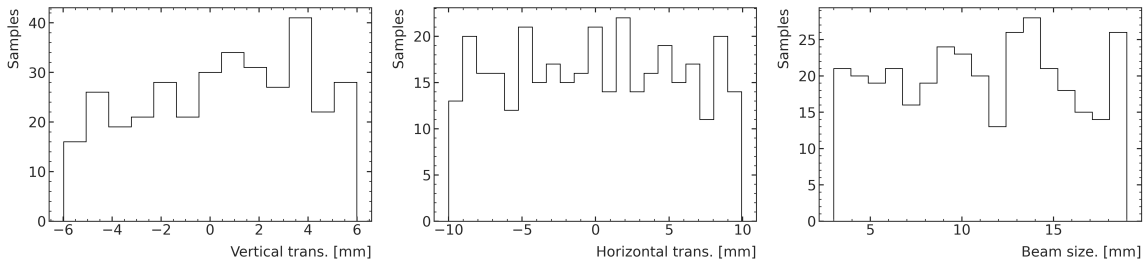


Figure 5.35: Schematic of the rat head simulation together with an exemplary incident MRT field and the scoring volume.



(a) (b) (c)

Figure 5.36: Three schematic visualizations of different data samples with varying lateral and vertical phantom translation as well as different MRT field (green) sizes.



(a) (b) (c)

Figure 5.37: Horizontal (a) and vertical (b) phantom translation together with the quadratic field sizes (c) comprising the datasets.

samples) and test dataset (20%, 69 samples). The resulting distribution of data samples in the parameter space is shown in Figure (5.38). In addition to those data samples, an additional test dataset is generated. This second test dataset comprises simulated samples with MRT field sizes of 3.5 mm, 4.5 mm, 5.5 mm, ... 9.5 mm. Those beam sizes are not contained in the whole training process and are designed to investigate more closely the generalization capability of the network after being trained with only integer-sized MRT fields.

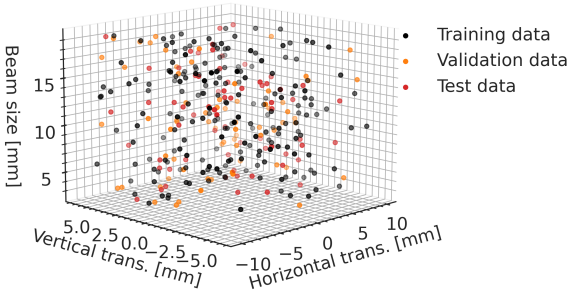
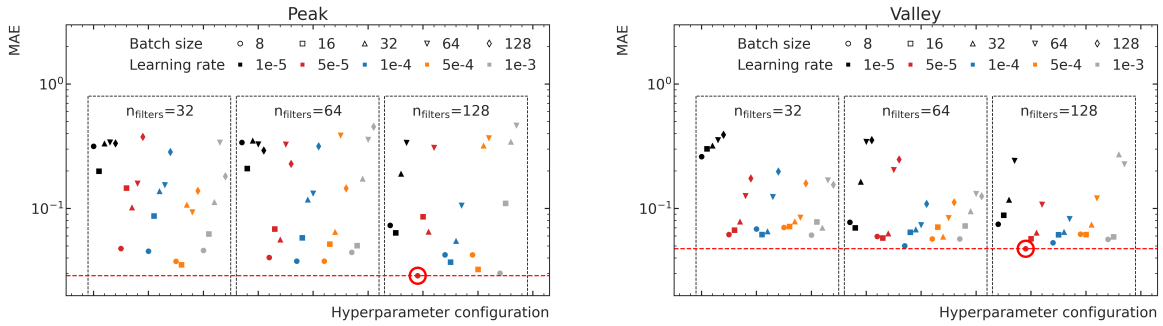


Figure 5.38: 3D scatter plot showing the distribution of samples in the parameter space. The colour shows the three datasets for training, validation, and test.

5.4.3 Grid search for optimal model hyperparameters

Like in the previous section, two separate models are trained for the peak and valley dose prediction. The shape and size of the MRT field is passed to the model similar to the broadbeam scenario using a water-only simulation using an MRT field of the respectively same size.

In a first, preliminary search for suitable hyperparameter settings, significant dependencies of the models on the chosen set of hyperparameters, especially the learning rate, were found. Therefore, a grid search is performed to find optimal model setting, individually for the peak and valley dose prediction model. The explored grid and the results are shown in Figure (5.39).



(a) (b)

Figure 5.39: Validation dataset performance of the peak (a) and valley (b) dose prediction ML models, trained using different sets of hyperparameters. The respective best-performing models are highlighted using a red circle.

For the search, convolutional filter sizes of [32, 64, 128] are tested with batch sizes of [8, 16, 32, 64, 128] and learning rates of $[1 \cdot 10^{-5}, 5 \cdot 10^{-5}, 1 \cdot 10^{-4}, 5 \cdot 10^{-4}, 1 \cdot 10^{-3}]$. For all trainings, the Adam optimizer is used together with the MAE as loss function. Although the differences in model performances look very large, it should be noted that the MAE is shown on a logarithmic scale. Both the best peak (Figure (5.39a)) and valley (Figure (5.39b)) dose prediction models are found to be the one using 128 convolutional filters per layer (right box), being trained with a batch size of 8 (circle marker) and a learning rate of $5 \cdot 10^{-5}$ (red colour of the marker).

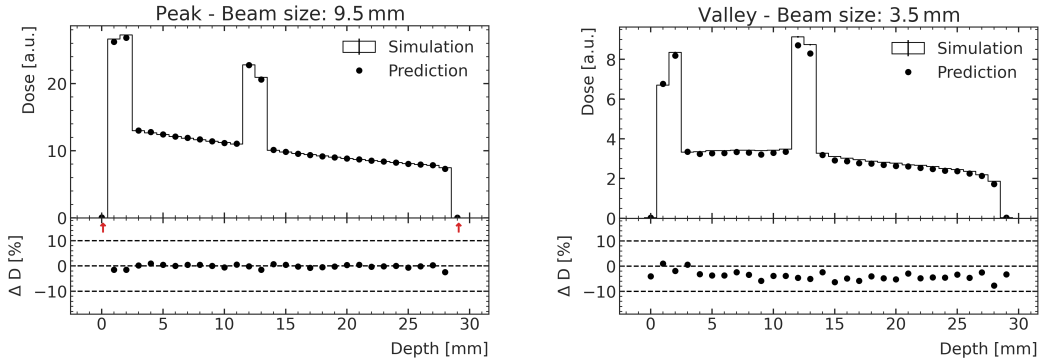
5.4.4 Generalization assessment

After the model is fully trained, it is executed for all training, validation and test data samples. To compare the performance on the different datasets and thereby assess the generalization, the average MAE between ML-predicted and MC-simulated dose deposition is calculated for each of the datasets. The results are shown in Table (5.1).
Overall, the MAE for the peak predictions is found to be lower than the ones for the valley doses. This is due to the peak dose distributions containing many very small values outside of the central beam area. The average MAE values for the different datasets are in agreement with respect to their uncertainty. This indicates that the model generalizes well in the frame of the split of the data into the three datasets.

Table 5.1: Average MAE for training, validation and test data predictions.

	Valley	Peak
Dataset	MAE $[1 \cdot 10^{-4}]$	MAE $[1 \cdot 10^{-5}]$
Training	2.2 ± 0.2	6.2 ± 0.1
Validation	2.1 ± 0.2	6.2 ± 0.3
Test	2.2 ± 0.3	6.4 ± 0.2

Next, the trained models are used to predict the additional test data with MRT field sizes which were not part of the training. Figure (5.40) shows two exemplary depth-dose curves at the respective centre of the scoring volume.

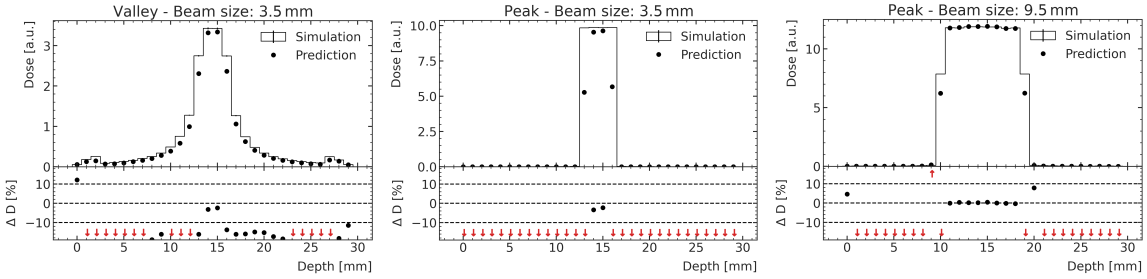


(a) (b)

Figure 5.40: Comparison of ML-predicted and MC-simulated depth-dose curves at the centre of the field for a peak dose prediction of a 9.5 mm edge length MRT field (a) and a valley dose prediction of a 3.5 mm edge length MRT field (b).

Figure (5.40a) shows the peak depth-dose curve at the centre of the scoring volume following the irradiation with a square 9.5 mm wide MRT field. The agreement is very good with only the first and last shown voxels which exhibit a very low absolute dose exceeds a deviation of 2%. The voxels within the skull deviate by at most 1.3%. Figure (5.40b) shows the valley depth-dose curve at the centre of the scoring volume following the irradiation with a square 3.5 mm wide MRT field. The overall agreement is still within 5% but a systematic underestimation of the dose by 3% on average can be seen. This trend of underestimating the

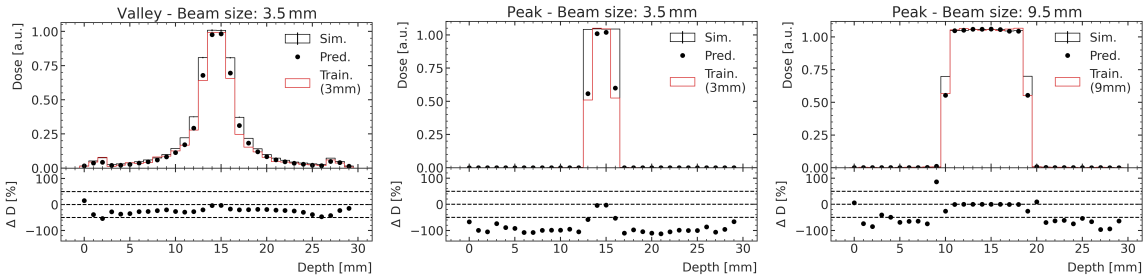
dose in the centre of the field can be observed for all tested fields. The smaller the additional test MRT field size is, the more significant the underestimation. The respective valley depth-dose curve at the centre of the field for the square 9.5 mm wide MRT field, as shown for the peak dose in Figure (5.40a), similarly does not show a notable systematic deviation.
The reason for the deviation in estimated dose at the centre of the phantom can be seen when inspecting the lateral dose profile, e.g. at the centre of the phantom as well, shown in Figure (5.41) for three different prediction scenarios: Figure (5.41a) shows the lateral valley dose profile for a squared MRT field with 3.5 mm edge length, Figure (5.41b) shows the lateral peak dose profile for a squared MRT field with 3.5 mm edge length, Figure (5.41c) shows the lateral peak dose profile for a squared MRT field with 9.5 mm edge length.



(a) (b) (c)

Figure 5.41: Comparison of ML prediction and MC simulation of the lateral profile of the (a) valley dose for a 3.5 mm MRT field, (b) peak dose for a 3.5 mm MRT field, and (c) peak dose for a 9.5 mm MRT field.

The deviation is a consequence of a lack of generalization with respect to the MRT field size. All three subfigures in Figure (5.41) exhibit significant deviations of the ML prediction from the MC simulation of the respective dose deposition. The discrepancy is especially notable in the case of the 3.5 mm MRT field peak dose prediction, shown in Figure (5.41b). Figure (5.41c) also gives an indication to why the depth-dose curves show better agreement for larger fields: the deviations occur at the edges of the predicted fields. For small field sizes, this also impacts the dose prediction at the centre of the field whereas for large fields, the centre is unaffected by the deviations at those field edges.
The lack of generalization can be seen to be a clear case of over-training on the available MRT field sizes in the training data sets when comparing the predicted dose profiles for a new field size with the respective MC simulation for a smaller field size which is part of the training dataset. This comparison is shown in Figure (5.42) for the same scenarios which were shown in Figure (5.41). The lower part of each plots shows the deviation of the ML prediction for the given MRT field size from the respective smaller MC simulated dose profile. Although the ML predictions for the 3.5 mm MRT fields (Figure (5.42a) and Figure (5.42b)) and the 9.5 mm MRT field (Figure (5.42c)) do not exactly coincide with the MC simulations for the respective smaller fields which are part of the training data (3 mm and 9 mm) the agreement is significantly better than with the MC simulations for the 3.5 mm and 9.5 mm MRT fields.
Three key learnings from this study should be noted. First, the developed model is conceptually capable of predicting peak or valley dose distributions in a simple rat head phantom for different MRT field sizes. Secondly, the over-fitting on the training data MRT field sizes was only found when using the additional test dataset comprising additional MRT field sizes. This highlights the importance of a well-designed training, validation, and test dataset in the first place to minimize the probability of not noticing a lack of generalization capability in a certain part of the parameter space which is expected to fall under the application range of



(a) (b) (c)

Figure 5.42: Comparison of ML prediction and MC simulation of the lateral profile of the (a) valley dose for a 3.5 mm MRT field, (b) peak dose for a 3.5 mm MRT field, and (c) peak dose for a 9.5 mm MRT field with the MC simulation of the closest smaller MRT field size which is included in the training dataset, shown in red. The deviation plot in the lower part shows the deviations with respect to the smaller, training data dose profile, shown in red.

the trained ML Model. Third, connected to the second one, it is found that using only discrete steps of parameters in the simulation dataset can lead to the model not generalizing well due to learning the limited variation of such a parameter from the dataset. Using continuously sampled values for the MRT field size, for example, would likely result in a better generalization. It is therefore found to be important for the generation of future datasets to sample the parameter space more evenly than it was done in this study. Due to long simulation times this can be a sever limitation of the ML dose prediction approach. A mitigation strategy is presented in the next section.

6 Towards treatment planning: retrospective dose prediction for a rat irradiation campaign

The previous section introduced the direct peak and valley dose distribution prediction using the macro voxel method in a simplified rat head phantom. This section investigates the application of the machine learning (ML) model in a preclinically more relevant scenario: the retrospective dose prediction for a preclinical microbeam radiation therapy (MRT) study with 16 rats with implanted glisarcoma, performed in April 2022 at the Imaging and Medical Beamline (IMBL). All rats were irradiated with an 8x8 mm^2 MRT field incident on the top of the head as described in the previous section. The prescribed valley doses for one group of rats were at least 8 Gy throughout the tumour, for the other one at least 16 Gy. The original purpose and the details of the preclinical study are beyond the scope of this thesis. Hence, only the dose prediction aspect will be discussed in the following. During the preclinical study, only the in-field peak and valley doses were used for treatment planning decisions. To allow for a first, simplified approach towards treatment planning using the developed ML model, the study presented in the following also only targets the dose predictions within the field.
Most parts of the results shown in this section have been already published prior to the submission of this thesis [34].

6.1 Creation of a CT-based dataset

The previous section used a simple rat head phantom modelled after a Computer Tomography (CT) rat scan. In this study, the Monte Carlo (MC) simulation and ML prediction is performed using a more realistic digital phantom directly derived from CT scans as discussed in the following.

6.1.1 Derivation of simulation phantoms from rat CT scans

The dataset comprises CT scans of 16 rats, which were taken two weeks after implanting gliosarcoma cells in their brain. The used CT scanner has a pixel pitch of 0.4-0.6 mm and a slice distance of 0.6 mm. In its original form, the scan file comprises a relatively large field of view showing in addition to the rat also the holding apparatus. Therefore, the CTs are cropped to the head of the rats and at least 5 mm padding to the bottom and top. In addition, the pixel values are transformed into Hounsfield units (HU) using the calibration function embedded into each of the CT files. An exemplary cropped and transformed CT is shown in Figure (6.1).

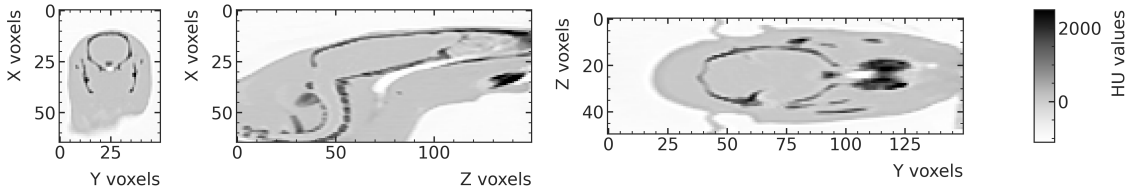


Figure 6.1: Three slices of an exemplary CT scan of a rat used in this study after being cropped and converted to HU.

As a next preprocessing step, the CT images are rotated so that the skull is centred and facing the positive z direction. An exemplary result is shown in Figure (6.2). Subsequently, the HU values are transformed to three discrete materials which are used in the

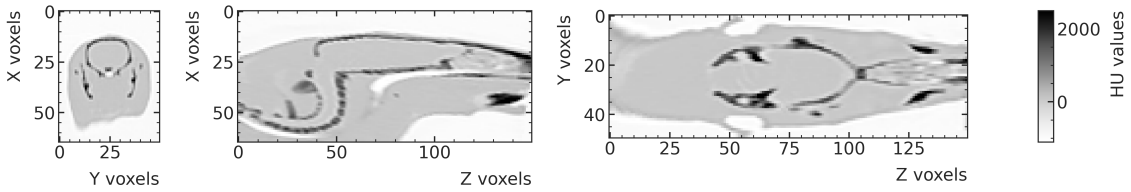


Figure 6.2: Three slices of an exemplary CT scan of a cropped and rotated rat used in this study. Reproduced from [34].

Geant4 MC simulations of the dose distributions. Voxels with less than -450 HU are defined to be air (G4_AIR [78]), voxels in the range [-450 HU, 350 HU) are defined to be water (G4_WATER [78]).
Voxels with higher HU values are defined to be bone (G4_BONE_COMPACT_ICRU [78]). During the irradiations, a nearly water-equivalent and 5 mm tick cushion (*Bolus*) was placed on top of the rat heads. To account for this during the MC simulation and the subsequent ML prediction, a 5 mm water layer is added to the rat head phantom. The result of the material digitization and Bolus addition to the scan shown in Figure (6.2) is shown in Figure (6.3)

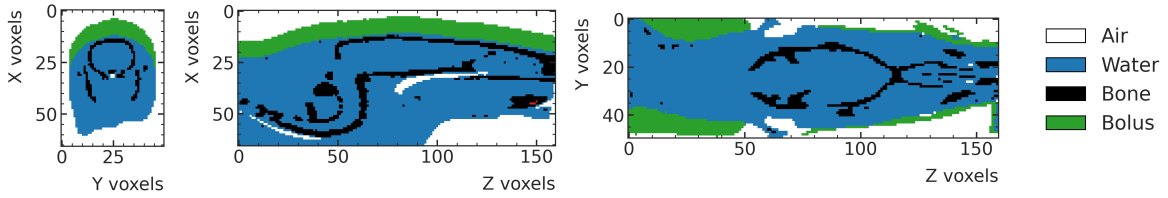


Figure 6.3: Exemplary CT scan of a cropped and rotated rat used in this study. The colour encodes the material assigned to the voxels. The Bolus is shown in green although it is also simulated as water. Reproduced from [34].

6.1.2 Using datasets with high statistical uncertainty

In Section (5.4) it was found that a large variety of irradiation scenarios is important for the model to generalize well. Due to limited computation resources, however, it is not possible to create both large and low-noise datasets. This study explores the use of high-noise MC simulations for training the ML models.
Previously, only lateral and vertical translations of the phantoms in front of the beam were performed to obtain different simulation samples. Those steps are performed to generate more variable samples and thereby minimize the risk of the ML model over-fitting to the training data. In this study, the same data augmentation of the CT images is performed as long as the centre of the 8x8 mm^2 MRT field still targets the brain of the rats. The maximum lateral translation distances (y and z, as shown in e.g. Figure (6.3)), are determined manually for each rat. In addition, the rat CT images are shifted by up to 5 mm up- or downstream with respect to the particle source. In addition, the CT scans are rotated up to 10 degrees around each axis. Finally, the individual voxels are scaled by a factor between 0.8 and 1.2. All translation distances, rotation angles and scaling factors are chosen randomly from a uniform distribution between the respective minimum and maximum. Six exemplary simulation geometries using the same CT scan (rat 2) but different augmentation parameters are shown in Figure (6.4).
The available total number of 16 rats is separated into the training, validation and test

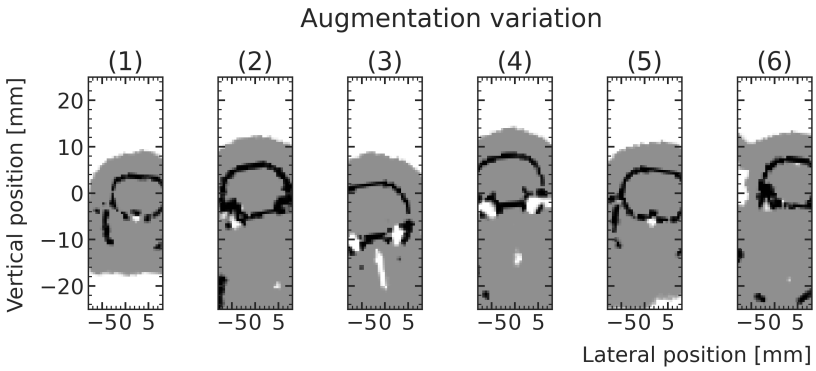
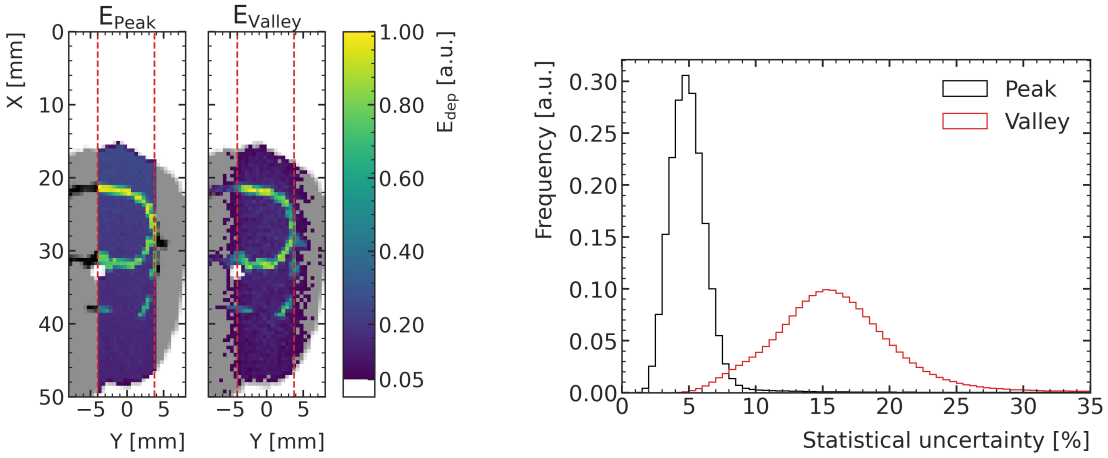


Figure 6.4: Examples of augmented data samples used for MC simulation. The colour encodes the simulation materials air (white), water (grey) and bone (black).

dataset. Following the augmentation steps described above, 4569 training data samples (rats 1-10), 1431 validation samples (rats 11-13) and 500 test data samples (rats 14-16) are simulated. Instead of the dose, the energy deposition is scored and also used as target for the ML prediction because it is found to be more suitable especially at the phantom-air-interface as doses in air scored in high-noise MC simulations tend to exhibit large fluctuations due to the deposited energy being divided by a very small mall. Using the densities of the voxels the energy deposition can be easily transformed to a dose for later analysis steps. All energy depositions are scored using the macro voxel method, resulting in a peak energy deposition distribution and a valley energy deposition distribution for each phantom geometry. Because all MRT fields in this study are 8x8 mm^2 in size, only this field size is simulated for all data samples. As the preclinical study was only concerned with the peak and valley energy deposition within the irradiation field, only this volume is predicted in this study. For this, the prediction voxel grid for the ML model is chosen to be 96x16x16 with a voxel edge length of 0.5 mm with no additional out-of-field volume included in contrast to previous studies in this thesis. To allow for a better visualization of the simulation results, the energy depositions are scored in a 96x32x32 voxel grid in the MC simulation although only the central 96x16x16 voxel grid is used for the later ML model training. An exemplary MC simulation result is shown in Figure (6.5a). It shows a 2D slice of both the peak and valley energy deposition at the centre of the simulated volume. The 96x16x16 ML prediction volume is highlighted using red dashed lines. The choice for shorter simulation times in favour of more simulation samples results in relatively large statistical uncertainties. Especially in the valley energy deposition distributions, the fluctuations are visible by eye. A more quantitative analysis of the uncertainties present in the training dataset is shown in Figure (6.5b). It shows the histograms of the voxel-wise statistical uncertainties obtained from the MC simulations. The uncertainties of the peak energy depositions are on average 5%, spreading from approximately 1% to over 11%, and the valley energy deposition average approximately 15% with a spread from less than 5% to more than 30%.

6.2 Machine learning model adaption and optimization

The ML model developed throughout this thesis is adapted for application in this study in two ways. First, the convolutional layers are adjusted so that their expected input and produced output matrices match the shape 96x16x16. In addition, the model is modified to only use the density matrix as input. No additional information, such as a water-only simulation or a distance matrix, is passed to the network because the prediction is only



(a) (b)

Figure 6.5: (a): Exemplary 2D slice of a high-noise MC simulation sample showing the simulated peak (left) and valley (right) energy deposition, normalized to their respective maximum value. The ML prediction volume is shown with red dashed lines. (b): Histograms of the voxel-wise uncertainties of the peak and valley energy depositions in the training data MC simulations. Reproduced from [34].

performed in the 8x8 mm^2 area covered by the same MRT field. Like in the previous studies, two independent networks are trained for the peak and the valley predictions. To find the best hyperparameters for the two models, a grid search is conducted as described in the previous section, using the Adam optimizer and the mean-absolute error (MAE) loss for all configurations. The respective models are trained using the high-noise training dataset, the performance is assessed by calculating the MAE on the high-noise validation dataset.

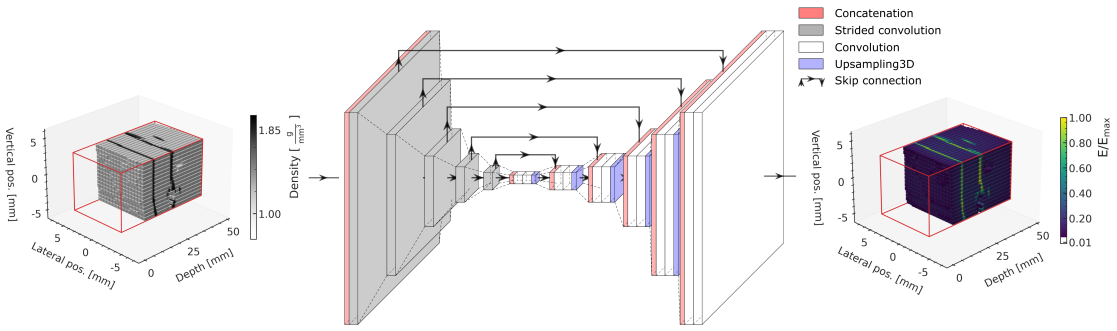


Figure 6.6: Schematic of the adapted ML model for this study. Reproduced from [34].

For the valley model, 32, 64, and 128 filters in the convolutional layers are combined with batch sizes of 4, 8, 16, 32, and 64, and learning rates of $1 \cdot 10^{-5}$, $5 \cdot 10^{-5}$, $1 \cdot 10^{-4}$, $5 \cdot 10^{-4}$, $1 \cdot 10^{-3}$, $5 \cdot 10^{-3}$, $1 \cdot 10^{-2}$, $5 \cdot 10^{-2}$. Trainings with a batch size of 4 and with the largest two learning rates did not converge. The best validation loss achieved by each model is shown in Figure (6.7). The best valley model, as indicated with the red circle, comprises 64 convolution filters, a batch size of 8 and a learning rate of $1 \cdot 10^{-3}$.
Following the results for the valley model, the peak model is investigated on an adjusted grid

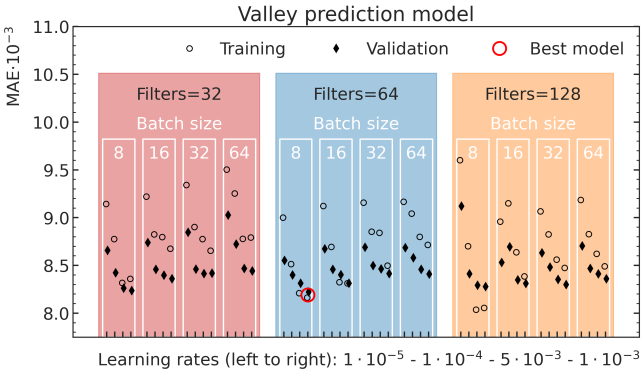


Figure 6.7: Validation (black diamonds) and training loss (open circles) of the hyperparameter search for the ML model predicting the valley energy depositions. The best model is indicated with a red circle. Reproduced from [34].

of 32, 64, and 128 filters in the convolutional layers, batch sizes of 4, 8, and 16, and learning rates of $1 \cdot 10^{-3}$, $5 \cdot 10^{-3}$, $1 \cdot 10^{-2}$, $5 \cdot 10^{-2}$. The best validation loss achieved by each model is shown in Figure (6.8). The best peak model, as indicated with the red circle, comprises 128 convolution filters, a batch size of 8 and a learning rate of $5 \cdot 10^{-3}$.

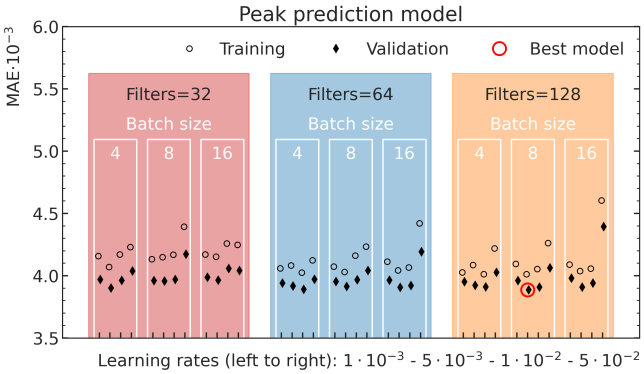


Figure 6.8: Validation (black diamonds) and training loss (open circles) of the hyperparameter search for the ML model predicting the peak energy depositions. The best model is indicated with a red circle. Reproduced from [34].

6.2.1 Performance and generalisation assessment for high-noise datasets

A notable observation in Figure (6.8) is that the training performance is lower (higher MAE) than the validation performance (lower MAE). As shown in Figure (6.7), this is also the case for several valley prediction models. Figure (6.9) allows for a closer investigation into this finding. It shows boxplots of the MAE computed using the ML predictions of the respective best peak and valley prediction model for all samples in the training (dark grey, rats 1-10), validation (medium grey, rats 11-13) and test (light grey, rats 14-16) dataset, separated by rat. For both the peak and the valley model, the performance varies between the rats.
Figure (6.9a) shows the MAE per rat for the peak ML model. Especially for the rats 1, 8 (both training dataset) and to some extend rat 13 (validation dataset), the prediction performance is lower (higher MAE) than for the other rats. This trend is also visible in Figure (6.9b), showing the valley ML model performances, though less pronounced.

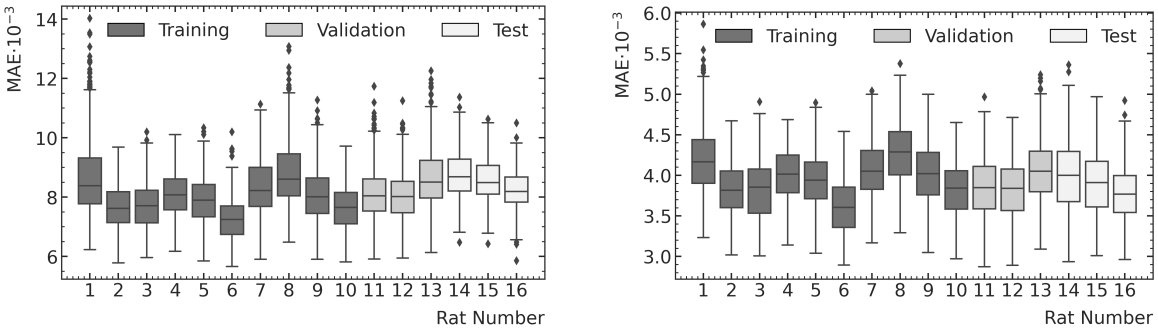


Figure 6.9: Boxplots of the MAE between the ML prediction and high-noise MC simulation of the peak (a) and valley (b) energy depositions, separated by rat. The colour indicates the training (dark grey), validation (medium grey) and test (light grey) dataset.

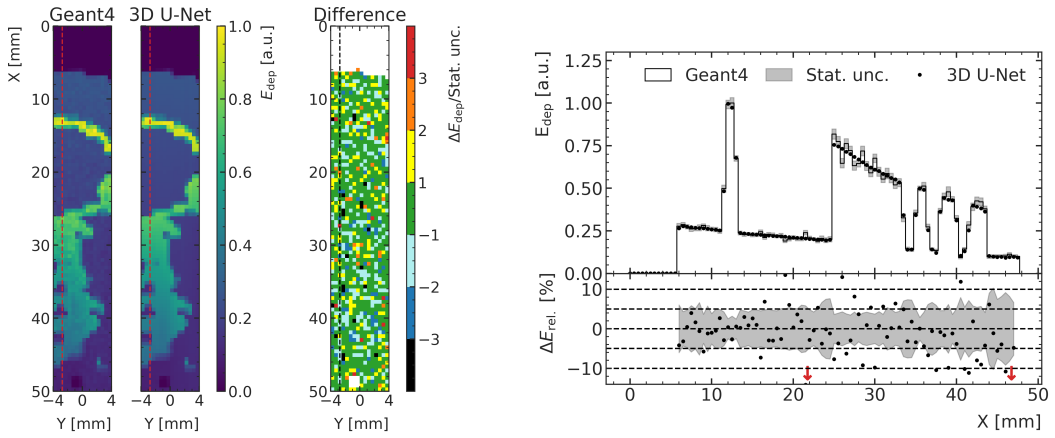


Figure 6.10: Comparison between the MC-simulated and ML-predicted energy deposition for a data sample derived from rat 1. (a): 2D slices at the centre of the prediction volume together with a colour encoded comparison in units of standard deviation. The black line indicates the location of the depth-wise comparison in (b). Reproduced from [34].

Looking into those rats, it is found that especially the data samples obtained using the CT scans of rats 1 and 8 exhibit a geometric feature which is not found in most other rats: the spinal cord as part of the prediction volume. Although a first thought could be that the network does not successfully predict the energy depositions in the spinal cord because it is a geometrical feature which is not present in enough data samples, this seems not to be the case: Figure (6.10) shows one exemplary peak energy deposition ML prediction compared to the MC simulation. Although being trained on high-noise MC data, the ML predictions are smooth and do not exhibit the noise found in the MC data samples.

In many cases, including the previous section of this thesis, comparisons between ML and MC estimates are made using the relative deviation between the simulated and ML-predicted values. In the case of high-noise simulations that is not useful because values are expected to deviate significantly only because of the statistical uncertainty of the MC simulation. Instead,

Figure (6.10a) shows the deviation between ML and MC in units of the MC standard deviation to give a more reasonable impression of occurring deviations. In fact, the energy deposition seems to be predicted without notable systematic deviations. The deviations are driven by the statistical uncertainties of the MC simulations. The reason for the lower performance, i.e. higher MAE values, follows from this as well: Because of the large energy deposition entries in the spine, larger contributions are made towards the MAE, resulting in larger MAE values for those rats comprising more samples with the spinal cord in the prediction volume.
Figure (6.11) shows two additional exemplary energy deposition ML predictions, one for the peak (Figure (6.11a)) and one for the valley (Figure (6.11b)) region, respectively.

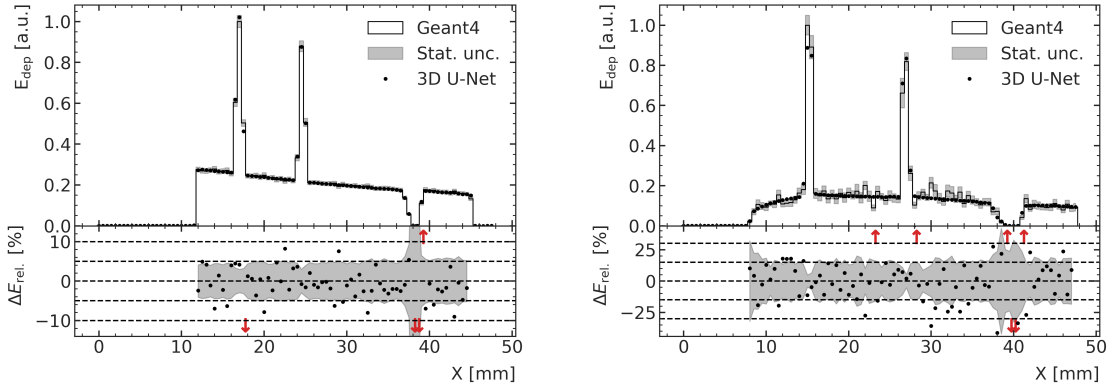


Figure 6.11: Comparison between the ML prediction and MC simulation of the peak (a) and valley (b) energy deposition for an exemplary test data sample. The grey band indicates the statistical uncertainties of the MC simulation. Reproduced from [34]

Apart from several voxels with larger relative deviations which are largest in the air cavity at the distal end of the phantom, most deviations can be seen to be within one standard deviation of the MC simulation which is shown as a grey band. In both cases, no systematic deviation is observed in the ML predictions.
To allow for a more relevant comparison of the smooth ML predictions with the noisy MC simulations, a different method is proposed. The central $\pm 1\sigma$ interval around the mean value of a normal distribution covers 68% of the area under the curve. This means that on average, 68% of random numbers which are sampled from such a normal distribution are expected to be within that $\pm 1\sigma$ interval. Assuming Gaussian uncertainties for the noise in the MC simulations, the simulated value in each voxel represents a sampled number with a given standard deviation and expectation value. If the ML model is trained to produce unbiased estimations of the expected values of the energy deposition for each voxel, 68% of the voxels exhibit a deviation between ML prediction and MC simulation of less than one standard deviation. If, on average, more than 68% of the voxels exhibit a smaller deviation, this would hint at over-fitting to the noise present in the data. A lower value than 68% indicates that the deviations can not be explained by the statistical uncertainties alone but there are additional deviations, for example due to systematic over- or underestimation of the energy deposition.
The presented expected value can be used to investigate the agreement between ML prediction and MC simulation for the training, validation, and test datasets. Figure (6.12) shows the histograms of the voxel-wise deviations in units of the respective statistical uncertainties. Figure (6.12a) shows the histograms computed on the peak ML prediction and MC

simulation, Figure (6.12b) for the valley. In both cases, all distributions of deviations are below the expected 68%, indicating a bias in the ML prediction. However, the performance on all three datasets is relatively close to the expected average of 68% agreement within one standard deviation. The mean values for all three datasets are shown in Figure (6.1) together with the standard error of the mean value and also the average mean absolute error computed for each dataset. The ML prediction and MC simulation agree for the training dataset in approximately 64% of the peak voxels and 65% of the valley voxels. The reported results on the validation datasets are on average approximately 1% lower but agrees with the training data performance within the uncertainties, despite the deviations seen earlier in the hyperparameter optimization due to the higher MAE values in the training dataset.

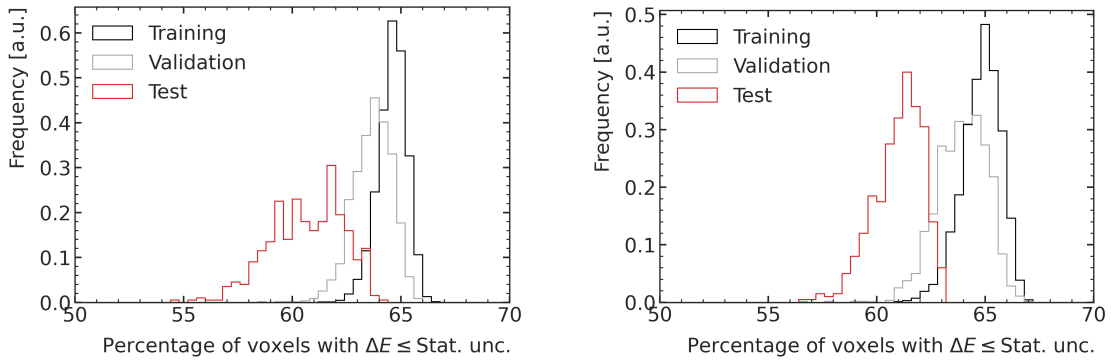


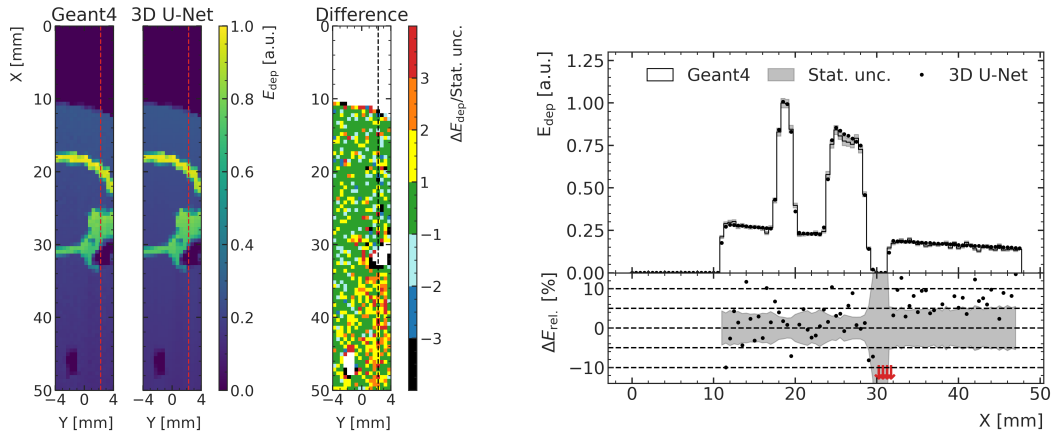
Figure 6.12: Histograms of the voxel-wise deviation between ML-predicted and MC-simulated energy deposition for the peaks (a) and valleys (b), computed for each dataset separately. Reproduced from [34].

Table 6.1: Average MAE and fraction of voxels in which the ML-predicted and MC-simulated energy deposition agrees within one standard deviation of statistical uncertainty for the three datasets and the peak and valley predictions.

	Valley		Peak	
Dataset	MAE $[1 \cdot 10^{-3}]$	ΔE < Stat. unc. [%]	MAE $[1 \cdot 10^{-3}]$	ΔE < Stat. unc. [%]
Training	8.2 ± 0.3	64.8 ± 0.9	4.0 ± 0.2	64.6 ± 0.7
Validation	8.2 ± 0.2	63.9 ± 1.2	3.9 ± 0.1	63.7 ± 0.9
Test	8.4 ± 0.1	61.0 ± 1.1	4.1 ± 0.1	60.7 ± 1.7

For the test data samples, on average 61% of both the peak voxels and valley voxels are in agreement between ML prediction and MC simulation following this method. The average MAE for the three datasets is in agreement with respect to its uncertainty, not indicating over-fitting to the data. However, a significant difference of the mean values of the discussed histograms can be found between the data set, which is an indicator for slight over-fitting. To investigate more closely whether the trained models are still suitable to proceed with, the respective worst examples from the test dataset are inspected. The respective lowest agreement within MC statistical uncertainty is achieved on the data samples shown in Figure (6.13) for the

peaks and in Figure (6.14) for the valleys.



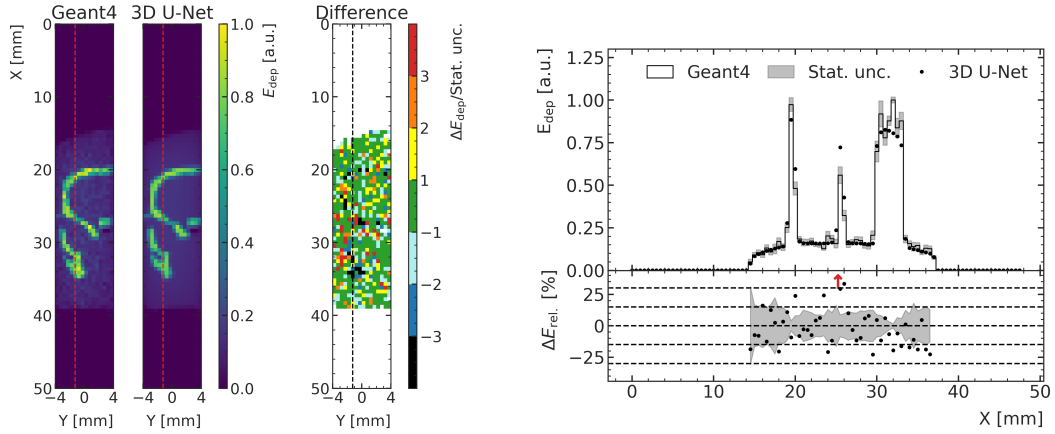
(a) (b)

Figure 6.13: (a): 2D comparison of the peak energy deposition data sample with the lowest agreements between ML prediction and MC simulation showing the respective estimation and the deviation in units of statistical uncertainty. (b): Respective depth-energy deposition curve obtained with ML and MC, located at the red dashed lines indicated in (a). Reproduced from [34].

The peak energy deposition prediction in Figure (6.13a) can be seen to agree relatively well with the MC simulation up to the air cavity located at approximately $x = 30\,\text{mm}$ which is the ear tunnel of the rat. On the distal side of this ear tunnel, the ML model systematically overestimates the peak energy deposition, leading to a low agreement with MC with respect to its standard deviation. The found deviations, however, are mostly smaller than 10% except in the air cavity where they are significantly larger due to the material being air only, leading to very small absolute numbers being recorded there.
Other data samples with low ML-MC agreement exhibit similar geometric features and prediction deviations. While this might require additional focus in future studies, a maximum deviation of approximately 10% for the worst-case scenarios are found suitable in this work.
The valley data sample with the lowest agreement between ML and MC, shown in Figure (6.14), exhibits larger deviations in voxels which are simulated as bone material with many voxels exhibiting deviations of more than 3σ statistical uncertainty.
Nevertheless, the trained ML models are found to produce suitable predictions both in the peak and valley regions when being compared to the high-noise MC simulations. The conducted analysis concludes that the ML predictions are close to unbiased estimations of the expected energy deposition values of the noisy MC simulations. To investigate the accuracy more closely, it is required to compare them to MC simulations with lower statistical uncertainty at this point.

6.3 Predictions for test rat patients in a preclinical treatment scenario

To allow for the closer inspection of the agreement of ML predictions with the MC simulations, three additional MC simulations with very low statistical uncertainties are created. Those three simulations, one each for the rats 14, 15, and 16, respectively, resemble the respective treatment scenarios in the preclinical study this study focuses on. This part of this study therefore also functions as a first practical test of the applicability of the model in an actual



(a) (b)

Figure 6.14: (a): 2D comparison of the valley energy deposition data sample with the lowest agreements between ML prediction and MC simulation showing the respective estimation and the deviation in units of statistical uncertainty. (b): Respective depth-energy deposition curve obtained with ML and MC, located at the red dashed lines indicated in (a). Reproduced from [34].

preclinical scenario. The number of test patients is very low, however, and future studies will have to be conducted such tests more thoroughly.

6.3.1 Test data samples with low statistical uncertainties

The rat heads are placed in front of the beam so that the gliosarcoma tumour is centred in the MRT field. During the preclinical study, the locations of the tumours were manually segmented on the CT scans, which were taken after injection of a contrast agent, facilitating the visual tumour detection. This location information was made available for this thesis. The resulting tumour locations (red) in the brains of the three test rats are shown in Figure (6.15) in which they are positioned with their respective centre of mass in the centre of the ML prediction volume (red lines). The tumours visibly vary in size and shape. The apparent satellite tumour in rat 14 (lower left corner or the segmented tumour) is actually connected to the rest of the tumour in a different slice of the CT. The resulting low-noise MC simulation for rat 15 is shown in Figure (6.16a). It shows a 2D slice of both the peak and valley energy deposition at the centre of the simulated volume. The 96x16x16 ML prediction volume is highlighted using red dashed lines. The 2D slices do not exhibit visible fluctuations anymore in these simulations. Figure (6.16b) shows the voxel-wise histograms of the uncertainties found in these three test rat samples. The peak energy depositions exhibit a statical uncertainty of 0.36% on average and the valley energy depositions exhibit a statistical uncertainty of 1.28% on average, both with a much narrower spread than in the high-noise simulations.

6.3.2 Performance and generalisation assessment for low-noise datasets

In this section, the ML predictions and MC simulations are reported in Gray by dividing the respective energy depositions by the voxel-wise density, because in a preclinical setting the dose is the main quantity of interest and the lower noise in the MC simulation allows a stable computation of it. Table (6.2) shows the fraction of voxels, for which the ML predictions agree within 3% with the MC simulation of the respective simulated peak and valley doses.

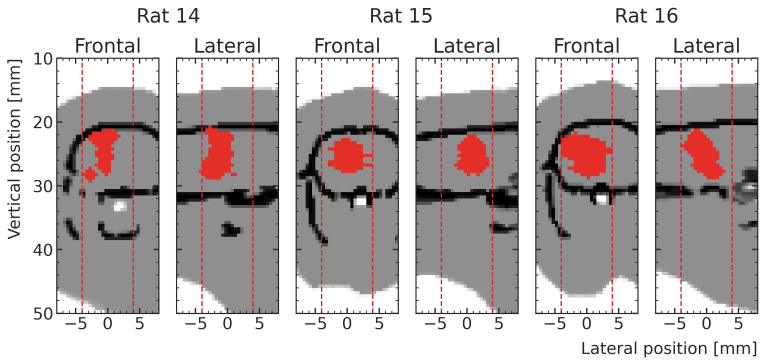
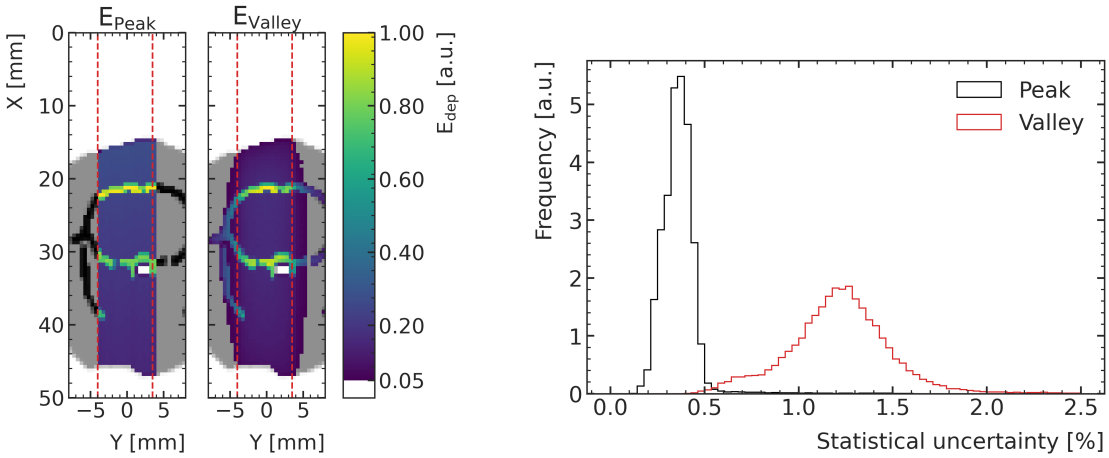


Figure 6.15: Visualization of the tumours (red) located in the brains of the three test patients. The colour encodes the simulation materials air (white), water (grey) and bone (black). The ML prediction volume is indicated with red dotted lines.



(a) (b)

Figure 6.16: (a): Exemplary 2D slice of a low-noise MC simulation sample showing the simulated peak (left) and valley (right) energy deposition, normalized to their respective maximum value. The ML prediction volume is shown with red dashed lines. (b): Histograms of the voxel-wise uncertainties of the peak and valley energy depositions in the low-noise test data MC simulations. Reproduced from [34].

As the numbers are obtained from counting the deviations of all voxels of the three test rats, the numbers do not exhibit uncertainties.

The peak doses are predicted with an accuracy of at least 3% for at least 93.9% of all voxels in the phantom for all three test rats. It is noteworthy that in the tumour volumes, the ML prediction deviates by less then 3% from the MC simulation for all voxels. The valley dose agreement is lower overall, with 77.6% (rat 14), 81.1% (rat 15) and 80.1% (rat 16) of predicted voxels deviating by less than 3% from the MC simulations. Comparing this to the ML model performance in the tumour volume, the agreement is a lot higher there with over 95.0% of voxels deviating less then 3% for all three rats with a maximum of 97.9% for rat 16. The reason for the larger deviation with respect to the whole phantom geometry as compared to the tumour region can be seen in Figure (6.17). While the agreement of the ML prediction with the MC simulation is very good inside of the brain, larger deviations can be seen around bone voxels and especially towards the distal ends of the phantoms: for the test rats 14 and

Table 6.2: Ratio of voxels in which the ML-predicted and MC-simulated peak and valley doses agree within 3%, shown for the full phantoms, only voxels comprising tissue and the tumour volumes. Reproduced from [34]

Rat ID	Peak/Valley	Voxel ratio with Δ D $< 3\%$ [%]		
		Full phantom	Tissue only	Tumour volume
14	Peak	93.9	95.0	100.0
	Valley	77.6	81.0	95.9
15	Peak	93.9	95.7	100.0
	Valley	81.1	85.0	97.7
16	Peak	94.6	96.1	100.0
	Valley	80.1	83.8	97.9

15, the ML model underestimates the dose there while it overestimates the dose for rat 16. The deviations, however, are mostly less than 5% (yellow and turquoise voxels).

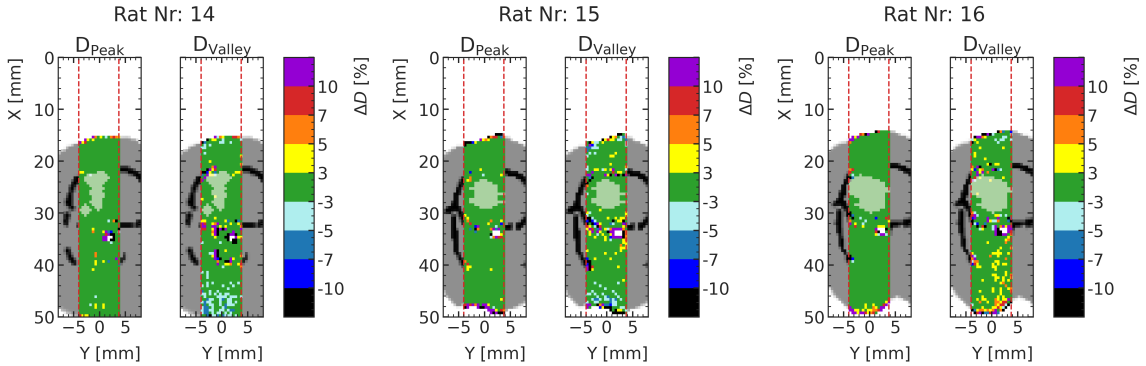
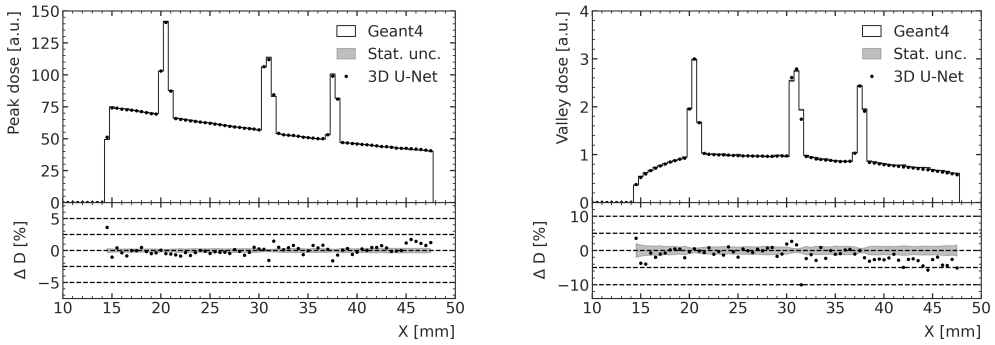


Figure 6.17: Comparison of ML-predicted and MC-simulated peak and valley doses for the three test rats, showing the relative difference $\Delta D = (D_{\mathrm{ML}} - D_{\mathrm{MC}})/D_{\mathrm{MC}}$. The tumour volume in the shown slice is indicated white. Reproduced from [34].

To allow for a better visual inspection of potential systematic deviations, Figure (6.18) shows the respective depth-dose curves for rat 14, exemplarily, at the centre of the prediction volume. Figure (6.18a) shows the depth-peak dose curve, exhibiting a larger deviation than 2.5% only at the very first voxel at the entrance of the MRT field into the phantom with a deviation of approximately 3%. All other voxels in this line agree with MC by less than 2.5%. The depth-valley dose curve in Figure (6.18b) shows the previously found tendency of the model to systematic underestimate the doses towards the distal end of the phantom. The deviations at the centre of the field are, however, found to have maximum values of around 5%. This is still found to be a very good agreement of the ML model with the MC simulation for the valley doses.

The overall goal of fast dose predictions with ML is, at this stage, not the complete replacement of highly accurate Geant4 dose simulations. Instead, the presented ML model is designed to provide a method for fast dose predictions suitable for e.g. treatment plan optimization. The



(a) (b)

Figure 6.18: Depth-wise comparison of ML-predicted and MC-simulated peak and valley doses for rat 14 at the centre of the prediction volume. Reproduced from [34].

previous section showed an agreement within 95% of the dose predictions within the tumour volumes of three exemplary treatment cases from a preclinical study. Due to the slightly reduced size of the ML model compared to earlier studies in this work, the prediction speed is found to be faster with approximately 50 ms per dose prediction in contrast to a simulation time of 20 hours for the low-noise MC data. The excellent agreement of the ML model with the MC simulations indicates that the ML model in fact learns an unbiased estimation of the expectation value for the voxels when being trained on high-noise data. These findings are found to make a strong case for further investigations considering the use of ML methods as part of the MRT workflow.

7 One end, many beginnings: what are the next steps?

During the previously discussed preclinical study only the in-field peak and valley doses using a fixed-size 8x8 mm^2 microbeam radiation therapy (MRT) field were used for treatment planning decisions and were thus the main goal of the machine learning (ML) model to predict. Future applications of a fast ML-based dose estimation engine will require more versatile ML models.
This section presents four research directions, which are currently explored to extend the developed model towards its applicability in future preclinical or even clinical MRT studies: 1) out-of-field predictions, 2) ML model robustness, 3) dose prediction for conformal MRT fields, and 4) a new, more complex evaluation of the portability to other fractionated therapies using the example of proton minibeam radiation therapy (pMBRT).

7.1 Expanding the prediction volume: out-of-field dose

Especially when considering the future application in treatment plan optimization, the accurate prediction of out-of-field doses, especially to organs at risk around the MRT field, will be increasingly important. This section presents a first extension study of the preclinical rat ML dose model from the last section towards a larger prediction volume. Only the valley dose is considered because it extends into the out-of-field region in contrast to the peak dose which is mostly limited to the in-field region and therefore does not necessarily require an extension of the prediction volume.
The Monte Carlo (MC) dataset and separation into training, validation and test data is the same as in the previous section. However, the ML model is adapted to be trained on and predict the whole MC scoring volume of 96x32x32 voxels. Figure (7.1) is the same as Figure (6.5a) but instead of the ML prediction volume, the red lines indicate now the in-field region of the MRT irradiation field, which was coincident with the ML prediction volume in the previous study. During training, the energy deposition is used again instead of the dose.

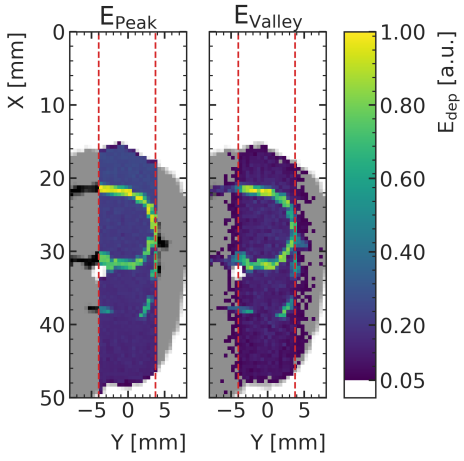


Figure 7.1: Exemplary 2D slice of a high-noise MC simulation sample showing the simulated peak (left) and valley (right) energy deposition, normalized to their respective maximum value. The MRT field region is shown with red dashed lines, the ML model predicts the entire shown volume.

After adapting the ML model to allow for the 96x32x32 voxel matrix input and output, a new grid search is performed to find optimal hyperparameters. Due to this study being an outlook towards future research directions however, the chosen grid is limited to 64 and 128 filters,

batch sizes of 4, 8, and 16 and learning rates of $1 \cdot 10^{-4}$, $1 \cdot 10^{-3}$, and $5 \cdot 10^{-3}$. As before, the Adam optimizer and the mean-absolute error (MAE) loss are used for all trainings. The results of the search are shown in Figure (7.2). The network with 64 convolutional filters, a batch size of 4 and a learning rate of $5 \cdot 10^{-3}$ is found to perform best on the validation dataset. It should be noted, however, that many different configurations lead to very similar network performances.

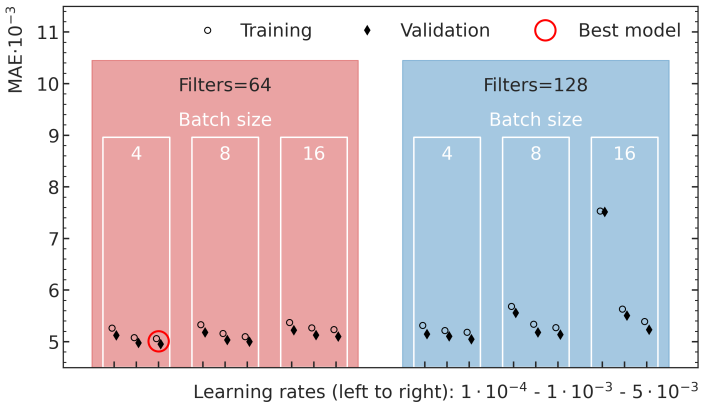


Figure 7.2: Results of the hyperparameter search for the ML model predicting the larger valley energy depositions, as determined by the best validation loss. The training data performance of the model performing best on the validation data is shown as a black open circle.

As in the previous study, the training performance is seen to be lower (higher MAE) than the validation performance. The trained model is subsequently used to predict all training, validation, and test datasets. To assess the accuracy and generalization with regard to the high-noise datasets, both the average MAE and the voxels in agreement between the ML prediction and MC simulation with regard to their respective statistical uncertainty of the MC simulation are reported in Table (7.1s).

Table 7.1: Average MAE and fraction of voxels in which the ML predicted, and MC simulated energy deposition agrees within one standard deviation of statistical uncertainty for the three datasets.

Dataset	MAE $[1 \cdot 10^{-3}]$	$\Delta E <$ Stat. unc. [%]
Training	5.3 ± 0.2	63.5 ± 0.8
Validation	5.0 ± 0.1	63.3 ± 0.8
Test	5.3 ± 0.1	60.9 ± 0.7

The MAE values compared between the training and test dataset are found to be in agreement with each other. The rate of voxels with less deviation between ML and ML than one sigma of statistical uncertainty is found to be very similar to the previous study, again revealing a lower agreement for the test dataset. The respective histograms of the distributions are shown in Figure (7.3). The test dataset performance is found to be acceptable upon investigating the worst performing samples from the test dataset.
The sample with the overall lowest ratio of voxels in agreement between ML prediction and MC simulation within one standard deviation (worst case) is shown in Figure (7.4).

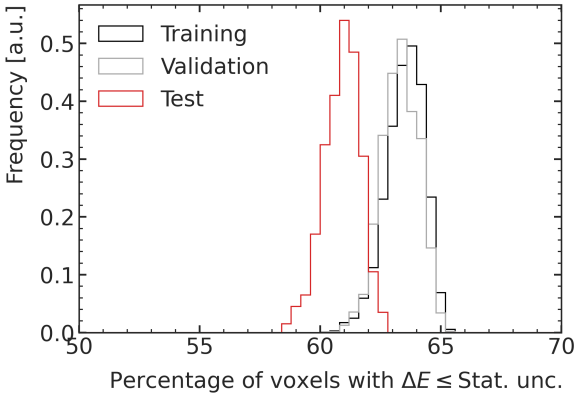
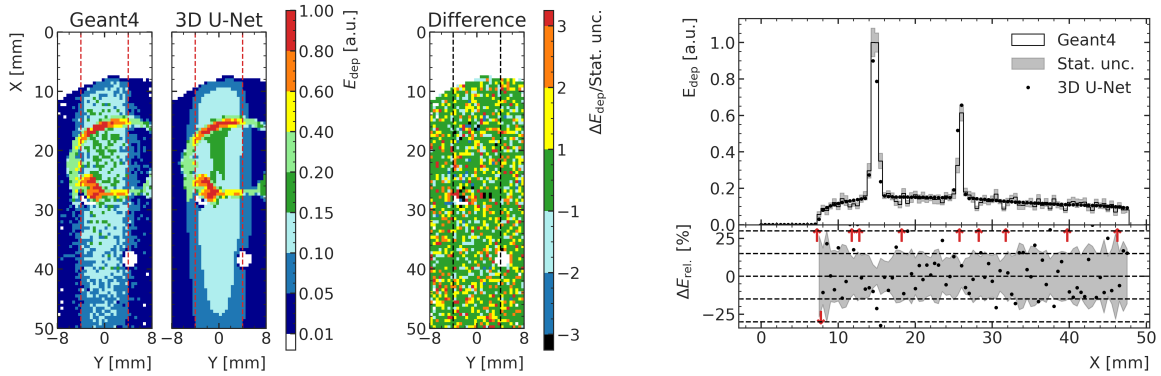


Figure 7.3: Histograms of the voxel-wise deviation between ML-predicted and MC-simulated energy deposition, computed for each dataset separately.



(a) (b)

Figure 7.4: (a): Test data sample with the lowest ratio of voxels in agreement between ML prediction and MC simulation. (b) Depth-wise comparison of ML prediction and MC simulation at the centre of the field.

While the agreement seems to be very good overall, a visible systematic in the depth-wise comparison shown in Figure (7.4b) is the more frequent occurrence of overestimations of energy deposition by the ML model in contrast to underestimations. This is shown by the majority of deviations larger than the shown area, indicated by red arrows in the lower part of the plots, especially, occurring towards larger values. The shown deviations exceed 30% which is the maximum on the y-axis of the lower plots. While this appears to be a large deviation, it should be kept in mind that the standard deviation averages 15% in the high-noise dataset. Therefore, it is useful to include again the additional three low-noise simulation samples. To first investigate the observed bias in the predictions, Figure (7.5) shows histograms of the relative deviations between ML prediction and MC simulation per voxel for all three test data samples.

Voxels with less than 1% of the maximum energy deposition are not considered for this comparison. The outermost bins are used as under- and overflow bins, respectively, showing all values which are lower or higher than the shown range from -10% to 10%. The mean values of the distributions and the percentage of voxels with more than -10% deviation (underflow) and more then +10% deviation (overflow) relative to the MC simulation are shown in Table (7.2) for a better overview of the results.

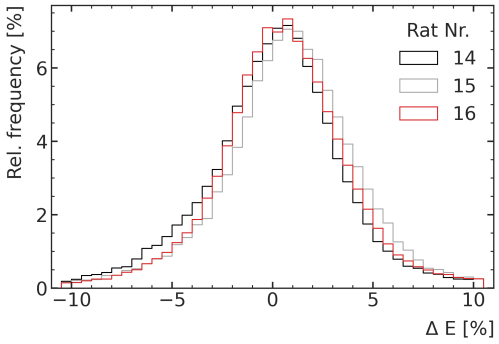


Figure 7.5: Histograms of the relative deviation between ML-predicted and MC-simulated energy deposition for the three low-noise test data samples. Only voxels with at least 1% of the respective maximum of MC simulated energy deposition are considered. The lowest and highest bin function as under- and overflow bins.

Table 7.2: Mean relative deviation between ML-predicted and MC-simulated energy deposition together with the percentage of voxels contained in the respective underflow and overflow bins.

Rat	Mean value [%]	Underflow [%]	Overflow [%]
14	1.08 ± 0.05	2.80	5.16
15	2.73 ± 0.06	2.36	6.87
16	1.78 ± 0.04	2.00	5.92

For all three predictions, more than 90% of the voxels (rat 14: 92.0%, rat 15: 90.7%, rat 16: 92.10%) exhibit less than 10% deviation between ML-predicted and MC-simulated energy deposition. The mean values are found to be larger than zero between +1.08% (rat 14) and +2.73% (rat 15), confirming the found systematic overestimation by the ML model. Similarly, nearly twice as many voxels exceed +10% deviation from ML to MC as compared to -10%.

Figure (7.6) shows a 2D comparison of the ML-predicted and MC-simulated energy depositions at the centre of the field for rat 15 which is found to exhibit the largest deviations. To allow for an easier visual assessment of the different regions, a discrete colour map was chosen. The main source of deviations is found to stem from the edge region outside of the field. In the right-side plot this can be seen in form of yellow and orange voxels signalising ML model overestimations between +2.5% and 7.5%. Deviations of more than 10% in any directions are mostly seen around the edges of bone structures which is attributed to the steep gradients there.

Figure (7.7) shows a depth-wise comparison of the ML prediction and the MC simulation along the line of the largest deviations between those as seen in Figure (7.6). A trend to overestimations by the ML model can be clearly seen but overall, the predictions are found to be in good agreement as deviations of up to 10% were assumed to be acceptable at this stage, especially taking into account that this study also considers out-of-field voxels.

The contribution of bone-tissue interfaces to the number of voxels with larger deviations can also seen in Figure (7.8a), which shows additional 2D slice comparisons right outside the field (Figure (7.8a)) and the last predicted slice, furthest from the centre (Figure (7.8b)).

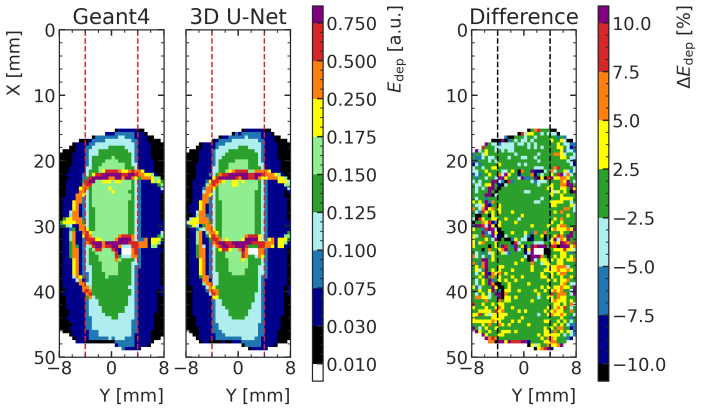


Figure 7.6: Comparison of ML-predicted and MC-simulated energy deposition for rat 15 at the centre of the field.

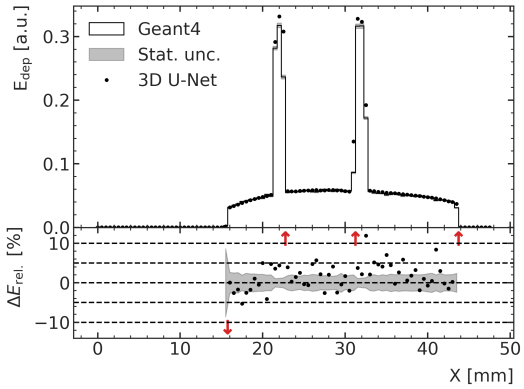
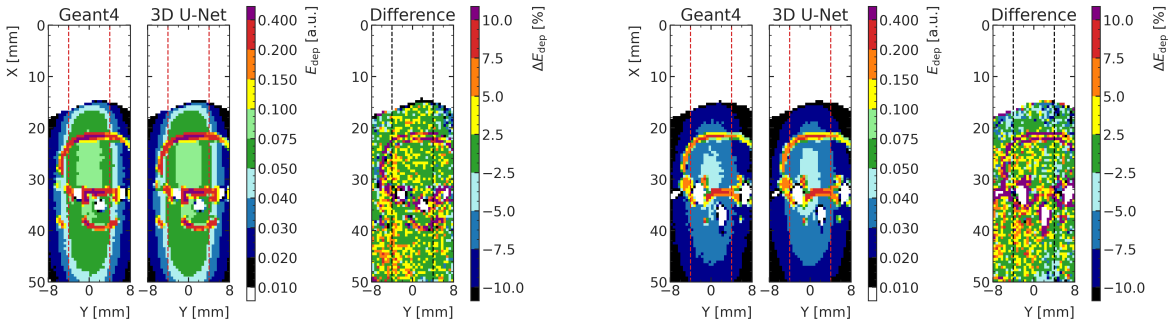


Figure 7.7: Depth-wise comparison of ML-predicted and MC-simulated energy deposition for rat 15 at the position of the largest deviations between them.

Overall, the deviations are significantly larger than in the centre, but this is partly due to larger statistical uncertainties of the MC simulation in those areas as well. However, voxels comprising bone-tissue interfaces nearly consistently exceed 10% deviation in the shown slices.



(a) (b)

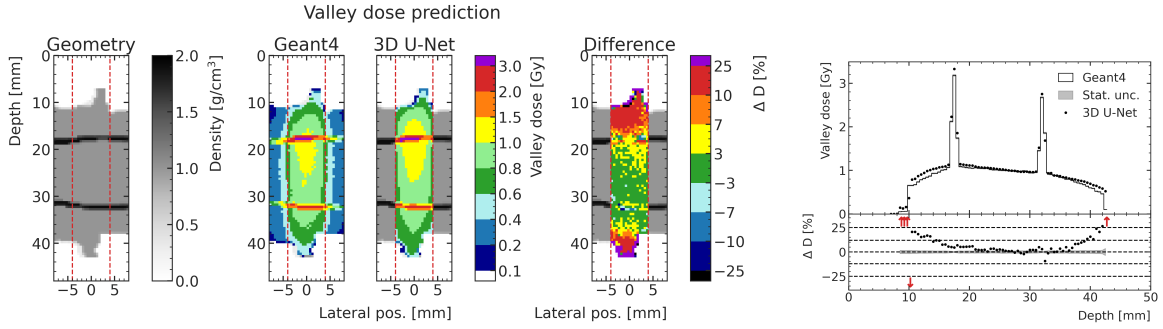
Figure 7.8: Comparison of ML-predicted and MC-simulated energy deposition for rat 15 at the edge of the MRT field (a) and at the outermost slice of the prediction volume (b).

Overall, the simplicity in expanding the predicted volume of the developed model, is an

encouraging finding of this outlook study. At this stage, the produced results are accepted as valuable fast method to produce preliminary dose distributions. Future studies on the applicability of the model for treatment planning will have to assess the capabilities of their models especially in those regions outside the field if doses in those regions are considered important for planning or optimization decisions. Investigations should include looking into ML models with a larger computational capacity (more trainable parameters), a loss function more adequately describing the learning goal than the currently used MAE, and also different, potentially better suitable ML architectures which have not been explored in the scope of this work yet.

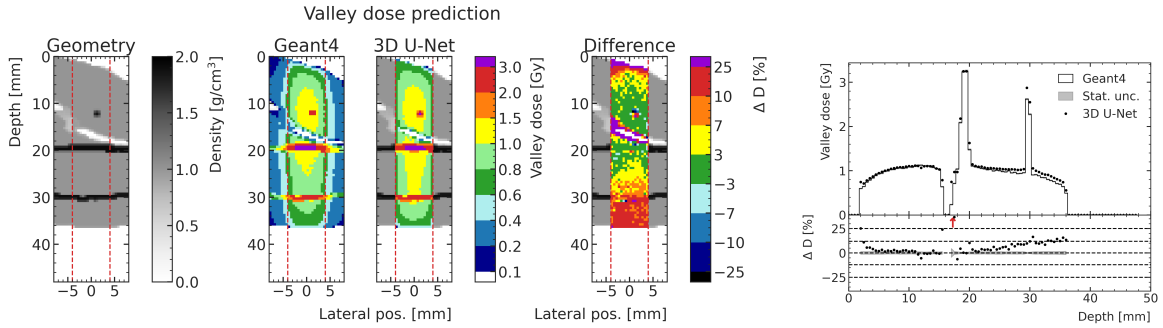
7.2 Data samples outside the training scope: analysis of model robustness

ML models are generally not necessarily expected to perform well on data samples which are too different from the training data, not presenting an interpolation of training data samples. For this reason, it is important to cover as many irradiation scenarios as possible during the creation of a training dataset to reduce the probability of future samples laying outside of the capabilities of the trained model. Nevertheless, it can be insightful to observe the model's behaviour when applied to data which it might be purposely or accidentally be applied on in the future but is expected to be substantially different from the training data.
Two example for such cases are shown in Figure (7.9) and Figure (7.10). They show the predicted valley doses using the ML model trained for the preclinical study at the Imaging and Medical Beamline (IMBL) discussed in Section (6), but applied to rat phantoms rotated by 90° (irradiation from the side of the head) and 180° (irradiation from the bottom of the head). Figure (7.9a) shows the density matrix of the first exemplary *extraordinary* prediction case where a rat head is irradiated from the side instead of from the top like in the training data samples. While the resulting deviations from the MC simulation are significantly larger than in the previous study, they mostly do not exceed 25%. The deviations are most notable in the entrance and exit region of the phantom. The ear of the rat seems to have an impact on the prediction quality as well as the model tends to overestimate the dose downstream from it until into the brain volume (group of yellow voxels extending into the brain region shown within the skull (black voxels)). Figure (7.9b) shows a depth-wise comparison in the centre of the field to give a better impression of the development of the ML prediction with respect to the MC simulation.
Figure (7.10b) shows the results of the second scenario comprising a rat head being irradiated from the bottom of the head. This case is especially noteworthy because for example the position of the skull is significantly different compared to the irradiation cases from the top of the head. The network nevertheless does not create unreasonably large predictions around the region where the skull was located in the training data, a problem which was observed similarly for an earlier version of the ML model in Section (3.2.3).
While the deviations in these examples are comparatively large it can be noted that the model does not *break down* i.e. produce extremely wrong or unexpected output values. This can be seen as a good sign of the stability of the model as the deviations especially within the brain are mostly within the self-set limit of 10%, but also potentially makes an assessment of what future input samples are valid and which ones are not more difficult as also unexpected input samples result in outputs which are relatively similar to those of expected inputs.
In the case of an ML model being used for dose predictions in preclinial or even clinical studies, this aspect requires a great deal of attention. Future studies on this field of application might investigate methods to detect data samples laying outside of the capability of their ML models and to potentially flag them e.g. as not predictable. Like this, it would be possible for a researcher to quickly identify problematic scenarios and either enhance the model or use a



(a) (b)

Figure 7.9: (a): 2D comparison of the ML-predicted and MC-simulated valley dose distribution following a rat head irradiation from the side, a scenario not covered by the training dataset shown in the left subfigure (white: air, grey: water, black: bone), together with the relative deviations on the right side. (b): Depth-wise comparison at the centre of the field.



(a) (b)

Figure 7.10: (a): 2D comparison of the ML-predicted and MC-simulated valley dose distribution following a rat head irradiation from the bottom, a scenario not covered by the training dataset shown in the left subfigure (white: air, grey: water, black: bone), together with the relative deviations on the right side. (b): Depth-wise comparison at the centre of the field.

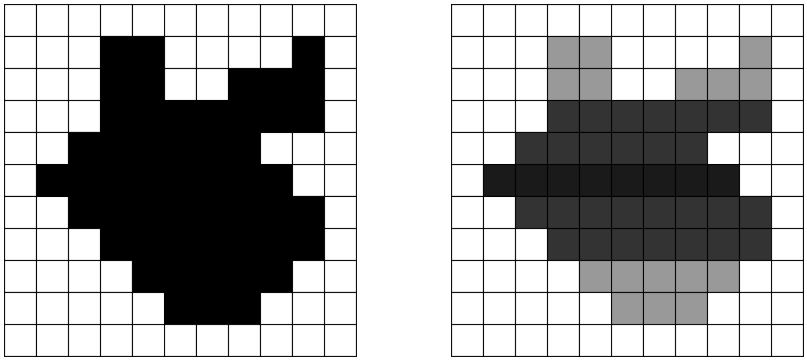
different method of dose calculation for those cases.

7.3 Dose predictions for conformal MRT irradiations

Conformal tumour irradiations with intensity-modulated radiation therapy (IMRT) and volumetric modulated arc therapy (VMAT) have led to a significant enhancement in treatment outcomes [14] by allowing for higher doses in the tumour volume while at the same time reducing the exposure of the surrounding healthy tissue. MRT research studies also have frequently included conformal masks or also leaf collimators to shape the MRT field (e.g. [17]). In addition, developments towards volumetric modulated MRT methods like spiral MRT [112] have gained attention in recent years.

All those methods usually require the prediction of doses following irradiations with different intensities and potentially also from different directions. Within the scope of this thesis, the microbeam superposition approach was presented in Section (5.3) which is highly capable of creating dose predictions for MRT fields of varying intensity. This line of research was

not continued in this thesis due to potentially accumulating systematic deviations in the predictions and the relatively long prediction times for many evaluations to generate a full field from individual microbeams. Instead, a study was conducted on directly predicting the dose following irradiations with full field of different sizes, however, only rectangular fields of different sizes Section (5.4). Resulting from this it was found to be important to include as many field size variations as possible within the realistic range of future fields as part of the training data due to over-fitting to the provided field sizes in the study.
The possibility of predicting more variable field shapes using binary masks describing their shape as additional input to the proposed ML model is investigated within the scope of a MSc project [113] supervised as part of the work on this thesis. An example of a binary mask describing an MRT field shape as input for an ML model is shown in Figure (7.11a). This approach could potentially be expanded towards also allowing for predicting intensity-modulated MRT fields, as is indicated in Figure (7.11b) which shows the mask in different shades of grey, representing different beam intensities.



(a) (b)

Figure 7.11: Binary (a) and intensity-modulated (b) mask as possible additional inputs for an ML model predicting conformal MRT fields.

In case of predictions in large Computer Tomography (CT)-based volumes, rotating the target phantoms as it has been done throughout this thesis might become infeasible due to a high computational cost to perform the rotation transformation to the CT data. Instead, the ML model might be adapted in the future to already incorporate direction-dependent predictions in a static CT-based phantom. This could be realised with a cubic prediction volume fitting the entire CT-based phantom, together with the projection of a mask on this volume as indicated in Figure (7.12).
The use of such masks, especially in the direction-dependent case, may lead to the requirement of very large datasets to mitigate over-fitting, which might render those solutions infeasible after all.
Future studies focusing on fast dose predictions for conformal MRT field dose predict will have to investigate more closely how to predict the dose distributions following conformal MRT irradiations. A re-evaluation of the use of superposition approaches may also be required as it allows for a more general approach of combining fields from individual beams, potentially resulting in fewer data samples being required for successful training of ML models.

7.4 Re-evaluation of the model portability

All studies in the scope of this thesis are performed using the same phase space file resulting from a simulation of the IMBL with two copper filter paddles in place. For future applications,

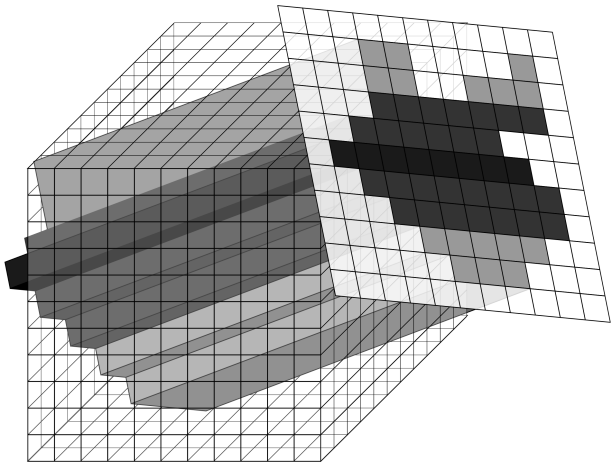
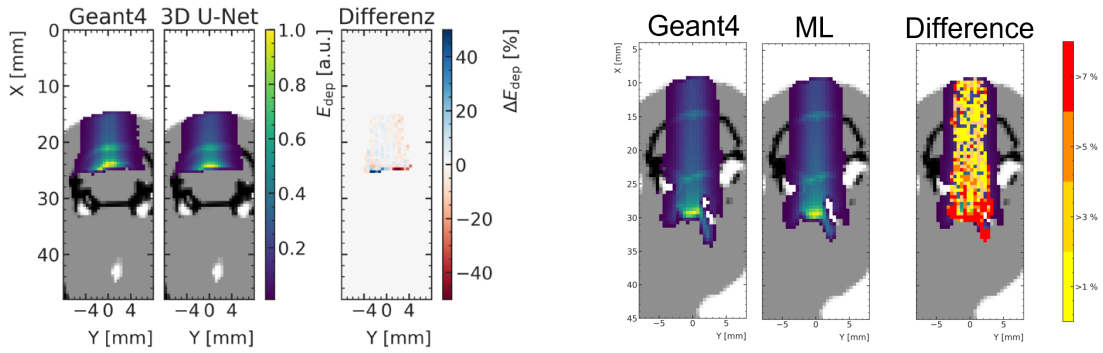


Figure 7.12: Directional intensity-modulated mask as possible input for an ML model predicting direction-dependent MRT fields.

different beam configurations, both with regard to the filtration but also e.g. the wiggler strength of the synchrotron, are likely to be part of irradiation campaigns and (pre-) clinical studies. Similarly, future studies on providing fast MRT dose prediction models will likely be extended to new MRT experiments at additional synchrotron sites (DESY, [114]) and with first non-synchrotron MRT sources (Line-focus x-ray tubes, e.g.[47]).
The naive method of providing an ML model for those scenarios would be to re-create datasets previously used for training for the given beam configuration and re-train the model. This approach was followed during a master's project [115] supervised as part of the work on this thesis, which investigated the transferability of the developed model to the prediction of the dose following irradiations with small proton beams, similar to the proof-of-concept study presented as part of this thesis in Section (4), in which the ML model was trained using the density matrix of a target phantom and the proton energy as input to predict the dose distribution. Figure (7.13), reproduced from [115], shows two exemplary comparisons between the prediction of the ML model and an MC simulation of proton beams with 42 MeV (Figure (7.13a)) and 50 MeV (Figure (7.13b)) incident on two of the rat head phantoms from the dataset used also in the study presented in Section (6). While the study found the ML predictions to follow most general features of the MC simulation, which can e.g. be seen in the apparent agreement between the main characteristics in Figure (7.13b), the study also found the model not to predict doses accurately within 10% for many scenarios. This would lead to a requirement to further investigate potential improvements to the proposed model for the use in proton therapy.
The full re-training method results in very large overhead for each additional ML model and therefore severely limit the applicability of such an approach. Instead, future studies might investigate the use of pre-trained models and transfer learning, allowing to potentially reduce the amount of needed training data by re-using an already trained ML model as base for a subsequent model for different beam configurations (e.g. [116, 117, 118]). This approach is especially promising to transfer a trained model using one filter configuration to a different filter configuration with potentially few data samples but might also reduce the number of required data samples when transferring a trained model to predict the doses in new or different target geometries. Both scenarios are expected to be frequent requirements for any future ML MRT dose prediction model.



(a) (b)

Figure 7.13: Comparison of ML-predicted and MC-simulated energy deposition following the irradiation of a rat head phantom with a proton beam with 42 MeV (left) and 50 MeV (b). Reproduced from [115].

8 Summary and Conclusion

This work presents the first machine learning-based dose prediction method for microbeam radiation therapy, starting from the initial proof-of-concept studies in Section (3) and finally showcasing the successful application in a retrospective preclinical study in Section (6).
Section (3) investigated the possibility of approaching the microbeam radiation therapy (MRT) dose prediction problem with generative adversarial networks (GAN) in a proof-of-concept study with synchrotron broadbeam data. The GAN-based approach produced relatively accurate predictions with at most 1% deviation of the maximum energy deposition in a simplified ellipsoid-based paediatric head phantom for over 96% of the voxels. With a computation time of approximately 100 milliseconds, the used 3D U-Net used as generator was found to be by far the fastest MRT-related dose prediction method reported at the time of writing this thesis. Moreover, when applied to a more complex Computer Tomography (CT)-based skull phantom, the dose within the brain was found to be accurate within 10%. However, the model was only trained on the simpler ellipsoid-based phantom. A more traditional regression training approach was compared to the GAN-based dose prediction model in an additional study using the simplified ellipsoid-based head model as target phantom. The regression model was found to be more accurate than the GAN-based model both by comparing the average test mean-absolute error (MAE) (GAN: $1.87 \pm 0.02 \cdot 10^{-3}$, regression model: $1.35 \pm 0.03 \cdot 10^{-3}$) and the ratio of voxels deviating by at most 3% of the respective voxel-wise dose value (GAN: $66.6 \pm 0.5\%$, regression model: $79.5 \pm 0.4\%$). It should be noted that the latter criterion is different from the previously used relative deviation concerning the overall maximum energy deposition. Following these results, the development focus was put on the regression machine learning (ML) 3D U-Net model for the subsequent studies.
Section (4) demonstrated that the proposed network is transferable to different beam modalities by investigating the performance of the network trained on data obtained using a proton minibeam simulation. A comparison to an adaption of the Dose Transformer (DoTA) [93] transformer-based dose prediction model showed that the proposed 3D-UNet was both faster and more accurate in the learning task posed in the study presented in Section (4). All models accurately predicted the Bragg peak position within 1 mm for nearly all test samples. Due to the steep dose gradients, however, the observed relative deviations between the ML predictions and Monte Carlo (MC) simulations were comparatively large. On the test dataset, the regression 3D U-Net model exhibited at most 3% relative deviation for an average of $61.0 \pm 0.1\%$ of the voxels with at least 1% of the maximum energy deposition. In comparison, the previously developed GAN-based 3D U-Net was found to exhibit at most 3% relative deviation for an average of $33.1 \pm 0.4\%$ of the qualifying voxels and the DoTA model for an average of $46.1 \pm 0.6\%$. A deeper investigation of the local generalization capabilities of the models showed slight overfitting of the regression model with respect to different proton energies, indicating the likely need for a denser sampling of the proton energies in the training data. Similarly, the DoTA model exhibited slight overfitting especially concerning the phantom translations. This finding is in agreement with the expectation that the sequence-based DoTA model exhibits strong results in the depth-wise interpolation between dose distributions resulting from different proton energies but has a disadvantage in generalizing on different geometries, indicating the need for more variable input geometries.
With the proposed 3D U-Net regression model being confirmed as the most promising candidate, Section (5) aimed at including the micrometre sub-structure of MRT into the ML model. For this, a novel scoring method was proposed, named the *macro voxel method*. The method allows scoring MRT fields over a large volume by selectively recording the energy depositions close to precomputed peak and valley locations. This approach was found to be

a requirement to generally enable microbeam superposition as each individual microbeam significantly contributes to valley doses in over 15 mm distance which is very large compared to the microbeam width of 50 μm. As a result, going forward, two different ML models were trained: one for the peak dose distributions and one for the valley dose distributions.
Using this method, a superposition approach to predicting MRT fields was investigated. Assuming a prediction time of 100 ms for a single microbeam over the height of 1 mm, this accumulates to 8 seconds for an 8x8 mm^2 MRT field. In addition, minor systematic deviations of the predictions were found to potentially accumulate due to the superposition. Therefore, the superposition of patches of individual microbeams was found not to be a favourable method for the goal of fast yet reasonably accurate dose predictions. Instead, a direct method for predicting doses for differently sized MRT fields was introduced by passing the shape of the predicted field as additional information to the network. The network was found to predict doses accurately within a few per cent, especially in the centre of the field. In addition the network was observed to over-fit to the field sizes contained in the training data when only a limited number of discrete MRT field sizes were used. This was shown most clearly by finding that the dose distribution predictions of the trained ML model were very close to the dose distributions resulting from MRT field sizes used for training the model, even when providing new field sizes to the model, which it was not trained on. This lack of generalization confirmed the earlier findings that a dense and continuous sampling of the desired prediction parameter space (different geometries, energies, ...) is important for training unbiased ML models for MRT dose prediction.
Following this finding, data augmentation in the form of the rotation and translation of simulated phantoms was implemented when the model was for the first time trained on and applied to preclinically relevant data simulated after a study conducted at the Imaging and Medical Beamline (IMBL) in April 2022 in which 16 rats were treated with an 8x8 mm^2 MRT field after implantation of gliosarcoma cells. The simulation data was obtained using the CT scans of the patient rats. To obtain a large and variable training dataset in accordance with the previous findings without increasing the simulation times to unfeasible lengths, the training data was obtained with high statistical noise of, on average, 15% for the valley doses and 5% for the peak doses. The dose predictions of the trained ML models were smooth even when being trained on noisy MC data. The smooth ML predictions were compared with the noisy MC simulations, expecting an average agreement within one standard uncertainty of 68% under the assumption of Gaussian statistical uncertainty of the voxel-wise dose values. For the training and validation data, an agreement within one standard deviation for, on average, about 64% (training: $64.6 \pm 0.7\%$, validation: $63.7 \pm 0.9\%$) of the voxels was found, indicating that the models were trained to provide nearly unbiased estimates of the voxel-wise dose distributions. The close agreement between training and validation data was found not to indicate overfitting. On the test data, however, the models were found to predict dose values which are in agreement with MC simulation within one standard deviation for, on average, $60.7 \pm 1.7\%$, which is slightly but significantly lower than for the training and validation datasets. Although this indicates some overtraining, an inspection of the worst-case test data predictions was ruled suitable for further development. After the statistical comparison of ML prediction and MC simulation, the predictions of the models were compared with respect to the voxel-wise prediction accuracy with three exemplary low-noise MC simulations of realistic treatment scenarios from the preclinical study. It was found that the ML models were able to predict the valley dose within the original tumour volumes of three test rats with an accuracy of at least 3% for over 95% of the voxels and 100% for the peak dose, although being trained on high-noise simulation data. This confirmed the drawn conclusion that the models were trained to provide unbiased dose predictions even when being trained on high-noise

simulation data. This finding is important because it allows significantly faster development of future ML models following this method. Generally, dose deviations for the low-noise test data were found to be at most 10%, which was ruled to be sufficient for potential future MRT treatment plan optimization purposes.
Section (7) discusses current research directions extending the presented model for its application in more preclinical and, at some point, potentially even clinical studies. An important aspect, e.g., is extending the prediction volume to the out-of-field region, allowing for dose predictions in organs at risk more distant from the MRT field.

This thesis comes to an end here. Within this thesis, multiple milestones towards treatment plan optimization for MRT were reached. The developed ML training methods using high-noise MC simulations pose a valuable finding to accelerate dose prediction models for new irradiation scenarios. The proposed macro voxel MRT scoring is found to be suitable to produce data which sufficiently describes the peak and valley dose distributions in target phantoms without the need for a large number of scoring voxels leading to large files for potential ML training datasets. The prediction speed on the order of 100 milliseconds, compared to about 30 minutes used by the HybridDC model currently used in preclinical planning and up to 20 hours for MC simulations, was found to be very promising for usage in fast dose prediction tasks. The reported accuracy of at least 10% within the studies in this work is ruled to be sufficient during treatment plan optimization. However, many following studies will have to showcase further the reliability, accuracy and versatility of this and similar ML models until they are expected to find their way into future treatment planning programs.

Acknowledgements

There are truly many people who have earned my most honest gratefulness during my time working as a PhD candidate. I could not have asked for more with you being such wonderful supervisors, team and working group members and friends. It has been some wonderful years which would not have been possible without you. If you are reading this and were part of my journey, I really hope you know that I have valued the time with you and everything you have done for and with me very much. Thank you for your supervision, guidance, friendship, laughter, discussions, support, love, so many more things, all the moments full of great memories and reasons to celebrate. Thank you, I could not have done this without you!

References

[1] E. H. Grubbé. Priority in the Therapeutic Use of X-rays. *Radiology*, 21(2):156–162, 1933.

[2] A. Brahme, J. E. Roos, and I. Lax. Solution of an integral equation encountered in rotation therapy. *Physics in Medicine and Biology*, 27:1221–1229, 1982.

[3] K. Otto. Volumetric modulated arc therapy: IMRT in a single gantry arc. *Medical Physics*, 35:310–317, 2008.

[4] R. Mohan. A review of proton therapy – Current status and future directions. *Precision Radiation Oncology*, 6(2):164–176, 2022.

[5] A. B. Jani, A. Su, and M. T. Milano. Intensity-modulated versus conventional pelvic radiotherapy for prostate cancer: Analysis of acute toxicity. *Urology*, 67(1):147–151, 2006.

[6] P. Deman, M. Vautrin, M. Edouard, V. Stupar, L. Bobyk, R. Farion, H. Elleaume, C. Rémy, E. L. Barbier, F. Estve, and J. F. Adam. Monochromatic minibeams radiotherapy: From healthy tissue-sparing effect studies toward first experimental glioma bearing rats therapy. *International Journal of Radiation Oncology Biology Physics*, 82 (4), 2012.

[7] W. Ulmer, J. Pyyry, and W. Kaissl. A 3D photon superposition/convolution algorithm and its foundation on results of Monte Carlo calculations. *Physics in Medicine and Biology*, 50(8):1767–1790, 2005.

[8] L. Hong, M. Goitein, M. Bucciolini, R. Comiskey, B. Gottschalk, S. Rosenthal, C. Serago, and M. Urie. A pencil beam algorithm for proton dose calculations. *Physics in Medicine and Biology*, 41(8):1305–1330, 1996.

[9] Y. Prezado, G. Jouvion, C. Guardiola, W. Gonzalez, M. Juchaux, J. Bergs, C. Nauraye, D. Labiod, L. De Marzi, F. Pouzoulet, A. Patriarca, and R. Dendale. Tumor Control in RG2 Glioma-Bearing Rats: A Comparison Between Proton Minibeam Therapy and Standard Proton Therapy. *International Journal of Radiation Oncology Biology Physics*, 104:266–271, 2019.

[10] E. Engels, N. Li, J. Davis, J. Paino, M. Cameron, A. Dipuglia, S. Vogel, M. Valceski, A. Khochaiche, A. O'Keefe, M. Barnes, A. Cullen, A. Stevenson, S. Guatelli, A. Rosenfeld, M. Lerch, S. Corde, and M. Tehei. Toward personalized synchrotron microbeam radiation therapy. *Scientific Reports*, 10:1–13, 2020.

[11] S. Agostinelli, J. Allison, K. Amako, J. Apostolakis, H. Araujo, P. Arce, M. Asai, D. Axen, S. Banerjee, G. Barrand, F. Behner, L. Bellagamba, J. Boudreau, L. Broglia, A. Brunengo, H. Burkhardt, S. Chauvie, J. Chuma, R. Chytracek, G. Cooperman, G. Cosmo, P. Degtyarenko, A. Dell'Acqua, G. Depaola, D. Dietrich, R. Enami, A. Feliciello, C. Ferguson, H. Fesefeldt, G. Folger, F. Foppiano, A. Forti, S. Garelli, S. Giani, R. Giannitrapani, D. Gibin, J. J. Gomez Cadenas, I. Gonzalez, G. Gracia Abril, G. Greeniaus, W. Greiner, V. Grichine, A. Grossheim, S. Guatelli, P. Gumplinger, R. Hamatsu, K. Hashimoto, H. Hasui, A. Heikkinen, A. Howard, V. Ivanchenko, A. Johnson, F. W. Jones, J. Kallenbach, N. Kanaya, M. Kawabata, Y. Kawabata, M. Kawaguti, S. Kelner, P. Kent, A. Kimura, T. Kodama, R. Kokoulin, M. Kossov,

H. Kurashige, E. Lamanna, T. Lampen, V. Lara, V. Lefebure, F. Lei, M. Liendl, W. Lockman, F. Longo, S. Magni, M. Maire, E. Medernach, K. Minamimoto, P. Mora de Freitas, Y. Morita, K. Murakami, M. Nagamatu, R. Nartallo, P. Nieminen, T. Nishimura, K. Ohtsubo, M. Okamura, S. O'Neale, Y. Oohata, K. Paech, J. Perl, A. Pfeiffer, M. G. Pia, F. Ranjard, A. Rybin, S. Sadilov, E. di Salvo, G. Santin, T. Sasaki, N. Savvas, Y. Sawada, S. Scherer, S. Sei, V. Sirotenko, D. Smith, N. Starkov, H. Stoecker, J. Sulkimo, M. Takahata, S. Tanaka, E. Tcherniaev, E. Safai Tehrani, M. Tropeano, P. Truscott, H. Uno, L. Urban, P. Urban, M. Verderi, A. Walkden, W. Wander, H. Weber, J. P. Wellisch, T. Wenaus, D. C. Williams, D. Wright, T. Yamada, H. Yoshida, and D. Zschiesche. GEANT4 - A simulation toolkit. *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 506:250–303, 2003.

[12] A. Dipuglia, M. Cameron, J. A. Davis, I. M. Cornelius, A. W. Stevenson, A. B. Rosenfeld, M. Petasecca, S. Corde, S. Guatelli, and M. L. F. Lerch. Validation of a Monte Carlo simulation for Microbeam Radiation Therapy on the Imaging and Medical Beamline at the Australian Synchrotron. *Scientific Reports*, 9:1–14, 2019.

[13] D. N. Slatkin, P. Spanne, F. A. Dilmanian, J. O. Gebbers, and J. A. Laissue. Subacute neuropathological effects of microplanar beams of x-rays from a synchrotron wiggler. *Proceedings of the National Academy of Sciences of the United States of America*, 92: 8783–8787, 1995.

[14] C. Von Neubeck, A. Seidlitz, H. H. Kitzler, B. Beuthien-Baumann, and M. Krause. Glioblastoma multiforme: Emerging treatments and stratification markers beyond new drugs. *British Journal of Radiology*, 88(1053), 2015.

[15] A. W. Stevenson, J. C. Crosbie, C. J. Hall, D. Häusermann, J. Livingstone, and J. E. Lye. Quantitative characterization of the X-ray beam at the Australian Synchrotron Imaging and Medical Beamline (IMBL). In *Journal of Synchrotron Radiation*, volume 24, pages 110–141. International Union of Crystallography, 2017.

[16] F. Mentzel, K. Kröninger, M. Lerch, O. Nackenhorst, J. Paino, A. Rosenfeld, A. Saraswati, A. C. Tsoi, J. Weingarten, M. Hagenbuchner, and S. Guatelli. Fast and accurate dose predictions for novel radiotherapy treatments in heterogeneous phantoms using conditional 3d-unet generative adversarial networks. *Medical Physics*, 49: 3389–3404, 2022.

[17] M. Donzelli, E. Bräuer-Krisch, C. Nemoz, T. Brochard, and U. Oelfke. Conformal image-guided microbeam radiation therapy at the ESRF biomedical beamline ID17. *Medical Physics*, 43(6):3157–3167, 2016.

[18] I. Martínez-Rovira, J. Sempau, and Y. Prezado. Development and commissioning of a Monte Carlo photon beam model for the forthcoming clinical trials in microbeam radiation therapy. *Medical Physics*, 39(1):119–131, 2012.

[19] F. Salvat and M. Fern. PENELOPE – A Code System for Monte Carlo Simulation of Electron and Photon Transport. *Nuclear Energy Agency*, page 406, 2015.

[20] S. Keshmiri, S. Brocard, R. Serduc, and J. F. Adam. A high-resolution dose calculation engine for X-ray microbeams radiation therapy. *Medical Physics*, 49(6):3999–4017, 2022.

[21] C. Debus, U. Oelfke, and S. Bartzsch. A point kernel algorithm for microbeam radiation therapy. *Physics in Medicine and Biology*, 62(21):8341–8359, 2017.

[22] M. Donzelli, E. Brauer-Krisch, U. Oelfke, J. J. Wilkens, and S. Bartzsch. Hybrid dose calculation: A dose calculation algorithm for microbeam radiation therapy. *Physics in Medicine and Biology*, 63:45013, 2018.

[23] K. Me. Kraus, J. Winter, Y. Zhang, M. Ahmed, S. E. Combs, J. J. Wilkens, and S. Bartzsch. Treatment Planning Study for Microbeam Radiotherapy Using Clinical Patient Data. *Cancers*, 14(3):685, 2022.

[24] L. R. J. Day, M. Donzelli, P. Pellicioli, L. M.L. Smyth, M. Barnes, S. Bartzsch, and J. C. Crosbie. A commercial treatment planning system with a hybrid dose calculation algorithm for synchrotron radiotherapy trials. *Physics in Medicine and Biology*, 66(5): 055016, 2021.

[25] J. Unkelbach, M. Alber, M. Bangert, R. Bokrantz, T. C. Y. Chan, J. O. Deasy, A. Fredriksson, B. L. Gorissen, M. Van Herk, W. Liu, H. Mahmoudzadeh, O. Nohadani, J. V. Siebers, M. Witte, and H. Xu. Robust radiotherapy planning. *Physics in Medicine and Biology*, 63(22):22TR02, 2018.

[26] S. Lim-Reinders, B. M. Keller, S. Al-Ward, A. Sahgal, and A. Kim. Online Adaptive Radiation Therapy. *Int J Radiat Oncol Biol Phys*, 99(4):994–1003, 2017.

[27] C. M. Poole, L. R. J. Day, P. A. W. Rogers, and J. C. Crosbie. Synchrotron microbeam radiotherapy in a commercially available treatment planning system. *Biomedical Physics & Engineering Express*, 3(2):025001, 2017.

[28] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

[29] V. Kearney, J. W. Chan, T. Wang, A. Perry, M. Descovich, O. Morin, S. S. Yom, and T. D. Solberg. DoseGAN: a generative adversarial network for synthetic dose prediction using attention-gated discrimination and generation. *Scientific Reports*, 10:11073, 2020.

[30] Y. Huang, Y. Pi, K. Ma, X. Miao, S. Fu, H. Chen, H. Wang, H. Gu, Y. Shao, Y. Duan, A. Feng, J. Wang, R. Cai, W. Zhuo, and Z. Xu. Virtual Patient-Specific Quality Assurance of IMRT Using UNet++: Classification, Gamma Passing Rates Prediction, and Dose Difference Prediction. *Frontiers in Oncology*, 11:2798, 2021.

[31] M. Lempart, H. Benedek, M. Nilsson, N. Eliasson, Sven Bäck, P. Munck a.osenschöld, L. E. Olsson, and C. Jamtheim Gustafsson. Volumetric modulated arc therapy dose prediction and deliverable treatment plan generation for prostate cancer patients using a densely connected deep learning model. *Physics and Imaging in Radiation Oncology*, 19:112–119, 2021.

[32] F. Mentzel, M. Barnes, K Kröninger, M. Lerch, O. Nackenhorst, J. Paino, A. Posenfeld, A. Saraswari, A. C. Tsoi, J. Weinfarten, M. Hagenbuchner, and S. Guatelli. A step towards treatment planning for microbeam radiation therapy: fast peak and valley dose predictions with 3d u-nets. accepted for publication in IFMBE Proceedings on the World Congress on Medical Physics and Biomedical Engineering 2022, 2022.

[33] F. Mentzel, K Kröninger, M. Lerch, O. Nackenhorst, A. Posenfeld, A. C. Tsoi, J. Weinfarten, M. Hagenbuchner, and S. Guatelli. Small beams, fast predictions - a comparison of machine learning dose prediction models for proton minibeam therapy. *Medical Physics*, pages 1–11, 2022.

[34] F. Mentzel, J. Paino, M. Barnes, M. Cameron, S. Corde, E. Engels, K. Kröninger, M. Lerch, O. Nackenhorst, A. Rosenfeld, M. Tehei, A. C. Tsoi, S. Vogel, J. Weingarten, M. Hagenbuchner, and S. Guatelli. Accurate and fast deep learning dose prediction for a pre-clinical microbeam radiation therapy study using low-statistics monte carlo simulations. Submitted for publication to Cancers, 2022.

[35] W. Yan, M. K. Khan, X. Wu, C. B. Simone, J. Fan, E. Gressen, X. Zhang, C. L. Limoli, H. Bahig, S. Tubin, and W. F. Mourad. Spatially fractionated radiation therapy: History, present and the future. *Clinical and Translational Radiation Oncology*, 20: 30–38, 2020.

[36] W. Zeman, H. J. Curtis, E. L. Gebhard, and W. Haymaker. Tolerance of Mouse-Brain Tissue to High-Energy Deuterons. *Science*, 130(3391):1760–1761, 1959.

[37] D. N. Slatkin, P. Spanne, F. A. Dilmanian, and M. Sandbora. Microbeam radiation therapy. *Medical Physics*, 19(6):1395–1400, 1992.

[38] J. W. Hopewell and K. R. Trott. Volume effects in radiobiology as applied to radiotherapy. *Radiotherapy and Oncology*, 56(3):283–288, 2000.

[39] M. D. Wright, P. Romanelli, A. Bravin, G. Le Duc, E. Brauer-Krisch, H. Requardt, S. Bartzsch, R. Hlushchuk, J. Laissue, and V. Djonov. Non-conventional Ultra-High Dose Rate (FLASH) Microbeam Radiotherapy Provides Superior Normal Tissue Sparing in Rat Lung Compared to Non-conventional Ultra-High Dose Rate (FLASH) Radiotherapy. *Cureus*, 13(11), 2021.

[40] H. Steel, S. C. Brüningk, C. Box, U. Oelfke, and S. Bartzsch. Quantification of differential response of tumour and normal cells to microbeam radiation in the absence of flash effects. *Cancers*, 13(13):3238, 2021.

[41] F. A. Dilmanian, Y. Qu, L. E. Feinendegen, L. A. Peña, T. Bacarian, F. A. Henn, J. Kalef-Ezra, S. Liu, Z. Zhong, and J. W. McDonald. Tissue-sparing effect of x-ray microplanar beams particularly in the CNS: Is a bystander effect involved? *Experimental Hematology*, 35(4 SUPPL.):69–77, 2007.

[42] A. Bouchet, R. Serduc, J. A. Laissue, and V. Djonov. Effects of microbeam radiation therapy on normal and tumoral blood vessels. *Physica Medica*, 31(6):634–641, 2015.

[43] M. Potez, C. Fernandez-Palomo, A. Bouchet, V. Trappetti, M. Donzelli, M. Krisch, J. Laissue, V. Volarevic, and V. Djonov. Synchrotron Microbeam Radiation Therapy as a New Approach for the Treatment of Radioresistant Melanoma: Potential Underlying Mechanisms. *International Journal of Radiation Oncology Biology Physics*, 105(5): 1126–1136, 2019.

[44] Y. Prezado, G. Jouvion, D. Hardy, A. Patriarca, C. Nauraye, J. Bergs, W. González, C. Guardiola, M. Juchaux, D. Labiod, R. Dendale, L. Jourdain, C. Sebrie, and F. Pouzoulet. Proton minibeam radiation therapy spares normal rat brain: Long-Term Clinical, Radiological and Histopathological Analysis. *Scientific Reports*, 7(1): 1–7, 2017.

[45] N. Matuszak, W. M. Suchorska, P. Milecki, M. Kruszyna-Mochalska, A. Misiarz, J. Pracz, and J. Malicki. FLASH radiotherapy: an emerging approach in radiation therapy. *Reports of Practical Oncology and Radiotherapy*, 27(2):344, 2022.

[46] L. Eling, A. Bouchet, C. Nemoz, V. Djonov, J. Balosso, J. Laissue, E. Bräuer-Krisch, J. F. Adam, and R. Serduc. Ultra high dose rate Synchrotron Microbeam Radiation Therapy. Preclinical evidence in view of a clinical transfer, 2019.

[47] J. Winter, M. Galek, C. Matejcek, J. J. Wilkens, K. Aulenbacher, S. E. Combs, and S. Bartzsch. Clinical microbeam radiation therapy with a compact source: specifications of the line-focus X-ray tube. *Physics and Imaging in Radiation Oncology*, 14:74–81, 2020.

[48] G. Margaritondo and J. Rafelski. The relativistic foundations of synchrotron radiation. *Journal of Synchrotron Radiation*, 24(4):898–901, 2017.

[49] M. J. Berger, J. H. Hubbell, S. M. Seltzer, J. Chang, J. S. Coursey, R. Sukumar, D. S. Zucker, and K. Olsen. Nist standard reference database 8 (xgam). NBSIR 87-3597. NIST, PML, Radiation Physics Division., 2010.

[50] L. M. L. Smyth, L. J. R. Day, K. Woodford, P. A. W. Rogers, J. C. Crosbie, and S. Senthi. Identifying optimal clinical scenarios for synchrotron microbeam radiation therapy: A treatment planning study. *Physica Medica*, 60:111–119, 2019.

[51] J. Apostolakis, M. Asai, A. Bagulya, J. M.C. Brown, H. Burkhardt, N. Chikuma, M. A. Cortes-Giraldo, S. Elles, V. Grichine, S. Guatelli, S. Incerti, V. N. Ivanchenko, J. Jacquemier, O. Kadri, M. Maire, L. Pandola, D. Sawkey, T. Toshito, L. Urban, and T. Yamashita. Progress in geant 4 electromagnetic physics modelling and validation. In *Journal of Physics: Conference Series*, page 072021. IOP Publishing, 2015.

[52] H. Miras, R. Jiménez, Á. Perales, J. A. Terrón, A. Bertolet, A. Ortiz, and J. Macías. Monte Carlo verification of radiotherapy treatments with CloudMC. *Radiation Oncology*, 13(1):1–9, 2018.

[53] L. R. J. Day, P. Pellicioli, F. Gagliardi, M. Barnes, L. M. L. Smyth, D. Butler, J. Livingstone, A. W. Stevenson, J. Lye, C. M. Poole, D. Hausermann, P. A. W. Rogers, and J. C. Crosbie. A Monte Carlo model of synchrotron radiotherapy shows good agreement with experimental dosimetry measurements: Data from the imaging and medical beamline at the Australian Synchrotron. *Physica Medica*, 77:64–74, 2020.

[54] EM Opt4 — PhysicsListGuide 11.0 documentation. https://geant4-userdoc.web.cern.ch/UsersGuides/PhysicsListGuide/html/electromagnetic/Opt4.html. Accessed: 2022-11-24.

[55] J. Spiga, Y. Prezado, E. Bräuer-Krisch, V. Fanti, P. Randaccio, and A. Bravin. The effect of beam polarization in Microbeam Radiation Therapy (MRT): Monte Carlo simulations using Geant4. In *IEEE Nuclear Science Symposium Conference Record*, pages 2170–2173, 2009.

[56] Geant4 Collaboration. Physics Reference Manual Documentation. https://geant4-userdoc.web.cern.ch/UsersGuides/PhysicsReferenceManual/html/index.html. Accessed May 18, 2021, 2017.

[57] A. Ahnesjö, P. Andreo, and A. Brahme. Calculation and application of point spread functions for treatment planning with high energy photon beams. *Acta Oncologica*, 26 (1):49–56, 1987.

[58] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional Networks for Biomedical Image Segmentation. arXiv:1505.04597 [cs.CV], 2015.

[59] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI), LNCS*, 9901:424–432, 2016.

[60] C. Kontaxis, G. H. Bol, J. J.W. Lagendijk, and B. W. Raaymakers. DeepDose: Towards a fast dose calculation engine for radiation therapy using deep learning. *Physics in Medicine and Biology*, 65:075013, 2020.

[61] R. N. Kandalan, D. Nguyen, N. H. Rezaeian, A. M. Barragán-Montero, S. Breedveld, K. Namuduri, S. Jiang, and M. H. Lin. Dose prediction with deep learning for prostate cancer radiation therapy: Model adaptation to different treatment planning practices. *Radiotherapy and Oncology*, 153:228–235, 2020.

[62] Y. Liu, Z. Chen, J. Wang, X. Wang, B. Qu, L. Ma, W. Zhao, G. Zhang, and S. Xu. Dose Prediction Using a Three-Dimensional Convolutional Neural Network for Nasopharyngeal Carcinoma With Tomotherapy. *Frontiers in Oncology*, 11:4674, 2021.

[63] D. D. Pham, M. Lausen, G. Dovletov, S. Serong, S. Landgraeber, M. Jäger, and J. Pauli. U-net in constraint few-shot settings: enforcing few-sample-fitting for faster convergence of u-net for femur segmentation in X-ray. In *Informatik aktuell*, pages 280–285, 2020.

[64] Y. Le Cun, B. Boser, J. S. Denker, D. Henderson, R.E. Howard, W. Hubbard, and L.D. Jackel. Handwritten Digit Recognition with a Back-Propagation Network. *Advances in Neural Information Processing Systems*, pages 396–404, 1990.

[65] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-Decem, pages 770–778. IEEE Computer Society, 2016.

[66] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *32nd International Conference on Machine Learning, ICML 2015*, volume 1, pages 448–456. International Machine Learning Society (IMLS), 2015.

[67] V. Nair and G. E. Hinton. Rectified linear units improve Restricted Boltzmann machines. In *ICML 2010 - Proceedings, 27th International Conference on Machine Learning*, pages 807–814, 2010.

[68] D. Scherer, A. Müller, and S. Behnke. Evaluation of pooling operations in convolutional architectures for object recognition. In K. Diamantaras, W. Duch, and L. S. Iliadis, editors, *Lecture Notes in Computer Science*, volume Part III, pages 92–101. Springer, Berlin, Heidelberg, 2010.

[69] V. Dumoulin and F. Visin. A guide to convolution arithmetic for deep learning. arXiv:1603.07285 [stat.ML], 2016.

[70] D. P. Kingma and J. L. Ba. Adam: A method for stochastic optimization. arxiv:1412.6980, 2015.

[71] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems (NIPS 2014)*, 27:2672–2680, 2014.

[72] M. Paganini, L. De Oliveira, and B. Nachman. CaloGAN: Simulating 3D high energy particle showers in multilayer electromagnetic calorimeters with generative adversarial networks. *Physical Review D*, 97(1):014021, 2018.

[73] D. Sarrut, N. Krah, and J. M. Létang. Generative adversarial networks (GAN) for compact beam source modelling in Monte Carlo simulations. *Physics in Medicine and Biology*, 64(21):215004, 2019.

[74] X. Zhang, Z. Hu, G. Zhang, Y. Zhuang, Y. Wang, and H. Peng. Dose calculation in proton therapy using a discovery cross-domain generative adversarial network (DiscoGAN). *Medical Physics*, 48(5):2646–2660, 2021.

[75] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. arxiv:1701.07875[stat.ML], 2017.

[76] R. L. Dobrushin. Prescribing a System of Random Variables by Conditional Distributions. *Theory of Probability & Its Applications*, 15(3):458–486, 1970.

[77] M. Mirza and S. Osindero. Conditional Generative Adversarial Nets. arXiv:1411.1784 [cs.LG], 2014.

[78] Geant4 Collaboration. Geant4 Material Database - Book For Application Developers 11.0 documentation. https://geant4-userdoc.web.cern.ch/UsersGuides/ ForApplicationDeveloper/html/Appendix/materialNames.html, Date accessed: 30.05.2022, 2017.

[79] D. A. Low, W. B. Harms, S. Mutic, and J. A. Purdy. A technique for the quantitative evaluation of dose distributions. *Medical physics*, 25(5):656–661, 1998.

[80] F. Chollet et al. Keras. `https://keras.io`, 2015.

[81] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Man'e, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Vi'egas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[82] D. Nguyen, X. Jia, D. Sher, M. H. Lin, Z. Iqbal, H. Liu, and S. Jiang. 3D radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected U-net deep learning architecture. *Physics in Medicine and Biology*, 64(6), 2019.

[83] R. Ayachi, M. Afif, Y. Said, and M. Atri. Strided Convolution Instead of Max Pooling for Memory Efficiency of Convolutional Neural Networks. In *Smart Innovation, Systems and Technologies*, volume 146, pages 234–243. Springer Science and Business Media Deutschland GmbH, 2020.

[84] A. Odena, V. Dumoulin, and C. Olah. Deconvolution and checkerboard artifacts. *Distill*, 2016.

[85] P. Ramachandran, B. Zoph, and Q. V. Le. Searching for activation functions. arXiv:1710.05941v2 [cs.NE], 2017.

[86] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

[87] E. Shelhamer, J. Long, and T. Darrell. Fully Convolutional Networks for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39 (4):640–651, 2017.

[88] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of wasserstein GANs. In *Advances in Neural Information Processing Systems*, volume 2017-Decem, pages 5768–5778, 2017.

[89] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, 2018.

[90] Menzel H. G., Clement C., and DeLuca P. ICRP Publication 110. Realistic reference phantoms: an ICRP/ICRU joint effort. A report of adult reference computational phantoms. *Annals of the ICRP*, 39(2):3–5, 2009.

[91] M. J. Large, A. Malaroda, M. Petasecca, A. B. Rosenfeld, and S. Guatelli. Modelling ICRP110 Adult Reference Voxel Phantoms for dosimetric applications: Development of a new Geant4 Advanced Example. In *Journal of Physics: Conference Series*, volume 1662, page 012021. IOP Publishing Ltd, 2020.

[92] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems 2017*, pages 5999–6009, 2017.

[93] O. Pastor-Serrano and Z. Perkó. Learning the Physics of Particle Transport via Transformers. arxiv:2109.03951, 2021.

[94] O. Pastor-Serrano and Z. Perkó. Millisecond speed deep learning based proton dose calculation with Monte Carlo accuracy. *Physics in Medicine and Biology*, 67:105006, 2022.

[95] Y. Prezado and G. R. Fois. Proton-minibeam radiation therapy: a proof of concept. *Medical physics*, 40:031712, 2013.

[96] O. Zlobinskaya, S. Girst, C. Greubel, V. Hable, C. Siebenwirth, D. W. M. Walsh, G. Multhoff, J. J. Wilkens, T. E. Schmid, and G. Dollinger. Reduced side effects by proton microchannel radiotherapy: Study in a human skin model. *Radiation and Environmental Biophysics*, 52:123–133, 2013.

[97] S. Girst, C. Greubel, J. Reindl, C. Siebenwirth, O. Zlobinskaya, D. W. M. Walsh, K. Ilicic, M. Aichler, A. Walch, J. J. Wilkens, G. Multhoff, G. Dollinger, and T. E. Schmid. Proton Minibeam Radiation Therapy Reduces Side Effects in an in Vivo Mouse Ear Model. *International Journal of Radiation Oncology Biology Physics*, 95:234–241, 2016.

[98] P. Lansonneur, H. Mammar, C. Nauraye, A. Patriarca, E. Hierso, R. Dendale, Y. Prezado, and L. De Marzi. First proton minibeam radiation therapy treatment plan evaluation. *Scientific Reports*, 10:1–8, 2020.

[99] C. Lamirault, E. Brisebard, A. Patriarca, M. Juchaux, D. Crepin, D. Labiod, F. Pouzoulet, C. Sebrie, L. Jourdain, M. Le Dudal, D. Hardy, L. De Marzi, R. Dendale, G. Jouvion, and Y. Prezado. Spatially Modulated Proton Minibeams Results in the Same Increase of Lifespan as a Uniform Target Dose Coverage in F98-Glioma-Bearing Rats. *Radiation Research*, 194:715–723, 2020.

[100] H. Bethe. Zur Theorie des Durchgangs schneller Korpuskularstrahlen durch Materie. *Annalen der Physik*, 397(3):325–400, 1930.

[101] F. Bloch. Zur Bremsung rasch bewegter Teilchen beim Durchgang durch Materie. *Annalen der Physik*, 408(3):285–320, 1933.

[102] W.H. Bragg and R. Kleeman. LXXIV. On the ionization curves of radium. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 8(48):726–738, 1904.

[103] P. Arce, D. Bolst, M. C. Bordage, J. M.C. Brown, P. Cirrone, M. A. Cortés-Giraldo, D. Cutajar, G. Cuttone, L. Desorgher, P. Dondero, A. Dotti, B. Faddegon, C. Fedon, S. Guatelli, S. Incerti, V. Ivanchenko, D. Konstantinov, I. Kyriakou, G. Latyshev, A. Le, C. Mancini-Terracciano, M. Maire, A. Mantero, M. Novak, C. Omachi, L. Pandola, A. Perales, Y. Perrot, G. Petringa, J. M. Quesada, J. Ramos-Méndez, F. Romano, A. B. Rosenfeld, L. G. Sarmiento, D. Sakata, T. Sasaki, I. Sechopoulos, E. C. Simpson, T. Toshito, and D. H. Wright. Report on G4-Med, a Geant4 benchmarking system for medical physics applications developed by the Geant4 Medical Simulation Benchmarking Group. *Medical Physics*, 48:19–56, 2021.

[104] S. Saravanan and K. Sudha. GPT-3 Powered System for Content Generation and Transformation. In *2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT)*, pages 514–519. Institute of Electrical and Electronics Engineers (IEEE), 2022.

[105] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, volume 1, pages 4171–4186. Association for Computational Linguistics (ACL), 2019.

[106] Z. Zheng, X. Yue, S. Huang, J. Chen, and A. Birch. Towards making the most of context in neural machine translation. In *IJCAI International Joint Conference on Artificial Intelligence*, volume 2021-Janua, pages 3983–3989, 2020.

[107] D. Khurana, A. Koli, K. Khatter, and S. Singh. Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*, pages 1–32, 2022.

[108] H. Chang, H. Zhang, L. Jiang, C. Liu, and W. T. Freeman. MaskGIT: Masked Generative Image Transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11305–11315. Institute of Electrical and Electronics Engineers (IEEE), 2022.

[109] O. Pastor-Serrano. Dose calculation via transformers. GitHub Repository https://github.com/opaserr/dota, access: 05.06.2022, 2022.

[110] Y. You, J. Li, S. Reddi, J. Hseu, S. Kumar, S. Bhojanapalli, X. Song, J. Demmel, K. Keutzer, and C. Hsieh. Large Batch Optimization for Deep Learning: Training BERT in 76 minutes. arxiv:1904.00962, 2019.

[111] P. Xu, D. Kumar, W. Yang, W. Zi, K. Tang, C. Huang, J. C. K. Cheung, S. J. D. Prince, and Y. Cao. Optimizing deeper transformers on small datasets. In *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 2089–2102. Association for Computational Linguistics (ACL), 2021.

[112] M. Donzelli, U. Oelfke, and E. Brauer-Krisch. Introducing the concept of spiral microbeam radiation therapy (spiralMRT). *Physics in Medicine and Biology*, 64(6), 2019.

[113] M. Schlimmbach. Master thesis at the TU Dortmund University, 2023.

[114] E. Schültke, S. Fiedler, C. Mewes, E. Gargioni, J. Klingenberg, G. Abreu Faria, M. Lerch, M. Petasecca, F. Prehn, M. Wegner, M. Scholz, F. Jaekel, and G. Hildebrandt. The Microbeam Insert at the White Beam Beamline P61A at the Synchrotron PETRA III/DESY: A New Tool for High Dose Rate Irradiation Research. *Cancers*, 14 (20):5137, 2022.

[115] L. Bussmann. Master thesis at the TU Dortmund University, 2022.

[116] L. Zimmermann, E. Faustmann, C. Ramsl, D. Georg, and G. Heilemann. Technical Note: Dose prediction for radiation therapy using feature-based losses and One Cycle Learning. *Medical Physics*, 48(9):5562–5566, 2021.

[117] M. Mashayekhi, I. R. Tapia, A. Balagopal, X. Zhong, A. S. Barkousaraie, R. McBeth, M. H. Lin, S. Jiang, and D. Nguyen. Site-agnostic 3D dose distribution prediction with deep learning neural networks. *Medical Physics*, 49(3):1391–1406, 2022.

[118] E. M. Ambroa, J. Pérez-Alija, and P. Gallego. Convolutional neural network and transfer learning for dose volume histogram prediction for prostate cancer radiotherapy. *Medical Dosimetry*, 46(4):335–341, 2021.