TECHNISCHE UNIVERSITÄT WIEN
Vienna | Austria

HEPHY
INSTITUTE OF HIGH ENERGY PHYSICS

DIPLOMARBEIT

Performance Improvements and Analysis of Image Reconstruction Techniques in Proton Computed Tomography

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Technische Physik

eingereicht von

Benjamin Kirchmayer BSc
Matrikelnummer 01611690

ausgeführt am Atominstitut
der Fakultät für Physik der Technischen Universität Wien
in Zusammenarbeit mit dem Institut für Hochenergiephysik

Betreuung
Betreuer: Assistant Prof. Dipl-Ing. Dr.techn. Albert Hirtl
Mitwirkung: Dipl.-Ing. Stefanie Kaser, BSc
Dipl.-Ing. Dr.techn. Thomas Bergauer

Wien, am 7. Dezember 2021

(Unterschrift Verfasser) (Unterschrift Betreuer)

Abstract

The interest in proton beam therapy for the treatment of tumorous diseases has increased due to the unique energy deposition properties of protons, which allows to reduce dose exposition of healthy tissue. To fully utilize the potential of proton therapy, emphasis has to be put on treatment planning, which includes measuring the energy deposition per unit length (stopping power) of the tissue to be treated. Therefore, a 3D map of the stopping power of the affected tissue must be generated. Currently, this stopping power (SP) map is extrapolated from Hounsfield units (HU) as determined form x-ray computed tomography (CT) images, thus introducing conversion errors. By making use of proton computed tomography (pCT), the SP is measured directly [1].

Difficulties arise as the particle trajectories deviate from straight lines due to multiple Coulomb scattering (MCS). Therefore the algorithms for image reconstruction commonly used in x-ray CT have to be adapted accordingly. For most implementations, this is done by introducing path estimates which account for the non-linear trajectory. Several methods have been introduced, including the most likely path (MLP) and cubic spline (CS) approximation [2]. One of the main drawbacks introduced by the path estimates is the additional computational effort that has to be dealt with. As the reconstruction by itself is already computationally expensive, the need to calculate path estimates for pCT additionally increases computing time. One common approach to improve performance in scientific computing is to parallelize computations within the problem set. Since the computations in pCT image reconstruction are well suited to be executed in parallel, this approach has also gained momentum for application to pCT [3].

In this thesis, the proton paths are estimated by a cubic spline, with the paths being used for generating improved radiographies as input for image reconstruction [4]. The performance of the computations is increased by conducting parts of the calculations in parallel on a GPGPU (general purpose graphic processing unit). As the improved radiographies can be used as input for well established x-ray CT reconstruction methods, the software written in the process is used to expand the open source x-ray CT image reconstruction framework TIGRE [5]. Once implemented, the framework is used to perform image reconstruction for simulation data of several phantoms.

Zusammenfassung

Das Interesse an Protontherapie zur Behandlung maligner Tumorerkrankungen nimmt aufgrund der vorteilhaften Dosisdeposition von Protonen stetig zu. Das charakteristische Dosisprofil geladener Teilchen in Materie ermöglicht eine Bestrahlungstherapie mit verringerter Dosis für den Tumor umgebendes Gewebe. Um das volle Potential der Teilchentherapie zu nutzen, muss die Therapieplanung verbessert werden. Dazu wird die 3D-Verteilung des relativen Bremsvermögens, genannt "Relative Stopping Power" (RSP), des zu bestrahlenden Gewebes aus den Daten einer Computertomographie (CT) berechnet. Durch Verwendung von Protonencomputertomographie (pCT) können Umrechnungsfehler die bei der Konvertierung von Hounsfield units, welche bei der bisherigen Therapieplanung durch X-Ray CT, in relatives Bremsvermögen induziert werden, vermieden werden [1].
Aufgrund mehrfacher Coulombstreuung kann die Trajektorie der Protonen im Rahmen der Bildrekonstruktionsverfahren, welche von der konventionellen X-Ray Computertomographie bekannt sind, nicht ohne Adaptionen verwendet werden. Deshalb werden verschiedene Algorithmen zur Approximation des stochastischen Protonpfades verwendet. In dieser Arbeit wird ein phänomenologischer Ansatz für einen kubischen Spline für die Trajektorieapproximation verwendet [2]. Da in der Protonencomputertomographie für jedes detektierte Proton eine solche Pfadapproximation durchgeführt werden muss, steigt der Bedarf an Rechenressourcen der Bildrekonstruktion erheblich. Um die Laufzeit der Berechnungen in einer für die klinische Anwendung interessanten Größenordnung zu gewährleisten, wird versucht einen Großteil der Berechnungen die im Rahmen der Algorithmen der Protonencomuptertomographie ausgeführt werden, zu parallelisieren [6]. In dieser Arbeit werden die Rechenschritte zur Erstellung verbesserter Radiographien nach Collins-Fekete et al. [4] durch Ausführung der parallelisierbaren Programmteile auf einer Grafikkarte (GPU) beschleunigt. Da die verbesserten Radiographien bereits die Pfadabschätzung beinhalten, können konventionelle Rekonstruktionsalgorithmen, die aus dem X-Ray CT bekannt sind, verwendet werden. Daher werden die in dieser Arbeit entwickelten Programme in Form einer Erweiterung des bereits etablierten X-Ray Rekonstruktionsframework TIGRE [5] implementiert. Anschließend wird die Erweiterung dazu verwendet, Bildrekonstruktionen für verschiedene Phantome durchzuführen.

Contents

1. Introduction

Radiation therapy takes up a crucial part in the treatment of malign cancerous diseases [7]. Since the breakthrough of conventional x-ray therapy in cancer treatment, the dose deposition in healthy tissue surrounding the tumor region has been a significant concern [8]. Proton therapy offers the opportunity to reduce the dose to healthy tissue (see figure 1) due to the characteristic energy deposition of protons, reaching its maximum energy deposition per unit path length at the end of the particle's path through a material (Bragg peak) [9]. Therefore, the range of the protons in the patient has to be known so that the region of maximum energy loss can be aligned with the position of the tumor. To deliver a spatially constant dose to the whole tumor region (spread-out Bragg peak, SOBP), either the proton beam energy is modulated passively, or the energy is adapted while raster scanning tumor slices with a pencil beam [10].

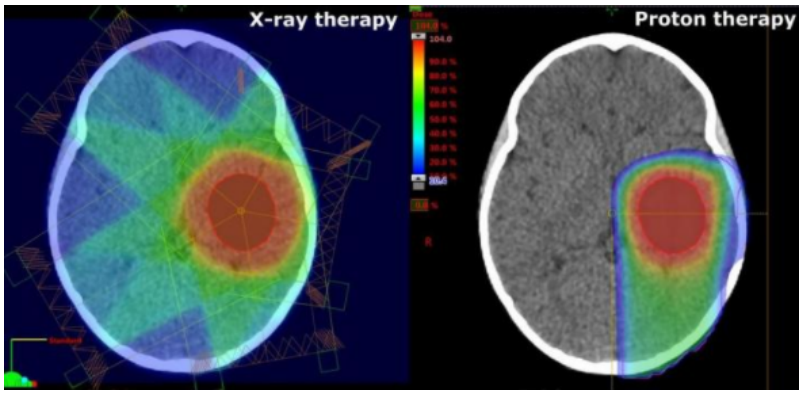


Figure 1. Dose delivery to the tumor region for x-rays (left) and protons (right). To deliver high doses (red) to the tumor region in x-ray therapy, the patient has to be irradiated from several angles. In proton therapy, less irradiation directions are necessary to deliver the same dose to the tumor, which leads to reduced overall dose delivered to healthy tissue [11].

The relative stopping power (RSP), which is the stopping power of a material relative to the stopping power of water at the same energy, of the region to be irradiated has to be known with high precision for accurate treatment planning. Currently, for treatment planning, the RSP is obtained by measuring the mass attenuation coefficient of the material with x-ray computed tomography (CT) and successively converting the attenuation coefficient to RSP via conversion tables [1]. Converting the Hounsfield units obtained from x-ray CT to RSP introduces errors concerning the range of the particles in the material. Proton computed tomography (pCT) aims to reduce these errors by directly measuring the three-dimensional RSP map, instead of extrapolating from Hounsfield units from an x-ray CT. Additionally, while gathering the necessary material characteristics for particle therapy, the overall

dose deposited in the patient can be crucially reduced using pCT [3]. Since the physics of charged particles traversing matter differs from the physics of x-rays, image reconstruction techniques known to be applicable in x-ray CT have to be adapted for the pCT reconstruction problem. A charged particle's path deviates from a straight line between the entry and exit point of the patient due to multiple Coulomb scattering (MCS). Because of that, path estimates have to be introduced to account for the curved path in image reconstruction (see section 2.2.5). Several path estimates were already developed, showing different results in accuracy and computational effort [2]. The data acquisition in pCT is based on tracking single protons by measuring their entry and exit positions as well as their residual energy for several projection angles. Due to a varying number of collisions along the proton path through the sample and the stochastic nature of the energy transfer to atomic electrons, an initially monoenergetic proton beam yields an energy distribution at the rear tracking detector. This is referred to as energy straggling (see section 2.1.5) [12]. To reduce the influence of energy straggling on the measurement, a reasonable number of particles have to be detected. For typical image sizes in medical imaging, e.g., 512 x 512 pixels [13], this leads to a number of tracked particles in the range of a few million to a few ten million particles per projection angle. As path estimation has to be performed for every single particle and several particles contribute to one pixel of the detector, the additional computational effort exceeds the expenditure known from conventional x-ray CT drastically. Since image reconstruction in x-ray CT is already a computationally demanding task [6], pCT requires a special focus on computational implementation to generate high-quality images in a reasonable amount of time.

Structurally, this thesis is subdivided into two parts. In the first part of this thesis, the method for generating improved radiographies developed by Collins-Fekete et al. [4] is implemented using the CUDA framework to accelerate the computation by utilizing Graphics Processor Units (GPU) for data processing. These improved radiographies can be used as input for conventional x-ray reconstruction algorithms. Therefore, the implementation is provided in the form of a toolbox that is intended to be used with the x-ray image reconstruction toolbox TIGRE [5]. In the second part of the thesis, the implemented algorithm is used in combination with the already implemented reconstruction algorithms within the TIGRE framework to perform image reconstructions from simulated data. Khellaf et al. [14] provided a performance comparison of direct image reconstruction techniques for pCT. Their study also featured the method developed by Collins-Fekete et al. [4], which is combined with a filtered-back projection (FBP) algorithm for image reconstruction. In this work, the improved radiographies are used in combination with various iterative reconstruction algorithms and the FBP algorithm. Therefore, Geant4 simulations are performed through GATE for two phantoms, which are based on the ones shown in Khellaf et al. [14]. Measures of spatial resolution and RSP accuracy are calculated for the images generated within this thesis and compared to the findings in their study (and among each other). Possible performance improvements obtained by using iterative

algorithms instead of the direct reconstruction approach are discussed.

Section 2 contains the physical and mathematical background of this thesis. This includes the nature of the interaction of charged particles with matter, the principles of x-ray and proton computed tomography, and path estimation methods and measures to deal with the stochastic proton paths in image reconstruction. Section 3 describes the software frameworks used for Monte Carlo simulations of the pCT imaging setup, the CUDA programming model, the TIGRE toolbox and the reconstruction algorithms used within this thesis, and the theory on the image quality metrics used to evaluate the reconstruction results. In section 4, the computation time of the GPU accelerated code extension is discussed, and reconstructed images are shown and analyzed in terms of spatial resolution and RSP accuracy. Conclusively, section 5 summarizes the findings and gives an outlook on further steps.

2. Physical and mathematical background

The aim of this section is to give an overview of the physical background of proton computed tomography and the techniques that are implemented in the scope of this thesis.

2.1. Interactions of ions with matter

The different properties of proton imaging in comparison to conventional x-ray imaging methods originate from the different interaction of photons and charged particles with matter. The following sections focus on the interaction mechanisms of charged particles.

2.1.1. Interaction mechanisms

To discuss the interaction of charged particles with matter in the scope of proton therapy and proton imaging, it is important to consider the energy range of the interacting particles, as the cross sections of certain interactions are energy dependent. For example, the substructure of protons, consisting of quarks and gluons, does not have to be considered in the clinical energy range. The interaction mechanisms dominating in the energy range of clinical application are shown in figure 2. Figure 2 (a) shows inelastic Coulomb interaction with the atomic electrons, which is the main interaction responsible for the continuous energy loss of the penetrating charged particle. If the incident proton passes close by the atomic nuclei of a material (figure 2 (b)), the proton experiences a repulsive elastic Coulomb interaction. Due to the large mass of the atomic nuclei this kind of interaction is mainly responsible for the lateral deflection of the incident charged particle. When protons pass through a material, they undergo a high number of these interactions (MCS). For thin materials, the resulting angular distribution can be described by the Highland formula [9]. In addition to the already discussed interaction processes, inelastic nuclear interactions between the initial particles and the atomic nuclei have to be considered. Inelastic nuclear interactions lead to reduction of the initial proton flux and generate secondary particles like photons, neutrons or secondary protons. These proton induced nuclear reactions have a small but non-negliable effect on the spatial dose distribution in the patient [9].

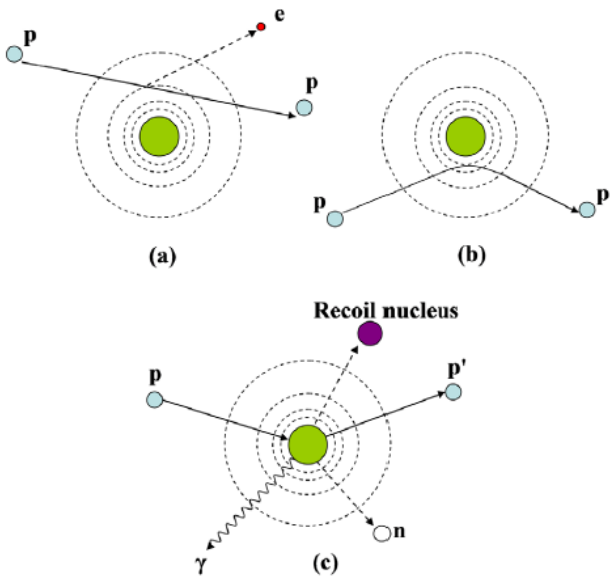


Figure 2. Dominant interactions for charged particles at clinical energies: (a) inelastic Coulomb scattering of the incoming proton with atomic electrons, (b) deflection of the proton trajectory by elastic Coulomb scattering with the nucleus, (c) removal of primary protons with successive liberation of secondary particles via non-elastic nuclear interaction [9].

2.1.2. Stopping power

The mean energy deposition per unit length, referred to as stopping power (SP), can be described with the Bethe-Bloch equation [10]

$$S = -\frac{dE}{dx} = 4\pi^2 \frac{\rho N_A e^4}{mc^2} \frac{Z}{A} \frac{Z_\mathrm{p}^2}{\beta^2} \left[\ln \frac{2m_e c^2 \gamma^2 \beta^2}{I} - \beta^2 - \frac{\delta}{2} - \frac{C}{Z} \right], \tag{2.1}$$

where ρ describes the material density, N_A is Avogadro's number, Z and A the atomic number and the mass number. C represents the shell correction term and δ the density correction term, m is the mass of the electron, Z_p is the charge of the projectile and

$$\gamma = \frac{1}{\sqrt{1-\beta^2}} \text{ with } \beta = \frac{v}{c}, \tag{2.2}$$

where v is the velocity of the proton and c is the speed of light. The energy loss of the projectile is proportional to the inverse square of the velocity and to to the square of the projectile's charge whereas the mass of the projectile does not occur in equation (2.1). For biological tissues the main contributions come from the varying mass density which lead to changes in SP of about three orders of magnitude due to the variety of different tissues in the human body and the ion velocity which alters the SP by a factor of 60 for energies in the range of 1 MeV to 250 MeV [9]. The largest contributions to the uncertainty of the SP are introduced by the mean excitation

energy I [15]. A simplified expression for the SP found by Bragg and Kleeman [16] which delivers accurate results for clinical energies and can be described as

$$-\frac{dE}{dx} = -\frac{E^{1-p}}{p\alpha}, \tag{2.3}$$

with p being a constant the defining the dependence of the energy loss on the particles velocity and α being a material constant.

2.1.3. Bragg curve

Due to the energy loss rate described in the previous section, charged particles deposit most of their energy per unit path length at the end of their path, resulting in the characteristic dose-depth curve, also referred to as Bragg curve (see figure 3). In the following, the characteristic regions of the Bragg peak are described in detail based on [9].

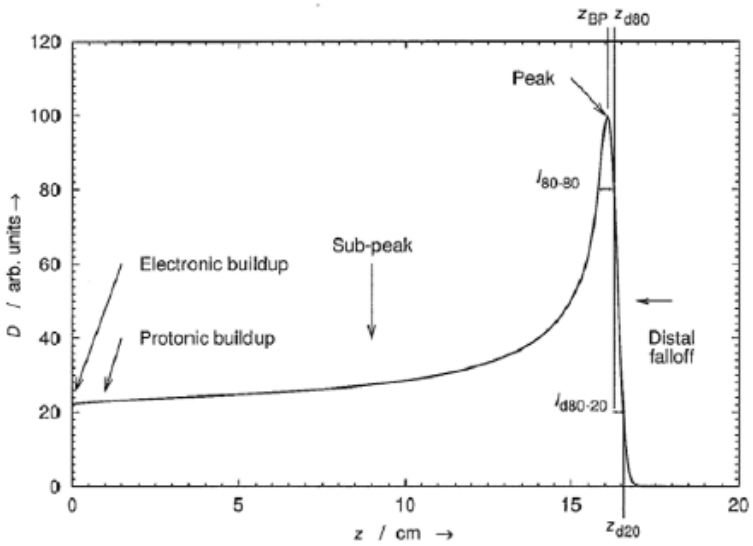


Figure 3. Characteristic depth dose profile for ions in matter [9]. z_{BP} describes the position of the peak of the depth-dose profile (Bragg peak) whereas z_{d80} and z_{d20} describe the distance where the dose deposited is reduced to 20% and 80 % of the peak value respectively.

Electronic buildup region The electronic build up can be observed in a small region near the surface, where the proton beam first enters the absorber. High-energy protons can liberate delta-rays (recoil electrons) which can travel up to a few millimeters. Under certain circumstances, the dose can therefore increase with depth for this region. Effects that influence the charge equilibrium at the surface of the absorber may cause a reduction or a complete disappearance of this effect.

Protonic buildup region The protonic build up can be observed near the surface of the absorber. The absorbed dose increases with depth due to secondary protons originating from non-elastic nuclear interactions.

Sub-peak region The sub-peak region extends from the surface of the absorber close to the depth of the Bragg peak. The dominating physical processes in this region are the SP's dependence of the inverse square of the protons velocity, the attenuation of the proton flux due to nuclear interactions, the liberation of secondary particles and the accumulation of lateral deflections caused by MCS.

Bragg peak region The Bragg peak is the maximum dose deposited near the end of a proton's path. The position of the Bragg peak is governed mainly by the material's SP and the energy straggling of the protons (see section 2.1.5).

Distal fallof region The distal fallof region extends from depths greater than that the Bragg peak's location. The width of this region is not restricted but can be truncated at the depth where the dose falls below 1% of the Bragg peaks dose maxima.

2.1.4. Range

From the SP of a material, the range R, where on average half of the initial protons are at rest, can be calculated

$$R = \int_{E_{\text{in}}}^{0} \frac{dE}{\frac{dE}{dx}}. \tag{2.4}$$

As there are small variations in energy loss for single protons the range given by equation (2.4) is an inherently average quantity. The range can be calculated numerically from equation (2.4)

$$R \approx \sum_{0}^{E_{\text{in}}} \left(\frac{dE'}{dx} \right)^{-1} \Delta E'. \tag{2.5}$$

This approach for calculating the range is not suitable for applications where computation time is crucial. Therefore, the numerical approach is not feasible for application to proton imaging. Consequently, an analytical model describing the charged particle's energy loss has to be found. Using equation (2.3) together with (2.4) yields for the initial energy E_0

$$R(E_0) = \alpha E_0^p. \tag{2.6}$$

The limitations to using this formula is their limited energy intervall of applicability, i.e. 10 - 200 MeVu^{-1} [17]. For samples large enough to completely stop the particle, the Bragg-Kleeman model seems to break down for the biologically relevant interval

below 10 MeVu^{-1} [17]. Donahue et. al [17] propose a model applicable for an energy interval of 1 keVu^{-1} - 450 MeVu^{-1}. Using this model the range is given by

$$R(E) = \frac{u}{\kappa}\left(\alpha E^p + \beta E^q + \frac{h}{g}\left(\exp(-gE) + gE - 1\right)\right) \text{ with } \kappa = 2\pi r_0^2 n_e c^2 z_1^2, \quad (2.7)$$

where h, g, α, β, p and q are fitting parameters, u is the atomic mass number, r_0 is the classical electron radius, m_e is the mass of the electron, c is the speed of light, z_1 is the charge of the projectile and n_e is the electron density of the material.

2.1.5. Energy and range straggling

The accumulation of stochastic differences in energy loss of single particles along their trajectory is referred to as energy straggling. Therefore, not all particles do have the same range which ultimately leads to a change in the shape and the location of the Bragg peak. In figure 4, the energy loss probability density functions (PDF) for protons traversing bulks of water with varying thickness are depicted. A

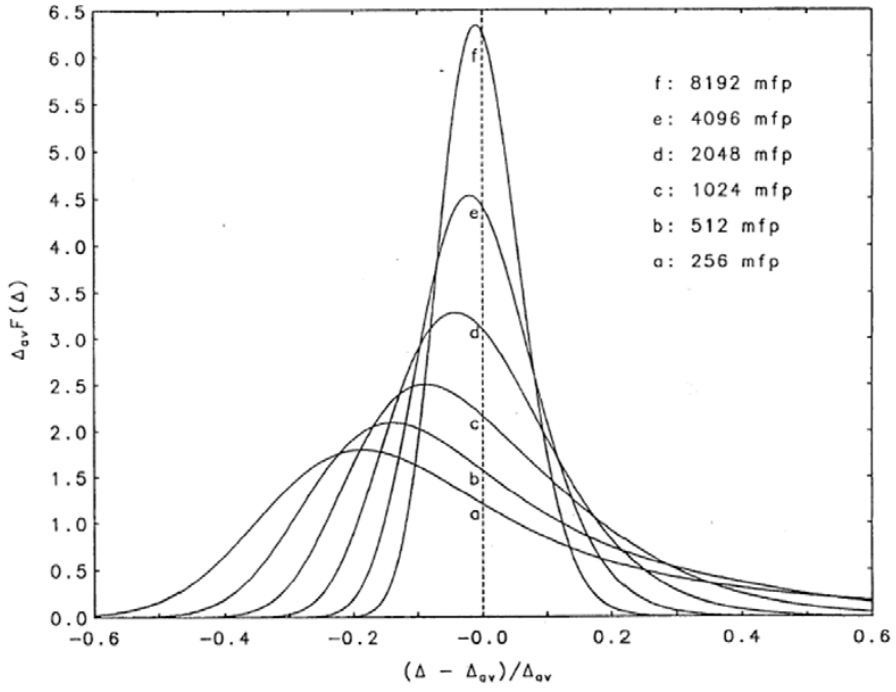


Figure 4. Relative energy loss probability density functions for different water absorber thicknesses. Due to a varying number of interactions along the proton's path and the stochastic nature of the energy loss of protons, caused by inelastic scattering with atomic electrons, a monoenergetic proton beam will show an energy distribution after passing the absorber. For larger sample thicknesses the energy loss distribution is approximately Gaussian, whereas for thin samples the energy distribution shows a tail towards larger energy losses. For thin absorbers the energy loss can be modelled by Vavilov's or Landau's theory [9].

larger sample thickness leads to an increased number of energy loss events along the trajectory of the protons. as the total energy loss is expected to be an accumulation the single stochastic energy loss events, the Gaussian form of the energy loss PDF is expected due to the central limit theorem. Within pCT measurements, besides the energy distribution of the beam and the uncertainty of the energy detector, energy straggling is the main contributor to noise and a limiting factor for the accuracy of the reconstructed RSP [12].

2.1.6. Multiple Coulomb scattering

Protons passing nearby positively charged nuclei are elastically scattered due to the Coulomb interaction. The energy transfer in such interaction is negligible for the traversing particle, but the lateral displacement can be of utter importance. Generally, Coulomb scattering is classified by the number of successive scattering events that are expected. For a single scattering event the interaction is best described by Rutherfords scattering formula [18]. In Multiple Coulomb scattering (MCS), the combined influence of 20 or more such single scattering events has to be accounted for [9]. Since the number of possible scattering events is inherently coupled to the thickness of the observed sample, MCS only appears for samples that are thick enough for this amount of scattering events [9]. The most complete theory known to date is Molieres theory [19], where the probability function of a net deflection angle is modelled. Approximations to Molieres theory done by Highland [20] yield the following expression for the distribution width of the net angle

$$\sigma_\theta = \frac{14.1}{\beta pc} Z_\mathrm{p} \sqrt{\frac{l}{X_0}} \left[1 + 0.038 \cdot \ln\left(\frac{l}{X_0} \right) \right], \tag{2.8}$$

where X_0 is the radiation length, l the thickness of the sample, and p is the momentum of the incident particle. c is the speed of light and β can be calculated by equation (2.2). As a single proton path is affected by successive stochastic scattering, the path of the proton can not be modelled exactly. For image reconstruction in CT it is necessary to have knowledge about the particle's path to perform the inversion of the path integral in the tomographic equation (see section 2.2.4). In fact, MCS is the most limiting factor in terms of spatial resolution in proton imaging [21]. Therefore, MCS has to be considered by introducing path estimates which are discussed in more detail in section 2.2.5.

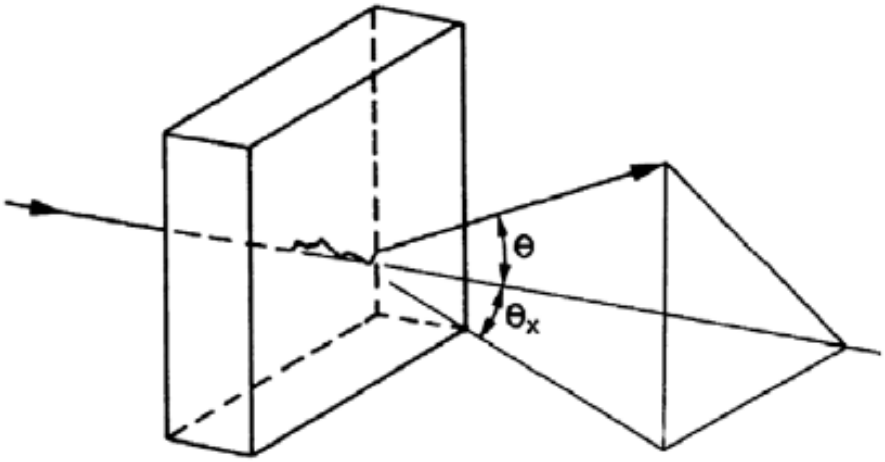


Figure 5. Lateral displacement by net angle θ, due to multiple Coulomb scattering events [9].

2.1.7. Water equivalent thickness

Water is known to behave similar to human tissue in terms of energy loss and MCS. Therefore, it is common to calculate the water equivalent path length (WEPL), describing the travelled distance of a proton in a water reference medium with initial energy E_{in} and residual energy E_{out} [22]. The WEPL can be calculated by

$$\text{WEPL} = \int_{E_{\text{in}}}^{E_{\text{out}}} \frac{dE}{\frac{dE}{dx}\Big|_{\text{H}_2\text{O}}}. \tag{2.9}$$

Another relevant quantity in the scope of proton therapy and imaging is water-equivalent-thickness (WET), which is closely related to WEPL. In contrast to WEPL, WET is a quantity which is assigned as property to the target material. A proton beam with initial energy E_{in} gradually loses energy while propagating through matter and in consequence, leaves the target with the final energy E_{out}. WET is now defined as the length of a water absorber yielding the exact same energy loss as the initial target when penetrated with a proton beam at energy E_{in}. This can be written as

$$\text{WET} = \frac{\rho_{\text{m}}\overline{S_{\text{m}}}}{\rho_{\text{w}}\overline{S_{\text{w}}}} t_{\text{m}}, \tag{2.10}$$

where t_{m} is the thickness of the material, ρ_{m} the density of the material, ρ_{w} the density of water and $\overline{S_{\text{m}}}$ and $\overline{S_{\text{m}}}$ are the respective mean SPs.

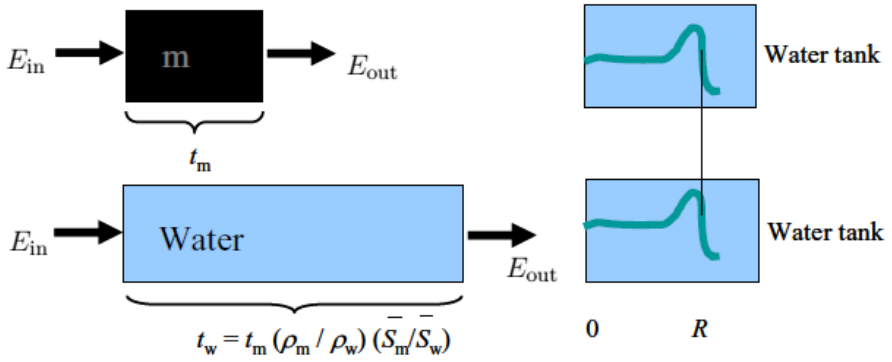


Figure 6. Schematic representation of the concept of WET and how it can be obtained from knowledge of the depth of the proton Bragg curve in a water tank (figure adapted from [9]).

2.2. Principles of proton radiography and proton computed tomography

In this section, the fundamentals of CT with protons are described, with emphasis being laid on the differences to conventional x-ray CT. Furthermore, the image reconstruction algorithms used within this thesis and path estimation methods accounting for MCS of the charged particles are discussed.

2.2.1. X-ray computed tomography

In principal, the aim of CT is to obtain a 3-dimensional image of the human body [23]. For x-ray beams, the quantity defining the image contrast is hereby the mass attenuation coefficient $\mu(x, y)$. In CT, images of cross sections are obtained by assigning a gray value proportional to the mass attenuation coefficient to the pixels of the image. This values are referred to as CT numbers and are given in Hounsfield units (HU)

$$\mathrm{HU} = \frac{\mu_{\mathrm{mat}} - \mu_{\mathrm{water}}}{\mu_{\mathrm{water}}} \times 1000, \tag{2.11}$$

where μ_{mat} and μ_{water} are the attenuation coefficients of water and the sample material. To perform a CT scan, multiple radiographies are obtained by measuring the attenuation of x-rays along a large number of lines from different angles φ (see figure 7) [23]. X-ray attenuation for photons traversing matter along straight lines is governed by Beer-Lambert's law

$$I = I_0 \exp\left(\int \mu(x, y) ds\right), \tag{2.12}$$

where $\int \mu(x, y) ds$ is the integration of the attenuation coefficient of the cross section along a straight line, I is the measured intensity of the x-rays that passed through the sample and I_0 is the intensity obtained from a calibration measurement where

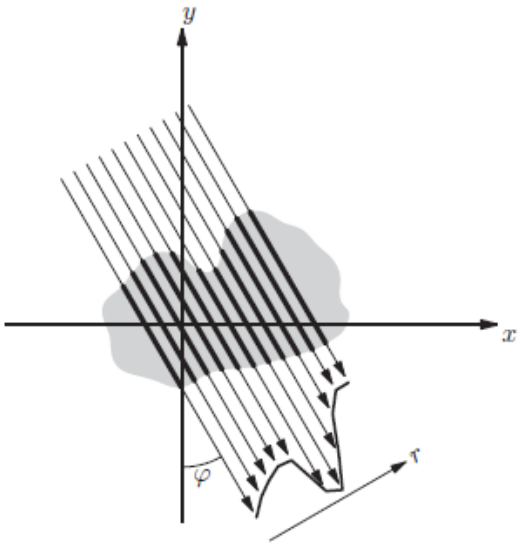


Figure 7. Measurement of the attenuation of x-rays along straight lines for entry angle φ [24].

the sample is removed from the beam's path. Therefore projection values can be obtained by

$$R_\mu(r,\phi) = \int ds\mu(x,y)|_{r=x\cos(\varphi)+y\sin(\varphi)} \approx \ln\left(\frac{I}{I_0}\right). \tag{2.13}$$

Measuring the attenuation of the x-rays therefore corresponds to the summation over the material's mass attenuation coefficient along a straight line. These measurements are repeated from many different angles φ, where each set of the line integrals is called a projection. The collection of the projections for all angles φ are called sinograms and are mathematically described by the Radon transform (equation (2.13)). A sinogram for a simple example is depicted in figure 8. Equation (2.13) is the tomographic

Figure 8. Simple example of the sinogram of a line [24]. The abscissa of the sinogram is the orthogonal distance r to the center and the ordinate is the projection angle φ [24].

equation of x-ray CT. Inverting the tomographic equation is summarized as image reconstruction and is discussed further in section 2.2.4.

2.2.2. Proton computed tomography

In pCT images of cross sections can be obtained in a similar fashion. Cormack first introduced the idea of using the energy loss of charged particles for imaging purposes [25]. In pCT, the residual energy of charged particles is measured while the initial energy characteristics of the beam are known. The counterpart to the Radon transformation in x-ray CT is obtained by integrating equation (2.1)

$$-\int_{E_{\text{in}}}^{E_{\text{out}}} \frac{dE}{S_{\text{water}}(E)} = \int dx \ \text{RSP}(x), \tag{2.14}$$

where E_{in} is the initial energy, E_{out} is the residual energy and $\text{RSP} = S_{\text{material}}/S_{\text{water}}$. S_{water} is the SP of water and S_{material} is the SP of the material the particle propagates through. It has to be mentioned that the RSP is approximately constant over a wide energy range [26]. Therefore, the RSP values obtained from pCT can be used for treatment planning although the beam energies differ for imaging and therapy.

2.2.3. pCT imaging set-up

Since protons undergo multiple scattering events while traversing matter, estimating the proton paths as straight lines as done for photons in x-ray CT is not sufficient to produce images of high spatial resolution. To account for MCS, the imaging setup in pCT is extended in contrast to x-ray CT, to gather vital information for path estimation. A typical pCT imaging setup is depicted in figure 9 [27]. This set-up is called ”single-tracking” setup as it collects information about the particle trajectories for each proton individually. The spatial information is gathered using the position sensitive detectors (PSD) upstream and downstream of the patient/phantom where the detector technology is based on silicium strip detectors (SSD) in the typical single tracking setup [27]. To not only obtain information on the entry and exit particle positions but also the directions, two detector planes per tracker are necessary. For detecting the residual energy, various approaches were already proposed: The residual energy range detector (RERD) can be designed as range telescope [28] which directly measures the WEPL of the particle, as calorimeter based on scintillator technology [29] or using time-of-flight measurements [30]. The data obtained from single-tracking pCT setups are referred to as list-mode data, since information is collected for each proton individually. These list-mode data obtained are used to generate appropriate path estimates that are used in reconstruction to reduce the negative effect of MCS on spatial resolution (see section 2.2.5).

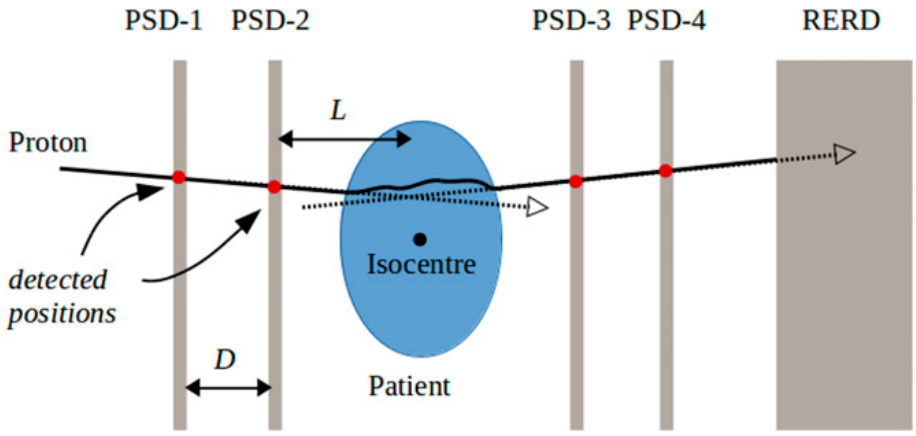


Figure 9. Single tracking pCT system consisting of two position sensitive detectors (PSD) upstream and downstream the patient and a residual energy range detector (RERD) [26].

2.2.4. Image reconstruction

The main objective of image reconstructions algorithms is to provide an approximation to the mass attenuation coefficient $\mu(x, y)$ in x-ray CT or the RSP in pCT, given sinogram data stemming either from measurements or simulations. These reconstruction algorithms can be sub-divided into two categories: direct and iterative reconstruction, which are discussed in the following.

2.2.4.1. Filtered back projection

The most popular algorithm for image reconstruction due to its computational efficiency is the FBP algorithm, which represents a direct reconstruction technique. This algorithm makes use of the projection slice theorem which connects the Fourier transform to the Radon transform [24]. For a function f and projection angle $\varphi = 0$, the projection slice theorem reads as

$$\hat{f}(k, 0) = \widehat{R_f(0, \cdot)}(k). \tag{2.15}$$

The theorem states that taking the slice at $s = 0$ of the 2D Fourier transformation of a real valued function $\hat{f}(k, 0)$ is equivalent to performing the Radon transform of the function $f(x, y)$ and subsequent computation of the Fourier transform, yielding $\widehat{R_f(0, \cdot)}(k)$. Figure 10 illustrates the Fourier slice theorem for a simple function f.

It has been shown in [23], that using equation (2.15), the function of interest $f(x, y)$ can be obtained by

$$f(x, y) = \frac{1}{2\pi} \int_0^{2\pi} \left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \widehat{R_f(\varphi, \cdot)}(r) |r| \exp\left(ir\left(x\cos\varphi + y\sin\varphi\right)\right) dr \right] d\varphi, \tag{2.16}$$

where the term in the brackets is the inverse Fourier transform of $\widehat{R_f(\varphi, \cdot)}(r)|r|$, r is the distance of the point (x, y) from the origin and φ is the projection angle. Since

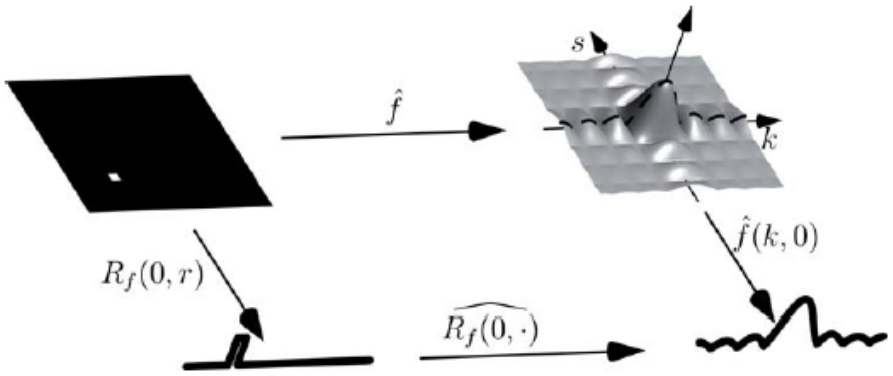


Figure 10. Fourier slice theorem for a simple function f. Taking the slice $s = 0$ of the two-dimensional Fourier transform of f is equivalent to performing the one dimensional Fourier transform of the Radon transform of f [24].

in practice, projections are available only for a finite number of angles,the expression within the brackets has to be evaluated for every angle φ_i yielding an approximation to the original function

$$f(x,y) \approx \frac{1}{2\pi} \sum_{i=1}^{N} R_f(\varphi_i, \cdot) \star |r|(t_i), \tag{2.17}$$

where t_i is the orthogonal distance of the line of the i^{th} projection going through the point (x,y) (see figure 11). The summand $R_f(\varphi_i, \cdot) \star |r|(t_i)$ contributes to all points

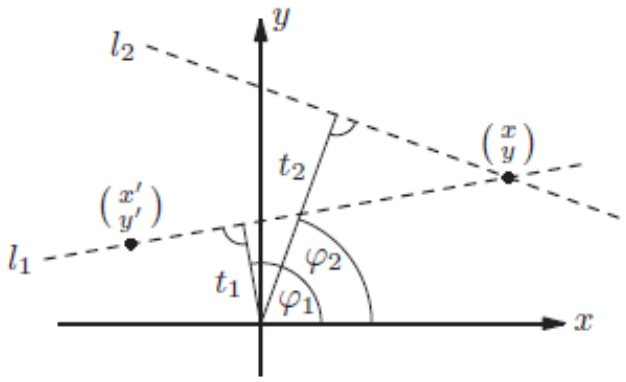


Figure 11. Orthogonal distance t_i of lines l_i intersecting the point (x,y) [24].

intersected by the line l_i with orthogonal distance t_i. To evaluate the expression in (2.17) efficiently, the summands $R_f(\varphi_i, \cdot) \star |r|(t_i)$ are smeared along the lines l_i with corresponding orthogonal distance t_i. This operation is called back projection, which is usually highly optimized in recent image reconstruction frameworks [5, 6, 31].

2.2.4.2. Iterative reconstruction

A different approach for image reconstruction consists of assuming that an image slice consists of an array of unknowns, and then setting up the algebraic equations for the unknowns with respect to the measured projection data. Recording energy measurements for N different protons and discretizing the irradiated volume into M voxels, the tomographic equation for pCT can be rewritten as

$$\sum_{j=0}^{N} = A_{ij}x_j = b_i, \tag{2.18}$$

where A_{ij} is the system matrix, with its elements describing the length of the intercept of the i^{th} proton with the j^{th} pixel. The M-dimensional vector x contains the RSP of the voxels and b is the N-dimensional vector containing the proton's WEPL calculated from the energy measured at the rear detector. The dimensions of the system matrix are large, as the number of proton histories N for all projection angles are found to be in the magnitude of at least a few hundred millions and the number of voxels M is of the order of a few million for typical medical image sizes ($512 \times 512 \times 512$ [13]). The goal of iterative reconstruction is to solve the linear system of equations for x, yielding a three-dimensional map of the volume's RSP. Since the system of equations is overdetermined ($N \gg M$), conventional matrix inversion methods can not be applied to solve for x. Solutions to equation (2.18) can be found in the following form

$$\hat{x_i} = \text{argmin} \parallel A_{ij}x_j - b_i \parallel + R(x_i). \tag{2.19}$$

$R(x_i)$ describes a regularizer function that allows for adding prior knowledge [13]. For these types of systems of linear equations, the iterative projection based methods first introduced by Karczmarz [32] can be utilized for finding a reasonable solution. Algorithms based on Karczmarz [32] are often referred to as Algebraic Reconstruction Techniques (ART). The basic idea is to successively project the solution vector onto the hyperplanes defined by the equations in (2.18). This principle is illustrated for a trivial example in figure 12. The ART algorithm yields

$$x_i^{k+1} = x_i^k + \lambda_k \frac{b - \langle a_i, x^k \rangle}{\parallel a_i \parallel^2} a_i^T, \tag{2.20}$$

where a_i denotes the i^{th} row of the system matrix and $\langle ., . \rangle$ is the scalar product. The ART approach has an significant disadvantage as for the image x^{k+1} the previous image has to be updated i times each iteration. This introduces uncertainties, as pixels altered by the i^{th} equation contribute to the update based on the $(i+1)^{\text{th}}$ equation [33]. This further amplifies the *salt and pepper* noise originally introduced by the inconsistencies in the measured projection data. In practice, the direct computation of the system matrix is not possible since it would not fit into the main memory of a CPU . Therefore, the matrix product operations Ax (forward projection) and $A^T b$

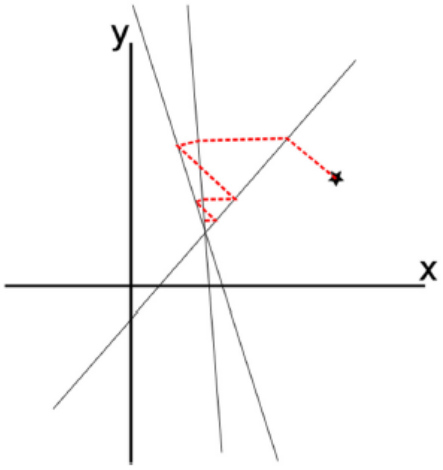


Figure 12. Trivial example illustrating the basic principle behind the projection based ART algorithm [3]. Starting from an arbitrary RSP vector (star), the vector is successively projected onto the hyperplanes (lines). Each line represents a line of a system of linear equations such as (2.18). If the hyperplanes do have an intersection the successive projection converges to the solution vector [32].

(back projection) are implemented instead [13]. The forward projection, Ax, describes the summation of the RSP along the proton paths, whereas the back projection can be understood as smearing a detector pixel's value over the voxels intersected by the path of the detected proton. Using this interpretation, equation (2.20) can be understood as updating the voxel values by backprojecting the difference of the measured projection values and the forward projection of the current voxel values. The iterative algorithms used within this thesis are based on the ART algorithm with certain adoptions introduced to overcome the infirmities of standard ART. In this thesis, the SART [34], the SIRT [34], the OS-SART [35] and the ASD-POCS [36] algorithm are used for image reconstruction.

2.2.5. Path estimates and image reconstruction for pCT

As the path of charged particles traversing through matter is affected by multiple scattering events (see section 2.1.6), assuming their trajectories as straight lines is not sufficient as the path discrepancies limit spatial resolution [37]. Considering that these scattering events are of stochastic nature implies that the path of a proton can not be calculated exactly. However, the proton paths can be estimated with the most common approaches discussed in more detail in the following subsections. In figure 13, the performance of the path estimation formalism is depicted. The most likely path (MLP) formalism shows the smallest deviation from the reference path. However, the cubic spline used for the fit depicted in figure 13 represents a standard implementation, not the optimized approach by Collins-Fekete et al. [38] that is used within this thesis. By improving the the cubic spline estimation the performance gap to the MLP method can be reduced (see section 2.2.5.2) [38]. Additionally, the evaluation of the

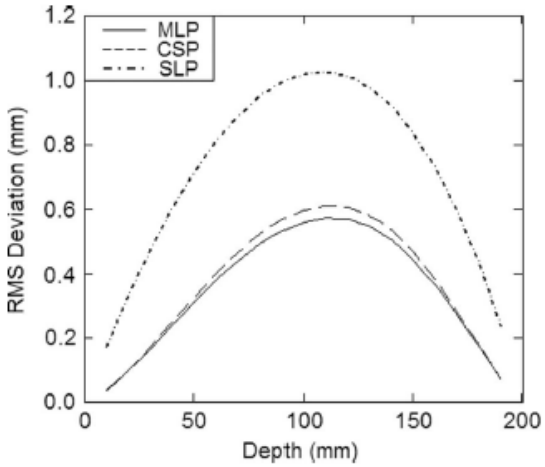


Figure 13. Deviation of the path estimates from the reference trajectory obtained by Monte Carlo simulation [2]. The most likely path (MLP) estimation has the best performance and is the Gold standard of path estimation in pCT.

MLP is computationally expensive. The cubic spline estimation delivers comparable results in terms of estimation performance but can be implemented computationally less demanding. Therefore, the cubic spline estimation delivers a reasonable trade-off between predicting accuracy and computational efficiency [38].

2.2.5.1. Most likely path

Several authors have introduced a mathematical formalism to describe the MLP of protons traversing a single tracking setup [39, 40, 41, 42]. In the following, the approach of Schulte et al. [40] is described in more detail. The variables needed for the derivation of the MLP formalism are depicted in the schematic representation of a pCT scanner geometry shown in figure 14. Since scattering in lateral and vertical direction are assumed to be independent processes [40], the ideas behind the MLP derivation are shown in 2D-space. The variable u describes the depth in direction of the proton beam whereas t describes the lateral displacement. The angle θ describes the direction in the $u-t$ plane at a given depth u. For any depth u the location and the direction can be summarized as parameter vector $y_i = (u_i)$. The MLP formalism provides an expression for the most likely y_i given prior knowledge from the tracking measurements. A proton that enters the tracking volume at y_{in} undergoes MCS and passes y_1 before it leaves the tracking volume at y_{out}. Using Bayes theorem, the joint likelihood $L(y1, y_{\text{out}}|y_{\text{in}})$ of the particle traversing y_1 and y_{out} when the particle entered the volume at y_{in} can be written as

$$L(y1, y_{\text{out}}|y_{\text{in}}) = L(y_{\text{out}}|y_1; y_{\text{in}}) \cdot L(y_1|y_{\text{in}}), \tag{2.21}$$

where $L(y_{\text{out}}|y_1; y_{\text{in}})$ describes the probability of a proton being scattered from y_1 to y_{out} and $L(y_1|y_{\text{in}})$ is the likelihood of being scattered from the entry point y_{in} to y_1. The transport from one point y_i to the successive point y_j is goverened by MCS. As

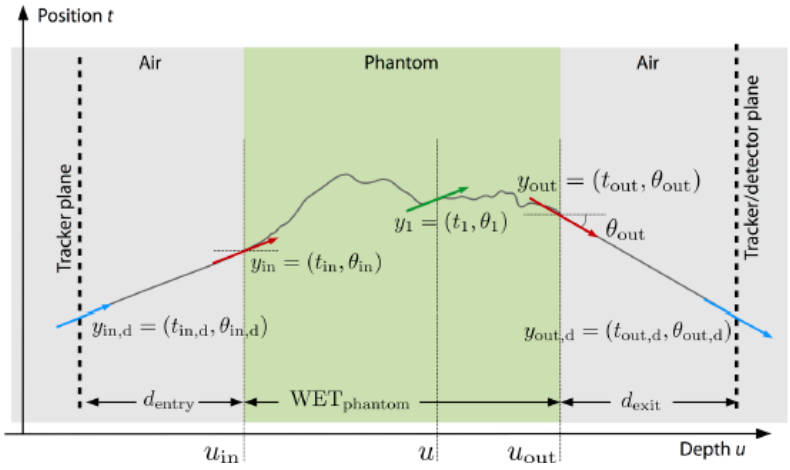


Figure 14. Geometry of the pCT scanner used for the derivation of the MLP formalism [40].

large angle scattering events typically are cut out by applying 3σ cuts [40], MCS can be modelled by the Gaussian approximation to the Fermi-Eyges scattering model [43]. Since the distributions of both, the scattering angle and the lateral deflection are assumed to be Gaussian the likelihood functions can be modelled as bivariate Gaussians respectively.

The MLP is then obtained by solving

$$\nabla_{y_1} L(y_1|y_{\text{in}}) \cdot L(y_{\text{out}}|y_1) \mid_{y_{\text{MLP}}} \stackrel{!}{=} 0. \tag{2.22}$$

2.2.5.2. Cubic spline

Another approach to account for the curved path of the protons is by approximating the trajectory by a cubic spline, using entry and exit information gathered by the single-tracking setup for parametrization. In this thesis, the path is approximated by a cubic spline based on the phenomenological fitting procedure suggested by Collins-Fekete in [38]. The protons trajectories are approximated by

$$\mathbf{S}(t) = (2t^3 - 3t^2 + 1)\mathbf{X}_0 + (t^3 - 2t^2 + t)\mathbf{P}_0 + (-2t^3 + 3t^2)\mathbf{X}_1 + (t^3 - t^2)\mathbf{P}_1, \tag{2.23}$$

where $\mathbf{S}(t)$ represents the position vector as function of the temporal parameter $t \in [0, 1]$. Within this description the first tracking detector is located at $t = 0$. $\mathbf{X}_0$ and $\mathbf{X}_1$ are the entry and the exit position vectors whereas $\mathbf{P}_0$ and $\mathbf{P}_1$ represent optimized entry and exit direction vectors of a proton

$$\begin{aligned} \mathbf{P}_0 &= \hat{\mathbf{P}}_0 \Lambda_0 \mid \mathbf{X}_1 - \mathbf{X}_0 \mid \\ \mathbf{P}_1 &= \hat{\mathbf{P}}_1 \Lambda_1 \mid \mathbf{X}_1 - \mathbf{X}_0 \mid, \end{aligned} \tag{2.24}$$

where Λ_i are optimization parameters and $\hat{\mathbf{P}}_i$ are the normalized direction vectors. Both of these quantities can be measured using two detectors upstream and another

two detectors downstream. Analyzing the spline coefficients in equation (2.23), it can be seen, that for small direction magnitudes $\mathbf{P}_i$ the spline converges to a straight line as the coefficients for higher orders in t tend to cancel out. Since the maximal deflection of the protons trajectory increases with the WET crossed by the particle, [38] suggests to parametrize the direction vectors $\mathbf{P}_0$ and $\mathbf{P}_1$ in dependence of the WET passed. This is done by introducing the optimization factors Λ_1 and Λ_2. The parameters are modelled as functions of the initial energy E_{in} and the WET crossed by the particle. The WET is calculated by integrating equation (2.14) for the measured energy loss, whereas the initial energy is converted to range in water via

$$R_{\text{w}} = aE_{\text{in}}^{b}, \tag{2.25}$$

where $a = 0.00244$ and $b = 1.75$ are the material dependent parameters for water [44, 38]. The optimization parameters Λ_i are modelled as

$$\Lambda_{0,1} = \text{A} + \text{B} \cdot \text{X}^2 \text{ with } X = \frac{R_{\text{w}}}{\text{WEPL}}, \tag{2.26}$$

where A and B are two-dimensional fitting parameters. The first entry of the fitting vectors describes the optimization parameter Λ_0, whereas the second entry describes Λ_1. In [38] this approach has been used to minimize the root-mean-square error of splines deviation from the reference path obtained by Monte Carlo (MC) simulations.The optimized parameter is obtained by [38]

$$\Lambda_{0,1}^{\text{opt}} = \text{A}_{\text{opt}} + \text{B}_{\text{opt}} \cdot \text{X}^2 \text{ with } \text{A}_{\text{opt}} = \begin{pmatrix} 1.01 \\ 0.99 \end{pmatrix} \text{ and } \text{B}_{\text{opt}} = \begin{pmatrix} 0.43 \\ -0.46 \end{pmatrix}. \tag{2.27}$$

2.2.6. Improved radiographies

Several approaches for improving the results of image reconstruction by considering the curved paths of the protons were proposed in the last years [45, 46, 47]. In this thesis, the maximum likelihood approach for the generation of improved radiographies developed by Collins-Fekete et al. [4] is implemented and described in the following. In this method, the space between front detector and rear detector is discretized into channels (see figure 15). $S(t)$ describes the trajectory of an individual proton obtained from a path estimation, $l_{k,n}$ is the length the n^{th} proton travelled inside the k^{th} channel. The approach aims for determining the most likely WET that can be assigned to a channel k, by maximizing the likelihood of the proton energy [4]. Therefore, the error measure $v_{k,n}$, which describes the difference between WET_k of the k^{th} channel and the WEPL_n of the n^{th} is introduced as

$$v_{k,n} = \text{WET}_k - \text{WEPL}_n. \tag{2.28}$$

The error $v_{k,n}$ will yield smaller values for protons that travel longer distances $l_{k,n}$ inside a specific channel k. To retrieve the dependence of $v_{k,n}$ on $l_{k,n}$ the corrected measure $\epsilon_{k,n}$ is proposed as [4]

$$\epsilon_{k,n} = \frac{l_{k,n}}{L_n} \left(\text{WET}_k - \text{WEPL}_n\right), \tag{2.29}$$

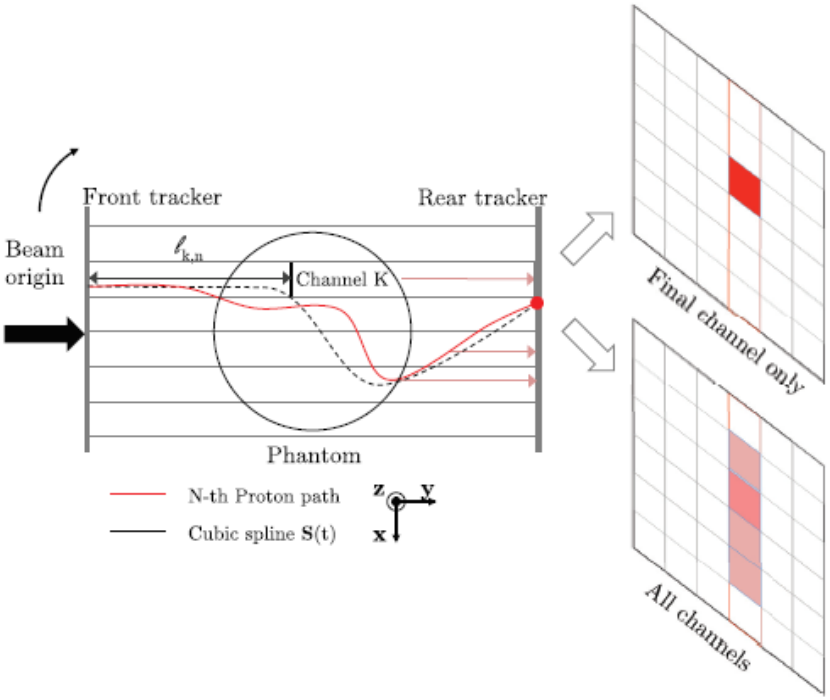


Figure 15. Illustration of the idea behind the Maximum likelihood formalism enabling the generation of improved radiographies [4].

where $l_{k,n}$ is the length a proton n has spent in channel k, and L_n is the length between entry and exit detectors. Furthermore it is assumed, that $\epsilon_{k,n}$ is a single realization of a distribution ϵ_k specific to channel k. Under the assumption that the $\epsilon_{k,n}$ follow a normal distribution

$$\mathcal{N}(\epsilon_k; 0, \sigma_k^2) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left\{ -\left(\frac{l_{k,n}}{\sqrt{2}L_n\sigma_k}\right)^2 (\mathrm{WEPL}_n - \mathrm{WET}_k)^2 \right\}, \tag{2.30}$$

the total likelihood can be written as

$$\mathcal{L} = \prod_{n=1}^{N} \mathcal{N}(\epsilon_{k,n}0, \sigma_{k,n}). \tag{2.31}$$

Maximizing the likelihood for the WET, the most likely WET for a specific channel k can be calculated by

$$\mathrm{WET}_k = \frac{\sum_{n=1}^{N} \left(\frac{l_{k,n}}{L_n}\right)^2 \mathrm{WEPL}_n}{\sum_{n=1}^{N} \left(\frac{l_{k,n}}{L_n}\right)^2}. \tag{2.32}$$

Improved 2D radiographies can be obtained by assigning the WET of the k^{th} channel to a pixel defined by the extension of the channel to the rear detector plane. These

improved radiographies can be used in combination with x-ray image reconstruction techniques. This is due to the fact, that the contribution of a proton travelling on a curved path is assumed to consist of several weighted straight line contributions to multiple pixels (see figure 15).

3. Materials and methods

3.1. Software frameworks

3.1.1. ROOT

ROOT [48] is a C++ based software framework mainly used in particle physics and data analysis. It was developed to match the increasing demand for a data analysis tool being capable of dealing with growing data sets from simulations and experiments. ROOT provides many data structures, like the TTree, which perfectly match the requirements for a data container from simulation frameworks like Geant4. The structures allow storing physical properties on an event-to-event basis. This thesis used the software framework for storing Geant4/GATE simulation data in the sophisticated data structures embedded within the framework [48].

3.1.2. CUDA

CUDA, short for *Compute Unified Device Architecture*, is a parallel computing platform or programming model developed by NVIDIA that leverages the parallel computing engine in NVIDIA's Graphics Processing Units (GPU) [49]. CUDA enables the utilization of typical programming languages, like C/C++ or Fortran, to program NVIDIA GPUs. In general, GPUs provide a much higher throughput and memory bandwidth than CPUs. While the CPU is designed to excel at executing sequential operations as fast as possible, the GPU enables the execution of thousands of operations in parallel [49]. Hardware-wise, this is accomplished by using a large number of processing units, referred to as streaming processors. These processors are grouped into multiprocessors, which, compared to CPUs, have relatively small clock speeds. However, devoting many processors to a task that would require a large amount of queuing if executed on a CPU can drastically improve the overall performance.

Programs designed with CUDA are executed in a Single Instruction Multiple Data (SIMD) fashion [6]. A special function called *kernel* is launched with N copies on the GPU, where each copy is referred to as a thread. These threads are grouped into blocks, which are then enumerated and executed on the multiprocessors. In the CUDA programming model, the threads are enumerated with a three-dimensional thread index based on the illustration in figure 18. The index i of a vectorized loop can be assigned to a thread index

$$i = \text{blockIdx.x} \cdot \text{blockdim.x} + \text{threadIdx.x}. \tag{3.1}$$

Since all threads contained within one block are executed on a single multiprocessor, the size of the block is physically limited to 1024 threads per block [49].

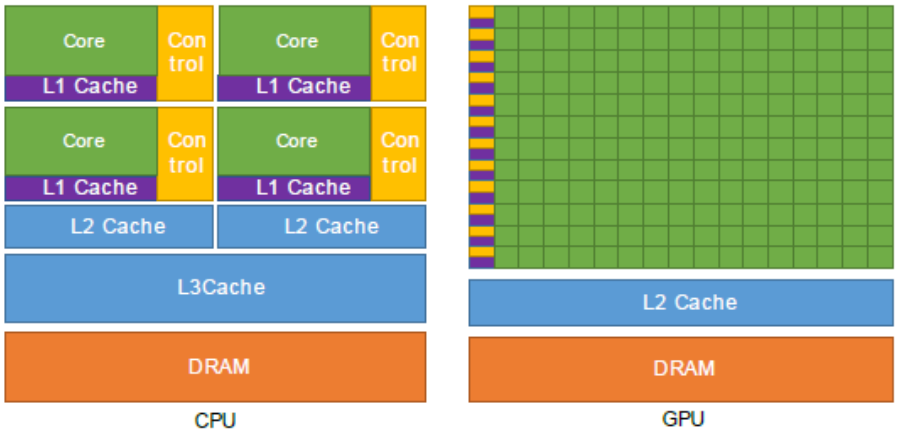


Figure 16. Compared to the CPU architecture, on a GPU, more transistors are devoted to data processing than flow control [49].

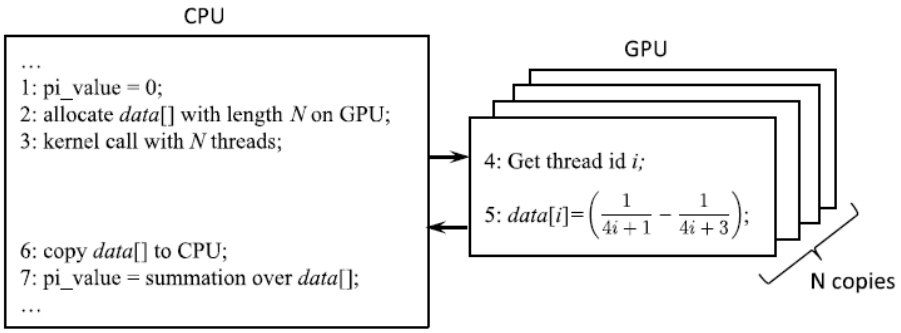


Figure 17. Comparison of the execution of a vectorizable loop on the CPU and the GPU [6].

In figure 17, the difference between executing a program on a CPU and a GPU is highlighted. Each iteration of the loop is executed within one of the threads on the GPU. Another difference to conventional CPU computation can be found in the memory hierarchy of the GPU. Each thread has access to local memory, which is realized in the form of registers on a single processing unit. Additionally, each streaming processor has access to shared memory, which allows for sharing memory within a thread block [49]. Furthermore, there is global memory which is analogous to the CPUs RAM on the GPU. Data stored in global memory can be accessed from all processors and is therefore available for every thread.

Performance limitations Although GPU computing promises a significant boost in computing performance, the executed program must be modeled appropriately to fully utilize the GPU acceleration. If the problem set can not be converted into an algorithm compatible with the SIMD model, executing the program on a GPU will yield no benefit [6]. However, even if the algorithm solving the problem set of interest is compliant with the SIMD model, GPU code can be inefficient if designed

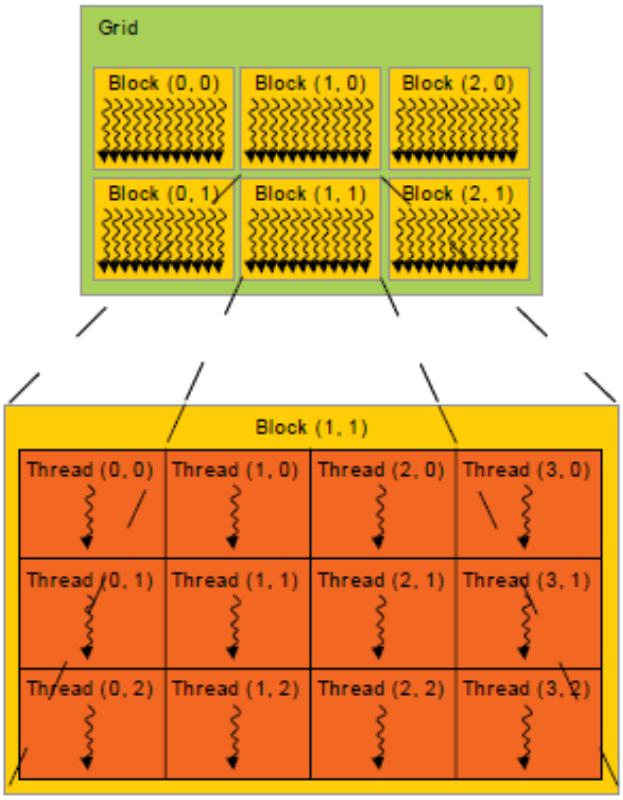


Figure 18. Organization of thread indices [49].

improperly. One key element limiting efficiency is thread divergence. If the kernel implementing the algorithm needs many control operations for execution, resulting in program sections of different lengths, efficiency decreases. This is due to the fact that GPUs are not designed to excel in control flow operations and some threads may have to wait for the thread with the largest execution time caused by thread divergence [6]. The memory hierarchy of GPUs also has an impact on program efficiency. Depending on the type of memory the kernel needs to access, the program encounters different memory bandwidths. The data transfer between the CPUs RAM and the GPUs global memory, as well as the individual threads access, are governed by low memory band width. These operations have to be minimized in efficient GPU code. Although GPU code optimization can improve performance by an order of magnitude or more [6], the underlying problem set the code is developed for ultimately determines the efficiency of the optimization efforts.

3.1.3. TIGRE

TIGRE, short for *Tomographic Iterative GPU-based Reconstruction* toolbox, is a Matlab/CUDA based toolbox originally developed for image reconstruction of cone-beam x-ray CT. The toolbox is created to provide an open-source tool for iterative

image reconstruction that emphasizes flexibility in the geometry of typical x-ray CT setups. The geometry is based on the illustration in figure 19. The distance between

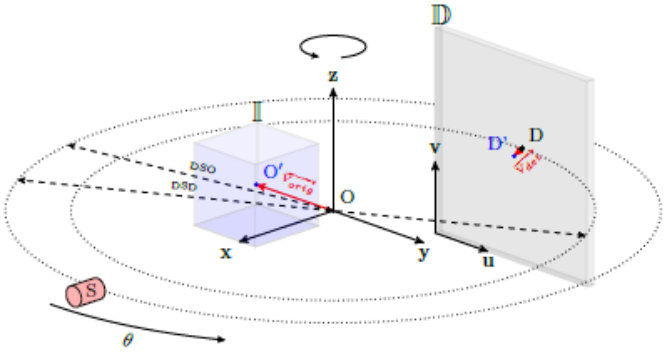


Figure 19. Geometry of the imaging setup in TIGRE [5].

particle source and origin (center of rotation) is referred to as distance source-origin (DSO) in the plot. The TIGRE toolbox also supports helical CT and parallel beam geometries in addition to the cone-beam geometry shown in figure 19. As all the elements shown in figure 19 can be shifted and rotated, the toolbox provides maximal flexibility for the imaging setup, allowing to consider uncertainties introduced by mechanical misalignment in the imaging setup.

TIGRE's design philosophy is based on reducing the gap between image reconstruction research and the end-users of tomographic imaging [13]. This philosophy is recapitulated in the layered software architecture of the TIGRE toolbox, which can be seen in figure 20. The top layer contains the user scripts and the algorithms which

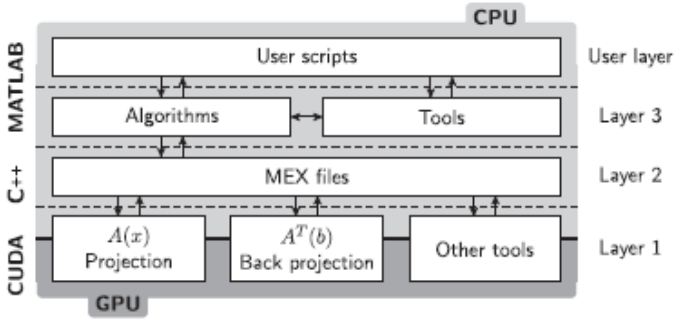


Figure 20. Abstraction of the software architecture of the TIGRE framework [5].

are implemented in Matlab. This allows the end-user to use a high-level scripting language to perform image reconstructions within a few lines of code. To boost image reconstruction performance in terms of computing time, the computationally demanding operations,i.e., forward projection and back projection, are implemented within the CUDA framework. This allows for launching computationally demanding tasks on the GPU, which reduces the computation time immensely [5]. The interface

between the CUDA implementations and the Matlab top layer is provided by Mex functions. Due to TIGRE's modular structure, implementing new CUDA accelerated modules is possible without changing the source code of the toolbox. Therefore, an extension to the TIGRE toolbox for pCT can be implemented. TIGRE introduces four main reconstruction families: the filtered-back projection family, the SART family, the Kryvlov subspace method family, and the total variation regularization family [5]. Except for the Kryvlov subspace method family, algorithms from each family are used for image reconstruction within this thesis. For a more detailed description of the algorithms used within this thesis, see section 2.2.4.2.

3.1.4. Monte Carlo simulations with Geant4 and GATE

Researchers rely on Monte Carlo (MC) particle transport calculations to simulate the effects of the stochastic deflection events a proton undergoes while traveling through matter. MC simulations provide a useful tool to investigate the influence of different tracking geometries. The influences of different tracking and energy measurement setups and the influence of the choice of the reconstruction algorithm on image quality can be studied without having access to a pCT scanner [37]. Particle transport simulations are fundamentally based on MC methods. Usually, in MC methods, mathematical or physical problems are solved by setting up stochastic models using random numbers that fit the problem of interest, e.g., particle transport [50]. A particle traversing matter is influenced by a large number of events $E_1, ...E_n$, such as scattering or particle decay. Depending on the type of particle, different scattering or decay events E_i have to be taken into account. Assigning probabilities $p_1, ..., p_n$ to the events $E_1, ..., E_n$ affecting the particle, a particular event E_i can be represented by a random number $\xi \leqslant 1$ if

$$p_1 + ... + p_{i-1} \leqslant \xi < p_1 + ... + p_i \text{ with } \sum_{i=0}^{n} p_i = 1. \tag{3.2}$$

The probability of the events p_i is governed by the underlying physics and can be obtained by

$$p_i = \frac{\sigma_i}{\sigma_t} \text{ with } \sigma_t = \sum_{i=1}^{n} \sigma_i, \tag{3.3}$$

where σ_i is the cross section of the interaction assigned to event E_i and σ_t is the total cross section. Based on this stochastic model, the physics affecting the particle transport, such as electromagnetic or hadronic interactions can be modelled. In the following sections, the MC software frameworks used within this thesis are described.

3.1.4.1. GEANT4

Geant4 [51] is a C++ software toolkit developed by research groups at CERN and KEK (High Energy Accelerator Research Organization). At the core, the simulation toolkit consists of a collection of physics models that describe the interaction of particles with matter over a wide energy range. Geant4 provides several physics lists yielding different levels of abstraction to interactions mechanisms, depending on particle type and energy range. By choosing an appropriate physics list, the user decides in which level of detail certain aspects of physics are modeled. The physics list mostly used for pCT studies is the *QGSPC_BIC_HP* model [37]. In this thesis, the *QGSPC_BIC* model is used instead. The *QGSPC_BIC_HP* model only extends the former by modelling thermal neutrons more accurately, whose influence on this work is neglected [52].

In the following, elements of the Geant4 simulation framework which are crucial to understand the simulation toolchain are highlighted. Emphasis will not be laid on software details but the software's representation of physical actions. Therefore, the transport of a single particle from its origin until the end of tracking and the representation of the underlying processes by the software are summarized in the following. The event (G4Event) is the main unit of the simulation. It provides an abstract interface to the physics event generators for the generation of primary particles. At the beginning of every event, primary tracks (G4Track) are generated. Tracks resemble "snapshots" of particles that contain current information like energy, position, and momentum. In Geant4, particle propagation is invoked via processes (G4Process). Each interaction that is contained in the physics model of the particle propagated proposes a physical step length, where the smallest of these steps is used to propagate the particle. A process is the only instance that can change the properties of a track. Necessary transient information of the updated track is contained in the step class (G4Step), which contains spatial information for the starting and the endpoint as well as the change in track properties (momentum, energy, etc...). The physical information of the step instance can be accessed via the hit class. The hit class represents the snapshot of the data of the G4Step class if the step occurred in a volume that is assigned with a so-called "sensitive detector".

Generally, all volumes where particle information has to be extracted are therefore modeled as sensitive detectors. For example, in a pCT simulation, this would correspond to tracking detectors and the calorimeter. For each event, the sensitive detectors obtain a collection of hits. By processing the collection of hits at the end of each event, single-particle tracking information can be extracted at the location of the tracking detectors. In this way, the list-mode data gathered from a single tracking pCT imaging setup can be generated.

3.1.4.2. GATE

GATE, short for Geant4 Application for Tomographic Emission, encapsulates Geant4 libraries to achieve a modular, scripted simulation toolkit optimized for the field of

nuclear medicine [53]. MC simulations can be performed in an intuitive matter as an extension to the native Geant4 command line interpreter is embedded within GATE. The simulation of whole imaging setups can be conducted by supplying GATE with a few macros containing GATE command line instructions. The simulation of a pCT imaging setup in GATE can be done in eight steps [53]:

I. Setting up the geometry of the pCT scanner
II. The geometry of the phantom has to be described.
III. The data acquisition system has to be described. In GATE, "actors" are tools for collecting information during the simulation. GATE provides several actors that allow to extract different types of information, for example, dose deposition.
IV. The time parameters have to be set. In GATE, the time-slice for which the simulation is assumed to be static can be set. The simulation's geometry is updated after each time-slice, which enables simulating dynamic objects, i.e., moving detector geometries.
V. The data output format has to be chosen. GATE provides several output formats for storing the collected data with the default formats being ROOT [48] and ASCII.
VI. The physics governing he simulation has to be determined via choosing an appropriate physics list.
VII. The properties of the particle source have to defined in terms of shape and energy profile.
VIII. The level of verbosity has to be set.

Although some flexibility is lost by switching from Geant4 to GATE for particle transport simulation, GATE's simplicity imposes an advantage for the purpose of simulating a pCT imaging set-up.

3.2. Simulation of the experimental setup

The experimental set-ups which are simulated with GATE are based on the common single particle tracking pCT scanner discussed in section 2.2.3. However, idealized conditions are assumed:

- infinite energy resolution of the tracking detectors
- infinite spatial resolution of the tracking detectors
- monoenergetic parallel proton beam
- energy measurement is conducted within the last tracking detector which cancels uncertainties which would be introduced through additional scattering or energy straggling within the calorimeter. Additionally, the result is independent of the technology used for the energy measurement.

For all simulations conducted within this thesis, 3 σ cuts are applied to eliminate the contribution of secondary particles on the path estimation [40]. In the scope of the simulation's post-processing, the WEPL for each proton is calculated by computing

the difference of range in water given the initial energy and the measured residual energy using equation (2.7).

The geometry of the pCT imaging setup for all simulations conducted within this thesis is based on figure 21. The particle source is chosen to be monoenergetic with a rectangular cross section. The physics model used for the simulation is the *QGSP_BIC* model which is recommended for medical physics. The distances between

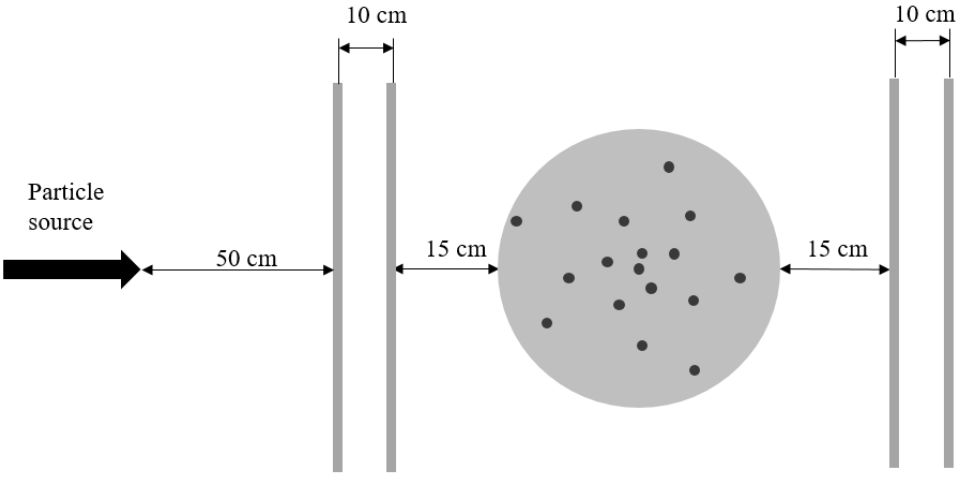


Figure 21. Geometry of the pCT scanner simulated within this thesis. The spiral phantom depicted in this figure acts as place holder for arbitrary samples.

the components of the experimental setup are the same for all simulations, with the only variable being the size of the particular phantom and the initial energy.

3.2.1. Phantoms

In this section, the phantoms which are used in the MC simulations of the pCT setup are described.

Water phantom

Since the pCT TIGRE toolbox has to be calibrated for biological samples of varying thickness, Geant4 simulations of water blocks, approximating human tissue, are conducted (see section 3.3.2). As the range of protons in water depends on the initial energy, adoptions to the initial energy have to be done to ensure that a large amount of particles can reach the rear detectors if the sample size is increased. The initial energies chosen for the respective water absorber thickness t_{water} are depicted in table 1.

t_{water} / cm	E_{in} / MeV
0	150
10	200
20	250
30	300
40	350
50	400

Table 1. Parameters for pCT calibration experiment.
t_{water}... phantom thickness, E_{in}... initial energy

Gammex phantom

The Gammex 467 tissue characterization phantom consists of a cylinder made of water which has a diameter of 33 cm and several tissue inserts embedded within the cylinder. As Khellaf et al. [14] use the same phantom for characterizing the RSP accuracy, the results of this thesis and their findings can be compared. Therefore, the material of the tissue inserts is chosen to match their setup. However, not all of the tissue materials used by Khellaf et al. [14] could be identified. Some tissues are replaced by standard tissues contained in the GATE material data base. In the simulations using the Gammex 467 as phantom, the initial beam energy is set to 250 MeV to ensure that the protons reach the rear detector. The phantom can be seen in figure 22 with the material abbreviations described in table 2.

Abbreviation	**Material**	**RSP$_{\text{ref}}$**
AP6	adipose tissue	0.976
SB3	spinal bone	1.341
BR12	brain	1.043
MB7	marrow bone	1.377
LM1	lung moby	0.298
BR	breast	1.034
LV	liver	1.065
MB5	marrow bone	1.285
LN5	lung	0.505
LM2	lung moby	0.298
CT	cortical bone	1.094
RB	rib bone	1.723

Table 2. Abbreviations and RSP reference values RSP$_{\text{ref}}$ for the Gammex 476 tissue inserts obtained from the R80 measurement (see section 3.5.1).

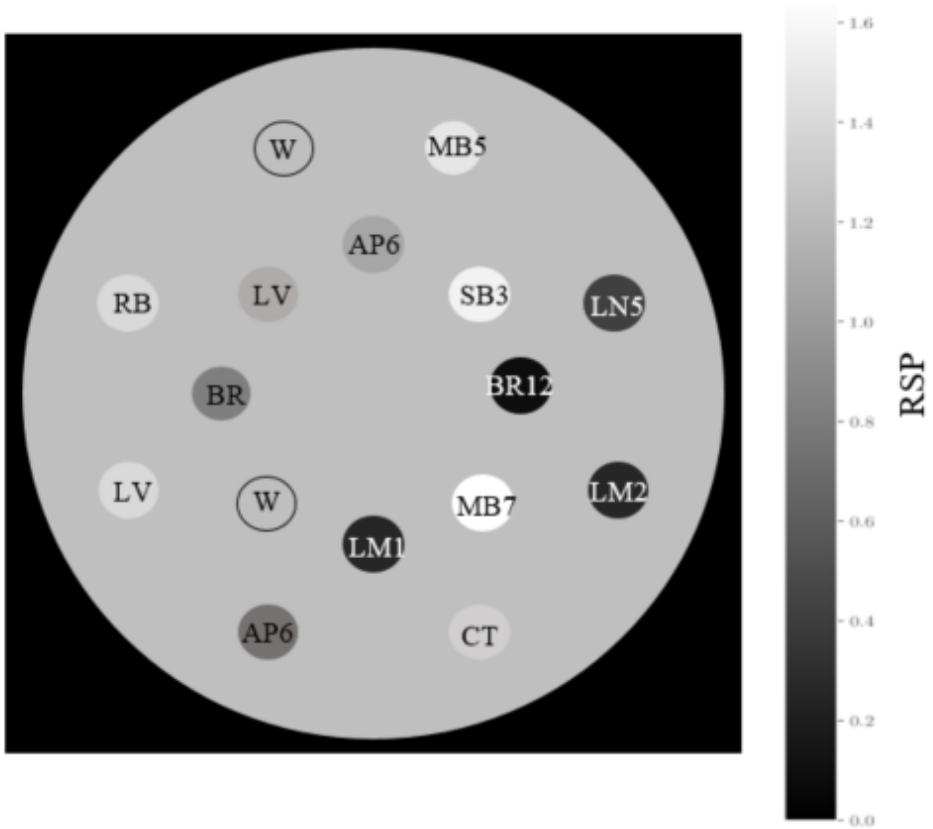


Figure 22. The Gammex 467 tissue characterization phantom consists of a 33 cm solid water disk approximating the average size of a pelvis [54]. Sixteen holes with a diameter of 28 mm enable inserting various interchangeable tissue or other biological probes. Within the simulations the solid water is approximated by a cylinder of liquid water. The inner inserts are located on a circle with a diameter of 11 cm whereas the outer insertions reside on a circle with a diameter of 21 cm.

Spiral phantom

To characterize the spatial resolution, a phantom typically used in literature is the spiral phantom [14, 45, 46] as it allows to evaluate the spatial resolution for several inserts in dependence of the distance to the phantom center. The phantom shown in figure 23 consists of a cylinder of water with a diameter of 14 cm, representing human tissue, and several aluminium inserts. Compared to the spiral phantom used in the literature, the phantom in this thesis consists of a reduced set of inserts. The furthest insert of each spiral arm is removed to downsize the phantom, improving the computing time of the MC simulation. The energy of the proton beam is set to 150 MeV in the MC simulations.

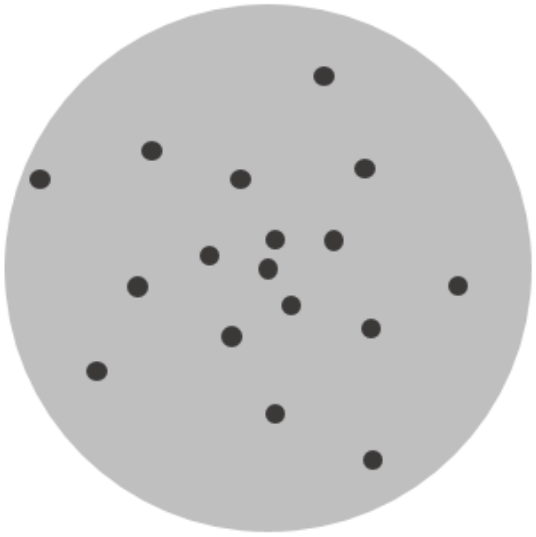


Figure 23. The phantom consists of a cylinder of water with 2 mm diameter aluminium inserts. Since there are inserts all over the cross section, the influence of the depth-dependent uncertainty of the path estimation can be evaluated.

3.3. TIGRE pCT toolbox

The field of proton therapy and proton imaging is multidisciplinary and therefore contains research from several disciplines such as medicine, mathematics, physics, and biology. To make pCT and proton imaging more applicable and more accessible for researchers with different expertise in scientific computing, it is necessary to reduce the programming overhead. Therefore, the development of a user-friendly software framework is beneficial. For this purpose, a collection of CUDA-based files with a Matlab header (following the structure of the TIGRE toolbox [5]), further referred to as pCT toolbox, is created [5]. The methods embedded in the pCT toolbox generate improved proton radiographies from data obtained either by MC simulations or by experiment.

For the development of the pCT toolbox the same design principles as in the TIGRE framework are applied. The application code is implemented in Matlab, and the computationally demanding but parallelizable calculations are implemented in C++ utilizing CUDA. Within the algorithms used in the pCT toolbox, the path estimation of the proton paths is done by cubic spline estimation based on [38], whereas the method for generating improved radiographies from the known paths of the single particles is based on the findings in Collins-Fekete et al. [4] (see section 2.2.6). To calculate the weighting factors needed to generate the improved radiographies, the intercept of the proton path and the channels the proton has crossed have to be found. Compared to the MLP, the cubic spline approach is superior in terms of computational efficiency. Additionally, using the optimized cubic spline estimation proposed by Collins-Fekete et al. [38] further reduces the path

estimation performance discrepancies. For a single particle trajectory, the cubic spline coefficients have to be calculated only once, whereas the MLP has to be re-evaluated at every step. Additionally, finding the intercepts of the estimated path with the channel boundaries can be done analytically for a third-order polynomial (see section A). Since the channel intercepts have to be found for every individual proton, GPU acceleration is utilized to overcome the run time limitations imposed by the cumulative computational efforts (see section 3.3.1).

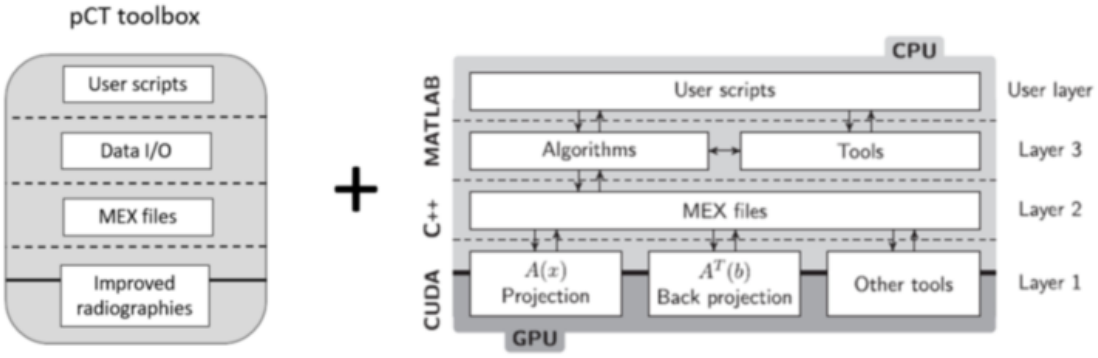


Figure 24. Implementation of the complete pCT image reconstruction toolchain based on improved radiographies [4] and the TIGRE reconstruction framework [5].

3.3.1. Performance considerations

To reduce the influence of noise introduced to the measurements by energy straggling and to keep the therapeutic dose deposited in the probe as low as possible, the rate of detected particles per mm^3 irradiated volume is suggested to be 100 [22, 55]. This results in $\mathcal{O}(100)$ detected particles per mm^2 detector area for pCT. Considering the spatial dimensions relevant in particle therapy, particle numbers are in the range of a few million up to a few ten million particles per projection, depending on the size of the area of interest. Since complex calculations have to be done for all these particles to generate an improved radiography, performance considerations have to be made to deal with the computational effort.

As these calculations can be done for each particle separately, the generation of proton radiographies based on the methods used within this thesis is perfectly fitted for being executed in parallel as the program's structure fits into the SIMD scheme (see section 3.1.2). The algorithm has already been implemented in [56] for sequential execution. Since calculating the intercepts of particle trajectories and the channel boundaries is done independently, GPU acceleration can be utilized by distributing the workload on the GPU's streaming processors. Since NVIDIA currently has the largest share in providing add-in GPGPU's for scientific computing [57], the CUDA parallel computing platform has been chosen for algorithm development. It enables the utilization of the hardware provided by NVIDIA. For the implementation of the

algorithm to be compatible with the CUDA framework, some key changes in the algorithmic design had to be made, which are described in detail in the following.

Dynamic memory allocation The most significant drawback in algorithmic design for this specific purpose is, that it is not possible to allocate memory within the threads local memory dynamically [58]. This makes it impossible to use data structures that are able to grow dynamically, like the vector of the C++ standard library. Therefore, only fixed-size arrays can be used as container to store the intercepts of the protons with the channel boundaries. Consequently, the size of the array already has to be defined at compile time. As shown in figure 25 for a 30 cm water block as phantom (see section 3.2.1), the number of channel intercepts ranges up to approximately 200 intercepts varying from particle to particle.

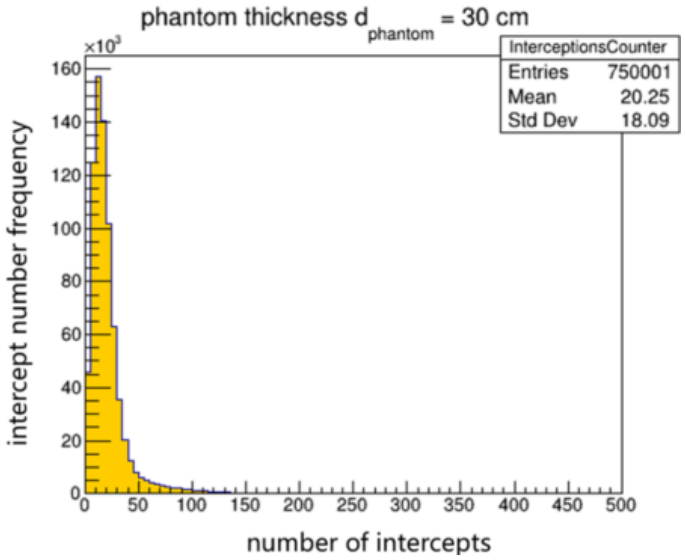


Figure 25. Distribution of the intercept count for a pCT simulation of a 30 cm water block.

Since the parametrized cubic spline in equation (2.23) is described by independent x and y coordinates, the channel intercepts can be found by solving a cubic equation for the x and the y component, independently. The decision logic is kept as simple as possible to avoid thread divergence. Therefore, the intercepts for each coordinate dimension are stored in the container array in sequence. For the weighting factor $\frac{l_{k,n}}{L_n}$ (see equation (2.29)) to be calculated correctly, the array has to be sorted beforehand due to the simplified decision logic. As the performance of sorting an array strongly depends on the chosen algorithm, emphasis has been put on selecting the optimal method. Figure 25 shows that if the array size is chosen to contain the intercepts of every proton, most initialized arrays are filled very sparsely. To reduce the influence of the sorting process on the runtime, the best performance is expected by using a partial sort algorithm.

3.3.2. Intercept number calibration

The pCT toolbox developed in the scope of this thesis is designed in the same structure as the TIGRE toolbox since its results are intended to be used with this framework. To reduce the scientific computing overhead, the toolbox should be provided so that no source code alteration from the user is necessary. Nevertheless, the size of the array that stores the intercepts has to be defined before compilation due to the dependence of the number of channel intercepts on the thickness of the sample, the chosen channel size and the distance between the inner detectors (see figure 26). As the the focus of this thesis is on improving the performance of image reconstruction for a clinical context, the pCT toolbox is calibrated for biological tissues. Therefore, MC simulations of several water blocks with varying thickness (see section 3.2.1) are performed, since water can be used to approximate biological tissues. For each water block, the distribution of the number of intercepts is extracted.

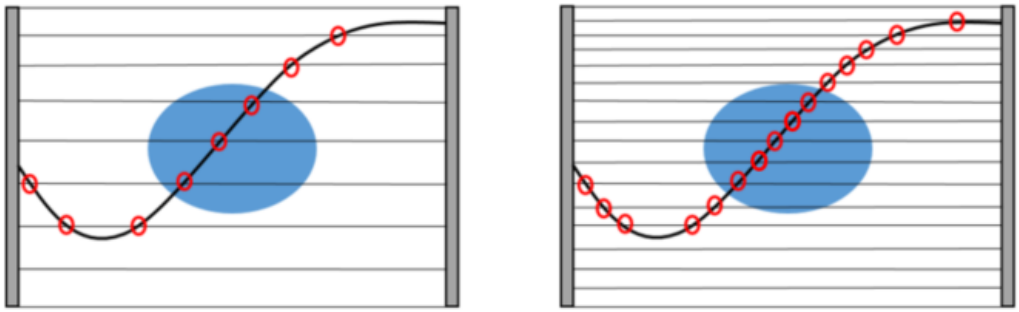


Figure 26. Sketch of pCT measurement with discretized channels showcasing the influence of the chosen pixel size on number of channel intercepts (red) with the channels being restricted by the inner tracking detectors (grey).

Depending on the thickness of the phantom, the number of intercepts is expected to increase. The water block thicknesses, as depicted in table 1, are chosen to cover a range representing typical patient body parts (head, torso). The dependence of the intercept number on the sample thickness is used to calculate a calibration curve that is used to define the optimal array size for the user-specified sample thickness. As the optimized array size also prevents the user from choosing a data container too large for the problem set, positive effects on runtime are expected but not discussed in this thesis.

3.4. Reconstruction parameters

In the second part of this thesis, the improved proton radiographies generated with the pCT toolbox are used to perform image reconstruction on two phantoms. In Khellaf et al. [14], several direct reconstruction algorithms for pCT are compared. Improved proton radiographies as they are implemented in the pCT toolbox in this work, used in combination with the FBP reconstruction algorithm are also analyzed

within the comparison. Compared to the other algorithms mentioned in the paper, the method delivers sub-par outcomes in terms of RSP accuracy and spatial resolution for idealized detectors (infinite energy resolution, infinite spatial resolution). However, for a realistic detector setup, the difference between the methods is negligible [14]. In this thesis, the performance of the pCT toolbox is compared to the performance of the same method [4] as evaluated in Khellaf et al. [14]. Additionally, the influence of choosing iterative algorithms for image reconstruction is observed. Therefore, RSP accuracy and spatial resolution of the central slices of the reconstructed images are compared to Khellaf et al. [14]. The irradiated phantoms and further experimental details are discussed in the respective sections 3.5.1 and 3.5.2. The algorithms used within this thesis and the necessary information about their parameters are listed for the reconstruction of the respective phantom in table 3. The number of iterations for the SIRT, the SART and the OS-SART algorithm is chosen heuristically. This is done by increasing the number of iterations until no further performance improvements can be observed. The ASD-POCS algorithm is only supplied with an upper threshold regarding the iteration count, as the algorithm is aborted if convergence is reached. In table 4, the most important parameters for the image reconstruction based on

Algorithms	**Parameters**	
	Spiral phantom	Gammex phantom
FBP	-	-
SART	iterations: 25	iterations: 25
OS-SART	iterations: 50 / blocksize: 20	iterations: 50 / blocksize: 20
SIRT	iterations: 100	iterations: 100
ASD-POCS	iterations: 15	iterations: 15

Table 3. Algorithmic parameters.

the improved radiographies generated for the spiral and the Gammex phantom are displayed. Additionally, the reconstruction parameters of Khellaf et al. [14] are shown to highlight the differences to the reconstruction conditions of this thesis. It has to be mentioned that in this thesis the channel sizes for the optimized radiographies (see section 2.2.6) are chosen to be equal to the pixel size of the image reconstruction.

	Spiral phantom		**Gammex phantom**	
	this thesis	Khellaf et al. [14]	this thesis	Khellaf et al. [14]
image size [pixels]	600×600	1000×1000	700×700	800×800
pixel size [mm]	0.25	0.25	0.5	0.5
N_{proj}	720	720	90	720

Table 4. The parameters shown in this table, describe the data fed to the reconstruction algorithms, with N_{proj} being the number of improved radiographies used for the reconstruction.

3.5. Image quality metrics

The purpose of this section is to describe the image quality metrics used to evaluate the performance of the image reconstruction results obtained within this thesis. The main performance objectives are the spatial resolution and the RSP accuracy. These quantities are the most crucial parameters concerning the range uncertainty in particle therapy based on treatment planning with pCT [59].

3.5.1. RSP accuracy

To evaluate the performance of the image reconstruction in terms of spatial resolution, reference RSP values for comparison are vital. Therefore, additional Geant4 simulations, where the water equivalent thickness is extracted from R80 measurements are conducted [60]. R80 describes the range in the falloff region of the Bragg peak where the dose has reduced to 80 % of its maximum amplitude. These ranges are measured in water blocks downstream of material blocks consisting of the same material as the tissue inserts of the Gammex phantom. A reference range without absorber is obtained by replacing the material block by air. Subsequent subtraction of the obtained ranges for the tissue materials from the reference range enables the computation of the WET of the respective tissue. From the WET of the material, the respective RSP can be calculated by dividing the WET by the thickness of the material block used in the simulation [61].

For the evaluation of the RSP accuracy the relative RSP error for a region of interest (ROI) is calculated as

$$\epsilon^k_{\text{rel}} = \frac{\text{RSP}^k_{\text{mean}} - \text{RSP}^k_{\text{ref}}}{\text{RSP}^k_{\text{ref}}} \text{ with } \text{RSP}^k_{\text{mean}} = \frac{1}{N}\sum_{i=1}^{N} \text{RSP}^k_i, \tag{3.4}$$

where RSP^k_i is the RSP value of the i^{th} pixel in the ROI of insert k, N is the number of pixels embedded in the ROI, $\text{RSP}^k_{\text{mean}}$ is the mean RSP of the k^{th} insert and $\text{RSP}^k_{\text{ref}}$ is the reference RSP from the material of the k^{th} insert.

3.5.2. Spatial resolution

One of the most comprehensive metrics used to evaluate spatial resolution is the modulated transfer function (MTF) [62]. The MTF describes how well an imaging system transfers contrast across spatial-frequencies [63]. Fundamentally, images can be viewed as a summation of impulses, e.g. scaled and shifted delta functions [64]. The imaging process itself is referred to as a linear system. In this description, "system" refers to any process that produces an output signal in response to an input signal. The output image $G(x, y)$ of an imaging process is given by the convolution of the input image $f(x, y)$ with the impulse response, also called point spread function (PSF), of the system

$$G(x, y) = \int\int f(x - x\prime y - y\prime)\text{PSF}(x\prime, y\prime)dx'dy\prime, \tag{3.5}$$

with the additional constraint

$$\int\int \mathrm{PSF}(x,y)dxdy = 1. \tag{3.6}$$

The PSF describes how much the contrast information contained within one point of the source material $f(x_0, y_0)$ is spread spatially through the imaging process (see figure 27). As the spreading described by the PSF contains all the information on the spatial resolution, obtaining the PSF from the image would be sufficient for performance characterization. The MTF ultimately is defined as the modulus of the 2D- Fourier transformation of the PSF. However, the PSF is experimentally hardly obtainable since it would require the imaging of an infinitesimal small object [63]. A more convenient method is to is to calculate the MTF with the use of the edge spread function (ESF). The ESF can be obtained from sharp contrast edges embedded in the image of interest. In this thesis, these edges are represented by the inserts of the analyzed phantoms (see section 3.1.4). The connection of the ESF to the PSF is given by

$$\mathrm{LSF}(x) = \int \mathrm{PSF}(x,y)dy = \frac{\partial}{\partial x}\mathrm{ESF}(x). \tag{3.7}$$

As the line spread function (LSF) is the Radon transform of the PSF, the MTF can be obtained using the Fourier slice theorem

$$\mathrm{MTF}(f) = \frac{1}{N}\int \frac{\partial}{\partial x}\mathrm{ESF}(x)\exp\left(-2\pi i f x\right)dx, \tag{3.8}$$

where N is the normalization of the LSF and f is the spatial frequency.

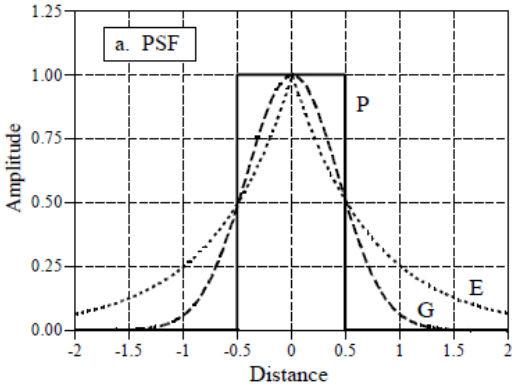


Figure 27. Illustration of the PSF in 2 dimensions with the PSF being depicted in various functional forms (G - Gaussian, E- Exponential, P - pillbox) [64].

In this thesis, the spatial resolution of the reconstructed image is evaluated by calculating the MTF for the inserts of the spiral phantom. Therefore, the ESF is calculated by sorting the pixel's RSP values according to the distance of the pixel from the center of the insert, as depicted in figure 28. Since the ESF is expected to look like a smeared step function with the step being located at the radius of the

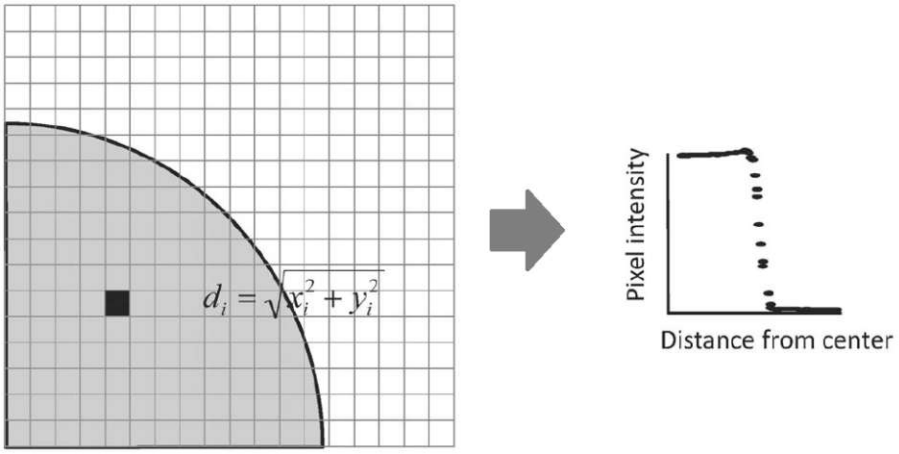


Figure 28. Calculating the edge spread function based on [63].

insert, a function mimicking the ESF has to be found so that a nonlinear fit can be performed. In this thesis, the ansatz as suggested in [14] is used:

$$\mathrm{ESF}(x) \propto a \cdot \mathrm{erf}\left(-\left(\frac{x-b}{c}\right)^2\right) + d. \tag{3.9}$$

The measure to characterize spatial frequency in the literature is the $f_{10\%}$ value which is the spatial frequency where the MTF evaluates to 10 % of its amplitude. Since the MTF is the Fourier transform of the derivative of the ESF the spatial frequency $f_{10\%}$ can be calculated as

$$f_{10\%} = \frac{1}{\pi \cdot c}\sqrt{\frac{\ln(10)}{2}}. \tag{3.10}$$

Therefore, $f_{10\%}$ solely depends on the parameter c, obtained from the nonlinear fit of the ESF of the reconstructed image. $f_{10\%}$ is calculated for all but the two innermost inserts, considering the RSP values in circular regions of interest with diameters of 10 mm around the inserts as outlined in figure 23. Consequently, the spatial resolution of the image can be calculated in dependence on the distance to the phantom's isocenter. This methodology is described for a single cylindrical insert in Appendix B in more detail.

4. Results and discussion

4.1. CUDA - performance improvements

In this section, the speedup that could be achieved by switching from single-core computing to GPU accelerated computing is shown. Therefore, the system computing time for the generation of one improved radiography is compared for different computing systems (see table 5). For each radiography approximately $1.1 \cdot 10^7$ particles were detected, for the Gammex phantom [54]. One must consider that the

Computing device	Computing time	Speed up
Intel(R) Xeon(R) E5 - 2650 v4 (CPU)	(163.35 ± 0.25) s	1
NVIDIA Quadro K620 (GPU)	(18.90 ± 0.1) s	9
NVIDIA RTX 4000 (GPU)	(1.25 ± 0.02) s	131

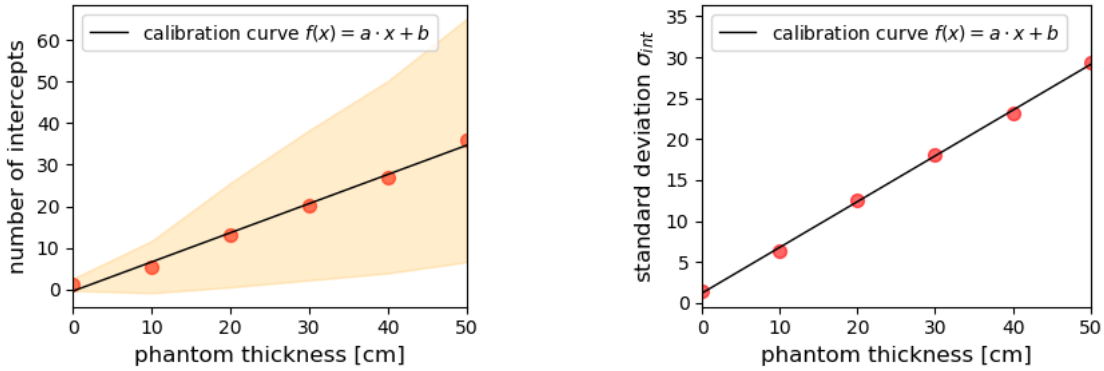
Table 5. Average computation time for the generation improved radiographies. The speed up in the third column is calculated by computing the fraction of the single core computing time of the Intel Xeon CPU and the respective GPU times.

computation time referenced in table 5 is measured for the execution of the algorithm creating the improved radiographies only, as this is the part of the calculations executed in parallel on the GPU. For total computing time, latency for reading the simulation data into memory (data I/O) must also be accounted for. Since the latency in data I/O depends on the size of the files that have to be fetched into memory, the number of particle trajectories simulated per projection ultimately strongly influences the total computation time. In this thesis, the simulation data, i.e., the detected positions, directions, and the WEPL for each proton, is provided in the form of plain text files.

For the simulation data of the CATPHAN, which was used for measuring the performance in table 5, the time needed for fetching the data into memory was measured to be on average 50 s for all systems. For smaller spatial dimensions and 225 particles per mm^2 detector area, leading to about 10^6-10^7 particles per projection, the data I/O time is in the order of a few seconds. However, for more extensive simulations, the data I/O is the bottleneck for overall computation time. Although the speedup achieved by the GPUs is quite remarkable, the CUDA code can certainly be improved as advanced code optimization techniques have not been considered within this thesis. For example, no emphasis has been laid on choosing the optimal type of memory for the data structures embedded in the implemented algorithm. Storing the arrays containing the simulation data in texture memory instead of ordinary global memory may improve the speedup additionally.

4.1.1. Calibration

As discussed in section 3.3.2, the number of intercepts is acquired for several phantom thicknesses. Figure 30 shows how the mean number of intercepted channel boundaries evolves with increasing sample thickness. Furthermore, an increase in the variance of the number of intercepts can be observed. While for a water block with a thickness of 10 cm the standard deviation is given by ≈ 6 intercepts, a water block of a thickness of 50 cm yields a standard deviation of ≈ 30 intercepts.



(a) **(b)**

Figure 29. (a) Linear fit of the intercept numbers with function parameters shown in table 6. The orange dots represent the mean number of intercepts for the corresponding thickness of the water block, whereas the orange area illustrates the standard deviation of the number of intercepts. (b) Linear fit of the standard deviation of the intercept numbers σ_{int}.

Parameter	Value
a	$(0.70 \pm 0.03)\frac{1}{\text{cm}}$
b	-0.45 ± 0.90

Table 6. Linear fit of the intercept numbers $f(x) = ax + b$.

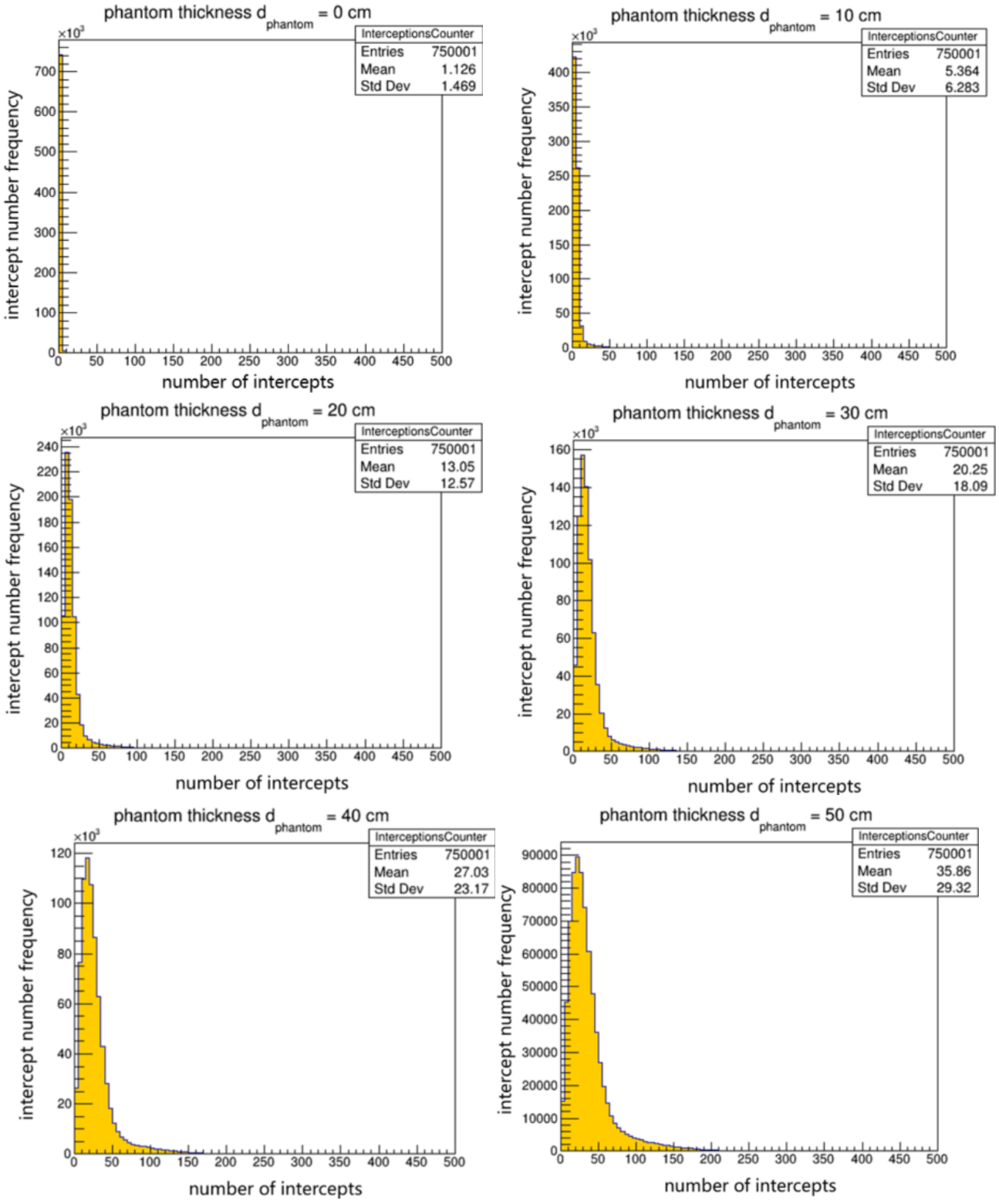


Figure 30. Distribution of channel intercepts for varying phantom thicknesses.

Using the fitting data, the optimal array size N_{opt} is calculated at pre-compile time using The mean of the intercept numbers and their variance is depicted in figure 29a for each phantom thickness. Figure 30 shows that the variance increases for larger sample dimensions, therefore it would not be sufficient to choose the optimal array size by evaluating the obtained calibration function for the mean number of intercepts as this would results in a truncation of many particles. To choose the array size optimally, the value obtained by evaluating the calibration function depending on the sample thickness is increased by three standard deviations (3 σ). As the standard deviation increases with the sample thickness, there has to be found a functional dependence as well. The dependence of the variation on the sample thickness is approximated to be linear (see figure 29b), which is expected to be sufficient for the calibration purpose. The ideal number of intercepts is then calculated by

$$N_{\text{opt}}(x) = \mu(x) + 3\sigma(x), \tag{4.1}$$

with $\mu(x)$ being the linear fit of the mean of the intercept numbers, $\sigma(x)$ the calibration curve of the standard deviation, and x the sample thickness. Using this approach for all improved radiographies generated within this thesis, only about 0.2 % of the particles are rejected from the data set due to their number of channel intercepts exceeding the allowed number N_{opt}.

4.2. Image reconstruction

In this section, the central slices of the image reconstructions of the Gammex 467 and the spiral phantom are analyzed in terms of spatial resolution and RSP accuracy, respectively. In table 7, the computation time of the image reconstruction algorithms used within this thesis is listed for both phantoms. The input paramters for the algoithms were already discussed in section 4.2. The computation time

Algorithm	**Computation time**	
	Spiral phantom	Gammex phantom
FBP	2.5 s	1.4 s
SART	12 min 5 s	2 min 39 s
OS-SART	7 min 21 s	39 s
SIRT	2 min 15 s	31 s
ASD-POCS	20 min	2 min 3 s

Table 7. The computation time of the image reconstruction for each algorithm.

of the algorithms is mainly determined by the size of the reconstructed volume (total number of voxels) and the number of radiographies used for reconstruction. Comparing the computation times, the reconstruction algorithms are faster for the Gammex phantom. The large computation time of the iterative algorithms for the spiral phantom can be explained by the large number of proton radiographies used, as the difference in reconstructed volume between the spiral and the Gammex phantom is negligible (see table 4).

4.2.1. RSP accuracy

In figure 31, the central slice of the Gammex phantom obtained with different reconstruction algorithms is presented. In figures 33 - 34, the mean RSP values of the tissue inserts obtained from the image reconstructions are compared to their reference value for the FBP and the ASD-POCS algorithm respectively. For the SIRT, the SART and the OS-SART algorithm the results can be found in appendix C. The mean RSP is calculated for a circular ROI with a diameter of 20 mm located at the center of the corresponding inserts. Although the tissue inserts have diameters of 26 mm, the circular ROI is chosen to be slightly smaller to assure that the spatial resolution does not affect the resulting mean RSP. The black bars, depicted in figures 33 - 34, describe the standard deviation of the RSP values in the corresponding ROI. A small standard deviation indicates better performance in terms of noise. For further noise analysis, the mean standard deviation is calculated for each algorithm. As depicted in figure 32, the ASD-POCS algorithm shows the best performance in terms of noise. This does not come by surprise as the algorithm implements a regularizer that minimizes the total variation of the RSP image [13].

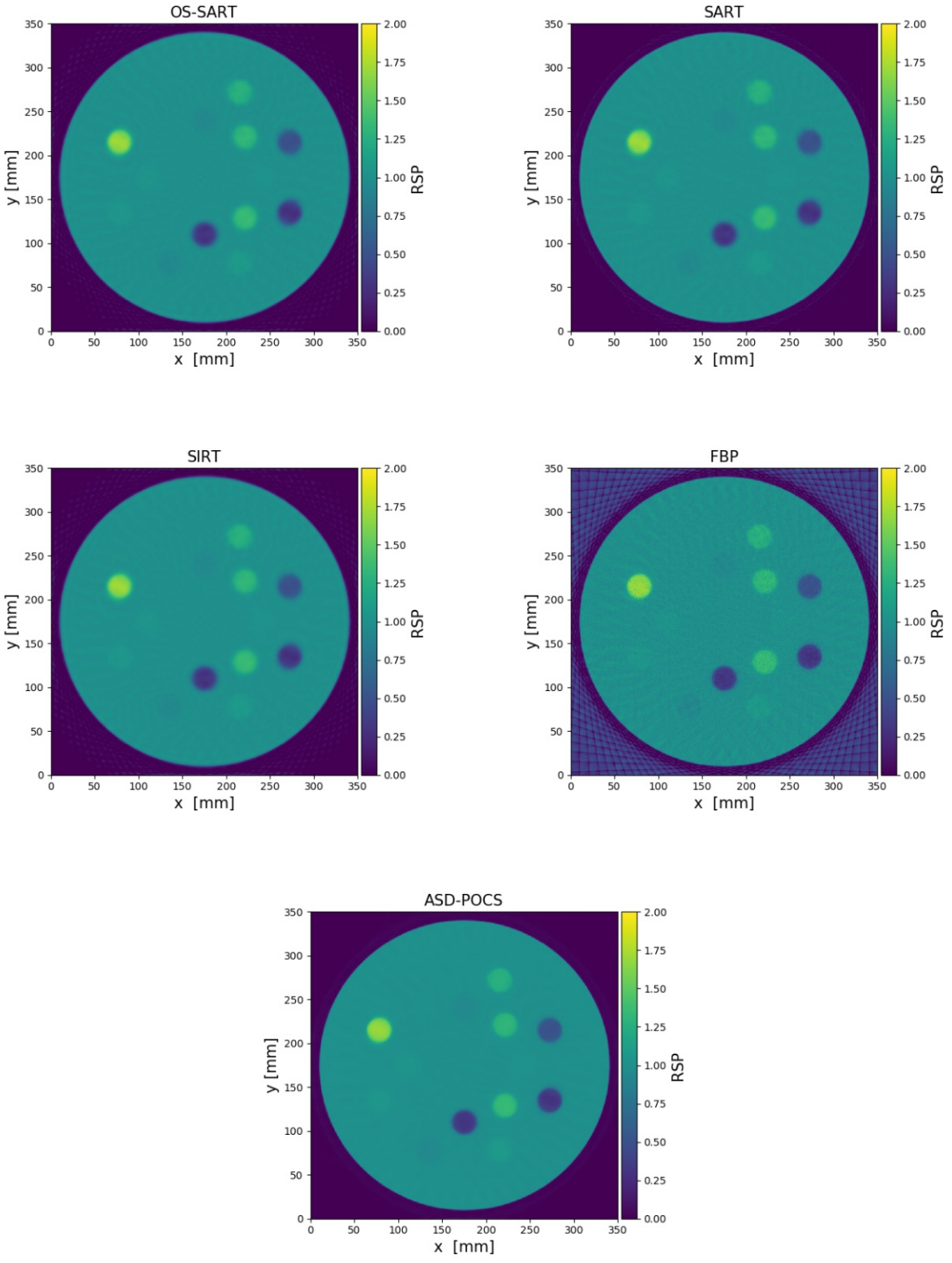


Figure 31. Central slice of the Gammex phantom reconstructed with different image reconstruction algorithms.

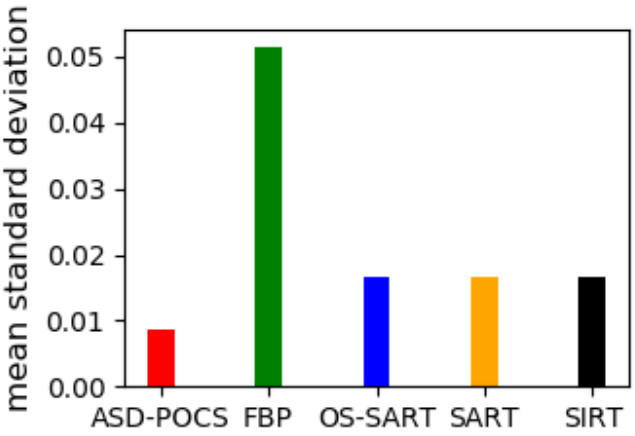


Figure 32. The mean standard deviation is obtained by calculating the mean from the standard deviations of the corresponding tissue inserts. The iterative algorithms perform better than the FBP approach.

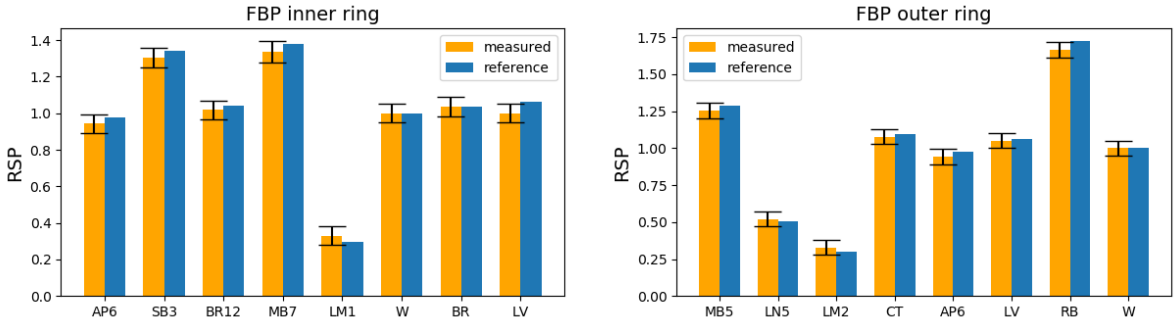


Figure 33. Comparison of the RSP values obtained from image reconstruction via FBP algorithm to the corresponding reference values for the tissue inserts. On the left side the RSP values correspond to the tissue inserts of the inner ring, whereas the right side depicts the RSP accuracy for the outer ring of the inserts.

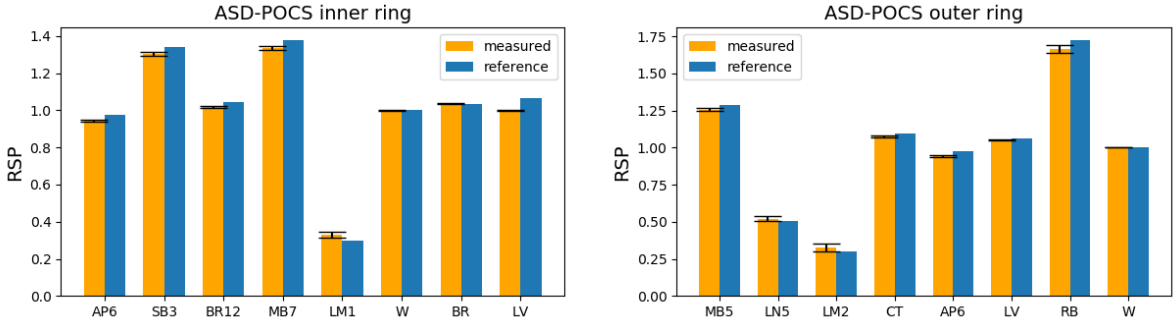


Figure 34. Comparison of the RSP values obtained from image reconstruction via ASD-POCS algorithm to the corresponding reference values for the tissue inserts. On the left side the RSP values correspond to the tissue inserts of the inner ring, whereas the right side depicts the RSP accuracy for the outer ring of the inserts.

In figure 35, the relative RSP error of the inserts is depicted for the image reconstruction algorithms used in this thesis. For most inserts, the choice of algorithm seems to be irrelevant, as no significant difference between the relative errors of the respective algorithms can be observed. However, for the tissue inserts with the largest errors, the iterative algorithms perform better than the FBP.

Generally, the algorithms show the best performance for inserts with RSP values close to the RSP of water ($\mathrm{RSP}_{\mathrm{water}} = 1$). Consequently, the largest relative error can be observed for the lung insert LM1 ($\mathrm{RSP}^{\mathrm{ref}}_{\mathrm{LM1}} = 0.298$), ranging from 2.9% for the SIRT algorithm to 8.2% for the FBP algorithm. The iterative algorithms perform better on the lung insert than the FBP algorithm. Interestingly, the ASD-POCS algorithm shows the worst performance on the lung inserts compared to the other iterative algorithms. Comparing the RSP accuracy of the outer ring and the inner ring for adipose tissue (AP6) and lungmoby (LM1 and LM2), the distance of insert to the center seems to have a negligible influence on the result, as nearly the same relative errors can be observed for the insert pairs.

The RSP values of the lung insert calculated with the FBP algorithm in Khellaf et al. [14] show the largest relative error compared to the other inserts. The same behaviour could be observed in this thesis. Although the results in Khellaf et al. [14] show better RSP accuracy, the performance is still comparable as it shows the same dependence on the insert material. As mentioned before, the iterative algorithms perform better on the lung inserts than the FBP. Therefore, the influence of the insert's material on the RSP accuracy can be improved by choosing an iterative algorithm over the FBP. In contrast to the findings for the RSP accuracy in Khellaf et al. [14], the relative errors of the insert's RSP in this thesis indicate inferior accuracy. This is not surprising as many influences are contributing to the observed discrepancies. First of all, in this thesis, projections have only been generated for 90 angles instead of 360 angles reported in Khellaf et al. [14]. Furthermore, the partly idealized proton imaging setup simulated in this thesis underlies severely different idealization assumptions than the setup in their study. In Khellaf et al. [14], the

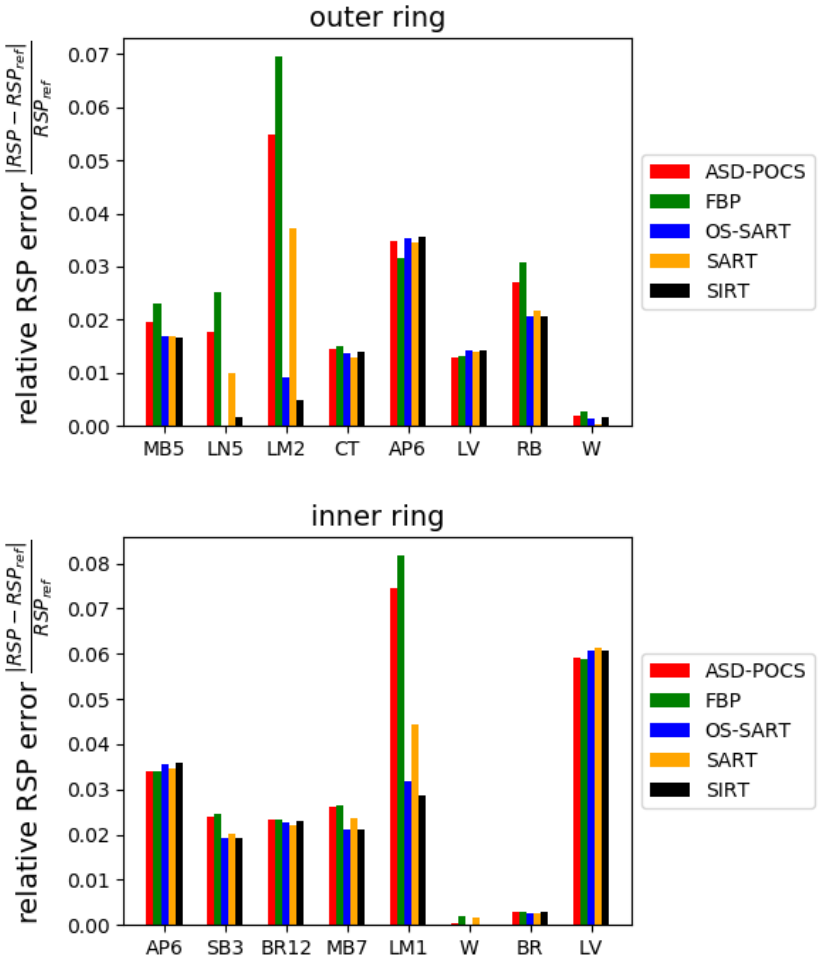


Figure 35. Comparison of the relative error of the RSP values within the Gammex phantom for the different algorithms.

idealized setup consisted of one tracking detector upstream and downstream with infinite spatial resolution. Contrary, the imaging setup in this thesis consists of two tracking detectors upstream and downstream with finite spatial dimensions (300 µm). Since the parameters of the cubic spline used to estimate the paths of the protons depend on the position measurements conducted with the tracking detectors, additional uncertainties are introduced by the finite extension of the detectors.

Using improved radiographies as suggested by Collins-Fekete et al. [4], the non-linear path of the particles is considered via redistribution of WET on several pixels by a weighting factor. The weighting factor ultimately depends on the length of each particle spent in a channel extending to the respective detector pixel. Therefore, uncertainties in the path estimate will undoubtedly translate to errors in the WET of the channels. Although the accuracy of the path estimation affects the estimated WET of certain channels, limitations on the RSP accuracy due to using improved radiographies for reconstruction will still prevail. This is due to the fact that particles

traversing regions of large stopping power can still contribute to channels with low WET. This intrinsic bias can not be eliminated by choice of the image reconstruction algorithm. Additionally, estimating the WEPL by using the approximation of Donahue, Newhauser, and Ziegler [17] also introduces uncertainties concerning RSP accuracy. Nevertheless, the choice of algorithm can improve the quality of the reconstructed image as the results obtained by the iterative algorithms, e.g., the ASD-POCS algorithm, show clearly better signal-to-noise ratio (figure 33 - figure 34).

4.2.2. Spatial resolution

In figure 36, the central slices of the image reconstructions of the spiral phantom are depicted. Figure 36 again shows that the image obtained with the FBP approach suffers the most from noise compared to the reconstruction results obtained with iterative algorithms. From visual observation the SIRT algorithm seems to generate less noise which can be observed in the less grainy water hull surrounding the aluminium inserts. On the other hand, the SIRT algorithm produces the most blurry inserts as there is no sharp contrast edge observable at the inserts boundaries. Smooth edges ultimately translate to flat edge spread functions, which is an indicator for poor spatial resolution. Visually, the SART and the OS-SART algorithm seem to produce similar outcomes. The most significant observable difference is the more apparent graininess in the image generated with the SART algorithm. In figure 37, the spatial resolution, calculated according to equation (3.10), in dependence of the distance to the center of the phantom is shown. It can be seen that the spatial resolution increases for the inserts located further away from the phantom's center. The same coherence could be observed in Khellaf et al. [14]. This is due to the fact that the uncertainty of the path estimates is the largest in the center of the phantom which is shown in Li et al. [2] and Collins-Fekete et al. [38].

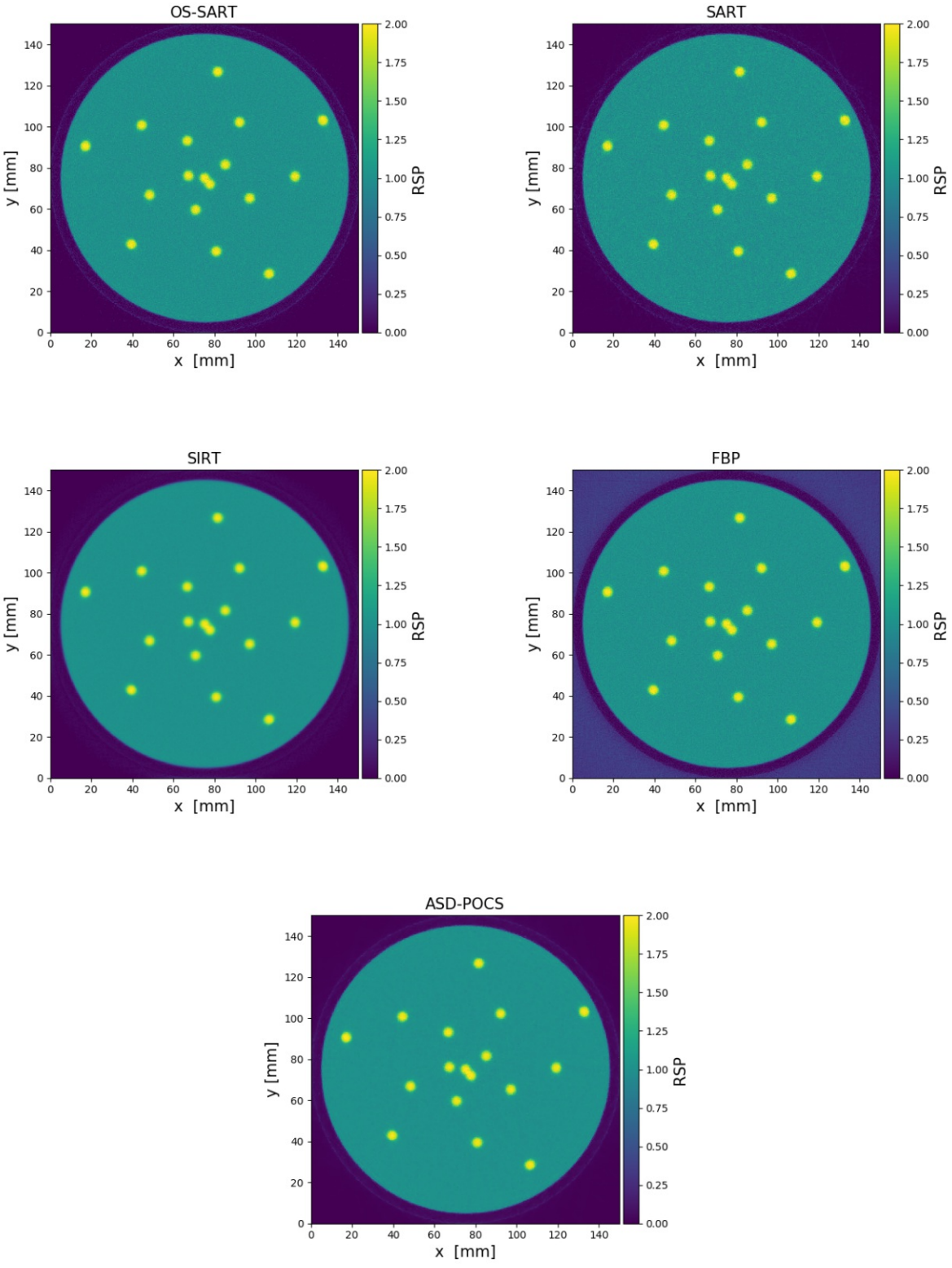


Figure 36. Central slice of the spiral phantom reconstructed with different image reconstruction algorithms.

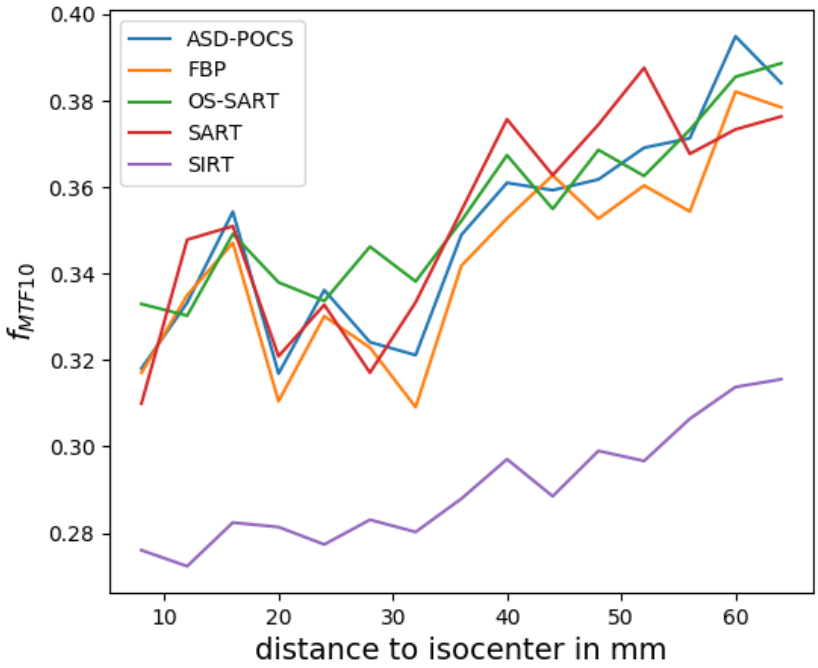


Figure 37. Evolution of the spatial resolution measure in dependence of the distance to the phantoms center.

The difference in spatial resolution of the images generated with FBP, SART, ASD-POCS and OS-SART are mostly negligible. As already discussed before, the image generated by the SIRT algorithm does perform the worst in terms of spatial resolution. This may be caused by the number of iterations chosen to execute the SIRT reconstruction as described in table 3. However, the SIRT algorithm is generally expected to deliver subpar results compared to the SART and the OS-SART algorithm [13].

Although the results show the same dependence of the spatial resolution on the distance of the insert to the phantom's center as the findings in Khellaf et al. [14], the image reconstruction performance seems to be worse for the spatial resolution as well due to the reasons already discussed in section 4.2.1. Additional discrepancies may be introduced by different assumptions in the calculation of the MTF, which includes background reduction techniques [63] in the implementation of Khellaf et al. [14]. This leads to a reduced overall noise in the obtained projection, which ultimately results in a less noisy ESF. As the measure for spatial frequency $f_{10\%}$ depends solely on parameter c of equation (3.10), reduced noise significantly impacts the spatial resolution as reduced RSP uncertainty around the edges of the insert leads to steeper decrease of the RSP in these areas. Therefore, the results for the spatial frequencies calculated in this thesis are not directly comparable to the findings in Khellaf et al. [14] but the spatial frequencies should be of the same magnitude. In this work the

spatial frequencies for the better performing algorithms (SART, OS-SART, FBP, ASD-POCS) are ranging from $(0.3 - 0.4)\frac{\text{lp}}{\text{mm}}$, which is about the same range that could be observed in the realistic setup described in Khellaf et al. [14].

5. Conclusion and outlook

In the scope of this thesis, a collection of CUDA-based files, referred to as pCT toolbox, that enables GPU accelerated generation of improved radiographies, was developed. It was designed in the same layered structure as the TIGRE toolbox [5] and is intended to be used together with the already implemented algorithms within this framework. Implementational details of the pCT toolbox were discussed in section 3.3. The performance improvement achieved by utilizing GPUs via CUDA was discussed in 4.1. The computing time for the generation of improved radiographies could be reduced drastically by the CUDA utilization as shown in section 4.1. Although the computing time could be reduced drastically, the overall computing performance can still be improved by optimizing the data I/O interface.

In this thesis, the simulation data were provided to the pCT toolbox via plain text files. As the data I/O is implemented in the Matlab header of the pCT toolbox, a file format more suitable to the Matlab environment could reduce the latency introduced by reading the simulation data into memory. Redesigning the processing of the simulation or experimental data to provide the necessary input to the pCT toolbox in the *.mat* format, which is optimized for Matlab loading operations, would be a suitable approach for improving overall computing performance. In section 4.1.1, the pre-compile process of the pCT toolbox, aiming to reduce source code alteration by the user, was developed. A calibration function, designed for biological tissues, that connects the estimated sample thickness with the expected number of channel intercepts was established in this context. This function is used within the start-up scripts of the pCT toolbox to provide a workaround for the missing ability of the algorithm to provide dynamic adjustments of the arrays storing the varying number of channel intercepts of individual protons. This restriction is ultimately implied by the CUDA framework, which forbids dynamic memory allocation of local thread memory [58]. The calibration function sets the size of the arrays storing the intercept data at the pre-compile time, which ensures that the user of the pCT toolbox will not receive subpar results. Choosing a small storage size would remove a significant amount of protons from the path estimation process, whereas an unreasonably large array size would unnecessarily increase the computing time. The estimating function found in this thesis can undoubtedly be improved by adding additional parameters which enable the consideration of further properties of the pCT imaging setup. It has to be mentioned that the current implementation can only be used for the geometries equal to the one discussed in this thesis as the approach lacks flexibility in terms of geometrical changes of the pCT apparatus. The distance of the particle source to the tracking detectors, the distance between the tracking detectors, and the distance of the tracking detectors to the sample could be included within the model to further improve the accuracy of the suggested array size.

In the second part of this thesis, Monte Carlo simulations of the proton imaging setup were performed with two different phantoms to provide the necessary list mode

data for image reconstruction (see section 3.1.4). The choice of the phantoms was based on the image reconstruction quality analysis in Khellaf et al. [14]. Therefore, the spiral phantom and the Gammex 476 tissue characterization phantom were used in the simulation of the proton imaging setup. In this context, the spiral phantom was used to evaluate spatial resolution, whereas the Gammex phantom was used to evaluate RSP accuracy. The results for the spatial resolution and RSP accuracy can be found in sections 4.2.1 and 4.2.2. Although the phantoms chosen in this thesis do match the ones in Khellaf et al. [14], the results found in this thesis can not be directly compared to the findings in Khellaf et al. [14]. As discussed in sections 4.2.1 and 4.2.2, the performance discrepancy can be mainly explained by the simulation conditions for the pCT scanner chosen within this thesis (see section 4.2).

However, it could be shown in section 4.2.1 that choosing an iterative algorithm in combination with the method of Collins-Fekete et al. [4] for image reconstruction reduces the influence of the material on RSP accuracy. Additionally, RSP accuracy may be improved by a more sophisticated way of evaluating the left side of equation (2.14). In this thesis, the evaluation is based on calculating the range difference by the analytic approximation of Donahue, Newhauser, and Ziegler [17]. Regarding the spatial resolution, no significant improvements could by achieved by using an iterative algorithm (see section 4.2.2) since the most substantial limitations on spatial resolution are imposed by the accuracy of the path estimation [2].

Future improvements of the pCT toolbox include the implementation of a cone beam geometry in the code. Currently, the pCT toolbox can only deal with parallel beam geometries. Since parallel beams are only suitable for describing pencil beam delivery, further adaptions of the algorithms used within the pCT toolbox have to be made to include, for example, the cone-beam geometry. It has to be mentioned that the pre-compilation script for estimating the optimal array size developed within this thesis would have to be redesigned for cone-beam geometry. As for now, the pCT toolbox is stored in a private repository, but actions are planned to adapt and refine it to make it available as an actual code extension to the TIGRE toolbox.

To conclude, the computation time for the generation of improved radiographies could be drastically reduced by the pCT TIGRE toolbox. The obtained high resolution proton radiographies can be used as input data for any conventional CT algorithm. Since the implementation of new algorithms to the TIGRE toolbox can be done modularly, further pCT image reconstruction techniques, which go beyond the scope of this work, (e.g. [45, 46, 47]), could also be added to the TIGRE image reconstruction toolbox.

6. Acknowledgements

First of all, I want to thank Dr. Albert Hirtl and Dr. Thomas Bergauer for making this thesis possible.

Most notably, I want to thank DI Stefanie Kaser for her endless patience and relentless effort in answering and discussing my questions from the origin to the end of this thesis. I also want to thank DI Felix Ulrich-Pur for giving insight on the hardware part of pCT by taking me to a shift to MedAustron. Additionally, I am very grateful to HEPHY for letting me use their computing infrastructure.

I also want to thank my family, who supported me through my studies from the beginning to the end. On a personal level, I especially want to thank Kristin, who always has been on my side through all ups and downs over the past few years. Without her consistent support, this journey would not have generated the wonderful memories I thankfully can keep in my mind now forever.

Appendices

A. General solution of the cubic equation and numerical implementation

The core task in the generation of the improved radiographies is to find the intercepts of the particle paths with the shannel boundaries of the discretized volume. As the paths are represented by a cubic spline and the channel boundaries can ultimately be treated as constants, the intercepts can be found by solving equation (1)

$$a \cdot x^3 + b \cdot x^2 + c \cdot x + d \cdot x^3 = d_{\text{pix}}, \tag{1}$$

with d_{pix} being the orthogonal distance between origin and the respective channel boundary in x and y direction.

The properties of the solutions to algebraic equations are classified by analyzing the discriminant of the polynomial which is given by

$$D = a^4(r_1 - r_2)^2(r_2 - r_3)^2(r_3 - r_1)^2, \tag{2}$$

with r_i being the roots of the polynomial and a the coefficient of the third-order term in equation (1). In this section the focus is on the special case $N = 3$.

$D \geqslant 0$ If the roots are real and distinct, the product in equation (2) has to be positive and therefore hints at three real roots.

$D < 0$ As complex solutions always occur pairwise, the assumption of for example r_1 to be a real number implies that r_2 and r_3 are complex conjugates. Therefore $(r_2 - r_3)$ is purely imaginary and contributes a negative sign to the expression. But as $r_2 = r_3^*$ and

$$(r_1 - r_2) = (r_2 - r_3)^* \cdot (r_1 - r_2) \cdot (r_2 - r_3) > 0, \tag{3}$$

the discriminant $D < 0$ and therefore the polynomial has only one real root. Evaluating the discriminant is crucial for developing a program that provides a general solution to the cubic equation. To simplify the problem, a coordinate transformation $x = t - \frac{3b}{a}$ is introduced and the cubic equation reduces to its "depressed form"

$$t^3 + pt + q = 0, \tag{4}$$

with $p = \frac{3ac - b^2}{3a^2}$ and $q = \frac{2b^3 - 9abc + 27a^2d}{27a^3}$. Using this transformation the cubic discriminant can be written as

$$D = -(4p^3 + 27q^2). \tag{5}$$

Since D can be calculated from the parameters of the depressed cubic using 5 the program designed to calculate the roots has to consider $D > 0$ and $D < 0$.

Purely real roots $D > 0$ For purely real roots Viete [65] suggested to use the trigonmetric identity

$$4\cos^3\theta - 3\cos\theta - \cos(3\theta) = 0 \tag{6}$$

and the substitution

$$t = u\cos\theta \tag{7}$$

as an ansatz for finding the roots of the cubic equation. The goal is to bring the depressed cubic into the same form as equation (6) by choosing p, q accordingly. Rewriting the depressed cubic by using the suggested substitution and dividing the equation by $\frac{u^3}{4}$ yields

$$4\cos^3 + \frac{4p}{u^2}\cos\theta + \frac{4q}{u^3} = 0. \tag{8}$$

Comparing the coefficients of equation (6) and (8), choosing $u = 2\sqrt{-\frac{p}{3}}$ further simplifies the equation to

$$4\cos^3 - 3\cos\theta + \frac{3q}{2p}\sqrt{-\frac{3}{p}} = 0. \tag{9}$$

Doing another coefficient comparison yields

$$\cos(3\theta) = \frac{3q}{2p}\sqrt{-\frac{3}{p}}. \tag{10}$$

The solution to the angles θ fulfilling equation (10) is given by

$$\theta_k = \frac{1}{3}\arccos\left(\frac{3q}{2p}\sqrt{-\frac{3}{p}}\right) - \frac{2\pi k}{3} \text{ with } k = 1, 2, 3..., \tag{11}$$

since phase shifts $\cos(\theta + 2\pi) = \cos(\theta)$ yield the same value for the cosine. By substituting equation (11) into equation (7), the real roots can be expressed as

$$t_k = 2\sqrt{-\frac{p}{3}}\cos\left[\frac{1}{3}\arccos\left(\frac{3q}{2p}\sqrt{-\frac{3}{p}}\right) - \frac{2\pi k}{3}\right] \text{ with } k = 1, 2, 3... \,. \tag{12}$$

One real root $D < 0$ For finding the only real root of the cubic polynomial [66] suggests a different parametrization of the cubic equation (13)

$$x^3 + ax^2 + cx + d = 0. \tag{13}$$

By choosing $Q = \frac{a^2 - 3b}{9}$ and $R = \frac{2a^3 - 9ab + 27c}{54}$ the discriminant criteria for classification of the roots is given by $R^2 < Q^2$ for three real roots and $R^2 > Q^2$ for the other cases. Following the algorithm in [66], A and B are calculated by

$$A = -sgn(R)\left[|R| + \sqrt{R^2 - Q^3}\right]^{\frac{1}{3}} \text{ and } B = \begin{cases} \frac{Q}{A} & (A \neq 0) \\ 0 & (A = 0) \end{cases}. \tag{14}$$

The single real root is then calculated by

$$x_0 = (A + B) - \frac{a}{3}. \tag{15}$$

B. Computation of the modulated transfer function for cylindrical inserts

In section 3.5.2, the modulated transfer function (MTF) is introduced as measure for the spatial resolution of images. The main quantity, that has to be extracted from the reconstructed image is the edge spread function (ESF). Therefore, RSP values of the pixels in a circular ROI (see figure 23) surrounding the centers of the inserts are sorted according to their distance from the insert's center. The aligned RSP values in dependence of their distance to their center can be seen in figure A1.

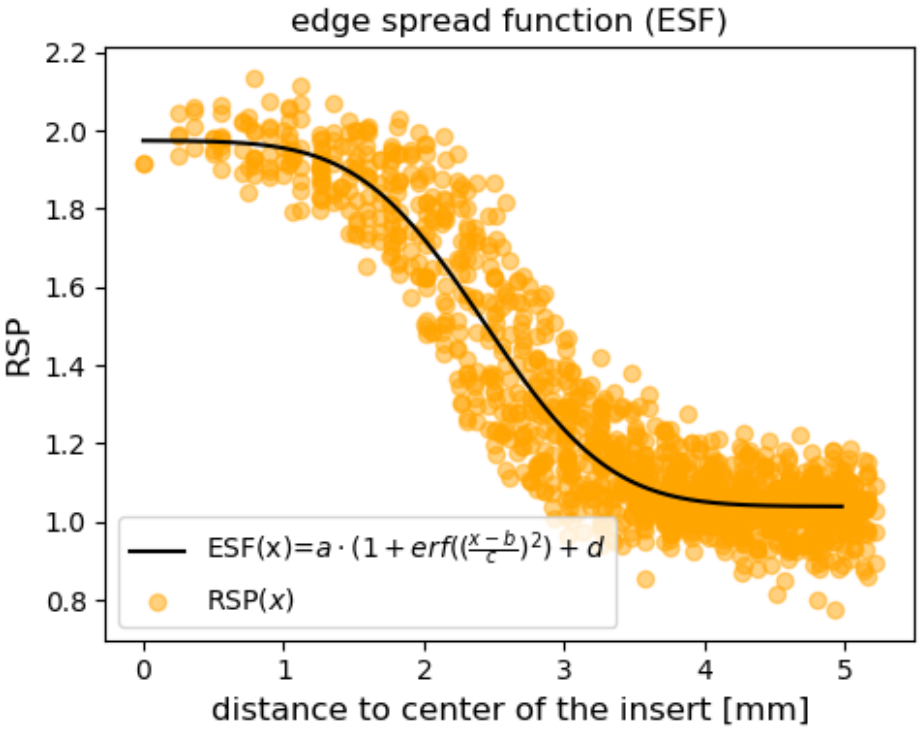


Figure A1. RSP values of the pixels in dependence of their distance x to the insert's center.

Since the MTF is obtained by applying a Fourier transformation to the derivative of the ESF, the RSP values have to be fitted so that equidistant values can be sampled. This is necessary since numerical Fourier transformations can only be performed for equidistant sampling points. In principal, the ESF can be fitted by a LOESS (locally estimated scatterplot smoothing [67]) fit whereas [63] suggests a non-linear fitting procedure based on the analytic ansatz already shown in equation (3.9). The LOESS fit can be used for evaluating an estimated ESF at equidistant points. In this thesis, the function obtained has been observed to suffer from oscillations due to the high degree polynomials necessary to fit the noisy pCT data. As the derivative of the gaussian error function, yields a gaussian function the Fourier transformation of the derivative of the ESF can be performed analytically

$$\frac{\partial}{\partial x}\mathrm{ESF}(x) = \frac{a}{c^2}\exp\left(-\left(\frac{x-b}{c}\right)^2\right) = \mathrm{LSF}(x). \tag{16}$$

Performing the Fourier transformation of equation (16), further normalizing and taking it's modulus yields

$$\mathrm{MTF}(f) = \frac{\int dx\ \mathrm{LSF}(x) \exp\left(2\pi i f\right)}{\int dx\ \mathrm{LSF}(x)} = \exp(-2\pi^2 c^2 f^2), \tag{17}$$

where f is the spatial frequency and c a fitting parameter of equation (3.9). The modulated transfer function is depending solely on the fitting parameter c, which controls the slope of the ESF at the edge of the insert. The measure of of spatial resolution, used in the literature [14, 4], is the spatial frequency value for which the MTF is decreased to 10% of its value at the origin (figure A2). The spatial frequency $f_{10\%}$ can be calculated by using (3.10).

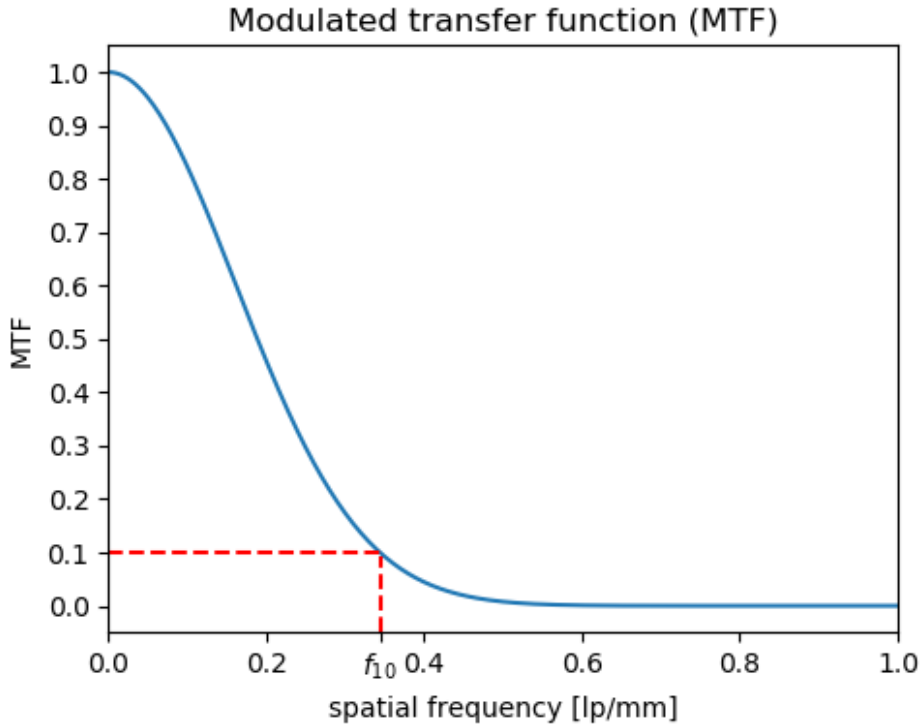


Figure A2. Modulated transfer function - The spatial frequency $f_{10\%}$ is obtained by finding the frequency for which the MTF decreased to 10% of the amplitude at MTF(0).

C. Relative stopping power accuracy

In this section the relative error RSP_{err} of the mean RSP within the ROI of a specific insert is tabulated for different tissue inserts and algorithms. Additionally, the mean RSP values of the tissue inserts of the Gammex 467 phantom are compared to their reference values for the SIRT, the SART and the OS-SART algorithm.

Algorithm comparison

In figure A3 - A5, the reconstructed RSP values of the inserts of the Gammex 467 phantom are depicted for the SIRT, the SART and the OS-SART algorithm respectively.

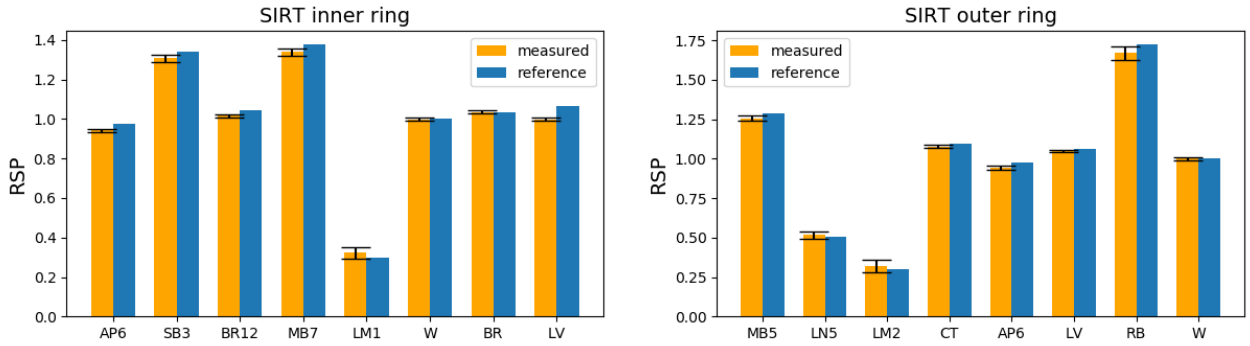


Figure A3. Comparison of the RSP values obtained from image reconstruction via SIRT algorithm to the corresponding reference values for the tissue inserts. On the left side the RSP values correspond to the tissue inserts of the inner ring, whereas the right side depicts the RSP accuracy for the outer ring of the inserts.

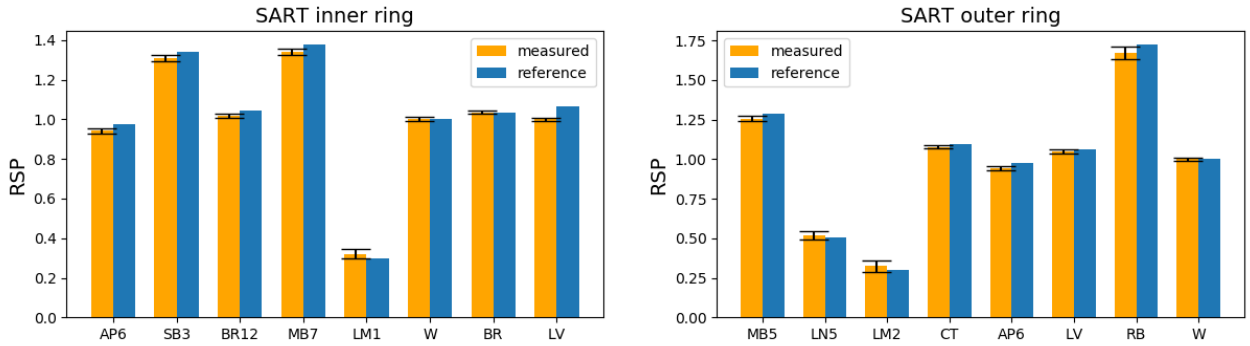


Figure A4. Comparison of the RSP values obtained from image reconstruction via SART algorithm to the corresponding reference values for the tissue inserts. On the left side the RSP values correspond to the tissue inserts of the inner ring, whereas the right side depicts the RSP accuracy for the outer ring of the inserts.

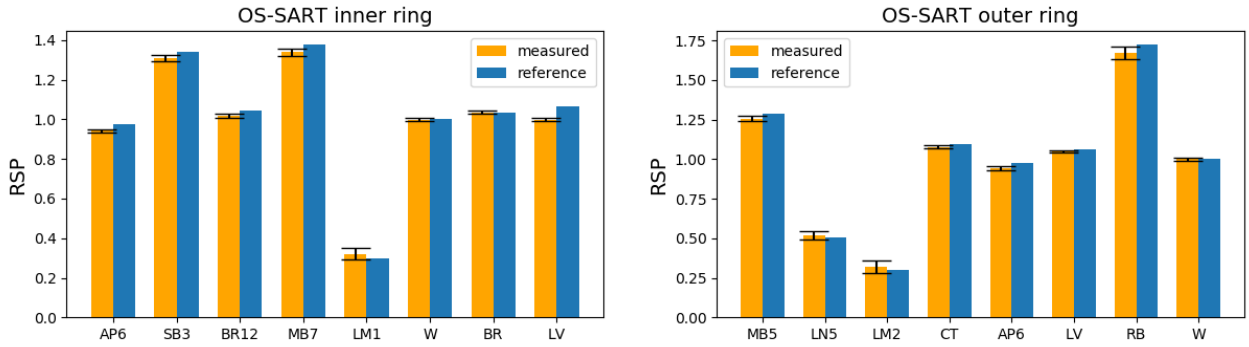


Figure A5. Comparison of the RSP values obtained from image reconstruction via OS-SART algorithm to the corresponding reference values for the tissue inserts. On the left side the RSP values correspond to the tissue inserts of the inner ring, whereas the right side depicts the RSP accuracy for the outer ring of the inserts.

Outer ring

In table A.1, the relative error of the RSP is shown for the tissue inserts of the outer ring of the Gammex 467 phantom.

	$\mathbf{RSP}_{\mathrm{err}}$				
insert	ASD-POCS	FBP	OS-SART	SART	SIRT
MB5	0.020 ± 0.005	0.023 ± 0.040	0.017 ± 0.011	0.017 ± 0.012	0.017 ± 0.011
LN5	0.018 ± 0.020	0.025 ± 0.093	0.000 ± 0.032	0.010 ± 0.032	0.002 ± 0.030
LM2	0.055 ± 0.057	0.069 ± 0.162	0.009 ± 0.092	0.037 ± 0.086	0.005 ± 0.092
CT	0.015 ± 0.004	0.015 ± 0.047	0.014 ± 0.008	0.013 ± 0.009	0.014 ± 0.008
AP6	0.035 ± 0.004	0.032 ± 0.053	0.035 ± 0.014	0.034 ± 0.015	0.036 ± 0.014
LV	0.013 ± 0.003	0.013 ± 0.049	0.014 ± 0.009	0.014 ± 0.011	0.014 ± 0.009
RB	0.027 ± 0.009	0.031 ± 0.031	0.021 ± 0.016	0.022 ± 0.016	0.021 ± 0.016
W	0.002 ± 0.003	0.003 ± 0.049	0.001 ± 0.010	0.000 ± 0.011	0.002 ± 0.010

Table A.1. Relative error of the RSP $\mathrm{RSP}_{\mathrm{err}}$ of the tissue inserts of the outer ring of the Gammex 476 phantom tabulated for different algorithms.

Inner ring

In table A.1, the relative error of the RSP is shown for the tissue inserts of the inner ring of the Gammex 467 phantom.

	$\mathbf{RSP}_{\mathrm{err}}$				
insert	ASD-POCS	FBP	OS-SART	SART	SIRT
AP6	0.034 ± 0.002	0.034 ± 0.050	0.036 ± 0.008	0.035 ± 0.012	0.036 ± 0.008
SB3	0.024 ± 0.004	0.025 ± 0.039	0.019 ± 0.009	0.020 ± 0.011	0.019 ± 0.009
BR12	0.023 ± 0.003	0.023 ± 0.049	0.023 ± 0.008	0.022 ± 0.009	0.023 ± 0.008
MB7	0.026 ± 0.004	0.026 ± 0.041	0.021 ± 0.008	0.024 ± 0.010	0.021 ± 0.008
LM1	0.075 ± 0.024	0.082 ± 0.163	0.032 ± 0.062	0.044 ± 0.065	0.029 ± 0.059
W	0.000 ± 0.003	0.002 ± 0.052	0.000 ± 0.009	0.002 ± 0.011	0.000 ± 0.009
BR	0.003 ± 0.002	0.003 ± 0.049	0.002 ± 0.007	0.003 ± 0.008	0.003 ± 0.007
LV	0.059 ± 0.002	0.059 ± 0.049	0.061 ± 0.007	0.061 ± 0.009	0.061 ± 0.008

Table A.2. Relative error of the RSP $\mathrm{RSP}_{\mathrm{err}}$ of the tissue inserts of the inner ring ring of the Gammex 476 phantom tabulated for different algorithms.

List of Figures

Bibliography

[1] Uwe Schneider, Eros Pedroni, and Antony Lomax. "The calibration of CT Hounsfield units for radiotherapy treatment planning." In: *Physics in Medicine and Biology* 41.1 (Jan. 1996), pp. 111–124. DOI: 10.1088/0031-9155/41/1/009.

[2] Tianfang Li et al. "Reconstruction for proton computed tomography by tracing proton trajectories: A Monte Carlo study." In: *Medical Physics* 33.3 (Feb. 2006), pp. 699–706. DOI: 10.1118/1.2171507.

[3] Robert P Johnson. "Review of medical radiography and tomography with proton beams." In: *Reports on Progress in Physics* 81.1 (Nov. 2017), p. 016701. DOI: 10.1088/1361-6633/aa8b1d.

[4] Charles-Antoine Collins-Fekete et al. "A maximum likelihood method for high resolution proton radiography/proton CT." In: *Physics in Medicine and Biology* 61.23 (Nov. 2016), pp. 8232–8248. DOI: 10.1088/0031-9155/61/23/8232.

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

[6] Xun Jia, Peter Ziegenhein, and Steve B Jiang. "GPU-based high-performance computing for radiation therapy." In: *Physics in Medicine and Biology* 59.4 (Feb. 2014), R151–R182. DOI: 10.1088/0031-9155/59/4/r151.

[7] Rajamanickam Baskar et al. "Cancer and radiation therapy: current advances and future directions." eng. In: *International journal of medical sciences* 9 (3 2012), pp. 193–9.

[8] Marco Durante and Jay S Loeffler. "Charged particles in radiation oncology." In: *Nature reviews Clinical oncology* 7.1 (Dec. 2010), pp. 37–43. DOI: 10.1038/nrclinonc.2009.183.

[9] Wayne D Newhauser and Rui Zhang. "The physics of proton therapy." In: *Physics in Medicine and Biology* 60.8 (Mar. 2015), R155–R209. DOI: 10.1088/0031-9155/60/8/r155.

[10] Marco Durante and Harald Paganetti. "Nuclear physics in particle therapy: a review." In: *Reports on Progress in Physics* 79.9 (Aug. 2016), p. 096702. DOI: 10.1088/0034-4885/79/9/096702.

[11] Mikaela Dell'Oro et al. "Influence of Target Location, Size, and Patient Age on Normal Tissue Sparing- Proton and Photon Therapy in Paediatric Brain Tumour Patient-Specific Approach." In: 12.9 (Sept. 2020), p. 2578. DOI: 10.3390/cancers12092578.

[12] Reinhard W. Schulte et al. "Density resolution of proton computed tomography." In: *Medical Physics* 32.4 (2005), pp. 1035–1046. DOI: https://doi.org/10.1118/1.1884906.

[13] Ander Biguri. "Iterative reconstruction and motion compensation in computed tomography on GPUs." PhD thesis. University of Bath, 2018.

[14] Feriel Khellaf et al. “A comparison of direct reconstruction algorithms in proton computed tomography.” In: *Physics in Medicine & Biology* 65.10 (June 2020), p. 105010. DOI: 10.1088/1361-6560/ab7d53.

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

[16] W. H. Bragg M.A. and R. Kleeman B.Sc. “On the α particles of radium, and their loss of range in passing through various atoms and molecules.” In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 10.57 (1905), pp. 318–340. DOI: 10.1080/14786440509463378.

[17] William Donahue, Wayne D Newhauser, and James F Ziegler. “Analytical model for ion stopping power and range in the therapeutic energy interval for beams of hydrogen and heavier ions.” In: *Physics in Medicine and Biology* 61.17 (Aug. 2016), pp. 6570–6584. DOI: 10.1088/0031-9155/61/17/6570.

[18] Professor E. Rutherford. “The scattering of α and β particles by matter and the structure of the atom.” In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 21.125 (1911), pp. 669–688. DOI: 10.1080/14786440508637080.

[19] Gert Moliere. “Theorie der Streuung schneller geladener Teilchen II Mehrfach-und Vielfachstreuung.” In: *Zeitschrift für Naturforschung A* 3.2 (1948), pp. 78–97. DOI: doi:10.1515/zna-1948-0203.

[20] Virgil L. Highland. “Some practical remarks on multiple scattering.” In: *Nuclear Instruments and Methods* 129.2 (1975), pp. 497–499. ISSN: 0029-554X. DOI: https://doi.org/10.1016/0029-554X(75)90743-0.

[21] Uwe Schneider and Eros Pedroni. “Multiple Coulomb scattering and spatial resolution in proton radiography.” In: *Medical Physics* 21.11 (1994), pp. 1657–1663. DOI: https://doi.org/10.1118/1.597212.

[22] C. Bopp. “The proton as a dosimetric and diagnostic probe.” PhD thesis. 2014.

[23] Gabor T. Herman. *Fundamentals of Computerized Tomography*. Springer London, 2009. DOI: 10.1007/978-1-84628-723-7.

[24] Wolfgang Birkfellner. *Applied medical image processing: a basic course*. CRC Press, 2016. ISBN: 1466555572.

[25] A. M. Cormack. “Representation of a Function by Its Line Integrals, with Some Radiological Applications.” In: *Journal of Applied Physics* 34.9 (1963), pp. 2722–2727. DOI: 10.1063/1.1729798.

[26] G Poludniowski, N M Allinson, and P M Evans. “Proton radiography and tomography with application to proton therapy.” In: *The British Journal of Radiology* 88.1053 (Sept. 2015), p. 20150134. DOI: 10.1259/bjr.20150134.

[27] N Krah et al. “A comprehensive theoretical comparison of proton imaging set-ups in terms of spatial resolution.” In: 63.13 (July 2018), p. 135013. DOI: 10.1088/1361-6560/aaca1f.

[28] P Pemler et al. “A detector system for proton radiography on the gantry of the Paul-Scherrer-Institute.” In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 432.2-3 (1999), pp. 483–495.

[29] R. Schulte et al. “Conceptual design of a proton computed tomography system for applications in proton radiation therapy.” In: *IEEE Transactions on Nuclear Science* 51.3 (2004), pp. 866–872. DOI: 10.1109/TNS.2004.829392.

[30] Felix Ulrich-Pur et al. “Feasibility study of a proton CT system based on 4D-tracking and residual energy determination via time-of-flight.” In: *arXiv preprint arXiv:2109.05058* (2021).

[31] S Rit et al. “The Reconstruction Toolkit (RTK), an open-source cone-beam CT reconstruction toolkit based on the Insight Toolkit (ITK).” In: 489 (Mar. 2014), p. 012079. DOI: 10.1088/1742-6596/489/1/012079.

[32] S Karczmarz. “Angenäherte Auflösung von Systemen linearer Gleichungen.” In: *Bull. Int. Acad. Pol. Sic. Let., Cl. Sci. Math. Nat.* (1937), pp. 355–357.

[33] Avinash C Kak and Malcolm Slaney. *Principles of computerized tomographic imaging*. Philadelphia: SIAM, 2001. ISBN: 089871494X.

[34] A.H. Andersen and A.C. Kak. “Simultaneous Algebraic Reconstruction Technique (SART): A superior implementation of the ART algorithm.” In: *Ultrasonic Imaging* 6.1 (1984), pp. 81–94. ISSN: 0161-7346. DOI: https://doi.org/10.1016/0161-7346(84)90008-7.

[35] Huihua Kong and Jinxiao Pan. “An Improved Ordered-Subset Simultaneous Algebraic Reconstruction Technique.” In: *2009 2nd International Congress on Image and Signal Processing*. 2009, pp. 1–5. DOI: 10.1109/CISP.2009.5302899.

[36] Emil Y Sidky and Xiaochuan Pan. “Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization.” In: 53.17 (Aug. 2008), pp. 4777–4807. DOI: 10.1088/0031-9155/53/17/021.

[37] George Dedes et al. “The role of Monte Carlo simulation in understanding the performance of proton computed tomography.” In: *Zeitschrift für Medizinische Physik* (2020). ISSN: 0939-3889. DOI: https://doi.org/10.1016/j.zemedi.2020.06.006.

[38] Collins-Fekete et al. “Developing a phenomenological model of the proton trajectory within a heterogeneous medium required for proton imaging.” In: *Physics in Medicine and Biology* 60.13 (June 2015), pp. 5071–5082. DOI: 10.1088/0031-9155/60/13/5071.

[39] D C Williams. “The most likely path of an energetic charged particle through a uniform medium.” In: 49.13 (June 2004), pp. 2899–2911. DOI: 10.1088/0031-9155/49/13/010.

[40] R. W. Schulte et al. “A maximum likelihood proton path formalism for application in proton computed tomography.” In: *Medical Physics* 35.11 (2008), pp. 4849–4856. DOI: https://doi.org/10.1118/1.2986139.

[41] B. Erdelyi. “A comprehensive study of the most likely path formalism for proton-computed tomography.” In: *Physics in Medicine and Biology* 54.20 (Sept. 2009), pp. 6095–6122. DOI: 10.1088/0031-9155/54/20/005.

[42] Nils Krah, Jean-Michel Létang, and Simon Rit. "Polynomial modelling of proton trajectories in homogeneous media for fast most likely path estimation and trajectory simulation." In: *Physics in Medicine and Biology* 64.19 (2019). DOI: 10.1088/1361-6560/ab3d0b.

[43] Leonard Eyges. "Multiple Scattering with Energy Loss." In: *Phys. Rev.* 74 (10 Nov. 1948), pp. 1534–1535. DOI: 10.1103/PhysRev.74.1534.

[44] Joe Deasy. "ICRU Report 49, Stopping Powers and Ranges for Protons and Alph Particles." In: *Medical Physics* 21.5 (1994), pp. 709–710. DOI: https://doi.org/10.1118/1.597176.

[45] Simon Rit et al. "Filtered backprojection proton CT reconstruction along most likely paths." In: *Medical Physics* 40.3 (2013), p. 031103. DOI: https://doi.org/10.1118/1.4789589.

[46] Simon Rit et al. "List-mode proton CT reconstruction using their most likely paths via the finite Hilbert transform of the derivative of the backprojection." In: *The 13th International Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine*. Fully3D 2015 Conference Proceedings. Newport, Rhode Island, United States, May 2015, pp. 324–327.

[47] Scott Penfold and Yair Censor. "Techniques in Iterative Proton CT Image Reconstruction." In: *Sensing and Imaging* 16.1 (2015), p. 19. ISSN: 1557-2072. DOI: 10.1007/s11220-015-0122-3.

[48] Rene Brun et al. *ROOT: data analysis framework v6.18/02*. 2019. DOI: 10.5281/ZENODO.848818.

[49] NVIDIA, Péter Vingelmann, and Frank H.P. Fitzek. *CUDA: Programming Guide*. 2021.

[50] Leland Lavele Carter and Edmond Darrell Cashwell. *Particle-transport simulation with the Monte Carlo method*. Tech. rep. Los Alamos Scientific Lab., N. Mex.(USA), 1975.

[51] S. Agostinelli et al. "Geant4—a simulation toolkit." In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 506.3 (July 2003), pp. 250–303. DOI: 10.1016/s0168-9002(03)01368-8.

[52] *Geant4 - Reference Physics Lists*. [Online, accessed 20.11.2021]. URL: https://geant4.web.cern.ch/node/155.

[53] S Jan et al. "GATE V6: a major enhancement of the GATE simulation platform enabling modelling of CT and radiotherapy." In: *Physics in Medicine and Biology* 56.4 (Jan. 2011), pp. 881–901. DOI: 10.1088/0031-9155/56/4/001.

[54] Chris Constantinou, James C. Harrington, and Larry A. DeWerd. "An electron density calibration phantom for CT-based treatment planning computers." In: *Medical Physics* 19.2 (1992), pp. 325–327. DOI: https://doi.org/10.1118/1.596862.

[55] U. Amaldi et al. "Construction, test and operation of a proton range radiography system." In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 629.1 (2011), pp. 337–344. ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2010.11.096.

[56] Benjamin Kirchmayer. "Analysis of optimized proton radiographies for proton CT." Project Thesis, Institute of High energy physics (HEPHY), 2021.

[57] Ritesh Nandgaonkar. *Global GPU Market Share (2010 – 2021)*. 2021. URL: https://businessquant.com/global-gpu-market-share.

[58] Jens Breitbart. "CuPP-a framework for easy CUDA integration." In: *2009 IEEE International Symposium on Parallel & Distributed Processing*. IEEE. 2009, pp. 1–8.

[59] Harald Paganetti. "Range uncertainties in proton therapy and the role of Monte Carlo simulations." In: (May 2012), R99–R117. DOI: 10.1088/0031-9155/57/11/r99.

[60] Stefanie Kaser et al. "First application of the GPU-based software framework TIGRE for proton CT image reconstruction." In: *Physica Medica* 84 (2021), pp. 56–64. ISSN: 1120-1797. DOI: https://doi.org/10.1016/j.ejmp.2021.03.006.

[61] Gregory Michalak et al. "A comparison of relative proton stopping power measurements across patient size using dual- and single-energy CT." In: *Acta Oncologica* 56.11 (2017). PMID: 28885130, pp. 1465–1471. DOI: 10.1080/0284186X.2017.1372625.

[62] C E Metz and K Doi. "Transfer function analysis of radiographic imaging systems." In: 24.6 (Nov. 1979), pp. 1079–1106. DOI: 10.1088/0031-9155/24/6/001.

[63] Samuel Richard et al. "Towards task-based assessment of CT performance: System and object MTF across different reconstruction algorithms." In: *Medical Physics* 39.7Part1 (2012), pp. 4115–4122. DOI: https://doi.org/10.1118/1.4725171.

[64] Steven Smith. *The scientist and engineer's guide to digital signal processing*. California Technical Pub, 1999. ISBN: 0966017668.

[65] R. W. D. Nickalls. "Viète, Descartes and the Cubic Equation." In: *The Mathematical Gazette* 90.518 (2006), pp. 203–208. ISSN: 00255572.

[66] William Press. *Numerical recipes : the art of scientific computing*. Cambridge, UK New York: Cambridge University Press, 2007. ISBN: 9780511335556.

[67] William G. Jacoby. "LOESS: a nonparametric, graphical tool for depicting relationships between variables." In: *Electoral Studies* 19.4 (2000), pp. 577–613. ISSN: 0261-3794. DOI: https://doi.org/10.1016/S0261-3794(99)00028-1.