A beamlet-free approach to proton therapy treatment planning

Reducing computational costs by combining Monte Carlo dose calculation and spot weight optimization in a single efficient algorithm

Danah Pross

A thesis presented for the degree of
Doctor of Biomedical and Pharmaceutical Sciences

Molecular Imaging, Radiotherapy and Oncology (MIRO)
Institut de Recherche Expérimentale et Clinique (IREC)
Université catholique de Louvain (UCLouvain)
Belgium

Danah Pross was supported by the F.R.S-FNRS through a Télévie grant.

Acknowledgements

This thesis is the result of four long, exciting, frustrating, fun, incredible years that at the time of writing feel somehow far too short. I don't know that I would have arrived at this point if it weren't for all the amazing people who supported and accompanied me on this journey. So thank you to everyone who helped me take it.

First of all, I want to thank my supervisors, John, Kevin and Edmond. John, you were always there for me, from the very beginning of this project to its end. Thank you for all your advice, the fruitful discussions and not letting the focus slip too far off the math. But maybe even more than that, thank you for being a true mentor. I always felt like I had a lot of freedom to conduct my research, to make my own decisions and to grow as a researcher and as a person. Kevin, thank you for your neverending support. It is always difficult when a supervisor leaves the lab, but you never abandoned it and I'm immensely grateful that you continued to take the time to be there for me. Your advice was truly invaluable. Edmond, when you get passionate about something, you're like a steam train, pushing forward full speed, infecting others with your enthusiasm, but never losing sight of the goal. Thank you for the lively discussions, believing in the project and putting into perspective what is relevant. And I also want to extend my thanks to Ana. Even though you were not officially my supervisor, it at times really felt like you were. You were always there, endlessly supportive and always willing to lend a listening ear or give a soft nudge when it was needed. You are the true backbone of MIRO, you are there for everyone, always aware of what is going on in the lab. You never seemed to have a single doubt that we can achieve whatever we set our minds to and reminded me of that when I doubted myself.

I also wish to thank all my committee and jury members for lending their expertise to this project and for their patience with the always challenging date finding.

It sounds so trivial and yet I was at some point genuinely worried if there would be a defense for the simple reason that finding a date that worked for everyone seemed impossible. Thank you all for making it work in the end.
A big thank you to all the lovely people at MIRO, past and present. It is said that the factors that contribute to the success of a PhD thesis are the colleagues and research group it is done in, the supervisors and the topic in precisely this order and I couldn't agree more. Every single PhD student I have spoken to (and I've spoken to a lot) has had a point in their PhD journey where they doubted everything. I've found that the best medicine in these cases is a beer with colleagues after a long conference and an honest discussion of our own doubts and each others successes. Beyond that, it is an amazing skill to make going to the office feel like fun. I couldn't have asked for a better PhD experience.

I would also like to thank all my friends and family who have supported and endured me during this exciting time. Thank you for your help, for believing in me, for listening to me complain that my code isn't working, for proofreading, for listening to me complain that I really want to be done with writing this thesis, for listening to my gloating about how amazing MIRO is, for distracting me when I needed it, and for simply being the best! A special thank you to Fabi and Stephen for all the help proofreading!

Thank you all for not making me regret the decision to start this journey for even a moment, for giving me the joy of being able to look back at the last four years and say that I thoroughly enjoyed them and thank you for giving me the courage to look towards the future with determination.

Abstract

Proton therapy is able to deliver highly conformal dose distributions with a significant reduction in dose to healthy tissue compared to conventional radiotherapy. To turn these theoretical advantages into clinical reality, however, a complex time- and resource-intensive treatment planning process is needed, including highly accurate Monte Carlo dose calculations and robust optimization. Robust optimization can already be challenging for clinics due to the long runtime and high memory usage that often require a reduction of plan complexity. These issues are amplified in novel treatment techniques such as adaptive proton therapy and proton arc therapy, which are highly dependent on fast, memory-efficient treatment planning tools. This work seeks to address these challenges by proposing a deviation from the conventional treatment planning workflow. The conventional workflow requires a separate beamlet dose calculation for each spot, the concatenation of the individual beamlet dose distributions to a large-scale dose influence matrix and matrix operations in the optimization. The proposed beamlet-free treatment planning approach avoids the computation and handling of the huge dose influence matrix by performing the dose calculation and optimization in a concurrent fashion. Over the course of this thesis this new method is developed, refined, extended and validated. This cumulates in a novel robust optimization method for proton therapy that addresses current issues and opens new pathways for even more elaborate treatment modalities.

PhD committee

Promoters

Prof. John A. Lee, PhD, Ir
Molecular Imaging, Radiotherapy and Oncology, Université catholique de Louvain
École polytechnique de Louvain, Université catholique de Louvain
Dr. Kevin Souris, PhD
Ion Beam Applications S.A. (IBA)
Molecular Imaging, Radiotherapy and Oncology, Université catholique de Louvain
Prof. Edmond Sterpin, PhD, Ir
Molecular Imaging, Radiotherapy and Oncology, Université catholique de Louvain
Laboratory of Experimental Radiotherapy, Katholieke Universiteit Leuven
Particle Therapy Interuniversity Center Leuven

President of the jury

Prof. Frédéric Lecouvet PhD, MD
Radiology, Cliniques Universitaires Saint Luc
Pôle d'Imagerie Médicale (IMAG), Université catholique de Louvain

Members

Prof. Nick Reynaert, PhD, Ir
Medical physics department, Institut Jules Bordet
Radiophysics and MRI physics laboratory, Université libre de Bruxelles

Prof. Ana Maria Barragán Montero, PhD
Molecular Imaging, Radiotherapy and Oncology, Université catholique de Louvain

Prof. François Glineur, PhD, Ir
Institute of Information and Communication Technologies, Electronics and Applied Mathematics, Université catholique de Louvain

Prof. Laurent Jaques, PhD, Ir
Institute of Information and Communication Technologies, Electronics and Applied Mathematics, Université catholique de Louvain

Invited members

Prof. Zoltán Perkó, PhD
Faculty of Applied Sciences, Department of Radiation Science and Technology, Section Medical Physics and Technology, TU Delft

Dr. Niklas Wahl, PhD
Medical Physics in Radiation Oncology, Radiotherapy Optimization, DKFZ Heidelberg

List of publications

Publications

Beamlet-free optimization for Monte-Carlo-based treatment planning in proton therapy
D. Pross, S. Wuyckens, S. Deffet, E. Sterpin, J. A. Lee and K. Souris. Medical Physics, 1-9, 2023.

Scalable, memory-efficient robust proton therapy optimization through beamlet-free treatment planning
D. Pross, V. Dormal, K. Souris, A. Barragán, E. Sterpin and J. A. Lee. Submitted to Medical Physics, 2024.

Conference communications

D. Pross, K. Souris, J. Lee, *Optimization strategies for beamlet-free treatment planning in proton therapy*, Annual meeting of Belgian Hospital Physicists Association (BHPA), 2023, Brussels (Belgium)

D. Pross, K. Souris, J. Lee, *Comparison of beamlet-based and beamlet-free spot weight optimization for treatment planning in proton therapy*, 62th Annual Conference of the Particle Therapy Cooperative Group (PTCOG 62), 2023, Madrid (Spain)

D. Pross, K. Souris, J. Lee, *Optimization strategies for beamlet-free treatment planning in proton therapy*, Belgian Symposium on Transversal Radiation Oncology (B-STRO), 2023, Brussels (Belgium)

D. Pross, K. Souris, J. Lee, *Fast and efficient beamlet-free robust optimization in proton therapy planning*, 63th Annual Conference of the Particle Therapy Cooperative Group (PTCOG 63), 2024, Singapore (Singapore)

D. Pross, V. Dormal, K. Souris, A. Barragán, E. Sterpin, J. Lee, *Reduced time and memory requirements for robust proton therapy planning through a beamlet-free approach*, Annual meeting of Belgian Hospital Physicists Association (BHPA), 2025, Brussels (Belgium)

List of abbreviations

AC ..... Anal canal
ADMM ..... Alternating direction method of multipliers
ADMM-BB ..... ADMM with a Barzilai–Borwein step size
AI ..... Artificial intelligence
APS ..... Adaptive particle sampling
BD ..... Biological surrogate dose
BFGS ..... Broyden-Fletcher-Goldfarb-Shanno algorithm
BS ..... Brain stem
CBCT ..... Cone beam computed tomography
CI ..... Conformity index
CPU ..... Central processing unit
CT ..... Computed tomography
CTV ..... Clinical target volume
DIR ..... Deformable image registration
DNA ..... Deoxyribonucleic acid
DVH ..... Dose volume histogram
GPU ..... Graphics processing unit
GTV ..... Gross tumor volume
HI ..... Homegeneity index
HU ..... Hounsfield units
ID ..... Integral dose
IMAT ..... Intensity modulated arc therapy
IMPT ..... Intensity modulated proton therapy
IMRT ..... Intensity modulated radiotherapy
L-BFGS ..... Limited-memory BFGS algorithm
LET ..... Linear energy transfer

LFH ..... Left femoral head
LON ..... Left optical nerve
MC ..... Monte Carlo
MRI ..... Magnetic resonance imaging
MSE ..... Mean square error
NTCP ..... Normal tissue complication probability
OAR ..... Organ(s) at risk
OC ..... Optical chiasm
PAT ..... Proton arc therapy
PBS ..... Pencil beam scanning
PET ..... Positron emission tomography
PDD ..... Percentage depth dose
PGD ..... Projected gradient descent
PT ..... Proton therapy
PTV ..... Planning target volume
QA ..... Quality assurance
RBE ..... Relative biological effectiveness
RFH ..... Right femoral head
ROI ..... Region of interest
RON ..... Right optical nerve
SGD ..... Stochastic gradient descent
SCD ..... Stochastic coordinate descent
TCP ..... Tumor control probability
TPS ..... Treatment planning system
VMAT ..... Volumetric modulated arc therapy

Contents

Chapter 1

Overview

1.1 Research motivation

Cancer is one of the leading causes of death worldwide, with almost 10 million cancer deaths in 2022 [1]. This number is only expected to increase in years to come due to increases in both population size and longevity.
Over the years, different cancer treatment modalities have been discovered and refined. The different modalities vary strongly in how they affect and target specific forms of cancer. The choice of which modalities are used depends on the type of cancer, the stage it is in, whether it is localized and its location in the body. Some, like chemotherapy and immunotherapy, are non-localized and are able to act across the body, while others such as surgery and radiotherapy are only able to target localized tumors in accessible locations, but also limit the impact of side effects. In many cases, a combination of several treatment modalities is used.
External beam radiotherapy is one such modality and is recommended for about half of all cancer patients [2, 3, 4]. It works by using beams of ionizing radiation, which can cause DNA damage and cell death. A combination of factors is exploited to increase the damage within certain regions, the tumor, while reducing it in others, the healthy tissue and organs.
While the most common form of radiotherapy utilizes photon beams as their ionizing radiation, charged particles such as electrons, protons and ions have the potential to create preferential dose distributions due to their unique physical characteristics. Instead of traversing the patient like high-energy photons do, they have a finite range and the additional advantage of losing most of their energy close to the end of this range in a phenomenon known as the Bragg peak. Current clinical standard is intensity modulated proton therapy, where several thin, monoenergetic proton pencil beams are combined to create a homogeneous dose region covering the tumor. To achieve this, the individual pencil beams are assigned weights, which are optimized in a computational pre-treatment process, the treatment planning.
Treatment planning for proton therapy involves several steps, including imaging, segmentation and plan creation, as described in section 2.2.2. After patient image data in the form of a voxelized geometry has been obtained, the tumor and relevant organs are contoured. A number of beams, each defined through a unique set of angles, are determined. Each beam is made up of several thin pencil beams. The tumor is then devided into energy layers orthogonal to the beam direction and a spot grid is placed across each energy layer to cover the tumor. A spot refers to the location of the Bragg peak of a pencil beam. The dose distribution created by a single pencil beam is known as a beamlet. Before the spot weights can be optimized, their individual impact on the total dose distribution d is determined by calculating their beamlet dose distribution and the concatenation of all 3D beamlet dose distributions in the so-called beamlet matrix $D_{i,j}$, which contains

the dose contribution of each beamlet j to each voxel i.
In radiotherapy with photons, dose calculations are often performed using analytical dose calculation engines, that predict the 3D dose distributions based on the well known shape of the photon beam.
Analytical dose calculation engines are available for protons as well and model the shape of individual pencil beams, however, they are often insufficient to accurately model the dose, due to the sharp dose gradients involved, especially around the Bragg peak, and the strong dose distortions caused by heterogeneous media.
The most accurate way of calculating dose distributions are Monte Carlo (MC) particle transport methods, which are presented in section 2.4. Monte Carlo methods are comparatively much slower than analytical dose calculation tools though, which greatly increases the time required for treatment planning.
Another challenge in the treatment planning process is the size of the beamlet matrix D. Since both the number of spots and the number of voxels are high, this is a huge matrix with considerable memory needs.
In the subsequent spot weight optimization step, the beamlet matrix is multiplied with the spot weights x_j to obtain the total dose distribution

$$\mathbf{d} = \sum_j D_{ij} x_j \ . \tag{1.1}$$

Large scale matrix multiplications and optimizations based on them are computationally heavy and increase the computation time considerably.
Both the time required for beamlet dose calculation and the time and memory requirements during optimization depend on the number of voxels and spots. To decrease them, one can reduce the number of spots by placing them further apart, which might lead to less homogeneous dose distribution, or reduce the number of voxels by rescaling the voxel grid of the patient data or using an independent dose grid with a coarser voxel distribution. Both of these strategies, while effective in reducing computational costs, reduce accuracy and optimization quality.
It would be desirable, especially for big plans with multiple beams, to remove the dependency on the number of spots entirely. However, that would require the omittance of the calculation of and optimization based on the individual beamlet dose distributions and the beamlet matrix.
The aforementioned problems are amplified when using methods such as robust optimization, as described in section 2.5.4. Novel treatment approaches, such as proton arc therapy (PAT) or online adaptive proton therapy, are also affected by the high computational costs during treatment planning. Arc proton therapy is the proton therapy equivalent to volumetric modulated arc therapy (VMAT), where photons are delivered by a gantry rotating around the patient. Different strategies in optimal Arc therapy optimization are being explored, but they usually

require a significant increase in the number of spots. Online adaptive therapy on the other hand does not necessarily require an increase in the number of spots, but it does rely on very fast reoptimizations, including dose recalculations. Online real-time adaptations seems an impossibility based on current treatment planning standards.

Because of these challenges and hurdles, new strategies in proton therapy treatment planning are needed and are an open field of research that the present work seeks to contribute to.

Relevant works in the field of proton therapy treatment planning typically address either the dose calculation or the optimization aspects. In dose calculation, relevant advances include the development of faster Monte Carlo tools through the use of parallelization or approximations [5, 6, 7, 8, 9, 10, 11, 12, 13, 14], or the development of more accurate analytical models [15]. More details on current dose calculation tools can be found in section 2.4.

With regards to the optimization, there are papers investigating how to reduce the computational impact of the optimizer in particularly challenging cases like robust optimization [16, 17, 18, 19], as discussed in section 2.5.4.

There are altogether few papers investigating a deviation from this established workflow. Yang et al. [20] proposed a concurrent MC dose calculation and fluence optimization algorithm for photon therapy and Li et al. [21, 22] developed a similar approach both for photon and proton therapy. In their method referred to as adaptive particle sampling, they simulate a total of n_j^k protons for each spot j proportional to the spot weight x_j at iteration k. This results in higher statistical accuracy in high intensity spots and allows the reduction of particles simulated towards low intensity spots. However, all of these methods still rely on a beamlet matrix and consequentially need to exert computation time and memory for the resulting matrix operations. The reported benefits, an efficiency increase of $41.6 \pm 15.3\%$ [22], result from improved spot targeting during the Monte Carlo simulation.

Over the course of this thesis, a novel treatment planning approach is investigated, that seeks to remove the need of individual beamlets altogether. A beamlet-free treatment planning approach was proposed, implemented and compared against conventional methods with an emphasis on plan quality, computation time and memory requirements. The suitability of this approach to computation-heavy robust optimization was investigated and a robust beamlet-free optimizer implemented and validated.

1.2 Main contributions

This project aims to address the various challenges in proton therapy treatment planning and propose a new beamlet-free treatment planning approach to mitigate them. Over the course of the thesis, a beamlet-free optimizer was implemented and compared against conventional methods.

1.2.1 Beamlet-free optimization

A beamlet-free optimization algorithm was developed and integrated within the CPU-parallelized Monte Carlo code MCsquare [12]. It follows a concurrent dose calculation and spot weight optimization, similar to the methods proposed by Yang et al. [20] for photons and Liu et al. [23] for protons, but differs in its concrete implementation. In particular, it is unique in that it doesn't require the calculation of a beamlet matrix at all.
The new method is validated against conventional treatment planning approaches on several patient cases.
This work lead to a publication which chapter 3 is based on [24].

1.2.2 Expected-value beamlet-free robust optimization

In a second part, the beamlet-free framework is adapted for use in robust optimization. The challenges that the beamlet computation and optimization pose on the treatment planning process are exacerbated in robust optimization. There are different approaches to robust optimization that are not equally suitable to a beamlet-free approach. The main robust optimization approaches include worst-case scenario, or minimax, robust optimization [25] and expected-value robust optimization [26, 27]. An expected-value beamlet-free optimizer is presented and compared to conventional worst-case and expected-value robust optimization methods.
A publication resulting from the work on robust optimization presents the basis for chapter 4.

1.2.3 Contributions to the open-source treatment planning system OpenTPS

As described in Chapter 2.3.3, treatment planning is typically performed using a treatment planning system (TPS) combining several individual treatment planning tools. Clinical TPS are commercially available from several vendors, however, they are based on proprietary source code and are hard to modify as that is not their intended use. Researchers working on the development of new treatment

planning tools need accessible, open-source systems that allow them to smoothly integrate and test their proposed methods.
The open-source treatment planning system OpenTPS [28] was developed specifically for proton therapy, but was recently expanded to include photon dose calculations as well. It is written in python making it completely open and easily accessible even for people without a strong programming background. It has served as basis for several proton therapy related research projects so far and was the TPS used for the validation and testing of the presented beamlet-free optimizer in Chapter 3 and robust beamlet-free optimizer in Chapter 4 as well as the respective comparisons.
The beamlet-free optimizer, and its robust successor, have been directly integrated within the code of the open-source MC dose engine MCsquare [12]. Additionally, expected-value robust optimization has been implemented by the author to allow conclusions about the comparative performance of different robust optimization methods to be drawn.

1.3 Structure of the thesis

This thesis is constructed around two main papers and is structured as follows:

Chapter 1: Introduction
The present chapter outlines the research motivation for the thesis project, as well as a list of the main contributions of the author to the research field and the present section on the structure of the thesis.

Chapter 2: Background
The second chapter is dedicated to introducing some key concepts to the reader in order to provide context and orientation. The base concepts of external beam radiotherapy are presented and a motivation for charged particle therapy is provided. The basics of proton therapy are outlined, including an overview of the current proton therapy treatment planning workflows. Monte Carlo methods as a dose calculation tool are examined in more detail due to their importance for the thesis project. Finally, an introduction to optimization theory provides the reader with the necessary mathematical background knowledge to understand the algorithms and methods mentioned later on.

Chapter 3: Beamlet-free optimization for Monte Carlo based treatment planning in proton therapy
The third chapter is centered around the first main publication, which introduces the beamlet-free algorithm. The beamlet-free algorithm is a new, proposed treat-

ment planning approach that seeks to address current challenges with respect to high computational requirements in proton therapy treatment planning. It includes a description of the algorithms, as well as its validation on phantom and patient cases.

Chapter 4: Robust beamlet-free optimization
The fourth chapter is centered around the second main publication, which investigates an expected-value beamlet-free robust optimization approach and compares it to conventional worst-case scenario beamlet-based robust optimization and expected-value beamlet-based robust optimization.

Chapter 5: Discussion and conclusion
The last chapter of this thesis reviews the findings presented in the chapters above in the context of proton therapy research at large. Future perspectives for the proposed methods are discussed and conclusions drawn.

Chapter 2

Background

The following chapter introduces some base concepts related to radiotherapy and optimization theory and can be skipped by readers already well-acquainted with the topics at hand. Since large parts of this background chapter consist of well-established tenants of radiotherapy, the use of citations is reserved for specific works that are to be highlighted or examined in more detail. All other information can be easily verified or extended with additional details by perusing any of the many available resources on medical physics, including

- *An Introduction to Medical Physics* by Maqbool [29],
- *Introduction to Medical Physics* by Keevil et al. [30],
- *Hadron Therapy Physics and Simulations* by Nunes [31],
- *Basic Clinical Radiobiology* by Joiner et al. [32] and
- *Proton therapy physics* by Paganetti [33].

2.1 External beam radiotherapy

External beam radiotherapy is one of several cancer treatment modalities at our disposal. It is used instead of, or in combination with, other treatment modalities, including surgery, chemotherapy, immunotherapy and brachytherapy.
Which modality is chosen for a specific case depends on the type of cancer, the stage it is in, whether it is localized and its location in the body. Radiotherapy is recommended for about half of all cancer patients [2, 3, 4].

2.1.1 Basics of radiobiology

External beam radiotherapy uses one or several beams of ionizing radiation directed at localized tumors. The ionizing radiation traverses the body and transmits energy to particles it encounters along its path. The transferred energy causes ionizations and the creation of free radicals which cause DNA strand breaks. While human cells have extensive DNA repair mechanisms, too many strand breaks at once, and especially more difficult to repair double strand breaks, can lead to cell death [34].
These ionizations do not occur everywhere equally. They depend on the amount of energy that is transferred from the incident ionizing particle to the surrounding tissue. The amount of energy transferred depends in turn, on the type of radiation and particle used, its initial energy and the material it is traversing. The effect that the particle has on the tissue is quantified and reported in terms of dose. The

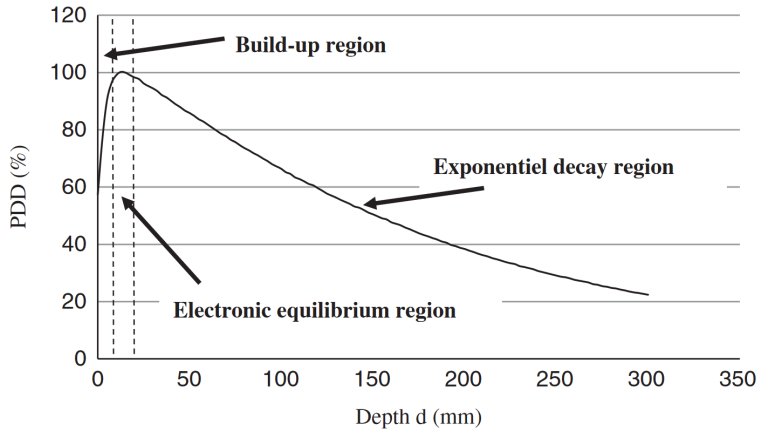


Figure 2.1: Depth dose curve for 6 MV photons. (Figure taken from [35].)

dose d is defined as the energy E that is actually absorbed per unit mass m,

$$d = \frac{E}{m} \ , \tag{2.1}$$

and is measured in Gray.
Photons are by far the most common choice of particle used in radiotherapy and the energy they transfer to surrounding tissue depends on their initial energy and their current depth. A depth dose curve as in Figure 2.1 shows the dose absorbed at a certain depth integrated over the other dimensions. The percentage depth dose (PDD), is a normalization of the dose such that the maximum corresponds to 100.

The absorbed dose initially increases until a maximum is reached after which it decreases exponentially. The reason for the initial increase, also known as the build-up region, lies in the fact that the incident photons transfer most of their energy to electrons through particle interactions. These electrons are considered secondary particles and travel themselves through matter until they have lost all their energy in further particle interactions. The amount of energy transferred in a single interaction is small compared to typical beam energies in charged particle therapy and the secondary electrons only travel a short distance before having transferred their entire energy to the local tissue. The peak in the depth dose curve coincides with the point at which the number of electrons created by the incident photons equals the number of previously created secondary electrons reaching the end of their range. The exact depth of the maximum dose depends on the energy of the photon beam and the traversed material.

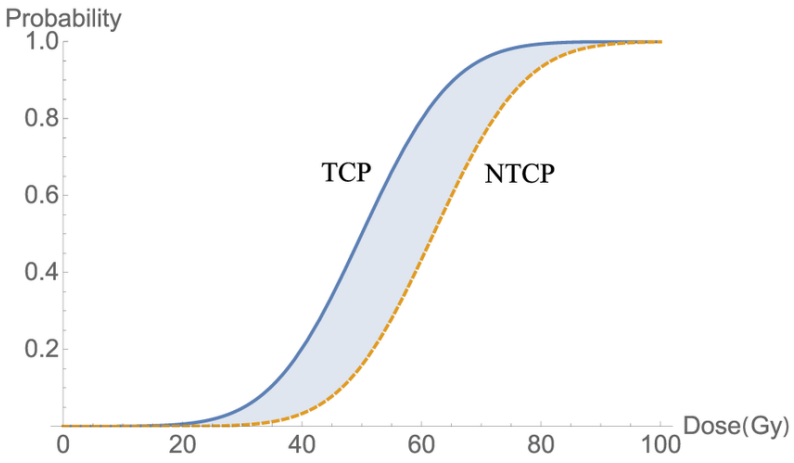


Figure 2.2: Therapeutic Window (Figure taken from [36].)

2.1.2 Therapeutic window and tumor control

Radiotherapy seeks to cause DNA damage and cell death through ionizing radiation. It is not possible to restrict the resulting damages to only tumor cells, however, a combination of different factors can be exploited to maximize damage to the tumor and limit damage to healthy surrounding tissue. Healthy cells are better able to repair damage to the DNA than cancer cells. This phenomenon can be exploited to deliver a dose that falls in the therapeutic window. The tumor control probability (TCP) describes how high a dose needs to be to reliably kill cancer cells. The normal tissue complication probability (NTCP) on the other hand gives probability that a healthy cell is killed at a certain dose, which allows conclusions to be drawn on how high the dose to healthy tissue may be while still allowing a good percentage of healthy cells to survive. The therapeutic window refers to the colored section between the TCP and NTCP curves, as visualized in Figure 2.2. A wider therapeutic window is desirable in radiotherapy, and there are strategies that actively try to increase it.
Typically a radiotherapy treatment is not delivered all at once, but over the course of several sessions that may stretch across multiple days or weeks. Such a treatment is referred to as fractionated radiotherapy, where each session a fraction of the total prescribed dose is delivered. Fractionation brings several advantages, among them the ability to limit the dose to healthy tissue per fraction to a level that can mostly be repaired, while exceeding these levels for the tumor region.
Fractionation also allows the exploitation of the oxygen effect [37]. About two thirds of the DNA damage caused by ionizing radiation is caused through free radicals which are created as a consequence of the ionization of oxygen atoms

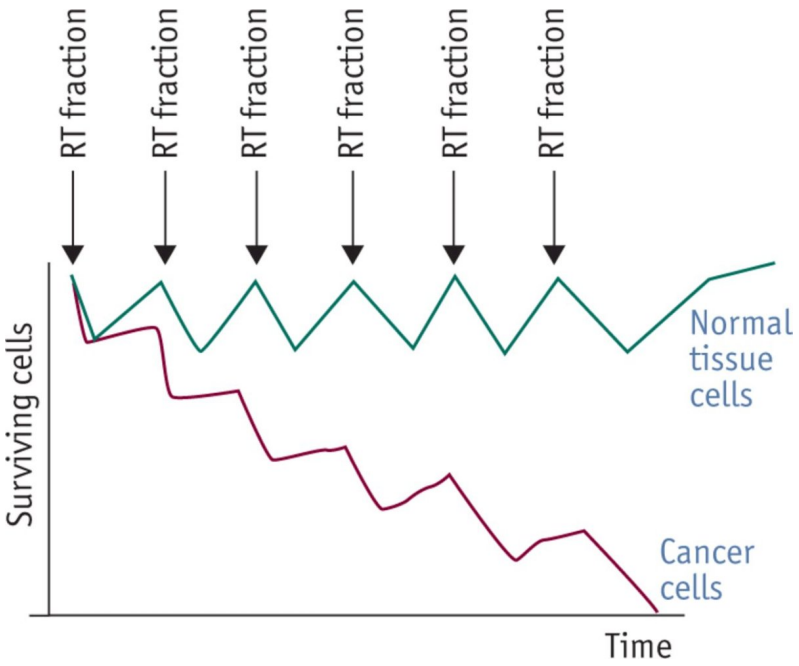


Figure 2.3: The impact of fractionation on cell survival. (Figure taken from [39].)

[38]. As a consequence oxygen levels of cells impact their radiosensitivity. Since the peripheral sections of tumors tend to be better oxygenated, they are first to be killed during radiotherapy. Fractionation allows the reoptimization of the new peripheral cells before the next treatment session and therefore the maximization of its impact.
The survival fraction (SP) describes the fraction of cells that survive after receiving a non-lethal amount of radiation between two fractions. It can be calculated with the linear quadratic model as

$$\begin{aligned} \text{SF} &= e^{-\alpha n d - \beta n d^2} \\ &= e^{-\alpha D - \beta d D} \end{aligned} , \tag{2.2}$$

where d is the dose per fraction, $D = nd$ the total dose over n fractions and α and β tissue dependent constants. The impact of fractionation on the cell survival of tumor and healty tissue cells is presented in Figure 2.3.

To further limit normal tissue complications, a combination of several beams is employed to create a high, homogeneous dose level in the tumor. By utilizing multiple beams overlapping within the tumor the dose delivered to the surrounding healthy tissue can be kept below the target prescription. The impact of different beam configurations on the dose distribution is demonstrated in Figure 2.4. It can

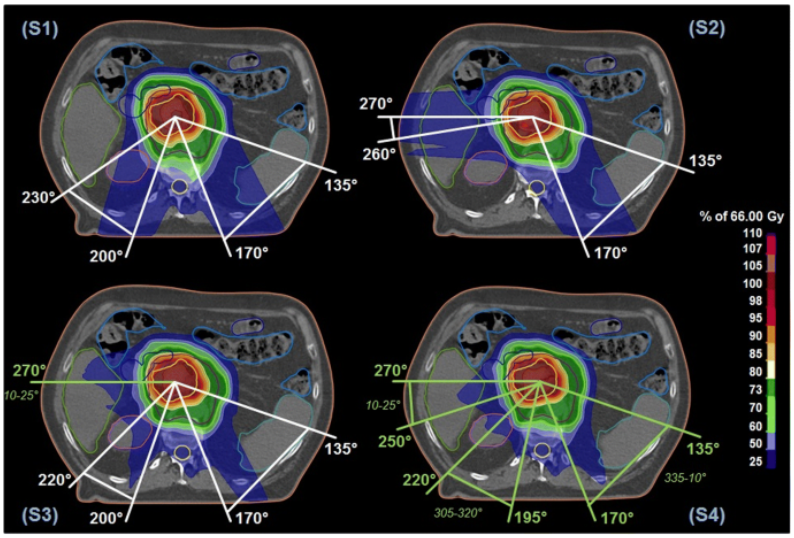


Figure 2.4: The impact of different beam configurations on the dose distribution. The used beam configurations are (S1) two posterior oblique beams, (S2) a lateral right beam and a left posterior oblique beam, (S3) two oblique posterior beams plus a right-sided non-coplanar beam, and (S4) three non-coplanar beams from posterior and the right side. The beam angles are indicated in white for coplanar beams and in green for non-coplanar beams. The beam angles are given in ranges utilized across several patients in the respective study by Stefanowics et al. (Figure taken from [40].)

be observed that using a three beam instead of a two beam configuration reduces the dose outside the target region. In intensity modulated radiotherapy (IMRT) the dose is sculpted through optimization to achieve the desired distribution.

Utilized together, these techniques result in a higher dose to the tumor region with a high chance of cell death and a comparative lower dose to healthy surrounding tissue that, while unable to completely prevent damage, modulates the damages to normal tissue and to critical organs.

Another option to reduce the dose to the surrounding tissue relative to the dose to the tumor is to switch from using photons, common in conventional radiotherapy, to charged particles, like electrons, protons or ions, which display a preferential dose profile. The dosimetric advantages and particularities of charged particle therapy are explored in the next section.

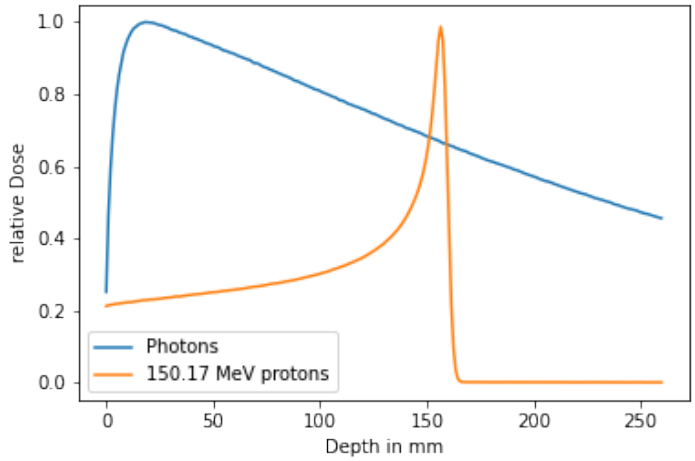


Figure 2.5: Relative depth dose profile of monoenergetic photon and proton beam in water.

2.2 Proton therapy

Proton therapy is a specialized form of radiotherapy that utilizes proton beams, instead of the more common photon beams. Both, photons as well as protons, are forms of ionizing radiation that can, in essence, be used to exploit the basic principles of radiotherapy. Where they differ is the manner in which they interact with matter and give off their energy. The number of photons travelling through matter decreases exponentially with the distance traveled. How much energy is deposited depends on the initial photon energy, the depth at which the measurement is taken and the material it traverses. Protons, and other charged particles, exhibit a different depth-dose behavior. They have a finite range that depends on initial energy and traversed material. In protontherapy, the particle energy can be chosen in such a way that the end of their range does not exceed the distal edge of the target region, drastically reducing the dose to surrounding healthy tissue. This almost completely eliminates any dose that would be deposited behind the tumor.
Additionally, the dose deposited by protons does not exponentially decrease like for photons. Instead, the deposited dose is relatively constant until the proton reaches the end of its range. Then, the protons start transferring more energy to the surrounding matter in a behavior known as the Bragg peak. This difference is easily demonstrated in a depth-dose profile like in Figure 2.5.

The reason for this difference lies in the physics of particle interactions. The

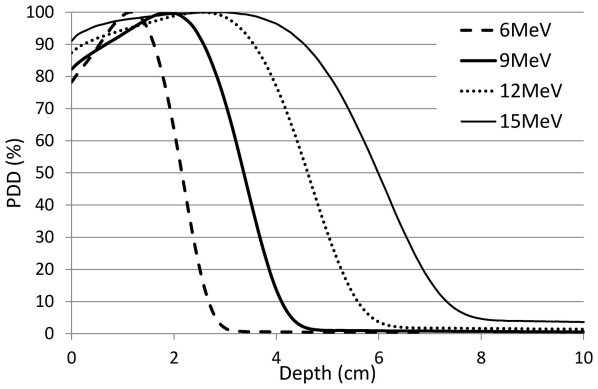


Figure 2.6: Electron depth dose curves at various energies. (Figure taken from [41].)

interactions a particle experiences depend on several factors and will be examined in more detail in section 2.4.

2.2.1 Why protons? - Other forms of charged particle therapy

These dosimetric advantages are not unique to protons and are shared between charged particles in general. However, not all kinds of charged particles are suitable for radiotherapy. Electrons are used in some cases, but they are typically only used for tumors at relatively shallow depths due to the high costs associated with accelerating them to higher energies and lead to high scattering and are therefore restricted to specific cases. The relative depth dose profiles of electrons, photons and protons are displayed in Figure 2.6.

Heavier ions can be used as well. The most common type are carbon ions which are considered the most effective particle type due to their superior radiobiological effectiveness, but the dosimetric differences between ions heavier than protons are relatively slim. Proton therapy is the most common, since they are easy to produce and were among the first particles studied.

2.2.2 Cost of proton therapy

Despite the seemingly obvious advantages from a physics perspective, proton therapy is still rare compared to conventional radiotherapy with photons. A big reason for this is the increased cost of proton therapy [42, 43]. Photons can be produced by a Cobalt-60 source or, as is more common today, a linear accelerator, a so called Linac. Complete treatment machines based on either technology can

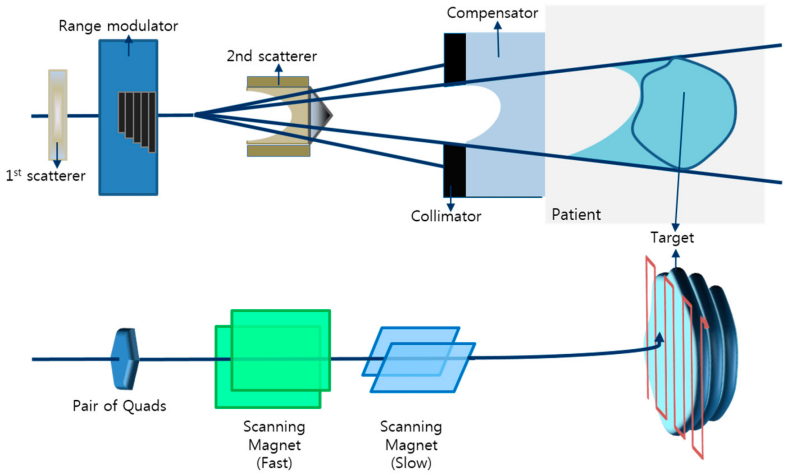


Figure 2.7: Visual representation of passive scattering (top) and active spot scanning (bottom) beam delivery. (Figure taken from [46]).

be directly purchased from manufacturers and placed into a treatment room. Protons require a much bigger accelerator to achieve the required energies for clinical applications. Clinics are typically outfitted with a cyclotron, or a synchrotron, which services one or several treatment rooms to minimize beam downtime. Due to this, particle therapy centers usually require considerable upfront building cost, as well as more space, specifically trained experts and, if desired, huge and complex gantries and higher maintenance costs.
Due to these high associated cost, proton therapy is typically only used for specific indications, and it is closely monitored that the theoretical dosimetric benefits can indeed be translated into improved patient outcomes for the selected groups [44]. These issues are only amplified for heavier particles such as carbon ions [45].

2.2.3 Beam delivery

The finite range and narrow Bragg peak necessitate a few adaptations in proton beam delivery compared to photons. While in radiotherapy with photons, high-energy photons are created and the beam is shaped and modulated through the so called multileaf collimators, proton therapy requires a different approach.

One way to modulate a proton beam is to accelerate the protons to a specific energy and then shoot the beam through an aperture equipped with a range modulator [47, 48]. The range modulator is a piece of matter, often ridges or pins, that force protons to loose part of their energy passing through it, thus decreasing their range. By varying the material or thickness, the proton range can be ad-

justed flexibly throughout the beam. This method is known as passive scattering. Modern proton therapy centers, usually opt for a technique called pencil beam scanning instead [49, 50]. Here, individual thin pencil beams are delivered. To this end, a spot grid is defined with each spot corresponding to the Bragg peak location of a pencil beam. These spots are arranged in distinct energy layers in beam direction. An energy layer corresponds to a specific depth within the patient. All spots in one energy layer are delivered and then the cyclotron accelerates a new batch of protons for the next energy layer. Magnets are used to adapt the Bragg peak position perpendicular to the beam direction.
Both beam delivery techniques are illustrated in Figure 2.7. When discussing the challenges of treatment planning for proton therapy in the following, pencil beam scanning is assumed.

2.3 Proton therapy treatment planning

2.3.1 Treatment planning workflow

Treatment planning is a necessary process at the beginning of each radiotherapy treatment. The treatment plan is comprised of information on the number and angle of beams, the spot placement and weights, and represents the relevant machine input parameters needed for delivery. In proton therapy, treatment planning consists of the following steps:

- Pre-treatment imaging
- Segmentation of regions of interest
- Definition of dose prescriptions and limits
- Plan creation
- Beamlet dose calculation
- Spot weight optimization
- Final dose calculation
- Plan evaluation
- Reoptimization/-evaluation (if necessary)

Pre-treatment imaging

Before the radiotherapy treatment can commence, high quality 3D images of the patient need to be obtained. This usually involves acquiring CT data, which allows conclusions to be drawn about the electron density or material composition. Information on the material composition and density is required for the dose computation later on. It is one of the key factors that determine the energy loss during particle transport. Magnetic resonance imaging (MRI) or positron emission tomography (PET) might be used in addition to CT data, since they are better able to visualize soft tissue contrasts. In conventional radiotherapy with photons, a workflow consisting of only MRI images is feasible, but requires the creation of a pseudo-CT [51, 52, 53] to still provide electron density information. The use of pseudo-CTs in proton therapy is under investigation [54, 55], but poses considerable challenges due to the considerable effects that density errors can have on the proton range and dose distribution.
The CT data comes in the form of a three-dimensional (3D) matrix, or rather a set of two-dimensional (2D) images called slices, that can be combined to form a 3D matrix, where each entry represents a so-called voxel and contains the matching Houndsfield unit (HU) value. Houndsfield units are defined as

$$\mathrm{HU} = 1000 \cdot \frac{\mu_{\mathrm{tissue}} - \mu_{\mathrm{water}}}{\mu_{\mathrm{water}} - \mu_{\mathrm{air}}} \ , \tag{2.3}$$

with μ_{tissue}, μ_{water} and μ_{air} being the total linear attenuation coefficient of the imaged tissue, water and air respectively. Relevant magnitudes in radiotherapy include the HU values of air, $\mathrm{HU}_{\mathrm{Air}} = -1000$, which appears black on greyscale CT images, and water, $\mathrm{HU}_{\mathrm{Water}} = 0$. Bone takes much higher values and appears white on CT images.

Segmentation of regions of interest

Within the CT data, the tumor is identified and delineated. The visible tumor region is referred to as the gross tumor volume (GTV). It is extended to the clinical target volume (CTV), which contains both the visible tumor and surrounding tissue that is likely to contain microscopic tumor-spread that cannot be identified on the imaging data. The CTV is itself often extended to the planning target volume (PTV) [56, 57], which takes uncertainties, like potential beam alignment errors, CT conversion errors or organ motion, into account. The PTV is supposed to ensure that no part of the CTV is affected from underdosage as a consequence of these uncertainties, however, it is a concept more relevant in conventional radiotherapy with photons. In proton therapy, uncertainties tend to have a much

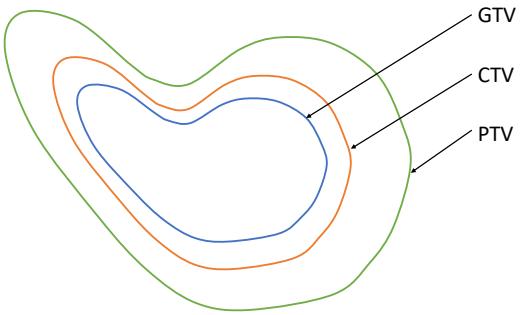


Figure 2.8: Representation of the Gross Tumor Volume (GTV), Clinical Target Volume (CTV) and Planning Target Volume (PTV) concept.

stronger impact on the resulting dose distribution, rendering conventional PTV margins insufficient to properly account for them. Robust optimization is used instead, which will be introduced in section 2.5.4.
The different target volume classifications are visualized in Figure 2.8. If different parts of the tumor receive different dose prescriptions, they need to be segmented independently. Besides the tumor, all relevant organs, the organs-at-risk (OAR), are segmented as well. The segmented regions relevant in the treatment planning process are referred to as regions of interest (ROI).
The segmentation can be performed manually by a radiation oncologist, or performed with segmentation software [58, 59] that is double checked by a clinician [60]. The contours of the ROI are delineated on each 2D slice of the imaging data and usually saved as a binary map.

Definition of dose prescriptions and limits

A dose prescription is set for the target, usually the CTV or PTV. The tumor region can be split into sections with different dose prescriptions, if required. Similarly, dose limits are defined for critical organs at risk (OAR) in the vicinity of the tumor or beam path. These dose prescriptions and limits are realized in dose objectives. There are different objective types including minimum dose, maximum dose, mean dose or dose-volume objectives, which control respectively the minimum dose that any voxel within a ROI should receive, the maximum dose that any voxel within a ROI should receive, the mean dose over all voxel that a ROI

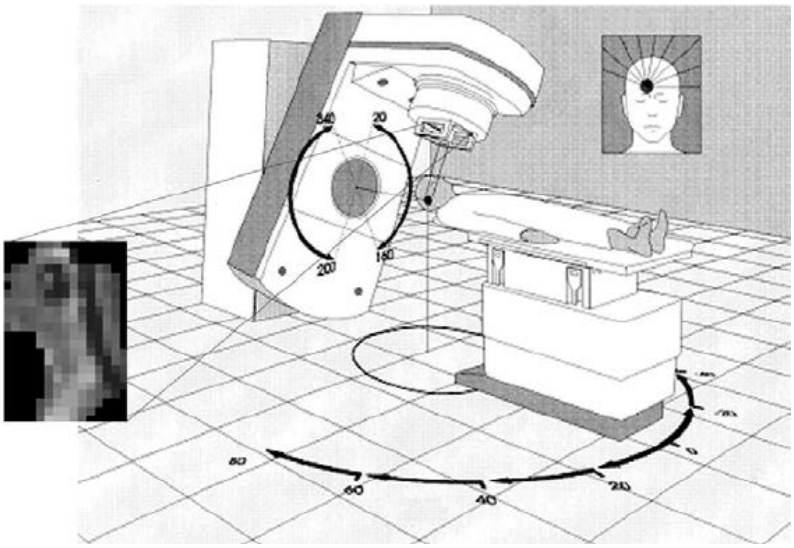


Figure 2.9: Representation of treatment room setup consisting of the rotating patient couch and the revolving gantry. (Figure taken from [65].)

should receive, or the dose that a certain percentage of a ROI should not exceed. The thus defined objectives regulate the aim of the optimization and should be set based on experience and guidelines for an optimal patient outcome.

Plan creation

The plan creation includes the definition of the beams, including their angle and number, and the spot placement.
Beams are described by two angles, the gantry angle and the couch angle. The gantry angle describes the orientation of the gantry, while the beam angle describes the orientation of the treatment couch perpendicular to the gantry angle. The basic setup is depicted in Figure 2.9. Usually the beam configuration is chosen based on clinical experience, but methods for automatic beam angle optimization are being investigated [61, 62, 63, 64].

Range shifters can be used to reduce the energy of the protons beyond the machine's capacity in order to better irradiate shallow regions. They are usually slabs of uniform matter placed in the beam path near the nozzle [66].
In pencil beam scanning IMPT, a 3 dimensional spot grid is placed across the target region. Each spot corresponds to the Bragg peak location of a thin pencil beam. For each beam, spots are placed independently. The spots are sorted in discrete energy layers in beam direction, the distance between layers can be defined

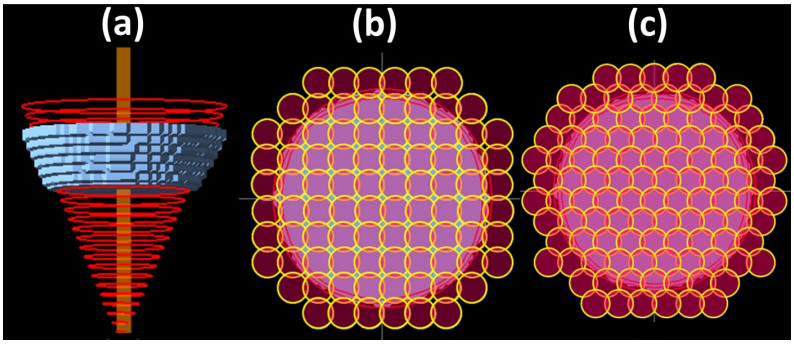


Figure 2.10: Representation of rectilinear (b) and hexagonal (c) spot placement on a conical target (a). (Figure taken from [69].)

by the operator. Within each layer, spots are placed to cover the target region and a user defined margin surrounding it. They are ordered in a grid-like pattern, the concrete shape of which might vary between machines. Most common are evenly spaced spots in a rectilinear or hexagonal pattern, as shown in Figure 2.10, but alternatives have been proposed to optimize dose distributions [67, 68, 66].

Beamlet dose calculation

Each spot corresponds to a thin proton pencil beam known as a beamlet. The beamlets are modulated with the spot weights to achieve the optimal dose distribution. To this end, each beamlet dose distribution must be calculated and stored individually, in order to quantify its impact on the total dose.
For each beamlet, a separate dose calculation is performed based on the patient CT and the resulting dose distribution stored. The beamlet dose distribution can be understood as a 3D matrix or vector whose elements represent the elementary dose contributions of the corresponding pencil beam on each voxel of the CT data. The beamlet dose distributions are combined into a sizeable matrix known as the beamlet or dose-influence matrix D, usually stored in sparse format due to the large number of zero dose voxels. A sparsity of over 90% is not uncommon. The number of voxels depends on the CT scanner utilized during image acquisition and the location of the tumor and relevant geometry. It can be reduced to decrease the required memory, but usually still ranges in the hundred thousands or more. The number of spots depends on the tumor size, spot resolution and number of beams and typically lies in the realm of several thousands. It is significantly bigger for robust optimization.

The total dose $\mathbf{d}$ can then be determined as

$$d_i = \sum_j D_{ij} x_j \ , \tag{2.4}$$

where d_i is the dose to voxel i and x_j is the spot weight of beamlet j, which is at this stage still unknown. It will be determined in the subsequent optimization step. For dose calculations in proton therapy, either analytical pencil beam algorithms [70] or Monte Carlo (MC) tools may be used. Analytical pencil beam algorithms have the advantage with speed, but Monte Carlo tools are considered the gold standard in dose calculation, since they are able to model even complex situations, like highly heterogeneous geometries, with good accuracy [71, 72, 73, 74, 75]. Dose calculation tools and their relevance in proton therapy will be discussed in more detail in section 2.4.
In the presented work, beamlet dose calculations are always performed with MC methods.

Spot weight optimization

The aim of the treatment planning process is to obtain a set of machine parameters that result in the best possible outcome for the patient. This includes the so called spot weights $\mathbf{x}$, which are modulation factors that define the impact of individual beamlets on the total dose distribution $\mathbf{d}$ defined in equation (2.4).
An optimal total dose distribution is one that fulfills the prescriptions and dose limits defined in the previous steps. Compliance with the prescriptions and dose limits is quantified in the objective function $f(\mathbf{x})$, which is applied to an optimization problem:

$$\begin{aligned} \min_{\mathbf{x}} \quad f(\mathbf{x}) &= \sum_k w_k f_k(\mathbf{d}) \\ &= \sum_k w_k f_k(D\mathbf{x}) \\ \text{s.t.} \quad x_j &\geq 0 \quad \forall j \ , \end{aligned} \tag{2.5}$$

where w_k is the weight of objective k and $f_k(\cdot)$ is the objective term.
This process is referred to as inverse planning where the spot weights are derived from the dose information contained in the dose influence matrix. It mostly replaced the so-called forward planning where plan parameters are chosen based on experience and a subsequent dose calculation is used to determine plan quality. Parameters are then manually adapted as necessary in a "trial and error" approach. More information on optimization in general, and the spot weight optimization in

particular, can be found in section 2.5.

Final dose calculation

Once a suitable set of spot weights $\mathbf{x}$ is identified, they are used as input in a final dose computation where the resulting dose distribution is calculated. Often, a MC simulation with a high particle number is used, to ensure a high degree of accuracy. Since only the total dose is computed, not individual beamlet dose distributions, the computation time is much lower than during the beamlet dose computation step.

Plan evaluation

The thus obtained treatment plan is evaluated to ensure sufficient plan quality. This usually includes the evaluation dose volume metrics of the target, such as the D_5 and D_{95}. These metrics refer to the highest dose that 5% and 95% of the target volume receive. Both values should ideally be close to the dose prescription. A D_5 higher than the prescription indicates overdosage, while a D_{95} lower than the prescription indicates underdosage within the target. It should be noted, that a fully homogeneous dose distribution is not usually attainable, and the treatment planning process focuses on keeping under- and overdosage of individual voxel within acceptable bounds.
Other dose volume metrics report the dose that a specific volume of the ROI receive, for example $D_{0.03\text{cm}^3}$, or the percentage of the volume of an ROI that receives at least a specific dose, like $V_{60\ \text{Gy}}$ or V_{95} for 95% of the prescription dose. Dose volume metrics can be visually portrayed in dose volume histograms (DVH). A DVH is a cumulative histogram in which the dose (x-axis) is plotted against the volume percentage of a ROI receiving at least this dose (y-axis). An example DVH with several ROI is presented in Figure 2.11.

Dose volume metrics can further be used to calculate secondary plan quality metrics. Other plan quality metrics including

- Homogeneity index (HI): The HI is a measure for how even the dose is distributed across a region, usually the target. The concrete definition varies across different papers, but the following definition will be assumed in this work:

  $$HI = D_5 - D_{95} \ . \tag{2.6}$$

  A lower HI corresponds to a more homogeneous dose distribution.

- Conformity index (CI): The CI is a measure for how well the dose distribution

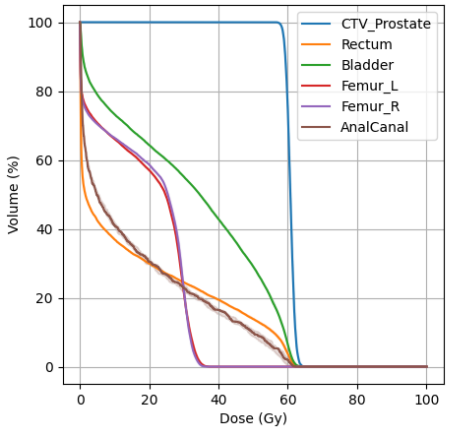


Figure 2.11: Example of a computed DVH for a prostate case showing a target region and several OARs optimized on OpenTPS.

conforms to the shape of a region, usually the target. The concrete definition varies across different papers [76], but the following definition, corresponding to the formulation proposed by Paddick in 2000 [77], will be assumed in this work:

$$CI = \frac{V_{TV,95}^2}{V_{p,95} V_{TV}} \ , \tag{2.7}$$

where $V_{TV,95}$ is the subvolume of the target receiving at least 95% of the prescribed dose, $V_{p,95}$ is the total volume in the patient receiving at least 95% of the prescription dose and V_{TV} is the target volume. A CI of 1 indicates perfect conformality, while a CI of 0 indicates no conformality whatsoever.

- Integral dose (ID): The ID refers to the dose integrated over the entire body. It can give a general measure for the exposure of healthy tissue to radiation.
- NTCP: The NTCP can be used as an additional metric to quantify the risk of complications on the OAR [78].

Additionally, 3D dose maps can be used to visually identify over- or under-dosage.

Reoptimization/-evaluation

If the optimized plan is unable to deliver an acceptable dose distribution, the plan needs to be reoptimized. Ideally this only requires the definition of new dose limits and objectives, proposed by the physician to achieve a compromise between

conflicting constraints. These types of changes require a repeat of the optimization, final dose computation and evaluation steps. If the beam or spot parameters need to be changed, the time expensive beamlet dose computation needs to be repeated as well. This cycle repeats until acceptable plan quality is reached.

2.3.2 Adaptive proton therapy

Most proton therapy treatments are delivered across several fractions. Anatomical differences like tumor shrinkage, weight loss, cavity fillings or differences in patient positioning can lead to vastly different conditions between sessions, and potentially even within one session. These differences can have a strong impact on the resulting dose distribution and can cause significant under- or over-dosage if not accounted for [79, 80, 81, 82, 83, 84]. To prevent this, parts of the treatment planning cycle can be repeated over the course of the treatment. This is referred to as *adaptive proton therapy*.
The field of adaptive PT can be divided into offline adaptive PT, online adaptive PT and real-time adaptive PT.
Offline adaptive PT is performed between sessions in a fractionated treatment. It is therefore not able to account for all types of geometry changes. Since it is performed between sessions, there is ample time to perform the computational plan adaptation. A similar workflow as for the original treatment plan can be followed.
Online adaptive PT is performed right before a treatment session while the patient is already on the treatment couch. It is therefore tailored to the anatomy of the day. In turn it is limited in how much time can be dedicated to the plan adaptation. The adaptation workflow can therefore differ. It typically consists of

- Daily imaging: The most common imaging modality used in APT is cone beam CT (CBCT). It provides good tissue-bone contrast and is suitable for precise image registration. On the downside, it provides poor soft tissue contrast and is prone to image artifacts [85].

- Contouring: Deformable image registration (DIR) can be used to to quickly and efficiently generate the updated ROI contours by reusing information from the original plan. They still require verification by a clinician. AI autosegmentation models start to emerge as an alternative [86], but further research is desirable to ascertain their accuracy.

- Plan adaptation: The plan adaptation can be quite time and resource expensive. A full MC-based replanning would be very costly with current methods, therefore several fast alternatives have been proposed [87, 88, 89,

90, 91, 92, 93]. A full analysis of these methods would exceed the scope of this work. A more comprehensive overview can be found in the review by Paganetti et al. [94].

If one also wants to take intrafraction motion into account, which is an issue for certain treatment sites, one has to go even faster than that. Real-time adaptive PT assumes that image data is obtained during an ongoing treatment session and plan adaptations are performed at such a speed that they can alter the plan during delivery [95]. There are still plenty of technical challenges to be solved on the path to real-time PT, both including hardware and software challenges [96].

2.3.3 Treatment planning system

The treatment planning is typically carried out in a computational system that is equipped to handle most or all the required steps. Such treatment planning systems are available commercially from different companies, both in combination with a treatment machine and separately.
Since these commercial systems are typically black-box systems, there are also some open-source treatment planning systems developed by researchers that allow for easy access and adaptation of the source code. Compared to commercial systems, these open-source systems are typically slower and the code less optimized since it is developed by researchers and intended for use in research, not in clinics. Well known examples for open-source, research treatment planning systems are matRad [97] and OpenTPS [28].

2.4 Monte Carlo dose calculation

Monte Carlo methods are a set of mathematical, usually computational, methods based on the sampling of random variables from a given probability distribution. They are highly flexible methods and can be used for all kinds of applications, including numerical integration and optimization. Their grand benefit lies in the fact that they are able to solve problems that are too complex to be solved by an analytical approach alone, or that are analytically unsolvable.
A drawback of Monte Carlo methods is that they require large numbers of sampled random variables to lower the statistical noise to acceptable levels. This results in substantial computation times and often a trade-off needs to be found between statistical accuracy and computational cost. The statistical noise σ correlates with the sampled number of random variables N with $\sigma \propto \frac{1}{\sqrt{N}}$. Luckily, Monte Carlo methods are inherently parallelizable in nature and significant improvements with respect to computation time can be achieved with modern parallelization

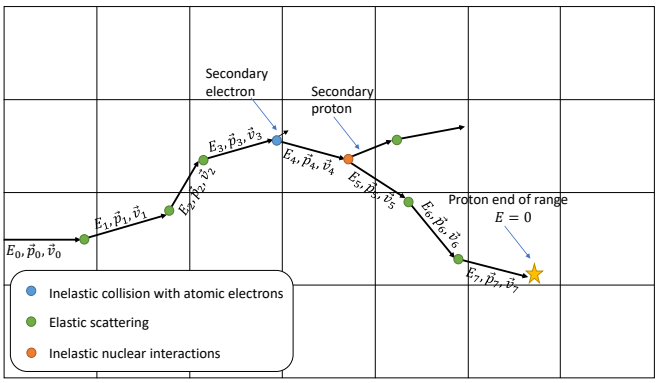


Figure 2.12: Simplified particle track across a voxelized geometry.

techniques.
In physics, their most prominent application is in particle simulations.

2.4.1 Particle transport Monte Carlo

Particle transport Monte Carlo methods are, usually quite extensive, software tools that simulate particle trajectories and model their interactions with other particles or surrounding matter with a high degree of accuracy. Their main uses are in particle physics, where they are used as a theoretical counterpart to large scale particle experiments and to predict particle behaviors. In medical physics they are mainly used as dose prediction tools.
Monte Carlo tools for medical physics applications typically model a phantom or patient as a voxelized geometry, often based on the already voxelized CT data. Particles are moved a certain distance until an interaction occurs. This distance, also referred to as step length, is determined by the interaction cross section of the simulated particle type at the current particle energy within the material of the voxel. A simplified particle trajectory across a voxelized geometry is shown in Figure 2.12. The shown trajectory shows how a charged particle is transported between interactions, at which its energy and direction may change. It should be noted that this visualization is strongly simplified and distances are not an accurate representation of reality.
Heavy charged particles like protons have three primary interaction types.

- **Elastic scattering** is a common type of interaction where the particles kinetic energy remains unchanged but its direction is changed, because the charged particle is affected by the magnetic field of particles in its path.

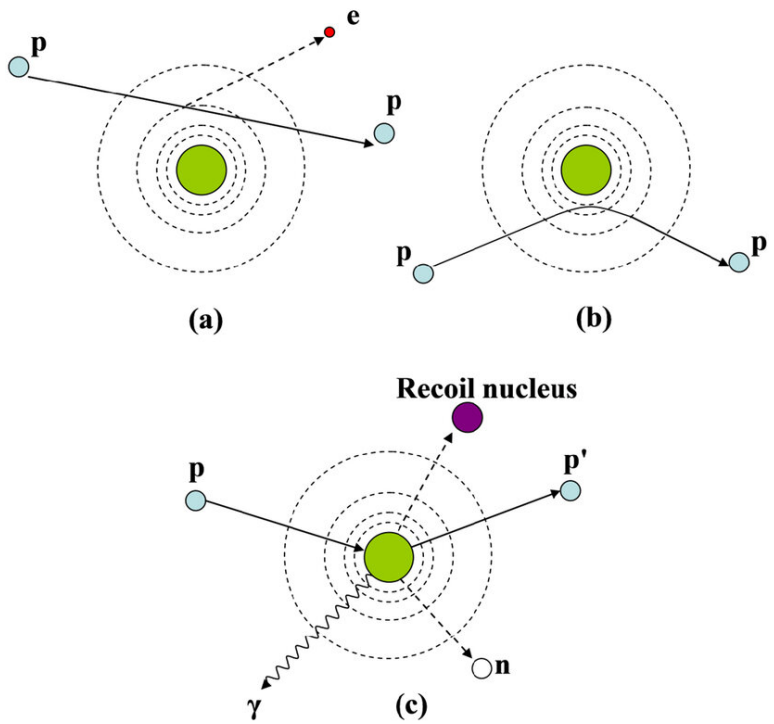


Figure 2.13: The three types of proton interactions with matter: inelastic collisions with atomic electrons (a), elastic scattering (b) and inelastic nuclear interactions (c). (Figure taken from [98].)

.

- **Inelastic scattering with atomic electrons** are another frequent interaction type, during which the particle scatters of an atomic electron, transferring part of its energy. The particle energy is reduced and the electron is knocked from its atomic bond and travels itself through the geometry until its energy is lost.

- **Inelastic nuclear interactions** are rarer. The particle scatters off the atomic nucleus, possibly creating secondary particles.

The three interaction types are depicted in Figure 2.13. How likely each interaction type is to occur depends on the energy and the traversed material.

Often, electromagnetic interactions are simulated in a condensed way. Only certain interactions are simulated explicitly at the end of each step, while others are combined in a stopping power describing the averaged energy loss over the

length of one step, and scattering models that take the combined effect of the soft electromagnetic interactions into account. The step length is usually giving an upper bound, a maximum step length after which a step may be capped if no interaction is occurring. Additionally, a step may be artificially capped at voxel boundaries to ensure accurate modeling at material boundaries. These type of Monte Carlo algorithms are called class-II algorithms, a classification that includes MCsquare [12], the Monte Carlo code used during this thesis.
The stopping power is the mean energy loss a particle experiences while traversing matter. For charged particles, this mean energy loss can be calculated using the Bethe-Bloch equation [99],

$$-\frac{\mathrm{d}E}{\mathrm{d}x} = \frac{4\pi}{m_e c^2} \cdot \frac{nz^2}{\beta^2} \cdot \left(\frac{e^2}{4\pi\epsilon_0}\right)^2 \cdot \left[\ln \frac{2m_e c^2 \beta^2}{I \cdot (1-\beta^2)} - \beta^2\right] \quad , \tag{2.8}$$

where $\mathrm{d}E$ is the energy lost while traversing distance $\mathrm{d}x$ of a material with electron density n and mean excitation energy I, c is the speed of light, e and m_e the electron charge and mass, ϵ_0 the vacuum permittivity, $\beta = \frac{v}{c}$, v the particle velocity and z the particle charge in units of electron charge, meaning $z = 1$ for protons.
Nuclear interactions are simulated explicitly at the end of the step. The cross sections of the relevant interaction types are calculated based on the particle energy and used to modulate the step length. A random variable is drawn to determine which type of interaction occurs. Energy loss, scattering angles and potential secondary particles, are calculated depending on the interaction type.

2.4.2 Applications in proton treatment planning

In proton therapy treatment planning, Monte Carlo tools are ideally used for both beamlet dose calculations, as well as for final dose calculations and evaluations. Since their long computation times pose an issue to efficient clinical workflows, research is done to improve the speed of Monte Carlo simulations.
By creating proton therapy specific tools that only model the interactions relevant for protons for the typical energies and materials in proton therapy, certain approximations can be introduced that significantly reduce computation times compared to general purpose Monte Carlo tools, such as GEANT4 [100] and FLUKA [101], without causing a huge loss of accuracy [5, 102]. For example, for proton therapy specific Monte Carlo tools, the simulation of secondary electrons, the δ-electrons, might be neglected, since their range is typically so short that their energy can be considered to be locally absorbed. The same consideration can be made for many ions and heavier particles created in inelastic nuclear interactions. If only protons,

and a few other particles, such as deuterons and α-particles, are simulated explicitly, certain physics models may be simplified, or look-up tables for stopping powers, may be used to speed up the simulation [102, 12].
Another promising strategy is the development of parallelized Monte Carlo tools. Both tools utilizing CPU-parallelization [12] and GPU-parallelization [8] have been developed.
There are strategies to simplify the Monte Carlo simulations even further to reduce computation times, such as the simplified MC (SMC) proposed by Kohno et al. [5] or track-repeating MC by Li et al. [103], both of which were also combined with GPU-parallelization [7, 6]. However, complex heterogeneities might still require a full Monte Carlo simulation.

2.5 Optimization

Since the core of this thesis project centers around the development and investigation of a novel spot weight optimization approach, this section introduces some basic concepts of optimization theory with a focus on their use and relevance for radiotherapy treatment planning.
Generally, optimization theory is a subfield of mathematics that investigates so-called optimization problems. An optimization problem consists of a vector of optimization variables $\mathbf{x}$, an objective function $f(\mathbf{x})$ and an optimization goal, typically to find the values $\mathbf{x}^*$ that minimizes or maximizes $f(\mathbf{x})$. The objective function may, but is not required to, depend on additional parameters $\mathbf{p}$ or be subjected to constraints.
An optimization algorithm is a method which seeks to solve an optimization problem by determining the values of variables in vector $\mathbf{x}$, which minimize or maximize the objective function involved in the problem while satisfying all constraints. Many optimization algorithms are numerical methods which iteratively approach one of the optimal solutions $\mathbf{x}^*$ to the optimization problem. While they can usually be shown mathematically to converge to an optimal solution, the methods are usually stopped once a preset condition, the convergence criterion, is fulfilled. This is to ensure that the optimization time is kept reasonable, and more importantly, finite. Not every optimization algorithm is suitable for every problem. Most of them can only be guaranteed to converge if the optimization function fulfills certain criteria.
In the following, some key optimization algorithms will be introduced. While there are many other optimization algorithms available, this introduction will be limited to those especially relevant in radiotherapy and in the scope of this thesis in particular.
This section was written primarily based on the lecture notes of the lecture "Opti-

mization models and methods II" as taught by Prof. François Glineur at UCLouvain (2020).
For easier readability, all examples of optimization algorithms below will assume an unconstrained optimization problem in the form of

$$\min_{\mathbf{x}} \quad f(x) \tag{2.9}$$

for a smooth objective function $f(\cdot)$ with $f \in C1$.
If f needs to fulfill certain requirements to guarantee convergence for individual methods it will be specified in the respective subsections.

2.5.1 First-order algorithms

First-order algorithms are a group of optimization algorithms based on the objective function gradient. Their most basic example is gradient descent, which is well known from, for example, machine learning applications.

Gradient descent

Gradient descent is a minimization algorithm which adjusts the optimization variables $\mathbf{x}$ in each iteration by subtracting the objective function gradient ∇f modulated by a step size h as represented in algorithm 1. The gradient descent algorithm is visualized on a simple one-dimensional example objective function in Figure 2.14.

Algorithm 1 Gradient descent

Input data: objective function f, initial vector of optimization variables $\mathbf{x}_0$
Parameters: step size h, stopping criterion
Output: optimized vector of optimization variables $\mathbf{x}$

1: $k \leftarrow 0$
2: **while** Stopping criterion not fulfilled **do**
3: $\quad \mathbf{x}_{k+1} \leftarrow \mathbf{x}_k - h\nabla f(\mathbf{x}_k)$
4: $\quad k \leftarrow k + 1$
5: **end while**

It is applicable on any multi-variate differentiable function f, but convergence to a global minimum can only be guaranteed if f is convex, ∇f is L-Lipschitz, which renders f L-smooth, and the step size h is well chosen. It should be noted that the step size h is not required to be a scalar value, but can instead be defined as a sequence. A typical choice for a constant step size applied to a convex,

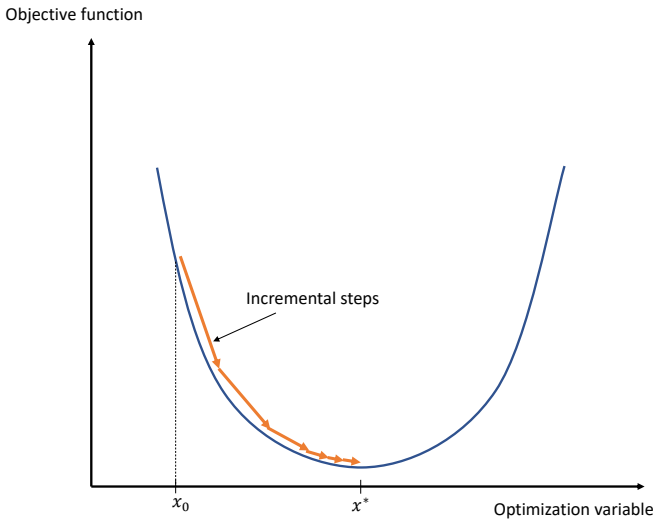


Figure 2.14: Visual representation of the Gradient Descent algorithm on a one-dimensional example objective function.

L-smooth objective function is $h = \frac{1}{L}$ which guarantees convergence of $\mathcal{O}(\frac{1}{N})$ by guaranteeing that after N iterations

$$f(\mathbf{x}_N) - f(\mathbf{x}^*) \leq \frac{L||\mathbf{x}_0 - \mathbf{x}^*||^2}{2N} \ , \tag{2.10}$$

where $\mathbf{x}_0$ is the initial vector of optimization variables, $\mathbf{x}_N$ the optimized vector after N iterations and $\mathbf{x}^*$ the optimal solution. In practice, the step size is often determined experimentally and even step sizes that theoretically do not guarantee convergence can lead to optimal solutions in practical applications.

Stochastic gradient descent

Stochastic gradient descent [104, 105] is an optimization algorithm based on the basic principles of gradient descent with some relevant differences. It requires an objective function of the form

$$f(\mathbf{x}) = \sum_{i \in [1,S]} f_i(\mathbf{x}) \ , \tag{2.11}$$

so a sum of S terms $f_i(\mathbf{x})$. This method is often applied in the field of machine learning where each summand $f_i(\cdot)$ typically corresponds to a generic objective function evaluated over a unique data point in a data set of N entries such that $i \in$

$\{1, N\}$. Instead of computing the full gradient, which could be computationally costly for large sums, only one randomly chosen term $f_i(\mathbf{x})$ is used in the gradient computation in each iteration, resulting in algorithm 2.

Algorithm 2 Stochastic gradient descent

Input data: objective function $f(\mathbf{x}) = \sum_{i\in[1,S]} f_i(\mathbf{x})$, initial vector of optimization variables $\mathbf{x}_0$

Parameters: step size h, stopping criterion

Output: optimized vector of optimization variables $\mathbf{x}$

1: $k \leftarrow 0$
2: **while** Stopping criterion not fulfilled **do**
3: Randomly pick $i \in S$
4: $\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k - h\nabla f_i(\mathbf{x}_k)$
5: $k \leftarrow k + 1$
6: **end while**

For convergence to a global minimum to be ensured, all independent terms f_i need to fulfill the conditions of L-smoothness and convexity and the step size h_n has to fulfill the Robbins-Monro conditions [106]:

$$\sum_n h_n = \infty \ , \qquad \sum_n h_n^2 < \infty \ . \tag{2.12}$$

Stochastic coordinate descent

Stochastic coordinate descent (SCD) is also based on gradient descent and is typically applied if the number of optimization variables is very large [107]. If $f(\mathbf{x}) = f(x^1, x^2, ..., x^n)$ with $n \gg 1$, then the computation of the full gradient can be very computationally costly. At each iteration, a coordinate $i \in \{1, 2, ..., n\}$ is picked at random and the derivative $\frac{\partial f}{\partial x^i}$ is used instead of the full gradient, resulting in algorithm 3.
Convergence can be guaranteed for smooth, convex functions for step lengths $h_k < \frac{1}{L_{\max}}$ where $L_{\max}$ is the biggest componentwise Lipschitz constant over all coordinates. The convergence rates for convex and strongly-convex functions are generally comparable to those of full-gradient methods when taking the increased number of iterations for coordinate descent and reduced cost per iteration into account. Individual problems may result in advantages with SCD, especially when $L_{\max}$ is significantly smaller than the full gradient Lipschitz-constant L [108, 109].

Algorithm 3 Stochastic coordinate descent

Input data: objective function f, initial vector of optimization variables $\mathbf{x}_0$
Parameters: step size h, stopping criterion
Output: optimized vector of optimization variables $\vec{x}$

1: $k \leftarrow 0$
2: **while** Stopping criterion not fulfilled **do**
3: Randomly pick $i \in \{1, 2, ..., n\}$
4: $x^i_{k+1} \leftarrow x^i_k - h_k \frac{\partial f}{\partial x^i_k}$
5: $k \leftarrow k + 1$
6: **end while**

Accelerated gradient descent

The performance of gradient descent based algorithms can be improved by adding acceleration terms. There are several proposed accelerated gradient descent schemes, the most commonly known is momentum-based acceleration first suggested by Polyak in 1964 [110]. Adding a momentum term changes the basic gradient descent algorithm to algorithm 4. The presented algorithm follows the formulation utilized by Sutskever et al. [111].

Algorithm 4 Gradient descent with momentum

Input data: objective function f, initial vector of optimization variables $\mathbf{x}_0$
Parameters: step size h, momentum parameter β, stopping criterion
Output: optimized vector of optimization variables $\mathbf{x}$

1: $k \leftarrow 0$
2: $y_0 \leftarrow 0$
3: **while** Stopping criterion not fulfilled **do**
4: $\mathbf{y}_{k+1} \leftarrow \beta \mathbf{y}_k - h \nabla f(\mathbf{x}_k)$
5: $\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k + \mathbf{y}_{k+1}$
6: $k \leftarrow k + 1$
7: **end while**

The momentum term carries over information from the previous iterations to the current one. It is especially beneficial in low curvature areas. It reduces oscillations, which often occur for large step sizes h, making the algorithm more resilient against bad parameter choices. A performance benefit as a consequence of the added momentum can be shown independently, while the increased computational

effort per iteration is small.
The hyper-parameter β can be tuned to modulate the impact of the momentum term. Like the step size h, it may be chosen as a scalar value β or as a sequence β_k.
Nesterov accelerated gradient descent [112] extends and refines the momentum method. Instead of calculating the gradient at the current position, the momentum term is used to calculate the gradient at a predicted future location and adjust accordingly. This reduces overshoot and smooths the convergence. For a convex, L-smooth function f, the convergence is reduced to $\mathcal{O}(\frac{1}{N^2})$, which is the optimal rate that a first-order method can achieve.
The formulation used in algorithm 5 was introduced by Sutskever et al. [111].

Algorithm 5 Nesterov accelerated gradient descent

Input data: objective function f, initial vector of optimization variables $\mathbf{x}_0$
Parameters: step size h, momentum parameter β, stopping criterion
Output: optimized vector of optimization variables $\mathbf{x}$

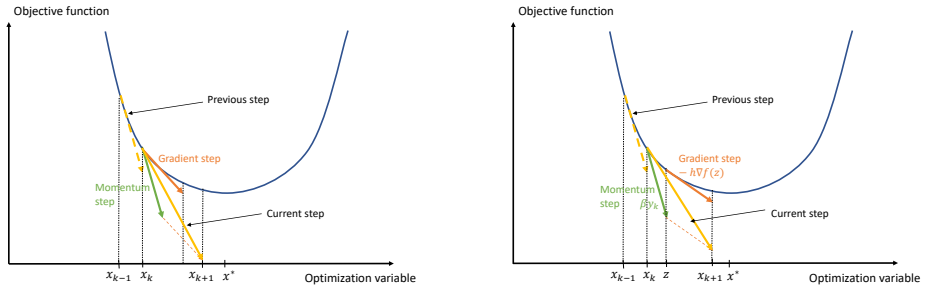
$k \leftarrow 0$
while Stopping criterion not fulfilled **do**
    $\mathbf{z} \leftarrow \mathbf{x}_k + \beta \mathbf{y}_k$
    $\mathbf{y}_{k+1} \leftarrow \beta \mathbf{y}_k - h \nabla f(\mathbf{z})$
    $\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k + \mathbf{y}_{k+1}$
    $k \leftarrow k + 1$
end while

Both momentum gradient descent and Nesterov accelerated gradient descent can also be combined with stochastic gradient descent and other gradient descent based methods [113]. They can provide considerable benefit in the case of noisy gradients [111].

2.5.2 Newton and Quasi-Newton methods

Newton's method

Newton's method is the standard example for second-order methods. Contrary to first-order methods, second-order methods also take information of the second-order objective function derivative into account. This increases the computational workload per iteration to $\mathcal{O}(n^3)$, n being the number of optimization variables, compared to $\mathcal{O}(n)$ for gradient descent [114, 115], but enables considerably faster convergence rates.



(a) Momentum (b) Nesterov accelerated gradient descent

Figure 2.15: Visual representation of momentum and Nesterov accelerated gradient descent on a one-dimensional example

Newton's method assumes that the objective function $f(\mathbf{x})$ is twice differentiable and that its Hessian is invertible. For a smooth, convex function, convergence can be guaranteed in the vicinity of the global minimum x^* with a quadratic convergence rate [116].

Algorithm 6 Newton's method

Input data: objective function f, initial vector of optimization variables $\mathbf{x}_0$
Parameters: stopping criterion
Output: optimized vector of optimization variables $\mathbf{x}$

1: $k \leftarrow 0$
2: **while** Stopping criterion not fulfilled **do**
3: $\quad \mathbf{x}_{k+1} \leftarrow \mathbf{x}_k - \nabla^2 f(\mathbf{x}_k)^{-1} \nabla f(\mathbf{x}_k)$
4: $\quad k \leftarrow k + 1$
5: **end while**

It does not generally guarantee global convergence. Newton's method does not require a step size parameter h, but a step size parameter may be introduced, leading to the damped Newton method, to reduce the risk of divergence. [117]

BFGS

The Broyden-Fletcher-Goldfarb-Shanno, or BFGS, algorithm is named after the four people who independently submitted it in a short period of time from one

another [118, 119, 120, 121]. It belongs to a group of methods known as quasi-Newton methods and it approximates the Hessian based on first-order information to reduce the computational burden per iteration. The BFGS algorithm is presented in algorithm 7, where H_k is the approximation of the inverse of the Hessian at the k^{th} iteration. The presented algorithm follows the scipy implementation [122, 123] as derived from the lecture notes of "Unconstrained Optimization" by P.E. Frandsen, K. Jonasson, H.B. Nielsen, O. Tingleff at the Technical University of Denmark in 1999 [124].

Algorithm 7 BFGS algorithm

Input data: objective function f, initial vector of optimization variables $\mathbf{x}_0$, initial approximation of inverse Hessian H_0

Parameters: stopping criterion

Output: optimized set of optimization variables $\mathbf{x}$

$k \leftarrow 0$
while Stopping criterion not fulfilled **do**
  $\mathbf{p}_k \leftarrow -H_k \nabla f(\mathbf{x}_k)$
  Perform line search to find suitable step size α_k
  $\mathbf{s}_k \leftarrow \alpha_k \mathbf{p}_k$
  $\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k + \mathbf{s}_k$
  $\mathbf{y}_k \leftarrow \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)$
  $b \leftarrow \frac{1}{\mathbf{s}_k^\top \mathbf{y}_k}$
  $a \leftarrow b(1 + b\mathbf{y}_k^\top H_k \mathbf{y}_k)$
  $H_{k+1} \leftarrow H_k + a\mathbf{s}_k\mathbf{s}_k^\top - b(H_k\mathbf{y}_k\mathbf{s}_k^\top - \mathbf{s}_k\mathbf{y}_k^\top H_k)$
  $k \leftarrow k + 1$
end while

The algorithm is guaranteed to converge towards a global minimum with a superlinear convergence rate [125], if the objective function is convex and twice differentiable and the step size α_k satisfies the Wolfe conditions [126]

$$\begin{aligned} f(\mathbf{x}_k + \mathbf{s}_k) - f(\mathbf{x}_k) &\leq \delta_1 \alpha_k \mathbf{p}_k^\top \nabla f(\mathbf{x}_k) \\ \mathbf{p}_k^\top \nabla f(\mathbf{x}_k + \alpha_k \mathbf{p}_k) &\geq \delta_2 \mathbf{p}_k^\top \nabla f(\mathbf{x}) \ , \end{aligned} \tag{2.13}$$

where $\delta_1 \leq \delta_2$ are constants in $(0, 1)$ [127].

L-BFGS

The Limited-memory Broyden-Fletcher-Goldfarb-Shanno, or L-BFGS, method [128, 129], is another Quasi-Newton method and approximates the BFGS method while

reducing the memory usage during optimization. Like the BFGS method, it approximates the inverse Hessian of the objective function, but instead of storing the full matrix, only a few vectors are stored to represent the approximation. Like other quasi-Newton methods its convergence rate is superlinear [130].

2.5.3 Optimization in radiotherapy treatment planning

There are several optimization algorithms available and suitable for treatment planning. Which one is most suitable for a specific task might depend on several factors, such as the characteristics of the used objective function and objective types, the hardware and software availabilities and the size and complexity of the optimization problem.
A commonly chosen objective function is a weighted sum of several sub-objectives of the form

$$f(\mathbf{x}) = \sum_{k \in K} f_{\text{obj},k}(\mathbf{x}) \ , \tag{2.14}$$

where K is the set of optimization objectives and $f_{\text{obj},k}$ is dependent on the objective type of objective k. Typical examples for objectives include maximum dose objectives,

$$f_{\text{max},k} = \frac{1}{N_k} \sum_{i \in I_k} \sum_{j \in J} \min(0, D_{i,j} x_j - d_{\text{lim},i,k})^2 \ , \tag{2.15}$$

and minimum dose objectives,

$$f_{\text{min},k} = \frac{1}{N_k} \sum_{i \in I_k} \sum_{j \in J} \min(0, d_{\text{lim},i,k} - D_{i,j} x_j)^2 \ , \tag{2.16}$$

where N_k is the number of voxels affected by objective k, I_k is the set of voxel affected by objective k, J the set of spots with spot weights $\mathbf{x}$, D the beamlet matrix and $d_{\text{lim},i,k}$ the upper or lower dose limit imposed by objective k on voxel i [131]. Minimum and maximum dose objectives are zero if all voxel within a ROI comply with the prescribed dose limits d_{lim}. If the voxel dose is below a minimum dose threshold or above a maximum dose threshold, the objective terms add a positive contribution to the total objective function value. This contribution is bigger the further below or above the threshold the voxel dose is, and is further enhanced by the square in equations (2.15) and (2.16). These types of objectives are also referred to as least squares optimization [132] and are a staple in radiotherapy treatment planning, integrated in most clinical and experimental TPS [97, 28].
An objective function of this type is smooth, convex and differentiable and consequently allows users a wide range of potential optimization methods.

Other types of objectives include mean dose objectives

$$f_{\text{mean},k} = \min(0, \hat{d}_k - d_{\text{lim},k})^2 \ , \tag{2.17}$$

with $\hat{d}_k = \frac{1}{N} \sum_{i \in I} \sum_{j \in J} D_{i,j} x_j$ is the mean dose in the ROI affected by objective k or DVH objectives, such as

$$f_{D_{95},k} = \min(0, d_{\text{lim},k} - D_{95,k})^2 \ , \tag{2.18}$$

regulating the dose that 95% of the ROI in question should receive. The $D_{95,k}$ and other DVH parameters are non-differentiable functions of the spot weights x and need to be computed at every iteration of the optimizer.
The corresponding optimization problem can be formulated as

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & x_j \geq 0 \quad \forall j \in J \ , \end{aligned} \tag{2.19}$$

with $\mathbf{x}$ representing the spot weights and J the set of spots. The spot weights cannot be negative and are therefore constrained to be so in the problem.
Both gradient based optimization methods and BFGS based methods are common in proton therapy planning. The clinical TPS Raystation employs BFGS-based optimization [133], while research TPS OpenTPS offers users a choice, including BFGS, L-BFGS and gradient descent [28].

2.5.4 Robust optimization

Robust optimization is an optimization concept that can be used during treatment planning and which describes the inclusion of potential error scenarios in the treatment planning and optimization process. While robust optimization is not necessary for treatment planning in general, it is highly recommended and often considered a requirement for proton therapy. In proton therapy, slight differences between the planning CT and the setup during actual delivery can lead to significant dose deformations and dramatic consequences for patients [26, 134, 79, 135]. These dosimetric distortions can often not be accounted for with a conventional PTV margin based approach as is common in radiotherapy with photons [136, 137, 138, 139].
Different error scenarios are defined to be taken into account. Which error sources are relevant may fluctuate between cases and centers, but they typically include beam alignment shifts and range errors [26, 25, 138, 140, 137, 141, 142, 27]. Beam alignment shifts can be caused during the positioning of the patient on the treatment couch or due to inaccuracies in the delivery system. They are modeled as a rigid shift of the beam isocenter compared to the patient. Range errors usually stem from errors during CT acquisition, the conversion of the HU values, or

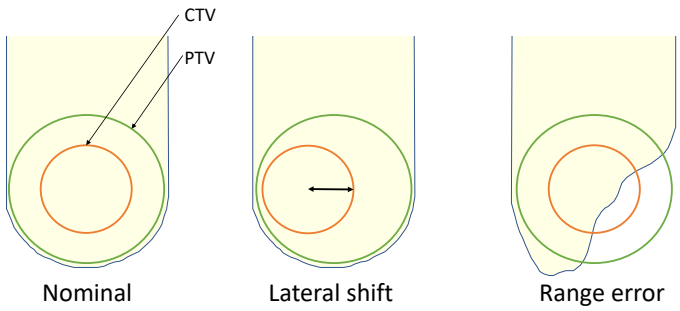


Figure 2.16: Representation of the effects of beam alignment shifts and range errors on proton therapy dose distributions. While the PTV concept as it is usually applied in conventional radiotherapy might be sufficient to account for lateral shifts, range errors may lead to significant distortions of the dose distribution, that require robust optimization.

biological uncertainties [134, 143, 144]. There are multiple options how to model them, including uniformly scaling the mass density [26] or HU values [27], modifying the stopping power ratios by a scaling factor [142], or shifting the proton range by a specified distance and modifying beam energies accordingly [138]. Other error sources, such as anatomical inter- and intrafractional changes, are often excluded due to the technical difficulty of assessing and modeling them. An exception is breathing motion, which is typically integrated in robust optimization, then known as 4D-robust optimization [145, 146], or accounted for in ITV margins [147] .
Range errors in particular are not present in photon therapy and can lead to severe dose distortions as illustrated in Figure 2.16. The potential consequences on real patients are demonstrated in Figure 2.17, where a robustly optimized plan is compared against IMPT on a paraspinal case. As can be seen, non-robust optimizations leads to underdosage in the CTV in the perturbed scenario.
There are different ways to implement robust optimization. The most common is the so called minimax optimization, which can be further distinguished in worst-case scenario [25], worst-case objective [142] and worst-case dose distribution optimization [148, 149, 23]. Worst-case scenario optimization, also known as compound worst-case, requires each scenario objective function to be evaluated at each iteration. The corresponding optimization problem can be formulated as

$$\min_{\mathbf{x}} \max_{s} f_s(\mathbf{x}) \quad , \tag{2.20}$$

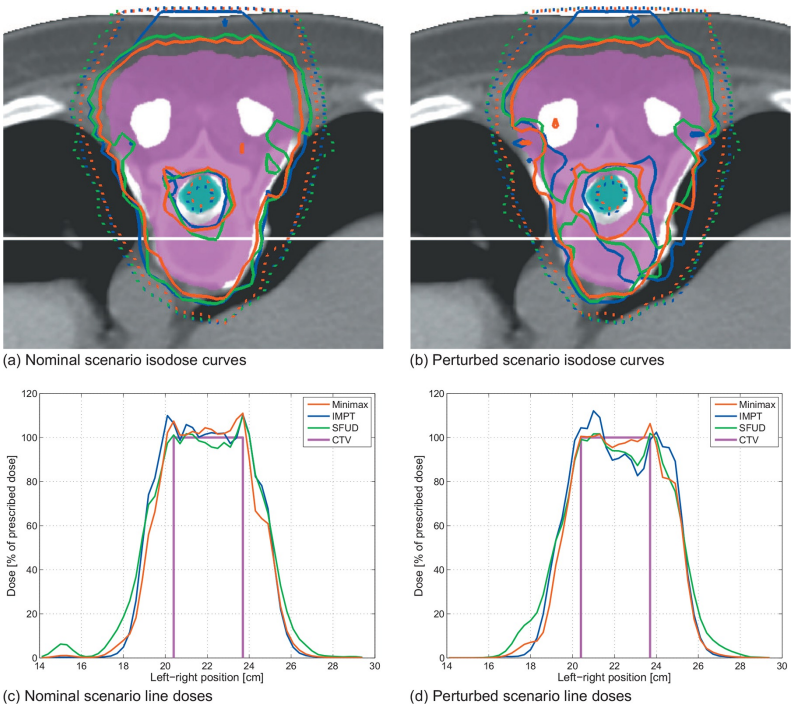


Figure 2.17: Comparison of the plan quality achieved with robust optimization (minimax, red), IMPT (blue) and single field unitary dose (SFUD, green) on a paraspinal case. The 95% and 55% isodose lines of all methods are presented for the nominal case (a) and in a perturbed scenario with a density scaling factor of 3% and a beam alignment shift of 2.5 mm applied (b). The dose is represented on a single slice against the prescription dose (CTV, purple) for the nominal (c) and perturbed case (d). (Figure taken from: [25].)

with $\mathbf{x}$ as the optimization variables, $s \in S$ a scenario from the pool of potential error scenarios S and f_s the scenario objective function. This is the most established form of robust optimization and current clinical standard.
Objective worst-case evaluates each objective separately and potentially optimizes an objective function consisting of different scenario contributions for each objective. The corresponding optimization problem is

$$\min_{\mathbf{x}} \sum_{k \in K} w_k \max_{s} f_{k,s}(\mathbf{x}) \ , \tag{2.21}$$

with K as the set of objectives and $f_{k,s}$ the objective function term for objective k and scenario s. It has been investigated specifically to aid in multicriteria optimization and could be shown to provide certain benefits [142]. However, it is still a comparatively niche method.
Worst-case dose distribution determines the worst-case scenario for each voxel separately. Because of this, it optimizes an unrealistically pessimistic dose distribution and is often considered too conservative.
Another common type of robust optimization is stochastic programming, also known as expected-value optimization [26, 27]. It should be noted that stochastic programming does not conform with the mathematical concept of robust optimization, but is commonly referred to as a robust optimization method in a radiotherapy context due to its ability to generate robust treatment plans. The expected-value objective function over all scenarios is minimized, as

$$\min_{\mathbf{x}} \sum_{s \in S} p_s f_s(\mathbf{x}) \ , \tag{2.22}$$

where p_s is a scenario weight representing the scenario probability. In the case of N_s uniformly sampled scenarios, $p_s = \frac{1}{N_s}$. A potential weakness of expected-value optimization compared to worst-case scenario robust optimization is that there might be outlier scenarios that are compensated by other scenarios during optimization. Since the expected-value over all scenarios is optimized, it is possible that some scenarios do not reach a desirable dose distribution, but do not continue to improve because the improvement of other, potentially already satisfactory, scenarios has a bigger impact on the global objective function. Eventually, one expects that the other scenarios reach a near optimum and the optimization focuses on the outliers to continue to improve, but that could potentially be quite inefficient.
The computational challenges during proton therapy treatment planning are exaggerated during robust optimization, where beamlet and dose calculations have to be repeated for each scenario and optimizers have to handle very large matrix operations. Often voxel sizes are increased or the number of spots decreased to

reduce Monte Carlo computation time and reduce matrix sizes.
Due to their high demand of computational resources and computation time, strategies are investigated to improve robust proton therapy planning workflows. Buti et al. [16] reduced both computational demands without sacrificing robustness against certain error types, by introducing what they refer to as dynamic minimax optimization. Within a worst-case scenario setting, only a subset of all scenarios, the "active pool", are considered during optimization, while other scenarios are moved into the "dead pool". The algorithm dynamically allocates scenarios to the "active pool" based on their probability of being the worst-case scenario. They reported time gain of up to 85% based on the respective size of the "active pool".
Xu et al. [17] developed a fast robust optimizer based on GPU-parallelization and a conjugate gradient approach [150], and compared it to the clinical treatment planning system Eclipse. Depending on the treatment site, their proposed optimizer was up to 25 times faster than Eclipse, but also had significantly higher memory needs. It should be noted that the two methods had different objective function definitions, which might have impacted resulting plan quality and optimization performance.
A different parallelized optimizer has been proposed by Fu et al. [18]. They used a CPU-parallelized optimization framework based on the alternating direction method of multipliers with a Barzilai–Borwein step size (ADMM-BB) [151] and compared it against a projected gradient descent (PGD) based optimizer. They reported a faster convergence and reduced computation time by a factor $6 - 7$ with their proposed framework.
A similar approach was suggested by Fan et al. [19] who likewise proposed a robust optimizer based on a parallel implementation and the ADMM method and reported a reduction of computation time by about 40%.
It should be noted that for all of these methods, the exact benefit depends on the investigated treatment site, the number of error scenarios taken into consideration and the utilized hardware. For all proposed fast robust optimizers, a reduction of computation time compared to conventional approaches was achieved, but other computational metrics like the memory needs were either unaffected or even negatively impacted. While memory restrictions are not a principle bottleneck in current proton therapy treatment planning, they might seriously impact the development of novel techniques with potentially higher numbers of spots. They might also limit the use of GPU-parallelization techniques, since GPUs usually have limited memory capacity.

2.5.5 Biological optimization

During proton therapy planning, it is often assumed that the biological effectiveness of protons is constant with a relative biological effectiveness (RBE) of 1.1 compared to photons. However, experiments indicate that the biological effectiveness of protons is not constant, but increases towards the end of their range [152, 153]. This increase in RBE is closely linked to the linear energy transfer (LET). RBE models often use a combination of dose, LET and tissue-specific parameters, usually based on the α/β ratio, to calculate variable RBE values.
Various RBE models have been proposed [154, 155, 156, 157, 158, 159, 160], however, they differ significantly in their predictions and add considerable uncertainty to the treatment planning process [161].
Due to the strong correlation between RBE and LET [162], LET-based optimization has been proposed as an alternative [163, 164, 165, 166, 167] to optimization with variable RBE. This includes the option to optimize the biological surrogate dose (BD) defined as

$$D_{\text{phys}} + BD = D_{\text{phys}} \times (1 + c \times LET) \ , \tag{2.23}$$

with D_{phys} as the physical dose and c as a scaling parameter [168]. Similar dose objectives can then be used to either directly optimize based on the BD, or to define objectives on the physical dose $f_k(\mathbf{d})$ and LET impacted "dirty dose" $f_k(\mathbf{LET} \times \mathbf{d})$ [165, 166].
Other methods incorporate LET objectives directly into the objective function [167, 169].

Chapter 3

Beamlet-free, Monte-Carlo-based treatment planning for proton therapy

The content presented in this chapter is adapted from the published article: Technical note: Beamlet-free optimization for Monte-Carlo-based treatment planning in proton therapy. Med Phys. 2024; 51: 485–493.

Authors: Danah Pross[1], Sophie Wuyckens[1], Sylvain Deffet[1,2], Edmond Sterpin[1,3,4], John A. Lee[1], Kevin Souris[1,2]

[1] *UCLouvain, Institut de recherche expérimentale et clinique, Molecular Imaging and Radiation Oncology Laboratory, Brussels, Belgium.*
[2] *Ion Beam Applications SA, Louvain-La-Neuve, Belgium.*
[3] *KULeuven, Department of Oncology, Leuven, Belgium.*
[4] *Particle, Leuven, Belgium.*

3.1 Introduction

In proton therapy, the state of the art treats patients with intensity modulated proton therapy [170] (IMPT) using pencil beam scanning. Treatment planning typically involves the precalculation of beamlet dose distributions for each pencil beam and subsequent optimization of their respective intensity. The total dose to voxel i is then given as the linear combination of these elementary dose distributions

$$d_i = \sum_j D_{ij} x_j \ , \tag{3.1}$$

with D_{ij} as the individual beamlet dose contribution of spot j to voxel i and x_j its corresponding intensity ('spot weight'). The dose influence matrix D is the concatenation of the beamlet dose distributions and usually reaches considerable size, due to the large number of spots in IMPT. The number of spots typically lies in the thousands for most clinical plans, while the number of voxel can easily lie in the millions.

While dose calculations can be performed using fast analytical dose calculation algorithms [70], their accuracy falls short of the more elaborate Monte Carlo (MC) methods [171, 74]. The drawback of MC dose engines is their high demand of computational resources that, especially for large numbers of spots, can prolong treatment planning time considerably. It is therefore an ongoing effort to reduce computation time for MC methods and several fast MC dose engines have been introduced, often based on CPU (Central Processing Unit) or GPU (Graphical Processing Unit) parallelization [102, 172, 7, 9, 12, 10]. Ma et al. even published a fast MC-based treatment planning system utilizing GPU-parallelization [173].

The storage of the beamlet dose distributions in the dose influence matrix, necessary for optimization, also puts considerable strain on the computer memory capacities [174].

Beside the time expended on the repeated calls to the dose calculation engine, the optimization algorithm also places high demands on the computational resources and prolongs computation time, due to the very costly multiplications with the large dose influence matrix involved in the gradient computation.

These high computational demands, both with respect to computation time and memory usage, are hindrances in clinical routines that require to be addressed to enable efficient workflows when a MC dose engine is involved. These problems become exacerbated in study cases that require a large number of beamlets, such as robust optimization, adaptive proton therapy or arc proton therapy. These increasingly elaborate methods seek to improve treatment quality by reducing dose to organs at risk (OAR) and ensuring target coverage, but they are limited to the currently available computational resources.

The objective of this work is to present a new method of treatment plan optimiza-

tion that merges dose calculation and optimization and substantially reduces their dependence on the number of spots in the plan. The proposed method aims to maintain plan quality while substantially reducing computation time and memory for many-spot plans.
The approach combines MC dose calculation and spot weight optimization into a single efficient algorithm. This algorithm is referred to as beamlet-free optimization since it does not require the computation of individual beamlet dose distributions. The method is applied here to proton therapy with pencil beam scanning, where the number of beamlets is typically higher than in photon therapy and where the superior accuracy of MC methods over analytical dose calculation algorithms is especially beneficial due to the steep dose gradients involved.

3.2 Methods

3.2.1 Beamlet-free algorithm

The idea of beamlet-free optimization is based on continually and incrementally evaluating the objective function while the dose is being simulated. Instead of computing each individual beamlet dose distribution, the dose is directly accumulated into a total distribution. Yang et al. [20] developed a similar method based on concurrent Monte Carlo simulation and fluence optimization for photon radiotherapy. Their method consists of dividing beamlets in sub-beamlets and the consecutive simulation of a small number of particles, between 10 and 1000, in each iteration for each sub-beamlet. The exact number of particles varies between sub-beamlets and depends on an estimation of the full objective function gradient that is evaluated at each iteration.
Li et al. [22] also developed a method based on concurrent MC dose calculation and optimization applied to proton therapy. They referred to their method as adaptive particle sampling (APS), because it samples the number of protons to be simulated to each spot based on its current spot weight determined during optimization.
Both approaches reduce the total number of simulated particles by reducing the number of protons simulated to, and therefore increasing the statistical uncertainty on, spots with low spot weights. However, they maintained a dose influence matrix containing the individual beamlet dose distributions that still imposes a substantial dependence on the number of spots with regard to memory requirements and computation time, even though the later could be reduced by about a factor 5 compared to the classic dose influence matrix precomputation and subsequent optimization scheme [22]. Our method, in contrast, combines this idea with an approach based on stochastic coordinate descent (SCD) [107].

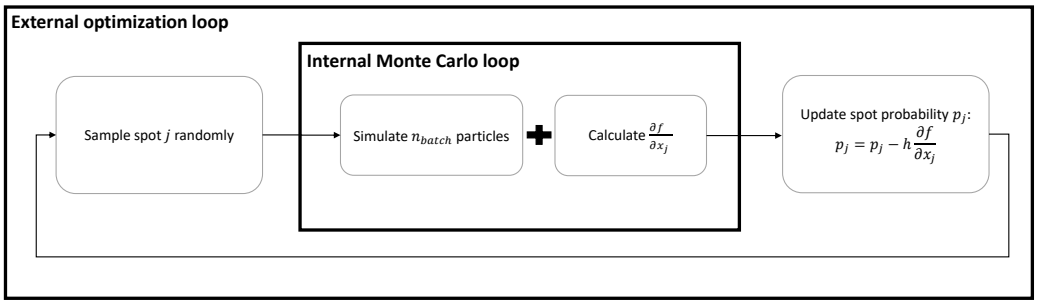


Figure 3.1: Visual representation of algorithm 8.

For the implementation of this method, spot weights in their original sense as modulation factors cannot exist anymore. The equivalent of the original spot weights with respect to the final dose output would be the number of protons simulated towards each spot. To optimize this quantity during the ongoing dose simulation, a set of non-normalized probabilities p is introduced, that are used to generate the spot probability distribution P.
In each iteration of the proposed algorithm, a spot is randomly sampled from the probability distribution P. A batch of protons is simulated towards the chosen spot and the impact on the objective function is calculated to estimate the gradient.
After all protons of the batch have been simulated, the estimated gradient, with respect to the sampled spot, is used to update the spot probability using SCD.
This yields Algorithm 8, which is visually represented in Figure 3.1, where p_j is the probability for a spot j, P is the corresponding probability distribution, h is a step size, c is the batch cost calculated as

$$c = \sum_k F(d_{i_k,k}) - F(d_{i_k,k-1}) \ , \tag{3.2}$$

with k an index over all Monte Carlo steps. A Monte Carlo step refers to an incremental movement of a particle through the voxelized geometry under the consideration of relevant particle interactions, usually resulting in energy loss,

change of direction and potentially the creation of secondary particles or absorption of the particle. The index i_k indicates the voxel actively receiving dose in Monte Carlo step k, $F(\cdot)$ represents the objective function, and $d_{i_k,k-1}$ and $d_{i_k,k}$ the normalized doses to voxel i_k before and after the kth Monte Carlo step of a proton in the batch. The particle cost calculated in equation (3.2) corresponds to the finite difference approximation of the partial derivatives of the objective function. It sums over the MC steps instead of the voxel for reasons of computational efficiency. The MC contributions are computationally very cheap, while the voxel contributions would require iterating over the entire voxel grid again at every iteration of the optimization loop. All relevant voxels that contribute significantly to the objective function derivative are still covered. The distribution P is defined as the probability distribution associated with the non-normalized probability $\boldsymbol{p}$ and is initialized as a uniform distribution at the beginning of algorithm 8.
The normalization of the dose during the cost computation is necessary to evaluate the relative impact of simulating particles to a given spot j. After each Monte Carlo step, the total, unnormalized dose is multiplied with a normalization coefficient $n(\boldsymbol{d}) = \frac{b}{\hat{d}}$ where $\hat{d}$ is the mean target dose and b is the prescribed target dose.

Algorithm 8 Beamlet-free algorithm

Input data: CT image data, plan data without spot weights, objective data
Parameters: step size h, batch size n_{batch}, total number of primaries n_{total}
Output: 3D dose image data, optimized plan data with spot weights $\mathbf{x}$

Initialize spot probability vector $\mathbf{p}$ to unit values
Generate probability distribution $P(\mathbf{p})$
while Less than n_{total} protons have been simulated **do**
    Sample spot j from P
    Initiate a batch of n_{batch} particles for spot j
    Initialize cost $c \leftarrow 0$
    Initialize normalization coefficient $n \leftarrow 1$
    while Particles in batch not fully simulated **do**
        Perform Monte Carlo step k and calculate deposited dose Δd_k
        Normalize previous dose: $d_{i_k}^{n,k-1} \leftarrow d_{i_k} \cdot n$
        Score deposited dose: $d_{i_k} = d_{i_k} + \Delta d_k$
        if Current voxel i_k in target region **then**
            Update mean target dose: $\hat{d} \leftarrow \hat{d} + d_k$
        end if
        if Mean target dose $\hat{d} \neq 0$ **then**
            Update normalization coefficient: $n \leftarrow \frac{b}{\hat{d}}$
        end if
        Normalize updated dose: $d_{i_k}^{n,k} \leftarrow d_{i_k} \cdot n$
        Update cost $c \leftarrow c + F(d_{i_k}^{n,k}) - F(d_{i_k}^{n,k-1})$
    end while
    Update spot weight: $x_j \leftarrow x_j + n_{\text{batch}}$
    Update spot probability: $p_j \leftarrow p_j - \frac{h}{n} c$
    Normalize spot probability vector s.t. $\sum p = 1$
end while

The quantity to be optimized in this procedure is the spot probability $\boldsymbol{p}$, which has a stochastic link to the number of protons simulated toward the corresponding spots, the spot weight $\boldsymbol{x}$. The sampling of a spot j is equivalent to the occurrence of an event with probability p_j, leads to an increase of the respective event counter x_j. Per the law of large numbers [175], the distribution of sampled events $\mathbf{x}$ is expected to approach the underlying probability distribution $\mathbf{p}$. The objective function is evaluated over the total dose distribution $\boldsymbol{d}$, which depends on $\boldsymbol{x}$, and only indirectly on $\boldsymbol{p}$. The finite difference approximation in equation (3.2) concerns the gradient of the objective function with respect to $\boldsymbol{x}$.

Therefore, two separate stochastic processes are taking place, the MC simulation generating the dose distribution and the spot sampling at the beginning of each iteration which causes the distribution of protons per spot, $\boldsymbol{x}$, and the probability distribution P to converge towards the same final distribution.
The beamlet-free algorithm also eliminates the need for a separate, final dose calculation, since the outputted dose distribution already reflects a MC simulation with low statistical uncertainty. While the precise statistical uncertainty depends on the simulated number of protons, the number of iterations required for the optimizer to converge is usually high enough to ensure a much larger number of protons is simulated than what is needed to achieve sufficiently low uncertainty in the MC simulation. After the joint optimization and dose calculation are completed, the optimized spot weights $\boldsymbol{x}$ are outputted and can be used for the treatment delivery.
The beamlet-free algorithm was implemented in C around the existing, fast Monte Carlo code MCsquare [12].

3.2.2 Phantom data

The beamlet-free method is expected to improve both computation time and memory usage when large numbers of beamlets are involved. To highlight its benefits and limitations, a complexity analysis was performed. Simulations are performed in a water phantom containing a target region with varying size. The phantom is represented in Figure 3.2. The multiple target sizes result in geometries with varying numbers of spots (650, 1482, 2794, 4769, 7307). Spot and layer spacing are set to 4 mm and a target margin of 6 mm is applied. The influence of the number of spots on computation time and memory usage are investigated. A single beam with couch and gantry angles of 0° is used for the analysis. The water phantom has dimensions of $150 \times 150 \times 150$ mm^3 with a voxel size of $1 \times 1 \times 1$ mm^3 and the target region is cubic, with the center located at a depth of 110 mm in beam direction and side lengths of 20, 30, 40, 50 and 60 mm.
A double Gaussian beam model is used, with spot sizes ranging from 1.80 to 3.99 mm for the inner Gaussian and from 2.84 to 5.59 mm for the outer Gaussian.
Simulations are repeated 5 times to estimate the impact of statistical fluctuations.

3.2.3 Patient data

To demonstrate the high potential of the beamlet-free algorithm, we chose to benchmark it against a conventional, beamlet-based gradient descent optimization algorithm.

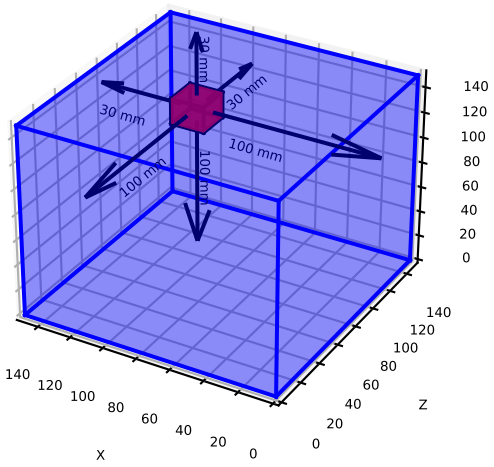


Figure 3.2: Water phantom used for the scaling analysis. An exemplar target region with side length of 20 mm is represented in red.

The treatment plan optimization is performed using the in-house developed, research treatment planning system OpenTPS [28] which can incorporate both the beamlet-free method as well as established optimization algorithms, like the BFGS algorithm [118, 121, 120, 119].
A brain case (base-of-skull tumor) and a prostate case were chosen for the comparison. The local ethics committee approved the use of the patient data.
The same beam model is used as for the water phantom.

Brain case

A treatment plan was generated using a three-beam combination at a fixed gantry angle G $= 90^{\circ}$ and rotating the couch angle C at 0°, 270° and 315°. The spot and layer spacing were set to 4 mm and a margin of 6 mm for the spot spacing was applied around the target. A total of 9812 spots compose the plan. The ROIs upon which optimization objectives are placed are the planning target volume (PTV), optical chiasm (OC), brain stem (BS), right optical nerve (RON) and left optical nerve (LON). The optimization objectives are listed in Table 3.1. The resolution of the dose grid was set to the resolution of the CT data of 0.468 mm $\times\, 0.468$ mm $\times\, 1.000$ mm and covers a total of $512 \times 512 \times 229$ voxels.

Table 3.1: Optimization objectives brain case

ROI	Objective Type	Dose limit	Weight
PTV	D_{min}	54.0	5
PTV	D_{max}	54.0	5
OC	D_{max}	60.0	1
BS	D_{max}	55.0	1
RON	D_{max}	55.0	1
LON	D_{max}	55.0	1

Prostate case

Two opposing beams with a fixed gantry angle of 0° and couch angles of 90° and 270°. Spot and layer spacing were set to 6 mm and a margin of 6 mm was applied around the target. The plan contains of 2732 spots and the regions of interest consist of the the CTV, rectum, bladder, right and left femoral head (RFH/LFH), and anal canal. The optimization objectives are listed in Table 3.2. The dose grid was set to CT resolution of 1.074 mm $\times 1.074$ mm $\times 2.000$ mm and covers a total of $512 \times 512 \times 209$ voxel.

Table 3.2: Optimization objectives prostate case

ROI	Objective Type	Dose limit	Weight
CTV	D_{min}	60.0	20
CTV	D_{max}	60.0	20
Rectum	D_{max}	60.0	5
Bladder	D_{max}	60.0	5
Anal canal	D_{max}	60.0	5
RFH	D_{max}	55.0	1
LFH	D_{max}	55.0	1

3.2.4 Optimization parameters

The total number of particles that are needed in the Monte Carlo simulation depends on the number of spots. For the water phantom, a number of 40 million primary protons was chosen, which was shown to result in good plan quality, evaluated on the basis of comparable target homogeneity to the conventional optimizer, for all cases. For the patient cases, which were more complex, the suitable

number of particles was determined to ensure low statistical uncertainty and sufficient iterations. For the brain case, which was much larger, a higher number of 200 million particles was chosen. For the prostate case 50 million particles were deemed sufficient. All other parameters were chosen identically for the different simulations.
The beamlet-free algorithm simulates the primary protons in batches of $n_{\text{batch}} = 16$.
The batch size was chosen as a compromise between optimization quality per iteration, which is improved for larger batch sizes, and overall number of iterations and computation time, which increases for comparable numbers of total primaries. The value $n_{\text{batch}} = 16$ is also the most efficient value hardware-wise for the size of particle vectors in MCsquare. The AVX512 instruction set available on modern CPU can indeed support 512 bit data vectors and thus up to 16 float32 values.
The step size h was chosen experimentally to lead to fast and reliable convergence in different scenarios. It is modified by a normalization coefficient n that artificially inflates it as the dose accumulation progresses. This method was experimentally determined to deliver better optimization results than could be obtained with a constant step size without modification through the normalization coefficient. This unconventional approach was chosen to account for the fast decrease of the gradient that can be expected after a few iterations, as the dose starts to increase. Theoretically, a decreasing step size can be shown to be optimal for Stochastic Gradient Descent (SGD) [104] and SCD [107], however, in practice other step size strategies often lead to superior results. The step size is set to $h = \frac{10^{10}}{n}$ with the normalization coefficient defined as $n = \frac{\hat{\boldsymbol{d}}}{b}$, where $\hat{\boldsymbol{d}}$ is the mean target dose at that point and b is the prescribed target dose.
For the comparison, a beamlet matrix with $5 \cdot 10^4$ protons per beamlet is simulated and subsequently optimized using the scipy [122] FORTRAN implementation (wrapped in Python) of the second-order optimization algorithm L-BFGS [129] (Limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm). The optimizer continues until it meets the tolerance criterion which is defined as the absolute difference in objective function values between iterations and set to 10^{-4}. A final dose computation is performed to match the statistical accuracy of the beamlet-free method. Both the beamlet calculation and the final dose computation are performed using the MCsquare dose engine that is also utilized in the beamlet-free algorithm. The number of primaries for the MC dose calculation in the beamlet-free algorithm was intentionally picked so that the optimized plan quality would reach the one obtained with the beamlet-based method and consequently ease the comparison.
Simulations were performed on a computation server with a total of 80 CPU cores and 526 GB RAM.

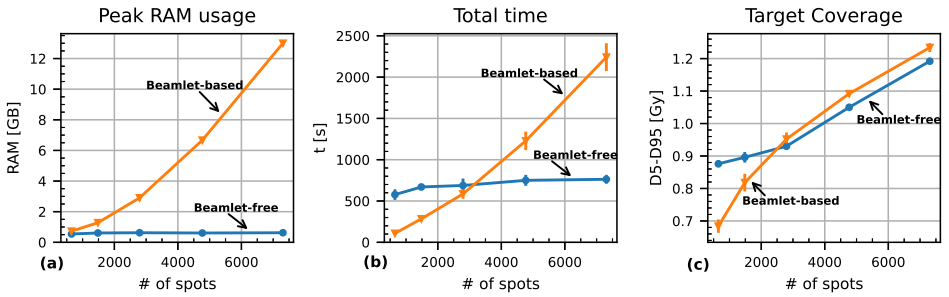


Figure 3.3: Memory usage (a) and computation time (b) required and target homogeneity (c), $D_5 - D_{95}$, achieved by the beamlet-based and beamlet-free methods versus the number of spots for a water phantom with different target sizes.

3.3 Results

3.3.1 Scaling analysis on water phantom

Computation time and peak memory usage are plotted against the number of spots for both methods and compared in Figure 3.3. The values are calculated as the mean over 5 runs and the error bar represents the standard deviation. The difference in target homogeneity, $D_5 - D_{95}$, was less than 0.20 Gy in all cases.

3.3.2 Patient data

Brain case

Dose volume histograms (DVH) obtained by both methods are depicted in Figure 3.4. In both cases all OAR objectives could be met and the target D95/D5 is within 5% of the prescription.
For the OAR, D_{max} objectives are considered as met, if the D_5 value of the ROI is below the dose limit.
The target objectives are considered as met if the D_{95}/D_5 value is within $\pm 5\%$ of the prescription dose.
Both methods ran 5 times. The fluctuations on the target homogeneity, $D_5 - D_{95}$, were less than 4% for both methods.
The objective function convergence over the course of the optimization is shown in Figure 3.5. The objective function values are plotted against the number of

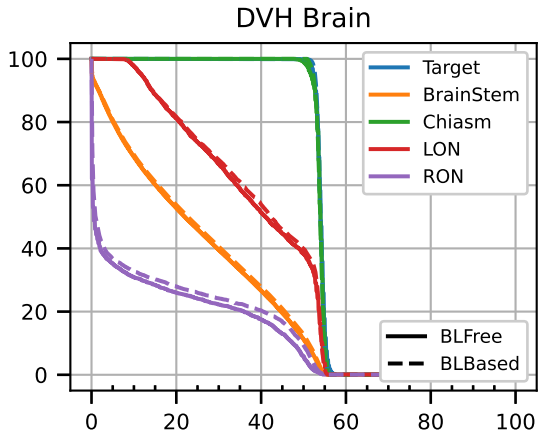


Figure 3.4: Computed dose volume histogram for beamlet-based (dashed) and beamlet-free (solid) for the brain case.

iteration of the respective optimization algorithm. The relative behaviour of the objective function can be compared, and a similar behaviour of rapidly decreasing objective function values can be observed. The decrease exhibited by the beamlet-free method is slightly larger due to decreasing levels of MC noise as the simulation progresses. The decreasing MC noise on the dose distribution causes the objective function to keep decreasing even after the spot weights have mostly converged.

The memory usage of both methods is compared in Table 3.3. The peak memory required during the dose calculation and optimization, could be reduced from 29 GB for the beamlet-based method to about 1.561 GB for the beamlet-free method. It is worth noticing that the peak memory for the beamlet-based method was recorded during the import of the individual beamlet dose distributions, compressed as sparse matrices, into the dose influence matrix (stacked as columns). Other steps in the treatment planning process, however, required higher peak memory usage than the beamlet-free method needed during dose calculation and optimization. This peak memory was recorded as about 3.569 GB. It depends on the used treatment planning system and might be improved independently from the presented algorithm. The required storage disk space of the files needed for the dose calculation and optimization is 8.8 GB for the beamlet-based and 4.6 GB for the beamlet-free case. The difference between the two methods with respect to utilized disk space can be explained through the storage of the large dose influence matrix that requires 4.4 GB alone. Temporary files during optimization required an additional 37 GB for the beamlet-based case.

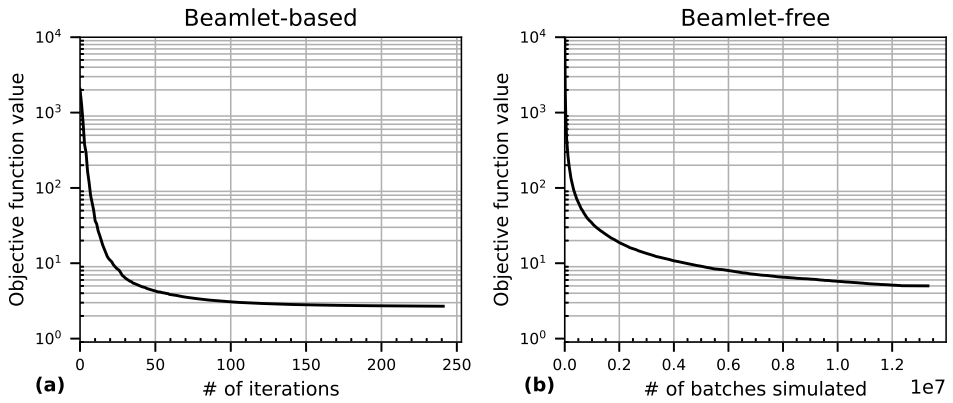


Figure 3.5: (a) Convergence plot of the beamlet-based algorithm (scipy L-BFGS) and (b) convergence plot of the beamlet-free algorithm.

Table 3.3: Comparison of memory usage on the brain case for the beamlet-based and beamlet-free method.

	Beamlet-based	Beamlet-free
Peak RAM usage [GB]	29	1.561
Storage [GB]	8.8	4.6

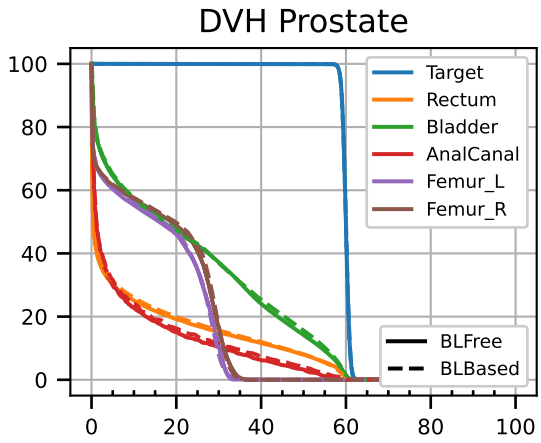


Figure 3.6: Computed dose volume histogram for beamlet-based (dashed) and beamlet-free (solid) for the prostate case.

The total computation times averaged over 5 runs are reported in Table 3.4.

Table 3.4: Comparison of computation time on the brain case for the beamlet-based and beamlet-free method averaged over 5 runs.

Process		t [s]
BL-based	Plan Creation	299 ± 2
	Beamlet Comp.	3356 ± 24
	Optimization	12601 ± 1714
	Final Dose Comp.	301 ± 1
	Total	$\mathbf{16558 \pm 1714}$
BL-free	Plan Creation	299 ± 2
	Optimization	4697 ± 129
	Total	$\mathbf{4996 \pm 129}$

Prostate case

As for the brain case, the DVH obtained by both methods are depicted in Figure 3.6 and all objectives could be met. After 5 runs, the fluctuations on the target homogeneity, $D_5 - D_{95}$, were less than 3% for both methods.

The objective function convergence is shown in Figure 3.7.

Memory usage (Peak RAM), could be reduced from 5.97 GB for the beamlet-based method to 1.72 GB for the beamlet-free method and disk space from 5.1

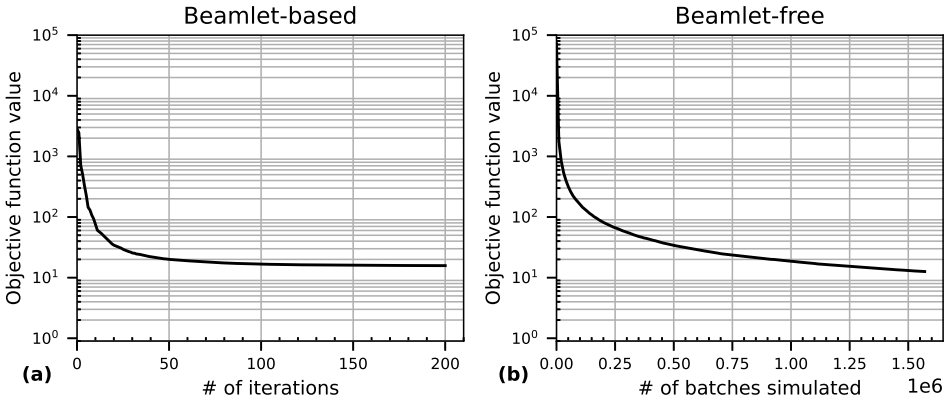


Figure 3.7: (a) Convergence plot of the beamlet-based algorithm (scipy L-BFGS) and (b) convergence plot of the beamlet-free algorithm for the prostate case.

GB for the beamlet-based to 4.5 GB for the beamlet-free method, as reported in Table 3.5. The beamlet matrix has a total size of 728 MB. Temporary files during optimization required an additional 1.7 GB of disk space for the beamlet-based case.

Table 3.5: Comparison of memory usage on the prostate case for the beamlet-based and beamlet-free method.

	Beamlet-based	Beamlet-free
Peak RAM usage [GB]	5.97	1.72
Storage [GB]	5.1	4.5

The total computation times averaged over 5 runs are reported in Table 3.6.
Additional tables and figures can be found in section 3.6.

3.4 Discussion

A novel treatment planning method, not limited to proton therapy, has been introduced. It merges dose calculation and optimization into a single process, resulting in improved computational efficiency and accuracy. The so-called beamlet-free algorithm was able to achieve comparable plan quality to a conventional beamlet-based algorithm.
The empirical complexity analysis revealed that, in beamlet-based treatment planning, computation time and peak memory usage increase almost linearly with the

Table 3.6: Comparison of computation time on the prostate case for the beamlet-based and beamlet-free method averaged over 5 runs.

Process		t [s]
BL-based	Plan Creation	68 ± 1
	Beamlet Comp.	339 ± 2
	Optimization	1930 ± 73
	Final Dose Comp.	135 ± 1
	Total	$\mathbf{2472 \pm 73}$
BL-free	Plan Creation	67 ± 1
	Optimization	1315 ± 88
	Total	$\mathbf{1382 \pm 88}$

number of spots. On the other hand, the beamlet-free method can achieve comparable plan quality without being affected linearly by the number of spots, except where the statistical uncertainty is concerned. The peak memory usage from the beamlet-free method was recorded below that of the beamlet-based simulation in all cases. With respect to computation time, for low numbers of spots, the beamlet-based method outperformed the beamlet-free method, but for increasing numbers of spots, the beamlet-based method gets slowed down considerably. It is possible to determine roughly from which number of spots it would be beneficial to utilize the beamlet-free instead of a beamlet-based method. However, elements like the machine running the computations, the exact beamlet-based optimizer and implementation chosen for the comparison, as well as the depth of the target and number of simulated particles, can impact the two methods differently and make it difficult to fix any specific number.

For very low numbers of spots, the beamlet-free method might be disadvantageous, although this possibility vanishes already for numbers of spots in the low thousands that are regularly encountered in clinical routines [176], where the beamlet-free method starts to outperform conventional beamlet-based treatment planning. The benefits of the beamlet-free method are particularly highlighted in settings that require a large number of beamlets, like robust optimization.

Two examplary patient cases showed that the proposed beamlet-free method is capable of optimizing complex clinical cases to an adequate plan quality. The optimized plans were compared to ones optimized with a classical beamlet-based approach and were of comparable quality. Memory usage could be reduced by 95% and 71% respectively and computation time by 70% and 44% through usage of the beamlet-free method. Especially the benefit with respect to computation time, depends on various factors such as individual optimization parameters, implementation and hardware, which the two methods do not draw the same benefit

from.

3.5 Conclusion

The proposed beamlet-free algorithm could successfully generate treatment plans of comparable quality to conventional treatment planning methods. Contrary to conventional treatment planning methods, computation time and memory usage are substantially less dependent of the number of spots, except for the increased number of particles that might be required to ensure low statistical uncertainty. The benefit to be drawn from the beamlet-free method increases for higher numbers of spots, which makes the method promising for new emerging treatment techniques such as robust optimization, adaptive, and arc proton therapy.

3.6 Additional Materials

3.6.1 Scaling Analysis

Figure 3.8 shows how much time was required for the different treatment planning steps by the beamlet-free and beamlet-based method respectively. For the beamlet-free method, dose calculation and optimization are combined in a single step, labelled "optimization". It can be noted that the time required for the beamlet calculation and optimization steps both increase with rising number of spots.

The total number of simulated protons in each approach is compared in Table 3.7.

Table 3.7: Total number of protons simulated (for the beamlet-based approach this includes only the beamlet calculation, but no final dose calculation).

Number of spots	Beamlet-based	Beamlet-free
650	32.5×10^6	40×10^6
1482	74.1×10^6	40×10^6
2794	139.7×10^6	40×10^6
4769	238.45×10^6	40×10^6
7307	365.35×10^6	40×10^6

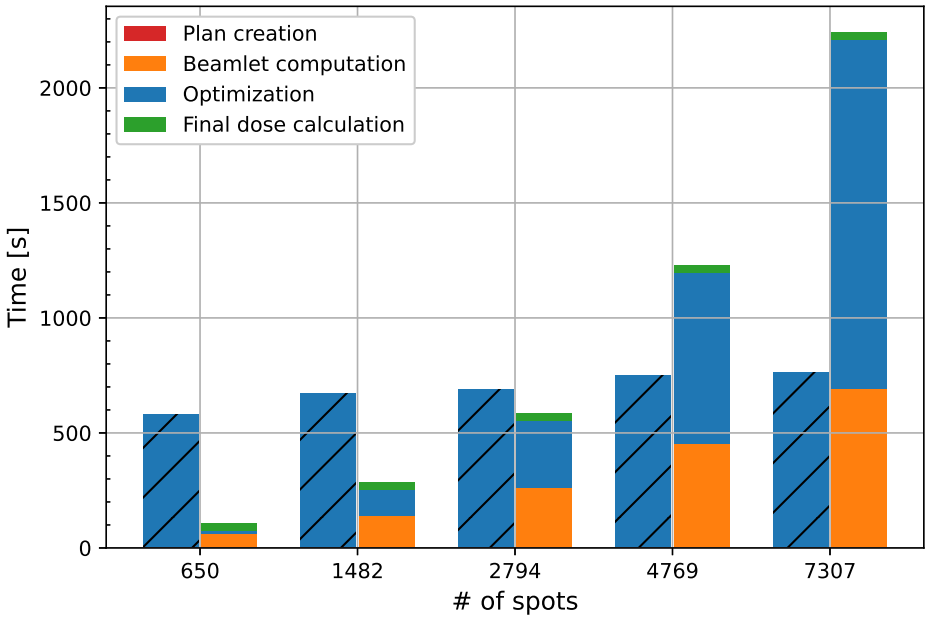


Figure 3.8: Time contributions of the individual treatment planning processes required for the beamlet-free method (left bar, striped) and the beamlet-based method (right bar, blank) for varying numbers of spots.

3.6.2 Patient data

Table 3.8 and Table 3.9 contain the evaluation of the clinical goals for the brain case and prostate case respectively. Figure 3.9 shows slices of the corresponding dose distributions.

The total number of simulated protons in each approach is compared in Table 3.10.

Table 3.8: Clinical goals evaluation of the brain case for the beamlet-based and beamlet-free method.

ROI	Criterium	Dose limit (Gy)	$D_{\text{Criterium}}$ (Gy) Beamlet-based	$D_{\text{Criterium}}$ (Gy) Beamlet-free
PTV	D_{95}	51.3	52.73	52.28
PTV	D_5	56.7	55.24	55.36
PTV	D_{98}	52.98	52.35	51.53
PTV	D_2	55.08	55.58	55.87
OC	D_5	60.0	55.08	55.09
OC	D_2	60.0	55.43	55.36
BS	D_5	55.0	53.11	52.55
BS	D_2	55.0	54.03	53.82
RON	D_5	55.0	51.74	49.99
RON	D_2	55.0	53.10	51.79
LON	D_5	55.0	54.57	54.58
LON	D_2	55.0	54.96	55.09

Table 3.9: Clinical goals evaluation of the prostate case for the beamlet-based and beamlet-free method.

ROI	Criterium	Dose limit (Gy)	$D_{\text{Criterium}}$ (Gy) Beamlet-based	$D_{\text{Criterium}}$ (Gy) Beamlet-free
CTV	D_{95}	57.0	58.73	58.73
CTV	D_5	63.0	61.25	61.23
CTV	D_{98}	58.8	58.37	58.37
CTV	D_2	61.2	61.57	61.53
Rectum	D_5	60.0	55.58	56.12
Rectum	D_2	60.0	59.00	59.34
Bladder	D_5	60.0	58.31	58.21
Bladder	D_2	60.0	59.67	59.73
AC	D_5	60.0	46.47	44.03
AC	D_2	60.0	54.86	53.37
RFH	D_5	55.0	32.23	32.62
RFH	D_2	55.0	33.25	33.92
LFH	D_5	55.0	30.96	31.13
LFH	D_2	55.0	31.84	31.95

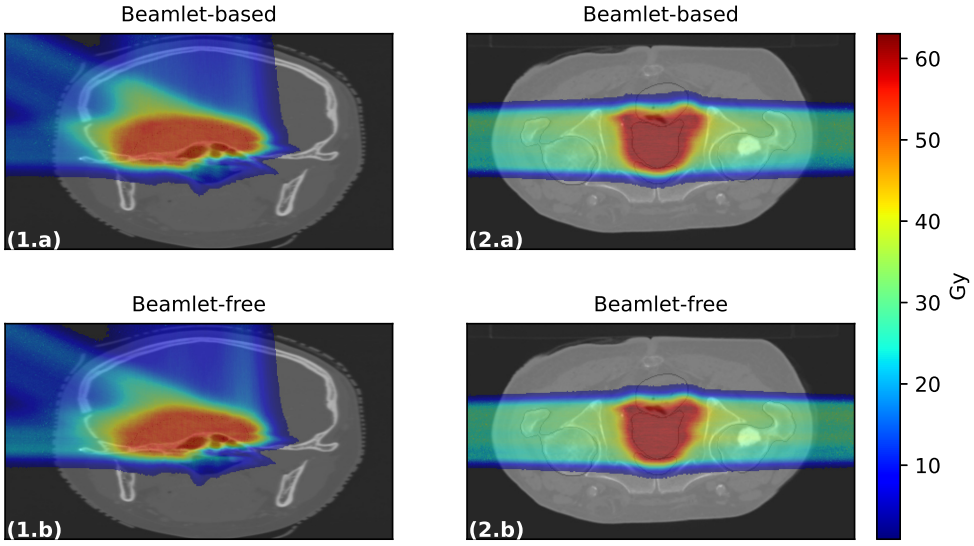


Figure 3.9: Comparison of the optimized dose distribution of the brain case (1) and of the prostate case (2) obtained by the beamlet-based (a) and beamlet-free method (b).

Table 3.10: Total number of protons simulated (for the beamlet-based approach this includes only the beamlet calculation, no final dose calculation).

Number of spots	Beamlet-based	Beamlet-free
Brain	490.6×10^6	200×10^6
Prostate	136.6×10^6	50×10^6

Chapter 4

Beamlet-free robust optimization

The content presented in this chapter corresponds an article submitted to Medical Physics Journal: Scalable, memory-efficient robust proton therapy optimization through beamlet-free treatment planning.

Authors: Danah Pross[1], Valentine Dormal[1], Kevin Souris[1,2], Ana Maria Barragan Montero[1], Edmond Sterpin[1,3,4], John A. Lee[1]

[1] *UCLouvain, Institut de recherche expérimentale et clinique, Molecular Imaging and Radiation Oncology Laboratory, Brussels, Belgium.*
[2] *Ion Beam Applications SA, Louvain-La-Neuve, Belgium.*
[3] *KULeuven, Department of Oncology, Leuven, Belgium.*
[4] *Particle, Leuven, Belgium.*

4.1 Introduction

Intensity-modulated proton therapy (IMPT) offers great advantages due to the physical properties of the protons. A considerable part of the dose is deposited at the end of their energy-dependent range in the Bragg peak, resulting in low entrance and, especially, no or low exit doses. However, due to a high number of treatment uncertainties, such as misalignment during patient positioning, changes to the patient geometry, errors in the translation of Houndsfield units (HU) from the computed tomography (CT) into proton stopping powers, and similar potential error sources, those steep dose gradients may quickly become a double edged sword, since even small errors can lead to significantly altered dose distributions.
In conventional radiotherapy with photons, margins are used to account for the effects of uncertainties, but in proton therapy, they may be insufficient [136], since they fail to accurately account for range errors and resulting distortions of the dose distribution. Robust optimization is used instead and is considered by many to be an integral part of proton therapy. Robust optimization refers to a group of techniques integrated in the treatment planning process to account for uncertainties and guide the optimizer towards a final dose distribution that is more robust with respect to them.
There are two main approaches utilized in robust optimization for proton therapy treatment planning, the most common of which is minimax optimization. It can be further separated into worst-case scenario optimization, also referred to as composite worst-case [25], voxel-wise worst-case dose distribution optimization [177, 149, 23] and objective-wise worst-case optimization [142].
Worst-case scenario optimization is the most common one, since voxel-wise worst-case optimization is based on unrealistic dose distributions, leading to an even more conservative approach than the worst-case scenario optimization, and objective-wise worst-case optimization was investigated specifically to aid in multi-criteria optimization and was found to deliver less robust plans than other minimax methods in other settings [178].
The second main approach is expected-value robust optimization, also referred to as stochastic programming. As for the other approaches, separate error scenarios are defined and corresponding scenario dose distributions calculated. Each scenario is associated with a probability meant to represent the likelihood of its occurrence. The expected value objective function over all scenarios is then calculated and optimized [26, 27].
Strategies to combine the principles of minimax and probabilistic optimization also exist [179, 180].
Independently of which robust optimization approach is used, computation time and memory usage are increased considerably. This especially limits the develop-

ment of new methods, such as online adaptive therapy [181].
The current computational limitations posed by robust optimization methods also limit the number of scenarios that can realistically be taken into account and often require approximations in their implementation [135]. This leads to a coarse sampling of the error space. Combinations of different error sources are not, or only to a limited extent, taken into consideration. Often, only systematic set-up and range errors are employed during optimization, when the inclusion of additional error sources such as anatomical changes was shown to lead to significantly more robust plans [182, 183, 184, 185].
Possibilities to speed-up robust optimization have been investigated, like a dynamic minimax robust optimization to decrease the number of scenarios to be evaluated at each iteration [16] or the scalable robust optimization method, recently developed by Fu et al., which is based on the division of the optimization problem into multiple subproblems and their distribution across multiple processors to be solved in parallel [18]. These methods can achieve impressive speed-up rates compared to conventional robust optimizers, but ultimately they can only reduce the optimization time itself, while the considerable time needed for beamlet computation remains unchanged.
Xu et al. developed a fast robust optimizer based on a small set of 9 scenarios, a fast analytical dose calculation tool and a combined CPU-GPU-parallelization approach [17]. While this approach performed well on three clinical cases when compared to the robust optimizer in the Varian Eclipse system, the small number of considered scenarios and analytical dose calculation instead of Monte Carlo put severe limitations on this method.
Cristoforetti et al. proposed a robust optimization method based on the expected dose distribution across several scenarios [186]. This method reduces both computation time and memory usage during optimization. However, it still requires the timewise usually much more impactful beamlet calculation and a precomputation step to compute the expected dose and variance.
A beamlet-free treatment planning approach, as developed previously [24], can help reduce computation time and memory usage compared to conventional beamlet-based methods. It relies on simultaneous dose calculation through Monte Carlo (MC) simulations and spot weight optimization with stochastic coordinate descent [107] as the optimization scheme. The beamlet-free planning approach simulates particles in small batches to one randomly selected spot at a time and stores only the total 3D dose matrix, not individual beamlet contributions. The gradient is calculated from the impact of a single batch on the dose distribution and used to adjust the probability of the spot being picked again in the future. The exact performance benefits depend on the specific case, size of the plan, and hardware used. The present work seeks to propose a new robust optimization

strategy based on beamlet-free optimization and to investigate whether such an approach can yield a reduction in computation time and memory usage. Robust optimization is especially impacted by the high computational burden of conventional treatment planning and can at the same time expect higher potential benefits from a beamlet-free approach due to decreased noise levels. Both the feasibility of a beamlet-free robust optimization approach and the potential benefit over conventional methods are investigated.

4.2 Methods

4.2.1 Robust optimization approaches

Beamlet-based worst-case scenario approach

In robust treatment planning, error scenarios are typically represented as lateral shifts of the beam in relation to the patient geometry, or a scaling factor applied to the density of the CT data during dose calculation.
For each error scenario that is defined to be used during robust optimization, an individual beamlet matrix needs to be computed, containing the beamlet contributions to the total dose distribution and, therefore, time and memory required for its computation scale with the number of voxels and the number of spots in the plan. Since the dose calculation is required for each scenario, time and memory additionally scale with the number of scenarios. This requires the total number of scenarios taken into consideration to be kept low, since otherwise computational requirements would breach the bounds of clinical feasibility.
At each iteration during optimization, the objective function of each scenario needs to be evaluated to determine the worst-case scenario:

$$\begin{aligned} &\min_{x} \max_{s \in S} f_s(\mathbf{x}) \\ &\text{subject to } \mathbf{x} \geq \mathbf{0} \ . \end{aligned} \tag{4.1}$$

The optimization problem is defined in equation (4.1), where $\mathbf{x}$ is the optimization parameter (e.g., the spot weights), S the scenario space and $f_s(\mathbf{x})$ the objective function of scenario s, which depends on the scenario beamlet matrix D_s. The problem is convex whenever f is.

Beamlet-based expected-value approach

Probabilistic robust optimization seeks to minimize the expected value of the sum over the scenario objective functions multiplied by a scenario weight:

$$\begin{aligned} \min_x f &= \sum_{s \in S} p_s f_s(x) \\ \Leftrightarrow \min_x f &= \sum_{s \in S} w_s f_s(D_s, x) \\ \text{subject to } x &\geq 0 \ . \end{aligned} \tag{4.2}$$

The optimization problem is formulated in equation (4.2), where w_s is the scenario weight of scenario s.
One obvious disadvantage of expected-value robust optimization is that it does not guarantee an acceptable worst-case scenario. Since the expected value over all scenarios is optimized, the difference between scenarios can potentially be big and outliers can persist. It can be considered less conservative than minimax robust optimization approach.
Expected-value robust optimization is less common in clinical practice and the number of publications studying its performance compared to minimax robust optimization is small. Fredriksson et al. [179] compared worst-case scenario optimization with expected-value robust optimization on a 2D phantom. They showed that for systematic errors, target coverage robustness is increased with the conservativeness of the method, while dose to organ-at-risk robustness is decreased.

Beamlet-free expected-value approach

In this approach, expected-value robust optimization, as it is defined above, is combined with a beamlet-free optimization approach [24]. To adapt the beamlet-free optimization workflow for robust optimization, a separate dose scoring grid needs to be defined for each scenario. In each iteration of the algorithm, one batch of particles is simulated for each scenario. The error scenarios are applied on an individual proton level, meaning that density scaling factors differ from proton to proton and need to be applied at each Monte Carlo step. A separate gradient is calculated for each scenario based on equation (4.2). As for the non-robust implementation of the beamlet-free algorithm [24], a batch is simulated towards a single, randomly chosen spot j. It is the same spot for each scenario, resulting in the gradient containing a single entry $\frac{\partial f}{\partial x_j}$. The spot weights x_j correspond to the number of particles simulated towards each spot, while the quantity optimized in the optimization step is the probability p_j of spot j being picked again, as

$$p_j^{k+1} = p_j^k - h \frac{\partial f}{\partial x_j^k} \tag{4.3}$$

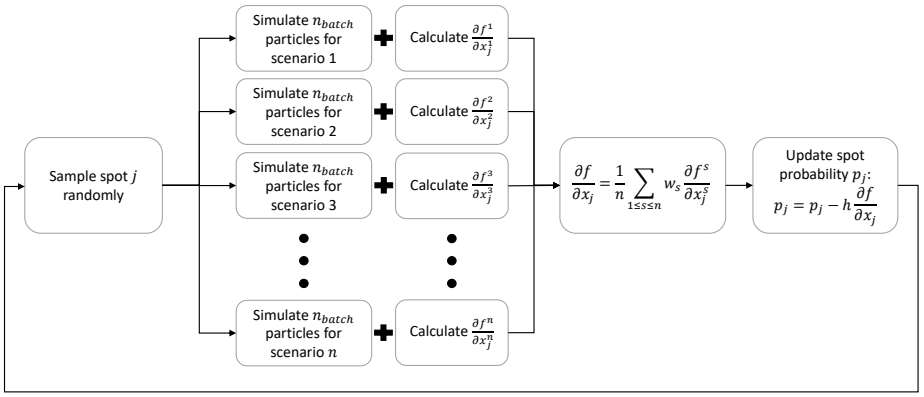


Figure 4.1: Workflow diagram of the beamlet-free expected-value robust optimization process.

with h being an experimentally determined step size. The spot weights are indirectly optimized through the updated probability distribution.
After the batch has been fully simulated, spot weights are adjusted before the next iteration. The workflow is presented in Figure 4.1.

One important particularity of the beamlet-free algorithm that needs to be adapted for robust optimization is the dose normalization. Over the course of the Monte Carlo simulation, the dose is normalized after each step to scale the absolute values of the scored dose to the dose prescription. This is important, since the premise of the beamlet-free algorithm centers on achieving the best relative dose distribution by stacking on more particles where needed. It cannot freely decrease spot weights like a beamlet-based optimizer can.
The scenario gradient is approximated as

$$(\nabla f_s^{k+1})_j = (\nabla f_s^k)_j + f(n^{k+1} d_s^{k+1}) - f(n^k d_s^k) \ , \tag{4.4}$$

where f_s is the objective function for scenario s, k is the current Monte Carlo step, d_i^k is the current dose distribution for scenario i at Monte Carlo step k and n^k is the normalization coefficient at Monte Carlo step k. At $k = 0$, the dose in each voxel i is $d_{i,s}^k = 0$.
For the non-robust beamlet-free optimization, the normalization factor is defined

as

$$n^k = \frac{b_{TV}^{pres}}{\hat{d}_{TV}^k} \ , \tag{4.5}$$

where b_{TV}^{pres} is the target prescription and $\hat{d}_{TV}^k$ is the mean target dose at Monte Carlo step k. For robust beamlet-free optimization, this cannot be done for each scenario separately, because at the end, only a single set of normalized spot weights needs to be exported. Instead, the mean target dose for all scenarios is averaged and used instead.

Expected-value robust optimization is well suited to this joint approach and was chosen over a worst-case scenario approach because it better supports the stochastic nature of the beamlet-free method. The beamlet-free approach calculates the gradient based on a very small batch of particles which makes it very noisy. In non-robust beamlet-free optimization, a batch size of 16 protons was chosen as a compromise between sufficiently small noise and sufficiently small contributions per batch so as not to inflate computation time. For a robust optimization approach, the number of particles to be simulated scales with the number of scenarios n. The stochastic noise of the gradient is reduced by a factor $\frac{1}{\sqrt{n}}$ in the case of expected-value robust optimization, while it maintains the noise level of the non-robust case for worst-case robust optimization. This additional noise benefit is not as relevant in beamlet-based robust optimization, where the noise at the time of the gradient calculation is much lower.

Additionally, it should be pointed out that the calculation of the objective function during beamlet-free optimization is not as straightforward. The usual calculation of the objective function iterates over the voxel of the total dose distribution. In beamlet-based optimization, this is done at each iteration, but does not have a big impact on the overall computation time since the number of iterations usually does not surpass a few hundred. In beamlet-free optimization, the number of iterations is much larger and computing the objective function at every iteration would significantly prolong computation time and interfere with the parallelization of the code.

Overall, a beamlet-free worst-case robust optimization approach has been investigated, but was deemed too inefficient compared to an expected-value approach and not pursued further. An expected-value robust optimization approach, should theoretically be able to achieve better relative performance benefits compared to beamlet-based methods, than the non-robust implementation, because it exploits these noise benefits at no additional computational cost.

Variance-guided beamlet-free optimization

The beamlet-free expected-value approach presented above has a decisive disadvantage compared to worst-case robust optimization methods. Since it minimizes the expected-value objective function over all scenarios but does not constrain the variance between the individual scenarios, it is possible that some scenarios emerge as outliers.
The feasibility of an expected-dose based variance term to successfully guide robust optimization was recently shown by Cristoforetti et al. [186] who used an objective function of the form

$$F(\mathbf{x}) = \bar{f}(\mathbf{x}) + \mathbf{x}^\top \Omega \mathbf{x} \tag{4.6}$$

where $\bar{f}$ is the objective function evaluated over the expected dose influence matrix $E[D]$ and $\mathbf{x}$ is the vector over the spot weights. The variance term is calculated as

$$\Omega = (E[D^\top D] - E[D]^\top E[D]) \ . \tag{4.7}$$

In the beamlet-free robust implementation, a variance term is introduced to reduce inter-scenario variability. It takes the form

$$\Omega(\mathbf{x}) = \sum_s (f_s(\mathbf{x}) - \hat{f}(\mathbf{x})^2) \tag{4.8}$$

with f_s as the scenario objective function and $\hat{f}$ as the objective function evaluated over the expected dose distribution $\bar{d}$. The expected dose distribution is computed by accumulating the dose contributions from all scenario simulations in a single dose scoring grid and normalizing over the number of scenarios. The full objective function can then be formulated as

$$F(\mathbf{x}) = \frac{1}{n_s} \sum_s f_s(\mathbf{x}) + p_\Omega \Omega(\mathbf{x}) \tag{4.9}$$

for the spot weight vector $\mathbf{x}$ and the tunable parameter p_Ω.
This expected dose distribution is used instead of the expected-value objective function, because it facilitates the computation of the gradient. Differences to the formulation used by Cristoforetti et al. result from the differences in purpose and implementation.
The variance term is evaluated in addition to the beamlet-free expected-value implementation without variance term. It was investigated specifically to observe if it can lead to improvements for challenging cases prone to develop outlier scenarios.

4.2.2 Optimization parameters

The treatment plan optimization is performed using the in-house developed, research treatment planning system OpenTPS [28].
All simulations involve 21 distinct error scenarios, one nominal scenario and systematic setup errors in all 6 spatial directions, each calculated with positive and negative range errors, as well as without.
For the beamlet-based methods, a beamlet matrix with $5 \cdot 10^4$ protons per beamlet is simulated and subsequently optimized using the scipy [122] FORTRAN implementation (wrapped in Python) of the second-order optimization algorithm L-BFGS [129] (Limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm). The optimizer terminates once a predefined objective function value is reached for the worst performing scenario. This cut-off objective function value was determined by performing 5 separate optimizations with a robust worst-case optimizer and a classic convergence criterion of $f_{\text{tol}} = 10^{-4}$ between successive evaluations. A final dose computation is performed with a number of particles chosen to match the statistical accuracy of the beamlet-free method. Both the beamlet calculation and the final dose computation are performed using the MCsquare dose engine [12] that is also utilized in the beamlet-free algorithm.
For the beamlet-free algorithm, the number of particles used in the MC dose calculation and the number of iterations performed by the optimization algorithm are tied together through the batch size of 16 protons per iteration [24]. The number of simulated particles, and consequently also the number of iterations, required by the beamlet-free algorithm depend on the plan size and complexity. The simulated number of particles and iterations was set for each patient individually and can be found in section 4.3. The iteration number for each patient is manually adjusted so that the worst-case scenario objective function is equal to or lower than that achieved with worst-case scenario optimization.
All objectives are optimized robustly and the scenario probabilities are set to $w_s = 1$ for all scenarios.
Simulations were performed on a computation server with a total of 120 CPU cores and 256 GB RAM.

4.2.3 Plan comparison metrics

To compare these different methods, the convergence of the scenario objective function has been observed and was used to determine adequate cut-off points for the optimizers.
After the robust optimization, a robustness evaluation based on good practice scenario selection is performed as defined by Sterpin et al. [187]. For the robustness evaluation, $n_s = 81$ error scenarios are taken into consideration. These include

the 21 scenarios already defined in section 4.2.2, as well as scenarios containing combinations of systematic setup errors with one another.
The dose volume histogram (DVH) is calculated for each method. The robustness of the methods can be qualitatively asserted by comparing the robustness bands of the different ROIs.
The $D98$ for the nominal and worst-case scenario are reported. The reported worst-case scenario is the scenario with the $D98$ that shows least agreement with the dose prescription. In all following metrics, the reported worst-case is always the worst performing scenario for that metric, which means that it may differ from one metric to the next.
For a quantitative analysis, the target homogeneity index for each scenario is calculated as

$$HI = D5 - D95 \tag{4.10}$$

and the nominal, worst-case scenario and expected value target homogeneity,

$$E[HI] = \frac{1}{n_s} \sum_s HI_s \tag{4.11}$$

over all scenarios, for each method are compared. The same is done for the target conformity index

$$CI = \frac{V_{TV,95}^2}{V_{p,95} V_{TV}} \ , \tag{4.12}$$

where $V_{TV,95}$ is the volume of the target covered by more than or equal to 95% of the target prescription dose, referred to as the 95th percentile isodose, $V_{p,95}$ is the total volume of the patient covered by the 95th percentile isodose and V_{TV} is the target volume. Equation (4.12) defines the Paddick conformity index [77], which takes target coverage into account alongside conformity. It gives a measure for the extend of the 95th percentile isodose beyond the target with a perfect conformity corresponding to $CI = 1$. Additionally, the mean square error (MSE), was computed and reported for the nominal, worst-case scenario and the expected value over all scenarios. It is defined as

$$MSE = \frac{1}{N} \sum_n^N (d_n - d^{\text{pres}})^2 \tag{4.13}$$

with N the number of voxel in the target, d_n the dose in voxel n and d^{pres} the target prescription. The MSE cannot give an objective measure on plan quality per se, but it can be used to quantify the differences between different plans on the same geometry and prescription.
The objective function is recalculated on the resimulated scenario dose distributions. While it highly depends on the respective plan and objectives and is not a

meaningful quantity for the plan quality by itself, it is the property that is actually optimized and gives valuable insight into the quality of the optimization.

4.2.4 Patient data

The different algorithms are compared on six patient cases with varying numbers of spots.
Two brain cases, two prostate cases, and two head and neck cases were chosen for the comparison. The local ethics committee approved the use of the patient data.
The beam model assumes a double Gaussian spot shape with energy-dependent parameters.

Brain cases

Identical plan parameters and objectives were used for both brain cases. Tumor size and location varied, as did the number of spots.
A treatment plan was generated using a three-beam combination at a fixed gantry angle G $= 90°$ and rotating the couch angle C at $0°$, $270°$ and $315°$. The spot and layer spacing were set to 4 mm and a margin of 16 mm for the spot placement was applied around the target to account for setup error scenarios. A total of 17541 spots for brain case 1 and 14832 spots for brain case 2 compose the plans. The ROIs upon which optimization objectives are placed are the gross target volume (GTV), optical chiasm (OC), brain stem (BS), right optical nerve (RON) and left optical nerve (LON). The optimization objectives are listed in Table 4.1.
For brain case 1, the resolution of the dose grid and the CT data was scaled to 1.872 mm $\times$ 1.872 mm $\times$ 1.000 mm and cover a total of $128 \times 128 \times 229$ voxel.
For brain case 2, the resolution of the dose grid and the CT data was scaled to 2.148 mm $\times$ 2.148 mm $\times$ 1.000 mm and cover a total of $256 \times 256 \times 267$ voxel.

Prostate cases

Identical plan parameters and objectives were used for both prostate cases. Tumor size and location varied, as did the number of spots.
The plans consist of two opposing beams with a fixed gantry angle of $0°$ and couch angles of $90°$ and $270°$. Spot and layer spacing were set to 6 mm and a margin of 24 mm was applied around the target for spot placement. The plan for prostate case 1 contains 9416 spots and that of prostate case 2 contains 8812.
The regions of interest consist of the the CTV, rectum, bladder, right and left

Table 4.1: Optimization objectives of the brain cases.

ROI	Objective Type	Dose limit	Weight
GTV	D_{min}	54.0	5
GTV	D_{max}	54.0	5
OC	D_{max}	60.0	1
BS	D_{max}	55.0	1
RON	D_{max}	55.0	1
LON	D_{max}	55.0	1

femoral head (RFH/LFH), and anal canal. The optimization objectives are listed in Table 4.2.
For both cases, the dose grid and CT resolution were scaled to 2.148 mm $\times$ 2.148 mm $\times$ 2.000 mm and cover a total of $256 \times 256 \times 209$ voxel.

Table 4.2: Optimization objectives of the prostate cases.

ROI	Objective Type	Dose limit [Gy]	Weight
CTV	D_{min}	60.0	100
CTV	D_{max}	60.0	100
Rectum	D_{max}	60.0	5
Bladder	D_{max}	60.0	5
Anal canal	D_{max}	60.0	5
RFH	D_{max}	55.0	1
LFH	D_{max}	55.0	1

Head and neck cases

Identical plan parameters and objectives were used for both head and neck cases. Tumor size and location varied, as did the number of spots.
The plans consist of four beams whose gantry and couch angles are listed in Table 4.3. Spot and layer spacing are set to 5 mm and a margin of 12 mm is applied around the target. The plan consists of 36064 spots for case 1 and of 41347 spots for case 2. The ROIs in question are the spinal cord, brain stem and several regions within the CTV with different prescriptions and weights as listed in Table 4.4, including the primary tumor (CTVp7000) and the nodal region having infiltrated tumor (CTVn7000) with a prescription of 70 Gy, and the prophylactic nymph nodes (CTVn5425) with a prescription of 54.25 Gy. The maximum dose objectives on the prophylactic lymph nodes are applied on a cropped ROI structure (CTVncrop) due to overlap with the infiltrated tumor.

Table 4.3: Beam configuration of the head and neck cases.

	Gantry angle [°]	Couch angle [°]
Beam 1	60	10
Beam 2	120	350
Beam 3	240	10
Beam 4	300	350

Table 4.4: Optimization objectives of the head and neck cases.

ROI	Objective Type	Dose limit [Gy]	Weight
Spinalcord	D_{max}	44.0	10
Brainstem	D_{max}	53.0	10
CTVp7000	D_{max}	70.0	2000
CTVp7000	D_{min}	70.0	2000
CTVnL5425	D_{min}	54.25	400
CTVnR5425	D_{min}	54.25	400
CTVnLcrop	D_{max}	54.75	400
CTVnRcrop	D_{max}	54.75	400
CTVnR7000	D_{max}	70.0	400
CTVnR7000	D_{min}	70.0	400

For both cases, the dose grid and CT resolution were scaled to 2.148 mm $\times$ 2.148 mm $\times$ 2.000 mm and covers a total of $256 \times 256 \times 149$ voxel.

4.3 Results

4.3.1 Brain case 1

The results for brain case 1 are summarized in Table 4.5. The corresponding dose volume histograms (DVH) can be seen in Figure 4.2. The DVH show good agreement between methods for the target, but reveal slight differences with respect to the OARs, notably the brain chiasm and brain stem. Since the dose limits for both OARs are relatively high, the objective function had no incentive to specifically target these OARs in the respective dose regions.
For all investigated methods, the D98 agrees with the target prescription within 5% for the nominal case, but not for all scenarios. For all three methods the pass rate of the target D98 was above 97%. While the average over 5 runs performed slightly better with the beamlet-based expected value method, the lowest pass rate across all 5 runs was identical for all methods.

The beamlet-free approach matches or outperforms the beamlet-based worst-case method with respect to all investigated metrics except for the target conformity index where the beamlet-based worst-case performed slightly better than either expected-value approach. The beamlet-based expected value approach managed to achieved better D98 and Target HI values for the worst scenario during evaluation, but performed very similar to the beamlet-free approach with respect to the other metrics. Both expected-value methods achieved better metrics for the nominal case and averaged over all scenarios, as was to be expected.
The beamlet-free approach required considerably less computation time than either beamlet-based method and achieved a reduction of about 81% compared to the beamlet-based worst-case method. For both beamlet-based methods, the majority of the computation time was required for the beamlet calculation. The beamlet-based expected value approach overall took longer to optimize. For all methods, the computation time between individual runs fluctuated considerably.
With respect to peak memory usage, the beamlet-free approach required a consistent 1.1 GB, which respresents a reduction of 89% compared to the beamlet-based method. The 9.7 GB recorded for the beamlet-based methods represent the peak memory during beamlet calculation and optimization. The actual memory usage was considerably lower for the majority of the beamlet calculation, then increased considerably for the optimization.
Results with and without the additional variance term were very similar and either agreed within error bounds or didn't reveal a clear advantage for either one over the other.

4.3.2 Brain case 2

The results for the brain case 2 are summarized in Table 4.6. The corresponding dose volume histograms can be seen in Figure 4.3. As for brain case 1, the DVH show good agreement with respect to the target, but slight differences between the beamlet-free and beamlet-based approaches for the brain chiasm and brain stem. Again, these differences occur in dose regions below the respective threshold values and were not actively targeted during optimization.
For all investigated methods, the D98 agrees with the target prescription within 5% for the nominal case, but not for all scenarios. For all three methods the pass rate of the target D98 was above 97%. While the average pass rate over 5 runs was slightly better with the beamlet-based expected value method, the lowest pass rate across 5 runs was identical for all methods.
The metrics for the worst-scenario of the beamlet-based worst-case scenario and beamlet-free methods are in agreement within 3σ except for the CI and D98 which are slightly lower for the beamlet-free approach. The beamlet-based expected-value implementation achieved better worst-case objective function and target

Table 4.5: Optimization statistics on brain case 1.

	Beamlet-based worst-case	Beamlet-based expected-value	Beamlet-free expected-value	Beamlet-free + variance
# iterations	169 ± 8	82 ± 3	2500000	2500000
Beamlet computation time [s]	20610 ± 83	20610 ± 83	0	0
Optimization time [s]	1638 ± 139	5281 ± 214	4234 ± 813	5059 ± 608
Total computation time [s]	22249 ± 222	25891 ± 297	4234 ± 813	5059 ± 608
Peak RAM [GB]	9.7 ± 0.2	9.7 ± 0.1	1.1 ± 0.1	1.1 ± 0.1
Objective function nominal	0.216 ± 0.004	0.170 ± 0.006	0.162 ± 0.004	0.157 ± 0.003
Objective function worst-case	0.424 ± 0.014	0.356 ± 0.006	0.405 ± 0.013	0.378 ± 0.012
E[Objective function]	0.265 ± 0.001	0.194 ± 0.002	0.187 ± 0.001	0.178 ± 0.001
Target D98 nominal [Gy]	52.46 ± 0.04	52.58 ± 0.04	52.64 ± 0.01	52.67 ± 0.05
Target D98 worst-case [Gy]	50.84 ± 0.12	51.33 ± 0.07	50.91 ± 0.01	51.03 ± 0.11
E[Target D98] [Gy]	52.25 ± 0.04	52.44 ± 0.01	52.49 ± 0.01	52.52 ± 0.03
Target HI nominal	2.52 ± 0.03	2.26 ± 0.04	2.23 ± 0.03	2.17 ± 0.02
Target HI worst-case	3.05 ± 0.05	2.69 ± 0.02	2.81 ± 0.01	2.71 ± 0.01
E[Target HI]	2.70 ± 0.01	2.35 ± 0.01	2.30 ± 0.01	2.24 ± 0.01
Target CI nominal	0.231 ± 0.001	0.223 ± 0.001	0.217 ± 0.001	0.231 ± 0.003
Target CI worst-case	0.203 ± 0.001	0.196 ± 0.001	0.195 ± 0.002	0.215 ± 0.003
E[Target CI]	0.241 ± 0.001	0.233 ± 0.001	0.224 ± 0.002	0.239 ± 0.003
Target MSE nominal	0.62 ± 0.01	0.46 ± 0.01	0.45 ± 0.01	0.44 ± 0.01
Target MSE worst-case	1.18 ± 0.06	1.01 ± 0.03	1.17 ± 0.03	1.06 ± 0.03
E[Target MSE]	0.72 ± 0.01	0.54 ± 0.01	0.52 ± 0.01	0.49 ± 0.01

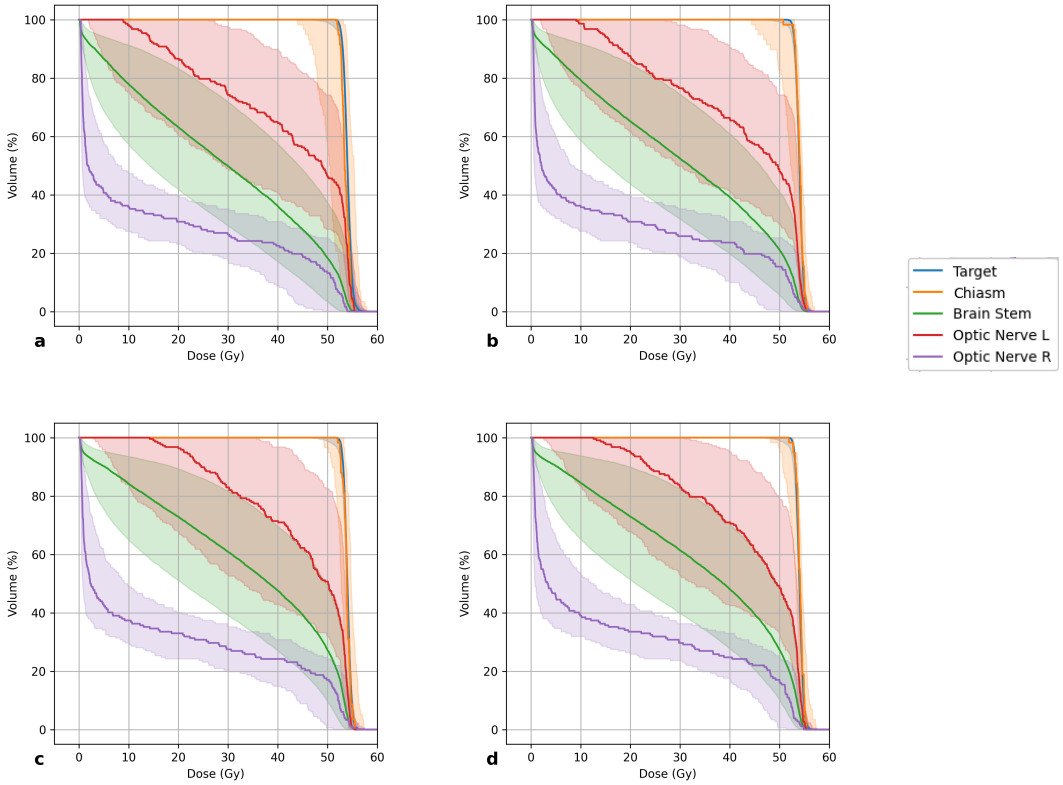


Figure 4.2: Dose volume histograms (DVH) for brain case 1 with the a) beamlet-based worst case, b) beamlet-based expected-value, c) beamlet-free expected value approach and d) beamlet-free expected value approach with additional variance term.

MSE values than the other methods. For the nominal and average scenarios, the two expected-value methods achieve better metrics than the worst-case scenario approach, as is to be expected, with the noticable exception of the CI. The beamlet-based expected-value approach consistently delivered equal or better results than the beamlet-free method.
The computation time of the beamlet-free method showed a decrease of 80% compared to the beamlet-based worst-case scenario approach, while peak memory could be reduced by 90%.
As for brain case 1, the results with and without additional variance term are very similar to each other.

Table 4.6: Optimization statistics on brain case 2.

	Beamlet-based worst-case	Beamlet-based expected-value	Beamlet-free expected-value	Beamlet-free + variance
# iterations	138 ± 5	57 ± 2	2500000	2500000
Beamlet computation time [s]	36310 ± 217	36310 ± 217	0	0
Optimization time [s]	2843 ± 176	7143 ± 393	7767 ± 1138	8940 ± 666
Total computation time [s]	39153 ± 393	43453 ± 610	7767 ± 1138	8940 ± 666
Peak RAM [GB]	14.0 ± 0.3	14.0 ± 0.1	1.4 ± 0.1	1.4 ± 0.1
Objective function nominal	0.445 ± 0.012	0.236 ± 0.021	0.262 ± 0.007	0.279 ± 0.011
Objective function worst-case	0.603 ± 0.008	0.538 ± 0.010	0.614 ± 0.007	0.514 ± 0.015
E[Objective function]	0.507 ± 0.004	0.368 ± 0.021	0.404 ± 0.003	0.363 ± 0.003
Target D98 nominal [Gy]	51.37 ± 0.02	52.27 ± 0.06	51.93 ± 0.06	51.84 ± 0.07
Target D98 worst-case [Gy]	50.70 ± 0.14	50.68 ± 0.08	50.28 ± 0.10	50.80 ± 0.07
E[Target D98] [Gy]	51.22 ± 0.08	51.66 ± 0.03	51.37 ± 0.11	51.55 ± 0.02
Target HI nominal	3.74 ± 0.06	2.59 ± 0.06	2.87 ± 0.07	3.00 ± 0.06
Target HI worst-case	4.30 ± 0.08	4.25 ± 0.09	4.54 ± 0.05	4.06 ± 0.10
E[Target HI]	3.92 ± 0.02	3.34 ± 0.01	3.59 ± 0.01	3.41 ± 0.02
Target CI nominal	0.290 ± 0.003	0.275 ± 0.001	0.238 ± 0.002	0.248 ± 0.003
Target CI worst-case	0.238 ± 0.002	0.229 ± 0.002	0.200 ± 0.002	0.212 ± 0.002
E[Target CI]	0.292 ± 0.002	0.276 ± 0.001	0.240 ± 0.001	0.253 ± 0.003
Target MSE nominal	1.24 ± 0.03	0.62 ± 0.02	0.73 ± 0.02	0.77 ± 0.03
Target MSE worst-case	1.68 ± 0.03	1.49 ± 0.04	1.71 ± 0.02	1.43 ± 0.04
E[Target MSE]	1.40 ± 0.01	1.01 ± 0.01	0.92 ± 0.40	1.01 ± 0.01

4.3.3 Prostate case 1

The results for prostate case 1 are summarized in Table 4.7. The corresponding dose volume histograms can be seen in Figure 4.4. The DVH show good agreement for the target and slight differences with respect to the OARs, with all OAR showing slightly better sparing with the beamlet-free method except for the anal canal.
The D98 agrees with the target prescription within 5% for all scenarios and all methods.
The two beamlet-based methods performed similarly for most metrics with the expected-value approach showing slightly better results for the nominal and average scenario. The beamlet-free method outperformed both beamlet-based methods in all metrics and scenarios.
Compared to the beamlet-based worst-case scenario approach, the beamlet-free method managed to reduce computation time by 63% and peak memory usage by 92%.
Unlike for the two brain cases, adding the variance term during optimization lead to a clear deterioration of optimization results. The parameter p_Ω regulating the impact of the variance on the global objective function, was fine tuned for head and neck case 2, better results could likely have been obtained with patient specific fine tuning.

4.3.4 Prostate case 2

The results for prostate case 2 are summarized in Table 4.8. The corresponding dose volume histograms can be seen in Figure 4.5. Again, the DVH showed good agreement for the target and slight differences with respect to the OARs, with slight sparing benefits for the beamlet-free method.
The D98 agrees with the target prescription within 5% for all scenarios and all methods.
As for prostate case 1, the two beamlet-based methods show similar results for all metrics with the expected-value approach performing slightly better for the nominal and average scenarios with respect to objective function value, HI and target MSE. The beamlet-free method achieved better results with respect to all metrics.
Compared to the beamlet-based worst-case scenario method, the beamlet-free approach resulted in a reduction of 71% in computation time and a reduction of 91% in peak memory usage.
Similarly to prostate case 1, the addition of the variance term during optimization leads to considerably worse optimization results after the same number of iterations.

Table 4.7: Optimization statistics on prostate case 1.

	Beamlet-based worst-case	Beamlet-based expected-value	Beamlet-free expected-value	Beamlet-free + variance
# iterations	273 ± 18	53 ± 1	1250000	1250000
Beamlet computation time [s]	15989 ± 95	15989 ± 95	0	0
Optimization time [s]	7836 ± 957	7479 ± 586	8789 ± 787	9609 ± 1530
Total computation time [s]	23825 ± 1052	23468 ± 681	8789 ± 787	9609 ± 1530
Peak RAM [GB]	16.5 ± 0.1	16.5 ± 0.1	1.3 ± 0.1	1.3 ± 0.1
Objective function nominal	0.444 ± 0.016	0.385 ± 0.009	0.312 ± 0.005	1.161 ± 0.109
Objective function worst-case	0.594 ± 0.007	0.600 ± 0.012	0.513 ± 0.018	2.313 ± 0.276
E[Objective function]	0.498 ± 0.005	0.454 ± 0.003	0.371 ± 0.005	1.476 ± 0.123
Target D98 nominal [Gy]	58.03 ± 0.03	58.15 ± 0.01	58.34 ± 0.02	56.77 ± 0.21
Target D98 worst-case [Gy]	57.37 ± 0.08	57.26 ± 0.06	57.52 ± 0.05	54.53 ± 0.34
E[Target D98] [Gy]	57.93 ± 0.02	57.96 ± 0.04	58.14 ± 0.03	56.32 ± 0.22
Target HI nominal	3.20 ± 0.07	3.00 ± 0.03	2.70 ± 0.04	5.17 ± 0.28
Target HI worst-case	3.63 ± 0.02	3.62 ± 0.04	3.29 ± 0.04	6.62 ± 0.41
E[Target HI]	3.33 ± 0.02	3.21 ± 0.02	2.90 ± 0.01	5.58 ± 0.27
Target CI nominal	0.324 ± 0.003	0.329 ± 0.001	0.343 ± 0.005	0.381 ± 0.021
Target CI worst-case	0.229 ± 0.001	0.230 ± 0.001	0.301 ± 0.005	0.352 ± 0.017
E[Target CI]	0.345 ± 0.003	0.349 ± 0.001	0.356 ± 0.005	0.358 ± 0.267
Target MSE nominal	0.95 ± 0.03	0.83 ± 0.02	0.67 ± 0.01	2.48 ± 0.23
Target MSE worst-case	1.28 ± 0.02	1.30 ± 0.05	1.11 ± 0.04	4.97 ± 0.58
E[Target MSE]	1.05 ± 0.01	0.97 ± 0.01	0.80 ± 0.01	3.16 ± 0.26

Table 4.8: Optimization statistics on prostate case 2.

	Beamlet-based worst-case	Beamlet-based expected-value	Beamlet-free expected-value	Beamlet-free + variance term
# iterations	181 ± 7	53 ± 2	937500	937500
Beamlet computation time [s]	17574 ± 34	17574 ± 34	0	0
Optimization time [s]	3927 ± 208	5942 ± 383	6153 ± 755	6305 ± 980
Total computation time [s]	21501 ± 242	23516 ± 417	6153 ± 755	6305 ± 980
Peak RAM [GB]	13.0 ± 0.1	13.1 ± 0.1	1.2 ± 0.1	1.2 ± 0.1
Objective function nominal	0.399 ± 0.013	0.366 ± 0.015	0.292 ± 0.010	1.357 ± 0.345
Objective function worst-case	0.533 ± 0.017	0.519 ± 0.011	0.463 ± 0.029	2.399 ± 0.576
E[Objective function]	0.444 ± 0.004	0.403 ± 0.002	0.339 ± 0.007	1.716 ± 0.413
Target D98 nominal [Gy]	58.11 ± 0.04	58.21 ± 0.04	58.40 ± 0.03	56.67 ± 0.37
Target D98 worst-case [Gy]	57.52 ± 0.09	57.53 ± 0.06	57.58 ± 0.11	54.59 ± 0.73
E[Target D98] [Gy]	58.05 ± 0.03	58.11 ± 0.02	58.23 ± 0.03	56.28 ± 0.42
Target HI nominal	3.05 ± 0.05	2.92 ± 0.07	2.61 ± 0.06	5.54 ± 0.73
Target HI worst-case	3.46 ± 0.01	3.41 ± 0.02	3.11 ± 0.07	6.84 ± 0.80
E[Target HI]	3.19 ± 0.01	2.92 ± 0.07	2.76 ± 0.02	5.93 ± 0.69
Target CI nominal	0.370 ± 0.004	0.374 ± 0.001	0.418 ± 0.005	0.417 ± 0.014
Target CI worst-case	0.268 ± 0.002	0.268 ± 0.001	0.384 ± 0.006	0.389 ± 0.015
E[Target CI]	0.394 ± 0.004	0.398 ± 0.001	0.432 ± 0.006	0.414 ± 0.017
Target MSE nominal	0.86 ± 0.03	0.79 ± 0.04	0.63 ± 0.02	2.90 ± 0.73
Target MSE worst-case	1.15 ± 0.01	1.12 ± 0.02	1.00 ± 0.06	5.17 ± 1.24
E[Target MSE]	0.95 ± 0.1	0.87 ± 0.01	0.73 ± 0.01	3.66 ± 0.87

4.3.5 Head and neck case 1

The results for head-and-neck case 1 are summarized in Table 4.9. The target metrics refer to the primary tumour (CTVp7000). The corresponding dose volume histograms can be seen in Figure 4.6. Overall, all DVH show good agreement with each other.
The $D98$ agrees with the target prescription within 5% for all scenarios and all methods.
The objective function for the worst-case scenario agrees within 3σ for all methods, but other metrics reveal differences between the methods. Individual metrics, such as the worst-case scenario $D98$ and MSE were significantly worse for the beamlet-free method and the target CI was lower for the nominal, worst-case and average scenario, while others such as the target HI and the nominal and average scenario objective function values were comparable or better than the beamlet-based worst-case results, but not as good as the beamlet-based expected value ones. Overall this case showed the biggest discrepancies between which methods achieved better plan quality for different metrics which might be explained through the presents of several target regions with different prescriptions, all of which were optimized robustly.
Compared to the beamlet-based worst-case approach, the beamlet-free method managed to reduce computation time by 93% and peak memory usage by 86%.
The results with and without additional variance term are similar to each other.

4.3.6 Head and neck case 2

While the two beamlet-based methods achieved decent plan quality, outliers emerged when using the beamlet-free method resulting in several scenarios performing significantly worse during evaluation and overall a non-robust plan. Using a variance term in the objective function, as described in section 4.2, significantly improved optimization results and delivered a robust plan. The target metrics in Table 4.10 refer to the primary tumor (CTVp7000). The corresponding dose volume histograms can be seen in Figure 4.7. Overall, the DVH show good agreement between methods.
The D98 agrees with the target prescription within 5% for all scenarios and all methods.
For the worst-case scenario, the beamlet-free and beamlet-based worst-case methods agree for all metrics within 3σ except for the target CI which was slightly lower for the beamlet-free approach. For the nominal and average scenario, the beamlet-free method was able to achieve significantly better objective function, target D98, HI and target MSE values, but lower target CI. The beamlet-based expected-value method showed good agreement with the beamlet-free results with the exception

Table 4.9: Optimization statistics on head and neck case 1.

	Beamlet-based worst-case	Beamlet-based expected-value	Beamlet-free expected-value	Beamlet-free + variance
# iterations	325 ± 28	79 ± 2	1875000	1875000
Beamlet computation time [s]	69281 ± 279	69281 ± 279	0	0
Optimization time [s]	7543 ± 734	12510 ± 133	5333 ± 1397	6900 ± 377
Total computation time [s]	76824 ± 1013	81791 ± 412	5333 ± 1397	5333 ± 1397
Peak RAM [GB]	12.9 ± 1.8	14.0 ± 0.1	1.8 ± 0.1	1.8 ± 0.1
Objective function nominal	0.660 ± 0.015	0.538 ± 0.007	0.564 ± 0.008	0.546 ± 0.014
Objective function worst case	0.910 ± 0.015	0.946 ± 0.015	0.954 ± 0.006	0.961 ± 0.073
E[Objective function]	0.766 ± 0.003	0.645 ± 0.006	0.705 ± 0.007	0.686 ± 0.024
Target D98 nominal [Gy]	67.84 ± 0.05	67.87 ± 0.06	67.78 ± 0.03	67.85 ± 0.02
Target D98 worst-case [Gy]	67.36 ± 0.05	67.20 ± 0.06	66.99 ± 0.09	67.04 ± 0.25
E[Target D98] [Gy]	67.70 ± 0.04	67.76 ± 0.03	67.61 ± 0.03	67.66 ± 0.07
Target HI nominal	3.55 ± 0.10	3.40 ± 0.04	3.54 ± 0.05	3.44 ± 0.06
Target HI worst case	3.94 ± 0.06	3.48 ± 0.28	3.98 ± 0.02	3.97 ± 0.08
E[Target HI]	3.67 ± 0.01	3.48 ± 0.01	3.70 ± 0.01	3.65 ± 0.03
Target CI nominal	0.286 ± 0.001	0.289 ± 0.001	0.269 ± 0.003	0.274 ± 0.004
Target CI worst-case	0.241 ± 0.001	0.246 ± 0.001	0.221 ± 0.002	0.241 ± 0.004
E[Target CI]	0.295 ± 0.002	0.298 ± 0.004	0.279 ± 0.002	0.281 ± 0.004
Target MSE nominal	1.17 ± 0.05	1.08 ± 0.03	1.17 ± 0.03	1.08 ± 0.03
Target MSE worst-case	1.46 ± 0.03	1.46 ± 0.05	1.64 ± 0.02	1.60 ± 0.12
E[Target MSE]	1.26 ± 0.01	1.14 ± 0.01	1.32 ± 0.02	1.26 ± 0.03

of the target CI and the MSE of the worst-case scenario, where it managed to achieve better results. The results obtained with the beamlet-free method showed overall a higher variance between runs compared to the other patient cases. While all runs achieved good plan quality and robustness, the exact benefit with respect to computation time may vary, if a different termination condition were to be chosen.
Compared to the beamlet-based worst-case method, the beamlet-free approach reduced computation time by 74% and peak memory usage by 90%.

4.4 Discussion

The feasibility of a beamlet-free robust optimization approach is demonstrated and its computational requirements are compared to conventional robust optimization methods.
The beamlet-based methods optimize until a predefined objective function value is reached. For the sake of fairness, the iterations for the beamlet-free method are manually adjusted until the highest scenario objective function value after evaluation is lower or equal to those achieved with beamlet-based worst-case optimization. Overall plan quality is verified by comparing different metrics, including objective function values, target D98, target HI, target CI, and target MSE for the nominal, worst and average scenarios out of 81 investigated during plan evaluation. While the performance of the individual methods varies between metrics and patients, all methods can achieve robust plans without any method performing considerably better or worse than the others. DVH bands are very cohesive across all methods. The D98 matches the prescription within 5%, and is therefore considered to be met, for all scenarios and methods, except some scenarios of the brain cases. For the brain cases, none of the methods were able to achieve agreement with the prescription within 5% for all scenarios due to the challenging nature of the cases. The pass rate varied between runs, but never fell below 97% for any method.
A frequent worry is that expected-value robust optimization methods are theoretically less stringent with outliers and the worst-case scenario could differ significantly for comparable quality of the nominal scenario. For this reason, the objective function of the worst-case scenario is used as termination criterion. The beamlet-based expected value method, for most metrics and cases, achieves equal or better results for the nominal and average scenario compared to the worst-case scenario approach, but also required longer computation time. The beamlet-free method performed similarly, except for the head-and-neck case 2, where the method struggles to achieve robustness at all. This is obviously a significant issue

Table 4.10: Optimization statistics on head and neck case 2.

	Beamlet-based worst-case	Beamlet-based expected-value	Beamlet-free expected-value	Beamlet-free + variance
# iterations	295 ± 13	69 ± 2	3125000	3125000
Beamlet computation time [s]	79463 ± 863	79463 ± 863	0	0
Optimization time [s]	11198 ± 606	18380 ± 635	13190 ± 2326	14097 ± 1511
Total computation time [s]	90661 ± 1469	97843 ± 1498	13190 ± 2326	14097 ± 1511
Peak RAM [GB]	21.8 ± 0.1	22.0 ± 0.1	2.1 ± 0.1	2.1 ± 0.1
Objective function nominal	0.763 ± 0.012	0.543 ± 0.009	0.702 ± 0.016	0.526 ± 0.005
Objective function worst-case	1.071 ± 0.012	0.987 ± 0.008	2.194 ± 0.070	0.905 ± 0.008
E[Objective function]	0.938 ± 0.005	0.723 ± 0.006	1.122 ± 0.009	0.655 ± 0.006
Target D98 nominal [Gy]	67.59 ± 0.02	67.82 ± 0.02	67.59 ± 0.04	67.84 ± 0.06
Target D98 worst-case [Gy]	66.86 ± 0.04	66.68 ± 0.03	64.70 ± 0.10	66.62 ± 0.06
E[Target D98] [Gy]	67.33 ± 0.01	67.52 ± 0.03	67.14 ± 0.03	67.57 ± 0.01
Target HI nominal	4.04 ± 0.04	3.49 ± 0.02	3.95 ± 0.03	3.52 ± 0.02
Target HI worst case	4.49 ± 0.04	4.31 ± 0.05	5.13 ± 0.05	4.06 ± 0.06
E[Target HI]	4.11 ± 0.04	3.76 ± 0.01	3.95 ± 0.01	3.67 ± 0.01
Target CI nominal	0.400 ± 0.001	0.404 ± 0.001	0.368 ± 0.001	0.376 ± 0.002
Target CI worst-case	0.342 ± 0.001	0.350 ± 0.001	0.310 ± 0.001	0.337 ± 0.002
E[Target CI]	0.407 ± 0.001	0.413 ± 0.001	0.373 ± 0.001	0.385 ± 0.001
Target MSE nominal	1.51 ± 0.02	1.14 ± 0.01	1.44 ± 0.03	1.14 ± 0.01
Target MSE worst-case	1.91 ± 0.04	1.85 ± 0.04	5.02 ± 0.15	1.85 ± 0.04
E[Target MSE]	1.63 ± 0.02	1.34 ± 0.01	2.14 ± 0.01	1.30 ± 0.01

for a robust optimization method. Introducing a variance term into the objective function that reduces the variance between scenarios, enables the beamlet-free optimizer to achieve plan robustness for this patient case as well. The modified objective function with variance does lead to worse optimization results for two of the other patient cases. A dynamic allocation of the variance term might be possible based on the properties of the distribution of scenario objective functions, like its kurtosis, but further research on large numbers of patient cases is necessary to make definitive claims. Otherwise patient-specific parameter selection is required.

The use of beamlet-free robust optimization could reduce computation times by 63% to 93% and peak memory by 86% to 92%.

The memory consumption for the beamlet-based methods depends primarily on the number of voxels, spots, and scenarios. A beamlet matrix needs to be computed for each scenario. During beamlet calculation, these beamlet matrices are exported as files by the MC engine and stored on the disc. They are stored as sparse matrices to save memory and need to be loaded and kept in memory for the optimization. This leads to considerably higher memory consumption than non-robust optimization. Both beamlet-based methods have comparable memory requirements.

Memory consumption during beamlet-free optimization is primarily determined by the number of voxels. The number of spots is almost entirely irrelevant and impact by the number of scenarios can be largely avoided through an efficient implementation. The peak RAM usage is therefore considerably lower than what can be achieved with beamlet-based methods. Additionally, the beamlet-free method does not require any files to be stored on the disc, which is in itself a great advantage over the beamlet-based methods.

A memory benefit can be achieved for any plan size and increases for larger plans and higher numbers of scenarios.

During beamlet calculation, the number of particles, and consequentially also the computation time increases linearly with the number of spots and the number of scenarios. Additionally, the time required for matrix conversions and operations increase with the number of scenarios, voxels, and spots and can take a considerable amount of time. For all patient cases, the beamlet computation was the aspect that required the most time. For worst-case beamlet-based optimization, an objective function evaluation needs to be performed at every step, which increases the computation time in relation to the number of scenarios compared to non-robust optimization. The gradient computation time only increases by a factor of two maximum, for the nominal and the worst-case scenario, and usually less than that since objectives are only evaluated for one of the two. For expected-value beamlet-based optimization, the time per iteration is longer, since

the gradient needs to be computed for each scenario. However, the total number of iterations could be reduced while maintaining comparable plan quality.
For the expected-value beamlet-free method, the number of protons, and therefore the computation time, increases with the number of scenarios. However, there is no direct causal relationship between the number of particles and the number of spots. For bigger targets, a higher number of protons is required to reduce statistical noise and ensure sufficient coverage. A finer spot grid resolution or larger spot placement margin, however, have little impact on particle number or computation time. For the gradient computation, the expected value over all scenario gradients is taken, which reduces noise in the gradient compared to non-robust optimization.
Especially for big plans with many spots, a huge benefit could be achieved by using beamlet-free robust optimization. These benefits would even grow if the treatment planning were performed on a simple computer instead of a many-core computation server, where the beamlet computations would be considerably slowed down. It should also be noted that the plan parameters, including the number of spots and voxels, had to be kept sufficiently low for the beamlet-based methods to be able to compute. In all patient cases, voxel sizes were increased compared to the original CT data to reduce the size of the beamlet matrix. While the beamlet-free method also benefited from the increase in voxel size, since it speeds up the MC simulation, it was to a far lesser extent and not doing so would not have impeded the ability to calculate the plans. The number of spots has no impact on the beamlet-free method beyond the need for sufficient particles to ensure low statistical noise, which is usually achieved by the required number of iterations anyway. This means that the beamlet-free method could have likely produced plans of significantly better quality, for no or little increase in computational cost, had it not been limited by the parameters available for beamlet-based methods for the sake of comparability.
It should be mentioned that the computation times depend on various factors, including the used hardware, treatment planning system, dose calculation engine, and optimizer. The general trends should, however, be observable for different implementations as well, since they are caused by the dependence of the method on the number of spots and scenarios.
The reduced computational requirements make beamlet-free robust optimization an interesting candidate in the exploration of novel, complex treatment techniques, including proton arc therapy and adaptive proton therapy. Although they remain still too large to facilitate processes such as real-time robust plan adaptation, they are contributing to the breaking-down of hard computational limits and they open up new pathways for exploration.

4.5 Conclusion

Comparable plan quality could be achieved by all three methods, beamlet-based worst-case, beamlet-based expected-value and beamlet-free expected-value robust optimization. Computation time and memory usage between the beamlet-based worst-case scenario and expected-value methods were comparable, making expected-value robust optimization a viable candidate for clinical applications.
The expected-value beamlet-free implementation managed to achieve the same levels of plan quality while reducing computation time by $51 - 69\%$ and peak memory usage by $87 - 92\%$ of what could be achieved with beamlet-based methods, depending on plan size. Beamlet-free robust optimization enables the use of more error scenarios, finer voxel resolutions and more spots at little or no increase of computational burden. This makes it capable of delivering higher quality treatment plans and makes it a promising candidate for novel treatment strategies and techniques.

4.6 Additional Materials

The total number of simulated protons in each approach is compared in Table 4.11.
For a full picture, the comparison of robust optimization methods should consider the convergence of the objective function. To this end, convergence plots for the three methods, beamlet-based worst-case, beamlet-based expected-value and beamlet-free expected-value robust optimization are included.

Table 4.11: Total number of protons simulated (for the beamlet-based approach this includes only the beamlet calculation, no final dose calculation).

Number of spots	Beamlet-based	Beamlet-free
Brain 1	1841.805×10^7	84×10^7
Brain 2	1557.36×10^7	84×10^7
Prostate 1	988.68×10^7	42×10^7
Prostate 2	925.26×10^7	31.5×10^7
Head and Neck 1	3786.72×10^7	63×10^7
Head and Neck 2	4341.4350×10^7	105×10^7

4.6.1 Monte Carlo noise contributions during objective function evaluation

Before discussing the convergence plots, it is worth to consider the impact of the MC noise on the convergence.
For the two beamlet-based methods, the reported objective function is calculated during optimization based on the beamlet dose distributions in the dose-influence matrix. It is reported after every iteration. They are therefore affected by the same amount of MC noise stemming from the beamlet dose calculation rendering them directly comparable. For the beamlet-free method, the objective function is calculated based on the total dose distributed as simulated at a given time. It is evaluated every 100.000 iterations. This ensures that the computation time is not increased dramatically, as the objective function requires an evaluation over the entire normalized dose distribution, which is more costly to compute than the gradient approximation used during optimization.
The dose distribution during the beamlet-free optimization is subjected to MC noise. The correlation between the number of sampled random variables N and the stochastic noise σ is

$$\sigma \propto \frac{1}{\sqrt{N}} \ , \tag{4.14}$$

which can be assumed to reflect the noise on the dose distribution σ_d in relation to the number of simulated primaries N_p. Since the number of simulated particles increases as the optimization progresses, the noise σ_d decreases. This impacts the objective function, however, the noise on the dose σ_d is not equivalent to the resulting noise on the objective function value σ_f. The impact of the noise on the objective function depends on the plan and objectives used, but it can generally be assumed that

$$\sigma_f \propto \frac{1}{N} \ , \tag{4.15}$$

since the dose terms in the objectives utilized for this comparison are squared. This means that the objective function f_{N_p} evaluated over a dose distribution $\mathbf{d}_{N_p}$ simulated with N_p primaries takes the form

$$f_{N_p}(\mathbf{d}_{N_p}) = a\frac{1}{N} + \hat{f} \ , \tag{4.16}$$

where $\hat{f}$ is the true objective function without noise contributions and a is a factor depending on the plan and objectives used. To estimate the impact of the MC noise on the objective function values reported during beamlet-free optimization, a optimized set of spot weights $\mathbf{x}'$ was used and dose distributions with increasing numbers of particles calculated based on it. Since the spot weights remain constant, the observed difference in objective function results purely from the MC

noise. The calculated objective function values were plotted against the number of primaries used in their calculation as represented in Figure 4.8. A function of the form presented in equation (4.16) was fitted against the simulated data and used to determine the factor a and the unbiased objective function $b = \hat{f}$ for each of the patient cases.
The optimized spot weights $\mathbf{x}'$ are obtained by performing a beamlet-based expected-value robust optimization. A forward dose calculation is then performed based on the optimized spot weights. Only the nominal scenario is computed, since the impact of the MC noise can be assumed to be similar across different scenarios. Before the objective function is evaluated, the dose distribution is normalized with normalization factor $n = \frac{\hat{d}_{\mathsf{TV}}}{d_{\mathsf{TV}}^{\mathsf{prescribed}}}$ with the mean target dose $\hat{d}_{\mathsf{TV}}$ and prescribed target dose $d_{\mathsf{TV}}^{\mathsf{prescribed}}$ to be consistent with the dose distribution during beamlet-free optimization.
It should be noted that the noise contribution to the objective function values is very consistent for a specific number of particles. They do not represent an uncertainty on the precision of the objective function calculation, but rather an additional contribution due to the uneven nature of the dose distribution.

The plots in Figure 4.8 show excellent agreement between the fitted curve in red and the measured data in blue. The only exception is the brain case, where a slight discrepancy can be observed for higher N_p. While the difference between fit and measurement is very small compared to the overall impact of MC noise as witnessed for low N_p, the fitted variables are affected by uncertainty. The fit variables are presented in Table 4.12 with their respective fit uncertainties.

Table 4.12: Fit parameters for the estimation of MC noise contributions on objective function values.

	a	b
Brain 1	5.27×10^6	0.110
Brain 2	2.32×10^6	0.420
Prostate 1	6.97×10^6	0.134
Prostate 2	6.88×10^6	0.123
Head and Neck	11.34×10^6	0.333
Head and Neck 2	11.64×10^6	0.508

4.6.2 Convergence plots

Utilizing the understanding of the MC noise impact on the objective function derived in the section above, the objective function obtained during beamlet-free

optimization can be split into a contribution of the "true" objective function for the current vector of weights **x** and a MC noise contribution. Both converge independently, the "true" objective function contribution towards the optimal solution $f(\mathbf{x}^*)$ and the MC noise contribution against zero, as the optimization progresses. A noise correction was not performed for the beamlet-based methods even though a slight discrepancy could be observed between the objective function during optimization and the one reported afterwards on a recalculated dose distribution with high numbers of particles. While this indicates that the number of protons during beamlet calculation was sufficient to eliminate the effects of MC noise completely, the difference is small and the benefit would be severely limited, especially when considering that the noise corrected objective function is subject to some uncertainty stemming from the fit quality. The impact of the MC noise on the dose-influence matrix is also more difficult to estimate, since the same number of protons was used to simulate each beamlet while the protons per spot in the forward dose calculation used in the evaluation are proportional to the spot weights. The convergence plots for all three robust optimization methods and all patient cases are compared in Figure 4.9, 4.10 and 4.11. The two beamlet-based methods are compared in the panel on the left with worst-case scenario robust optimization in blue and expected-value robust optimization in orange. The panel on the right shows the convergence of the beamlet-free method with and without MC noise correction. The blue line is the uncorrected objective function and the orange line is the objective function with the estimated MC noise contribution subtracted to present a more accurate picture of convergence.
Since the same objective function is used across all methods, they can be directly compared. The beamlet-based worst-case scenario and beamle-based expected value methods were terminated once a specific objective function value was reached. An equal objective function value indicates equal plan quality from an optimization point of view. This does not neccessarily translate to equal plan quality in a clinical sense, but provides a good measure to compare optimizers. As can be seen in the figures, the expected-value methods decreased more quickly and required fewer iterations to achieve the same final objective function value. Since the optimization was terminated based on a specific objective function value, not a convergence criterium, better solutions might have been reached if the optimization had been allowed to continue. The objective function value was used instead because the focus lies on the computational efficiency to reach an acceptable solution, not on the best possible plan quality. To first determine a suitable objective function value, worst-case scenario robust optimization was used with a convergence criterium. Several runs were observed to estimate when convergence is reached with this method. The reported worst-case robust optimization, as well as the expected-value robust optimization were terminated once the objective

function had fallen below this value. Since expected-value robust optimization has potentially a wider spread between the individual scenario objective functions, the scenario objective function of the worst performing scenario was chosen.
It should be noted that even the noise corrected objective function for the beamlet-free, while identical in its definition, does not necessarily converge to the same value as the beamlet-based methods. The reason for this is the normalization performed during beamlet-free optimization. The dose distribution is constantly normalized such that the mean target dose is equivalent to the prescribed target dose. This presents an additional constraint on the optimization and might cause it to not always be able to reach the solution that the beamlet-based methods find. However, after optimization, all plans are usually normalized in such a way that the D_{50}, D_{95} or D_{98} match the prescription dose. The objective function evaluated across these normalized plans was compared to ensure that the beamlet-free method achieved comparable plan quality.
The objective function for the beamlet-free function is first reported after 200.000 iterations at which point it is already more optimized than the beamlet-based counterpart at the very beginning. The reason for this is that the objective function is very high in the beamlet-free case as it starts from an empty dose distribution. It can clearly be seen that the noise corrected curve starts to flatten towards the end of the simulation while the uncorrected objective function keeps decreasing due to the reduced MC noise. However, even the noise corrected curve has not reached a plateau yet when simulation is terminated indicating that better results can be obtained with longer run times.

Time contributions

Figure 4.12 shows how much time was required for the different treatment planning steps by the beamlet-free and beamlet-based method respectively. Only the beamlet dose calculation and optimization steps are included. It can be noted that the time required for the beamlet calculation and optimization steps both increase with rising number of spots.

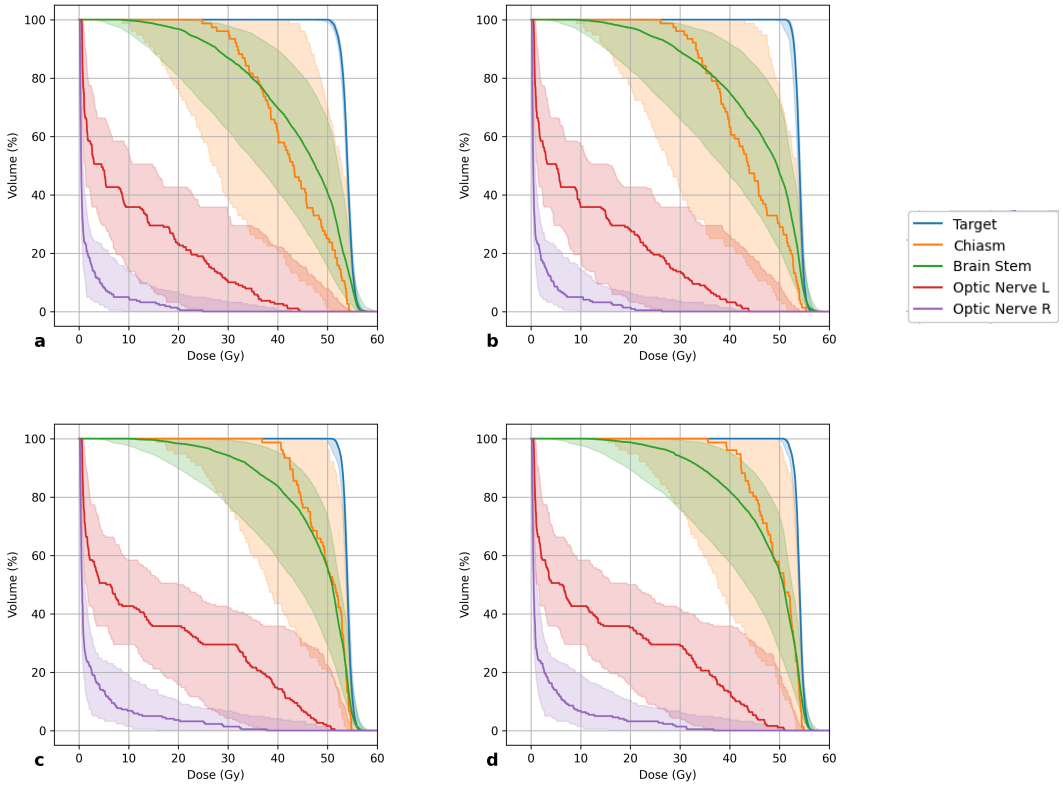


Figure 4.3: Dose volume histograms (DVH) for brain case 2 with the a) beamlet-based worst case, b) beamlet-based expected-value, c) beamlet-free expected value approach and d) beamlet-free expected value approach with additional variance term.

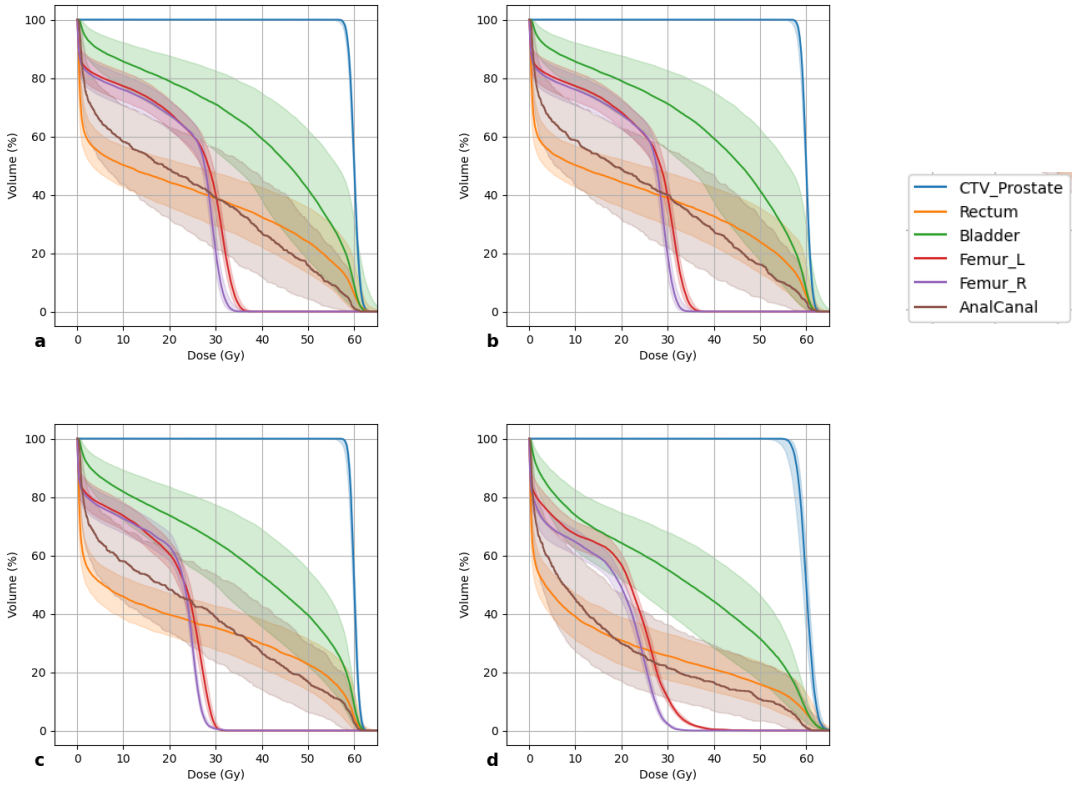


Figure 4.4: Dose volume histograms (DVH) for prostate case 1 with the a) beamlet-based worst case, b) beamlet-based expected-value, c) beamlet-free expected value approach and d) beamlet-free expected value approach with additional variance term.

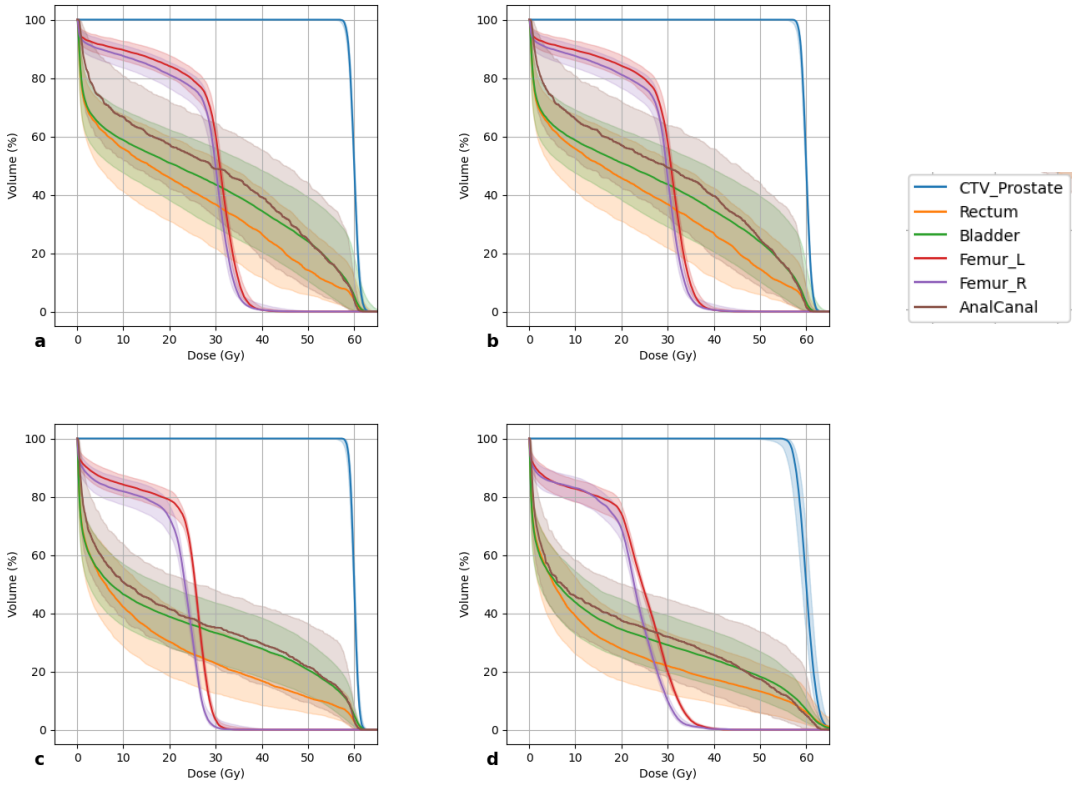


Figure 4.5: Dose volume histograms (DVH) for prostate case 2 with the a) beamlet-based worst case, b) beamlet-based expected-value, c) beamlet-free expected value approach and d) beamlet-free expected value approach with additional variance term.

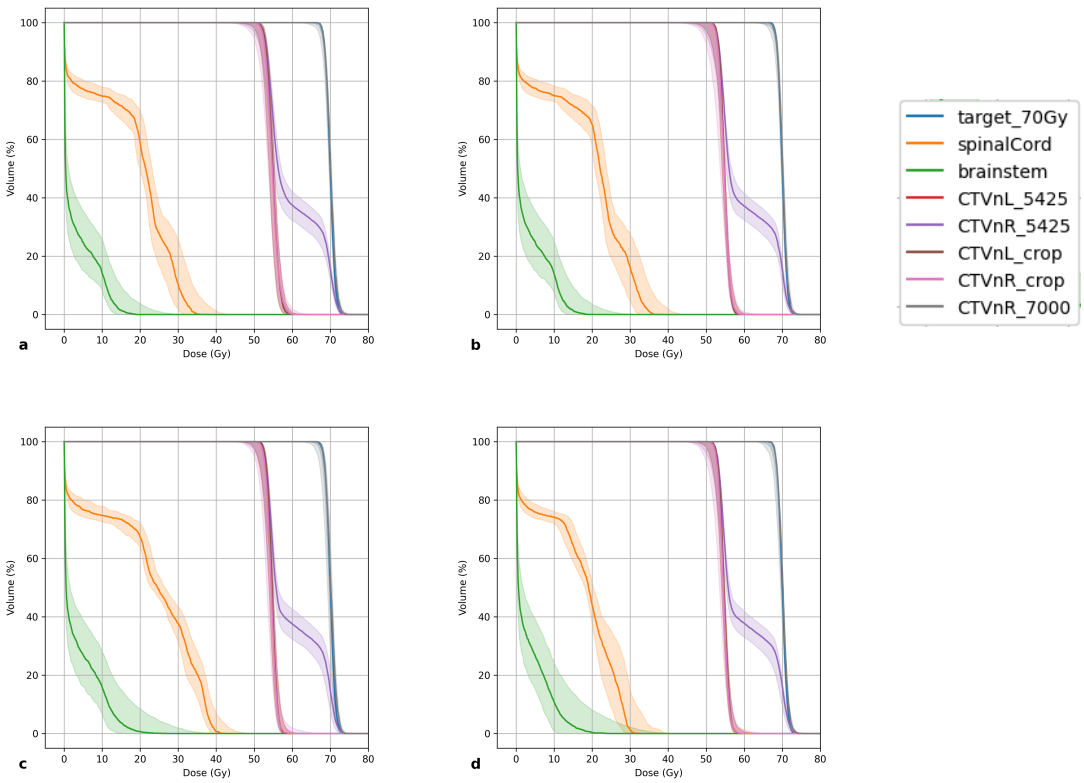


Figure 4.6: Dose volume histograms (DVH) for head and neck case 1 with the a) beamlet-based worst case, b) beamlet-based expected-value, c) beamlet-free expected value approach and d) beamlet-free expected value approach with additional variance term.

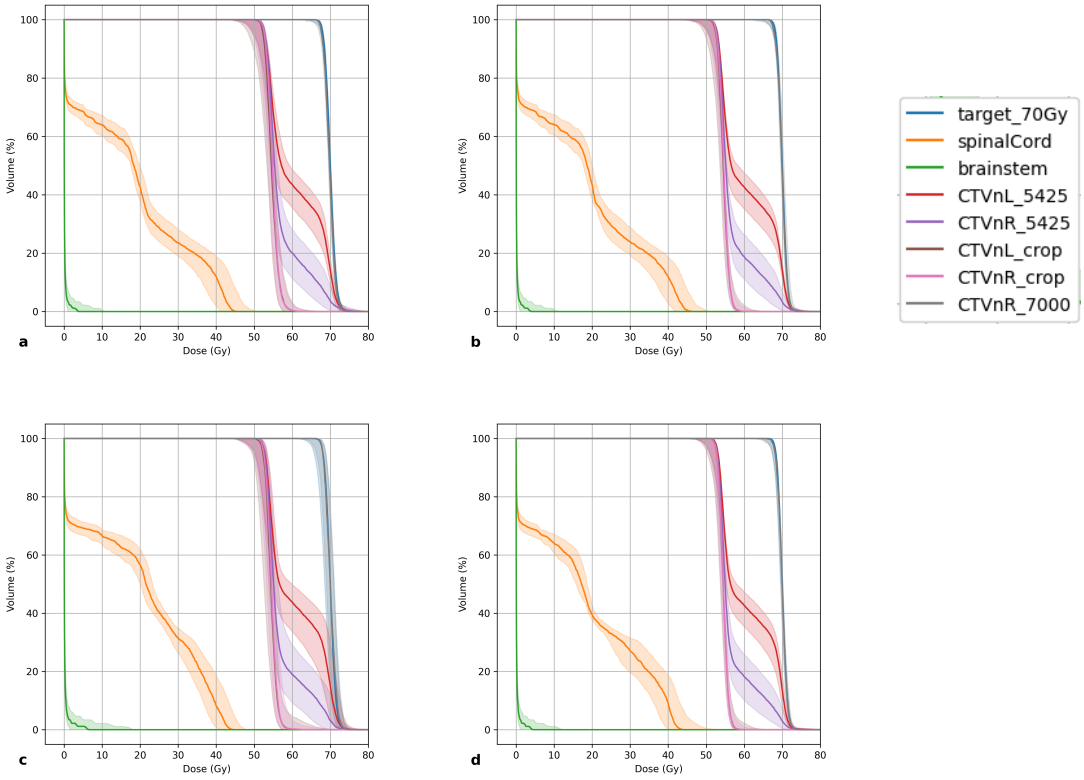
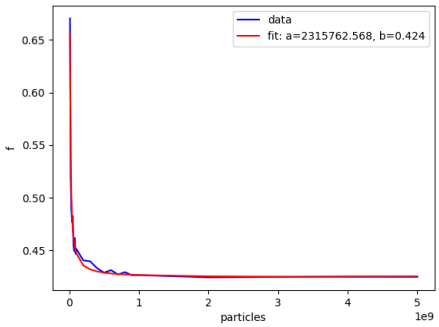
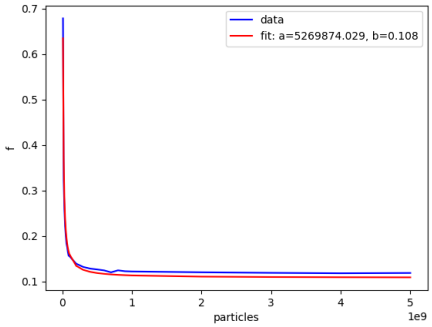


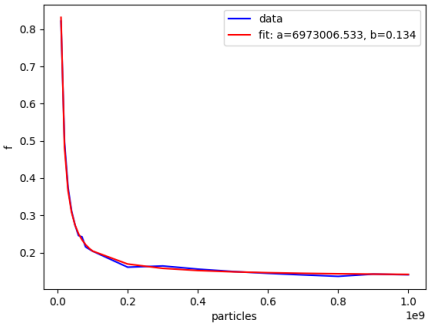
Figure 4.7: Dose volume histograms (DVH) for head and neck case 2 with the a) beamlet-based worst case, b) beamlet-based expected-value, c) beamlet-free expected value approach with variance term and d) beamlet-free expected value approach with additional variance term.



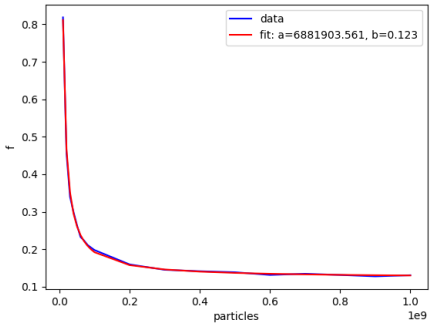
(a) Brain case 1



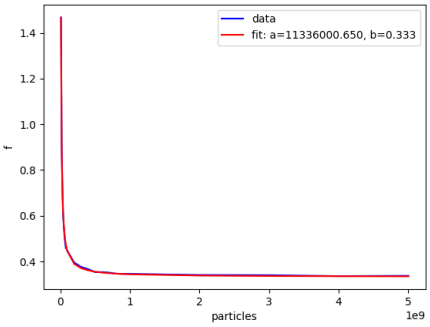
(b) Brain case 2



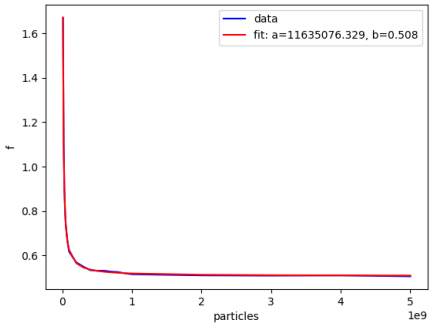
(c) Prostate case 1



(d) Prostate case 2

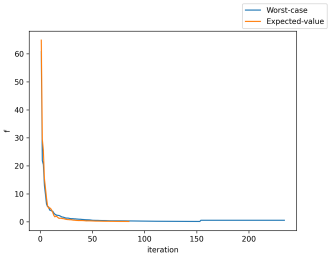


(e) Head and neck case 1

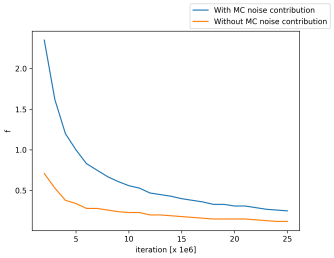


(f) Head and neck case 2

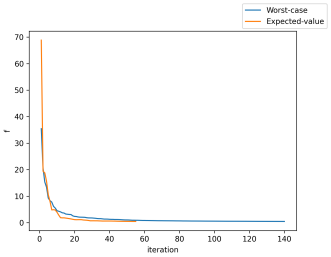
Figure 4.8: Impact of MC noise during dose calculation on the calculated objective function value. Data from a series of dose calculations and subsequent objective function evaluations (blue) is used as basis for a fit of function $f(N_p) = a\frac{1}{N_p} + b$ with fit variables a and b (red).



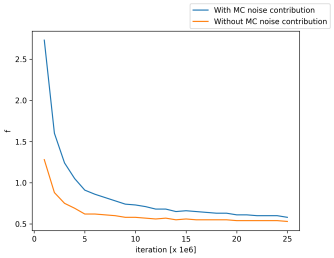
(a) Brain 1: Beamlet-based methods



(b) Brain 1: Beamlet-free method

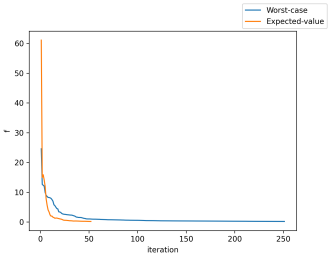


(c) Brain 2: Beamlet-based methods

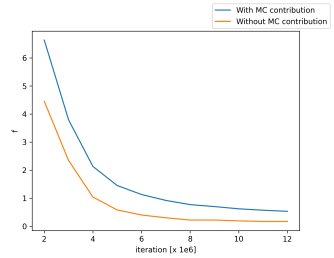


(d) Brain 2: Beamlet-free method

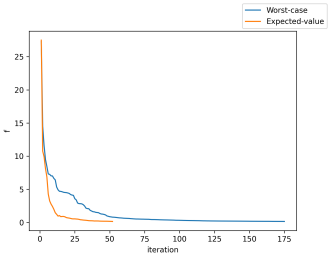
Figure 4.9: Convergence plots for the two brain cases cases.



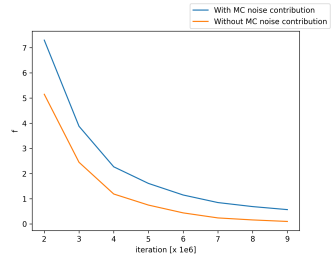
(a) Prostate 1: Beamlet-based methods



(b) Prostate 1: Beamlet-free method

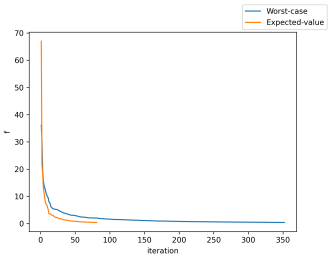


(c) Prostate 2: Beamlet-based methods

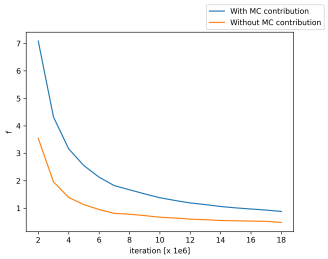


(d) Prostate 2: Beamlet-free method

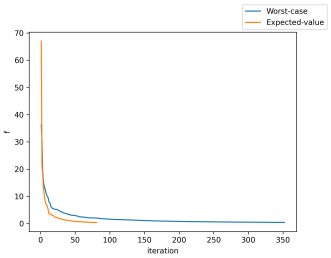
Figure 4.10: Convergence plots for the two prostate cases.



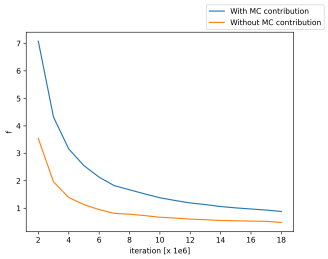
(a) Head and neck 1: Beamlet-based methods



(b) Head and neck 1: Beamlet-free methods



(c) Head and neck 2: Beamlet-based methods



(d) Head and neck 2: Beamlet-free methods

Figure 4.11: Convergence plots for the two head and neck cases.

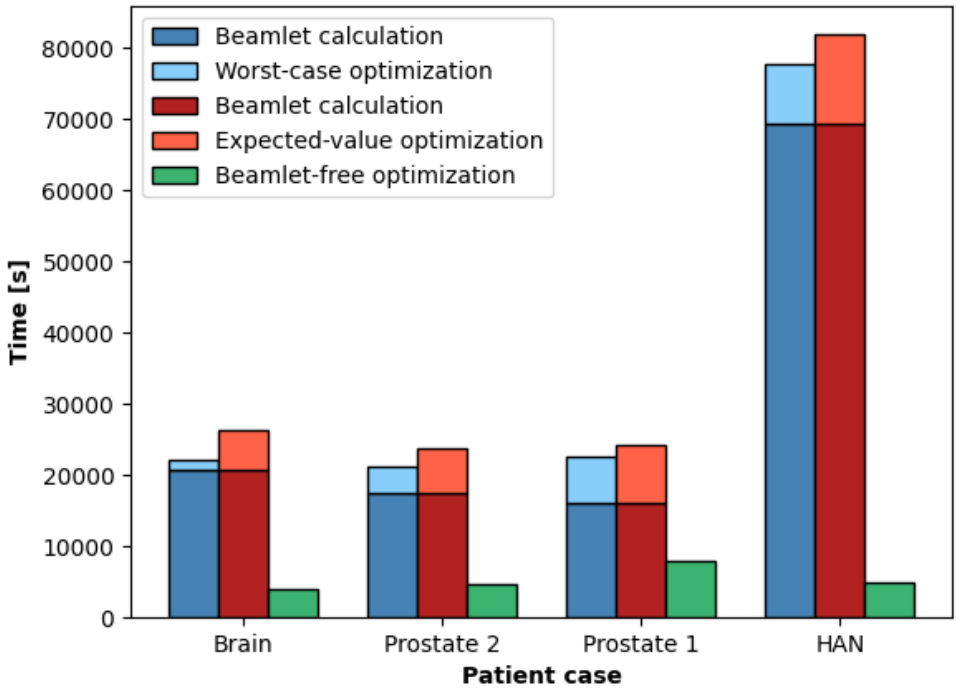


Figure 4.12: Time contributions of the individual treatment planning processes required for the beamlet-based worst-case optimization method (left bar, blue), the beamlet-based expected-value optimization (middle bar, red) and the beamlet-free method (right bar, green) for varying patients and numbers of spots.

Chapter 5

Discussion, conclusion and outlook

Proton therapy requires precise and accurate dose calculation and optimization tools in order to profit from the dosimetric benefits it promises. Those accurate treatment planning tools have high computational costs that impede efficient clinical workflows and put severe limitations on novel treatment techniques. In particular, the calculation, storage and use of the large scale dose influence matrix poses an issue regarding overall computation time and memory usage. This thesis seeks to address these challenges through the following contributions

- A beamlet-free treatment planning algorithm is proposed that greatly reduces computation time and memory usage for large IMPT plans. The novel methods proposes a deviation from the conventional IMPT treatment planning workflow that allows treatment planning without the generation of a dose-influence matrix. The beamlet-free optimizer is validated on several patient cases against conventional methods. The reduction of both computation time and memory usage that the beamlet-free optimizer achieves depends on the plan size and is bigger for larger plans.

- A beamlet-free robust optimizer is implemented and compared against conventional robust optimization methods. This expected-value based beamlet-free robust optimizer managed to reduce both computation time and memory usage compared to its beamlet-based counterparts and achieved bigger computational benefits than the non-robust beamlet-free algorithm by leveraging the noise reduction in the expected-value gradient.

The implications of this work and the perspectives it offers will be discussed in the following paragraphs.

The need for Monte Carlo based treatment planning in proton therapy

The entire premise of this thesis is based on the assumption that MC-based treatment planning is necessary, or at least highly preferential, for proton therapy treatment planning. If analytical dose calculation algorithms achieve an accuracy that is sufficient for even highly heterogeneous geometries, most of the issues this work seeks to solve suddenly evaporate into thin air, as analytical dose calculation engines require a fraction of the time that MC methods do.
The current state of research still finds analytical dose calculation tools unable to achieve the accuracy of MC methods in heterogeneous geometries [134, 171, 71, 72, 188, 73, 189, 190, 74, 191]. However, it could be argued wether this accuracy is truly required in all steps of the treatment planning workflow. Clinical treatment planning systems often employ analytical models in the beamlet calculation and perform only the final dose calculation with MC to verify the quality of optimiza-

tion. A reduced accuracy in the beamlet calculation might still be sufficient to guide optimization to an acceptable solution. However, recent studies revealed that plans optimized on analytically calculated dose data and then recalculated with MC demonstrate lower plan quality and reduced robustness compared to those optimized on the basis of MC dose data [75, 192]. This indicates that there is a benefit to fully MC-based treatment planning, at least for treatment sites with high levels of tissue heterogeneity such as the lung or head and neck region.
On another side, new dose calculation tools could emerge to rival MC in the future. New AI-based dose prediction tools have recently become available for photons [193, 194, 195, 196, 197, 198, 199, 200] and protons [201, 202], although their suggested use is so far mainly limited to initial assessments to aid the selection of treatment modalities. Just recently, several fast AI-based dose calculation models intended for use in real-time adaptive proton therapy have been developed [95, 203]. These models are able to calculate dose distributions in just a few seconds while achieving high accuracy and conformity with MC methods even in heterogeneous geometries. AI dose calculation models might be restricted to or biased against specific treatment sites, or input data formats depending on how they were trained, though those issues should be remedied by thorough training and testing of models. Additionally, there are some concerns with regards to quality assurance (QA), because AI models are blackbox algorithms. If and how AI methods can safely be included in clinical workflows is an ongoing discussion in the medical physics field. For other applications like segmentation algorithms, it is usually considered necessary to have a doctor check and if necessary correct the output of AI autosegmentation models. For dose calculation models, such a manual check is not practical, because dose distributions are not intuitive to the human eye. Thorough testing can achieve a high level of trust in specific models, but if their added benefit with respect to computation speed is worth the risk of some uncertainty is ultimately an ethics question.
Although there are prospects for the future, at the current time, MC is still the unrivaled master of accuracy, and even if it might not be necessary in all cases, it is necessary in a sufficient number of them to motivate further research in fast MC dose calculation tools and MC-based treatment planning workflows like they were assumed in this work.
Even if analytical or AI-based dose calculation tools were found to be suitable to replace MC beamlet calculation tools, rendering a concurrent approach unnecessary, there would still be high memory requirements resulting from the size of the dose-influence matrix. Consequently, there would still be a case to be made for a beamlet-free treatment planning approach, as the method not only reduced computation time, but also memory usage. Furthermore, the results of the robust optimization framework presented in Chapter 4 indicate that, for conventional

beamlet-based treatment planning, the time required for the optimization step alone is often of a similar magnitude to the total time required by the beamlet-free implementation. Therefore, the merit of a beamlet-free treatment planning approach is self-evident, independently of the necessity for a fully MC-based workflow.

Should the clinical treatment planning workflow be revised?

Since the proposed algorithm resulted in a significant reduction in computational resources, should the logical consequence be a revision of the clinical treatment planning workflow? To answer this question, an honest evaluation of the advantages and disadvantages of a revised beamlet-free treatment planning workflow need to be examined.
While the main advantages are obvious, the significant reduction of computation time and memory requirements, it should additionally be mentioned that the current implementation of the beamlet-free framework is quite hardware independent. The comparisons presented in this thesis were performed on many core CPU clusters against a workflow based on the CPU-parallelized MC code MCsquare [12]. The beamlet-free implementation is based on CPU-parallelized code, but only utilizes this parallelization to a low degree. This results in a similar performance independently of the number of cores. Had the comparisons been performed on a machine with fewer CPU cores, the relative benefit of the beamlet-free algorithm with respect to time would have been much bigger. A hospital could gain considerable flexibility with respect to the computers it uses for treatment planning.
But there are a row of drawbacks to the beamlet-free workflow as well. One of its main drawbacks concerns reoptimizations. It may occur that after the treatment plan optimization, the evaluations reveal that the resulting plan is insufficient and can not be used to treat the patient. In the conventional workflow, clinicians can adjust the objective types, weights and dose limits and reoptimize without needing to recompute the dose-influence matrix. How high the impact of the beamlet computation is on the total computation time varies depending on several factors, but usually 50% or more of the total computation time can be saved by not having to recompute it. This means that the relative benefit of the beamlet-free method is reduced in the case of reoptimization without beamlet calculation or in a MCO setting. For the non-robust implementation, especially for smaller plans, this may even result in longer overall treatment planning time. On the other hand, for the robust optimization, which is generally advised in proton therapy, the total computation time of the beamlet-free method is still lower than just the optimization part in the beamlet-based workflow. In these cases the beamlet-free method is still overall beneficial, even in the case of several reoptimizations.

Further considerations to take into account before attempting a revision of clinical workflows are the impact of the TPS on the relative computational benefit and the available objectives types which will be discussed below.

The choice of TPS and its impact on comparisons of beamlet-based and beamlet-free optimization strategies

While this indicates that a clinical application of the beamlet-free method would result in immediate benefits, all comparisons have been performed utilizing the experimental TPS OpenTPS [28] which is considerably less efficient than commercial TPS. The reasons for the utilization of OpenTPS are the accessibility of its source code and its easy adaptability, whereas the source code of commercial treatment planning software is not generally intended to be modified or perused by users. Access to the source code allows to ensure that the optimization performed by the TPS in the beamlet-based case corresponds exactly with the implementation in the beamlet-free case, which is imperative for a fair comparison of the two. However, this means that both methods are impacted by the efficiency of the TPS, but to different degrees. It is not trivial to predict how this would be reflected in a different TPS.
General TPS functionalities related to patient import, segmentation, plan creation and evaluation would affect both methods equally. Aspects of the TPS that could impact the relative computational benefit of a beamlet-free framework include

- the MC dose calculation engine,
- methods that reduce the plan complexity,
- the optimization algorithm,
- and the way data is handled, converted and transferred between tools.

There are several different fast MC codes available for proton therapy [12, 8, 5, 7, 9, 10, 11, 13, 14]. Clinical TPS often utilize their own in-house implementations which are often capable of reaching very competitive computation times. For example, Janson et al. [133] reported that the Raystation MC dose calculation engine often outperforms analytical algorithms. The efficiency of these fast MC codes is tweaked through paralellization [12, 8, 10, 11, 13], simplified physics models and approximations [5, 7, 6, 9, 14] and optimized implementations. Simplified physics models and approximations would benefit both the beamlet dose calculation in a beamlet-based approach and the beamlet-free method, as would an optimized implementation. While the computation time of MC dose engines can be significantly reduced through CPU- or GPU-parallelization, our beamlet-free

approach is only partially able to profit from parallelization. The reason for this are optimization variables that need to be shared between all threads. However, the comparisons presented in this work already compare a fast, CPU-parallelized MC code on a many-core computation server against the beamlet-free optimizer. Hypothetically, if hardware keeps becoming more accessible, fast MC codes could reduce their computation time further, while the beamlet-free method, at least in its current implementation, would hit a ceiling, but beyond that there is no reason to assume that utilizing a different MC dose calculation engine would significantly affect the relative benefit of the beamlet-free method.

Clinical TPS often improve their computational efficiency by reducing the problem complexity. This can for example be achieved by reducing the number of voxels [204] or using non uniform resolutions [205, 206]. While these attempts to reduce plan complexity have some impact on the beamlet-free algorithm through the speed of the MC simulation, it is very small compared to the achievable benefit for the beamlet-based approach where it would affect not just the MC simulation, but also the optimization. This benefit during the optimization is achieved through the reduced size of the dose-influence matrix, reducing both the memory usage and time required for matrix conversions and multiplications. On the flip side, the computation time and memory usage of the beamlet-free optimizer being nearly independent from the voxel resolution means that far finer voxel resolutions are an option in the first place and might allow better optimization quality at little computational cost. The comparisons presented in Chapter 4 were performed on a resampled voxel grid to reduce the problem size, since the size of the dose-influence matrix with the original CT resolution would have required too much memory. A finer scoring grid size might increase the impact of MC noise in early iterations, but also offers the chance of achieving better plan quality. This is due to the averaging effect that coarser voxel have on the dose distribution.

Different optimization algorithms are used in proton therapy optimization and they have different advantages and disadvantages. Some optimization algorithms have been proposed specifically due to their efficient use of computational resources [207, 208]. Other optimization methods try to achieve shorter computation times by adapting the optimization problem or by employing parallelization [17, 180, 18]. Many were developed to help in specific cases like robust optimization [186, 17, 209, 19, 210, 16] or multicriteria optimization [211]. Truly comprehensive comparisons of different optimization algorithms are unfortunately rare. Some concepts could be easily adaptable for use in a beamlet-free framework, like the usage of momentum terms or accelerated gradient methods, others like the robust optimization method developed by Cristoforetti et al. [186] are not. Overall, the algorithm used in the presented comparisons, L-BFGS, is common in proton therapy optimization and clinical TPS like Raystation also utilize

BFGS based optimization methods [133]. For both methods, beamlet-based and beamlet-free optimization, further research might lead to improvements in computational efficiency.
Lastly, the implementation of clinical TPS concerning data import, export, transfer and conversion are usually optimized for computational efficiency and carefully developed and maintained by a set of skilled software developers. Open-source TPS for use in clinics are typically written and maintained by researchers, often in addition to their main responsibilities or even in their free time, and they are built with accessability and modifiability in mind, not necessarily efficiency. This is a big part of why research TPS like OpenTPS [28] are usually unable to achieve the same computational efficiency as clinical TPS. This data handling affects beamlet-based implementations more than it does our beamlet-free one, where the dose-influence matrix is never created and both dose calculation and optimization are performed in the same external tool written in C. In contrast, in a beamlet-based implementation, the beamlet calculation is usually performed in an external MC tool, in this case MCsquare, then the results are imported into OpenTPS written in Python, converted to sparse format and exported again to the optimizer.

Objective types

Another important factor for clinical applications are the types of objectives that can be utilized in the objective function. For the time being, only minimum dose, maximum dose and mean dose objectives are supported in the beamlet-free framework. DVH-objectives however are not supported, since their use requires frequent DVH evaluations which are too computationally costly to perform after every iteration in the beamlet-free framework. The beamlet-free algorithm requires considerably more iterations than a conventional optimization algorithm such as L-BFGS, but is able to profit from the minimal time that is required per iteration. DVH objectives are highly desirable in a clinical context, since for many organs individual high dose voxels are managable, as long as only a small part of the volume is compromised.
Another type of optimization objectives that are of interest are biological and LET based objectives. MCsquare is already capable of calculating the LET during the MC dose simulation. It can be used to either calculate the biological effective dose BD as defined in equation (2.23) and optimize directly on the BD instead, or to define a combination of dose and LET terms and objectives. This would likely require little additional computational cost.

Expected-value robust optimization

A beamlet-free robust optimization strategy was presented based on expected-value robust optimization. This optimization approach is not considered "robust" according to the mathematical concept of robust optimization, but is typically referred to as such in the field of radiotherapy [135] due to its proven ability to generate treatment plans that are robust against uncertainties [27]. While expected-value robust optimization was proven an efficient tool at maintaining dose prescriptions under the face of errors and uncertainties, most clinical TPS opt for worst-case scenario robust optimization instead. The main reason for that is that worst-case optimization guarantees control even for the most challenging scenarios, while expected-value robust optimization may allow individual scenarios to perform significantly worse than the others.
However, our tests did not confirm this fear. The conventional expected-value based optimizer used in comparisons achieved very similar plan quality for the best, average and worst scenarios in the robustness evaluation. The expected-value robust optimizer achieved a similar plan quality as the worst-case scenario robust optimizer. However, the worst-case scenario robust optimizer converged to a higher objective function value than the expected-value robust optimizer. Had the expected-value robust optimizer not been terminated once the same objective function value was reached, it likely would have been able to achieve better expected-value, nominal and worst-case scenario objective functions at higher iterations.
The same can be said about the proposed beamlet-free robust optimizer. It showed only small differences between the average, nominal and worst case, comparable to those achieved by the two beamlet-based methods.
The proposed beamlet-free robust optimizer did not struggle with outliers in five out of the six investigated patient cases. In one patient case, however, robustness could not be achieved. An additional variance term was then introduced and successfully mitigated the issues and produced a robust plan.

Variance-guided robust optimization

Cristoferetti et al. [186] performed robust optimization based on an expected dose distribution, which in this case means a dose-influence matrix averaged over all n_s scenarios

$$\bar{D} = \frac{1}{n_s} \sum_s D_s \ , \tag{5.1}$$

where D_s refers to the dose-influence matrix of scenario s. To achieve robustness, a variance term Ω is added to the objective function which can then be written as

$$F(\mathbf{x}) = f(\bar{D}, \mathbf{x}) + \mathbf{x}^\top \Omega \mathbf{x} \ . \tag{5.2}$$

The variance term itself was defined as

$$\Omega = \sum_s p_s \Omega_s = \sum_s p_s \mathbb{E}[D_s^\top D_s] - \mathbb{E}[D_s]^\top \mathbb{E}[D_s] \ , \tag{5.3}$$

with p_s representing a scenario modulation factor.
The motivation for this "scenario-free" robust optimization approach was very similar to the idea behind the beamlet-free framework: the reduction of the dose-influence matrix. In robust optimization, it can be understood as a $m \times n \times k$ dimensional matrix where m is the number of voxel, n the number of spots and k the number of scenarios. The beamlet-free approach removes the dimension of spots, while the "scenario-free" approach removes the dimension of scenarios. The notable difference is that the number of spots usually far exceeds the number of scenarios and that the "scenario-free" approach still requires the full beamlet dose calculation in addition to a matrix preprocessing step. Nevertheless, the fact that a variance term is sufficient to ensure robust optimization is potentially quite relevant for other expected-value based approaches, including the beamlet-free one.
A variance term to drive optimization to reduce differences between scenarios was tested on the beamlet-free robust optimizer. Instead of calculating the variance of the dose distributions, the variance over the scenario objective functions was computed instead. It can be formulated as

$$\begin{aligned} \Omega &= \sum_k \Omega_k \\ &= \sum_k (f(d_k) - f(\bar{d}))^2 \ , \end{aligned} \tag{5.4}$$

where $\bar{d}$ is the averaged dose distribution. It is computed directly from the scoring contributions of all protons during dose calculation and therefore requires almost no effort to compute.
This beamlet-free implementation could successfully generate robust plans, but contrary to the implementation by Cristoforetti et al. brought no computational benefits compared to conventional expected-value robust optimization. Since no dose-influence matrix is computed in the first place, its size can obviously not be reduced to speed up optimization.
It was further investigated whether adding a variance term to an expected-value objective function of the form

$$F = \frac{1}{n_s} \sum_s f_s(\mathbf{x}) + \mathbf{x}^\top \Omega \mathbf{x} \ , \tag{5.5}$$

would provide a benefit. For most investigated patient cases, the addition of a variance term had little impact on plan quality compared to beamlet-free expected-value optimization without it after the same number of iterations. For two cases,

the additional variance term even worsened plan robustness. However, one case showed significant improvement. Further investigations would be required to draw definitive conclusions.

When to stop optimizing?

The comparisons made over the course of this thesis between the new beamlet-free algorithm and the established methods all suffer from the strong dependence of the computation time on several factors, such as the plan size, hardware used, but also the termination criteria of the optimization.
Usually in treatment planning, the optimization is terminated once a predefined convergence criterion has been reached. These convergence criteria are usually defined as the absolute value of the gradient norm or the change of the objective function between one or several iterations. A smaller value then leads to longer optimization times, closer proximity to the optimum and a higher resulting plan quality. A small caveat of these types of convergence criteria is that the total value of the objective function is fairly meaningless and depends on, for example, the concrete objective function used, the individual objective weights and whether they are normalized, how easily the objectives are satisfied etc. As a consequence, choosing the same convergence criterium for different plans does not necessarily imply that they are equally close to the optimal solution.
With respect to the comparisons run during the course of the thesis, an additional problem was that the methods that were compared had very different updating mechanisms. The conventional optimizer with L-BFGS typically only required a few hundred iterations to reach decent results and utilized multidirectional information. The beamlet-free optimizer on the other hand required several million iterations and only updated a single direction of the gradient in each one. Any convergence criteria comparing the change in objective function or gradient from one iteration to a set number later would terminate the two methods at very different stages of the optimization. Any increase in the number of iterations used for such a comparison would be just as arbitrary as choosing an entirely different convergence criterion for one of them. Additionally, the objective function value depends on the Monte Carlo noise of the underlying dose data. Since the beamlet-free optimizer continually simulates new particles and thereby reduces the Monte Carlo noise between iterations, any comparison of the objective function between iterations would be impacted.
For the sake of comparability, the optimization was terminated once a convergence criterion had been reached with the conventional beamlet-based method and the number of iterations of the beamlet-free optimizer were manually adapted to achieve comparable plan quality as specified in the respective sections. The plan quality was evaluated on dose data resimulated with matching particle numbers.

In the case of a severe discrepancy in optimization quality, the optimization was repeated with an adapted number of iterations, until comparable plan quality was reached. What constitutes a comparable plan quality was determined in each case individually, for example in the robust optimization comparison a similar scenario objective function value of the worst performing scenario was used.
This approach has its drawbacks as it leads to a conclusion of how plan quality varies after a set computation time instead of how much computation time varies to achieve a set plan quality. Since both qualities are obviously related this was deemed an acceptable concession, since any attempt to identify comparable convergence criteria and then repeat simulations with those multiple times would have added an undue amount of computation time to each comparison. If one considers that the comparisons of computation time can in the best case only give a vague idea about the actual performance of these methods in comparison to each other due to their dependence on multiple factors, this would have been excessive. After all, the plan size, voxel resolution, used hardware and desired convergence, all have a huge impact on how much computation time these methods require.
Beyond the scope of the presented comparisons, this raises the issue of what would constitute a good termination criterion for the beamlet-free optimizer in general. A suitable iteration number varies from plan to plan and could be identified from a convergence plot, the creation of which would already require one run, or through trial and error. Both approaches are inconvenient and time consuming and negate all benefits with respect to the speed of the treatment planning process. A convergence criterium based on the change in objective function could be identified just like for other optimization methods, however, it would need to take the impact of decreasing Monte Carlo noise into account. The Monte Carlo noise on the dose distribution can be quantified more easily, however it propagates to the objective function, where the relation between MC noise and number of simulated particles is considerably less straightforward. The MC noise affecting the dose distribution is proportional to $\sigma_{\mathsf{dose}} \propto \frac{1}{\sqrt{N}}$ for N simulated particles. The dose contributed quadratically to the utilized objective function f, which results in a noise contribution $\sigma_{\mathsf{f}} \propto \frac{1}{N}$. The exact impact of MC noise on the objective function will differ between plans and patients, however, so does the objective function value. A convergence criterion based on the difference in objective function between iterations in a beamlet-free framework would need to

- increase the number of iterations between comparisons and
- be larger than a comparable criterion for a beamlet-based optimizer in order to account for the contribution of the noise reduction.

This requires some intuition and awareness by the treatment planner, but not

necessarily more than is already the case for conventional methods.
Other options would be to observe the gradient norm instead of the change in objective function between iterations. This sounds unintuitive since the gradient is always calculated for just one spot making it noncomparable between different iterations. This is easily remedied however, by taking the combined or averaged gradient over a larger number of iterations. Statistical variations due to the impact of different spots can then be averaged out. However, while the impact of MC noise is very small on a single-iteration gradient, it would be present in the gradient computation as well.
The dependence on MC noise might be removed if the impact on spot weights between iterations is measured instead. In an optimized plan, the spot weights should theoretically not change between iterations. However, the definition of spot weights in a beamlet-free context is highly relevant. The quantities optimized in the stochastic coordinate step are spot probabilities. These translate to spot weights through the random sampling of spots in future iterations. While spot weights in a conventional optimizer ideally converge against a stable solution, the spot probabilities might always fluctuate slightly. The reason for this is that, assuming that an optimal dose distribution has been reached in iteration $k-1$, in iteration k another spot j will be sampled and a batch of n_{batch} particles will be simulated towards it. The spot probability p_j will be decreased to counter the departure from an optimal dose distribution. Overall, the stability of the spot probabilities is hard to guarantee or predict. The spot weights that are exported at the end of the optimization corresponds to the number of particles simulated towards each spot, normalized to meet the dose prescription. These spot weights could also be observed over the course of the optimization and should converge towards an optimal solution. They face the same basic issue as the spot probabilities that for as long as the optimization continues, the spot weight of sampled spot j will be increased in each iteration. The increase per iteration decreases as the optimization continues, due to the normalization of the weights. The more particles are simulated, the lower the impact of each batch. However, compared across a sufficiently high number of iterations, the change in the normalized spot weight distribution should converge against zero and might present a suitable candidate for a beamlet-free convergence criterium.

Implications for other radiotherapy modalities

The presented beamlet-free optimizer has been implemented within the proton-specific MC code MCsquare [12]. However, the underlying method is equally applicable to other forms of charged particle therapy, for example Carbon Ions.
It could theoretically be applied to radiotherapy with photons as well, however, the method was developed specifically to address challenges in the proton therapy

treatment planning workflow. Most of these are simply not an issue in conventional radiotherapy planning and a beamlet-free approach would provide little benefit.
Since photons do not have a Bragg peak, there are no spots. The photon beam is typically divided into thin wavelets and a fluence optimization is performed to modulate the individual wavelets. However, the number of wavelets in IMRT tends to be a lot smaller than the number of spots in IMPT, such that memory usage does not become burdensome.
Another consequence of their lack of finite range is that the usage of MC methods during dose computation is not a requirement as it is in IMPT. Though there are studies that show that the increased accuracy of MC methods can still bring a benefit during optimization [212, 213], their usage is not considered mandatory in IMRT and sufficient plan quality can be reached without them. Since the beamlet-free optimization approach builds on a continuous MC particle simulation, its usage is likely much slower than a typical IMRT workflow.
The big selling point of the beamlet-free optimizer, robust optimization, where it can achieve the highest benefit compared to beamlet-based optimization approaches, is likewise not necessary in IMRT. The PTV concept is usually adequate to achieve plans that are robust against treatment uncertainties.

Adaptive proton therapy

Adaptive proton therapy seeks to take changes to the patient geometry over the course of a fractionated treatment into account. This includes offline adaptations, where the treatment plan is recalculated based on updated CT and delivery information at a certain point.
More and more, online adaptive therapy moves into the center of attention. Online refers to the fact that imaging and plan adaptations are performed while the patient is already on the treatment couch. This allows for setup and positioning errors as well as the most up-to-date geometry information to be taken into account. However, it also requires fast and efficient plan reoptimizations in order to avoid blocking treatment rooms and couches at proton therapy facilities without an actively ongoing treatment. Xing et al. [214] identified computational challenges, including but not limited to plan reoptimizations, as the vital bottleneck in the dawn of adaptive radiotherapy with photons. Adaptive proton therapy needs to reexamine the challenge of fast reoptimizations taking the unique challenges of proton therapy into account, including the need for MC dose computations, the calculation, storage and multiplication with the large dose-influence matrix and robust optimization.
Plan reoptimizations require an updated dose influence matrix which is costly to compute. Analytical dose calculation algorithms are less accurate than MC dose calculation engines, however Nenoff et al. [215] showed that the benefit of adap-

tation, even on less accurate dose data obtained by an analytical algorithm, is still considerable. A fully MC-based adaptive workflow was presented by Botas et al. [87] utilizing the fast GPU-parallelized MC code gPMC [8]. Even though they were using a fast MC code, the recomputation of the dose influence matrix was deemed too computationally demanding for online applications. Instead they adapted the original beamlet information to fit the altered geometry, used the MC engine for verification of plan quality and only optimized if necessary. For the optimization, they identified a small subset of beamlets, calculated during the verification with sufficiently low uncertainty, to use as base for the optimization. This workflow required on average 16.9 s for the geometric adaptation, $115.6 - 419.2$ s for the MC dose calculation and $12 - 198$ s for the spot weight optimization, with potential for reduction depending on the number of GPUs and the used voxel resolution.

The pinnacle of adaptation would be reached in real-time online adaptive therapy. During the delivery itself images are taken and used to adapt the treatment plan on the spot. This could provide an alternative to gating or breathhold techniques in the treatment of lung tumors, but could also help to account for other intrafraction organ motion, for example due to the filling of cavities. Real-time online adaptive therapy still has a lot of questions to answer before becoming a realistic option in clinical routines. These challenges do include the imaging itself, but also the reoptimizations of the treatment plan which may require at most seconds of computation time. How this can be accomplished with the required accuracy is still mostly an open question.

A beamlet-free treatment planning approach might provide a solution, both for online, as well as real-time adaptive therapy. The benefits with respect to computation time that could be achieved are far from sufficient to eradicate all reason for worry, but they could already present a good step in the right direction. A more interesting aspect to beamlet-free reoptimization might be the possibility to reuse spot weight and dose information to potentially shortcut the treatment planning by a good margin. If one assumes that the geometry underlying the reoptimized plan still resembles the original planning CT, information from the previously optimized plan could be used to hotstart the beamlet-free algorithm with preexisting spot weight and dose values. Of course, reusing spot weights in reoptimization is also an option with conventional methods, however, there they could only affect the optimization, not the beamlet dose calculation, which would still be required for each new CT image.

Proton Arc therapy

Arc therapy refers to radiotherapeutic methods in which a continuously rotating gantry is used instead of individual fixed-angle beams to deliver dose to the tumor. Arc techniques have been used in radiotherapy with photons for some time now, including VMAT [216], IMAT [217], Tomotherapy [218] and RapidArc [219]. They could be shown to offer more flexibility by allowing more degrees of freedom during optimization, lower monitor units and integrated dose, reduced delivery time, and reduced dose to OARs [220, 221, 222, 223, 224, 225, 226, 227, 228, 229, 230].
These benefits are expected to translate to proton arc therapy as well and might even be enhanced because of the unique properties of protons and the highly conformal dose distributions they allow to be achieved. Different approaches have been proposed on how to perform proton arc therapy, including proton tomotherapy with distal edge tracking [231], and passive scattering approaches [232, 233, 234, 235]. However, their translation into clinical application proved to be very challenging. More recently, spot scanning proton arc therapy [236] has been introduced and a clinical delivery system based on its principle called DynamicARC® is under development by IBA.
The benefits of PAT plans have since been demonstrated for various treatment sites [237, 238, 239, 240, 169] and include reduced NTCP [241], improved LET distribution [242, 243] and increased radiobiological effectiveness [244] compared to IMPT plans.
In addition to significant challenges related to hardware, delivery, and quality assurance, PAT also introduces unique problems and challenges to the treatment planning workflow. If all potential beam angles in the arc were to be treated like fixed-angle beams in IMPT, with comparable numbers of energy layers and spots, the delivery time would become an issue due to the long layer switching times mandated by available hardware. A clinical PAT solution will need to take the delivery time into account during planning and optimize accordingly. Additionally, PAT plans require much higher numbers of spots due to the many beam angles, which poses a challenge for computational efficiency.
These two core challenges in PAT optimization can be addressed separately or jointly. A joint approach allows full profit of the available degrees of freedom to find the optimal plan, however, the number of energy layers is initially very large and poses high computational demands during optimization. A separate approach typically includes an energy layer preselection [245, 246] followed by a classical, usually robust, beamlet dose calculation and spot weight optimization workflow. The accessible degrees of freedom during optimization are reduced and plan quality might be inferior to plans optimized in a joint approach, but optimization time can be significantly reduced and final plan quality is still comparable or even superior to IMPT plans.

Other groups have proposed combined optimization approaches that take advantage of more degrees of freedom and are able to deliver optimal plans [239, 247, 248, 249, 250]. However, these methods come at the cost of high computational demand during treatment planning.
A beamlet-free optimization approach could be beneficial to both types of PAT optimization approaches. For energy layer preselection based methods, the expected benefit is comparable to that of regular robust optimization in IMPT with comparable numbers of spots. This already presents a considerable reduction in hardware requirements, computation time and memory usage.
For joint energy layer and spot weight optimization approaches which typically have high numbers of potential spots in the initial phases of optimization, the expected benefits could be significantly larger. These methods can be able to fully leverage the available degrees of freedom in PAT to produce optimal plans, but are often discarded in favor of energy layer preselection based methods. The primary reason for energy preselection based methods is the increased computational efficiency during optimization and Engwall et al. [245] even presented their algorithm as the trade-off between plan quality and computational efficiency. If the computational efficiency of other PAT optimizers could be significantly improved, there might not be a need to consider such a trade-off at all.
While the beamlet-free optimizer has not been tested specifically for PAT optimization, the benefits of the beamlet-free optimizer increase with the number of spots and can therefore be expected to be quite significant.

Final conclusion

The dosimetric benefits and steep dose gradients that proton therapy is able to achieve necessitate high degrees of accuracy and precision during treatment planning and delivery in order to succesfully translate into actual clinical benefits with respect to patient outcome. To minimize errors and uncertainties, MC dose calculations and robust optimization need to be integrated into proton therapy treatment planning. These methods create considerable computational costs, both with respect to time and memory usage.
A beamlet-free treatment planning approach is able to significantly reduce computational requirements during treatment planning and consistently achieves shorter computation times and lower memory usage. The extent of the expected benefits depends on the size of the plan and is especially relevant in the case of robust optimization.
Novel proton therapy techniques such as adaptive proton therapy and proton arc therapy bring additional challenges to treatment planning. Computational hurdles currently restrict their use in clinics. While a beamlet-free treatment planning

approach will not be sufficient to address all concerns, it is a considerable step in the direction of turning them into clinical reality.

References

[1] J. Ferlay, M. Ervik, F. Lam, M. Laversanne, M. Colombet, L. Mery, M. Piñeros, A. Znaor, I. Soerjomataram, and F. Bray, International Organization for Research on Cancer: Global Cancer Observatory: Cancer Today., 2024, https://gco.iarc.who.int/media/globocan/factsheets/cancers/39-all-cancers-fact-sheet.pdf; accessed: 20.01.2025.

[2] G. Delaney, S. Jacob, C. Featherstone, and M. Barton, The role of radiotherapy in cancer treatment, Cancer **104**, 1129–1137 (2005).

[3] M. Abdel-Wahab, S. S. Gondhowiardjo, A. A. Rosa, Y. Lievens, N. El-Haj, J. A. Polo Rubio, G. B. Prajogi, H. Helgadottir, E. Zubizarreta, A. Meghzifene, V. Ashraf, S. Hahn, T. Williams, and M. Gospodarowicz, Global Radiotherapy: Current Status and Future Directions—White Paper, JCO Global Oncology , 827–842 (2021), PMID: 34101482.

[4] M. Abdel-Wahab et al., Radiotherapy and theranostics: a Lancet Oncology Commission, The Lancet Oncology **25(11)**, e545–e580 (2024).

[5] R. Kohno, Y. Takada, T. Sakae, T. Terunuma, K. Matsumoto, A. Nohtomi, and H. Matsuda, Experimental evaluation of validity of simplified Monte Carlo method in proton dose calculations, Physics in Medicine Biology **48**, 1277 (2003).

[6] P. P. Yepes, D. Mirkovic, and P. J. Taddei, A GPU implementation of a track-repeating algorithm for proton radiotherapy dose calculations, Physics in Medicine Biology **55**, 7107 (2010).

[7] R. Kohno, K. Hotta, S. Nishioka, K. Matsubara, R. Tansho, and T. Suzuki, Clinical implementation of a GPU-based simplified Monte Carlo method for a treatment planning system of proton beam therapy, Physics in Medicine

Biology **56**, N287 (2011).

[8] X. Jia, J. Schümann, H. Paganetti, and S. B. Jiang, GPU-based fast Monte Carlo dose calculation for proton therapy, Physics in Medicine Biology **57**, 7783 (2012).

[9] K. Jabbari and J. Seuntjens, A fast Monte Carlo code for proton transport in radiation therapy based on MCNPX, Journal of Medical Physics **39**, 156–163 (2014).

[10] A. Schiavi, M. Senzacqua, S. Pioli, A. Mairani, G. Magro, S. Molinelli, M. Ciocca, G. Battistoni, and V. Patera, Fred: a GPU-accelerated fast-Monte Carlo code for rapid treatment plan recalculation in ion beam therapy, Physics in Medicine Biology **62**, 7482 (2017).

[11] H. Wan Chan Tseung, J. Ma, and C. Beltran, A fast GPU-based Monte Carlo simulation of proton transport with detailed modeling of nonelastic interactions, Medical Physics **42**, 2967–2978 (2015).

[12] K. Souris, J. A. Lee, and E. Sterpin, Fast multipurpose Monte Carlo simulation for proton therapy using multi- and many-core CPU architectures, Medical Physics **43**, 1700–1712 (2016).

[13] F. Fracchiolla, E. Engwall, M. Janson, F. Tamm, S. Lorentini, F. Fellin, M. Bertolini, C. Algranati, R. Righetto, P. Farace, M. Amichetti, and M. Schwarz, Clinical validation of a GPU-based Monte Carlo dose engine of a commercial treatment planning system for pencil beam scanning proton therapy, Physica Medica **88**, 226–234 (2021).

[14] J. Shan, H. Feng, D. H. Morales, S. H. Patel, W. W. Wong, M. Fatyga, M. Bues, S. E. Schild, R. L. Foote, and W. Liu, Virtual particle Monte Carlo: A new concept to avoid simulating secondary particles in proton therapy dose calculation, Medical Physics **49**, 6666–6683 (2022).

[15] C. Wu, D. Nguyen, Y. Xing, A. B. Montero, J. Schuemann, H. Shang, Y. Pu, and S. Jiang, Improving proton dose calculation accuracy by using deep learning, Machine Learning: Science and Technology **2**, 015017 (2021).

[16] G. Buti, K. Souris, A. M. Barragán Montero, M. Cohilis, J. A. Lee, and E. Sterpin, Accelerated robust optimization algorithm for proton therapy treatment planning, Medical Physics **47**, 2746–2754 (2020).

[17] Y. Xu, J. Chen, R. Cao, H. Liu, X. G. Xu, and X. Pei, A fast robust optimizer for intensity modulated proton therapy using GPU, Journal of Applied Clinical Medical Physics **21**, 123–133 (2020).

[18] A. Fu, V. T. Taasti, and M. Zarepisheh, Distributed and scalable optimization for robust proton treatment planning, Medical Physics **50**, 633–642 (2023).

[19] Q. Fan, L. Zhao, X. Li, J. Hu, X. Lu, Z. Yang, S. Zhang, K. Yang, X. Ding, G. Liu, and S. Dai, A novel fast robust optimization algorithm for intensity-modulated proton therapy with minimum monitor unit constraint, Medical Physics **51**, 6220–6230 (2024).

[20] Y. M. Yang, M. Svatos, C. Zankowski, and B. Bednarz, Concurrent Monte Carlo transport and fluence optimization with fluence adjusting scalable transport Monte Carlo, Medical Physics **43**, 3034–3048 (2016).

[21] Y. Li, Z. Tian, F. Shi, T. Song, Z. Wu, Y. Liu, S. Jiang, and X. Jia, A new Monte Carlo-based treatment plan optimization approach for intensity modulated radiation therapy, Physics in Medicine Biology **60**, 2903 (2015).

[22] Y. Li, Z. Tian, T. Song, Z. Wu, Y. Liu, S. Jiang, and X. Jia, A new approach to integrate GPU-based Monte Carlo simulation into inverse treatment plan optimization for proton therapy, Physics in Medicine Biology **62**, 289 (2016).

[23] W. Liu, X. Zhang, Y. Li, and R. Mohan, Robust optimization of intensity modulated proton therapy, Medical Physics **39**, 1079–1091 (2012).

[24] D. Pross, S. Wuyckens, S. Deffet, E. Sterpin, J. A. Lee, and K. Souris, Technical note: Beamlet-free optimization for Monte-Carlo-based treatment planning in proton therapy, Medical Physics **51**, 485–493 (2024).

[25] A. Fredriksson, A. Forsgren, and B. Hårdemark, Minimax optimization for handling range and setup uncertainties in proton therapy, Medical Physics **38**, 1672–1684 (2011).

[26] J. Unkelbach, T. C. Y. Chan, and T. Bortfeld, Accounting for range uncertainties in the optimization of intensity modulated proton therapy, Physics in Medicine Biology **52**, 2755 (2007).

[27] J. Unkelbach, T. Bortfeld, B. C. Martin, and M. Soukup, Reducing the

sensitivity of IMPT treatment plans to setup errors and range uncertainties via probabilistic treatment planning, Medical Physics **36**, 149–163 (2009).

[28] S. Wuyckens, D. Dasnoy, G. Janssens, V. Hamaide, M. Huet, E. Loÿen, G. R. de Hertaing, B. Macq, E. Sterpin, J. A. Lee, K. Souris, and S. Deffet, OpenTPS – Open-source treatment planning system for research in proton therapy, 2023.

[29] M. Maqbool, editor, *An Introduction to Medical Physics*, Biological and Medical Physics, Biomedical Engineering, Springer International Publishing, Cham, 2017.

[30] S. Keevil, R. Padovani, S. Tabakov, T. Greener, and C. Lewis, editors, *Introduction to Medical Physics*, CRC Press, Boca Raton, 2022.

[31] M. d. Nunes, *Hadron Therapy Physics and Simulations*, SpringerBriefs in Physics, Springer, New York, NY, 2014.

[32] M. Joiner and A. van der Kogel, *Basic Clinical Radiobiology*, CRC Press, 2018.

[33] H. Paganetti, *Proton therapy physics*, CRC press, 2016.

[34] R. Baskar, K. Lee, R. Yeo, and K. Yeoh, Cancer and radiation therapy: current advances and future directions, Int J Med Sci **9**, 193–9 (2012).

[35] M. Bencheikh, A. Maghnouj, and J. Tajmouati, Percentage depth dose fragmentation for investigating and assessing the photon beam dosimetry quality, Journal of Radiotherapy in Practice **18**, 280–284 (2019).

[36] H.-H. Tseng, Y. Luo, R. Ten Haken, and I. El Naqa, The Role of Machine Learning in Knowledge-Based Response-Adapted Radiotherapy, Frontiers in Oncology **8** (2018).

[37] A. Rakotomalala, A. Escande, A. Furlan, S. Meignan, and E. Lartigau, Hypoxia in Solid Tumors: How Low Oxygenation Impacts the "Six Rs" of Radiotherapy, Frontiers in endocrinology **12**, 742215 (2021).

[38] H. E.J and G. A.J, *Radiobiology for the Radiologist*, volume 7, 2011.

[39] Z. Klaassen and A. Raldow, Urotoday: SUO 2024: Novel Approaches to Radiotherapy to Reduce Patient Burden, 2024, https://www.urotoday.com/conference-highlights/suo-2024/suo-2024-

prostate-cancer/156763-suo-2024-novel-approaches-to-radiotherapy-to-reduce-patient-burden.html; accessed: 08.03.2025.

[40] S. Stefanowicz, K. Stützer, S. Zschaeck, A. Jakobi, and E. G. C. Troost, Comparison of different treatment planning approaches for intensity-modulated proton therapy with simultaneous integrated boost for pancreatic cancer, Radiation Oncology **13** (2018).

[41] J. Mott and N. West, Essentials of Depth Dose Calculations for Clinical Oncologists, Clinical Oncology **33**, 5–11 (2021).

[42] Y. Lievens and W. V. den Bogaert, Proton beam therapy: Too expensive to become true?, Radiotherapy and Oncology **75**, 131–133 (2005).

[43] T. Bortfeld and J. Loeffler, Three ways to make proton therapy affordable, Nature **549**, 451–453 (2017).

[44] Z. Chen, M. M. Dominello, M. C. Joiner, and J. W. Burmeister, Proton versus photon radiation therapy: A clinical review, Frontiers in Oncology **Volume 13 - 2023** (2023).

[45] A. Peeters, J. P. Grutters, M. Pijls-Johannesma, S. Reimoser, D. De Ruysscher, J. L. Severens, M. A. Joore, and P. Lambin, How costly is particle therapy? Cost analysis of external beam radiotherapy with carbon-ions, protons and photons, Radiotherapy and Oncology **95**, 45–53 (2010).

[46] J. Son, S. B. Lee, Y. Lim, S. Y. Park, K. Cho, M. Yoon, and D. Shin, Development of Optical Fiber Based Measurement System for the Verification of Entrance Dose Map in Pencil Beam Scanning Proton Beam, Sensors **18** (2018).

[47] A. Koehler, R. Schneider, and J. Sisterson, Range modulators for protons and heavy ions, Nuclear Instruments and Methods **131**, 437–440 (1975).

[48] A. M. Koehler, R. J. Schneider, and J. M. Sisterson, Flattening of proton dose distributions for large-field radiotherapy, Medical Physics **4**, 297–301 (1977).

[49] T. Kanai, K. Kawachi, Y. Kumamoto, H. Ogawa, T. Yamada, H. Matsuzawa, and T. Inada, Spot scanning system for proton radiotherapy, Medical Physics **7**, 365–369 (1980).

[50] A. J. Lomax, T. Böhringer, A. Bolsi, D. Coray, F. Emert, G. Goitein, M. Jer-

mann, S. Lin, E. Pedroni, H. Rutz, O. Stadelmann, B. Timmermann, J. Verwey, and D. C. Weber, Treatment planning and verification of proton therapy using spot scanning: Initial experiences, Medical Physics **31**, 3150 – 3157 (2004), Cited by: 253.

[51] J. Edmund and T. Nyholm, A review of substitute CT generation for MRI-only radiation therapy, Radiation Oncology **12** (2017).

[52] E. Johnstone, J. J. Wyatt, A. M. Henry, S. C. Short, D. Sebag-Montefiore, L. Murray, C. G. Kelly, H. M. McCallum, and R. Speight, Systematic Review of Synthetic Computed Tomography Generation Methodologies for Use in Magnetic Resonance Imaging–Only Radiation Therapy, International Journal of Radiation Oncology*Biology*Physics **100**, 199–217 (2018).

[53] M. A. Bahloul, S. Jabeen, S. Benoumhani, H. A. Alsaleh, Z. Belkhatir, and A. Al-Wabil, Advancements in synthetic CT generation from MRI: A review of techniques, and trends in radiation therapy planning, Journal of Applied Clinical Medical Physics **25**, e14499 (2024).

[54] G. Pileggi, C. Speier, G. C. Sharp, D. Izquierdo Garcia, C. Catana, J. Pursley, F. Amato, J. Seco, and M. F. Spadea, Proton range shift analysis on brain pseudo-CT generated from T1 and T2 MR, Acta oncologica **57**, 1521–1531 (2018).

[55] L. Tian and A. Luehr, Proton range uncertainty caused by synthetic computed tomography generated with deep learning from pelvic magnetic resonance imaging, Acta oncologica **62**, 1461–1469 (2023).

[56] J. C. Stroom, H. C. de Boer, H. Huizenga, and A. G. Visser, Inclusion of geometrical uncertainties in radiotherapy treatment planning by means of coverage probability, International Journal of Radiation Oncology*Biology*Physics **43**, 905–919 (1999).

[57] M. van Herk, P. Remeijer, C. Rasch, and J. V. Lebesque, The probability of correct target dosage: dose-population histograms for deriving treatment margins in radiotherapy, International Journal of Radiation Oncology*Biology*Physics **47**, 1121–1135 (2000).

[58] K. Harrison, H. Pullen, C. Welsh, O. Oktay, J. Alvarez-Valle, and R. Jena, Machine Learning for Auto-Segmentation in Radiotherapy Planning, Clinical Oncology **34**, 74–88 (2022), Artificial Intelligence in Radiation Therapy.

[59] A. C. Erdur, D. Rusche, D. Scholz, J. Kiechle, S. Fischer, Óscar Llorián-Salvador, J. A. Buchner, M. Q. Nguyen, L. Etzel, J. Weidner, M.-C. Metz, B. Wiestler, J. Schnabel, D. Rueckert, S. E. Combs, and J. C. Peeken, Deep learning for autosegmentation for radiotherapy treatment planning: State-of-the-art and novel perspectives, Strahlentherapie und Onkologie **201**, 236–254 (2025).

[60] J. Wong, V. Huang, D. Wells, J. Giambattista, J. Giambattista, C. Kolbeck, K. Otto, E. P. Saibishkumar, and A. Alexander, Implementation of deep learning-based auto-segmentation for radiotherapy planning structures: a workflow study at two cancer centers, Radiation Oncology **16** (2021).

[61] V. T. Taasti, L. Hong, J. S. Shim, J. O. Deasy, and M. Zarepisheh, Automating proton treatment planning with beam angle selection using Bayesian optimization, Medical Physics **47**, 3286–3296 (2020).

[62] W. Cao, H. Rocha, R. Mohan, G. Lim, H. Goudarzi, B. Ferreira, and J. Dias, Reflections on beam configuration optimization for intensity-modulated proton therapy, Physics in medicine and biology **67** (2022), Publisher Copyright: © 2022 Institute of Physics and Engineering in Medicine.

[63] H. Shen, G. Zhang, Y. Lin, R. L. Rotondo, Y. Long, and H. Gao, Beam angle optimization for proton therapy via group-sparsity based angle generation method, Medical Physics **50**, 3258–3273 (2023).

[64] H. Rocha and J. Dias, On polling directions for randomized direct-search approaches: application to beam angle optimization in intensity-modulated proton therapy, Journal of Global Optimization **91**, 371–392 (2025).

[65] K.-H. Küfer, M. Monz, A. Scherrer, P. Süss, F. Alonso, A.-S. Azizi-Sultan, T. Bortfeld, and C. Thieke, *Multicriteria Optimization in Intensity Modulated Radiotherapy Planning*, volume 26, pages 1–45, 2009.

[66] B. Lin, S. Fu, Y. Lin, R. L. Rotondo, W. Huang, H. H. Li, R. C. Chen, and H. Gao, An adaptive spot placement method on Cartesian grid for pencil beam scanning proton therapy, Physics in Medicine Biology **66**, 235012 (2021).

[67] G. Meier, D. Leiser, R. Besson, A. Mayor, S. Safai, D. C. Weber, and A. J. Lomax, Contour scanning for penumbra improvement in pencil beam scanned proton therapy, Physics in Medicine Biology **62**, 2398 (2017).

[68] M. ur Rehman, K. Erhart, J. Kielbasa, S. L. Meeks, Z. Li, T. Willoughby, N. Ramakrishna, K. Stephenson, T. S. Rahman, P. Kelly, and O. Zeidan, An optimized approach for robust spot placement in proton pencil beam scanning, Physics in Medicine Biology **64**, 235016 (2019).

[69] M. ur Rehman, O. A. Zeidan, T. Willoughby, S. L. Meeks, P. Kelly, and K. Erhart, Dosimetric Comparison of Various Spot Placement Techniques in Proton Pencil Beam Scanning, International Journal of Particle Therapy **9**, 54–63 (2022).

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

[71] J. Schuemann, J. Shin, J. Perl, C. Grassberger, J. Verburg, B. Faddegon, and H. Paganetti, SU-E-T-500: Pencil-Beam versus Monte Carlo Based Dose Calculation for Proton Therapy Patients with Complex Geometries. Clinical Use of the TOPAS Monte Carlo System, Medical Physics **39**, 3820–3820 (2012).

[72] C. Grassberger, J. Daartz, S. Dowdell, T. Ruggieri, G. Sharp, and H. Paganetti, Quantification of Proton Dose Calculation Accuracy in the Lung, International Journal of Radiation Oncology*Biology*Physics **89**, 424–430 (2014).

[73] J. Schuemann, D. Giantsoudi, C. Grassberger, M. Moteabbed, C. H. Min, and H. Paganetti, Assessing the Clinical Impact of Approximations in Analytical Dose Calculations for Proton Therapy, International Journal of Radiation Oncology*Biology*Physics **92**, 1157–1164 (2015).

[74] P. A. Taylor, S. F. Kry, and D. S. Followill, Pencil Beam Algorithms Are Unsuitable for Proton Dose Calculations in Lung, International Journal of Radiation Oncology*Biology*Physics **99**, 750–756 (2017).

[75] B. K. Sasidharan, S. Aljabab, J. Saini, T. Wong, G. Laramore, J. Liao, U. Parvathaneni, and S. R. Bowen, Clinical Monte Carlo versus Pencil Beam Treatment Planning in Nasopharyngeal Patients Receiving IMPT, International Journal of Particle Therapy **5**, 32–40 (2019).

[76] J. Stanley, K. Breitman, P. Dunscombe, D. P. Spencer, and H. Lau, Evaluation of stereotactic radiosurgery conformity indices for 170 target volumes in patients with brain metastases, Journal of Applied Clinical Medical Physics

12, 245–253 (2011).

[77] I. Paddick, A simple scoring ratio to index the conformity of radiosurgical treatment plans, Journal of Neurosurgery **93**, 219–222 (2000).

[78] L. B. Marks, E. D. Yorke, A. Jackson, R. K. Ten Haken, L. S. Constine, A. Eisbruch, S. M. Bentzen, J. Nam, and J. O. Deasy, Use of Normal Tissue Complication Probability Models in the Clinic, International Journal of Radiation Oncology*Biology*Physics **76**, S10–S19 (2010), Quantitative Analyses of Normal Tissue Effects in the Clinic.

[79] A. J. Lomax, Intensity modulated proton therapy and its sensitivity to treatment uncertainties 2: the potential effects of inter-fraction and inter-field motions, Physics in Medicine Biology **53**, 1043 (2008).

[80] J. Kim, Y.-K. Park, G. Sharp, P. Busse, and B. Winey, Water equivalent path length calculations using scatter-corrected head and neck CBCT images to evaluate patients for adaptive proton therapy, Physics in Medicine Biology **62**, 59 (2016).

[81] M. Moteabbed, A. Trofimov, G. C. Sharp, Y. Wang, A. L. Zietman, J. A. Efstathiou, and H.-M. Lu, Proton therapy of prostate cancer by anterior-oblique beams: implications of setup and anatomy variations, Physics in Medicine Biology **62**, 1644 (2017).

[82] A. Trofimov, P. L. Nguyen, J. A. Efstathiou, Y. Wang, H.-M. Lu, M. Engelsman, S. Merrick, C.-W. Cheng, J. R. Wong, and A. L. Zietman, Interfractional Variations in the Setup of Pelvic Bony Anatomy and Soft Tissue, and Their Implications on the Delivery of Proton Therapy for Localized Prostate Cancer, International Journal of Radiation Oncology*Biology*Physics **80**, 928–937 (2011).

[83] Y. Wang, J. A. Efstathiou, G. C. Sharp, H.-M. Lu, I. Frank Ciernik, and A. V. Trofimov, Evaluation of the dosimetric impact of interfractional anatomical variations on prostate proton therapy using daily in-room CT images, Medical Physics **38**, 4623–4633 (2011).

[84] Y. Z. Szeto, M. G. Witte, S. R. van Kranen, J.-J. Sonke, J. Belderbos, and M. van Herk, Effects of anatomical changes on pencil beam scanning proton plans in locally advanced NSCLC patients, Radiotherapy and Oncology **120**, 286–292 (2016).

[85] S. Sajja, Y. Lee, M. Eriksson, H. Nordström, A. Sahgal, M. Hashemi, J. G. Mainprize, and M. Ruschin, Technical Principles of Dual-Energy Cone Beam Computed Tomography and Clinical Applications for Radiation Therapy, Advances in radiation oncology **5**, 1–16 (2019).

[86] G. R. Sarria, F. Kugel, F. Roehner, J. Layer, C. Dejonckheere, D. Scafa, M. Koeksal, C. Leitzen, and L. C. Schmeel, Artificial Intelligence–Based Autosegmentation: Advantages in Delineation, Absorbed Dose-Distribution, and Logistics, Advances in Radiation Oncology **9**, 101394 (2024).

[87] P. Botas, J. Kim, B. Winey, and H. Paganetti, Online adaption approaches for intensity modulated proton therapy for head and neck patients based on cone beam CTs and Monte Carlo simulations, Physics in Medicine Biology **64**, 015004 (2018).

[88] T. Jagt, S. Breedveld, S. van de Water, B. Heijmen, and M. Hoogeman, Near real-time automated dose restoration in IMPT to compensate for daily tissue density variations in prostate cancer, Physics in Medicine Biology **62**, 4254 (2017).

[89] T. Jagt, S. Breedveld, R. van Haveren, B. Heijmen, and M. Hoogeman, An automated planning strategy for near real-time adaptive proton therapy in prostate cancer, Physics in Medicine Biology **63**, 135017 (2018).

[90] T. Z. Jagt, S. Breedveld, R. van Haveren, R. A. Nout, E. Astreinidou, B. J. M. Heijmen, and M. S. Hoogeman, Plan-library supported automated replanning for online-adaptive intensity-modulated proton therapy of cervical cancer, Acta Oncologica **58**, 1440–1445 (2019).

[91] K. Bernatowicz, X. Geets, A. Barragan, G. Janssens, K. Souris, and E. Sterpin, Feasibility of online IMPT adaptation using fast, automatic and robust dose restoration, Physics in Medicine Biology **63**, 085018 (2018).

[92] M. Bobić, A. Lalonde, G. C. Sharp, C. Grassberger, J. M. Verburg, B. A. Winey, A. J. Lomax, and H. Paganetti, Comparison of weekly and daily online adaptation for head and neck intensity-modulated proton therapy, Physics in Medicine Biology **66**, 055023 (2021).

[93] M. Matter, L. Nenoff, G. Meier, D. C. Weber, A. J. Lomax, and F. Albertini, Intensity modulated proton therapy plan generation in under ten seconds, Acta Oncologica **58**, 1435–1439 (2019).

[94] H. Paganetti, P. Botas, G. C. Sharp, and B. Winey, Adaptive proton therapy, Physics in Medicine Biology **66**, 22TR01 (2021).

[95] A. Neishabouri, J. Bauer, A. Abdollahi, J. Debus, and A. Mairani, Real-time adaptive proton therapy: An AI-based spatio-temporal mono-energetic dose calculation model (CC-LSTM), Computers in Biology and Medicine **188**, 109777 (2025).

[96] F. Albertini, A. McWilliam, and B. Winey, Editorial: Advances in online and real-time adaptive radiotherapy, Physics in Medicine Biology **70**, 070301 (2025).

[97] H.-P. Wieser, E. Cisternas, N. Wahl, S. Ulrich, A. Stadler, H. Mescher, L.-R. Müller, T. Klinge, H. Gabrys, L. Burigo, A. Mairani, S. Ecker, B. Ackermann, M. Ellerbrock, K. Parodi, O. Jäkel, and M. Bangert, Development of the open-source dose calculation and optimization toolkit matRad, Medical Physics **44**, 2556–2568 (2017).

[98] W. Newhauser and R. Zhang, The physics of proton therapy, Physics in medicine and biology **60**, R155–R209 (2015).

[99] H. A. Bethe, Molière's Theory of Multiple Scattering, Phys. Rev. **89**, 1256–1266 (1953).

[100] S. Agostinelli et al., Geant4—a simulation toolkit, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **506**, 250–303 (2003).

[101] A. Ferrari, P. Sala, A. Fasso, and J. Ranft, FLUKA: a multi-particle transport code, CERN-2005-10 (2005), INFN/TC_05/11, SLAC-R-773.

[102] M. Fippel and M. Soukup, A Monte Carlo dose calculation algorithm for proton therapy, Medical Physics **31**, 2263–2273 (2004).

[103] J. S. Li, B. Shahine, E. Fourkal, and C.-M. Ma, A particle track-repeating algorithm for proton beam dose calculation, Physics in Medicine Biology **50**, 1001 (2005).

[104] H. E. Robbins, A Stochastic Approximation Method, Annals of Mathematical Statistics **22**, 400–407 (1951).

[105] J. Kiefer and J. Wolfowitz, Stochastic Estimation of the Maximum of a Regression Function, Annals of Mathematical Statistics **23**, 462–466

(1952).

[106] J. R. Blum, Approximation Methods which Converge with Probability one, The Annals of Mathematical Statistics **25**, 382 – 386 (1954).

[107] Y. Nesterov, Efficiency of Coordinate Descent Methods on Huge-Scale Optimization Problems, SIAM Journal on Optimization **22**, 341–362 (2012).

[108] S. Wright, Coordinate descent algorithms, Mathematical Programming **151**, 3–34 (2015).

[109] S. J. Wright and B. Recht, *Coordinate Descent*, page 100–117, Cambridge University Press, 2022.

[110] B. Polyak, Some methods of speeding up the convergence of iteration methods, USSR Computational Mathematics and Mathematical Physics **4**, 1–17 (1964).

[111] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, On the importance of initialization and momentum in deep learning, in *Proceedings of the 30th International Conference on Machine Learning*, edited by S. Dasgupta and D. McAllester, volume 28 of *Proceedings of Machine Learning Research*, pages 1139–1147, Atlanta, Georgia, USA, 2013, PMLR.

[112] Y. Nesterov, A method for solving the convex programming problem with convergence rate $O(1/k^2)$, Proceedings of the USSR Academy of Sciences **269**, 543–547 (1983).

[113] Y. Liu, Y. Gao, and W. Yin, An Improved Analysis of Stochastic Gradient Descent with Momentum, 2020.

[114] H. Fawzi, Lecture Notes: Topics in Convex Optimisation.

[115] H. Ye, L. Luo, and Z. Zhang, Approximate Newton Methods, Journal of Machine Learning Research **22**, 1–41 (2021).

[116] D. E. Steward, *Numerical Analysis: A Graduate Course*, chapter Optimization, pages 537–600, CMS/CAIMS Books in Mathematics, Springer Cham, 1 edition, 2022.

[117] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Applied Optimization, Springer, 1 edition, 2003.

[118] C. G. BROYDEN, The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations, IMA Journal of Applied Mathematics **6**, 76–90 (1970).

[119] R. Fletcher, A new approach to variable metric algorithms, The Computer Journal **13**, 317–322 (1970).

[120] D. Goldfarb, A family of variable-metric methods derived by variational means, Math. Comp. **24**, 23–26 (1970).

[121] D. F. Shanno, Conditioning of quasi-Newton methods for function minimization, Math. Comp. **24**, 647–656 (1970).

[122] P. Virtanen et al., SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, Nature Methods **17**, 261–272 (2020).

[123] Github, Scipy implementation of the BFGS algorithm.

[124] O. T. Poul Erik Frandsen, Kristian Jonasson Hans Bruun Nielsen, Unconstrained optimization, 1999.

[125] M. J. D. Powell, Convergence Properties of Algorithms for Nonlinear Optimization, SIAM Review **28**, 487–500 (1986).

[126] P. Wolfe, Convergence Conditions for Ascent Methods, SIAM Review **11**, 226–235 (1969).

[127] Y.-H. Dai, Convergence Properties of the BFGS Algoritm, SIAM Journal on Optimization **13**, 693–701 (2002).

[128] J. Nocedal, Updating Quasi-Newton Matrices with Limited Storage, Mathematics of Computation **35**, 773–782 (1980).

[129] D. C. Liu and J. Nocedal, On the limited memory BFGS method for large scale optimization, Mathematical Programming **45**, 503–528 (1989).

[130] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer Series in Operations Research and Financial Engineering, Springer, 2 edition, 2006.

[131] U. Oelfke and T. Bortfeld, Inverse planning for photon and proton beams, Medical Dosimetry **26**, 113–124 (2001).

[132] C. Wu, G. Olivera, H. Keller, T. McNutt, and T. Mackie, Optimization

of importance factors and desired doses in a quadratic dose objective function, in *Proceedings of the 22nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (Cat. No.00CH37143)*, volume 1, pages 123–126 vol.1, 2000.

[133] M. Janson, L. Glimelius, A. Fredriksson, E. Traneus, and E. Engwall, Treatment planning of scanned proton beams in RayStation, Medical Dosimetry **49**, 2–12 (2024), Treatment Planning in Proton Therapy.

[134] A. J. Lomax, Intensity modulated proton therapy and its sensitivity to treatment uncertainties 1: the potential effects of calculational uncertainties, Physics in Medicine Biology **53**, 1027 (2008).

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

[136] W. Liu, S. J. Frank, X. Li, Y. Li, R. X. Zhu, and R. Mohan, PTV-based IMPT optimization incorporating planning risk volumes vs robust optimization, Medical Physics **40**, 021709 (2013).

[137] W. Liu, R. Mohan, P. Park, Z. Liu, H. Li, X. Li, Y. Li, R. Wu, N. Sahoo, L. Dong, X. R. Zhu, and D. R. Grosshans, Dosimetric benefits of robust treatment planning for intensity modulated proton therapy for base-of-skull cancers, Practical Radiation Oncology **4**, 384–391 (2014).

[138] H. Li, X. Zhang, P. Park, W. Liu, J. Chang, Z. Liao, S. Frank, Y. Li, F. Poenisch, R. Mohan, M. Gillin, and R. Zhu, Robust optimization in intensity-modulated proton therapy to account for anatomy changes in lung cancer patients, Radiotherapy and Oncology **114**, 367–372 (2015).

[139] F. Albertini, E. B. Hug, and A. J. Lomax, Is it necessary to plan with safety margins for actively scanned proton therapy?, Physics in Medicine Biology **56**, 4399 (2011).

[140] W. Liu, S. J. Frank, X. Li, Y. Li, P. C. Park, L. Dong, X. Ronald Zhu, and R. Mohan, Effectiveness of robust optimization in intensity-modulated proton therapy planning for head and neck cancers, Medical Physics **40**, 051711 (2013).

[141] W. Liu, X. Zhang, Y. Li, and R. Mohan, Robust optimization of intensity

modulated proton therapy, Medical Physics **39**, 1079–1091 (2012).

[142] W. Chen, J. Unkelbach, A. Trofimov, T. Madden, H. Kooy, T. Bortfeld, and D. Craft, Including robustness in multi-criteria optimization for intensity-modulated proton therapy, Physics in Medicine Biology **57**, 591 (2012).

[143] B. Schaffner and E. Pedroni, The precision of proton range calculations in proton radiotherapy treatment planning: experimental verification of the relation between CT-HU and proton stopping power, Physics in Medicine Biology **43**, 1579 (1998).

[144] O. Jäkel and P. Reiss, The influence of metal artefacts on the range of ion beams, Physics in Medicine Biology **52**, 635 (2007).

[145] C. Graeff, Motion mitigation in scanned ion beam therapy through 4D-optimization, Physica Medica **30**, 570–577 (2014), Particle Radiosurgery Conference.

[146] E. Engwall, A. Fredriksson, and L. Glimelius, 4D robust optimization including uncertainties in time structures can reduce the interplay effect in proton pencil beam scanning radiation therapy, Medical Physics **45**, 4020–4029 (2018).

[147] R. Canters, K. van der Klugt, V. T. Taasti, J. Buijsen, B. Ta, I. Steenbakkers, R. Houbena, G. Vilches-Freixas, and M. Berbee, Robustness of intensity modulated proton treatment of esophageal cancer for anatomical changes and breathing motion, Radiotherapy and Oncology **198**, 110409 (2024).

[148] A. J. Lomax, E. Pedroni, H. P. Rutz, and G. Goitein, The Clinical Potential of Intensity Modulated Proton Therapy, Zeitschrift für Medizinische Physik **14**, 147–152 (2004).

[149] D. Pflugfelder, J. J. Wilkens, and U. Oelfke, Worst case optimization: a method to account for uncertainties in the optimization of intensity modulated proton therapy, Physics in Medicine Biology **53**, 1689 (2008).

[150] M. R. Hestenes and E. Stiefel, Methods of Conjugate Gradients for Solving Linear Systems, Journal of Research of the National Bureau of Standards **49**, 409 (1952).

[151] M. Zarepisheh, L. Xing, and Y. A. Ye, A computation study on an integrated

alternating direction method of multipliers for large scale optimization, Optimization Letters **12**, 3–15 (2018).

[152] H. Paganetti, Relative biological effectiveness (RBE) values for proton beam therapy. Variations as a function of biological endpoint, dose, and linear energy transfer, Physics in Medicine Biology **59**, R419 (2014).

[153] U. T. S. H. L. D. M. M. D. K. X. R. Z. J. D. M. S. C. S. C. R. P. R. T. R. M. . D. R. G. Fada Guan, Lawrence Bronk, Spatial mapping of the biologic effectiveness of scanned particle beams: towards biologically optimized particle therapy, Scientific Reports **5** (2015).

[154] Y. Chen and S. Ahmad, Empirical model estimation of relative biological effectiveness for proton beam therapy†, Radiation Protection Dosimetry **149**, 116–123 (2011).

[155] A. Carabe, M. Moteabbed, N. Depauw, J. Schuemann, and H. Paganetti, Range uncertainty in proton therapy due to variable biological effectiveness, Physics in Medicine Biology **57**, 1159 (2012).

[156] M. Wedenberg, B. K. Lind, and B. Hårdemark, A model for the relative biological effectiveness of protons: The tissue specific parameter / of photons is a predictor for the sensitivity to LET changes, Acta Oncologica **52**, 580 – 588 (2013), Cited by: 223.

[157] B. Jones, A simpler energy transfer efficiency model to predict relative biological effect for protons and heavier ions, Frontiers in Oncology **5** (2015), Cited by: 42; All Open Access, Gold Open Access, Green Open Access.

[158] A. L. McNamara, J. Schuemann, and H. Paganetti, A phenomenological relative biological effectiveness (RBE) model for proton therapy based on all published in vitro cell survival data, Physics in Medicine Biology **60**, 8399 (2015).

[159] A. Mairani, I. Dokic, G. Magro, T. Tessonnier, J. Bauer, T. T. Böhlen, M. Ciocca, A. Ferrari, P. R. Sala, O. Jäkel, J. Debus, T. Haberer, A. Abdollahi, and K. Parodi, A phenomenological relative biological effectiveness approach for proton therapy based on an improved description of the mixed radiation field, Physics in Medicine Biology **62**, 1378 (2017).

[160] E. Rørvik, S. Thörnqvist, C. H. Stokkevåg, T. J. Dahle, L. F. Fjæra, and K. S. Ytre-Hauge, A phenomenological biological dose model for proton

therapy based on linear energy transfer spectra, Medical Physics **44**, 2586–2594 (2017).

[161] E. Rørvik, L. F. Fjæra, T. J. Dahle, J. E. Dale, G. M. Engeseth, C. H. Stokkevåg, S. Thörnqvist, and K. S. Ytre-Hauge, Exploration and application of phenomenological RBE models for proton therapy, Physics in Medicine Biology **63**, 185013 (2018).

[162] J. O. Brita Singers Sørensen and N. Bassler, In vitro RBE-LET dependence for multiple particle types, Acta Oncologica **50**, 757–762 (2011), PMID: 21767171.

[163] W. Deng, Y. Yang, C. Liu, M. Bues, R. Mohan, W. W. Wong, R. H. Foote, S. H. Patel, and W. Liu, A Critical Review of LET-Based Intensity-Modulated Proton Therapy Plan Evaluation and Optimization for Head and Neck Cancer Management, International Journal of Particle Therapy **8**, 36–49 (2021).

[164] D. Giantsoudi, C. Grassberger, D. Craft, A. Niemierko, A. Trofimov, and H. Paganetti, Linear Energy Transfer-Guided Optimization in Intensity Modulated Proton Therapy: Feasibility Study and Clinical Potential, International Journal of Radiation Oncology*Biology*Physics **87**, 216–222 (2013).

[165] J. Unkelbach, P. Botas, D. Giantsoudi, B. L. Gorissen, and H. Paganetti, Reoptimization of Intensity Modulated Proton Therapy Plans Based on Linear Energy Transfer, International Journal of Radiation Oncology*Biology*Physics **96**, 1097–1106 (2016).

[166] Y. An, J. Shan, S. H. Patel, W. Wong, S. E. Schild, X. Ding, M. Bues, and W. Liu, Robust intensity-modulated proton therapy to reduce high linear energy transfer in organs at risk, Medical Physics **44**, 6138–6147 (2017).

[167] W. Cao, A. Khabazian, P. P. Yepes, G. Lim, F. Poenisch, D. R. Grosshans, and R. Mohan, Linear energy transfer incorporated intensity modulated proton therapy optimization, Physics in Medicine Biology **63**, 015013 (2017).

[168] M. McIntyre, P. Wilson, P. Gorayski, and E. Bezak, A Systematic Review of LET-Guided Treatment Plan Optimisation in Proton Therapy: Identifying the Current State and Future Needs, Cancers **15** (2023).

[169] G. Liu, X. Li, A. Qin, W. Zheng, D. Yan, S. Zhang, C. Stevens, P. Kab-

olizadeh, and X. Ding, Improve the dosimetric outcome in bilateral head and neck cancer (HNC) treatment using spot-scanning proton arc (SPArc) therapy: a feasibility study, Radiation oncology (London, England) **15**, 21 (2020).

[170] H. M. Kooy and C. Grassberger, Intensity modulated proton therapy, The British Journal of Radiology **88**, 20150195 (2015), PMID: 26084352.

[171] H. Paganetti, Range uncertainties in proton therapy and the role of Monte Carlo simulations, Physics in Medicine Biology **57**, R99 (2012).

[172] A. Tourovsky, A. J. Lomax, U. Schneider, and E. Pedroni, Monte Carlo dose calculations for spot scanned proton therapy, Physics in Medicine Biology **50**, 971 (2005).

[173] J. Ma, C. Beltram, H. S. W. C. Tseung, and M. G. Herman, A GPU-accelerated and Monte Carlo-based intensity modulated proton therapy optimization system, Medical physics **41**, 121707 (2014).

[174] P. Ziegenhein, J. J. Wilkens, S. Nill, T. Ludwig, and U. Oelfke, Speed optimized influence matrix processing in inverse treatment planning tools, Physics in Medicine Biology **53**, N157 (2008).

[175] J. Bernoulli, *Wahrscheinlichkeitsrechnung (Ars conjectandi, 1713)*, Ostwalds Klassiker der exakten Wissenschaften, 1899.

[176] D. C. Weber, H. P. Rutz, A. Bolsi, E. Pedroni, A. Coray, M. Jermann, A. J. Lomax, E. B. Hug, and G. Goitein, Spot Scanning Proton Therapy in the Curative Treatment of Adult Patients With Sarcoma: The Paul Scherrer Institute Experience, International Journal of Radiation Onkology **69**, 865–871 (2007).

[177] A. J. Lomax, T. Boehringer, A. Coray, E. Egger, G. Goitein, M. Grossmann, P. Juelke, S. Lin, E. Pedroni, B. Rohrer, W. Roser, B. Rossi, B. Siegenthaler, O. Stadelmann, H. Stauble, C. Vetter, and L. Wisser, Intensity modulated proton therapy: A clinical example, Medical Physics **28**, 317–324 (2001).

[178] A. Fredriksson and R. Bokrantz, A critical evaluation of worst case optimization methods for robust intensity-modulated proton therapy planning, Medical Physics **41**, 081701 (2014).

[179] A. Fredriksson, A characterization of robust radiation therapy treatment

planning methods—from expected value to worst case optimization, Medical Physics **39**, 5169–5181 (2012).

[180] J. Ma, H. S. Wan Chan Tseung, M. G. Herman, and C. Beltran, A robust intensity modulated proton therapy optimizer based on Monte Carlo dose calculation, Medical Physics **45**, 4045–4054 (2018).

[181] F. Albertini, M. Matter, L. Nenoff, Y. Zhang, and A. Lomax, Online daily adaptive proton therapy, The British Journal of Radiology **93**, 20190594 (2020).

[182] X. Wang, H. Li, X. R. Zhu, Q. Hou, L. Liao, B. Jiang, Y. Li, P. Wang, J. Lang, and X. Zhang, Multiple-CT optimization of intensity-modulated proton therapy – Is it possible to eliminate adaptive planning?, Radiotherapy and Oncology **128**, 167–173 (2018).

[183] S. van de Water, F. Albertini, D. C. Weber, B. J. M. Heijmen, M. S. Hoogeman, and A. J. Lomax, Anatomical robust optimization to account for nasal cavity filling variation during intensity-modulated proton therapy: a comparison with conventional and adaptive planning strategies, Physics in Medicine Biology **63**, 025020 (2018).

[184] M. Cubillos-Mesías, E. G. Troost, F. Lohaus, L. Agolli, M. Rehm, C. Richter, and K. Stützer, Including anatomical variations in robust optimization for head and neck proton therapy can reduce the need of adaptation, Radiotherapy and Oncology **131**, 127–134 (2019).

[185] Z. Yang, X. Zhang, X. Wang, X. Zhu, G. Gunn, S. Frank, Y. Chang, Q. Li, K. Yang, G. wu, L. Liao, Y. Li, M. Chen, and H. Li, Multiple-CT optimization: An adaptive optimization method to account for anatomical changes in intensity-modulated proton therapy for head and neck cancers, Radiotherapy and Oncology **142** (2019).

[186] R. Cristoforetti, J. J. Hardt, and N. Wahl, Scenario-free robust optimization algorithm for IMRT and IMPT treatment planning, Medical Physics **n/a**.

[187] E. Sterpin, S. T. Rivas, F. V. den Heuvel, B. George, J. A. Lee, and K. Souris, Development of robustness evaluation strategies for enabling statistically consistent reporting, Physics in Medicine Biology **66**, 045002 (2021).

[188] J. Schuemann, S. Dowdell, C. Grassberger, C. H. Min, and H. Paganetti,

Site-specific range uncertainties caused by dose calculation algorithms for proton therapy, Physics in Medicine Biology **59**, 4007 (2014).

[189] C. Geng, J. Daartz, K. Lam-Tin-Cheung, M. Bussiere, H. A. Shih, H. Paganetti, and J. Schuemann, Limitations of analytical dose calculations for small field proton radiosurgery, Physics in Medicine Biology **62**, 246 (2016).

[190] J. Saini, D. Maes, A. Egan, S. R. Bowen, S. St James, M. Janson, T. Wong, and C. Bloch, Dosimetric evaluation of a commercial proton spot scanning Monte-Carlo dose algorithm: comparisons against measurements and simulations, Physics in Medicine Biology **62**, 7659 (2017).

[191] P. Yepes, A. Adair, D. Grosshans, D. Mirkovic, F. Poenisch, U. Titt, Q. Wang, and R. Mohan, Comparison of Monte Carlo and analytical dose computations for intensity modulated proton therapy, Physics in Medicine Biology **63**, 045003 (2018).

[192] D. Maes, J. Saini, J. Zeng, R. Rengan, T. Wong, and S. R. Bowen, Advanced proton beam dosimetry part II: Monte Carlo vs. pencil beam-based planning for lung cancer, Translational Lung Cancer Research **7** (2018).

[193] S. Shiraishi and K. L. Moore, Knowledge-based prediction of three-dimensional dose distributions for external beam radiotherapy, Medical Physics **43**, 378–387 (2016).

[194] C. McIntosh and T. G. Purdie, Contextual Atlas Regression Forests: Multiple-Atlas-Based Automated Dose Prediction in Radiation Therapy, IEEE Transactions on Medical Imaging **35**, 1000–1012 (2016).

[195] C. McIntosh and T. G. Purdie, Voxel-based dose prediction with multi-patient atlas selection for automated radiotherapy treatment planning, Physics in Medicine Biology **62**, 415 (2016).

[196] V. Kearney, J. W. Chan, S. Haaf, M. Descovich, and T. D. Solberg, DoseNet: a volumetric dose prediction algorithm using 3D fully-convolutional neural networks, Physics in Medicine Biology **63**, 235022 (2018).

[197] D. Nguyen, T. Long, X. Jia, W. Lu, X. Gu, Z. Iqbal, and S. Jiang, A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning, Scientific reports **9**, 1076 (2019).

[198] A. M. Barragán-Montero, D. Nguyen, W. Lu, M.-H. Lin, R. Norouzi-Kandalan, X. Geets, E. Sterpin, and S. Jiang, Three-dimensional dose prediction for lung IMRT patients with deep neural networks: robust learning from heterogeneous beam configurations, Medical Physics **46**, 3679–3691 (2019).

[199] X. Chen, K. Men, Y. Li, J. Yi, and J. Dai, A feasibility study on an automated method to generate patient-specific dose distributions for radiotherapy using deep learning, Medical Physics **46**, 56–64 (2019).

[200] J. Fan, J. Wang, Z. Chen, C. Hu, Z. Zhang, and W. Hu, Automatic treatment planning based on three-dimensional dose distribution predicted from deep learning technique, Medical Physics **46**, 370–381 (2019).

[201] F. Guerreiro, E. Seravalli, G. Janssens, J. Maduro, A. Knopf, J. Langendijk, B. Raaymakers, and C. Kontaxis, Deep learning prediction of proton and photon dose distributions for paediatric abdominal tumours, Radiotherapy and Oncology **156**, 36–42 (2021).

[202] F. Pirlepesov, L. Wilson, V. P. Moskvin, A. Breuer, F. Parkins, J. T. Lucas Jr., T. E. Merchant, and A. M. Faught, Three-dimensional dose and LETD prediction in proton therapy using artificial neural networks, Medical Physics **49**, 7417–7427 (2022).

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

[204] M. Mohammadisiahroudi, W. Zou, L. Dong, Y. Zinchenko, and T. Terlaky, On Computational Challenges in Proton Therapy Treatment Planning, 2024, XXth International Conference on the use of Computers in Radiation therapy.

[205] A. Niemierko and M. Goitein, The use of variable grid spacing to accelerate dose calculations, Medical Physics **16**, 357–366 (1989).

[206] M. Wiesmeyer, M. Miften, and S. Monthofer, A multigrid method for accelerating superposition dose calculations, in *The Use of Computers in Radiation Therapy*, edited by W. Schlegel and T. Bortfeld, pages 221–223, Berlin, Heidelberg, 2000, Springer Berlin Heidelberg.

[207] D. Pflugfelder, J. J. Wilkens, S. Nill, and U. Oelfke, A comparison of

three optimization algorithms for intensity modulated radiation therapy, Zeitschrift für Medizinische Physik **18**, 111–119 (2008).

[208] X. Zhang, H. Liu, X. Wang, L. Dong, Q. Wu, and R. Mohan, Speed and convergence properties of gradient algorithms for optimization of IMRT, Medical Physics **31**, 1141 – 1152 (2004), Cited by: 53.

[209] G. Buti, K. Souris, A. M. B. Montero, J. A. Lee, and E. Sterpin, Towards fast and robust 4D optimization for moving tumors with scanned proton therapy, Medical Physics **46**, 5434–5443 (2019).

[210] H. Mahmoudzadeh, T. G. Purdie, and T. C. Chan, Constraint generation methods for robust optimization in radiation therapy, Operations Research for Health Care **8**, 85–90 (2016).

[211] W. Chen, D. Craft, T. M. Madden, K. Zhang, H. M. Kooy, and G. T. Herman, A fast optimization algorithm for multicriteria intensity modulated proton therapy planning, Medical Physics **37**, 4938–4945 (2010).

[212] N. Dogan, J. V. Siebers, P. J. Keall, F. Lerma, Y. Wu, M. Fatyga, J. F. Williamson, and R. K. Schmidt-Ullrich, Improving IMRT dose accuracy via deliverable Monte Carlo optimization for the treatment of head and neck cancer patients, Medical Physics **33**, 4033–4043 (2006).

[213] R. Jeraj, P. J. Keall, and J. V. Siebers, The effect of dose calculation accuracy on inverse treatment planning, Physics in Medicine Biology **47**, 391 (2002).

[214] L. Xing, J. Siebers, and P. Keall, Computational Challenges for Image-Guided Radiation Therapy: Framework and Current Research, Seminars in Radiation Oncology **17**, 245–257 (2007), Image-Guided Radiation Therapy.

[215] L. Nenoff, M. Matter, A. G. Jarhall, C. Winterhalter, J. Gorgisyan, M. Josipovic, G. F. Persson, P. Munck af Rosenschold, D. C. Weber, A. J. Lomax, and F. Albertini, Daily Adaptive Proton Therapy: Is it Appropriate to Use Analytical Dose Calculations for Plan Adaption?, International Journal of Radiation Oncology*Biology*Physics **107**, 747–755 (2020).

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

[217] C. X. Yu, Intensity-modulated arc therapy with dynamic multileaf collima-

tion: an alternative to tomotherapy, Physics in Medicine Biology **40**, 1435 (1995).

[218] T. R. Mackie, T. Holmes, S. Swerdloff, P. Reckwerdt, J. O. Deasy, J. Yang, B. Paliwal, and T. Kinsella, Tomotherapy: A new concept for the delivery of dynamic conformal radiotherapy, Medical Physics **20**, 1709–1719 (1993).

[219] C. C. Ling, P. Zhang, Y. Archambault, J. Bocanek, G. Tang, and T. LoSasso, Commissioning and Quality Assurance of RapidArc Radiotherapy Delivery System, International Journal of Radiation Oncology*Biology*Physics **72**, 575–581 (2008).

[220] M. Teoh, C. H. Clark, K. Wood, S. Whitaker, and A. Nisbet, Volumetric modulated arc therapy: a review of current literature and clinical use in practice, British Journal of Radiology **84**, 967–996 (2011).

[221] G. Macchia, F. Deodato, S. Cilla, S. Cammelli, A. Guido, M. Ferioli, G. Siepe, V. Valentini, A. G. Morganti, and G. Ferrandina, Volumetric modulated arc therapy for treatment of solid tumors: current insights, OncoTargets and Therapy **Volume 10**, 3755–3772 (2017).

[222] S. O. Hunte, C. H. Clark, N. Zyuzikov, and A. Nisbet, Volumetric modulated arc therapy (VMAT): a review of clinical outcomes—what is the clinical evidence for the most effective implementation?, British Journal of Radiology **95**, 20201289 (2022).

[223] P. Zhang, L. Happersett, M. Hunt, A. Jackson, M. Zelefsky, and G. Mageras, Volumetric Modulated Arc Therapy: Planning and Evaluation for Prostate Cancer Cases, International Journal of Radiation Oncology*Biology*Physics **76**, 1456–1462 (2010).

[224] F. Kjær-Kristoffersen, L. Ohlhues, J. Medin, and S. Korreman, RapidArc volumetric modulated therapy planning for prostate cancer patients, Acta Oncologica **48**, 227–232 (2009).

[225] S. Vieillot, D. Azria, C. Lemanski, C. L. Moscardo, S. Gourgou, J.-B. Dubois, N. Aillères, and P. Fenoglietto, Plan comparison of volumetric-modulated arc therapy (RapidArc) and conventional intensity-modulated radiation therapy (IMRT) in anal canal cancer, Radiation Oncology **5**, 92 (2010).

[226] W. Duthoy, W. De Gersem, K. Vergote, M. Coghe, T. Boterberg, Y. De

Deene, C. De Wagter, S. Van Belle, and W. De Neve, Whole abdominopelvic radiotherapy (WAPRT) using intensity-modulated arc therapy (IMAT): first clinical experience, International Journal of Radiation Oncology*Biology*Physics **57**, 1019–1032 (2003).

[227] W. Duthoy, W. De Gersem, K. Vergote, T. Boterberg, C. Derie, P. Smeets, C. De Wagter, and W. De Neve, Clinical implementation of intensity-modulated arc therapy (IMAT) for rectal cancer, International Journal of Radiation Oncology*Biology*Physics **60**, 794–806 (2004).

[228] E. Wong, D. P. D'Souza, J. Z. Chen, M. Lock, G. Rodrigues, T. Coad, K. Trenka, M. Mulligan, and G. S. Bauman, Intensity-modulated arc therapy for treatment of high-risk endometrial malignancies, International Journal of Radiation Oncology*Biology*Physics **61**, 830–841 (2005).

[229] L. Cozzi, K. A. Dinshaw, S. K. Shrivastava, U. Mahantshetty, R. Engineer, D. D. Deshpande, S. Jamema, E. Vanetti, A. Clivio, G. Nicolini, and A. Fogliata, A treatment planning study comparing volumetric arc modulation with RapidArc and fixed field IMRT for cervix uteri radiotherapy, Radiotherapy and Oncology **89**, 180–191 (2008).

[230] J. L. Bedford, Treatment planning for volumetric modulated arc therapy, Medical Physics **36**, 5128–5138 (2009).

[231] J. Deasy, T. Mackie, and P. DeLuca, Conformal proton tomotherapy using distal-edge tracking, Radiotherapy and Oncology **37** (1995).

[232] G. Sandison, E. Papiez, C. Bloch, and J. Morphis, Phantom assessment of lung dose from proton arc therapy, International Journal of Radiation Oncology*Biology*Physics **38**, 891–897 (1997).

[233] J. Seco, G. Gu, T. Marcelos, H. Kooy, and H. Willers, Proton Arc Reduces Range Uncertainty Effects and Improves Conformality Compared With Photon Volumetric Modulated Arc Therapy in Stereotactic Body Radiation Therapy for Non-Small Cell Lung Cancer, International Journal of Radiation Oncology*Biology*Physics **87**, 188–194 (2013).

[234] D. Sánchez-Parcerisa, J. C. Pourbaix, C. G. Ainsley, D. Dolney, and A. Carabe, Fast range switching of passively scattered proton beams using a modulation wheel and dynamic beam current modulation, Physics in Medicine Biology **59**, N19 (2014).

[235] X. Ding, A. Witztum, X. Liang, M. Reiche, H. Lin, B. Teo, L. Yin, J. Fiene, J. McDonough, and A. Kassaee, WE-F-16A-03: 3D Printer Application in Proton Therapy: A Novel Method to Deliver Passive-Scattering Proton Beams with a Fixed Range and Modulation for SRS and SRT, Medical Physics **41**, 514–514 (2014).

[236] X. Ding, X. Li, J. M. Zhang, P. Kabolizadeh, C. Stevens, and D. Yan, Spot-Scanning Proton Arc (SPArc) Therapy: The First Robust and Delivery-Efficient Spot-Scanning Proton Arc Therapy, International Journal of Radiation Oncology*Biology*Physics **96**, 1107–1116 (2016).

[237] X. Ding, X. Li, A. Qin, J. Zhou, D. Yan, C. Stevens, D. Krauss, and P. Kabolizdeh, Have we reached proton beam therapy dosimetric limitations? - A novel robust, delivery-efficient and continuous spot-scanning proton arc (SPArc) therapy is to improve the dosimetric outcome in treating prostate cancer, Acta oncologica (Stockholm, Sweden) **57**, 435–437 (2018).

[238] X. Li, P. Kabolizadeh, D. Yan, A. Qin, J. Zhou, Y. Hong, T. Guerrero, I. Grills, C. Stevens, and X. Ding, Improve dosimetric outcome in stage III non-small-cell lung cancer treatment using spot-scanning proton arc (SPArc) therapy, Radiation oncology (London, England) **13**, 35 (2018).

[239] X. Ding, J. Zhou, X. Li, K. Blas, G. Liu, Y. Wang, A. Qin, P. Chinnaiyan, D. Yan, C. S. I. Grills, and P. Kabolizadeh, Improving dosimetric outcome for hippocampus and cochlea sparing whole brain radiotherapy using spot-scanning proton arc therapy, Acta oncologica (Stockholm, Sweden) **58**, 483–490 (2019).

[240] S. Chang, G. Liu, L. Zhao, J. Dilworth, W. Zheng, S. Jawad, D. Yan, P. Chen, C. Stevens, P. Kabolizadeh, X. Li, and X. Ding, Feasibility study: spot-scanning proton arc therapy (SPArc) for left-sided whole breast radiotherapy, Radiat Oncol **15**, 232 (2020).

[241] B. A. de Jong, C. Battinelli, J. Free, D. Wagenaar, E. Engwall, G. Janssens, J. A. Langendijk, E. W. Korevaar, and S. Both, Spot scanning proton arc therapy reduces toxicity in oropharyngeal cancer patients, Medical Physics **50**, 1305–1317 (2023).

[242] A. Bertolet and A. Carabe, Proton monoenergetic arc therapy (PMAT) to enhance LETd within the target, Physics in Medicine Biology **65**, 165006 (2020).

[243] X. Li, X. Ding, W. Zheng, G. Liu, G. Janssens, K. Souris, A. M. Barragán-Montero, D. Yan, C. Stevens, and P. Kabolizadeh, Linear Energy Transfer Incorporated Spot-Scanning Proton Arc Therapy Optimization: A Feasibility Study, Frontiers in Oncology **11** (2021).

[244] A. Carabe, I. V. Karagounis, K. Huynh, A. Bertolet, N. François, M. M. Kim, A. Maity, E. Abel, and R. Dale, Radiobiological effectiveness difference of proton arc beams versus conventional proton and photon beams, Physics in Medicine Biology **65**, 165002 (2020).

[245] E. Engwall, C. Battinelli, V. Wase, O. Marthin, L. Glimelius, R. Bokrantz, B. Andersson, and A. Fredriksson, Fast robust optimization of proton PBS arc therapy plans using early energy layer selection and spot assignment, Physics in Medicine Biology **67**, 065010 (2022).

[246] S. Wuyckens, G. Janssens, M. Chocan Vera, J. Sundström, D. Di Perri, E. Sterpin, K. Souris, and J. A. Lee, Proton arc therapy plan optimization with energy layer pre-selection driven by organ at risk sparing and delivery time, Physics in Medicine Biology **70**, 045003 (2025).

[247] W. Gu, D. Ruan, Q. Lyu, W. Zou, L. Dong, and K. Sheng, A novel energy layer optimization framework for spot-scanning proton arc therapy, Medical Physics **47**, 2072–2084 (2020).

[248] S. Wuyckens, L. Zhao, M. Saint-Guillain, G. Janssens, E. Sterpin, K. Souris, X. Ding, and J. A. Lee, Bi-criteria Pareto optimization to balance irradiation time and dosimetric objectives in proton arc therapy, Physics in Medicine Biology **67**, 245017 (2022).

[249] G. Zhang, H. Shen, Y. Lin, R. C. Chen, Y. Long, and H. Gao, Energy layer optimization via energy matrix regularization for proton spot-scanning arc therapy, Medical Physics **49**, 5752–5762 (2022).

[250] G. Zhang, Y. Long, Y. Lin, R. C. Chen, and H. Gao, A treatment plan optimization method with direct minimization of number of energy jumps for proton arc therapy, Physics in Medicine Biology **68**, 085001 (2023).