Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

ACCEPTED MANUSCRIPT • OPEN ACCESS

A hybrid column generation and simulated annealing algorithm for direct aperture optimization

To cite this article before publication: Silvan Mueller *et al* 2022 *Phys. Med. Biol.* in press https://doi.org/10.1088/1361-6560/ac58db

A hybrid column generation and simulated annealing algorithm for direct aperture optimization

S Mueller[1], G Guyer[1], T Risse[1], S Tessarini[1], D M Aebersold[1], M F M Stampanoni[2], M K Fix[1] and P Manser[1]

[1] Division of Medical Radiation Physics and Department of Radiation Oncology, Inselspital, Bern University Hospital, and University of Bern, Switzerland

[2] Institute for Biomedical Engineering, ETH Zürich and PSI, Villigen, Switzerland

To be submitted as a research article to:

Physics in Medicine & Biology

Corresponding Author:

Silvan Mueller, PhD

Division of Medical Radiation Physics

Inselspital and University Hospital Bern

CH-3010 Bern

Switzerland

Tel: +41 31 632 44 02

Fax: +41 31 632 26 76

E-Mail: silvan.mueller@insel.ch

Abstract

The purpose of this work was to develop a hybrid column generation (CG) and simulated annealing (SA) algorithm for direct aperture optimization (H-DAO) and to show its effectiveness in generating high quality treatment plans for intensity modulated radiation therapy (IMRT) and mixed photon-electron beam radiotherapy (MBRT).

The H-DAO overcomes limitations of the CG-DAO with two features improving aperture selection (branch-feature) and enabling aperture shape changes during optimization (SA-feature). The H-DAO algorithm iteratively adds apertures to the plan. At each iteration, a branch is created for each field provided. First, each branch determines the most promising aperture of its assigned field and adds it to a copy of the current apertures. Afterwards, the apertures of each branch undergo an MU-weight optimization followed by an SA-based simultaneous shape and MU-weight optimization and a second MU-weight optimization. The next H-DAO iteration continues the branch with the lowest objective function value. IMRT and MBRT treatment plans for an academic, a brain and a head and neck case generated using the CG-DAO and H-DAO were compared.

For every investigated case and both IMRT and MBRT, the H-DAO leads to a faster convergence of the objective function value with number of apertures compared to the CG-DAO. In particular, the H-DAO needs on average half the apertures to reach the same objective function value as the CG-DAO for a specifically selected number of apertures. The average aperture areas are 27% smaller for H-DAO than for CG-DAO leading to a slightly larger discrepancy between optimized and final dose. However, a dosimetric benefit remains.

The H-DAO was successfully developed and applied to IMRT and MBRT. The faster convergence with number of apertures of the H-DAO compared to the CG-DAO allows to select a better compromise between plan quality and number of apertures.

Keywords: Direct aperture optimization, column generation, simulated annealing, IMRT, MBRT

1. Introduction

Treatment plans are generated and delivered in photon intensity modulated radiation therapy (IMRT) (Bortfeld 2006) to achieve a highly conformal dose distribution to the target volume. This was enabled through the introduction of the multileaf collimator (MLC) (Convery and Webb 1992), which collimates photon beams delivered from static beam directions in a step-and-shoot or dynamic movement manner such as sliding-window. Optimization algorithms were introduced to determine a suitable intensity modulation of the photon fields, which are discretized into beamlets (Webb 1989, Bortfeld 2006). These algorithms are classified in two categories: fluence map optimization (FMO) and direct aperture optimization (DAO). FMO optimizes the non-negative and independent weights of the beamlets in terms of monitor units (MUs) of each provided treatment field simultaneously. FMO is a large-scale convex optimization problem that can be efficiently solved using deterministic algorithms such as gradient descent. However, FMO results only in a non-deliverable fluence map for each field. Thus, a post-processing leaf-sequencing step is added translating the fluence maps to deliverable plans, resulting in a degraded plan quality (Mohan *et al* 2000). In contrast, DAO directly considers the machine constraints of the MLC such as leaf movement constraints or minimal gaps. Thus, DAO deals with MU weighted mechanically deliverable apertures that describe which beamlets are covered by the MLC leaves. However, this leads to a difficult large-scale non-convex optimization problem which cannot be solved efficiently. Several algorithms were developed for DAO:

- Column generation based DAO (CG-DAO) (Romeijn *et al* 2005, Preciado-Walters *et al* 2006, Men *et al* 2007, Carlsson 2008, Renaud *et al* 2017) iteratively adds the most promising aperture shape to the aperture pool of the plan using a pricing mechanism. The most promising aperture shape of all the provided fields is the one with the steepest gradient on the objective function, called price. After each aperture addition, the aperture MU weights are optimized (restricted master problem) using a deterministic algorithm. This approach is computationally efficient and has full freedom in the number of apertures per field. On the other side, the aperture shapes stay fixed as soon as appended to the aperture pool. Moreover, the selection of the most promising aperture is only based on the objective function gradient,

i.e. it is unknown by how much the objective function value can be decreased by increasing its MU weight and re-adjusting the MU weights of the already added apertures. Thus, a suboptimal aperture selection is given, especially if many fields are provided like this is the case for 4Pi (Dong *et al* 2013) or mixed beam radiotherapy (MBRT) (Palma *et al* 2012, Renaud *et al* 2017, Mueller *et al* 2017).

- Stochastic DAO approaches such as simulated annealing (SA-DAO) (Shepard *et al* 2002), quantum tunnel annealing (Pakela *et al* 2020) and genetic algorithms (Li *et al* 2003) randomly change the shapes and MU weights of apertures according to a scheme . In contrast to the CG-DAO algorithm, the number of apertures per field is pre-defined. Moreover, these approaches start with arbitrary aperture shapes such as conformal to the target or closed. Hence, numerous optimization steps are needed leading to a long computation time. On the other side, these algorithms have basically the ability to overcome local minima, because they also accept changes on the apertures leading to a worse objective function value with a certain probability. In comparison to the CG-DAO, another benefit is the ability to simultaneously optimize the shapes and MU weights of the apertures.
- Local gradient-based leaf refinement approaches such as the direct machine parameter optimization (DMPO) (Hårdemark *et al* 2004) and aperture shape optimization (ASO) (Cassioli and Unkelbach 2013) typically start with an initial aperture set generated through CG-DAO (Carlsson 2008, Cassioli and Unkelbach 2013) or FMO and leaf-sequencing. Subsequently, they refine the leaf positions locally within the current beamlet using a linear function of the leaf positions approximating the dose distribution. They are able to simultaneously optimize shapes and MU weights of the apertures. However, they can end up in a local minimum making them dependent on the starting conditions.
- Segmentation of fluence maps can also be directly integrated into FMO as shown by Nguyen *et al* 2017 using a multiphase piecewise constant Mumford-Shah formulation. Thus, the MLC constraints are directly included in the FMO formulation making it a DAO. Potential

limitations of the published approach are that it is designed to generate only non-overlapping apertures and that the maximal number of allowed apertures per field is pre-defined.

In this work, a novel algorithm called hybrid DAO (H-DAO) is developed to solve the DAO problem. The H-DAO follows the basic idea of the CG-DAO of adding apertures iteratively, but the H-DAO overcomes limitations of the CG-DAO with the following newly implemented features:

- The branch-feature exploring the most promising aperture of each field in a separate branch to identify the aperture improving the objective function value the most.
- The SA-feature applying the simulated annealing algorithm to enable continuous optimization of the aperture shapes and to enable continuous leaf positions not limited by the discrete beamlet grid resolution.

The H-DAO is applied to IMRT and MBRT. We demonstrate the effectiveness of the H-DAO in generating treatment plans of high quality compared to the baseline of a CG-DAO.

2. Methods

2.1. Treatment planning process

An implementation of the H-DAO is embedded in the treatment planning process (TPP) as illustrated in Figure 1 (top). The TPP considered here is used to create treatment plans for IMRT (photon apertures only) and MBRT (photon and electron apertures) deliverable in a step-and-shoot manner on a TrueBeam (Varian Medical Systems, Palo Alto, CA) treatment unit equipped with a Millennium MLC 120 (Varian Medical Systems, Palo Alto, CA) to collimate photon and electron beams. However, the whole TPP could also be conceptually applied to other MLC based and photon-electron beam supporting treatment units. In the first subprocess, the CT image set is imported into a research version of the Eclipse treatment planning system (TPS) 15.6 (Varian Medical Systems, Palo Alto, CA) and the contouring of the PTV, organs at risk (OARs) and normal tissue (body without PTV) is done using Eclipse.

The second subprocess consists of the manual setup of photon and optionally electron radiation fields within Eclipse. The definition of a field consists of the gantry, table and collimator angles, isocenter location, particle type, beam energy, secondary collimator jaw positions and the beamlet grid size. For photon beams, the secondary collimator jaw positions are equivalent to the maximum rectangular beamlet grid size, while for the electron beams, the secondary collimator jaws are always set to 15 x 35 cm^2 due to the electron beam model utilized (Henzen *et al* 2014). These field definitions stay fixed, but one can setup as many fields as wanted with different field aspects (e.g. for each electron beam direction, all possible beam energies can be setup to give the H-DAO more freedom in selecting the energy of apertures to be added).

To perform Monte Carlo (MC) beamlet dose calculations of the photon and electron fields, the Eclipse interfaced Swiss Monte Carlo Plan (SMCP) (Fix *et al* 2007) is used. The source of the beamlet dose calculations are pre-simulated phase-spaces located at the treatment head exit plane. The dose distributions of photon and electron beamlets are calculated using Voxel Monte Carlo (VMC++) (Kawrakow and Fippel 2000) and Macro Monte Carlo (MMC) (Neuenschwander and Born 1992, Neuenschwander *et al* 1995, Fix *et al* 2013), respectively. Both MC algorithms are embedded within the SMCP framework. The beamlet size is 0.5 x 0.5 cm^2 or 0.5 x 1.0 cm^2 depending whether the beamlet belongs to an inner 0.5 cm wide leaf pair or outer thicker 1 cm wide leaf pair. After beamlet dose calculation, the plan optimization is performed using the H-DAO as described in detail in the next section 2.2.

Next, a final dose calculation of the optimized apertures considering the impact of the MLC is performed within the SMCP framework. Source of the dose calculation for photon beams is a pre-simulated phase-space located on a plane above the secondary collimator jaws. For photon beams, VMC++ is used to simulate the patient-specific part of the treatment head including secondary collimator jaws and MLC as well as the dose calculation in the patient. For electron beams, the source is a multiple source beam model, called ebm70 (Henzen *et al* 2014), consisting of a foil and a jaw source. The dose calculation in the patient is performed using the MMC. All the utilized beam sources

and dose calculation algorithms are validated against measurements of photon (Magaddino *et al* 2011, Mueller *et al* 2017) and electron beams (Henzen *et al* 2014, Mueller *et al* 2018, 2017).

After final dose calculation, the MU weights of the apertures are re-optimized using a limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm to reduce the degradation from optimized to final dose distribution, called optimization convergence error (OCE) (Jeraj 2002).

2.2. Plan optimization

The goal of the optimization is to minimize the objective function F, which quadratically penalizes deviations of the plan's dose distribution to N_{DV} upper and lower dose-volume objectives (Wu and Mohan 2000) $f_{r,DV}$ weighted by $p_{r,DV}$, N_{gEUD} generalized equivalent uniform dose (gEUD) objectives (Niemierko 1999) $f_{s,gEUD}$ weighted by $p_{s,gEUD}$ and N_{NT} normal tissue objectives (Varian 2017) $f_{q,NT}$ weighted by $p_{q,NT}$:

$$F(D_i) = \sum_{r=1}^{N_{DV}} p_{r,DV} \cdot f_{r,DV}(D_i) + \sum_{s=1}^{N_{gEUD}} p_{s,gEUD} \cdot f_{s,gEUD}(D_i) + \sum_{q=1}^{N_{NT}} p_{q,NT} \cdot f_{q,NT}(D_i) \quad (1)$$

$$f_{r,DV}(D_i) = \frac{1}{V_{str,r}} \sum_{i=1}^{M_r} v_{i,r} \cdot \theta(a_r \cdot (D_i - D_r)) \cdot \theta(a_r \cdot (D(V_r) - D_i)) \cdot (D_i - D_r)^2 \quad (2)$$

$$f_{s,gEUD}(D_i) = \theta(gEUD(t,s) - gEUD_s) \cdot (gEUD(t,s) - gEUD_s)^2 \quad (3)$$

$$f_{q,NT}(D_i) = \frac{1}{V_{str,q}} \sum_{i=1}^{M_q} v_{i,q} \cdot \Theta(D_i - D_{i,q}) \cdot (D_i - D_{i,q})^2 \quad (4)$$

$$gEUD(t,s) = \left(\frac{1}{V_{str,s}} \cdot \sum_{i=1}^{M_s} v_{i,s} \cdot (D_i)^t\right)^{1/t} \quad (5)$$

$$D_{i,q} = \begin{cases} d_0 e^{-b(x_i - x_{start})} + d_\infty (1 - e^{-b(x_i - x_{start})}), & if\ x_i \geq x_{start} \\ d_0, & otherwise \end{cases} \quad (6)$$

where D_i is the dose of the plan to voxel i, $v_{i,r}, v_{i,s}$ and $v_{i,q} \in [0,1]$ are the fraction of voxel i overlapping with the considered structure, $V_{str,r}$, $V_{str,s}$and $V_{str,q}$ are the summed voxel fractions of the voxels overlapping with the considered structure, M_r, M_s and M_q are the number of voxels of the considered structure, θ is the Heaviside function, a_r is equal to 1 and -1 for upper and lower dose-volume objectives, respectively, D_r is the objected dose and $D(V_r)$ is the dose received by at least the

tolerated volume V_r of the considered structure for dose-volume objective r, gEUD(t,s) is the gEUD value with tissue-specific factor t, $gEUD_s$ is the objected gEUD value, $D_{i,q}$ is the objected normal tissue dose to voxel i which is a function of the nearest distance x_i of voxel i to the target, d_0 and d_∞ are the start and end dose parameters, respectively, x_{start} is the start distance of the fall-off and b is the fall-off factor.

Plan dose D_i can be formulated as

$$D_i = \sum_{k \in K} A_{ki} \cdot w_k \tag{7}$$

where A_{ki} is the dose delivered to voxel i with unit MU and w_k is the MU weight of aperture k. K is the set of all deliverable apertures of the provided fields. For clarification, deliverable does not make any statement about the accuracy of the delivered dose. In this work, deliverable just means mechanically deliverable by the MLC. A_{ki} can be further defined as

$$A_{ki} = \sum_{j \in F_k} B_{ji} \cdot t_j(G_{j,L},G_{j,R}) \tag{8}$$

where F_k is the set of all beamlets belonging to the field of aperture k and B_{ji} is the delivered dose to voxel i per unit MU of beamlet j. B_{ji} is calculated by the subprocess 3 of the TPP as described in the previous section 2.1. $t_j(G_{j,L},G_{j,R}) \in [0\%,100\%]$ is the transmission factor for beamlet j given by

$$t_j(G_{j,L},G_{j,R}) = \begin{cases} t_{Open} & ,if\ G_{j,L} \geq 1\ and\ G_{j,R} \geq 1 \quad (a) \\ t_{Open}(G_{j,L} + G_{j,R} - 1) + t_{Tip}(2 - G_{j,L} - G_{j,R}) & ,if\ 0 < G_{j,L} < 1\ and\ 0 < G_{j,R} < 1 \quad (b) \\ t_{Open}G_{j,L} + t_{Tip}(1 - G_{j,L}) & ,if\ 0 \leq G_{j,L} < 1\ and\ G_{j,R} > 1 \quad (c) \\ t_{Open}G_{j,R} + t_{Tip}(1 - G_{j,R}) & ,if\ 0 \leq G_{j,R} < 1\ and\ G_{j,L} > 1 \quad (d) \\ t_{Tip}(G_{j,L} + 1) - t_{MLC}G_{j,L} & ,if -1 < G_{j,L} < 0 \quad (e) \\ t_{Tip}(G_{j,R} + 1) - t_{MLC}G_{j,R} & ,if -1 < G_{j,R} < 0 \quad (f) \\ t_{MLC} & ,else \quad (g) \end{cases} \tag{9}$$

where $G_{j,L} = C_{j,L} - P_{k,j,L}$ and $G_{j,L} = C_{j,R} - P_{k,j,R}$. $C_{j,L}$ and $C_{j,R}$ are the position of beamlet j counted from the left and right beamlet grid border, respectively, and $P_{k,j,L}$and $P_{k,j,R}$ are the left and right leaf position of the leaf pair in aperture k in the line of beamlet j. t_{Open} is 100%, t_{Tip} is 12.9% (photon beam) and 0% (electron beam) and t_{MLC} is 1.3% (photon beam) and 0% (electron beam). These

transmission factors of photon beams are an empirically determined model to approximate the transmission through the MLC (t_{MLC}) and increased transmission through the leaf tip (t_{Tip}). This model aims to reduce the dose prediction error (DPE) and therefore, the OCE (Jeraj 2002, Bergman *et al* 2006, Men *et al* 2007, Mueller *et al* 2017). In other words, the discrepancy between optimized and final dose distribution is reduced, because the transmission of photon beams through the MLC is already considered during optimization. Beside of that, the transmission model defined above also allows for leaf positions between beamlet borders. The calculation of $t_j(G_{j,L},G_{j,R})$ is illustrated in Figure 2.

Note that equation 7 must be understood as a sum over all possible deliverable apertures. Apertures which are part of the aperture pool have a weight $w_k > 0$ and all other apertures have a weight of 0. The H-DAO algorithm starts with an empty aperture pool, iteratively adds apertures to the pool and optimizes the shapes and MU weights of the apertures at each iteration to minimize F given in equation 1. Figure 1 (middle) illustrates the workflow of the H-DAO algorithm. In the following the details of one iteration are described. One iteration creates as many branches as there are fields provided (branch-feature). The subprocess of these branches are performed in parallel in multiple computer threads. For the current work we used an AMD Epyc2 CPU featuring 2x64 CPU cores to have enough cores such that no thread needs to share a CPU core with another thread. A specific optimization example of the H-DAO algorithm workflow illustrating the concept of parallel branches is shown in Figure 1 (bottom). As visible there, a branch consists of four sequential subprocesses. First, the pricing is performed to find the most promising aperture shape of the field considered in this branch. This determination of the most promising aperture per field is the same as for the CG-DAO (Romeijn *et al* 2005, Men *et al* 2007). The most promising aperture is a set of beamlets that can be translated to a deliverable aperture and for which the summed gradient components of the beamlets not covered by the MLC on the objective function, called price, is minimal. The optimization problem to find the most promising aperture of a field can be formulated by

$$min_{k \in K}\left(\frac{\partial F(D_i)}{\partial w_k}\right) \quad (10)$$

For derivation and description of strategies to solve this problem, the work of Romeijn *et al* 2005 and Men *et al* 2007 is referenced here.

After determination of the most promising aperture, it is added to a copy of the current aperture pool. It follows an MU weight optimization of the apertures using a projected L-BFGS two loop recursion (Nocedal and Wright 2013, Bangert 2011), which is a quasi-Newton algorithm approximating the product of the inverse Hessian and the gradient. The length of the L-BFGS history to store objective function values and gradient values is set to 4. The algorithm is combined with a line search to find an appropriate step length satisfying Armijo's Rule. If any aperture MU weight is below 0 during line search, the MU weight is projected to 0. The L-BFGS terminates if the objective function value is not lowered more than 0.01% in three consecutive iteration.

The next subprocess is the SA-feature, which simultaneously optimizes shapes and weights of the apertures in the pool including the recently added aperture according to an SA cooling schedule. The SA-feature runs maximally for a total of 5000 iterations. It stops earlier, if the objective function value did not decrease more than 0.1% for 250 consecutive iterations. At every iteration, an aperture of the pool is selected randomly for which its shape or MU weight is changed with a probability of P_S or (1 - P_S), respectively. The change is accepted if the objective function value is decreased or otherwise with a probability of

$$P = 2 \cdot P_0 \cdot \frac{1}{1 + e^{\frac{log\left(\frac{n_S + n_W}{N_A} + 1\right)}{T_P}}} \tag{11}$$

where T_P is the cooling rate, P_0 is the initial value of P, N_A is the number of apertures and n_S and n_W are the number of previous total accepted shape and MU weight changes, respectively. If the shape of the aperture is aimed to be changed, a leaf is randomly selected, and its position is randomly changed according to a normal distribution around the current leaf position and a width of

$$\sigma_S = 1 + (\sigma_{S0} - 1) \cdot e^{-\frac{log\left(\frac{n_S}{N_L} + 1\right)}{T_S}} \tag{12}$$

in units of number of beamlets. T_S is the cooling rate and σ_{S0} is the initial width of the normal distribution and N_L is the total number of leaf pairs of all apertures in the pool. In case of a weight change, the weight is changed according to a normal distribution around the current aperture weight w_k and a width of

$$\sigma_W = 0.01 + (\sigma_{W0} - 0.01) \cdot e^{-\frac{\log\left(\frac{n_W}{N_A}+1\right)}{T_W}} \quad (13)$$

in relative units of w_k. T_W is the cooling rate and σ_{W0} is the initial width of the normal distribution. Following parameter values are used in this work: $P_S = 90\%$, $T_P = 1$, $P_0 = 3.5\%$, $T_S = 2$, $\sigma_{S0} = 0.3$, $T_W = 1$, $\sigma_{W0} = 0.1$. Note that σ_{S0} is only 0.3 beamlets wide. Thus, most of the shape changes are small such that the SA-feature follows its purpose of a leaf refinement.

When the SA-feature terminates, the MU weights of the apertures are again optimized using the same L-BFGS implementation. Due to performing an MU weight optimization before and after performing the SA-feature, the SA-feature starts with a better initial solution and the need for more iterations of the SA is also smaller. Note that the MU optimization is convex and involves many optimization variables less (i.e. leaf positions) than the whole DAO optimization problem. Thus, it can be solved very efficiently with the L-BFGS

When all branches are performed, the objective function value is evaluated for each branch and the aperture pool for the next iteration of the H-DAO is the one with the lowest objective function value among all branches. These H-DAO iterations are repeated until the desired number of apertures in the aperture pool is reached.

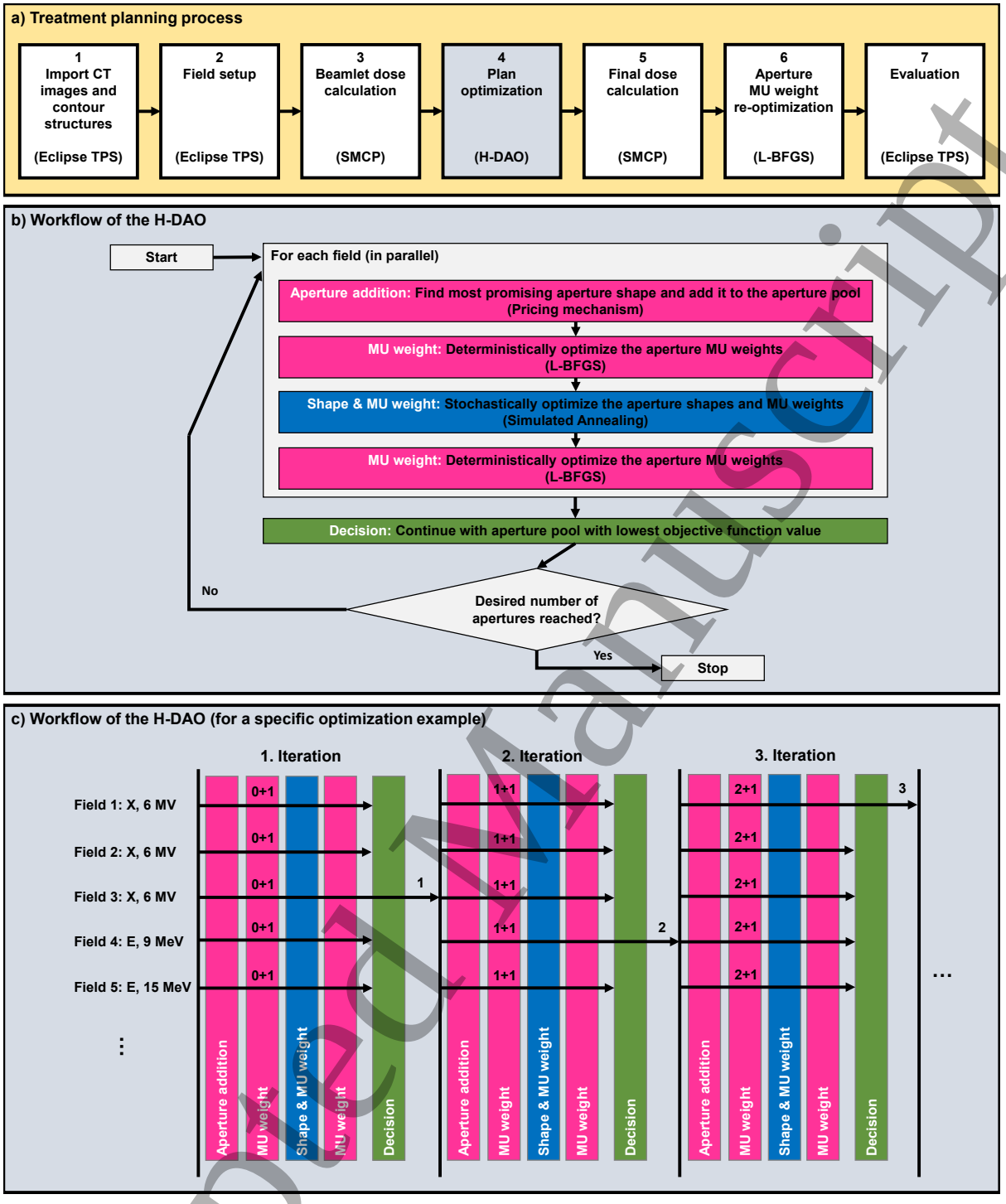

Figure 1: a): TPP in which the H-DAO is embedded. b): General workflow of the H-DAO. While the processes with magenta color are part of the CG-DAO algorithm, the processes with blue and green color are extensions described in this work. c): Again, the workflow of the H-DAO though presented for a specific optimization example to better illustrate the parallel execution. Each arrow represents a branch run in a separate computer thread. The numbers stand for the number of apertures currently present in the aperture pool.

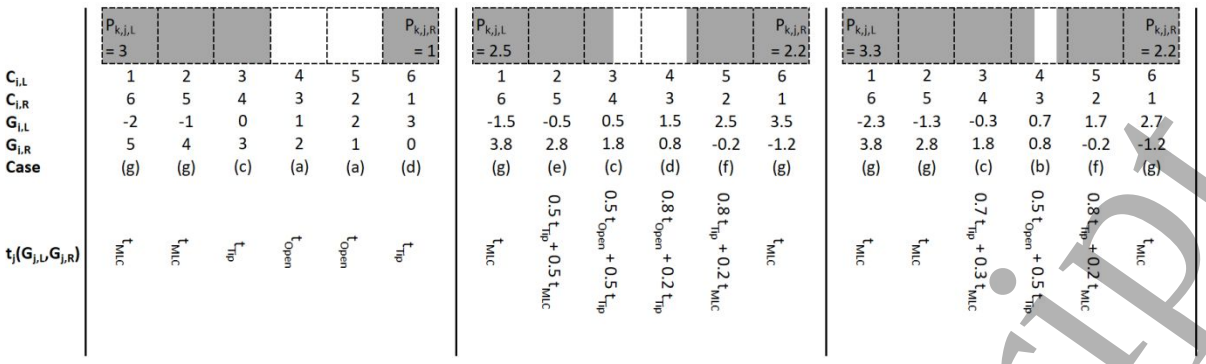

Figure 2: Illustration of the transmission calculation for three example leaf pairs (left, center and right). The gray areas are the areas blocked by the leaves and the dashed grid represents the beamlet grid seen from beams eye view. The cases (a)-(g) are related to equation 9.

2.3. Academic and clinical cases

In the computational study of this work, treatment plans of the treatment techniques IMRT and MBRT are generated for an academic case (Mueller *et al* 2017), a brain case and a head and neck case. Motivation behind the selection of these three cases and the corresponding field setup illustrated in Figure 3 is to have an increasing complexity in the geometrical treatment situation and number of fields. The collimator rotation of all the fields is 0°, except the photon fields for the two clinical cases, which have a collimator rotation such that the maximal field width parallel to leaf movement direction is minimized. Reason for this is to get a field width smaller than 15 cm, which is the maximal leaf separation distance between leaves of the same MLC bank.

The academic case is a cylindric homogeneous water phantom with a radius of 10 cm to be treated with a prescribed median dose of 50 Gy to the PTV in 25 fractions. A superficially located PTV including a deep-seated part and two OARs, OAR-lateral and OAR-distal, are contoured. These three structures are extended perpendicular to the transversal plane by 7.4 cm. The first clinical case is a glioblastoma brain case to be treated with a prescribed median dose of 60 Gy to the PTV in 30 fractions. The second clinical case is an oropharynx head and neck case to be treated with a prescribed D95% of 50 Gy to the PTV in 25 fractions. For each case, every optimization used the same dose objectives listed in Table 1.

Table 1: Objectives used for all the optimizations performed in the computational study. The parameters for the normal-tissue objective have the same values for every case: Start dose d_0 is 95%, end dose d_∞ is 10%, start distance x_{start} of the fall-off is 0.5 cm and fall-off factor b is 0.15. Note that gEUD(t=1) is equivalent to the mean dose.

Case	Structure	Objective type	Priority factor	Dose (%)	Volume (%)
Academic	PTV	Upper dose-volume	700	101	0
		Lower dose-volume	700	99	100
	OAR-distal	gEUD(t=1)	20	15	-
	OAR-lateral	gEUD(t=1)	20	15	-
	Normal tissue	Upper dose-volume	700	100	0
		Normal-tissue objective	50	See caption	-
Brain	PTV	Upper dose-volume	700	101	0
		Lower dose-volume	700	99	100
	Ipsilateral eye	Upper dose-volume	20	5	0
	Contralateral eye	Upper dose-volume	20	0	0
	Ipsilateral lens	Upper dose-volume	20	0	0
	Contralateral lens	Upper dose-volume	20	0	0
	Ipsilateral optical nerve	Upper dose-volume	20	40	0
	Contralateral optical nerve	Upper dose-volume	20	0	0
	Chiasma	Upper dose-volume	20	35	0
	Brainstem	Upper dose-volume	100	65	0
	Brain	gEUD(t=1)	20	18	-
	Normal tissue	Upper dose-volume	40000	98	0
		Normal-tissue objective	20	See caption	-
Head and neck	PTV	Upper dose-volume	5000	105	0
		Lower dose-volume	5000	100	100
		Lower dose-volume	5000	95	100
		Lower dose-volume	5000	100	95
	Brainstem	Upper dose-volume	20	20	0
	Spinal cord	Upper dose-volume	20	35	0
	Mandibula	Upper dose-volume	20	100	0
	Oral cavity	gEUD(t=1)	20	55	-
	Pharyngeal constrictors	gEUD(t=1)	20	35	-

Larynx	gEUD(t=1)	20	45	-
Esophagus	gEUD(t=1)	20	25	-
Left Parotid gland	gEUD(t=1)	20	55	-
Right Parotid gland	gEUD(t=1)	20	0	-
Right submandibular gland	gEUD(t=1)	20	7	-
Lips	gEUD(t=1)	20	17	-
Normal tissue	Upper dose-volume	50000	98	0
	Normal-tissue objective	30	See caption	-

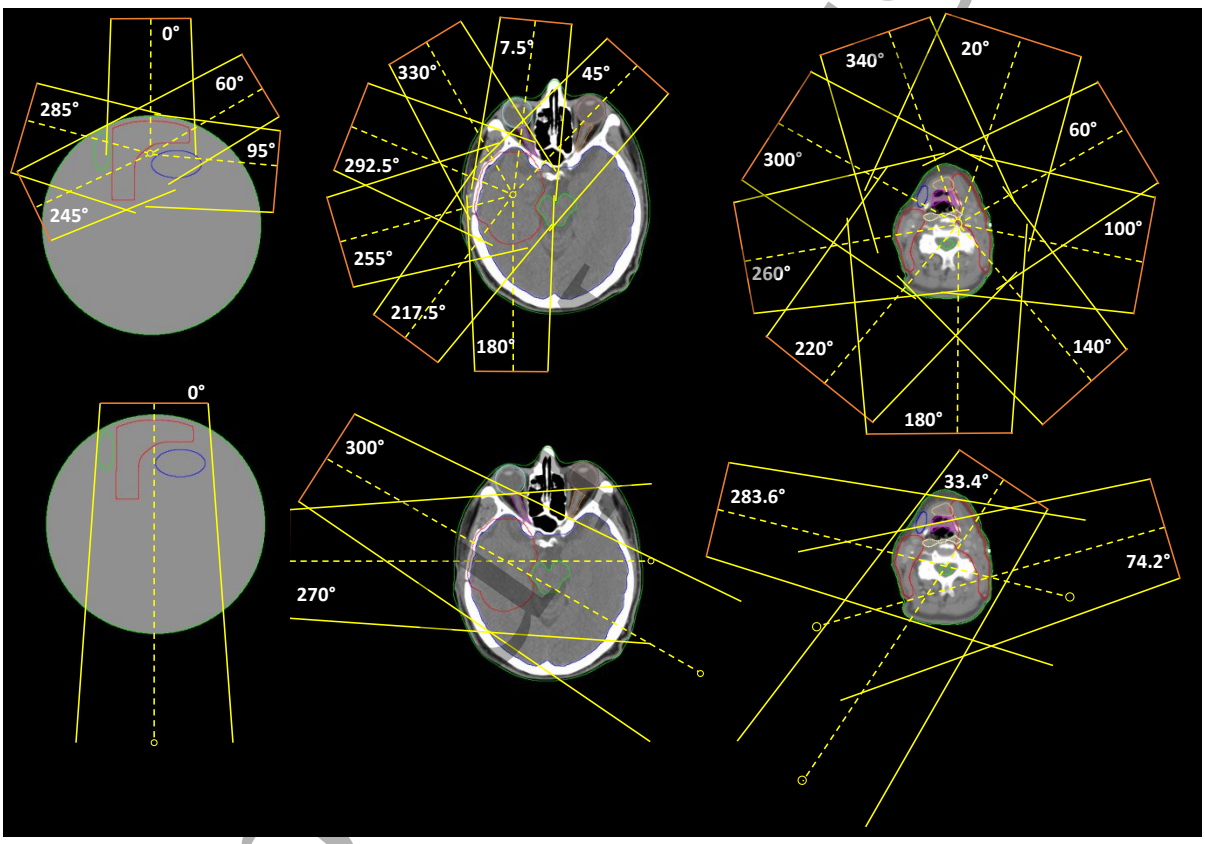

Figure 3: Photon (top row) and electron (bottom row) field setup for an academic case (left column), a brain case (center column) and a head and neck case (right column) used to create IMRT and MBRT treatment plans. Following structures are visible on the transversal views: PTV (red), OAR-distal (blue) and OAR-lateral (green) for the academic case, PTV (red), brainstem (green), brain (blue), left (cyan) and right (brown) eyes and left (magenta) and right (orange) optic nerves for the brain case, PTV (red), spinal cord (green), right submandibular gland (blue), oral cavity (light brown), larynx (magenta) and pharyngeal constrictors (light yellow) for the head and neck case.

2.4. Computational study

A computational study is performed to analyze the influence of the two features added to the CG-DAO algorithm, i.e. the branch- and SA-features, on the following optimization results and properties.

1. Convergence behavior with number of apertures
2. Final dose distribution
3. Convergence error (OCE)
4. Plan complexity
5. Field contributions
6. Computational performance
7. Statistical uncertainty

In this computational study, treatment plans of the treatment techniques IMRT and MBRT are generated for the introduced academic and clinical cases using the optimization algorithms listed in Table 2. We investigated both IMRT and MBRT to find out if the complexity of the field setup plays any role in the algorithm performance. Due to the choice of two particle types and multiple beam energies, the complexity of the field setup for MBRT can be judged substantially higher. In particular, the fields provided to IMRT are a subset of the fields provided to MBRT, because the same photon fields are provided to both IMRT and MBRT.

Table 2: The DAO algorithms compared in the computational study. The columns "Branch-feature" and "SA-feature" indicate whether the listed algorithms utilize these features.

Name	Branch-feature	SA-feature
CG-DAO	No	No
CG-DAO_Branch	Yes	No
CG-DAO_SA	No	Yes
H-DAO	Yes	Yes

To study the convergence with number of apertures of the investigated DAO algorithms, the objective function value after plan optimization (subprocess 4 in the TPP) is collected as a function of the apertures for up to 200 apertures assuming that more than 200 apertures would only lead to marginal improvements in plan quality. Furthermore, an FMO is performed with the L-BFGS using the same beamlets without any constraints on smoothness of the fluence map. This FMO optimized plan is used as an ideal benchmark. This is done for IMRT and MBRT for all three cases.

To also see the differences between the different DAO algorithms in the dosimetric space for a specific number of apertures, the TPP is further performed to the end utilizing 50 (academic case), 100 (brain case) and 150 (head and neck case) apertures for all DAO algorithms investigated and the final dose distribution is evaluated by dose-volume-histograms (DVHs) and following dosimetric quantities: PTV dose homogeneity index, mean dose to normal tissue D_{mean}^{NT}, average mean dose to parallel OARs $\hat{D}_{mean}$ and average $D_{2\%}$ to serial OARs $\hat{D}_{2\%}$. The OARs with a gEUD(t=1) objective are considered as parallel and the others as serial for the scope of this study. The PTV dose homogeneity index HI is defined as

$$HI = \frac{D_{2\%} - D_{98\%}}{D_p} \tag{14}$$

where D_p is the prescribed dose and $D_{2\%}$ and $D_{98}\%$ the dose receiving at least 2% and 98% of the PTV volume.

Using the results of these final re-optimized plans, OCE, plan complexity described in numbers of MU and average aperture area $\hat{A}$, field contributions (only academic case) and the computation time are also evaluated. The OCE is calculated as

$$OCE\ (\%) = \frac{F_F - F_O}{F_O} \cdot 100\% \tag{15}$$

where F_O is the objective function value after optimization (subprocess 4 of the TPP) and F_F is the objective function value after aperture MU weight re-optimization of the final dose distributions (subprocess 6 of the TPP). Connected to the investigation of the OCE, it is evaluated how a potential improvement in terms of objective function value after optimization of the extended DAO algorithms over the CG-DAO evolves after the aperture MU weight re-optimization. Therefore, the improvement after optimization is calculated by

$$\Delta F_o = \frac{F_O - F_O^{CG-DAO}}{F_O^{CG-DAO}} \cdot 100\% \tag{16}$$

and after aperture MU weight re-optimization by

$$\Delta F_F = \frac{F_F - F_F^{CG-DAO}}{F_F^{CG-DAO}} \cdot 100\% \quad (17)$$

where F_O^{CG-DAO} and F_F^{CG-DAO} are the objective function value of the reference CG-DAO after optimization and after aperture MU weight re-optimization, respectively. F_O and F_F are the analogous values for the DAO algorithm for which ΔF_o and ΔF_F are calculated for.

A field contribution is defined as the fraction of the PTV mean dose delivered by all the apertures belonging to the corresponding field. In case of MBRT, also photon and electron contributions are analyzed, which is the sum of all photon and electron field contributions, respectively.

To study the statistical uncertainty on the objective function value and the DVH of the DAO algorithms utilizing the SA-feature, the optimizations for the academic case utilizing 50 apertures are repeated 100 times using different seeds to initialize the random number generator.

3. Results

3.1. Convergence behavior with number of apertures

Figure 4 shows the convergence behavior of the objective function value as a function of the number of apertures for all the investigated DAO algorithms. It is visible that for each combination of case and treatment technique (IMRT or MBRT), the fastest convergence is always given by H-DAO, followed by CG-DAO_SA, CG-DAO_Branch and CG-DAO. All the DAO algorithms do not converge to the value given by the FMO, because all the DAO algorithms consider transmission through the MLC in case of photon apertures in contrast to FMO.

As the fields provided to IMRT is a subset of the fields provided to MBRT, it is reasonable to hypothesize that the objective function value for MBRT is at least as good as for IMRT. This is true for all the DAO algorithms for at least 10 apertures with the following two exceptions given for the head and neck case: CG-DAO and CG-DAO_SA.

It is noteworthy how many apertures the H-DAO could use less than the CG-DAO to reach the same objective function value, e.g. when the number of apertures of 50, 100 and 150 are selected for the

CG-DAO for the academic, brain and head and neck case, respectively, the H-DAO needs on average 44.3% (IMRT) and 55.3% (MBRT) less apertures to reach the same objective function value as the CG-DAO.

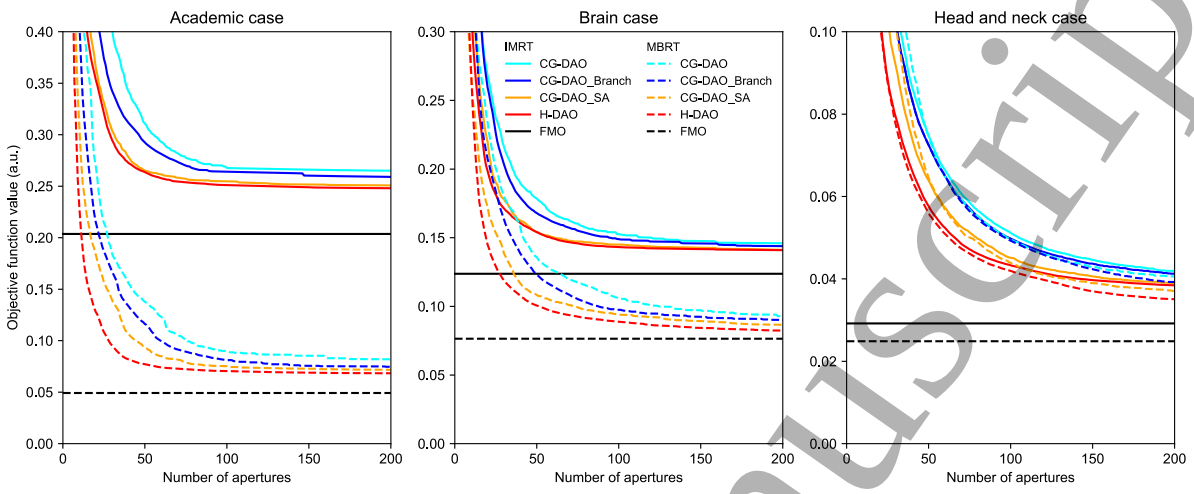

Figure 4: Convergence behavior of the objective function value as a function of the number of apertures created for all the investigated DAO algorithms applied to an academic case (left), a brain case (center) and a head and neck case (right). As the FMO does not generate apertures, its resulting objective function value is indicated as a constant line.

3.2. Specific number of apertures

Figure 5 compares DVHs and Table 3 dosimetric quantities of the IMRT and MBRT plans optimized with the different DAO algorithms after performing the whole TPP, i.e. including final dose calculation and MU weight re-optimization. For these plans with a specific number of apertures of 50 (academic case), 100 (brain case) and 150 (head and neck case), the DVHs confirm the finding of the convergence behavior analysis that H-DAO performed best followed by CG-DAO_SA, CG-DAO_Branch and CG-DAO. Only the PTV and selected OARs are shown in the DVH comparison for better visibility, but the mentioned finding is in general also confirmed by the DVHs of other structures considered during optimization.

Table 3 also shows that the OCE is higher for the DAO algorithms using the introduced branch- and SA-feature compared to the CG-DAO. However, the OCE is typically small enough such that the improvement after optimization ΔF_o over the CG-DAO is not vanished after aperture weight re-optimization. In case of H-DAO, the improvement ΔF_F is reduced by 15.4% (IMRT) and 14.5% (MBRT) compared to ΔF_O averaged over the three investigated academic and clinical cases.

Regarding plan complexity, Table 3 shows that the branch- and SA-features lead to smaller apertures and connected to it, also to higher number of MUs. This is true for every investigated case and both IMRT and MBRT. In case of H-DAO, the aperture areas are 21.8% (IMRT) and 31.2% (MBRT) smaller and the MUs are 11.3% (IMRT) and 30.1% (MBRT) higher compared to CG-DAO averaged over the three investigated academic and clinical cases.

Averaged over the three investigated academic and clinical cases, the CG-DAO_Branch required 19.3% (IMRT) and 27.4% (MBRT) longer computation time than the CG-DAO. Thus, the computation time is not substantially increased due to the use of a multi-core CPU being able to perform the branches in parallel. Again, averaged over the three cases and compared to the CG-DAO, CG-DAO_SA required 146.8% (IMRT) and 110.9% (MBRT) longer computation time and the H-DAO required 179.2% (IMRT) and 201.8% (MBRT) longer computation time.

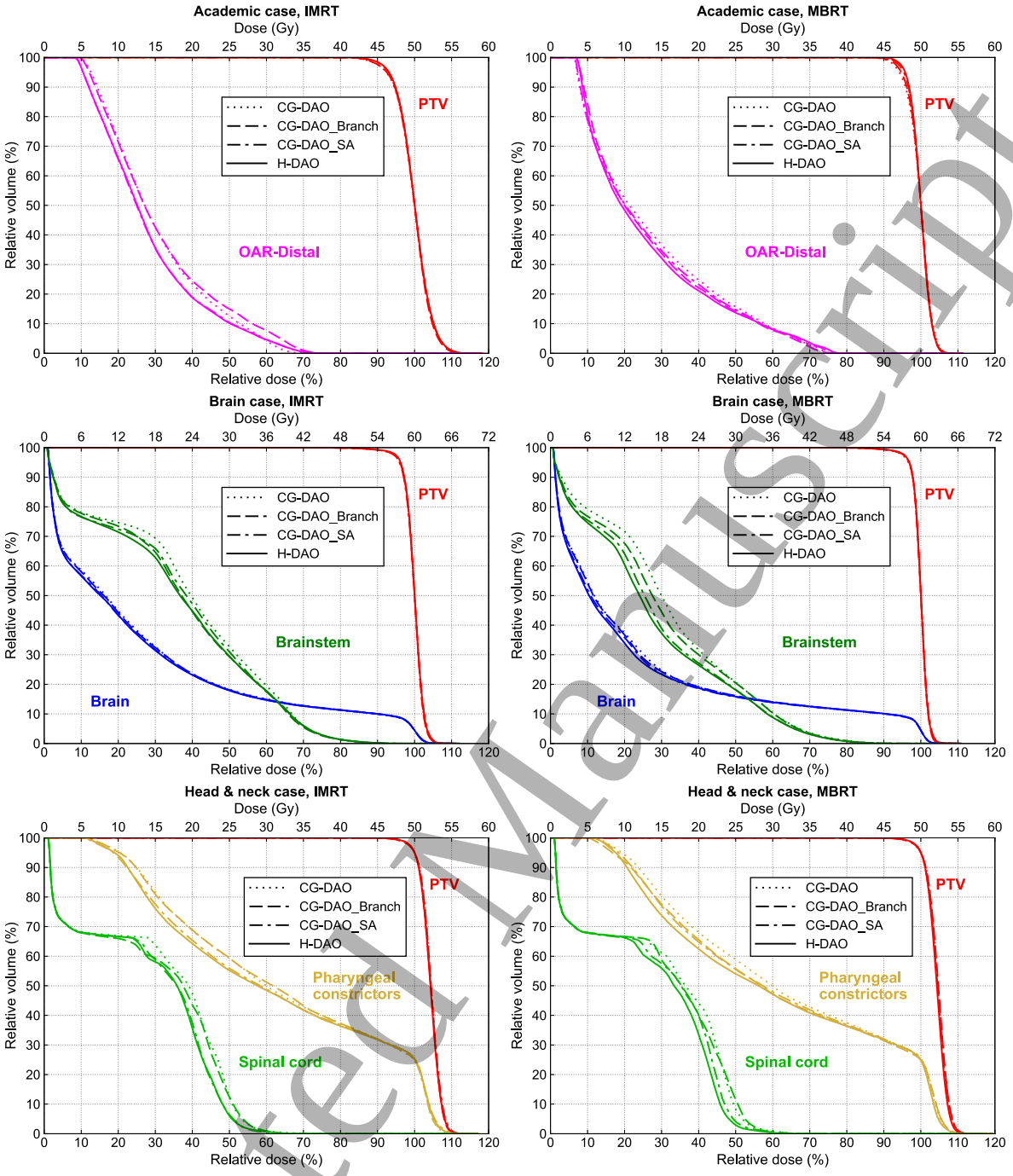

Figure 5: DVHs for the PTV and selected OARs of the IMRT (left column) and MBRT (right column) plans for the academic case with 50 apertures (top row), the brain case with 100 apertures (middle row) and the head and neck case with 150 apertures (bottom row) optimized by the different DAO algorithms.

Table 3: Dosimetric quantities and several plan characteristics are compared between the different DAO algorithms applied to IMRT and MBRT for the academic, brain and head and neck cases.

	IMRT	MBRT

	H-DAO	CG-DAO_SA	CG-DAO_Branch	CG-DAO	H-DAO	CG-DAO_SA	CG-DAO_Branch	CG-DAO
Academic case								
HI	0.179	0.182	0.197	0.192	0.106	0.111	0.118	0.129
D_{mean}^{NT} (Gy)	3.0	3.0	3.1	3.0	2.7	2.6	2.9	2.7
$\hat{D}_{mean}$ (Gy)	14.1	14.1	15.0	15.1	12.0	12.8	13.0	13.7
Electron contr. (%)	-	-	-	-	55.5	60.9	48.1	61.1
OCE (%)	11.6	12.3	10.5	9.2	41.8	37.6	29.3	27.5
ΔF_o (%)	15.2	14.7	5.9	-	43.9	32.3	15.6	-
ΔF_F (%)	13.3	12.3	4.8	-	37.6	26.9	14.4	-
MUs	884.4	859.3	809.2	785.5	925.9	838.5	799.6	751.3
$\hat{A}$ (cm^2)	17.6	19.1	21.0	23.9	22.9	31.9	29.8	38.8
Computation time (s)	94.9	81.3	62.4	60.6	107.2	92.5	76.9	77.1
Brain case								
HI	0.119	0.126	0.122	0.123	0.102	0.105	0.102	0.109
D_{mean}^{NT} (Gy)	5.3	5.3	5.5	5.5	4.9	4.9	5.1	5.1
$\hat{D}_{mean}$ (Gy)	16.0	16.1	16.3	16.4	23.7	23.9	24.4	24.6
$\hat{D}_{2\%}$ (Gy)	16.9	16.9	17.2	17.4	16.1	15.8	16.1	16.4
Electron contr. (%)	-	-	-	-	53.8	50.3	48.9	45.9
OCE (%)	23.1	25.1	21.3	20.6	21.5	17.0	18.2	16.1
ΔF_o (%)	6.1	4.9	2.3	-	16.1	11.1	8.0	-
ΔF_F (%)	4.2	1.4	1.7	-	12.2	10.4	6.3	-
MUs	623.7	582.4	583.6	569.7	672.0	578.1	591.4	509.4
$\hat{A}$ (cm^2)	13.9	15.4	16.0	16.8	14.8	16.3	18.0	19.1
Computation time (s)	187.0	146.9	79.1	61.5	196.8	154.4	80.0	71.7
Head and neck case								
HI	0.107	0.111	0.109	0.112	0.107	0.113	0.115	0.113

D_{mean}^{NT} (Gy)	9.6	9.6	9.8	9.8	9.4	9.5	9.4	9.6
$\hat{D}_{mean}$ (Gy)	22.8	23.0	23.4	23.6	22.7	22.9	22.8	23.5
$\hat{D}_{2\%}$ (Gy)	33.2	33.2	33.6	33.7	31.7	32.3	32.8	33.2
Electron contr. (%)	-	-	-	-	14.7	12.0	6.9	11.0
OCE (%)	22.3	23.0	20.6	20.7	21.4	23.3	21.7	20.8
ΔF_o (%)	11.0	9.0	2.0	-	13.8	9.1	1.3	-
ΔF_F (%)	9.8	7.3	2.1	-	13.3	7.2	0.5	-
MUs	682.2	676.0	643.4	609.3	669.0	572.5	548.1	495.0
$\hat{A}$ (cm^2)	40.8	44.9	49.6	52.1	44.9	55.0	50.6	64.2
Computation time (s)	1264.7	1233.1	424.0	335.8	2333.8	1410.8	810.7	474.6

Figure 6 compares the field contributions and photon and electron contribution for the plans generated by the different DAO algorithms utilizing 50 apertures for the academic case. In case of IMRT, the variations between the different DAO algorithms in field contributions are only a few percent. The field contributions also match well with those received by the FMO algorithm. For MBRT, the field contributions vary more between the different DAO algorithms, but they are still within 11% of the FMO field contributions. Overall, the field contributions of the H-DAO match closest with those of the FMO. The same is true for the photon and electron contribution.

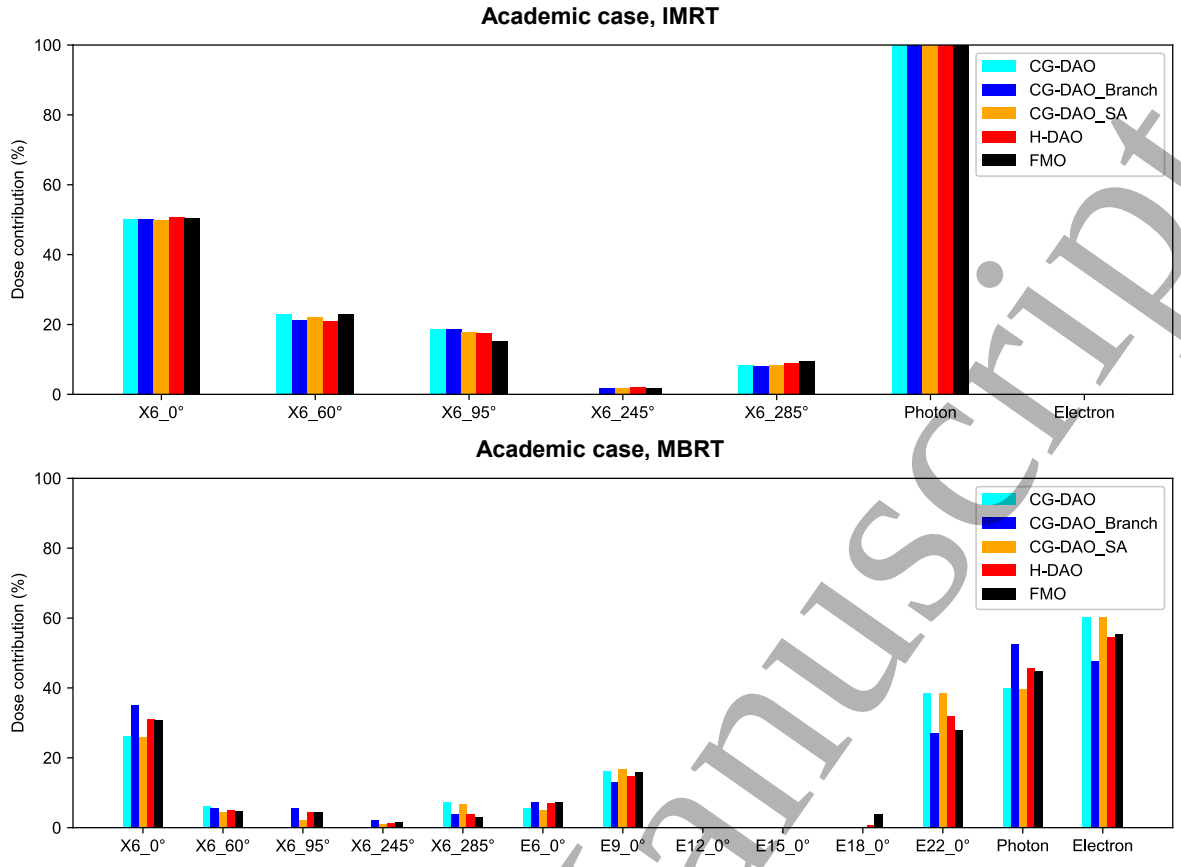

Figure 6: Dose contributions of the single fields and photon and electron dose contributions for the DAO plans s determined for the academic case utilizing 50 apertures and the corresponding FMO plan.

The statistical uncertainty of the CG-DAO_SA and H-DAO due to the seed to initialize the random number generator for the SA-feature is demonstrated in Figure 7. The standard deviations of the objective function value distributions are 0.0045 (IMRT) and 0.0079 (MBRT) for CG-DAO_SA and 0.0033 (IMRT) and 0.0028 (MBRT) for H-DAO. Thus, the branch-feature included in the H-DAO seems to lead to a lower statistical uncertainty. The dosimetric differences between the worst and best optimization run out of 100 runs are demonstrated in Figure 7 for the H-DAO applied to IMRT.

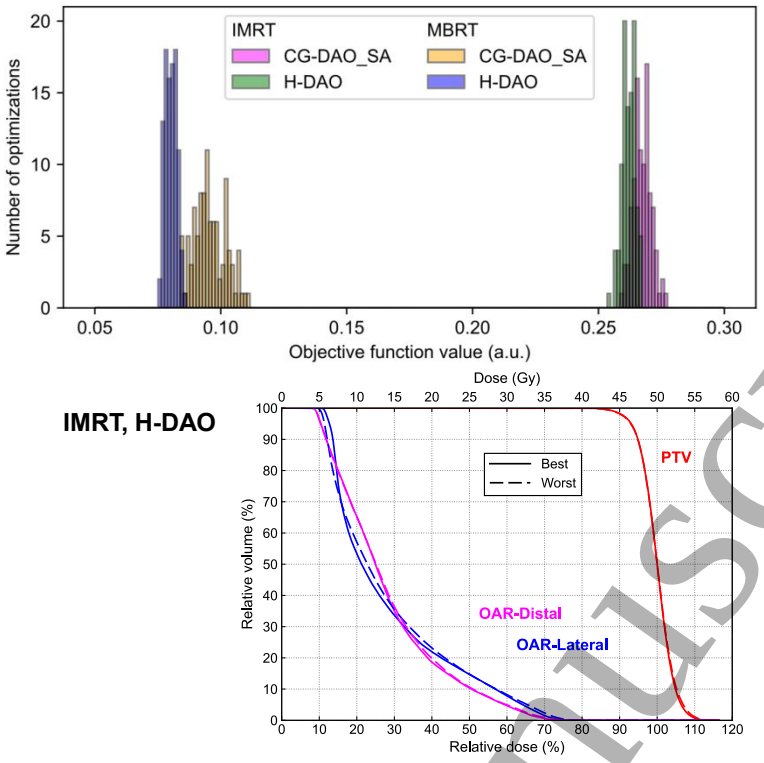

Figure 7: Top: Histograms of the objective function values received for the IMRT and MBRT plans generated with the CG-DAO_SA and H-DAO utilizing different seeds for the academic case. Bottom: DVH comparison of the two optimizations leading to the best and the worst result in terms of objective function value of the distribution shown for IMRT, H-DAO shown at the top (green histogram).

4. Discussion

The H-DAO algorithm extending the CG-DAO with the branch- and SA-features is successfully implemented and investigated. Each feature alone leads to a faster convergence with number of apertures than the CG-DAO and the combination of the two features also lead to an additional benefit in fast convergence. These statements are generally true for both IMRT and MBRT and for all three investigated cases. The faster convergence with number of apertures can be exploited to create treatment plans with higher dosimetric plan quality or with reduced number of apertures leading to shorter plan delivery time.

Interestingly, CG-DAO and CG-DAO_SA showed a slower convergence with number of apertures for IMRT than for MBRT for the head and neck case up to about 50 apertures, even though the fields provided to IMRT are a subset to those provided to MBRT. This was not observed for CG-DAO_Branch and H-DAO, which both use the branch-feature. Thus, utilizing the branch-feature fulfills better that the provided fields are exploited and not overburdening the DAO algorithm.

Therefore, the branch-feature is promising to be also used as a beam angle optimization feature for any CG based DAO algorithm, similarly as it was used for FMO algorithms (Papp *et al* 2015). Other treatment techniques than MBRT that also handle many fields such as 4Pi radiotherapy or treatment techniques including path-finding such as dynamic trajectory radiotherapy (DTRT) (Smyth *et al* 2019, Papp *et al* 2015, Fix *et al* 2018) could benefit from this feature.

Both branch- and SA-feature lead to smaller apertures. Thus, the finding by Cassioli and Unkelbach 2013 that the CG-DAO leads to large apertures, whose MU weight can only be reduced by generating more other apertures, is confirmed. At least for the SA-feature this was expected as it has the purpose of a leaf refinement similar to the aperture shape optimization (ASO) (Cassioli and Unkelbach 2013). Through the evaluation of the OCE, ΔF_o and ΔF_F, it was shown that the smaller apertures can be handled successfully through the whole TPP without substantial deterioration of the objective function value. Main reason for this is a feature considering transmission through the leaves and increased transmission through the leaf ends during optimization. The plans are also deemed to be accurately deliverable by the treatment unit with high accuracy in the delivered dose as this was shown by a recent work of Heath et al. about robust optimized MBRT utilizing the H-DAO algorithm for plan generation (Heath *et al* 2021). There, the delivery of MBRT plans was validated by dose measurements using gafchromic film placed in an anthropomorphic phantom. However, any uncertainties in the modeling of the MLC would be amplified by smaller apertures making an accurate consideration of the treatment head necessary. Furthermore, smaller apertures are usually connected to an increase in MUs and therefore to an increase in delivery time. However, only a minor increase in delivery time is expected when switching from CG-DAO to H-DAO, e.g. for the plans with equal number of apertures shown in this work, the additional delivery time for H-DAO compared to CG-DAO is expected to be about 12 s with 600 MU/min due to the additional 120 MUs on average.

The additional computational effort to perform the SA-feature is substantial and even enormous in case of the branch-feature. However, the branch-feature can be easily parallelized by running each branch in a separate thread like this is done in this computational study. Utilizing state-of-the-art CPUs with high number of CPU-cores are of high value for this task. The results of this

computational study show, that in the worst case the computation time to perform the optimization using H-DAO instead of CG-DAO was not more than 3.9 times increased. Multiple techniques were recently investigated to enhance computational efficiency for DAO that could be combined with the H-DAO.

- Yang *et al* 2018 replaced the original pricing mechanism with a combination of noise cancellation of the prices using a fuzzy controller followed by aperture generation using threshold segmentation. This allowed to reduce optimization time by 58.61%. This technique could be also used in combination with the H-DAO to determine the most promising aperture per field.
- Men *et al* 2009 developed a GPU based implementation of the column generation leading to optimization times below 3.8 s. A GPU implementation would be compliant with the branch-feature and the SA-feature could also be implemented to be performed on the GPU.
- MacFarlane *et al* 2019 reformulated the objective function using a second order Taylor series expansion allowing to find the global minimum by a fast matrix inversion. They applied it to a gradient-based optimization algorithm with at least 70-200 times faster execution and noted that it could be also applied to SA.
- Renaud *et al* 2017 applied the CG-DAO to MBRT and investigated different aperture addition-schemes adding multiple apertures per iteration. These aperture addition-schemes could be combined with the SA-feature but not directly with the branch-feature presented in this work.

A limitation of this work is the small number of test cases, which does not allow to state any treatment site specific benefits. The H-DAO algorithm could be tuned scenario specific leading potentially to further improvements, e.g. using different parameter values for the SA-feature depending on the treatment site or by adapting the parameter values according to characteristics of the clinical case such as the target size.

5. Conclusions

The H-DAO algorithm is successfully developed. It extends the CG-DAO algorithm by the branch-feature acting as a more founded decision on the aperture to be added to the aperture pool and by the SA-feature acting as a leaf refinement. This computational study shows that both features lead to a faster convergence of the objective function value with number of apertures. This allows to select a better compromise between dosimetric plan quality and number of apertures.

Acknowledgments

This work was supported by grant 200021_185366 of the Swiss National Science Foundation and by Varian Medical Systems. Calculations were performed on UBELIX (http://www.id.unibe.ch/hpc), the HPC cluster at the University of Bern.

References

Bangert M 2011 New concepts for beam angle selection in IMRT treatment planning 113 S.

Bergman A M, Bush K, Milette M-P, Popescu I A, Otto K and Duzenli C 2006 Direct aperture optimization for IMRT using Monte Carlo generated beamlets. *Med. Phys.* **33** 3666–79 Online: https://dx.doi.org/10.1118/1.2336509

Bortfeld T 2006 IMRT: a review and preview *Phys. Med. Biol.* **51** R363–79 Online: https://doi.org/10.1088/0031-9155/51/13/R21

Carlsson F 2008 Combining segment generation with direct step-and-shoot optimization in intensity-modulated radiation therapy *Med. Phys.* **35** 3828–38

Cassioli A and Unkelbach J 2013 Aperture shape optimization for IMRT treatment planning *Phys. Med. Biol.* **58** 301–18 Online: http://www.ncbi.nlm.nih.gov/pubmed/23257284

Convery D J and Webb S 1992 The generation of intensity-modulated fields for conformal radiotherapy by dynamic collimation *Phys. Med. Biol.* **37** 1359–74 Online: https://doi.org/10.1088/0031-9155/37/6/012

Dong P, Lee P, Ruan D, Long T, Romeijn E, Yang Y, Low D, Kupelian P and Sheng K 2013 4π non-

coplanar liver SBRT: A novel delivery technique *Int. J. Radiat. Oncol. Biol. Phys.* **85** 1360–6 Online: http://dx.doi.org/10.1016/j.ijrobp.2012.09.028

Fix M K, Cygler J, Frei D, Volken W, Neuenschwander H, Born E J and Manser P 2013 Generalized eMC implementation for Monte Carlo dose calculation of electron beams from different machine types *Phys. Med. Biol.* **58** 2841–59 Online: https://doi.org/10.1088/0031-9155/58/9/2841

Fix M K, Frei D, Volken W, Terribilini D, Mueller S, Elicin O, Hemmatazad H, Aebersold D M and Manser P 2018 Part 1: Optimization and evaluation of dynamic trajectory radiotherapy *Med. Phys.* **45** 4201–12

Fix M K, Manser P, Frei D, Volken W, Mini R and Born E J 2007 An efficient framework for photon Monte Carlo treatment planning *Phys. Med. Biol.* **52** N425–N437 Online: https://doi.org/10.1088/0031-9155/52/19/N01

Hårdemark B, Liander A, Rehbinder H and Löf J 2004 P3IMRT®: Direct machine parameter optimization *Pinnacle3® White Pap.*

Heath E, Mueller S, Guyer G, Duetschler A, Elicin O, Aebersold D, Fix M K and Manser P 2021 Implementation and experimental validation of a robust hybrid direct aperture optimization approach for mixed-beam radiotherapy *Med. Phys.* **48** 7299–312

Henzen D, Manser P, Frei D, Volken W, Neuenschwander H, Born E J, Vetterli D, Chatelain C, Stampanoni M F M and Fix M K 2014 Monte Carlo based beam model using a photon MLC for modulated electron radiotherapy *Med. Phys.* **41** 021714 Online: https://doi.org/10.1118/1.4861711

Jeraj R 2002 The effect of dose calculation accuracy on inverse treatment planning *Phys. Med. Biol.* **47** 391–407 Online: https://doi.org/10.1088/0031-9155/47/3/303

Kawrakow I and Fippel M 2000 VMC++, a fast MC algorithm for radiation treatment planning *Use Comput. Radiat. Ther. 8th Int. Conf. (Heidelberg, Ger. ed W Schlegel T Bortfeld (heidelb. Springer)* 126–8 Online: https://doi.org/10.1007/978-3-642-59758-9

Li Y, Yao J and Yao D 2003 Genetic algorithm based deliverable segments optimization for static intensity-modulated radiotherapy. *Phys. Med. Biol.* **48** 3353–74

MacFarlane M, Hoover D A, Wong E, Goldman P, Battista J J and Chen J Z 2019 A fast inverse direct aperture optimization algorithm for intensity-modulated radiation therapy *Med. Phys.* **46** 1127–39

Magaddino V, Manser P, Frei D, Volken W, Schmidhalter D, Hirschi L and Fix M K 2011 Validation of the Swiss Monte Carlo Plan for a static and dynamic 6 MV photon beam *Z. Med. Phys.* **21** 124–34 Online: https://doi.org/10.1016/j.zemedi.2010.10.010

Men C, Gu X, Choi D, Majumdar A, Zheng Z, Mueller K and Jiang S B 2009 GPU-based ultrafast IMRT plan optimization. *Phys. Med. Biol.* **54** 6565–73

Men C, Romeijn H E, Taşkin Z C and Dempsey J F 2007 An exact approach to direct aperture optimization in IMRT treatment planning. *Phys. Med. Biol.* **52** 7333–52

Mohan R, Arnfield M, Tong S, Wu Q and Siebers J 2000 The impact of fluctuations in intensity patterns on the number of monitor units and the quality and accuracy of intensity modulated radiotherapy *Med. Phys.* **27** 1226–37

Mueller S, Fix M K, Henzen D, Frei D, Frauchiger D, Loessl K, Stampanoni M F M and Manser P 2018 Electron beam collimation with a photon MLC for standard electron treatments *Phys. Med. Biol.* **63** 14pp Online: https://doi.org/10.1088/1361-6560/aa9fb6

Mueller S, Fix M K, Joosten A, Henzen D, Frei D, Volken W, Kueng R, Aebersold D, Stampanoni M and Manser P 2017 Simultaneous optimization of photons and electrons for mixed beam radiotherapy *Phys Med Biol* **62** 5840–60 Online: https://doi.org/10.1088/1361-6560/aa70c5

Neuenschwander H and Born E J 1992 A Macro Monte Carlo method for electron beam dose calculations *Phys. Med. Biol.* **37** 107–25 Online: https://doi.org/10.1088/0031-9155/37/1/007

Neuenschwander H, Mackie T R and Reckwerdt P J 1995 MMC--a high-performance Monte Carlo code for electron beam treatment planning. *Phys. Med. Biol.* **40** 543–74 Online: https://doi.org/10.1088/0031-9155/40/4/005

Nguyen D, O'Connor D, Ruan D and Sheng K 2017 Deterministic direct aperture optimization using multiphase piecewise constant segmentation *Med. Phys.* **44** 5596–609

Niemierko A 1999 A generalized concept of equivalent uniform dose (EUD) *Med Phys* **26** 1100 Online: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:A+Generalized+Concept+of+Equivalent+Uniform+Dose#0

Nocedal J and Wright S J 2013 *Numerical Optimization* vol 11 Online: http://www.biogeosciences-discuss.net/11/12733/2014/%5Cnhttp://www.deepdyve.com/lp/elsevier/the-reflex-project-comparing-different-algorithms-and-implementations-mUMVk0bx1v%5Cnhttp://books.google.co.uk/books/about/Numerical_Optimization.html?id=epc5fX0lqR

Pakela J M, Tseng H H, Matuszak M M, Ten Haken R K, McShan D L and El Naqa I 2020 Quantum-inspired algorithm for radiotherapy planning optimization *Med. Phys.* **47** 5–18

Palma B A, Sánchez A U, Salguero F J, Arráns R, Sánchez C M, Zurita A W, Hermida M I R and Leal A 2012 Combined modulated electron and photon beams planned by a Monte-Carlo-based optimization procedure for accelerated partial breast irradiation *Phys. Med. Biol.* **57** 1191–202 Online: https://doi.org/10.1088/0031-9155/57/5/1191

Papp D, Bortfeld T and Unkelbach J 2015 A modular approach to intensity-modulated arc therapy optimization with noncoplanar trajectories *Phys. Med. Biol.* **60** 5179–98

Preciado-Walters F, Langer M P, Rardin R L and Thai V 2006 Column generation for IMRT cancer therapy optimization with implementable segments *Ann. Oper. Res.* **148** 65–79

Renaud M-A, Serban M and Seuntjens J 2017 On mixed electron – photon radiation therapy optimization using the column generation approach *Med. Phys.* **44** 4287–98 Online: https://doi.org/10.1002/mp.12338

Romeijn H, Ahuja R, Dempsey J and Kumar A 2005 A column generation approach to radiation therapy treatment planning using aperture modulation *SIAM J. Optim.* **15** 838–62 Online:

https://doi.org/10.1137/040606612

Shepard D M, Earl M A, Li X A, Naqvi S and Yu C 2002 Direct aperture optimization: a turnkey solution for step-and-shoot IMRT. *Med. Phys.* **29** 1007–18 Online: https://doi.org/10.1118/1.1477415

Smyth G, Evans P M, Bamber J C and Bedford J L 2019 Recent developments in non-coplanar radiotherapy *Br. J. Radiol.* **92**

Varian 2017 *Eclipse Photon and Electron Algorithms 15.5 Reference Guide Document*

Webb S 1989 Optimisation of conformal radiotherapy dose distribution by simulated annealing *Phys. Med. Biol.* **34** 1349–70

Wu Q and Mohan R 2000 Algorithms and functionality of an intensity modulated radiotherapy optimization system. *Med. Phys.* **27** 701–11 Online: https://doi.org/10.1118/1.598932

Yang J, Gui Z, Zhang L and Zhang P 2018 Aperture generation based on threshold segmentation for intensity modulated radiotherapy treatment planning *Med. Phys.* **45** 1758–70