**Supplementary material**

*Description of GPU implementation*

The GPU implementation of our in-house developed treatment planning system was realized in Java (Oracle Corporation, Redwood Shores, CA, USA). Extensive parts of the calculations were implemented on Aparapi GPU kernels. In figure S1 we show a schematic of all computational steps of our treatment plan generation. The calculation steps executed on the GPU are depicted in grey boxes. The steps belonging to the field generation are depicted with a dark background. These steps are repeated for every field direction, other steps only once. The steps for the plan generation are described in detail in this supplement.

# Inputs

The first step of the treatment plan generation consisted of reading the input data, which took a non-negligible amount of time (1-3 seconds). Input data included CT data, targets and organs at risks (OARs) contours with respective dose constraints and their relative importance, as well as information for all fields such as gantry and couch angles and pre-absorber settings.

# Phase I

The water equivalent depth (WED) map on the CT grid was calculated for each field direction, but only in a Volume Relevant for Optimization (VRO). This VRO was a box including all volumes of interest with a margin of 3 cm. For this, the same algorithm as in the clinical implementation was used [1] and was implemented as a dedicated Aparapi kernel. HU to proton stopping power conversion was performed using the stoichiometric approach described by [2]. A distance to target map was then calculated also on the CT grid inside the VRO.

# Phase II

Using the WED map, possible positions of deliverable Bragg peaks (BP) were calculated with the same lateral spacing of 4 mm and discrete WED range separations in depth of between 2.5 - 5.0 mm as used in the clinical implementation. The distance to target map was used to find all BPs within the defined target volume, and up to 5 mm outside. This was repeated for every field and the selected BPs were used for the subsequent optimization.

Optimization points in the VRO were chosen on a grid independent of the CT grid, with points located in air (HU values *<* -970) not being considered. This use of an independent optimization matrix provided flexibility. For example, a higher density of optimization points could be placed in OARs with a small volume or in areas where a high dose gradient was expected. For this work however, a uniform grid with a regular lattice spacing of 3.4 mm was used. This spacing resulted in an optimization point density, similar to that used in our clinical TPS.

# Phase III

Initial BP weights were defined by assigning predefined spread out BP weights depending on the first and last energy of the BPs in each column along the beam direction [3]. The *Dij* dose deposition matrix (see main text) was then calculated with a kernel employing the ray-casting (RC) dose calculation algorithm as described by [4]. This algorithm was parallelized over the different pencil beams allowing for the dose distribution of every spot to every optimization point to be stored separately for use in the optimization. This dose deposition matrix, with dimensions of the number of BPs multiplied by the number of dose points in the VRO, will henceforth be referred to as *Dij*. After its calculation, the matrix was compressed, using a compressed sparse row algorithm, to free GPU memory.

Translating the dose-volume constraint of the OARs to optimization point constraints was then a straight forward process, which involved the determination if an optimization point is inside an OAR or not.

Finally, as the ray-casting algorithm used a single Gaussian lateral dose kernel (i.e. secondary particle contributions were ignored) absolute dose was corrected after the optimization using an empirical nuclear interaction model that corrected absolute dose for each field. This model determined a single scaling factor per field, which depended on the number of BP in the field, the range of the most distal BP, the pre-absorber setting and treatment location (e.g. cranial or extra cranial).

# Optimization

The optimization process consisted of two phases – a dose calculation and BP weight (fluence) update. To obtain the dose to the optimization points a simple matrix multiplication of the dose deposition matrix with the BP weight vector was required, after which the BP weights were updated according to an iterative procedure described by [3]. This is a quasi Newtonian gradient descent algorithm with a damping factor to assure convergence, and which tried to minimize the quadratic difference between the calculated dose and prescribed dose. The prescribed dose was taken to be uniform inside the target with a Gaussian fall-off outside. In addition, if doses in an OAR infringed a dose-volume constraint, a correction term in the weight update function was activated on optimization points inside this OAR. The importance of OAR constraints were controlled with a relative (to the target constraint) weight. All optimizations reported in this work were based on 60 iterations.

# Outputs

The output from the plan generation procedure was the final, optimized dose distribution within the VRO, together with the set of optimized BP fluences that can be delivered to the treatment machine. The dose calculation from the GPU implementation is limited to the VRO, since only this points were used within the optimization. The *Dij* matrix calculation could be extended to the whole patient geometry, then a dose calculation of the entire volume would be available after optimization. We estimate that this would prolong the plan generation by 2-4 seconds. This approach was not pursued, since a commissioned dose calculation needs to be performed, for the plan review. Commissioning of the GPU dose calculation and optional *Dij* matrix extension are possible further steps.

# GPU

The code was optimized to run on a single Nvidia Quadro P6000 GPU. This GPU has 3840 CUDA kernels and 24 GB working memory. The size of the working memory is critical, since the dose deposition matrix can become large for large target volumes. If the dose deposition matrix exceeds the available GPU memory, its calculation has to be split up in parts, which increases calculation duration considerably. For a two field plan for a cranio-spinal axis patient with a target volume of 2.8 l and an optimization point spacing of 3.4 mm, the dose deposition matrix reached a size of close to 24 GB before compression.

**References**

1. Siddon RL. Fast calculation of the exact radiological path for a three-dimensional CT array. Med Phys. 1985;12:252–255.
2. Schneider U, Pedroni E, Lomax A. The calibration of CT Hounsfield units for radiotherapy treatment planning. Phys Med Biol. 1996;41:111–24.
3. Lomax A. Intensity modulation methods for proton radiotherapy. Phys Med Biol. 1999; :185–205.
4. Schaffner B, Pedroni E, Lomax A. Dose calculation models for proton treatment planning using a dynamic beam delivery system: an attempt to include density heterogeneity effects in the analytical dose calculation. Phys Med Biol. 1999;44:27–41.

**Figure S1.** Schematic of the computational treatment plan generation. Computational steps are represented in boxes. Steps in grey boxes are executed on the GPU. Steps belonging to the field generation are marked with a dark background. WED: water equivalent depth, OAR: organ at risk, DVH: dose volume histogram.

Figure S1

