

PAPER


Fred: a GPU-accelerated fast-Monte Carlo code for rapid treatment plan recalculation in ion beam therapy

To cite this article: A Schiavi *et al* 2017 *Phys. Med. Biol.* **62** 7482

View the [article online](#) for updates and enhancements.

Related content

- [Nuclear physics in particle therapy: a review](#)
Marco Durante and Harald Paganetti
- [Initial development of goCMC: a GPU-oriented fast cross-platform Monte Carlo engine for carbon ion therapy](#)
Nan Qin, Marco Pinto, Zhen Tian et al.
- [Recent developments and comprehensive evaluations of a GPU-based Monte Carlo package for proton therapy](#)
Nan Qin, Pablo Botas, Drosoula Giantsoudi et al.



ASTRO
Booth 2616

**“QA in the era of
MR-SIM and MRgRT”**

Don't miss this
20 minute talk
by Dr. Carri Glide-Hurst

Register Now 



[modusQA]

Accuracy. Confidence.™

Fred: a GPU-accelerated fast-Monte Carlo code for rapid treatment plan recalculation in ion beam therapy

A Schiavi^{1,2}, M Senzacqua^{1,2}, S Pioli^{1,5}, A Mairani^{3,4},
G Magro³, S Molinelli³, M Ciocca³, G Battistoni⁶
and V Patera^{1,2}

¹ Dipartimento SBAI, University of Rome ‘La Sapienza’, Rome, Italy

² INFN, Sezione di Roma 1, Rome, Italy

³ CNAO, Pavia, Italy

⁴ HIT, Heidelberg, Germany

⁵ INFN, LNF, Frascati, Italy

⁶ INFN, Sezione di Milano, Milan, Italy

E-mail: angelo.schiavi@uniroma1.it

Received 7 February 2017, revised 7 July 2017

Accepted for publication 20 July 2017

Published 5 September 2017



Abstract

Ion beam therapy is a rapidly growing technique for tumor radiation therapy. Ions allow for a high dose deposition in the tumor region, while sparing the surrounding healthy tissue. For this reason, the highest possible accuracy in the calculation of dose and its spatial distribution is required in treatment planning. On one hand, commonly used treatment planning software solutions adopt a simplified beam–body interaction model by remapping pre-calculated dose distributions into a 3D water-equivalent representation of the patient morphology. On the other hand, Monte Carlo (MC) simulations, which explicitly take into account all the details in the interaction of particles with human tissues, are considered to be the most reliable tool to address the complexity of mixed field irradiation in a heterogeneous environment. However, full MC calculations are not routinely used in clinical practice because they typically demand substantial computational resources. Therefore MC simulations are usually only used to check treatment plans for a restricted number of difficult cases. The advent of general-purpose programming GPU cards prompted the development of trimmed-down MC-based dose engines which can significantly reduce the time needed to recalculate a treatment plan with respect to standard MC codes in CPU hardware. In this work, we report on the development of FRED, a new MC simulation platform for treatment planning in ion beam therapy. The code can transport particles through a

3D voxel grid using a class II MC algorithm. Both primary and secondary particles are tracked and their energy deposition is scored along the trajectory. Effective models for particle–medium interaction have been implemented, balancing accuracy in dose deposition with computational cost. Currently, the most refined module is the transport of proton beams in water: single pencil beam dose–depth distributions obtained with FRED agree with those produced by standard MC codes within 1–2% of the Bragg peak in the therapeutic energy range. A comparison with measurements taken at the CNAO treatment center shows that the lateral dose tails are reproduced within 2% in the field size factor test up to 20 cm. The tracing kernel can run on GPU hardware, achieving 10 million primary s^{-1} on a single card. This performance allows one to recalculate a proton treatment plan at 1% of the total particles in just a few minutes.

Keywords: graphics processing units, particle therapy, treatment planning, Monte Carlo codes

(Some figures may appear in colour only in the online journal)

1. Introduction

Charged particle therapy (CPT), i.e. radiotherapy performed with protons or light ions, aims to deliver a high precision treatment of solid tumors (Jäkel *et al* 2008, Durante and Loeffler 2010). The dose deposition of ions as a function of depth in the traversed matter exhibits a sharp peak (the Bragg peak) at the end of the particle range and CPT exploits this characteristic shape. Compared to standard x-ray radiotherapy, CPT can obtain accurate irradiation of the tumor and at the same time reduce the dose to surrounding healthy tissue, thus achieving a lower complication probability. According to recent statistics (PTCOG 2017), by the end of 2014, more than 137 000 patients worldwide had been treated with charged hadrons (about 86% with protons), and the number of clinical centers dedicated to CPT is rapidly increasing. The high spatial selectivity of CPT puts stringent requirements on the accuracy that has to be achieved. In order to preserve the intrinsic advantages of hadron therapy, fast and accurate dose calculation tools are necessary in order to check, verify, and eventually correct, initial treatment planning.

One of the most powerful and versatile strategies for dose optimization is the intensity modulated particle therapy (IMPT) (Lomax 1999, Albertini *et al* 2011) with active scanning where a flat depth–dose distribution throughout the target volume is generated by the superposition of many thousands of individually weighted, narrow beam Bragg peaks. The main advantage of this technique is that it can deliver highly-conformal dose distributions to arbitrarily-shaped tumor volumes. Sophisticated software tools, called treatment planning systems (TPS), have been developed to produce a patient-specific set of particle beams. Standard commercial solutions are typically based on a pencil beam algorithm, where an accurate dose–depth profile is remapped in the transverse direction using analytical functions. These semi-analytical algorithms are fine tuned on measurements taken at each treatment center. They are very effective and accurate in reproducing the dose distribution under homogeneous irradiation conditions, e.g. the spread-out Bragg peak distribution in liquid water. However, in the presence of large density gradients and non-uniform materials, several studies have indicated that the most accurate approaches for dose estimation are those based on Monte Carlo (MC)

methods (Paganetti *et al* 2008, Parodi *et al* 2012, Grassberger *et al* 2015). Well established general purpose MC codes (Ferrari *et al* 2005, Agostinelli *et al* 2003, Pelowitz 2011) can be used to perform a complete plan recalculation. Significant effort has recently been made to interface these MC tools with the accelerator machine and patient data to produce integrated software platforms, e.g. TOPAS (Perl *et al* 2012), which could be more easily introduced in the clinical environment. The large computing resources needed when compared to semi-analytical codes have prevented the use of MC simulations in clinical practice until recently. The advent of general programming graphics processing units (GPU) has prompted the development of MC codes that can dramatically reduce the plan recalculation time (Jia *et al* 2012, 2014, Giantsoudi *et al* 2015).

In this framework, our group developed FRED (Fast paRticle thErapy Dose evaluator): a dose engine on GPU to recalculate and optimize ion beam treatment plans. The purpose of the code is to rapidly recalculate a complete treatment plan within minutes, opening the way for many clinical applications where the time-factor is important. The recalculation of a patient verification plan, as described in section 7, will be the first testbed of the new tool.

While developing the tracing algorithm from scratch, we tried to balance accuracy, calculation time, and GPU execution guidelines. To this end, we chose the most effective physical models from the literature and tested their implementation. For many processes, FRED relies on a library of precomputed look-up tables instead of performing an explicit calculation. This approach performs extremely well on GPU cards, where hardware interpolation can be exploited using the so-called texture units.

2. Code structure and physical models

The FRED code has been designed to perform fast and accurate calculations of energy deposition in the patient's body during the delivery of a treatment plan. The plan consists of the whole set of beam directions, particle energy, spot shape and fluence produced by the TPS. A plan is hierarchically decomposed in *fractions*, each fraction having one or more *fields*, i.e. treatment directions, with each field consisting of a series of *pencil beams*. The pencil beam encapsulates all the accelerator parameters that can be controlled in a given treatment facility. It corresponds to a single source of particles, controlling direction, angular divergence, and the energy or momentum spectrum of the particles to be traced. During the calculation, each pencil beam generates a user-prescribed number of primary particles, which are traced one-by-one in parallel by the tracing kernel.

The simulation domain is called *phantom*, and it could represent both a uniform volume of material (e.g. liquid water or PMMA, for testing purposes) or a patient's 3D reconstruction obtained from a series of CT scans. The phantom is divided by a Cartesian uniform grid in voxels, each storing information on the local density and atomic composition. The translation of CT values from Hounsfield units (HU) to material properties is performed using a built-in (Schneider *et al* 1996, 2000) or user-supplied conversion table.

The simulation setup is performed by parsing a text file of input parameters, or it can be prepared by a series of python scripts which parse a DICOM file tree containing patient CT scans together with morphological decomposition, radiation therapy plans, and TPS-calculated dose maps. Once the complete geometry has been imported, the primary particles are generated using the plan prescription. Particles are produced inside the accelerator beamline in vacuum. The effects of the beam monitors and exit window are modeled by a water-equivalent layer placed inside the vacuum pipe. Propagation from the exit window to the phantom is performed in air using the energy loss and scattering routines implemented in the code. Each field could

have several *filters* (clipping filters, range shifters, ripple filters, etc), which can alter the propagation of the primary particle. The phase space distributions after each filter have been characterized using full-MC simulations. The particle position, direction and energy are sampled from these distributions for each traversed filter. If a particle reaches the phantom, it is queued to the *tracing kernel*. Depending on the hardware the code is running on, the kernel adopts different strategies for distributing the workload among the available computing resources. In any case, the finest granularity of the simulation is the evaluation of the history of a single particle in passing through the phantom. The particle track is generated step-by-step using a class II MC algorithm Berger (1963), i.e. the physical processes acting on the particle are divided into condensed-history or *continuous* models, and point-like or *discrete* interaction models. At the beginning of the step, all active physical models are requested to determine the $s_{\max,i}$ maximum allowed tracklength which is a function of particle energy, voxel composition, and requested accuracy. The actual step length is the minimum value $s = \min_i s_i$, expressed in areal density, i.e. g cm^{-2} . The continuous processes are then applied to the traced particle, determining, for instance, the mean energy loss, the energy fluctuation, and the mean scattering angle. The end of a step corresponds to a voxel boundary crossing or a discrete interaction point. In the former case, the position of the particle in the simulation grid is updated. In the latter case, the cross sections for discrete interactions are evaluated, and the occurring interaction is determined via a sampling procedure. In a discrete interaction, other particles might be produced by knock-on events or nuclear fragmentation of both the particle projectile or the target nucleus. These *secondary* particles are queued for later tracking if their energy is above threshold, or extinguished locally in the voxel as described below. At this stage, the original particle is updated in position, direction and energy, and the stepping process is restarted. The particle history ends if it exits the phantom or if its energy becomes smaller than the threshold.

The FRED code can transport several kinds of particles (e.g. protons, deuterons, light ions, electrons, photons) with different levels of detail and accuracy. In this paper, we will focus on the tracking features needed to simulate the proton beams only, which is the most refined module presently integrated in the code, and for which the first clinical applications of our tool are described in the final sections.

3. Continuous processes for protons

3.1. Mean energy loss

In the energy range relevant for CPT (10 MeV–300 MeV), protons lose energy in their passage through matter, mainly by collisions with atomic electrons. The average energy release is well reproduced by the Bethe–Bloch formula (Bethe and Ashkin 1953). At the very end of the proton range, the interaction with the target nuclei also becomes important. The mean energy loss \overline{dT} suffered by a proton of energy T in a step is given by

$$\overline{dT} = S(T) \rho dz = S(T) ds, \quad (1)$$

where S is the total stopping power for the considered material, ρ is the density, dz is the actual tracklength, and ds is the corresponding areal density. In passing through liquid water, the energy loss for the same tracklength would be

$$\overline{dT}_w = S_w(T) \rho_w dz, \quad (2)$$

so that we can define in relative terms

$$\overline{dT} = \frac{S(T)}{S_w(T)} \cdot \frac{\rho}{\rho_w} \cdot \overline{dT}_w = f_S \cdot \frac{\rho}{\rho_w} \cdot \overline{dT}_w \quad (3)$$

where the density ρ , being expressed in g cm^{-3} , is actually already given relative to water. The stopping power ratio $f_S = S/S_w$ depends mainly on the electronic properties of the material, and is obtained for human tissues using a calibration curve (Schneider *et al* 1996, 2000). There is also a slight dependence of f_S on the proton energy, which is computed using the approach adopted by Fippel and Soukup (2004)

$$f_S(T, \rho) = 1.0123 - 3.386 \cdot 10^{-5} T - 0.291(1 + T^{-0.3421})(1 - \rho^{-0.7}) \quad (4)$$

for material density $\rho > 0.9 \text{ g cm}^{-3}$.

The code uses *PSTAR* (Berger *et al* 2016) tabulated total stopping power for integrating equation (2) for a given step length s . In order to speed-up the calculation while retaining a high level of accuracy, the tabulated S_w is spline-interpolated and integrated at the beginning of the simulation for a series of initial proton energies T_i and step lengths s_j using

$$s_j = - \int_{T_i}^{T_{\text{end}}} \frac{dT}{S_w(T)}. \quad (5)$$

The results are stored in a look-up table (T_i, s_j) , which is bilinearly interpolated to obtain T_{end} , the mean proton energy at the end of the step.

3.2. Energy fluctuations

The distribution of the total energy loss for each step is described by an energy straggling probability function, the shape of which is dependent on several parameters relating to the particle and to material properties. Two different regimes can be identified. In the *thick absorber* regime, the number of collisions suffered by the incoming proton is large, and the energy fluctuations can be well approximated by a Gaussian distribution centered on the mean energy loss value given by the stopping power of the previous section. The standard deviation of the distribution is computed as follows (Seltzer and Berger 1964):

$$\sigma_E^2 = \xi T_e^{\text{max}} \left(1 - \frac{1}{2}\beta^2\right) \text{MeV}^2, \quad (6)$$

where $T_e^{\text{max}} = \frac{2m_e\beta^2\gamma^2}{1+2\gamma m_e/m_p+(m_e/m_p)^2}$ is the maximum energy transferable by a proton to an electron in a single collision. Here m_e and m_p are the electron and proton mass in MeV/c^2 ; β and γ are the relativistic factors. The ξ parameter is the characteristic energy loss corresponding to the leading term in the Bethe–Bloch formula

$$\xi = 4\pi\mathcal{N}_A r_e^2 m_e z^2 \frac{Z}{A} \frac{1}{\beta^2} ds, \quad (7)$$

where \mathcal{N}_A , r_e , z and Z/A are, respectively, the Avogadro's number, the classical electron radius, the particle atomic number, and the atomic number over atomic mass ratio of the medium.

For very thin layers of material, the number of collision events is not enough to lead to a normal distribution of energy losses. The energy fluctuation is well described by the model developed by Landau (1944) and Vavilov (1957). Because the evaluation of the Landau–Vavilov distribution function is demanding in terms of calculation time, in the code we approximated

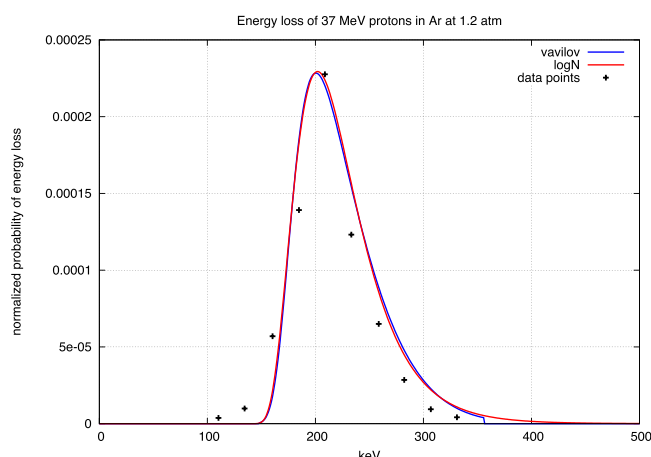


Figure 1. Comparison of Vavilov (blue line) and logarithmic normal (red line) distributions with experimental results of Gooding and Eisberg for 37 MeV protons through 10 cm of argon at 1.2 atm, corresponding to $k = 0.2$ and $\beta^2 = 0.07$. Data points were taken from Seltzer and Berger (1964).

the distribution for given pairs of k and β^2 parameters with a logarithmic normal function. This allowed us to greatly simplify the extraction procedure, and to maintain the accuracy level within the treatment plan requirements.

In order to determine the energy straggling regime for each step, the k parameter is evaluated

$$k = \frac{\xi}{T_e^{\max}}. \quad (8)$$

Following Seltzer and Berger (1964) we used the Gaussian approximation (thick absorber regime) for $k \geq 10$. For $k < 10$, the distribution of the energy loss was generated by sampling the Vavilov distribution, and it was then interpolated with a logarithmic normal function:

$$L_N(\lambda_L) = \frac{1}{(\lambda_L - \theta)\sqrt{2\pi}\sigma} \exp \left[-\frac{(\ln \frac{\lambda_L - \theta}{m})^2}{2\sigma^2} \right], \quad (9)$$

where λ_L is the Landau parameter (Vavilov 1957). The shape σ , location θ and scale m parameters of equation (9) were stored in a look-up table as a function of k and particle β^2 . Figure 1 shows the Vavilov distribution and the logarithmic normal fit implemented in FRED calculated for 37 MeV protons through 10 cm of argon at 1.2 atm (Seltzer and Berger 1964). The experimental conditions correspond to $k = 0.22$ and $\beta^2 = 0.074$.

3.3. Multiple Coulomb scattering

Fast charged particles passing through matter suffer a large number of small-angle deflections due to elastic Coulomb scattering, mainly with medium nuclei. The overall deflection angle is well described by the Molière (1948) theory of multiple Coulomb scattering (MCS). We adopted the small-angle approximation for the determination of the scattering angle θ , which is valid for most of the particle trajectory in the case of protons and heavier ions. The angle distribution is roughly Gaussian for small deflection angles, but at larger angles it behaves like

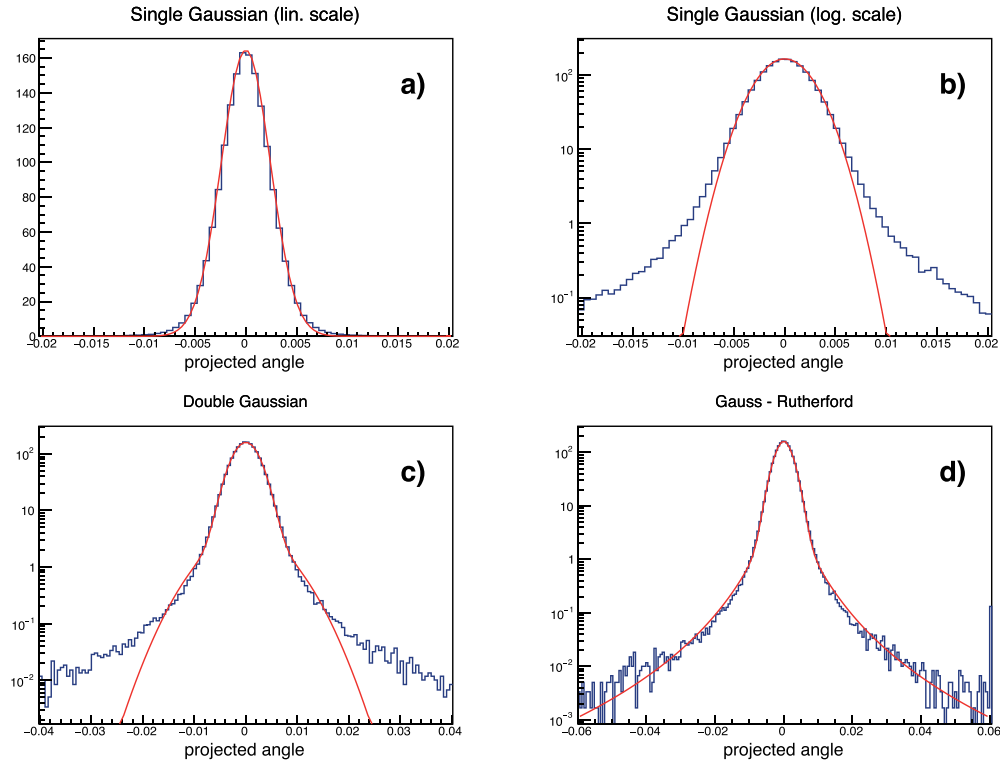


Figure 2. Single step angular distributions for 150 MeV protons through 1 mm of liquid water using Geant4 (blue) and the fit implemented in FRED (red). Top panels present the single Gaussian model in a) linear and b) logarithmic scale. Bottom panels show c) double Gaussian and d) Gauss–Rutherford models in logarithmic scale.

Rutherford scattering. Several effective distributions are implemented in the code, balancing accuracy with sampling calculation time.

3.3.1. Single Gaussian approximation. This is the simplest approach to capture the central part of the distribution. It is derived from the first approximation of the Molière formula (Molière 1948) and it is a Gaussian distribution with zero mean

$$f_G(\theta) = \frac{1}{\sqrt{2\pi}\theta_0} \exp\left[-\frac{1}{2}\frac{\theta^2}{\theta_0^2}\right]. \quad (10)$$

The width of the distribution is computed using Highland’s formula (1975)

$$\theta_0 = \frac{14.1 \text{ MeV}}{pv} z \sqrt{\frac{L}{L_R}} \left[1 + \frac{1}{9} \log\left(\frac{L}{L_R}\right)\right] \text{ rad} \quad (11)$$

where p , v and z are the momentum, velocity and charge of the particle, and L and L_R are the thickness and the radiation length of the material.

As shown in figure 2(a)), the single Gaussian is an excellent approximation for a single pencil beam, reproducing 98% of the angular distribution. In CPT, however, the superposition of thousands of pencil beams calls for a more accurate description of the lateral tails (see figure 2(b)).

3.3.2. Double Gaussian approximation. A slight improvement can be obtained by adding a second Gaussian distribution with a larger width to the core Gaussian (figure 2(c))

$$f_{2G}(\theta) = \frac{(1-w)}{\sqrt{2\pi}\sigma_1} \exp\left[-\frac{1}{2}\left(\frac{\theta}{\sigma_1}\right)^2\right] + \frac{w}{\sqrt{2\pi}\sigma_2} \exp\left[-\frac{1}{2}\left(\frac{\theta}{\sigma_2}\right)^2\right]$$

with $\sigma_1 < \sigma_2$ and $w \ll 1$. The central width σ_1 is very close to θ_0 .

3.3.3. Gauss–Rutherford approximation. The correction to the central Gaussian is here represented by a Rutherford-like distribution with wider tails

$$f_{GR}(\theta) = \frac{1-w}{\sqrt{2\pi}\sigma_1} \exp\left[-\frac{1}{2}\left(\frac{\theta}{\sigma_1}\right)^2\right] + w \frac{a}{(\theta+b)^c}, \quad c \sim 2.0. \quad (12)$$

Figure 2(d)) shows that this approach is by far superior to the previous ones for a single step. It is worth mentioning that it is also the only one that can retain a good degree of accuracy after a few hundred steps, close to the end of a particle track. The gain in accuracy has a cost in terms of calculation time which typically amounts to a 20% run time increase with respect to the other implementations.

In the code, at the end of each step, the scattering angle is sampled from the chosen distribution and the velocity direction is rotated accordingly. The parameters of the distributions, depending on the step length, material and particle energy, were extracted from a series of simulations performed with the Geant4 code (Agostinelli *et al* 2003). The results were stored in a pre-calculated look-up table for fast interpolation at run time.

4. Discrete processes for protons

With respect to the continuous processes, the discrete events are characterized by short-range hard interactions which lead to large variations in the energy and/or direction of the incoming particle. For hadrons and in particular for protons, the main discrete processes of interest in the CPT energy range are the nuclear elastic and inelastic interactions. In the model, each discrete process is statistically independent from the others, and each one contributes to the determination of the total macroscopic cross section Σ_{tot} . The cross section is computed at each step, since it depends on the particle type, energy and medium composition. The *on-the-fly* mean free path is then defined by $\lambda = 1/\Sigma_{\text{tot}}$, expressed in g cm^{-2} .

The number n_λ of mean free paths to the next discrete interaction point is sampled from an exponential distribution at the beginning of a track (and regenerated just after a discrete event) (Geant4 Collaboration 2016). This number is decreased step after step by the fraction δn of the mean free path travelled by the particle:

$$\delta n = \frac{ds}{\lambda} = \Sigma_{\text{tot}} ds. \quad (13)$$

When n_λ reaches zero, the code determines, by a sampling procedure, which of the active discrete processes is occurring, weighting each one by its contribution to the total cross section. The selected interaction is then processed, and finally the particle is reinserted in the stepping kernel.

In the following sections, we describe the two nuclear interaction models implemented for protons.

4.1. Nuclear inelastic interactions

The inelastic interactions in particle therapy are responsible for the production of secondary protons, deuterons and tritons and other fragments, generally emitted at a large angle and in a wide kinetic energy range. This process has a huge impact on the lateral tails of the pencil beam inside the patient, and must be taken into account carefully.

The ICRU Report 63 (ICRU 2000) and ENDF database (McLane 2001) of the double differential cross section of the interaction of $p - X$ between 10 and 250 MeV were used to compute the nuclear inelastic interaction probability and the emission angle and energy of the particles after the interaction. According to the ICRU46 (ICRU 1992), the materials included in the ICRU63 tables account for the vast majority of human body materials. The data for the elements not present in the ICRU tables are interpolated using a $A^{\frac{2}{3}}$ rule.

The macroscopic cross section Σ , i.e. the interaction probability per unit length for a given step, is computed from the microscopic cross section σ by

$$\Sigma(T_p) = \rho N_A \frac{w_X}{A_X} \sigma(T_p), \quad (14)$$

where ρ is the density, N_A is the Avogadro constant, w_X is the fractional weight of the X element, and A_X is the atomic mass. At the energy of interest, several kinds of secondary particles are produced in the $p - X$ inelastic nuclear interaction: protons, neutrons, deuterons, tritons, alpha particles and de-excitation photons. Even higher charge secondaries are generated with a limited rate. In the present FRED implementation, we explicitly take into account only the production of secondary protons and deuterons. The rest of the charged interaction products are considered as local dose deposition only due to their limited range.

The multiplicity of protons and deuterons in each event are interpolated from the ICRU63 data for the specific $p - X$ interaction. In the case of an inelastic interaction, the program first samples the secondary multiplicity (protons and deuterons) from a Poisson distribution with the interpolated value as average. Then the energy and finally the angle of the particle is sampled for each generated secondary.

The kinetic energy of the secondary T_s is obtained following the procedure described in Fippel and Soukup (2004), but substituting the original uniform sampling of the kinetic energy with a sampling from the distribution given by:

$$f(x) = 2 \frac{1 + ax}{2 + a} \quad x = \frac{T_s - T_s^{\min}}{T_s^{\max} - T_s^{\min}} \quad (15)$$

with $a = 0.2$, and $T_s^{\min, \max}$ defined as in Fippel and Soukup (2004). Figure 3 shows the kinetic energy distribution for the produced protons in the $p - C$ interaction with the kinetic energy of the primary proton $T_p = 200$ MeV.

Once the kinetic energy of the secondary has been assigned, the propagation angle with respect to the incoming primary proton is computed. The azimuthal angle around the parent particle direction is uniformly distributed while the aperture angle θ is sampled within the kinematic limit:

$$2 \frac{T_s}{T_p} - 1 \leq \cos(\theta) \leq 1 \quad (16)$$

and according to a parametric distribution function of $\sin(2\theta)$ defined by:

$$f(\theta) = B \exp[-B \sin(2\theta)]. \quad (17)$$

The B parameter is a function of the kinetic energy of the primary proton T_p and of the secondary kinetic energy T_s , given by:

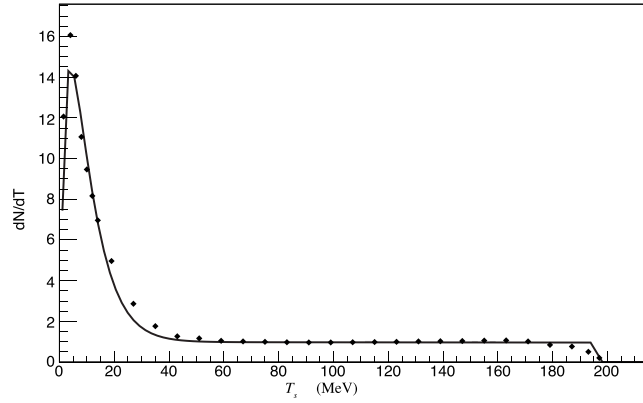


Figure 3. Kinetic energy distribution obtained in FRED for secondary protons produced by the interaction of a 200 MeV proton with a carbon nucleus. The solid line shows the ICRU63 data.

$$B = C_0 + \exp\left(-C_1 \frac{T_s}{T_p}\right) \quad (18)$$

$$C_0 = 0.1853 + 0.02157T_p - 5.442 \times 10^{-5}T_p^2 \quad (19)$$

$$C_1 = -2.476 - 4.224 \exp\left(-\frac{T_p - 7.878}{2.753}\right). \quad (20)$$

Figure 4 presents the distribution of $\sin(2\theta)$ of secondary protons with $T_s = 99$ MeV generated by a primary proton with $T_p = 200$ MeV, while in figure 5 the same distribution is shown for protons with $T_s = 35$ MeV and primary with $T_p = 120$ MeV. The points in the figure represent the ICRU63 data.

4.2. Nuclear elastic interactions

The model of nuclear elastic scattering adopted in FRED closely follows the approach presented in Fippel and Soukup (2004), to which we refer for implementation details. The macroscopic cross section is computed using the fit in Fippel and Soukup (2004) for $p-p$ and $p-O$ elastic interactions. For other nuclei, the relevant parameters of the $p-O$ interaction are rescaled using a $A^{\frac{2}{3}}$ rule.

The kinematics of $p-p$ scattering is solved in the center of mass, and then Lorentz transformed back to the laboratory frame. Both protons are treated as secondary particles and queued to the tracing kernel if they are above the tracking energy threshold.

In the case of $p-O$ elastic interaction, the knock-on oxygen deposits its energy locally, while the scattered proton is transported further by the tracing kernel.

5. Water model

The implemented physics models and the fast-MC performance were tested against the full-MC codes FLUKA and Geant4 for protons in liquid water.

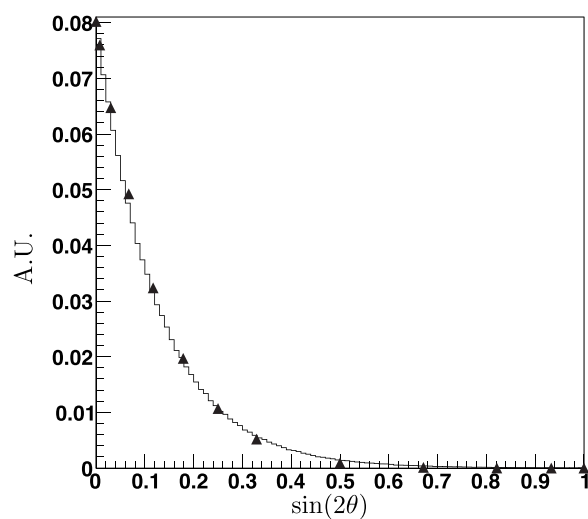


Figure 4. Distribution of $\sin(2\theta)$ for secondary protons of $T_s = 99$ MeV produced by the interaction of a 200 MeV proton with a carbon nucleus. The points show the ICRU63 data.

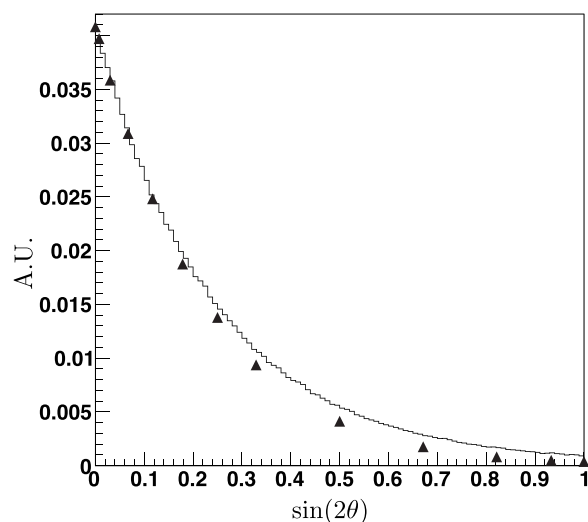


Figure 5. Distribution of $\sin(2\theta)$ for secondary protons of $T_s = 35$ MeV produced by the interaction of a 120 MeV primary proton with a carbon nucleus. The points show the ICRU63 data.

Figure 6 presents the Bragg curves for monoenergetic proton beams in a water phantom. The elastic and inelastic nuclear interactions were switched on and off in order to check each interaction model separately. The profiles closely overlap for most of the particle range. Just near the peak, slight differences can be spotted, as highlighted in the inset. The agreement with FLUKA is within 1.5% of the Bragg peak value for all models in the 50–250 MeV energy range. Using the complete interaction model, we scored the dose on a 12x12x40cm water box on a 1 mm grid, and compared the full-MC dose map with the FRED calculation.

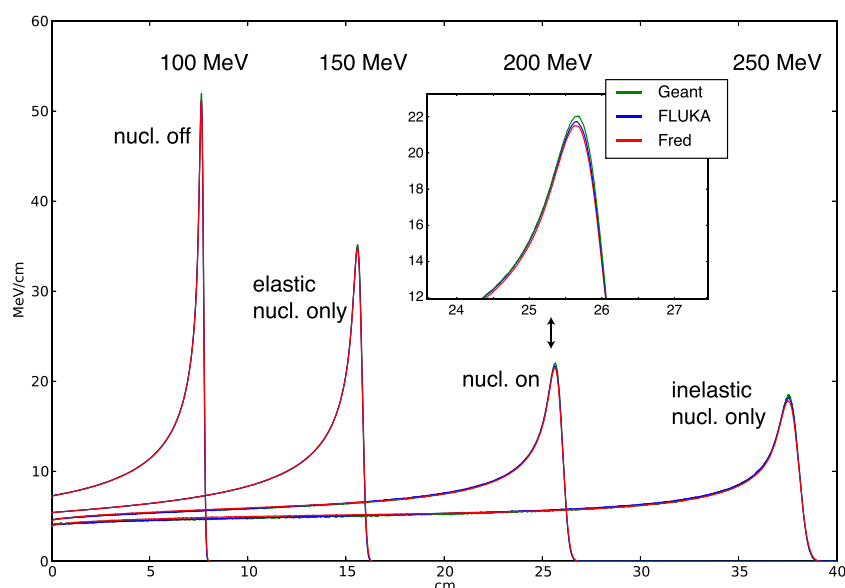


Figure 6. Bragg curves for protons in water with different energies. The nuclear interactions were switched on/off to check each model separately. FRED (red), FLUKA (blue) and Geant4 (green) calculations are presented. The inset shows the peak details for 200 MeV protons with the complete physical model.

We performed a γ -index test (Low *et al* 1998, Depuydt *et al* 2002) to compare the dose distributions, and we found a 100% pass-rate at 1 mm/1% for all voxels in the 50–250 MeV energy range. We also compared the dose maps using z -statistics analysis (Kawrakow and Fippel 2000), finding that systematic differences between FRED and FLUKA are well within 1% of the maximum dose.

The accuracy of the lateral dose distribution is very important for CPT applications, since the dose value in a single voxel is dependent on contributions from many thousands of pencil beams closely bundled in the transverse direction. The dose distribution for a single pencil beam of 150 MeV is presented in panel (a) of figure 7. The colormap is in logarithmic scale so that the tracks of secondary protons and deuterons are clearly visible. The dose line profile (b) along the beam axis shows the lateral beam spreading due to MCS. The transverse lineouts at the Bragg peak position, namely $z = 15.6$ cm, show in (c) linear and (d) logarithmic scale the tails of the distribution, mainly due to nuclear interactions. Comparison with FLUKA simulation (dashed lineouts), with the same scoring grid and the same statistics, shows good agreement up to four orders of magnitude in lateral dose distribution at a distance of 3 cm from the beam axis.

The field size factor (FSF) test is a technique that allows us to directly measure the contribution of long-range lateral tails in the dose distribution (see also Russo *et al* (2015) for more details). The test is performed in homogeneous conditions, placing a dose detector inside a liquid water phantom. A set of pencil beams with the same energy and fluence are delivered to the phantom. The scan spots are distributed on a square grid with constant spacing centered on the detector. The field size, i.e. the side of the spots square grid, is increased, and the dose recorded by the detector is scored at several depths in water. The test can be easily reproduced in simulations, therefore also allowing a comparison method for dose calculation codes.

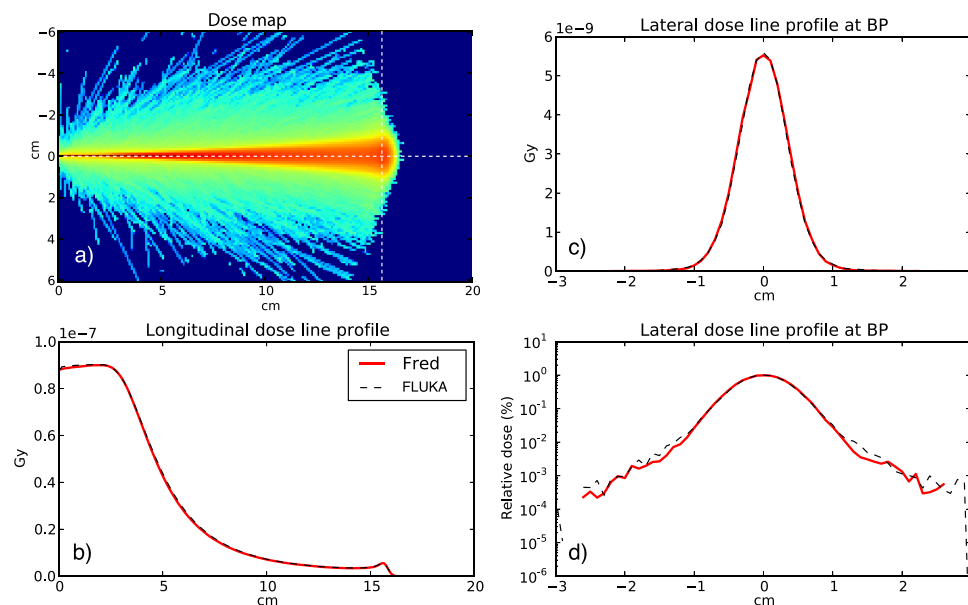


Figure 7. Dose in water for a 150 MeV proton beam with 0 FWHM. Dose map (a), central voxel lineout along beam axis (b), lateral lineout at 15.6 cm depth in linear (c), and logarithmic (d) scale.

The FSF test was run for a series of proton energies in water both with FRED and FLUKA. Figure 8 presents the case of a field made by 200 MeV pencil beams with 1 cm FWHM stacked with a transverse pitch of 2 mm. The dose was scored in a cylinder representing the active volume of a Markus ionization chamber placed at a depth of 20 cm in a water phantom. The integrated dose was normalized to the value of the maximum field size, namely 10 cm, for the FLUKA simulation. It was found that the lateral tails of the FRED dose distribution perfectly match the full-MC up to 3 cm away from the field axis. Further out, the signal is lower with respect to FLUKA since long range particles, such as neutrons or gamma-rays, are currently not transported. The overall agreement is within 1% for a field size up to 10 cm.

6. Comparison with TPS and clinical data

The accuracy of FRED dose recalculation was compared with results of the CNAO TPS. Since FRED, in principle, could achieve a higher accuracy in dose computation with respect to CNAO TPS, we also benchmarked the FRED results with data collected in the treatment room.

The commercial CE-marked Syngo RT Planning TPS (Siemens AG Healthcare, Erlangen, Germany) version VB10 was used for proton plan calculation and optimization. The software models each scan spot by a single pencil beam and calculates patient plans applying the water equivalent approach (Krämer *et al* 2000). The longitudinal dose profile is read from the database for the corresponding beam energy, whilst the dose is distributed across the plane perpendicular to the pencil beam central line according to a 2D distribution function using the weighted sum of two Gaussian distributions (Parodi *et al* 2011, Schwaab *et al* 2011). For proton effective dose calculation, a constant RBE value of 1.1 is currently applied (ICRU 2007). In order to allow a meaningful comparison of the FRED results with a clinical case, it is

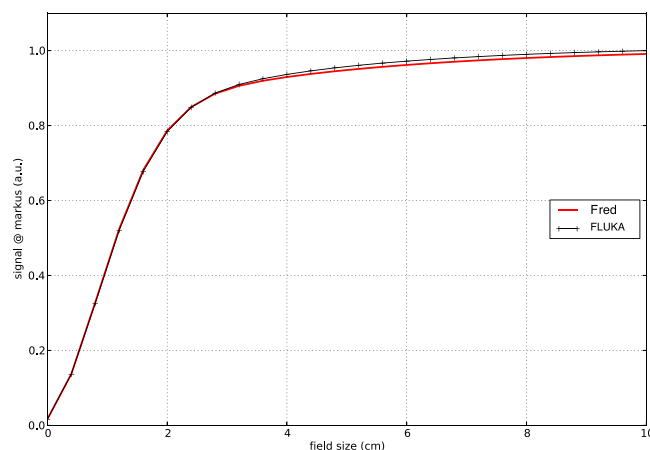


Figure 8. Field size factor test for a 200 MeV proton beam with 1 cm FWHM. The signal at the Markus ionization chamber at a depth of 20 cm is normalized to the maximum value of the FLUKA simulation.

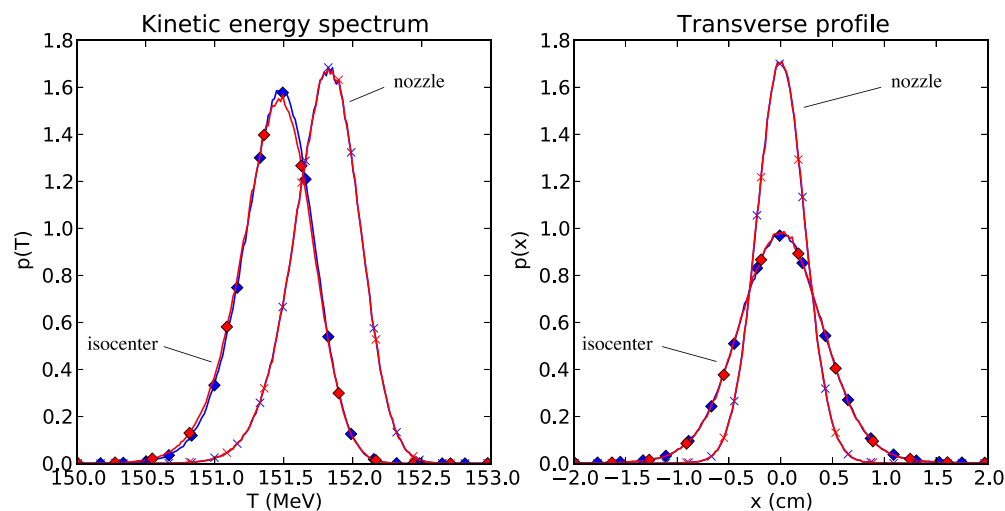


Figure 9. Phase space distributions of kinetic energy and lateral position at the nozzle exit (crosses) and at the isocenter position (diamonds). The Fluka simulations are in blue and the FRED results in red.

fundamental that a realistic beam model is implemented in FRED. This has been accomplished by importing the specific phase space description of a CNAO beam, which has been generated by MC simulations, constrained by experimental data, and developed at CNAO in order to compare MC simulations with treatment planning results. The successful implementation of such a phase space is exemplified in figure 9, where a comparison of energy spectra and transverse profiles at the nozzle exit and at the isocenter position in air, as calculated by FLUKA and FRED, is presented.

At CNAO, following plan approval, a patient-specific quality-assurance (*patient QA*) protocol is applied (Molinelli *et al* 2013, Magro *et al* 2015). The absorbed dose distribution is recalculated with the TPS in the PTW MP3-P water phantom independently for each treatment

field. The dosimetric system consists of a dedicated PMMA 3D block holder, able to host up to 24 pin-point ionization chambers (IC) aligned in six rows in such a way that each one is directly facing the beam. The holder is attached to the movable arm inside the water phantom, and optimal positions, in terms of dose sampling, are selected on the TPS. In our experimental settings, each measurement data set consists of the simultaneous use of 12 ICs (PTW model 31015) connected to a multi-channel precision electrometer (PTW Multidos). For each data set, at least two consecutive measurements are performed and the mean values taken.

The accuracy of the TPS dose calculation and the integrity of the whole treatment chain are also verified as part of the periodic QA protocol. The *SOBP QA* consists of a set of four 6 cm-sided cubic volumes, which are planned and verified in water with the same dosimetric system applied for the patient-specific QA. Each volume is centered at a different water depth, namely 9, 15, 21 and 27 cm, in order to cover a wide energy range. Additionally, a 3 cm cube centered at 21 cm in water is verified to test the dose calculation in small volumes at a large depth.

The level of agreement between the computed dose maps and measured dose points is evaluated using the relative mean deviation \bar{d} and absolute deviation $|\bar{d}|$

$$\bar{d} = \frac{1}{N} \sum_{i=1}^N \frac{d_{\text{meas},i} - d_{\text{calc},i}}{d_{\text{max}}} \quad ; \quad |\bar{d}| = \frac{1}{N} \sum_{i=1}^N \frac{|d_{\text{meas},i} - d_{\text{calc},i}|}{d_{\text{max}}} \quad (21)$$

calculated as the difference between measured d_{meas} and calculated dose d_{calc} , normalized to the maximum beam dose d_{max} and averaged over the N positions of the ionization chambers.

The actual N points included in the calculation are the measuring positions with a calculated gradient lower than 0.04 Gy mm^{-1} . This threshold value was determined by the sensitivity and limitations of the employed hardware. For the QA measurements in reference conditions, the applied acceptance threshold is 5% for both the mean deviation and its standard deviation over a data set. In addition, $\pm 7\%$ maximum deviation is accepted for any single IC point.

The SOBP QA was calculated with the TPS and then recalculated with FRED. Figure 10 presents (a) the irradiation geometry and dose maps in the (b) longitudinal and (c) transverse planes for the SOBP QA cube at a depth of 15 cm. Dose profiles for TPS and FRED passing through the cube center point are also shown in panels (d) and (e). It is worthwhile noting that the TPS does not score the dose in the entry channel in air, but only inside the phantom. The mean deviation and the standard deviation were computed for all cubes, both for the TPS and for FRED. Panels (f) and (g) in figure 10 show the lineouts corresponding to the pin-point IC number 11 together with the measured point. Differences in the 3D dose distributions were quantitatively assessed through a γ -index analysis extended to the cubic target volume. The 2 mm/2% pass-rate for mid-depth cubes was higher than 99%, whereas for the 9 cm and 27 cm depth cubes, the pass-rate was around 92%. Table 1 reports the evaluation of mean dose deviation with respect to the measured data set acquired for the SOBP QA. The deviation was computed for the nominal ‘zero-shift’ PPCH holder position. In order to take into account possible systematic errors in target alignment, as described in Molinelli *et al* (2013), the position of the holder was also varied in a 1 mm sided cube around the nominal position, and the ‘best-shift’ minimizing the dose deviation is presented in table 1. Even if the acceptance threshold of the protocol is 5%, the results of table 1 clearly show that both the TPS and FRED are well within that limit. The agreement with the experimental data significantly improves by exploring points close to the nominal position (see best-shift evaluations). FRED dose maps are equivalent to TPS-generated dose distributions, except for the cube at 27 cm, suggesting that at a greater depth, the effective models implemented in the fast-MC need further improvement.

The same dosimetric setup was also used for the field size factor measurements, and figure 11 shows a comparison of the measured dose levels with the synthetic curve generated

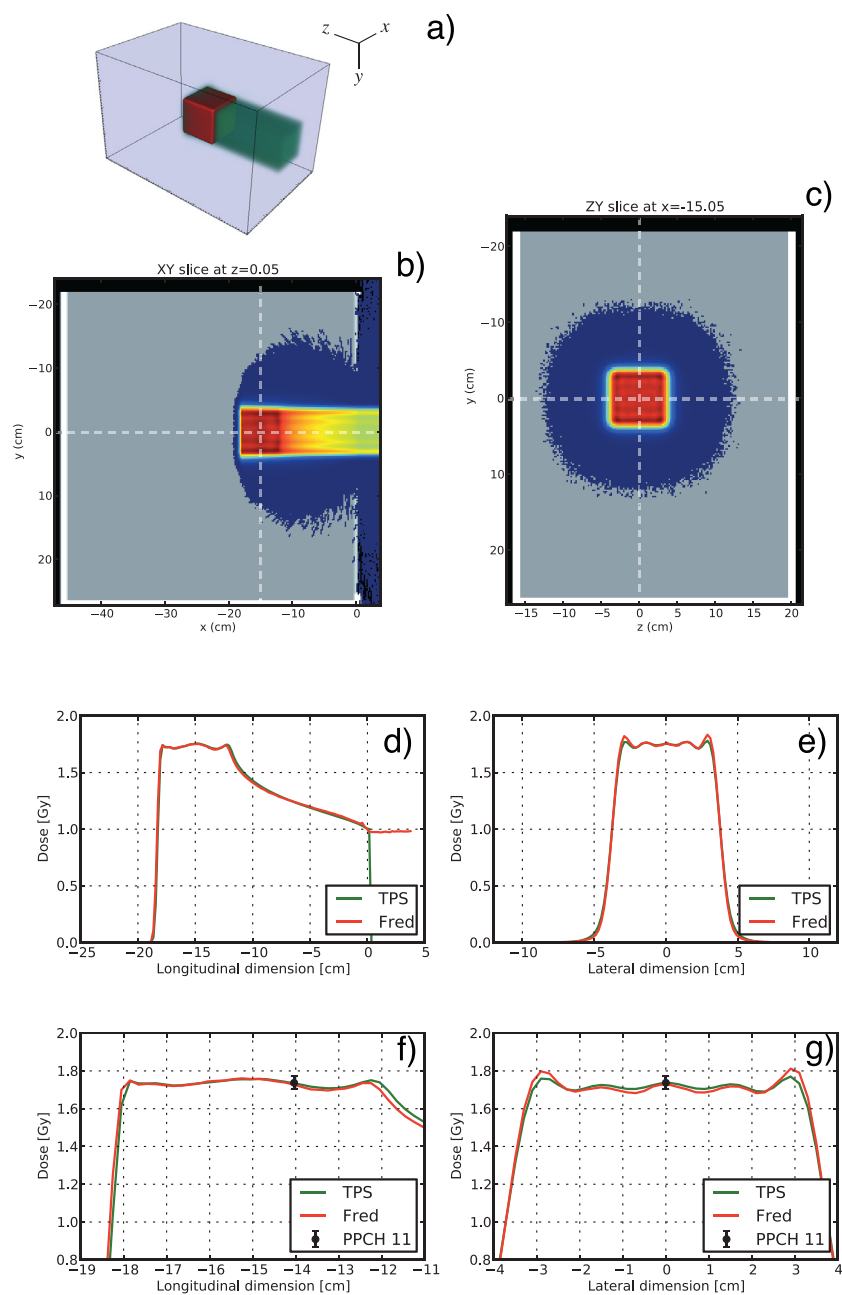


Figure 10. SOBP QA cube at a depth of 15 cm: irradiation geometry (a); CT scan of the water phantom with the overlaid FRED dose map in the longitudinal (b) and transverse (c) plane; TPS and FRED dose lineouts through the cube center ((d) and (e)). Enlarged version of the profiles together with the measured dose level for pin-point IC number 11 ((f) and (g)).

Table 1. Relative dose differences with respect to experimental data of TPS and FRED dose calculation in the SOBP QA protocol.

Cube	<i>N</i>	Zero shift				Best shift			
		TPS		FRED		TPS		FRED	
		Mean dose difference (%)	Standard deviation (%)	Mean dose difference (%)	Standard deviation (%)	Mean dose difference (%)	Standard deviation (%)	Mean dose difference (%)	Standard deviation (%)
$6 \times 6 \times 6 \text{ cm}^3$ 9 cm depth	11	−1.89	0.72	1.27	0.65	−1.60	1.08	−0.56	1.32
$6 \times 6 \times 6 \text{ cm}^3$ 15 cm depth	11	−0.44	0.55	0.06	0.57	0.0	0.62	0.0	0.58
$6 \times 6 \times 6 \text{ cm}^3$ 21 cm depth	11	−0.16	0.47	0.41	0.85	0.0	0.38	0.06	1.13
$3 \times 3 \times 3 \text{ cm}^3$ 21 cm depth	10	−0.32	1.78	0.65	2.03	0.0	1.27	0.0	1.48
$6 \times 6 \times 6 \text{ cm}^3$ 27 cm depth	11	−0.29	0.48	1.57	0.60	−0.15	0.88	1.4	0.8

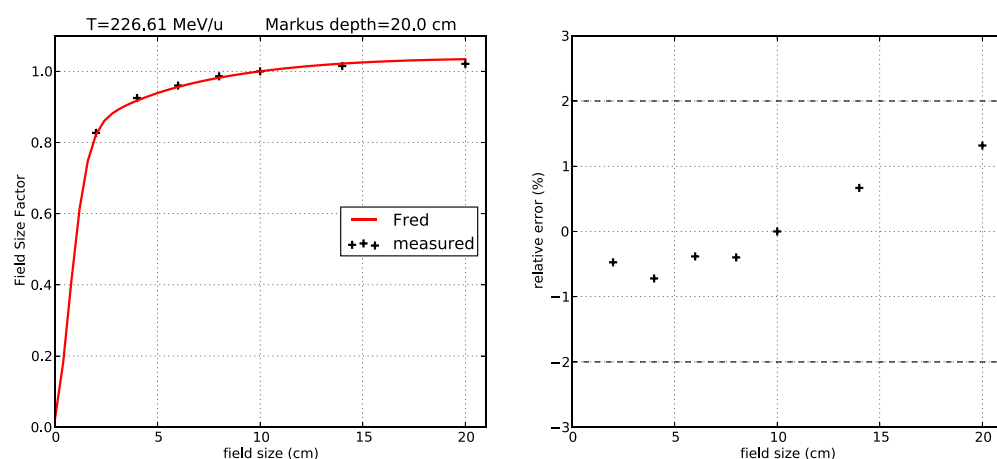


Figure 11. Field size factor test for a 226.61 MeV u^{-1} proton beam with the Markus chamber positioned at a depth of 20 cm in the CNAO water phantom. FSF is normalized to 10 cm size value. FSF computed with FRED and measured data (left), and relative error (right).

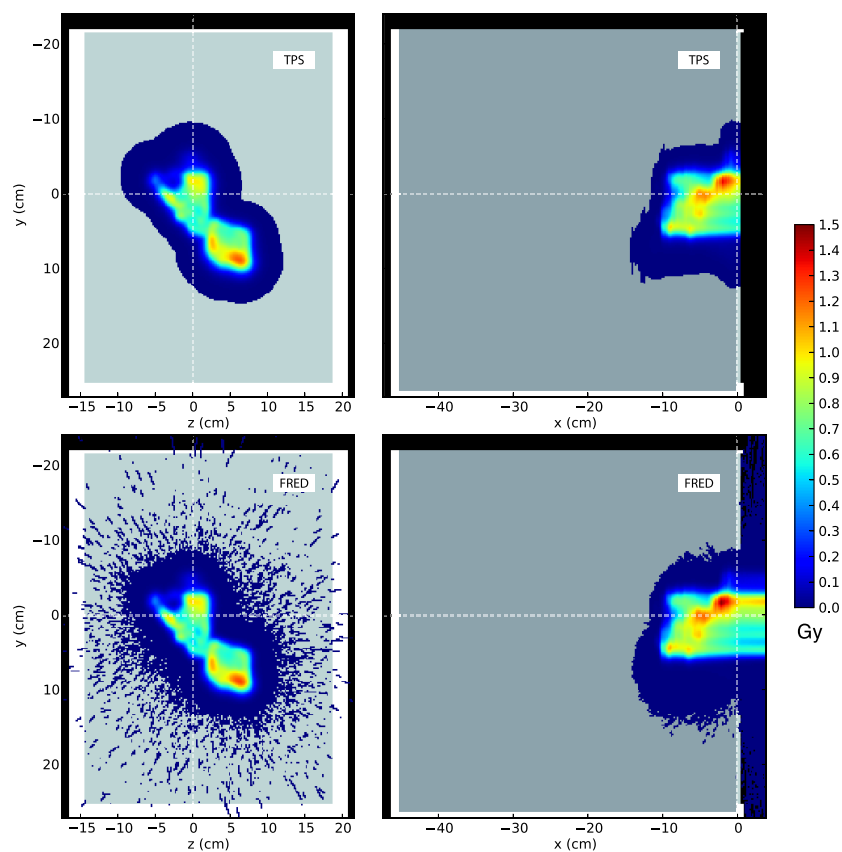


Figure 12. Transverse and longitudinal dose maps for a patient verification plan obtained with the TPS (top) and with FRED (bottom).

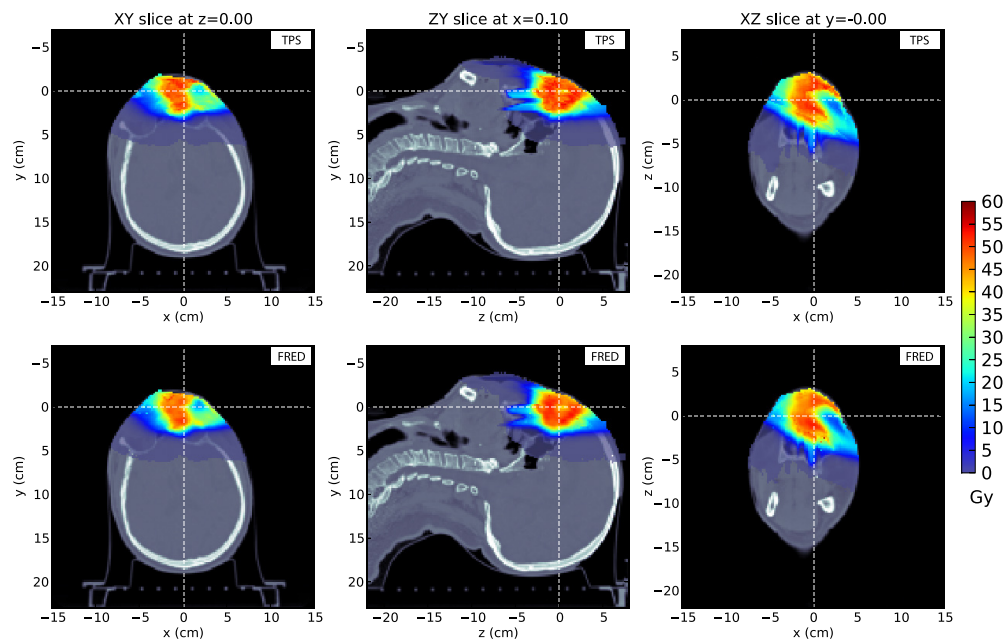


Figure 13. Physical dose maps for a three-field proton plan obtained with the TPS (top) and with FRED (bottom). Dose points within the patient's skin are shown only.

with FRED. For this data set, a single advanced Markus ionization chamber was collocated inside the CNAO water phantom (Russo *et al* 2015). Computed and collected data points were normalized to their respective value at 10 cm field size. FRED reproduces the experimental data within 2% for a field size up to 20 cm.

7. Verification plan at CNAO

As anticipated in section 6, the patient QA requires that for each field of an approved plan, the dose map is recalculated in water with the TPS and measured with the water phantom. Figure 12 presents the dose distribution for a verification plan using the TPS and FRED. The main features of the dose distribution are well predicted by FRED, and this is confirmed by a 99.6% pass rate for γ -index 2 mm/2%, and 96.7% for the 1 mm/1%. It is also possible to see the differences in the low dose regions between the pencil beam algorithm of the TPS and the discrete tracks of the fast-MC. The complete field recalculation at 1% of the total number of planned protons, namely 680 million primary particles, took 2 min on a 4-GPU workstation (see the appendix for details). This field was also measured and the mean dose deviation was calculated over 11 IC positions. The results are 1.7%/0.8% (absolute mean and standard deviation) for the TPS, and 1.9%/1.1% for FRED, both leading to a *verified* plan.

As a final check, the accuracy of the dose calculation was tested on a complete plan with the actual patient geometry. The plan consisted of a three-field proton irradiation of a superficial tumor. As such, each field had a range shifter to spread out the Bragg peaks. Figure 13 compares the results obtained with FRED and the TPS. The overall agreement of dose distributions is at the level of 97% for γ -index 2 mm/2%. However, hot and cold spots (up to 5%) can be identified at a finer scale. Fine-tuning of the simulation setup (e.g. using the same HU calibration curves) and conversion from dose-to-medium to dose-to-water (Paganetti 2009) for direct comparison will be included in a subsequent publication.

8. Conclusions

The development of a new platform for treatment planning in particle therapy has been presented. A few years ago, when the project began, MC simulations were used only for academic investigations or off-routine plan verification due to the large computing resources and/or execution time needed. The possibility of exploiting the computing power of multiple GPU cards opened up the way to bringing MC simulations into the treatment planning itself. The new hardware demanded for a new bottom-up development of a tracing kernel with the aim of balancing the speed of calculation with the accuracy of implemented physical models. The FRED code has reached a considerable speed-up in the plan recalculation, namely a typical run of about 72 h/core (Mairani *et al* 2013) can now be delivered in less than 2 min on a GPU-fitted workstation (see the appendix). The code can run on a variety of hardware configurations, and it can be used as a standalone application or it can be driven as an external library. Great effort has been made in order to match the dose deposition accuracy reached by state-of-the-art commercial tools and general purpose MC codes. The most refined model currently implemented in FRED is the transport of proton beams in liquid water. The dose maps calculated with FRED are in good agreement with *de-facto* standard MC codes and with the adopted TPS. The beamline QA and patient QA protocols of CNAO were reproduced with FRED, and the results are well within the acceptance thresholds set by the treatment center. We foresee the first clinical application of FRED in the recalculation of patient verification plans at CNAO. The code will run in parallel with the actual QA protocol during a commissioning phase. Once a significant statistical base is acquired, we will proceed to a clinical validation of the tool and progressively introduce FRED in the patient QA protocol, partly replacing the time-consuming measuring sessions with an *in silico* verification calculated with the FRED code. Future developments of the project consist of the introduction of RBE models for proton therapy, and the implementation of a carbon ion transport model in water for the patient QA with C at CNAO.

Appendix. Computation platform structure

The FRED fast-MC platform has been designed for maximum utilization of computing resources available to the user. The front-end is written in C++ and deals with the parsing of input data and parameters, the model of the accelerator machine, and the post-processing and output of the dose calculation. Typically, a complete treatment plan is stored in a DICOM file tree, including a series of patient CT scans, the definition of the volumes of interest (e.g. RTSTRUCT), the plan together with the irradiation geometry and the accelerator set-up (RTPLAN), and the evaluated dose (RTDOSE). A series of python scripts interface the code with DICOM data, allowing customization of the import routines.

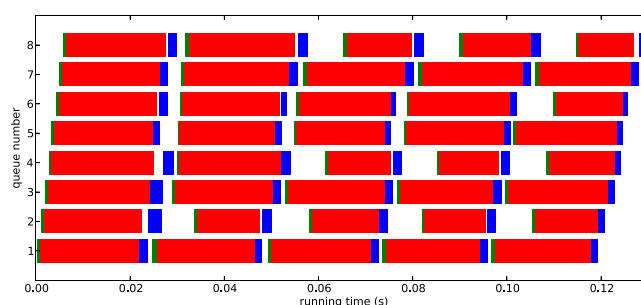
The particle tracing kernel is the computing core of the platform, and it gathers all available resources before laying out the computation plan. FRED can run on a distributed cluster of nodes using MPI communications. The intranode computation is split and balanced between the CPU cores and the connected GPU devices. High parallelism is achieved using POSIX threads on multicore CPU hardware and OpenCL threads on GPU hardware.

Using this paradigm, the code can run at high performance on laptops, workstations and mainframes as well. For treatment planning in a clinical center, the reference hardware is a workstation with multicore CPU and one or more GPU devices. The code performance was measured on a few hardware platforms and is reported in table A1. The benchmark case consists of a 150 MeV proton beam with 1 cm FWHM propagating through a water phantom. The scoring grid is a $20 \times 20 \times 20$ cm box with 2 mm cubic voxels. It is worthwhile noting

Table A1. Computing times for different hardware architectures.

CPU ^a	Threads	primary s ⁻¹ (k)	μ s/primary
Full-MC	1	0.75	1330
FRED	1	15	68
FRED	16	50	20
FRED	32	80	12.5

GPU	Cards	primary s ⁻¹ (k)	μ s/primary
AMD Radeon R9 M370X ^b	1	500	2
AMD FirePro D300 ^c	2	2000	0.5
NVIDIA GTX 1080	1	11 200	0.09
NVIDIA GTX 980 ^d	1	5350	0.2
NVIDIA GTX 980	2	10 200	0.1
NVIDIA GTX 980	3	15 600	0.064
NVIDIA GTX 980	4	19 900	0.05

^a Motherboard with two Intel® Xeon E5-2687 8-Core CPU at 3.1 GHz^b LAPTOP: Apple® MacBook Pro with one AMD® Radeon R9 M370X.^c DESKTOP: Apple® Mac Pro with two AMD® FirePro D300.^d WORKSTATION: Linux box with four NVIDIA® GTX 980.**Figure A1.** Execution timeline for eight queues on four GPUs using OpenCL. Double-queue buffering for each device ensures asynchronous kernel execution and data transfers, exploiting the double copy-engine available on GPU cards. Host-to-device transfers (green), kernel execution (red), and device-to-host (blue) transfers are shown for each queue.

that the sub-linear scaling with respect to CPU threads on the same hardware is mainly due to cache competition and/or automatic chipset overclocking of IntelTM processors. The ‘full-MC’ performance is representative of both FLUKA and Geant4 codes, which held almost the same result on the tested hardware.

Single-card GPU performance is very sensitive to the chipset model and the number of available computing cores. A $2\times$ performance increase has been observed for the latest generation of cards with respect to the previous one, e.g. the NVIDIA® GTX 1080 versus NVIDIA® GTX 980.

Multi-card GPU performance showed a high level of scalability. The lower part of table A1 reports performance scaling at a fixed workload for a 4 GPU system. Up to 93% linear scaling was obtained for concurrent execution on four cards.

In order to exploit the high-parallelism of the GPU hardware, the tracing kernel allocates one particle history per thread. By using double-queue asynchronous execution on each device

and host pinned memory for data exchange (see figure A1), it was possible to continuously stream particles to the GPU achieving a significant occupancy of the device. As a consequence, the GPU cards heat up rapidly, reaching 80 °C within seconds, when the temperature protection circuitry of the device downgrades the performance. For long runs, water-cooling the GPUs was necessary in order to obtain a 100% duty cycle during the calculation.

Taking into account the fact that the computing time is highly dependent on several factors (e.g. the size of the treatment volume, the size and spacing of the scoring grid, the energy transport threshold, the adopted hardware solution) the overall computing performance of FRED is similar to other GPU-based dose engines for proton therapy (Wan Can Tseung *et al* 2015, Quin *et al* 2016).

References

- Agostinelli S *et al* 2003 Geant4—a simulation toolkit *Nucl. Instrum. Methods A* **506** 250–303
- Allison J *et al* 2006 Geant4 developments and applications *IEEE Trans. Nucl. Sci.* **53** 270–8
- Allison J *et al* 2016 Recent developments in Geant4 *Nucl. Instrum. Methods A* **835** 186–225
- Albertini F, Casiraghi M, Lorentini S, Rombi B and Lomax A J 2011 Experimental verification of IMPT treatment plans in an anthropomorphic phantom in the presence of delivery uncertainties *Phys. Med. Biol.* **56** 4415–31
- Berger M J 1963 Monte Carlo calculation of the penetration and diffusion of fast charged particles *Methods in Computational Physics* vol 1, ed B Alder *et al* (New York: Academic) pp 135–215
- Berger M J, Coursey J S, Zucker M A and Chang J 2016 *ESTAR, PSTAR, and ASTAR: Computer Programs for Calculating Stopping-Power and Range Tables for Electrons, Protons, and Helium Ions (version 1.2.3)* (Gaithersburg, MD: National Institute of Standards and Technology) (www.nist.gov/pml/stopping-power-range-tables-electrons-protons-and-helium-ions (Accessed: 15 August 2017))
- Bethe H and Ashkin J 1953 *Experimental Nuclear Physics* ed E Segrè (New York: Wiley) p 253
- Depuydt T, Van Esch A and Huyskens D P 2002 A quantitative evaluation of IMRT dose distributions: refinement and clinical assessment of the gamma evaluation *Radiother. Oncol.* **62** 309–19
- Durante M and Loeffler J S 2010 Charged particles in radiation oncology *Nat. Rev. Clin. Oncol.* **7** 37
- Ferrari A, Sala P R, Fassò A and Ranft J 2005 FLUKA: a multi-particle transport code CERN-2005-10 (2005), INFN/TC-5/11, SLAC-R-773
- Böhlen T T, Cerutti F, Chin M P W, Fassò A, Ferrari A, Ortega P G, Mairani A, Sala P R, Smirnov G and Vlachoudis V 2014 The FLUKA code: developments and challenges for high energy and medical applications *Nucl. Data Sheets* **120** 211–4
- Battistoni G *et al* 2016 The FLUKA code: an accurate simulation tool for particle therapy *Front. Oncol.* (<https://doi.org/10.3389/fonc.2016.00116>)
- Fippel M and Soukup M 2004 A Monte Carlo dose calculation algorithm for proton therapy *Med. Phys.* **31** 2263
- Fippel M 2006 Monte carlo dose calculation for treatment planning *New Technologies in Radiation Oncology* (Berlin: Springer) pp 197–206
- GEANT4 Collaboration 2016 Physics Reference Manual, Version: Geant4 10.3 (9 December 2016) available at <http://geant4.web.cern.ch/geant4/support/index.shtml> (Accessed: 15 August 2017)
- Giantsoudi D, Schümann J, Jia X, Dowdell S, Jiang S B and Paganetti H 2015 Validation of a GPU-based Monte Carlo code (gPMC) for proton radiation therapy: clinical cases study *Phys. Med. Biol.* **60** 2257–69
- Grassberger C, Lomax A J and Paganetti H 2015 Characterizing a proton beam scanning system for Monte Carlo dose calculation in patients *Phys. Med. Biol.* **60** 633–45
- ICRU 1992 *Photon, Electron, Proton, Neutron Interaction Data for Body Tissues (ICRU-Report vol 46)* (Bethesda, MD: International Commission on Radiation Units and Measurements)
- ICRU 2000 *Nuclear Data for Neutron and Proton Radiotherapy and for Radiation Protection (ICRU-Report vol 63)* (Bethesda, MD: International Commission on Radiation Units and Measurements)
- ICRU 2007 Prescribing, recording, and reporting proton-beam therapy *J. ICRU* **7**
- Jäkel O *et al* 2008 The future heavy ion radiotherapy *Med. Phys.* **35** 5633

- Jia X, Schümann J, Paganetti H and Jiang S B 2012 GPU-based fast Monte Carlo dose calculation for proton therapy *Phys. Med. Biol.* **57** 7783–97
- Jia X, Ziegenhein P and Jiang S B 2014 GPU-based high-performance computing for radiation therapy *Phys. Med. Biol.* **59** R151–82
- Kawrakow I and Fippel M 2000 Investigation of variance reduction techniques for Monte Carlo photon dose calculation using XVMC *Phys. Med. Biol.* **45** 2163–83
- Knopf A and Lomax A 2013 *In vivo* proton range verification: a review *Phys. Med. Biol.* **58** R131–60
- Krämer M, Jäkel O, Haberer T, Kraft G, Schardt D and Weber U 2000 Treatment planning for heavy-ion radiotherapy: physical beam model and dose optimization *Phys. Med. Biol.* **45** 3299–317
- Landau L 1944 On the energy loss of fast particles by ionization *J. Phys. USSR* **8** 201
- Lomax A J 1999 Intensity modulated methods for proton therapy *Phys. Med. Biol.* **44** 185–205
- Low D A, Harms W B, Mutic S and Purdy J A 1998 A technique for the quantitative evaluation of dose distributions *Med. Phys.* **25** 656–61
- Magro G et al 2015 Dosimetric accuracy of a treatment planning system for actively scanned proton beams and small target volumes: Monte Carlo and experimental validation *Phys. Med. Biol.* **60** 6865
- Mairani A, Böhlen T T, Schiavi A, Tessonnier T, Molinelli S, Brons S, Battistoni G, Parodi K and Patera V 2013 A Monte Carlo-based treatment planning tool for proton therapy *Phys. Med. Biol.* **58** 2471–90
- McLane V 2001 ENDF-102 data formats and procedures for the evaluated nuclear data file ENDF-6 *Technical Report* BNL-NCS-44945-01/04-Rev (Upton NY, Brookhaven National Laboratory, National Nuclear Data Center)
- Molière G Z 1948 *Z. Naturforsch.* **3a** 78
- Molinelli S, Mairani A, Mirandola A, Vilches Freixas G, Tessonnier T, Giordanengo S, Parodi K, Ciocca M and Orecchia R 2013 Dosimetric accuracy assessment of a treatment plan verification system for scanned proton beam radiotherapy: one-year experimental results and Monte Carlo analysis of the involved uncertainties *Phys. Med. Biol.* **58** 3837–47
- Quin N et al 2016 Recent developments and comprehensive evaluations of a GPU-based Monte Carlo package for proton therapy *Phys. Med. Biol.* **61** 7347–62
- Paganetti H, Jiang H, Parodi K, Slopesma R and Engelsman M 2008 Clinical implementation of full Monte Carlo dose calculation in proton beam therapy *Phys. Med. Biol.* **53** 4825
- Paganetti H 2009 Dose to water versus dose to medium in proton beam therapy *Phys. Med. Biol.* **54** 4399–421
- Parodi K, Bauer J, Kurz C, Mairani A, Sommerer F, Unholtz D, Haberer T and Debus J 2011 Monte Carlo modeling and *in vivo* imaging at the Heidelberg Ion Beam Therapy *NSS/MIC: IEEE Nuclear Science Symp. and Medical Imaging Conf.* pp 2795–9
- Parodi K, Mairani A, Brons S, Hasch B G, Sommerer F, Naumann J, Jäkel O, Haberer T and Debus J 2012 Monte Carlo simulations to support start-up and treatment planning of scanned proton and carbon ion therapy at a synchrotron-based facility *Phys. Med. Biol.* **57** 3759–84
- Pelowitz D B (ed) 2011 MCNPX User's Manual, Version 2.7.0 *Los Alamos National Laboratory Report* LA-CP-11-00438
- Perl J, Shin J, Schumann J, Faddegon B and Paganetti H 2012 TOPAS: an innovative proton Monte Carlo platform for research and clinical applications *Med. Phys.* **39** 6818–37
- Particle Therapy Co-Operative Group 2017 <http://ptcog.web.psi.ch/> (Accessed: 15 August 2017)
- Russo G et al 2015 A novel algorithm for the calculation of physical and biological irradiation quantities in scanned ion beam therapy: the beamlet superposition approach *Phys. Med. Biol.* **61** 183
- Schneider U, Pedroni E and Lomax A 1996 The calibration of CT Hounsfield units for radiotherapy treatment planning *Phys. Med. Biol.* **41** 111–24
- Schneider W, Bortfeld T and Schlegel W 2000 Correlation between CT numbers and tissue parameters needed for Monte Carlo simulations of clinical dose distributions *Phys. Med. Biol.* **45** 459–78
- Schwaab J, Brons S, Fieres J and Parodi K 2011 Experimental characterization of lateral profiles of scanned proton and carbon ion pencil beams for improved beam models in ion therapy treatment planning *Phys. Med. Biol.* **56** 7813–27
- Seltzer S M and Berger M J 1964 Energy loss straggling of protons and mesons: tabulation of the vavilov distribution *Studies in Penetration of Charged Particles in Matter, Publication 1133* (Washington DC: National Academy of Sciences–National Research Council) pp 187–203
- Vavilov P V 1957 Ionization losses of high-energy heavy particles *Sov. Phys.-JETP* **5** 749
- Wan Chan Tseung H, Ma J and Beltran C 2015 A fast GPU-based Monte Carlo simulation of proton transport with detailed modeling of nonelastic interactions *Med. Phys.* **42** 2967–78