Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

PAPER • OPEN ACCESS

GPU accelerated Monte Carlo scoring of positron emitting isotopes produced during proton therapy for PET verification

To cite this article: Keegan McNamara *et al* 2022 *Phys. Med. Biol.* **67** 244001

IOP Publishing Phys. Med. Biol. 67 (2022) 244001 https://doi.org/10.1088/1361-6560/aca515

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

OPEN ACCESS

RECEIVED
3 May 2022

REVISED
11 November 2022

ACCEPTED FOR PUBLICATION
22 November 2022

PUBLISHED
12 December 2022

PAPER

GPU accelerated Monte Carlo scoring of positron emitting isotopes produced during proton therapy for PET verification

Keegan McNamara[1,2], Angelo Schiavi[3], Damian Borys[4,5], Karol Brzezinski[5], Jan Gajewski[5], Renata Kopeć[5], Antoni Rucinski[5], Tomasz Skóra[6], Shubhangi Makkar[1,2], Jan Hrbacek[1], Damien C Weber[1,7,8], Antony J Lomax[1,2] and Carla Winterhalter[1,2]

[1] Centre for Proton Therapy, Paul Scherrer Institute, Villigen, Switzerland
[2] Physics Department, ETH Zürich, Zürich, Switzerland
[3] Department of Basic and Applied Sciences for Engineering, Sapienza University of Rome, Rome, Italy
[4] Department of Systems Biology and Engineering, Silesian University of Technology, Gliwice, Poland
[5] Institute of Nuclear Physics Polish Academy of Sciences, Kraków, Poland
[6] Department of Radiotherapy, Maria Sklodowska-Curie National Research Institute of Oncology, Kraków Branch, Kraków, Poland
[7] Department of Radiation Oncology, Inselspital, Bern University Hospital, University of Bern, Switzerland
[8] Department of Radiation Oncology, University Hospital of Zürich, Switzerland

E-mail: carla.winterhalter@psi.ch

Keywords: PET range verification, GPU Monte Carlo, proton therapy

Abstract

Objective. Verification of delivered proton therapy treatments is essential for reaping the many benefits of the modality, with the most widely proposed *in vivo* verification technique being the imaging of positron emitting isotopes generated in the patient during treatment using positron emission tomography (PET). The purpose of this work is to reduce the computational resources and time required for simulation of patient activation during proton therapy using the GPU accelerated Monte Carlo code FRED, and to validate the predicted activity against the widely used Monte Carlo code GATE. *Approach.* We implement a continuous scoring approach for the production of positron emitting isotopes within FRED version 5.59.9. We simulate treatment plans delivered to 95 head and neck patients at Centrum Cyklotronowe Bronowice using this GPU implementation, and verify the accuracy using the Monte Carlo toolkit GATE version 9.0. *Main results.* We report an average reduction in computational time by a factor of 50 when using a local system with 2 GPUs as opposed to a large compute cluster utilising between 200 to 700 CPU threads, enabling simulation of patient activity within an average of 2.9 min as opposed to 146 min. All simulated plans are in good agreement across the two Monte Carlo codes. The two codes agree within a maximum of 0.95σ on a voxel-by-voxel basis for the prediction of 7 different isotopes across 472 simulated fields delivered to 95 patients, with the average deviation over all fields being $6.4 \times 10^{-3}\sigma$. *Significance.* The implementation of activation calculations in the GPU accelerated Monte Carlo code FRED provides fast and reliable simulation of patient activation following proton therapy, allowing for research and development of clinical applications of range verification for this treatment modality using PET to proceed at a rapid pace.

1. Introduction

In order to maximise the dosimetric and potential clinical benefits of proton therapy, robust and reliable verification of the proton range within patients is highly desirable (Knopf and Lomax 2013, Rucinski *et al* 2020). One of the most widely investigated verification techniques utilises positron emission tomography (PET) imaging of the patient, both post irradiation (Parodi *et al* 2005, 2007, Zhu *et al* 2011), and more recently on-line during treatment delivery (Piliero *et al* 2016, Buitenhuis *et al* 2017). Positron emitting isotopes (PEI) are generated within the body, and the subsequent PET image is correlated to the delivered dose. As such,

calculation of the expected activity distribution within the patient is required to connect the measured PET image to the dose delivered to the patient.

Analytical convolution techniques for calculation of the activation of the patient have been proposed in order to reduce dependency on computationally intensive Monte Carlo (MC) codes (Parodi and Bortfeld 2006, Attanasi *et al* 2009, Frey *et al* 2013), however these techniques may suffer from the same difficulties faced by analytic dose calculation engines in highly inhomogeneous regions (Schaffner *et al* 1999), and therefore may not provide the same verification accuracy as that of MC (Paganetti 2012). Many studies on patient activation utilise full MC toolkits based on the Geant4 code, such as GATE (Jan *et al* 2013, Robert *et al* 2013, Meiner *et al* 2019), and TOPAS (Perl *et al* 2012, Onecha *et al* 2022), or FLUKA (Parodi *et al* 2007, Augusto *et al* 2018). Such toolkits provide the most accurate modelling of all relevant physical processes, however are limited by large simulation times to achieve sufficient statistics for clinical plans. Such codes are CPU based, and speedup of simulations is possible by distributing calculations over a large number of computing nodes. Speedup is then limited by the available computing resources and the required statistics of the simulation. Flux based scoring techniques have previously been implemented in such codes using experimentally determined cross sections (Parodi *et al* 2007, Onecha *et al* 2022). Such techniques allow for simulation of fewer primary particles to calculate the PEI distribution, and can achieve better statistics thanks to variance reduction, however still rely on computationally intensive codes.

Daily adaptive therapy is of growing interest for proton therapy (Albertini *et al* 2020), however long computation times restrict the viability for adaptive use of PET activation calculations in treatment validation, which would ideally be performed on-line and while the patient is still on the treatment couch. As such, the possibility to produce accurate activation calculations in clinically compatible timeframes is essential for the future viability of range verification using PET alongside daily adaptive therapy.

The continuous improvement of general purpose GPU components has given rise to GPU based MC codes for radiotherapy dose calculations. FRED (Fast paRticle thErapy Dose evaluator) is one such tool enabling fast MC simulation of proton therapy plans in clinical settings (Schiavi *et al* 2017, De Simoni *et al* 2020), and is being expanded for use in carbon therapy (De Simoni *et al* 2022). Dose calculations in FRED were initially validated to existing general purpose MC codes Geant4 and FLUKA, and later against measurements in water and heterogeneous phantoms for commissioning as a TPS (Schiavi *et al* 2017, Garbacz *et al* 2019, Gajewski *et al* 2020, 2021). Calculation of isotope production is however heavily dependent on the flux of the protons as they pass through the patient, and so further investigation of the code is necessary to ensure that clinically reliable activation calculations are possible with FRED.

In this work we have utilised the plugin development tools of FRED to model continuous scoring of PEI within the patient on the GPU. GATE/Geant4 has been widely validated for clinical use in proton therapy, initially considering various physics processes in homogeneous materials and beam modelling (Grevillot *et al* 2010, 2011, Fuchs *et al* 2017, Resch *et al* 2019), before being applied in clinical practice (Aitkenhead *et al* 2020, Grevillot *et al* 2020, 2021). As the dose is directly dependent on the proton flux as a function of energy, as well as the stopping power of the protons, we assume that the determination of the proton flux within the patient predicted by GATE/Geant4 is sufficiently accurate for the simulation of PEI production. We therefore use GATE as a reference result against which we may validate FRED. We show that the production of PEI predicted by the two codes is in good agreement, allowing use of FRED for calculation of patient activation within clinically relevant timescales.

2. Methods

Throughout this study we consider the production of the PEI ^{10}C, ^{11}C, ^{13}N, ^{14}O, ^{15}O, ^{30}P, and ^{38}K, which have half lives between 19.3 and 1223 seconds, and are therefore relevant for imaging of patients following treatment delivery. In section 2.1 we describe how the GATE calculations were performed, and in section 2.2 we introduce the GPU scoring technique implemented in FRED. In section 2.3 we present the cross sections which are provided to FRED in order to match the available physics settings of GATE. In section 2.4 we present the geometries and fields which we simulated. In section 2.5 we discuss the Hounsfield Units (HU) to material conversion which were kept identical across the two simulations to provide a valid comparison. Following the recommendations for the reporting of Monte Carlo studies in medical physics made by Sechopoulos *et al* (2018) we report on the specific details of our simulations.

2.1. Isotope scoring using Geant4 and GATE

In a general physics MC code, such as Geant4, the production of isotopes is a discrete process (Allison *et al* 2006). An inelastic event occurs based on the inelastic nuclear scattering cross sections of the material for which the proton is stepping through. When an inelastic scattering event occurs, a nuclear cascade model is invoked, and

determines the resulting products of the event by sampling the various energetically accessible outcomes. Approximately 1% of protons undergo nuclear inelastic events per cm, and of these events only a fraction will go on to produce an isotope of interest. In order to produce a statistically reliable result for scoring of the production of isotopes a large number of primary protons must therefore be simulated. In this work 10% of the protons planned to be delivered were simulated.

GATE simulations were performed using version 9.0 with Geant4 version 10.6 patch 1. The simulations of patient plans were run on the Ziemowit HPC cluster, primarily utilising quanta nodes consisting of Intel E5-2660v3 CPUs, with 20 cores and 256 GB of RAM available per node. A number of simulations were also performed using IBM nodes consisting of Intel x5650 CPUs, with 12 cores and 36 GB of RAM per node. The GATE simulations were performed using the `QGSP_BIC_HP_EMY` physics settings. The `QGSP_BIC` hadronic model library is recommended for accurate production of secondary particles from proton and neutron interactions with nuclei. A step limiter of 10 mm was applied to all particles in the simulation. As the primary physics of interest were those of protons, a production cut for protons was set to 0.01 mm in the scoring region, while the production cut for photons, electrons, and positrons was set to 5 m. The scoring of the production of PEI within the simulations used the `ProductionAndStoppingActor` for each isotope of interest using the same voxel grid as the phantom or CT images.

2.2. GPU implementation of isotope scoring in FRED

In order to both reduce the number of primaries required for simulation, as well as take advantage of the GPU threading, a continuous generation of isotopes along each proton track was implemented. Instead of using the inelastic cross sections to determine the probability of an inelastic event occurring, the cross sections for individual reaction pathways are loaded into the GPU kernel. Along each step s of the proton track a fractional amount of isotope i is produced, we calculate the quantity

$$P_{i,v,s} = \sum_{e \in E_v} w_s x_{v,e} \sigma_{e \to i}(\bar{E}_s) l_s, \tag{1}$$

where E_v is the set of elements e in the voxel v, w_s is the weight of the current proton track, given by the number of protons in the beamlet divided by the number of protons simulated for that beamlet, $x_{v,e}$ is the molar fraction of the element within the voxel, and l_s is the step length. $\sigma_{e \to i}$ is the cross section for the reaction $e + p \to i + X$, where e accounts for the natural isotopic composition of the element and X indicates any possible fragmentation products, and is calculated at $\bar{E}_s$, the average energy between the start and end of the step. After tracking of all protons the quantity $P_{i,v} = \sum_s P_{i,v,s}$ has been scored for each isotope and voxel. This quantity is then scaled by the number of atoms within the current voxel, given by $N_A \rho_v / \bar{A}_v$, where $N_A = 6.022 \times 10^{23}$, ρ_v is the density of the voxel, and $\bar{A}_v$ is the average atomic weight of the elements in the voxel, giving the total production of isotope i in voxel v due to the delivered protons. We only consider production of PEI due to interactions from primary and secondary protons, in section 3.1 we discuss the validity of this assumption.

Our implementation removes the need for conditional statements within the GPU code, allowing the GPU threading to remain robust. As the implementation scores the average production expected from all steps, and not the discrete production from inelastic nuclear events, there is no change to the particle transport in the existing FRED code. We used the standard physics implementation of FRED, as described by Schiavi *et al* (2017).

2.3. Cross sections

Cross sections were calculated using Geant4, version 10.6 patch 1, with the `QGSP_BIC` model for protons incident on the most abundant elements in the human body; O, C, N, Ca, P, K, S, Cl, and Mg. For each element 10^8 protons were simulated incident at energies from 1 to 300 MeV with a spacing of 1 MeV in a uniform region of each element in its natural isotopic composition. Only inelastic nuclear scattering physics was enabled, with elastic and electromagnetic physics removed. The reaction products from each inelastic event were scored. Excited states of isotopes were scored as their ground state isotope. The production of isotopes such as ^{16}F, which immediately decays into ^{15}O, were included in the cross section calculation for their daughter isotope. The cross sections are then loaded into the GPU kernel as described in section 2.2, allowing a direct comparison between GATE simulations using `QGSP_BIC`, and FRED simulations.

2.4. Simulations

In order to validate the GPU accelerated code, we performed a number of simulations using both GATE and FRED. We first considered delivery of a single proton beam at the clinically relevant energy of 135 MeV delivered to homogenous phantoms consisting of water, PMMA, and Brain as defined by Woodard and White (1986), in blocks of size $5 \times 5 \times 20$ cm^3, with a voxel size of $1 \times 1 \times 1$ mm^3. The beam was directed along the z-axis.

A simulation using the CT of a CIRS head-and-neck phantom (model 731-HN) (Albertini *et al* 2011) with a cubic planning target volume (PTV) of 72 cm^3 and 2Gy RBE dose was performed. The CT was resampled to $2.5 \times 2.5 \times 2.5\ mm^3$ resolution.

Following this, the simulation of 95 head and neck cancer patients previously treated at Centrum Cyklotronowe Bronowice (CCB) within the Institute of Physics Polish Academy of Sciences was performed. The patient plans used within this work consisted of all consecutive head and neck plans treated from November 2016 to September 2018, and have previously been used to investigate biological range uncertainty by Garbacz *et al* (2022). The treatment plans each consisted of multiple fields, with some plans also including a boost, such that 472 fields in total were simulated. Fraction doses of 1.8-2Gy RBE, and one boost with a fraction dose of 1 Gy RBE, were delivered to volumes ranging from 11 to 1010 cm^3, covering a wide range of head and neck tumor cases. All CT scans were cropped to the external volume which contained all relevant structures for simulation, and resampled at $2.5 \times 2.5 \times 2.5\ mm^3$ resolution. The implementation of the beam models in FRED for two gantry rooms at CCB has been previously described by Gajewski *et al* (2021). The experimentally validated FRED beam model was adapted for use in GATE, i.e. the initial energy, energy spread, and Twiss parameters describing the lateral propagation of the beam (Twiss and Frank 1949), have been recalculated to the corresponding GATE parameters, giving an equivalent beam model in GATE. The equivalent beam models implemented in FRED and GATE have been previously used for dosimetric and radiation quality characterisation by Stasica *et al* (2020).

Each simulation in GATE was performed across multiple nodes, utilising between $N_C = 200 - 500$ threads per simulation. For the CIRS head and neck phantom we simulated 10% and 100% of the planned primary protons. For the patient plans we simulated 10% of the planned primary protons for each field. The PEI production for each thread was stored, allowing calculation of the total production within each voxel of the CT, as well as the standard deviation of these results. Each FRED simulation used 10^5 primaries per delivered pencil beam, such that on average for each field in the patient plans, 1.55% of the total number of planned primaries were simulated.

2.5. HU to material conversion

The CCB-specific clinical HU to material conversion was calculated based on the Schneider stoichiometric calibration method (Schneider *et al* 1996) for 93 human-tissue materials. The calibration contains information on HU, composition, density, relative proton stopping power, and the radiation length for each material. To reduce the total number of materials defined in GATE, HU values of similar density and composition were grouped into 421 bins of varying width. In order to compare the results between the two MC engines identical HU to material conversion tables were used.

3. Results

3.1. Simple geometries

In figure 1 an example of the integrated PEI production in a uniform Brain phantom from delivery of a single 135 MeV pencil beam is presented. Simulation of 10^5 primary protons in FRED converges rapidly to the limiting distribution, while in GATE simulation of 10^8 primaries, an increase by a factor of 1000, was necessary to achieve comparable results.

To quantify the differences between the simulations we considered the percentage difference of the integrated profiles of PEI production between FRED and GATE with respect to the maximum. We define the pass rate as the percentage of slices along the beam direction for which the agreement between FRED and GATE was within 5%. We consider only slices for which the production was at least 1% of the maximum. In table 1 we present the pass rate for all isotopes considered, and find that the pass rate increases as the number of primaries simulated in GATE increases. The low agreement of ^{30}P and ^{38}K is due to the overall lower number of isotopes produced, and the resulting statistical fluctuations, see figure 1.

We investigated whether simulation of only proton induced PEI production in FRED was sufficient to model the activation of the patient. Several secondary products from previous inelastic nuclear interactions, such as deuterons and neutrons, may undergo further reactions within the patient, contributing to PEI production. In table 2 we show the direct parent particle which produced the PEI of interest for the simulation of the brain phantom using 10^8 primaries. The relative contribution due to particles other than protons is considered to be negligible.

3.2. PEI production validation in a head phantom

The simulation of a cubic PTV delivered to a CIRS head and neck phantom was performed in GATE using $N_C = 400$ threads with 10% and 100% of the primary protons simulated, and in FRED using 10^5 primaries per delivered spot, allowing a thorough comparison of the results produced by the two MC codes. The planned field

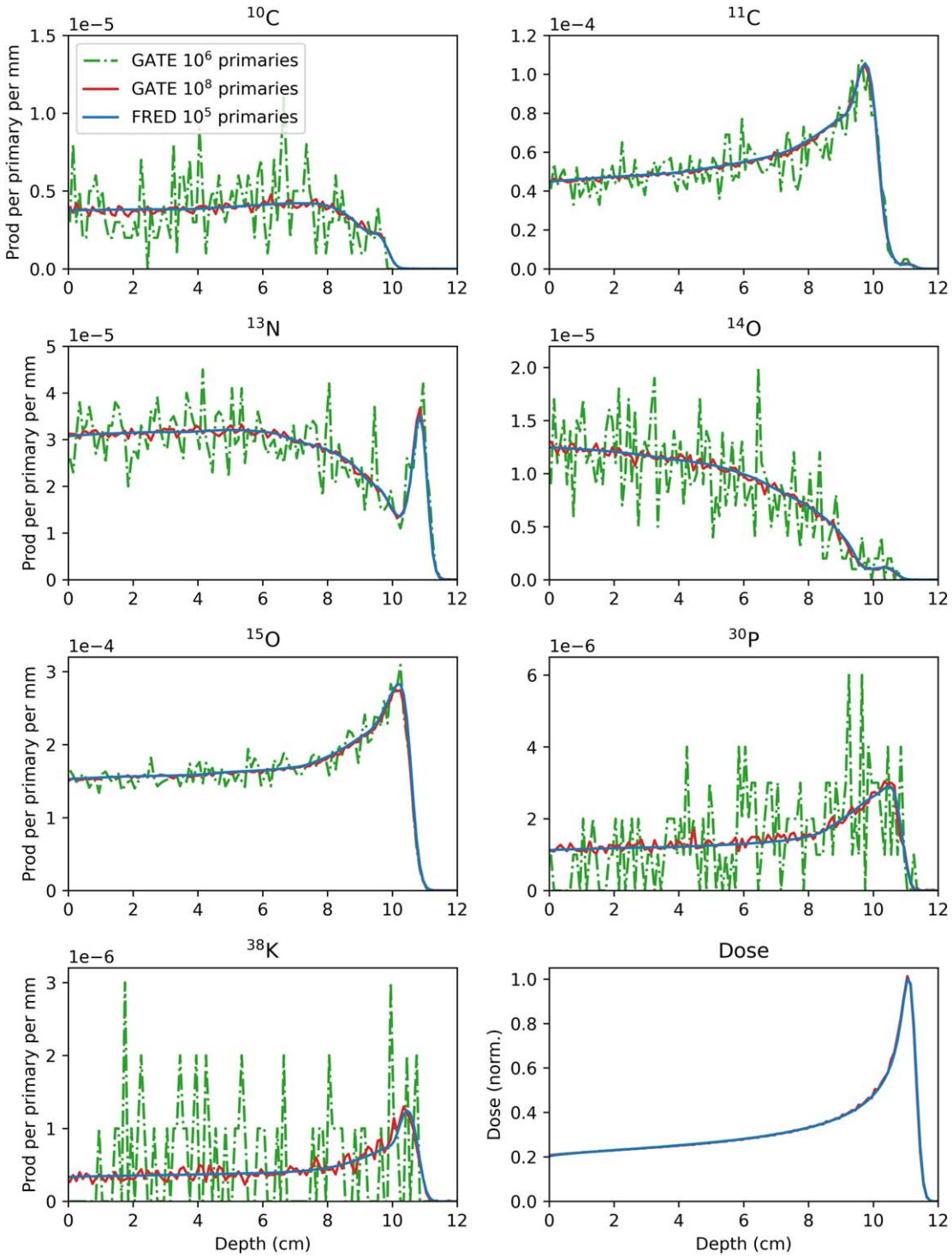

Figure 1. Integrated profiles of PEI production and dose for a 135 MeV beam in a uniform brain phantom. The FRED calculation was performed with 10^5 primary protons. Two GATE calculations are presented, with 10^6 and 10^8 primaries simulated.

consisted of 4.8×10^{10} protons delivered by 1676 spots, meaning that 0.35% of the total planned protons were simulated by FRED. An example of the production of ^{15}O, the isotope with the greatest overall production within the patient, is shown in figure 2. As indicated by table 2, there is some contribution to the total isotope production due to secondary particles such as neutrons. Such production is distributed over a broader spatial region, visible primarily past the distal edge in figure 2(b), and is not relevant to range verification.

In figure 2(c) we present the difference between the two slices shown in figures 2(a) and (b). The deviations between the two codes shows some structure, which is primarily caused by small deviations in the phase space of the beam upon entry into the phantom, and is more visible due to the regular shape of the field leading to an interference pattern. The relative difference between the two slices is at most 4% of the maximum. In figure 2(d)

Table 1. Pass rate for PEI prediction along the integrated profiles lying with 5% for a 135 MeV beam in brain, water, and PMMA. We compare FRED using 10^5 primaries to GATE using 10^6 and 10^8 primaries. ^{30}P and ^{38}K are not produced in water or PMMA.

	Brain		Water		PMMA	
Isotope	GATE 10^6	GATE 10^8	GATE 10^6	GATE 10^8	GATE 10^6	GATE 10^8
^{10}C	17.5%	76.7%	15.9%	72.9%	24.2%	97.8%
^{11}C	46.9%	100%	30.9%	99.1%	76.0%	99.0%
^{13}N	28.7%	99.1%	31.1%	97.5%	17.5%	91.3%
^{14}O	17.3%	96.4%	20.4%	94.2%	20.5%	81.8%
^{15}O	73.0%	97.3%	76.5%	96.5%	57.6%	96.0%
^{30}P	3.54%	77.0%				
^{38}K	1.77%	66.4%				

Table 2. Direct parent particles contributing to PEI production for a 135 MeV beam in a uniform brain phantom as predicted by GATE.

	Production per primary	p	n	d	α
^{10}C	3.70×10^{-4}	99.96%	0.037%		
^{11}C	5.96×10^{-3}	99.66%	0.34%	0.0006%	
^{13}N	3.22×10^{-3}	99.36%	0.59%	0.05%	
^{14}O	9.46×10^{-4}	99.92%	0.08%		
^{15}O	1.88×10^{-2}	99.66%	0.34%	0.001%	0.0004%
^{30}P	1.70×10^{-4}	98.61%	1.39%		
^{38}K	5.12×10^{-5}	99.25%	0.75%		

we show the total activity due to all scored isotopes along the profile indicated in figure 2. We see that the deviations between the predicted activity in all three cases is small. When considering the inherent uncertainties in measurement of the PET signal with a real scanner, the deviations may be considered to be negligible.

Every instance of GATE launched on the N_C threads produces an independent calculation of the isotope production for each field. We then calculate the mean production of isotope i in voxel v for a single GATE thread, $\lambda^g_{i,v}$, and the standard deviation, $\sigma^g_{i,v}$. The true result when simulating a large number of protons is given by drawing from a Poisson distribution with mean $\lambda_{i,v}$, such that $\lambda^g_{i,v}$ is the best estimate of $\lambda_{i,v}$. We scale the total value of the FRED calculation to give the comparable quantity $\lambda^f_{i,v} = f_p \mu^f_{i,v} / N_C$, where $\mu^f_{i,v}$ is the estimate of the total production of isotope i in voxel v by FRED, and f_p is the fraction of total primaries simulated for the GATE result which we compare to.

For most voxels $\lambda^g_{i,v} = 0$ and $\lambda^f_{i,v} < 1/N_C$. That is, the production of isotope i in voxel v is negligible in both simulations, and such voxels are considered to be in agreement. In order to validate the prediction of the FRED simulations we consider two other cases for each voxel; case 1 consists of all voxels for which $\lambda^g_{i,v} = 0$ and $\lambda^f_{i,v} \geqslant 1/N_C$, meaning that FRED predicts a non-negligible production of isotope within the given voxel while GATE does not. In case 2 GATE predicts a non-negligible production of isotope, $\lambda^g_{i,v} \geqslant 1/N_C$. Here $\sigma^g_{i,v} > 0$, and we calculate the ratio $(\lambda^f_{i,v} - \lambda^g_{i,v})/\sigma^g_{i,v}$ to assess how well the results agree given the statistical uncertainty. We note that here FRED may show a negligible production of isotope, $\lambda^f_{i,v} < 1/N_C$, in contrast to GATE. In such instances the production of PEI may be due to other reaction channels, as discussed in the previous section. We also consider that drawing from a Poisson distribution with mean $\lambda_{i,v} < 1/N_C$ may still produce a non-zero result when drawing N_C times, giving a value of $\lambda^g_{i,v} \geqslant 1/N_C$, hence such voxels are considered in agreement provided that $(\lambda^f_{i,v} - \lambda^g_{i,v})/\sigma^g_{i,v}$ is small. Throughout the following sections we report the average or maximum value of $r = (\lambda^f_{i,v} - \lambda^g_{i,v})/\sigma^g_{i,v}$ in the format $r\sigma$, where the average or maximum is calculated for each field and isotope over all relevant voxels.

As the production of PEI in a voxel is Poissonian, the statistical uncertainty $\sigma^g_{i,v}$ scales as $\sqrt{\lambda^g_{i,v}}$. Thanks to the variance reduction obtained by allowing every proton track to contribute to the PEI production, the uncertainty of $\lambda^f_{i,v}$ was negligible in comparison.

In figure 3(a) we show the distribution of $\lambda^f_{i,v}$ for all voxels and isotopes in the plan where $\lambda^g_{i,v}$ was 0. A low value of $\lambda^f_{i,v}$ suggests that when drawing N_C times from a Poisson distribution with mean $\lambda^f_{i,v}$, it is not unlikely to draw 0 every time. We perform a χ^2 test on the N_C observations of 0 compared to the expected result of $N_C e^{-\lambda^f_{i,v}}$ observations of 0. The p-value for the worst voxel in the 10% and 100% plans is 4.02×10^{-3} and 8.57×10^{-4}

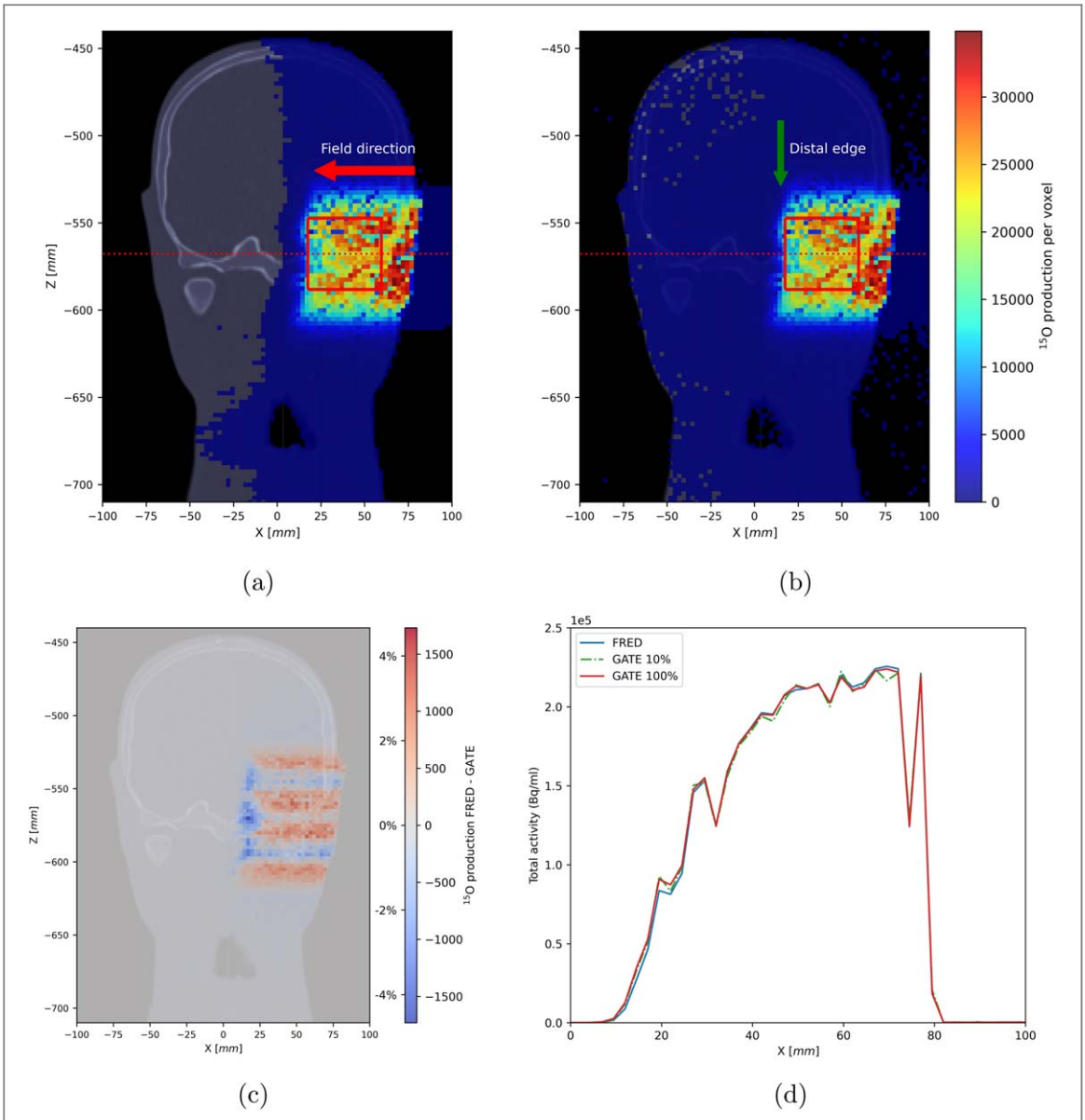

Figure 2. An example of the total production of ^{15}O per voxel calculated for a single field in a CIRS phantom in FRED (a) and GATE with 100% of primary protons simulated (b) with the cubic PTV contour shown in red. Past the distal edge of the field PEI production from neutrons is visible in GATE. This is not seen in FRED as we only consider the PEI production due to protons. The slice shown corresponds to the middle of the delivered field. (c) Voxel wise difference between (a) and (b), the colour bar shows the differences in absolute number of isotopes produced, as well as as a percentage of the maximum voxel. (d) An example of the total activity due to all isotopes along the line profile indicated in (a) and (b), as predicted by FRED, and GATE with 10% and 100% of the primary protons simulated.

respectively. This suggests that the over-prediction of FRED is indistinguishable from statistical chance for most of these voxels.

We then consider the voxels for which PEI production was observed in GATE. In figure 3(b) and (c) we show the distribution of $(\lambda^f_{i,v} - \lambda^g_{i,v})/\sigma^g_{i,v}$ for both ^{15}O and ^{11}C production for all voxels where isotope production was more than 1% of the maximum. For all voxels in both plans the deviation between the two results is within one standard deviation, 0.364σ in the 10% plan and 0.813σ in the 100% plan, suggesting that there is good agreement. We note that there is a small systematic over-prediction of the production by FRED when comparing to the simulation of 100% of primaries, most clearly visible in prediction of ^{11}C and ^{15}O. The mean result over all voxels is 0.031σ (0.049σ) for ^{11}C (^{15}O) with 10% of protons simulated, but increases to 0.092σ (0.151σ) when 100% of protons are simulated. The increase in statistics by a factor of 10 reduces $\sigma^g_{i,v}$ by a factor of $\sqrt{10}$, and so the increase in overall deviation is expected. The systematic deviation between the results is likely caused by small differences in the flux of the protons as simulated by the two MC engines, as well as possible differences in the beam model used. We also note that other uncertainties, such as that of the elemental composition of each voxel, have a larger impact on the prediction of the PEI, and outweigh uncertainties on this scale.

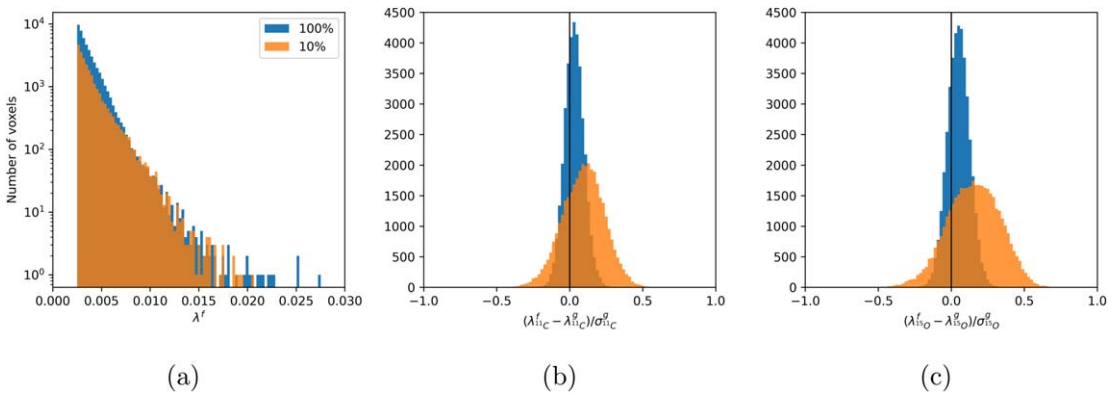

Figure 3. (a) Histogram of $\lambda^f_{i,v}$ for all voxels and isotopes for which $\lambda^g_{i,v} = 0$ and $\lambda^f_{i,v} > 1/N_C$. Histogram of $(\lambda^f_{i,v} - \lambda^g_{i,v})/\sigma^g_{i,v}$ for (b) $i = {}^{11}$C, and (c) $i = {}^{15}$O.

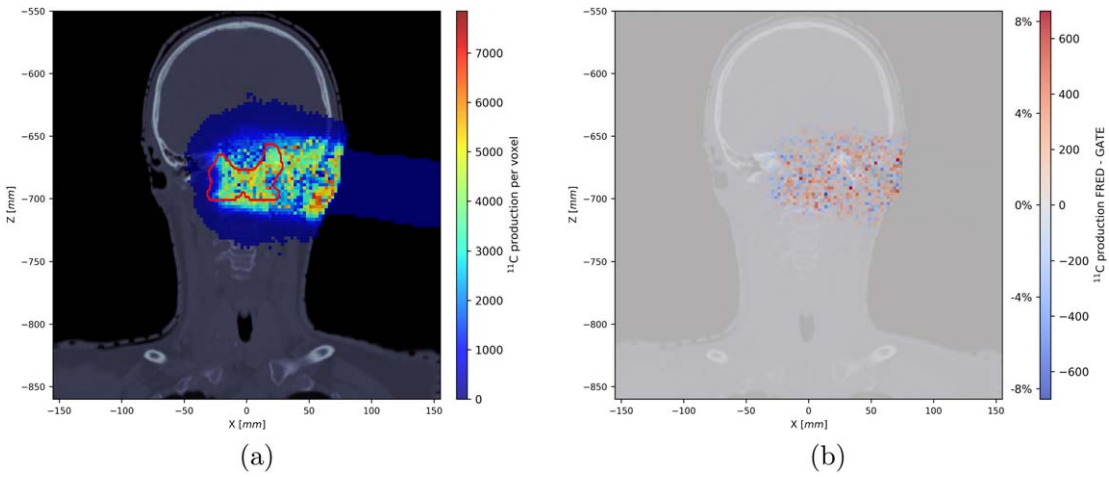

Figure 4. (a) Prediction of ^{11}C production in a patient as predicted by FRED, the PTV is indicated in red, (b) voxel wise difference between FRED and GATE for the same slice as in (a).

Given the high computational requirements for simulation of a single field, as well as the satisfactory convergence of the isotope production for most cases, simulation of 10% of the planned protons was assumed to provide sufficient statistics for comparison of all other plans. This is a factor of 100 larger than the number used for MC calculations of dose by Winterhalter *et al* (2019).

3.3. Head and neck cancer patients study

Following the analysis introduced in the preceding section, we now consider the agreement between the two MC codes for all 472 fields simulated for the 95 patient plans considered. We show an example of the ^{11}C production in a patient in figure 4(a), as predicted by FRED, as well as the voxel wise difference of this distribution to GATE in figure 4(b). In figure 5 we present the largest value of $\lambda^f_{i,v}$ for each field and each isotope for voxels where $\lambda^g_{i,v} = 0$. Figure 5(a) shows that the worst case voxel has generally low values of $\lambda^f_{i,v}$, suggesting that differences are in general minor. The maximum $\lambda^f_{i,v}$ for which $\lambda^g_{i,v} = 0$ from all 472 fields is 0.0603, with a p-value of 4.21×10^{-4}. We also consider a neighbourhood average of the $3 \times 3 \times 3$ voxels surrounding each worst case voxel in figure 5(b). The lack of any systematic impact when considering these surrounding voxels is clear, indicating that differences are statistical noise and not indicative of any measurable effects.

In figure 6 we present the average and maximum value of $(\lambda^f_{i,v} - \lambda^g_{i,v})/\sigma^g_{i,v}$ over all voxels for each field and isotope. It is immediately obvious that across all 472 fields and all 7 isotopes, the two codes show no meaningful differences. The average deviation across all fields is $6.4 \times 10^{-3}\sigma$, and the worst case over all voxels in all 472 plans is 0.951σ, suggesting that across all considered fields the two codes are in good agreement. As in the previous section, an increase in the simulated statistics would decrease $\sigma^g_{i,v}$, increasing the deviation between

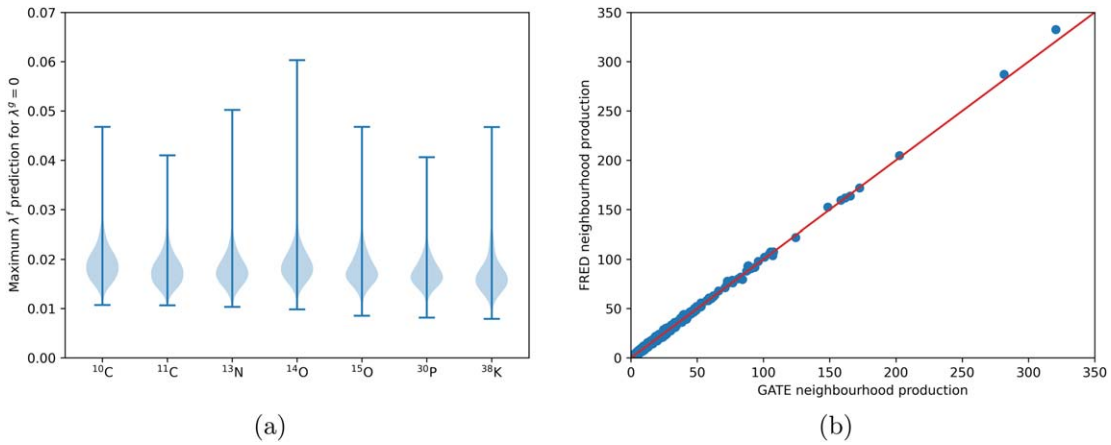

Figure 5. (a) Violin plot showing the distribution of the worst case prediction of λ^f when $\lambda^g = 0$ across all fields. (b) The average isotope production in the $3 \times 3 \times 3$ cube of voxels surrounding the worst case prediction in both FRED and GATE, the line of FRED = GATE is shown for comparison.

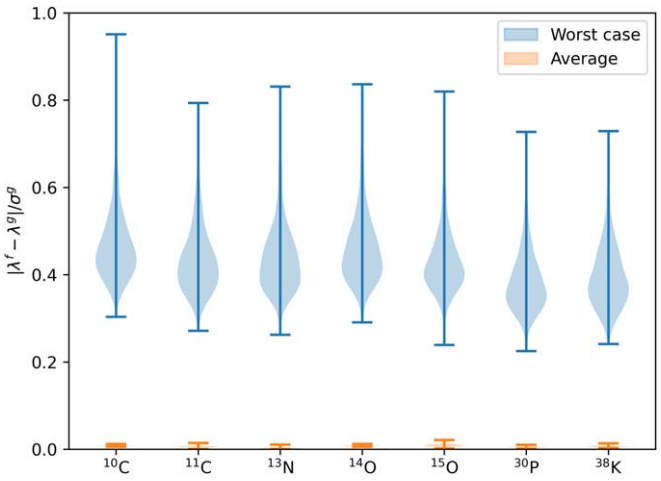

Figure 6. Violin plot of the distribution of the largest value of $|\lambda^f_{i,v} - \lambda^g_{i,v}|/\sigma^g_{i,v}$ (worst case) for each field, as well as the absolute value of the average of $(\lambda^f_{i,v} - \lambda^g_{i,v})/\sigma^g_{i,v}$ over all voxels for each field.

FRED and GATE, however it would not impact the spatial profile of the expected activity, and would require months to simulate.

3.4. Timing and resources

The main motivation for implementation of PEI production scoring capabilities using GPU accelerated codes is the significant decrease in total simulation time. We note that an equivalent variance reduction technique has previously been introduced into the FLUKA code (Parodi *et al* 2007), and was also recently implemented in TOPAS for use in generating a dictionary of activities for dose reconstruction (Onecha *et al* 2022), however the MC simulation was notably still the most time consuming step. Such an approach may also be implemented for the GATE toolkit. For a fair comparison, we therefore consider the overall speed improvements both from the consideration of variance reduction requiring simulation of fewer total primaries, as well as the increase in number of simulated primaries per second (PPS). For each field it was found that simulation of 10^5 primaries per delivered pencil beam in FRED provided rapid convergence to the final estimated production of all isotopes. This corresponds to, on average, simulation of 1.55% of the total primaries, a reduction by a factor of 6.4 compared to the 10% used in our GATE simulations. Simulation of 10% of primary particles in GATE results in larger statistical fluctuations than using 1.55% of primaries within FRED, such that comparable uncertainties would require simulation of an even greater number of primaries in GATE.

In figure 7 we present the distribution of the calculation time per field for the two codes. The average calculation time in FRED was 2.9 min, compared to 146 min for GATE. Since each GATE calculation was

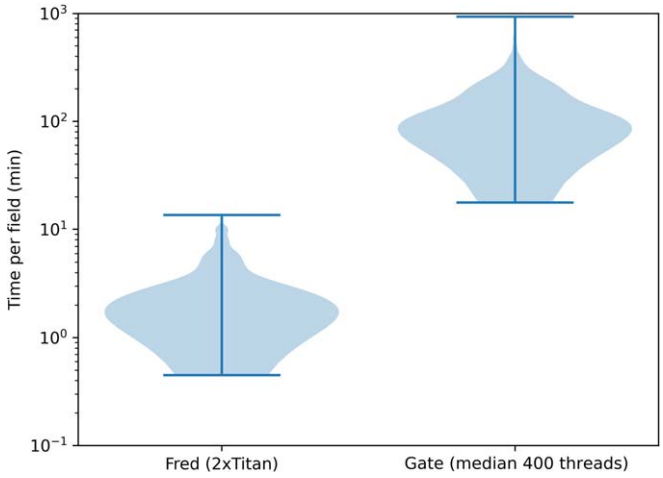

Figure 7. Distribution of the calculation time per field for all 472 fields.

Table 3. Estimated time for simulation of a single field consisting of 4.8×10^{10} primary protons using a variety of computational setups and differing percentages of primaries simulated.

	PPS	1.55%	10%	100%
GATE single CPU thread	870	9.6 days	64 days	1.7 years
GATE 16 CPU threads	1.39×10^4	14 h	4 days	40 days
GATE 400 CPU threads	3.46×10^5	34 min	3.9 h	38 h
FRED NVIDIA P2200 GPU	3.31×10^5	36 min	4 h	40 h
FRED 2×NVIDIA TITAN X GPUs	2.41×10^6	5.2 min	33.5 min	5.6 h
FRED NVIDIA RTX A5000	3.78×10^6	3.2 min	21.2 min	3.5 h

distributed over a large number of threads, the delay of one thread may increase the total calculation time. We therefore consider the time for each GATE calculation to be the average simulation time across all N_C threads of the simulation, instead of the maximum time. In table 3 we estimate the expected time requirements for simulation of an intermediate sized field using GATE on differing numbers of CPU threads, as well as FRED using three different GPU setups currently available to us. The number of PPS on a single GATE thread was, on average, 870. A single high end machine may have access to 16 CPU cores, allowing for a possible factor of 16 increase in computing speed. However even with such resources, inclusion of the given variance reduction, and optimisation of simulation settings for increased PPS, a typical plan may still take several hours to simulate. With access to a large compute cluster, which is both costly to maintain and energy intensive, we may simulate on the order of 3×10^5 PPS, giving calculations within reasonable time frames. However this is on par with the PPS possible on a lower cost NVIDIA P2200 GPU, enabling such simulations to be performed on a single computer with significantly lower energy and cost requirements. FRED simulations performed using two NVIDIA TITAN X GPUs, or a single NVIDIA RTX A5000, provide simulation of the whole field within 5.2 and 3.2 min respectively. Improvements in GPU technology over the past decade mean that simulations are already achievable within clinically useful time frames. Though improvements in the PPS of general purpose MC codes are achievable, the massive parallelism afforded to GPU based codes currently make this the most attractive solution.

As the implementation of PEI scoring did not alter the existing GPU threading of FRED, the scoring of dose as well as the seven PEI of interest took approximately 20% longer in comparison to scoring dose alone.

4. Discussion

We have implemented PEI scoring capabilities into the GPU accelerated MC code FRED through use of continuous scoring of isotope production along each step of the protons within the simulation. In sections 3.1, 3.2, and 3.3 we have shown that the prediction of production of PEI is in excellent agreement between FRED and GATE for a wide variety of phantoms and clinical cases. We showed for simple materials that FRED predicted PEI distributions in agreement with the limiting results of GATE. We then introduced a comparison of the two codes for a single field in a CIRS head phantom, showing that 10% of the primary protons simulated in GATE

provided a sufficient point of comparison for further simulations. We then continued the analysis of the code by showing that for 472 fields delivered to 95 head and neck cancer patients the two results showed no statistically relevant differences, with the average deviation between the predicted PEI on a voxel by voxel basis being $6.4 \times 10^{-3}\sigma$, with the worst case voxel in all simulated plans being 0.95σ.

The small differences that do exist between the two codes are presumably caused by differences in the underlying proton transport and scattering physics of the two codes, lack of neutron and other particle transport in FRED, and possible differences in the beam model definition. No deviations relevant for the application of the code to prediction of the PET activation within the patient occur. This indicates that implementation of FRED into clinical workflows for the purpose of calculating predicted PET activity distributions inside the patient, both on-line and post irradiation, is not only possible, but would be highly recommended if *in vivo* range verification should be introduced routinely in the clinical management of these patients. Further validation against PET measurements will be necessary and is ongoing. The validation performed in this work also indicates that the proton flux predicted by FRED is consistent with that of GATE, and extension to prediction of PEI production for other isotopes not considered in this work is possible. To our knowledge this is the first time such an approach has been implemented and validated in a GPU MC code.

Throughout this work we have compared FRED to the `QGSP_BIC` hadronic models of Geant4, however the scoring of activation within FRED may also be performed with any cross sections, which can be loaded into the GPU kernel at run time. This allows use of in house or experimentally measured cross sections as in (Parodi *et al* 2007, Onecha *et al* 2022). We have also only considered production of isotopes relevant for offline imaging, however scoring of any isotope of interest, for example ^{12}N which is important for on-line imaging (Ozoemelam *et al* 2020), is possible. The reaction cross sections for production of ^{12}N are less well known (Buitenhuis *et al* 2017), therefore the speed at which calculations are performed may allow for investigation of the validity of cross sections by comparison to experiments similar to work by Matsushita *et al* (2016).

In section 3.4 we reported the massive increase in the number of primaries simulated per second. The financial and energy cost of simulations are also significantly reduced, allowing for easier implementation of PEI activation calculations in treatment planning and validation workflows. The ease of performing calculations on a single machine as opposed to a large compute cluster is also of practical benefit. Distribution of MC jobs to a large number of nodes means that each node will produce output files which must, post simulation, be aggregated to give a final result. The memory and data transfer requirements, though not inherently challenging, must therefore also be taken into account. The ability to simulate and analyse the result of a simulation on a single, local, machine is therefore less demanding.

During the implementation and validation of this work FRED was integrated into the ProTheRaMon framework, which allows for the simulation of PEI production in the patient using GATE or FRED, as well as tools to simulate PET detector response and reconstruction, giving a complete simulation of the range verification process (Borys *et al* 2022). Inclusion of detector and reconstruction modelling will allow for validation against experimental measurements in future work.

The large improvement in speed, bringing accurate MC predictions of activity into clinically relevant time frames, may enable the use of PET verification for daily adaptive therapy going forward.

5. Conclusion

The implementation of PEI scoring in the GPU accelerated Monte Carlo code FRED has been validated with GATE for a number of simulations in phantoms as well as 472 fields delivered for 95 head and neck cancer patients at CCB. For each simulation the predicted PEI production was compared on a voxel by voxel basis. The deviation between the two results was within a maximum of 0.95σ, and was $6.4 \times 10^{-3}\sigma$ on average, showing good agreement for all fields and all isotopes considered. The good agreement for all simulations suggests that FRED can be reliably used to calculate the production of other isotopes of interest which are produced during proton therapy, and a reduction in the necessary computational time by a factor of 50 using significantly fewer computational resources is seen. The reduction in computational time and resource requirements while achieving high accuracy will allow use of FRED in prediction of PEI for both research and future clinical developments.

Acknowledgments

This work was supported by the Swiss National Science Foundation, Grant No. CRSII5189969, and the National Centre for Research and Development (NCBiR), grant no. LIDER/26/0157/L-8/16/NCBR/2017.

Calculations were performed on the Ziemowit computer cluster in the Laboratory of Bioinformatics and Computational Biology, Silesian University of Technology, created in the EU Innovative Economy Programme POIG.02.01.00-00-166/08 and expanded in the POIG.02.03.01-00-040/13 project.

ORCID iDs

Keegan McNamara https://orcid.org/0000-0002-2281-7121
Angelo Schiavi https://orcid.org/0000-0002-7081-2747
Damian Borys https://orcid.org/0000-0003-0229-2601
Karol Brzezinski https://orcid.org/0000-0002-9795-5158
Jan Gajewski https://orcid.org/0000-0002-7416-5145
Renata Kopeć https://orcid.org/0000-0002-0919-9859
Antoni Rucinski https://orcid.org/0000-0002-5815-4606
Tomasz Skóra https://orcid.org/0000-0001-6322-0615
Shubhangi Makkar https://orcid.org/0000-0002-0752-9295

References

Aitkenhead A H, Sitch P, Richardson J C, Winterhalter C, Patel I and Mackay R I 2020 Automated Monte-Carlo re-calculation of proton therapy plans using Geant4/Gate: implementation and comparison to plan-specific quality assurance measurements *Br. J. Radiol.* **93** 20200228

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

Albertini F, Matter M, Nenoff L, Zhang Y and Lomax A 2020 Online daily adaptive proton therapy *Br. J. Radiol.* **93** 20190594

Allison J *et al* 2006 Geant4 developments and applications *IEEE Trans. Nucl. Sci.* **53** 270–8

Attanasi F, Knopf A, Parodi K, Bortfeld T, Paganetti H, Rossoxy V and Guerra A Del 2009 Clinical validation of an analytical procedure for in vivo PET range verification in proton therapy *IEEE NSS Conf. Rec.* pp 4167–71

Augusto R *et al* 2018 An overview of recent developments in FLUKA PET tools *Phys. Med.* **54** 189–99

Borys D *et al* 2022 ProTheRaMon—a GATE simulation framework for proton therapy range monitoring using PET imaging *Phys. Med. Biol.* **67** 224002

Buitenhuis H J T, Diblen F, Brzezinski K W, Brandenburg S and Dendooven P 2017 Beam-on imaging of short-lived positron emitters during proton therapy *Phys. Med. Biol.* **62** 4654–72

De Simoni M, Fischetti M, Gioscio E, Marafini M, Mirabelli R, Patera V, Sarti A, Schiavi A, Sciubba A and Traini G 2020 FRED: a fast Monte Carlo code on GPU for quality control in Particle Therapy *J. Phys. Conf. Ser.* **1548** 012020

De Simoni M *et al* 2022 A data-driven fragmentation model for carbon therapy gpu-accelerated monte-carlo dose recalculation *Front. Oncol.* **12** 780784

Frey K, Bauer J, Unholtz D, Kurz C, Krämer M, Bortfeld T and Parodi K 2013 TPSPET—A TPS-based approach for in vivo dose verification with PET in proton therapy *Phys. Med. Biol.* **59** 1–21

Fuchs H, Vatnitsky S, Stock M, Georg D and Grevillot L 2017 Evaluation of GATE/Geant4 multiple Coulomb scattering algorithms for a 160 MeV proton beam *Nucl. Instrum. Methods Phys. Res. B: Beam Interact. Mater. At.* **410** 122–6

Gajewski J, Schiavi A, Krah N, Vilches-Freixas G, Rucinski A, Patera V and Rinaldi I 2020 Implementation of a compact spot-scanning proton therapy system in a GPU Monte Carlo code to support clinical routine *Front. Phys.* **8** 578605

Gajewski J *et al* 2021 Commissioning of GPU-accelerated monte carlo code fred for clinical applications in proton therapy *Front. Phys.* **8** 567300

Garbacz M *et al* 2019 Proton therapy treatment plan verification in CCB krakow using fred monte carlo TPS tool *World Congress on Medical Physics and Biomedical Engineering 2018* 68/1 ed L Lhotska *et al* pp 783–787

Garbacz M *et al* 2022 Quantification of biological range uncertainties in patients treated at the Krakow proton therapy centre *Radiat. Oncol.* **17** 50

Grevillot L, Bertrand D, Dessy F, Freud N and Sarrut D 2011 A Monte Carlo pencil beam scanning model for proton treatment plan simulation using GATE/GEANT4 *Phys. Med. Biol.* **56** 5203–19

Grevillot L, Boersma D J, Fuchs H, Bolsa-Ferruz M, Scheuchenpflug L, Georg D, Kronreif G and Stock M 2021 The GATE-RTion/IDEAL independent dose calculation system for light ion beam therapy *Front. Phys.* **9** 704760

Grevillot L, Frisson T, Zahra N, Bertrand D, Stichelbaut F, Freud N and Sarrut D 2010 Optimization of GEANT4 settings for Proton Pencil Beam Scanning simulations using GATE *Nucl. Instrum. Methods Phys. Res. B: Beam Interact. Mater. At.* **268** 3295–305

Grevillot L *et al* 2020 Technical Note: GATE-RTion: a GATE/Geant4 release for clinical applications in scanned ion beam therapy *Med. Phys.* **47** 3675–81

Jan S *et al* 2011 GATE V6: a major enhancement of the GATE simulation platform enabling modelling of CT and radiotherapy *Phys. Med. Biol.* **56** 881–901

Knopf A-C and Lomax A 2013 In vivo proton range verification: a review *Phys. Med. Biol.* **58** R131–60

Matsushita K, Nishio T, Tanaka S, Tsuneda M, Sugiura A and Ieki K 2016 Measurement of proton-induced target fragmentation cross sections in carbon *Nucl. Phys.* A **946** 104–16

Meiner H, Fuchs H, Hirtl A, Reschl C and Stock M 2019 Towards offline PET monitoring of proton therapy at MedAustron *Z. Med. Phys.* **29** 59–65

Onecha V V, Galve P, Ibáñez P, Freijo C, Arias-Valcayo F, Sanchez-Parcerisa D, España S, Fraile L M and Udías J M 2022 Dictionary-based software for proton dose reconstruction and submilimetric range verification *Phys. Med. Biol.* **67** 045002

Ozoemelam I, van der Graaf E, van Goethem M-J, Kapusta M, Zhang N, Brandenburg S and Dendooven P 2020 Feasibility of quasi-prompt PET-based range verification in proton therapy *Phys. Med. Biol.* **65** 245013

Paganetti H 2012 Range uncertainties in proton therapy and the role of Monte Carlo simulations *Phys. Med. Biol.* **57** R99–117

Parodi K and Bortfeld T 2006 A filtering approach based on Gaussian–powerlaw convolutions for local PET verification of proton radiotherapy *Phys. Med. Biol.* **51** 1991–2009
Parodi K, Ferrari A, Sommerer F and Paganetti H 2007 Clinical CT-based calculations of dose and positron emitter distributions in proton therapy using the FLUKA Monte Carlo code *Phys. Med. Biol.* **52** 3369–87
Parodi K, Ponisch F and Enghardt W 2005 Experimental study on the feasibility of in-beam PET for accurate monitoring of proton therapy *IEEE Trans. Nucl. Sci.* **52** 778–86
Parodi K *et al* 2007 Patient study of in vivo verification of beam delivery and range, using positron emission tomography and computed tomography imaging after proton therapy *Int. J. Radiat. Oncol. Biol. Phys.* **68** 920–34
Perl J, Shin J, Schümann J, Faddegon B and Paganetti H 2012 TOPAS: An innovative proton Monte Carlo platform for research and clinical applications *Med. Phys.* **39** 6818–37
Piliero M *et al* 2016 First results of the INSIDE in-beam PET scanner for the on-line monitoring of particle therapy treatments *J. Inst.* **11** C12011
Resch A F, Elia A, Fuchs H, Carlino A, Palmans H, Stock M, Georg D and Grevillot L 2019 Evaluation of electromagnetic and nuclear scattering models in GATE/Geant4 for proton therapy *Med. Phys.* **46** 2444–56
Robert C, Fourrier N, Sarrut D, Stute S, Gueth P, Grevillot L and Buvat I 2013 PET-based dose delivery verification in proton therapy: a GATE based simulation study of five PET system designs in clinical conditions *Phys. Med. Biol.* **58** 6867–85
Rucinski A *et al* 2020 Investigations on physical and biological range uncertainties in Kraków proton beam therapy centre *Acta Phys. Pol.* B **51** 9–16
Schaffner B, Pedroni E and Lomax A 1999 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.* **44** 27–41
Schiavi A, Senzacqua M, Pioli S, Mairani A, Magro G, Molinelli S, Ciocca M, Battistoni G and Patera V 2017 Fred: a GPU-accelerated fast-Monte Carlo code for rapid treatment plan recalculation in ion beam therapy *Phys. Med. Biol.* **62** 7482–504
Schneider U, Pedroni E and Lomax A 1996 The calibration of CT Hounsfield units for radiotherapy treatment planning *Phys. Med. Biol.* **41** 111–24
Sechopoulos I, Rogers D W O, Bazalova-Carter M, Bolch W E, Heath E C, McNitt-Gray M F, Sempau J and Williamson J F 2018 RECORDS: improved reporting of monte carlo radiation transport studies: report of the aapm research committee task group 268 *Med. Phys.* **45** e1–5
Stasica P *et al* 2020 A simple approach for experimental characterization and validation of proton pencil beam profiles *Front. Phys.* **8** 346
Twiss R Q and Frank N H 1949 Orbital stability in a proton synchrotron *Rev. Sci. Instrum.* **20** 1–17
Winterhalter C, Meier G, Oxley D, Weber D C, Lomax A J and Safai S 2019 Log file based Monte Carlo calculations for proton pencil beam scanning therapy *Phys. Med. Biol* **64** 035014
Woodard H Q and White D R 1986 The composition of body tissues *Br. J. Radiol.* **59** 1209–18
Zhu X, España S, Daartz J, Liebsch N, Ouyang J, Paganetti H, Bortfeld T R and Fakhri G E 2011 Monitoring proton radiation therapy with in-room PET imaging *Phys. Med. Biol.* **56** 4041–57