Enhancing image quality in fast neutron-based range verification of proton therapy using a deep learning-based prior in LM-MAP-EM reconstruction

Setterdahl, L. M.; Skjerdal, Kyrre; Ratliff, Hunter N.; Ytre-Hauge, Kristian Smeland; Lionheart, William R.B.; Holman, Sean; Pettersen, Helge E.S.; Blangiardi, Francesco; Lathouwers, Danny; Meric, Ilker

DOI
10.1088/1361-6560/ade198

Publication date
2025

Document Version
Final published version

Published in
Physics in medicine and biology

IOP Publishing Phys. Med. Biol. 70 (2025) 125011 https://doi.org/10.1088/1361-6560/ade198

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

RECEIVED
14 March 2025

REVISED
27 May 2025

ACCEPTED FOR PUBLICATION
5 June 2025

PUBLISHED
17 June 2025

PAPER

Enhancing image quality in fast neutron-based range verification of proton therapy using a deep learning-based prior in LM-MAP-EM reconstruction

Lena M Setterdahl[1,*], Kyrre Skjerdal[1], Hunter N Ratliff[1], Kristian Smeland Ytre-Hauge[2], William R B Lionheart[3], Sean Holman[3], Helge E S Pettersen[4], Francesco Blangiardi[5], Danny Lathouwers[6] and Ilker Meric[1,*]

[1] Department of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, Bergen, Norway
[2] Department of Physics and Technology, University of Bergen, Bergen, Norway
[3] Department of Mathematics, University of Manchester, Manchester, United Kingdom
[4] Cancer Clinic, Haukeland University Hospital, Bergen, Norway
[5] Fraunhofer Institute for Electronic Nanosystems, Chemnitz, Germany
[6] Department of Radiation Science and Technology, Delft University of Technology, Delft, The Netherlands
[*] Authors to whom any correspondence should be addressed.

E-mail: lena.marie.setterdahl@hvl.no and ilker.meric@hvl.no

Keywords: proton therapy, range verification, fast neutron imaging, organic scintillators, Monte Carlo simulation, generative adversarial network, LM-MAP-EM

Abstract

Objective. This study investigates the use of list-mode (LM) maximum *a posteriori* (MAP) expectation maximization (EM) incorporating prior information predicted by a convolutional neural network for image reconstruction in fast neutron (FN)-based proton therapy range verification. *Approach.* A conditional generative adversarial network (pix2pix) was trained on progressively noisier data, where detector resolution effects were introduced gradually to simulate realistic conditions. FN data were generated using Monte Carlo simulations of an 85 MeV proton pencil beam in a computed tomography-based lung cancer patient model, with range shifts emulating weight gain and loss. The network was trained to estimate the expected two-dimensional ground truth FN production distribution from simple back-projection images. Performance was evaluated using mean squared error, structural similarity index (SSIM), and the correlation between shifts in predicted distributions and true range shifts. *Main results.* Our results show that pix2pix performs well on noise-free data but suffers from significant degradation when detector resolution effects are introduced. Among the LM-MAP-EM approaches tested, incorporating a mean prior estimate into the reconstruction process improved performance, with LM-MAP-EM using a mean prior estimate outperforming naïve LM maximum likelihood EM (LM-MLEM) and conventional LM-MAP-EM with a smoothing quadratic energy function in terms of SSIM. *Significance.* Findings suggest that deep learning techniques can enhance iterative reconstruction for range verification in proton therapy. However, the effectiveness of the model is highly dependent on data quality, limiting its robustness in high-noise scenarios.

1. Introduction

Proton therapy has emerged as a promising cancer treatment modality, offering precise dose delivery and reduced side effects compared to conventional radiotherapy. This advantage stems from the sharp dose gradients of the Bragg peak, allowing for reduced entrance dose, virtually no exit dose, and highly localized deposition of maximum dose to the targeted tumor volume. However, proton therapy's full potential is constrained by uncertainties in the proton beam range, necessitating safety margins of several millimeters around the target. These uncertainties arise from various sources, including patient motion, setup errors,

anatomical changes, tissue density variations, and Hounsfield unit-to-stopping power conversion (Paganetti 2018).

Real-time, *in vivo*, verification of the beam range could potentially reduce these uncertainties (Knopf and Lomax 2013) and, consequently, minimize the dose delivered to healthy tissue. As such, *in vivo* proton therapy range verification has become an active field of research. One of the earliest approaches used positron emission tomography (PET)-based imaging of 511 keV coincident gamma rays produced by positron emitters generated by fragmentation of target nuclei in inelastic proton-nucleus collisions (Enghardt *et al* 2004). To address the intrinsic time delay in PET imaging due to the half-life of positron emitters, Stichelbaut and Jongen proposed imaging prompt gamma (PG) rays resulting from inelastic proton-nucleus interactions (Stichelbaut and Jongen 2003). These interactions leave the target nucleus in an excited state, which then de-excites by promptly emitting characteristic gamma rays. Several PG imaging (PGI) systems have been proposed, including PG spectroscopy (PGS), PG timing (PGT), slit cameras, and Compton cameras (CC) (Polf *et al* 2009, Moteabbed *et al* 2011, Bom *et al* 2012, Smeets *et al* 2012, Golnik *et al* 2014, Draeger *et al* 2018, Hueso-González *et al* 2018). More recently explored is the imaging of fast neutrons (FNs), also produced in proton-nucleus interactions, for range verification and assessment of neutron dose distribution in the patient (Marafini *et al* 2017, Ytre-Hauge *et al* 2019, Lerendegui-Marco *et al* 2022, Ma *et al* 2022).

The NOVO project, to which this work is related, proposes a novel approach for simultaneous detection of both PGs and FNs for dose and range verification in proton therapy (Meric *et al* 2023). This method's rationale lies in the increased event counts from observing two particle species, with FNs and PGs providing complementary information on the proton beam path. FNs are predominantly created earlier in the beam path, while PGs are generated later, with peak production near the Bragg peak, offering a more comprehensive view of the dose distribution. The conceptual design of the proposed system, named NOVCoDA (the NOVO compact detector array) is a densely stacked array of organic scintillator bars coupled to dual-ended light readout units.

For FNs, the system detects two consecutive elastic collisions with a hydrogen nucleus (i.e. proton), while for PG rays, it records three consecutive interactions, either three Compton scatters or two Compton scatters and one photoabsorption interaction. These interactions enable the reconstruction of cone-shaped surfaces representing potential particle origins. These 'event cones' are then back-projected into space to create a simple back-projection (SBP) image (Wilderman *et al* 1998b). In this image, areas where multiple cones intersect indicate a higher probability of particle creation from that region. This approach enables the creation of a spatial distribution map of the particle production, which can be used to infer information about the proton beam path and dose deposition.

While an image of the particles' production distribution can be generated through SBP, this method is significantly influenced by detector resolution and various sources of 'noise.' Noise sources include misidentified particle interactions (e.g. mistaking an elastic collision of a neutron with a carbon nucleus as an elastic collision with a proton), scattering in the surroundings, and background neutron/gamma detection unrelated to beam interactions. These factors can lead to blurring and artifacts in the reconstructed images, potentially compromising the accuracy of range verification. Therefore, more sophisticated reconstruction and post-processing techniques are necessary to overcome these limitations and fully exploit the range correlated information provided by FNs and PGs.

Maximum likelihood expectation maximization (MLEM) (Shepp and Vardi 1982) is one of the standard image reconstruction algorithms used in astronomy and medical imaging, including proton therapy range verification. To avoid information loss due to data binning, and when dealing with few observed events relative to the number of available data bins, the list-mode version of MLEM (LM-MLEM) is applied (Barret *et al* 1997, Wilderman *et al* 1998a). A key challenge in MLEM is its tendency to amplify high-frequency noise. While this characteristic enhances signal quality in low-noise scenarios, it becomes problematic when significant statistical noise is present (Nuyts and Matej 2014). In such cases, the algorithm indiscriminately amplifies both signal and noise, leading to reconstructions polluted by undesirable pixel-level fluctuations. To mitigate this issue, regularization—such as through the maximum *a posteriori* (MAP) EM algorithm (De Pierro 1995)—and post-processing techniques are applied. MAP modifies the underlying objective function by imposing constraints or prior knowledge on the solution to stabilize the reconstruction process while the latter modifies the image after reconstruction to enhance its quality.

Various methods have been proposed to mitigate noise amplification in MLEM reconstruction. Kim *et al* introduced the concept of point spread function (PSF) deconvolution for resolution recovery in CC imaging (2013). Feng *et al* improved CC image quality using smoothness *a priori* information expressed through total variation (TV) regularization and PSF deconvolution in LM-MLEM reconstruction (2018). TV regularization enhanced images for ideal data, while combining TV with PSF deconvolution proved more effective for realistic data, where detector resolution was accounted for. Kohlhase *et al* demonstrated recovery

of spatial information in CC images of PGs by Gaussian-kernel smoothing post-reconstruction (2020b). However, concerns were raised about the impact of smoothing kernels on the accuracy of the estimated image. Shortly after, they explored various methods to incorporate *a priori* information from a particle-tracking device to enhance image quality of a CC designed for proton therapy range verification (Kohlhase *et al* 2020a). They introduced a novel approach for regularized maximum likelihood reconstruction by incorporating observed beam characteristics in a weighted system matrix and restricted volume approach to LM-MLEM. Their results indicated that while the penalized reconstruction significantly improved image quality, it also introduced higher statistical noise in the background. This is the only study, to our knowledge, which introduces proton therapy beam specific information to regularize MLEM reconstruction.

Deep learning has emerged as a powerful tool for improving image reconstruction. Supervised pix2pix (Isola *et al* 2017), a generative adversarial network (GAN), is one of the models that dominate the field of medical image processing tasks, such as denoising, deblurring, modality translation, and reconstruction. A non-exhaustive list exemplifying the architecture's versatility in medical image processing tasks follows. Platscher *et al* applied pix2pix to synthesize magnetic resonance images (MRI) of brain volumes with and without ischemic stroke lesions as a learning-based data augmentation approach to improve lesion segmentation (2022), while Nakanishi *et al* used it to predict computed tomography (CT) images from PET scans in small animal systems (2021). Sun *et al* employed pix2pix for low-dose myocardial perfusion single-photon emission CT (SPECT) denoising (2022), and Wang *et al* utilized the model to generate synthetic CT images from MRI for intracranial tumor radiotherapy planning (2022). Fard *et al* synthesized SPECT from MRI and PET (2024), and Pandey *et al* used it to predict optimized dose plans and post-treatment multiparametric MRI for brain metastases (2024).

The potential of deep learning in enhancing image reconstruction has also been demonstrated in context of proton beam range verification. Jiang *et al* explored the use of deep learning for enhanced PG imaging, focusing on two-dimensional (2D) reconstructions (2023). Xiao *et al* developed a long short-term memory network for predicting three-dimensional (3D) PG emission profiles, incorporating relative stopping power and dose information as inputs (2024). Hu *et al* proposed a machine learning framework for 3D dose predictions using positron emitter activity PET-images, integrating anatomical information from CT images and analytically derived stopping power as supplementary input (2020).

Given the success of deep learning in medical image processing, we explore its application for predicting 2D FN production distributions from low-quality SBP images and demonstrate the potential of using the predicted FN distributions as a means for treatment-specific regularization of LM-MAP-EM. We train a pix2pix model to predict the underlying FN production distribution $\boldsymbol{\lambda}$ given a low-quality SBP-reconstructed image $\boldsymbol{a}$. The universal approximation theorem suggests that, with sufficient layers and nodes, neural networks can approximate any function, making them suitable for this complex mapping task (Scarselli and Chung Tsoi 1998). We formulate the mapping as $F(\boldsymbol{a}) = \boldsymbol{\lambda}$ where F is the mapping function to be approximated, and use the predicted $\boldsymbol{\lambda}$ to form a prior in LM-MAP-EM reconstruction.

This strategy aims to improve reconstruction quality of the imaged FNs, ultimately enabling more precise range verification with the NOVCoDA system. Building on previous work (Setterdahl *et al* 2024) in which a U-Net architecture was explored, we extend the application to a heterogeneous CT-based patient model, incorporate detector resolution effects, and explore the use of the pix2pix GAN architecture in this context. Pix2pix was chosen for its well-documented implementation and its demonstrated performance on image generation tasks, as discussed earlier in the introduction. The data used in this study was generated through Monte Carlo (MC) simulations of a proton pencil beam aimed at a patient model with a cancer lesion in the left lung. Range shift scenarios emulating weight gain/loss of the patient were artificially induced in the simulations. We demonstrate our method with FN data and train a pix2pix architecture on datasets of paired ground truth (GT) (target) and SBP (input) images of the simulated 2D FN production distribution. The pix2pix architecture was trained separately on three such datasets, each with progressively increased noise levels in the SBP images, introduced through accounting for detector resolution effects in the underlying FN data. We further apply our method by incorporating the generated image of a trained pix2pix model as a mean estimate of the prior in LM-MAP-EM.

2. Methods

2.1. Simulations

The FN scatter data used in this study was originally generated by the NOVO consortium for a previous investigation (Meric *et al* 2023). The simulations were based on a modified patient dataset from the Lung CT Segmentation Challenge, available through the Cancer Imaging Archive (Clark *et al* 2013). A single patient scan was used, with a proton pencil beam targeted at a central spot in a left lung tumor 3 cm in diameter.

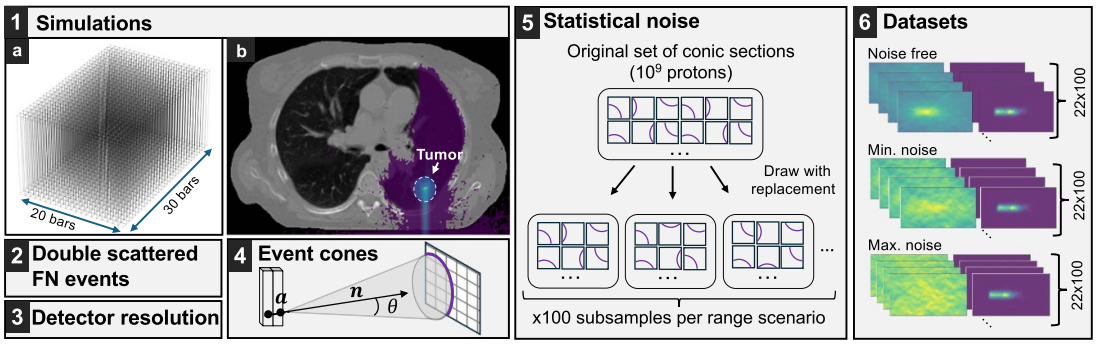

Figure 1. An illustration of the data generation workflow. (1) Monte Carlo simulations of fast neutron (FN) collisions with hydrogen nucleus within a plastic detector volume. (1a) The detector volume measures $20 \times 30 \times 20$ cm^3 and is divided into 20×30 subvolumes measuring $1 \times 1 \times 20$ cm^3. FNs are generated by interactions of an 85-MeV proton beam, with 10^9 protons per beam spot, traversing a patient model. The modeled patient is based on a computed tomography (CT) scan and has a tumor volume located in the left lung. (1b) Shows the dose deposition from a single proton beam entering the patient model, superimposed on the CT image. (2) 'Double-scattered FN' events (neutrons scattering twice in separate subvolumes) are identified and stored. (3) Detector resolution effects are incorporated into the double-scattered FN event data. (4) FN event cones, characterized by their vertex $\boldsymbol{a}$, axis $\boldsymbol{n}$, and half-opening angle θ, are constructed and projected onto an 400×400 mm^2 image plane. (5) To expand the dataset and introduce statistical noise, subsets of cone projections (conic sections) are sampled with replacement from the original dataset. The number of samples in a subset correspond to a beam intensity of 10^8 protons per beam sport and 100 subsets are created for each range scenario. (6) Two-dimensional (2D) images of the FN production distribution are reconstructed by simple back-projections (SBP), along with the corresponding 2D ground truth (GT) production distribution. A SBP and GT image make up an image pair. The final three datasets, accord with three levels of detector resolutions introduced in (3), all datasets consist of 22×100 image pairs, representing the 100 subsets sampled across 22 range scenarios.

Simulations were performed using Geant4 10.5 (Agostinelli *et al* 2003, Allison *et al* 2006, 2016)/GATE v. 8.2 (Sarrut *et al* 2014), configured with the QGSP_BIC_EMY physics list. The proton beam was simulated with 85 MeV energy, 4.7 mm full width at half maximum (FWHM), 2.5 mrad divergence, and 10^9 primary protons, and was directed posterior-anteriorly towards the tumor center. The number of primary protons reflects a realistic beam intensity, typically ranging from 10^6 to 10^{10} proton per beam spot (Pausch *et al* 2020).

Twenty-two range scenarios were created using the Medical Interactive Creative Environment (MICE) Toolkit v 1.1.3 (Nonpi Medical, Umeå, Sweden). The patient body was threshold-masked, then eroded or dilated in ± 1 mm steps up to ± 5 mm. Eroded areas were replaced with air, while dilated areas were filled with water. Additionally, due to voxel size limits, the proton beam energy was reduced by 405 keV to achieve intermediate 0.5 mm steps. Figure 1(b) illustrates a CT slice of the patient model with the dose deposition from a single proton beam superimposed on it.

The generated FN emission distributions were stored in phase-space files, which were then used as sources for subsequent detector simulations. A $20 \times 30 \times 20$ cm^3 volume, composed of 93% carbon and 7% hydrogen by mass, was used to represent the NOVCoDA detector volume. In detector simulations, energy deposition, time, position of interaction were stored for source FNs escaping the patient model and scattering within the detector volume. FNs scattering within the patient model before escaping were also considered. In post-processing of the output files, a 100 keV energy deposition threshold was applied and the detector volume was segmented into 20×30 subvolumes measuring $10 \times 10 \times 200$ mm^3 to emulate scintillator bars in a detector array, shown in figure 1(a). FNs scattering twice on protons, in unique subvolumes and passing the energy threshold, were kept while the remaining events were discarded. The number of registered double-scattered FN events was on order of 10^5 for a beam intensity of 10^9 protons per spot.

2.2. Datasets

Pix2pix models were trained on datasets of paired SBP and GT images of 2D FN production distributions. The dataset generation workflow is illustrated in figure 1. The generated datasets were based on the MC simulations described in section 2.1. Each step of the workflow following simulations and assembly of double-scatted FN events (steps 1 and 2) is described in detail in the following sections.

2.2.1. Detector resolution

Detector resolution effects were incorporated into the filtered FN scatter data by the method described in Setterdahl *et al* (2024). The detector resolution is determined by the combined effects of energy, timing, and spatial resolution of the detector element. The spatial resolution of a detector element is constrained by the transverse dimension of the bar, which in our case is 1×1 cm^2, and by the depth-of-interaction (DOI) resolution, which defines the precision of the scattering position along the longitudinal axis. Double

scattered FN events were assembled before introducing detector resolution effects, and event cones were constructed afterward. Three noise scenarios were considered: a noise-free, a minimum noise, and a maximum noise scenario. This comprehensive approach provides a more accurate representation of actual detector system limitations and uncertainties.

The 'noise-free' dataset represents an idealized scenario in FN scatter detection, containing MC truth values of FN scatters. This dataset provides a baseline for comparison, free from the uncertainties and limitations inherent in real detector systems. It includes precise information on scatter coordinates, energy deposits, and interaction times, as simulated by the MC method.

A 'minimum noise' dataset was generated by considering only the unavoidable spatial uncertainty inherent in the detector design. While original interaction coordinates along the longitudinal axis of the scintillator bar were kept, the coordinates in the transversal directions were adjusted to the center of the bars, reflecting the discrete nature of the detector elements. However, other resolution effects, including DOI (i.e. the longitudinal position), energy, and time resolutions, were disregarded. This approach isolates the fundamental spatial discretization of the detector while eliminating other sources of uncertainty that depend on the choice of scintillator material and light readout system.

The 'maximum noise' dataset simulates realistic detector performance by incorporating all relevant resolution effects into the originally noise-free data. Coordinates in the transversal directions were adjusted as in the 'minimum noise' dataset. The longitudinal-coordinate was resampled from a Gaussian distribution with a 10 mm FWHM to account for DOI resolution. Energy deposits and interaction times were resampled using Gaussian distributions with FWHMs corresponding to 10% energy resolution and 500 ps timing resolution, respectively. The chosen resolutions correspond to typical values of experimentally estimated resolutions for organic plastic scintillators (Sweany *et al* 2019, Turko *et al* 2024).

2.2.2. Event cone reconstruction

For a FN scattering twice by elastic collision with a proton within the detector volume, in separate subvolumes, the incidence angle θ of the FN in the laboratory frame can be estimated, applying conservation of momentum and energy, as:

$$\sin^2\theta = \frac{\Delta E}{E_n' + \Delta E}, \tag{1}$$

where ΔE is the energy deposited by the FN in the first scatter and E_n' is the scattered neutron energy, estimated by the time-of-flight (TOF) and distance d between the first and second scatter locations:

$$E_n' = \frac{1}{2} m_n \left(\frac{d}{\mathrm{TOF}}\right)^2, \tag{2}$$

where $m_n = 1.674\,9286 \cdot 10^{-27}$ kg is the neutron rest mass. With the azimuth angle ϕ unknown, and $\phi \in [0, 2\pi]$, possible locations of production of the double scattered FN is restricted to a conical surface, referred to as an event cone. An event cone can be described by its vertex, defined by the position coordinates of the first FN scatter; axis, defined by the vector connecting the second and first point of FN interaction, and half-opening angle θ, which are illustrated in figure 1(4).

2.2.3. Event cone projection

Given a cone with vertex coordinates (a_x, a_y, a_z), whose axis is described by vector components (n_x, n_y, n_z), and half opening-angle θ, and a plane positioned in x_0 and parallel to the yz-plane, the conic section defining the curve created by the intersecting cone and plane can be described by

$$n_x(x_0 - a_x) + n_y(y - a_y) + n_z(z - a_z) = \lambda^2 \left[(x - a_x)^2 + (y - a_y)^2 + (z - a_z)^2\right], \tag{3}$$

where $\lambda = \cos\theta$. Event cones were projected onto a 400×400 mm^2 pixilated image plane, with surface normal perpendicular to the proton beam axis and facing the detector volume. The pixel dimensions were 1×1 mm^2. For each range scenario, conic sections of FN event cones were computed. Pixels intersected by a conic section were assigned a value of 1 unit, and 0 otherwise.

2.2.4. Statistical noise

A SBP image is formed by projecting all event cones onto a common image plane, providing an estimate of the FN production distribution. Each range scenario is associated with a single set of conic sections (cone projections). As a result, the initial datasets described in section 2.2.1 yield only 22 SBP-GT image pairs per dataset. To generate a larger dataset for training, validation, and testing of the pix2pix networks, and to

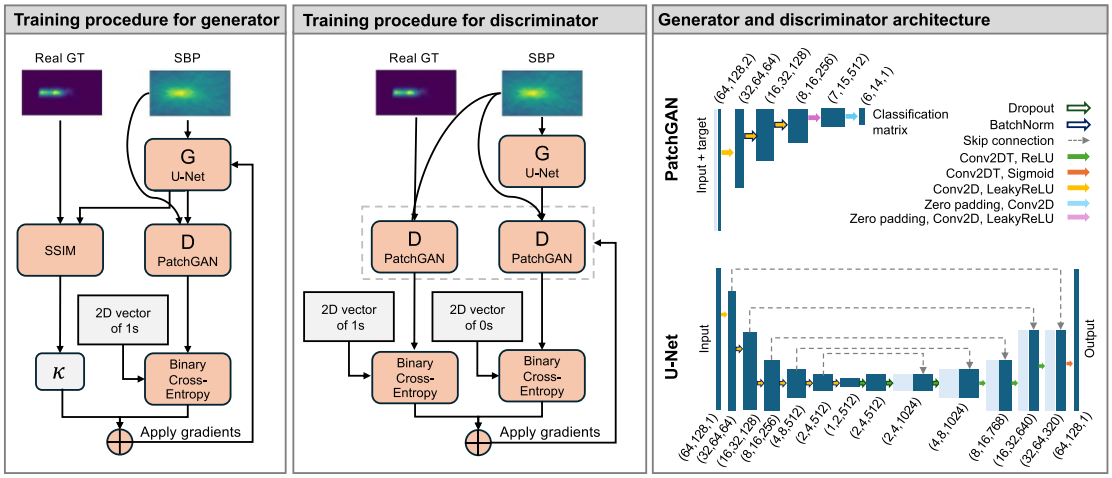

Figure 2. The training procedure and architecture of the pix2pix generator (G) and discriminator (D). Dimensions of network layers are (height, width, channel). In the U-Net, skip connections are a concatenation along the channel dimension.

introduce statistical noise corresponding to lower beam intensity, a conic section subsampling technique was employed, illustrated in figure 1(5). For each range scenario, 100 subsets of conic sections were generated; conic sections were randomly sampled with replacement (i.e. bootstrapped (Efron 1979)) from the original set, which represents a beam intensity of 10^9 protons per spot. For each sampled conic section, the GT production origin of the associated double-scattered FN event was stored and later used to generate GT images. Each subset contained on the order of 10^4 conic sections, reflecting a reduced beam intensity of 10^8 protons per spot. A SBP-GT image pair was then generated for each subset, yielding a total of 2200 SBP-GT image pairs (100 subsets $\times$ 22 range scenarios) for each of the three datasets described in section 2.2.1, with 100 distinct realizations of statistical noise for each range scenario.

2.3. Reconstruction with pix2pix

GANs, introduced by Goodfellow *et al* (2020), are a class of deep learning frameworks where two models are trained simultaneously, a generator and a discriminator. The generator learns to produce data that mimics the training set, while the discriminator learns how to distinguish between 'real' and 'fake' (generated) data, driving the generator to improve its outputs over time. Pix2pix, a conditional GAN (cGAN), extends this concept by conditioning both the generator and discriminator on additional input data (e.g. data labels) to enable more targeted predictions.

The applied pix2pix architecture comprises two main components, a U-Net-based generator and a PatchGAN discriminator, and is illustrated in figure 2. The U-Net generator employs an encoder–decoder structure with skip connections that directly link corresponding encoding and decoding layers. These skip connections enable better learning of spatial information. For the U-Net generator implemented in this work, the encoder consists of 6 consecutive down-sampling blocks, each comprising a 2D convolution (Conv2D), batch normalization, and LeakyReLU layer. All convolution layers use a kernel size of 4×4 and a stride of 2. The decoder consists of equally many up-sampling blocks, each containing a 2D transposed convolution (Conv2DT), batch normalization (BatchNorm), and ReLU layer. Dropout layers (with 50% dropout rate) are added to the first three up-sampling blocks after the BatchNorm layer. Following up-sampling is a final 2D transpose convolution with a single output channel, a kernel with size 4×4 and stride 2 and a sigmoid activation function to ensure output values between 0 and 1. The PatchGAN discriminator (introduced in Isola *et al* 2017) evaluates patches of an input image rather than the entire image, classifying each patch as 'real' or 'fake'. The discriminator of the presented work consists of three down-sampling blocks as described above, followed by 2D zero padding, 2D convolution with kernel size 4×4 and stride 1, batch normalization, LeakyReLU, 2D zero padding, and finally a last 2D convolution layer with same kernel parameters as the previous. With a receptive field size of 70×70 pixels, the PatchGAN outputs a $K \times L$ classification matrix where each entry represents the probability of the corresponding patch being real.

The global objective function for cGANs is:

$$\mathcal{L}_{\text{cGAN}}(G,D) = \mathbb{E}_{c,x}[\log D(c,x)] + \mathbb{E}_{c,z}[\log(1 - D(c,G(c,z)))], \tag{4}$$

where we let c, x, and z represent the SBP, GT, and random noise input, respectively; the generator and discriminator are denoted as G and D; where $\mathbb{E}$ represents the expected value; and, the generated, 'fake'

image is represented by $G(c,z)$. As in the original pix2pix implementation, the random noise is implemented by activating dropout in the generator layers at both training and test time, and an additional loss function that penalized differences in generated and GT images is added to the global objective function.

L1 and L2, which compute pixel-wise absolute or squared differences, are common choices of loss functions due to their computational efficiency, ease of differentiation, and widespread availability as built-in functions in most deep learning libraries, including TensorFlow (Abadi *et al* 2015). These loss functions do not explicitly favor either high- or low-frequency components, but because low-frequency content typically dominates images, they tend to produce smoothed results that capture broad transitions while neglecting fine details and textures. An alternative widely used in image generation is the structural similarity index measure (SSIM) (Wang *et al* 2004), which is considered a state-of-the-art metric for quantifying image quality. SSIM is valued for capturing perceptual similarity more effectively than traditional pixel-wise losses such as L1 and L2. SSIM compares images based on luminance, contrast, and structure using a sliding window approach and can be expressed as

$$\mathrm{SSIM}\left(x,y\right)=\frac{\left(2\mu_x\mu_y+C_1\right)\left(2\sigma_{xy}+C_2\right)}{\left(\mu_x^2+\mu_y^2+C_1\right)\left(\sigma_x^2+\sigma_y^2+C_2\right)}, \tag{5}$$

where C_1 and C_2 are constants, x and y represent the images being compared, μ_x and μ_y refer to the mean intensity, σ_{xy} refers to the covariance, and σ_x and σ_y represent the standard deviation of the pixel values within a considered window. In TensorFlow's implementation of SSIM, default parameters include a sliding window that is Gaussian weighted (with a standard deviation of 1.5 pixels) and has a size of 11×11 pixels, and the constants are $C_1=0.01$ and $C_2=0.03$, as in the original SSIM publication (Wang *et al* 2004). SSIM ranges from 0 to 1 where 1 indicates that x and y are identical.

In our work, we have chosen an SSIM-based loss function to penalize differences in generated and label images:

$$\mathcal{L}_{\mathrm{SSIM}}\left(G\right)=1-\mathrm{SSIM}\left(G\left(c,z\right),x\right), \tag{6}$$

where $\mathcal{L}_{\mathrm{SSIM}}(G)$ approaches 0 as a generated image $G(c,z)$ becomes increasingly similar to its label image x.

The global objective function of our model is:

$$G^*=\arg\min_G\max_D\left(\mathcal{L}_{\mathrm{cGAN}}\left(G,D\right)+\kappa\mathcal{L}_{\mathrm{SSIM}}\left(G\right)\right), \tag{7}$$

where κ is set to 100 (as in the original pix2pix publication) and balances the adversarial and structural loss terms. The generator and discriminator both have their own loss functions that they aim to minimize during training, respectively,

$$\mathcal{L}_G=BCE(1,D(c,G(c))+\kappa\mathcal{L}_{\mathrm{SSIM}}(G) \tag{8}$$

and

$$\mathcal{L}_D=BCE(0,D(c,G(c))+BCE(1,D(c,x)), \tag{9}$$

where BCE denotes binary cross entropy, and the 1 and 0 correspond to a 2D vector of the same shape as the PatchGAN output, consisting entirely of ones and zeros.

2.3.1. Dataset preparation

The original images, initially 400×400 pixels, were cropped to 128×64 pixels. This cropping was necessary to balance the ratio of signal-containing pixels to non-signal pixels in the GT images. Without cropping, the PatchGAN discriminator, which uses a 70×70 receptive field, would have produced a classification matrix with a disproportionately high number of entries corresponding to patches with no signal, as the receptive field covers only about 3% of the original 400×400 image area and the signal region is roughly 128×30 pixels large. To reduce high-frequency pixel noise, we applied Gaussian blurring to the images. This was achieved using a 2D Gaussian kernel with a standard deviation of 1 and a kernel size of 20×20 pixels. Following the blurring process, we normalized the images based on their maximum pixel value.

Datasets were split into training-validation and test subsets. Range shift scenarios $\{-5.5,-5.0,-3.5,-2.5,-1.0,-0.5,0.5,1.0,2.5,3.5,5.0\}$ mm and $\{-4.5,-4.0,-3.0,-2.0,-1.5,0.0,1.5,2.0,3.0,4.0,4.5\}$ mm were allocated for the training-validation and test subsets, respectively. Test scenarios were specifically chosen as intermediate range shifts between training values (e.g. -4.5 mm falls between the training shifts -5.0 mm and -3.5 mm) to evaluate interpolation capability across unseen gaps in the training distribution. The training-validation subset was randomly split into training and validation with a 70:30 ratio. The training, validation and test subset contained 770, 330, and 1100 image pairs respectively.

2.3.2. Training

The models were trained over a course of 10 epochs, utilizing an early stopping mechanism to prevent overfitting. An epoch represents a complete pass through the entire training dataset. We employed 'instance normalization' by using a batch size of 1, meaning the model updated its weights after processing each individual image. The training process performs N steps per epoch, where each step corresponded to a weight update of the model and $N =$ dataset size/batch size. For both the generator and discriminator components of the network, we used the Adam optimizer (Kingma and Ba 2017) with a learning rate of $2 \cdot 10^{-4}$ and an exponential decay rate of the first moment $\beta_1 = 0.5$. These values were set empirically as they provided stable results but were not systematically optimized. The remaining parameters of the optimizer were set to default values of the TensorFlow implementation.

2.3.3. Implementation

The models were built with TensorFlow 2.13.0 (Abadi *et al* 2015) and trained on a MacBook Pro equipped with an Apple M3 Pro chip featuring an 14-core GPU, enabled with the `tensorflow-metal` plug-in.

2.3.4. Evaluation

The models were evaluated using a combination of image quality metrics and range-landmark (RL) correlation analysis.

2.3.4.1. Image quality

Image quality was assessed using mean squared error (MSE) and SSIM over the test dataset. While MSE provides a pixel-level comparison of intensities, SSIM evaluates structural, luminosity, and contrast aspects of the images. An MSE $= 0$ and SSIM $= 1$ means compared images are identical.

2.3.4.2. RL correlation

Of particular importance was the evaluation of how well the network captured the range features from the input SBP images and accurately reflected these in the generated 'fake' GT images.

The longitudinal profile of a 2D FN production distribution is a one-dimensional (1D) representation obtained by summing pixel values along the direction perpendicular to the beam axis, capturing how the production distribution varies along the beam path. A RL metric was defined as the weighted average of a longitudinal profile, used in previous work (Meric *et al* 2023, Setterdahl *et al* 2024):

$$\mathrm{RL} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}, \tag{10}$$

where w_i are the weights and x_i are the pixel positions along the longitudinal profile. RLs were plotted against true range shift magnitude, and a linear regression model was then fitted to these data points. The coefficient of determination (R^2) was used to assess the strength of the correlation. An R^2 value of 1 indicates perfect linear correlation, while values below 0.5 are considered to represent weak correlation. This approach allowed us to gauge the model's ability to preserve crucial range information, which is essential if the model is to be useful for range verification. Note that the RL metric is not intended to determine the absolute position of the Bragg peak but is instead used to detect changes in the FN distribution that are associated with shifts in the proton beam range.

2.4. MAP EM with predicted prior

The MAP estimate of the emission distribution $\boldsymbol{\lambda}$ from the measured data vector $\boldsymbol{y}$ is given by

$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda} \geqslant 0} \frac{p(\boldsymbol{y}|\boldsymbol{\lambda})\, p(\boldsymbol{\lambda})}{p(\boldsymbol{y})}, \tag{11}$$

where $p(\boldsymbol{y}|\boldsymbol{\lambda})$ is the likelihood of measuring $\boldsymbol{y}$ given $\boldsymbol{\lambda}$, and $p(\boldsymbol{\lambda})$ is the prior probability of $\boldsymbol{\lambda}$.

The prior probability of $\boldsymbol{\lambda}$ is assumed to be proportional to a Gibbs distribution (Geman and Geman 1984), with the form

$$p(\boldsymbol{\lambda}) \propto \exp\left(-\beta U(\boldsymbol{\lambda})\right), \tag{12}$$

where β is a positive constant, and $U(\boldsymbol{\lambda})$ is the so-called energy function that characterizes the prior distribution, e.g. by incorporating spatial relationships between pixels.

We consider three energy functions. The first is a quadratic function that enforces smoothness between a pixel j and pixels in its neighborhood, N_j, indexed by c:

$$U_1(\boldsymbol{\lambda}) = \frac{1}{2}\sum_{j}^{J}\sum_{c\in N_j} w_{jc}\left(\lambda_j - \lambda_c\right)^2, \tag{13}$$

where w_{jc} is a weight defined as the inverse distance of evaluated pixels separated by a distance d, $w_{jc} = 1/d$. A second-order neighborhood evaluates direct neighbors and diagonals, and a fourth-order neighborhood evaluates two layers of direct neighbors and diagonals, i.e. a 3×3 and 5×5 pixel region surrounding the evaluated pixel, respectively. For a second-order neighborhood, direct and diagonal neighbors are evaluated with distance weights $w_{jc} = 1$ and $w_{jc} = 1/\sqrt{2}$, respectively.

The other two energy functions evaluate the similarity of $\boldsymbol{\lambda}$ and a prediction of the prior mean value $\boldsymbol{p}$, generated by the pix2pix model trained on the minimum noise dataset. The second energy function is similar to equation (13),

$$U_2(\boldsymbol{\lambda}) = \frac{1}{2}\sum_{j}^{J}\sum_{c\in N_j} w_{jc}\left(\lambda_j - p_c\right)^2, \tag{14}$$

where we have replaced a neighbor pixel value λ_c of the reconstructed image with the corresponding value p_c in the predicted image. The last and third energy function directly compares values of pixel j in the reconstructed image to the value of the corresponding pixel in the predicted image (as proposed by Levitan and Herman 1987),

$$U_3(\boldsymbol{\lambda}) = \frac{1}{2}\sum_{j}^{J}\left(\lambda_j - p_j\right)^2. \tag{15}$$

Partial derivatives of the presented energy functions with respect to λ_j, which will be used in later equations, are:

$$\frac{\partial}{\partial\lambda_j}U_1(\boldsymbol{\lambda}) = \sum_{c\in N_j} w_{jc}\left(\lambda_j - \lambda_c\right) \tag{16}$$

$$\frac{\partial}{\partial\lambda_j}U_2(\boldsymbol{\lambda}) = \sum_{c\in N_j} w_{jc}\left(\lambda_j - p_c\right) \tag{17}$$

$$\frac{\partial}{\partial\lambda_j}U_3(\boldsymbol{\lambda}) = \left(\lambda_j - p_c\right). \tag{18}$$

Maximizing equation (11) is equivalent to finding the $\boldsymbol{\lambda}$ that maximizes the log-posterior probability:

$$\log p(\boldsymbol{y}|\boldsymbol{\lambda}) - \beta U(\boldsymbol{\lambda}) - \log p(\boldsymbol{y}), \tag{19}$$

where $\log p(\boldsymbol{y})$ is considered a constant, since it is independent of $\boldsymbol{\lambda}$. Applying the Poisson noise model to equation (19), using Green's one-step-late (OSL) procedure (Green 1990), and considering LM data (Barret *et al* 1997, Parra and Barrett 1998) we derive the iterative LM-MAP-EM update equation:

$$\lambda_j^{k+1} = \frac{s_j\lambda_j^*}{s_j + \beta\frac{\partial}{\partial\lambda_j}U(\boldsymbol{\lambda})\big|_{\lambda_j=\lambda_j^k}}, \tag{20}$$

where

$$\lambda_j^* = \frac{\lambda_j^k}{s_j}\sum_{i=1}^{I}\frac{a_{ij}}{\sum_{j'}a_{ij'}\lambda_{j'}^k}. \tag{21}$$

The formula, in equation (20), for the updated estimate λ_j^{k+1} incorporates s_j, the sensitivity of pixel j, and the system matrix element a_{ij} which represents the probability of detecting an emission from pixel j in the ith event. For simplicity, we assign equal sensitivity to all pixels, letting $s_j = 1$ for all j. Through equation (21), the formula includes the sum over all detected events, indexed by i up to the total number of events I, and compares each a_{ij} to a forward projection $\sum_{j'}a_{ij'}\lambda_{j'}^k$ of the current estimate λ_j^k, summed over all pixels using the dummy index j'. The system matrix elements are defined as 1 if the conic section of event i intersects pixel j during back-projection, and is 0 otherwise. With the OSL method, the partial derivative of the energy

Table 1. Summary of image reconstruction approaches for seven different combinations of LM-MAP-EM reconstruction algorithm (alg.), energy function, size of the neighborhood (neigh.) evaluated by the energy function, and the constant β governing the emphasis put on the energy function during reconstruction. Energy functions U_1, U_2, and U_3 are defined by equations (13), (14) and (15), respectively. The latter two utilize a mean estimate of the prior. Neighborhood sizes 3×3 and 5×5 correspond to a second- and fourth-order neighborhood. Approach (1), for which $\beta = 0$, corresponds to reconstruction by the LM-MLEM algorithm. Approach (2) is LM-MAP-EM with a smoothing quadratic energy function evaluated over a second-order neighborhood. Approach (3)–(7) is LM-MAP-EM reconstruction where energy functions compare the current update of the reconstructed image to a mean estimate of the prior.

ID	Alg.	β	Energy	Neigh. size
1	Equation (20)	0	/	/
2	Equation (20)	0.8	U_1	3×3
3	Equation (22)	0.9	U_3	/
4	Equation (20)	0.1	U_2	3×3
5	Equation (20)	0.5	U_2	3×3
6	Equation (20)	0.1	U_2	5×5
7	Equation (20)	0.5	U_2	5×5

function $\frac{\partial}{\partial \lambda_j} U(\boldsymbol{\lambda})$ considers the current estimate λ_j^k. When $\beta = 0$, equation (20) reduces to $\lambda_j^{k+1} = \lambda_j^*$ which is the LM-MLEM update (Barret *et al* 1997, Parra and Barrett 1998).

When using the energy functions that consider neighborhoods, i.e. equation (13) and equation (14), we apply the LM-MAP-EM algorithm shown in equation (20). For the energy function in equation (15), which does not consider neighborhoods, the MAP objective function becomes a quadratic equation (as in Levitan and Herman 1987) to which the solution with respect to λ_j^{k+1} is

$$\lambda_j^{k+1} = \frac{2\lambda_j^* s_j}{(s_j - \beta p_j) + \sqrt{(s_j - \beta p_j)^2 + 4\beta \lambda_j^* s_j}}, \tag{22}$$

where λ_j^* is as defined in equation (21). Once again, we see that the expression for the LM-MAP-EM update reduces to $\lambda_j^{k+1} = \lambda_j^*$ when $\beta = 0$.

We perform reconstruction with seven different combinations of LM-MAP-EM algorithm (equations (20) and (22)), energy functions (equations (13)–(15)), the size of the neighborhood considered by the energy function (second- and fourth-order), and the constant β, summarized in table 1.

All reconstructions were performed for a predetermined number of iterations, using early stopping at 20 iterations. For LM-MLEM reconstruction (ID 1 in table 1), higher iterations caused increased pixel noise, and little improvement was seen in LM-MAP-EM updates (ID 2-7 in table 1). We evaluated the structural similarity between the longitudinal profiles of reconstructed images and the GT by SSIM. SSIM is sensitive to differences in intensity and contrast. To emphasize the shape of the reconstructed profiles, we perform max-min normalization before assessing similarity.

3. Results

Two distinct peaks are consistently present in the GT FN production distributions, shown in figure 3. The first peak arises from increased FN production as protons enter the patient model and interact with muscle (soft tissue), while the second peak occurs due to protons traversing and interacting with the ribcage/bone. The distributions differ mainly in the location of the rising edge and the shape of the tail, reflecting range shifts induced by adding/removing material off the patient model to emulate weight gain/loss.

Pix2pix was trained and tested on three datasets with increasing data quality degradation due to detector resolution effects. Its performance was evaluated on unseen range scenarios. An example of the 2D SBP-GT image pair, predicted image and the difference map of GT and prediction for the unseen -3.0 mm range shift scenario and their corresponding longitudinal profiles are shown in figure 4. In all scenarios, pix2pix produced visually realistic and plausible estimates of the underlying 2D GT FN production distribution. The two distinct 'bunny ear' peaks were consistently captured in the predicted images. However, closer inspection revealed that the placement of the rising edge and the tail shape of the FN production were often displaced, particularly in models trained and tested on the noisy datasets. Misalignment of the rising edge was more frequent and pronounced in the maximum noise dataset, as demonstrated in figure 4 and reflected in the smaller mean SSIM ($= 0.96$) and larger mean MSE ($= 1 \cdot 10^{-3}$) values compared to those of the noise-free (MSE $= 2 \cdot 10^{-4}$, SSIM $= 0.986$) and minimum noise datasets (MSE $= 8 \cdot 10^{-4}$, SSIM $= 0.97$), shown in figure 5. Note that the SSIM values in all cases are relatively close to 1, which indicates that the compared images are identical, due to the significantly large portion of zero-valued pixels in both the GT and predicted images.

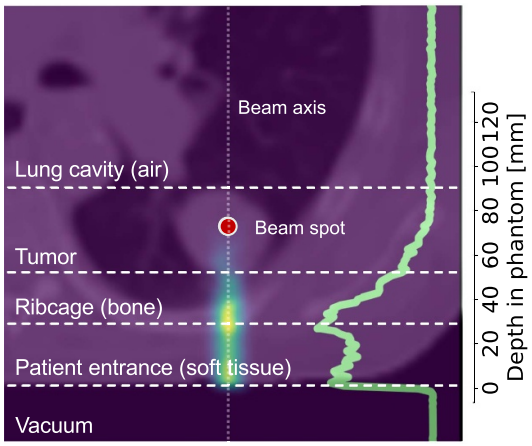

Figure 3. Simulated fast neutron (FN) production along the path of an 85 MeV proton pencil beam superimposed on the patient CT image with a 3 cm diameter tumor in the left lung. FN production peaks where the beam enters the patient and the section where the beam passes the ribcage and tapers off leading up to the beam's stopping position within the tumor.

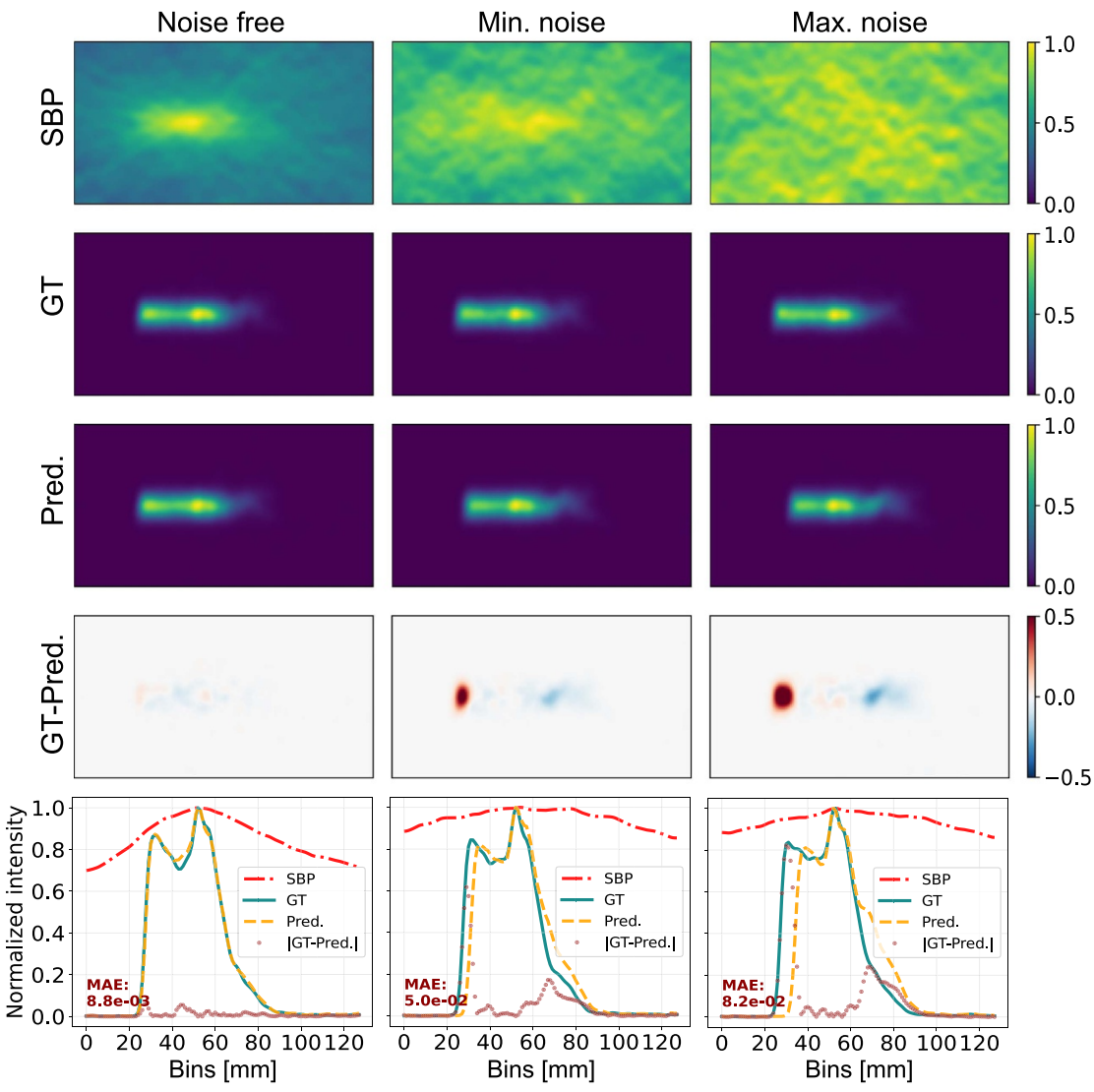

Figure 4. Longitudinal profiles corresponding to the example images of a simple back-projection (SBP)-ground truth (GT) image pair, the corresponding pix2pix predicted (Pred.) image and the mean absolute error (MAE) of the GT and Pred. profile.

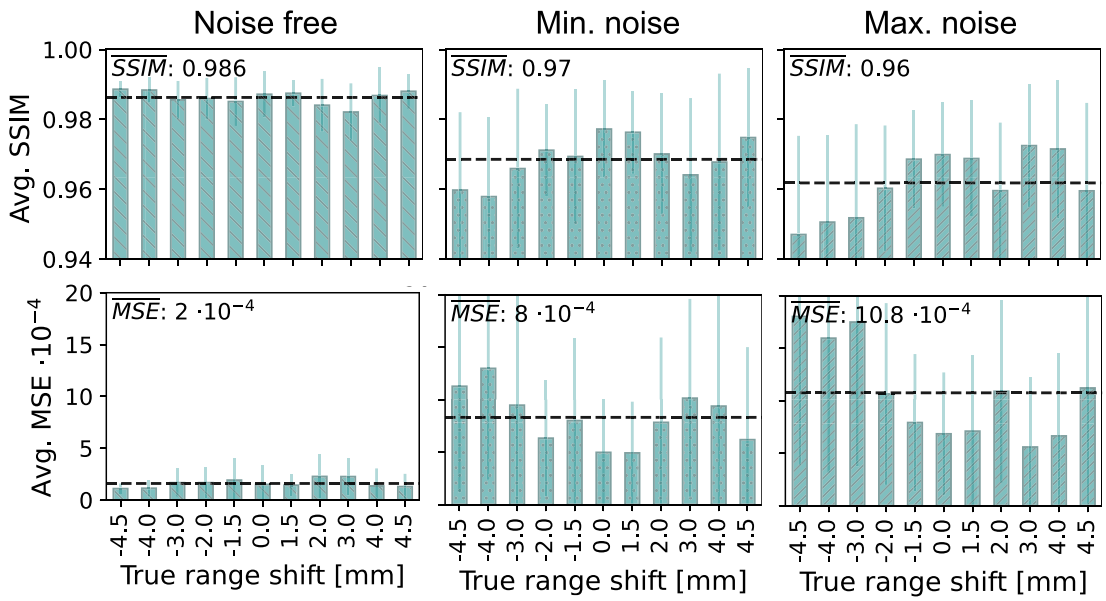

Figure 5. Mean squared error (MSE) and structural similarity index (SSIM) values for the noise-free, minimum noise, and maximum noise datasets. The metrics indicate a degradation in performance with increasing noise, with the maximum noise dataset exhibiting the highest MSE and lowest SSIM. Bias toward specific range shifts is observed, particularly in the minimum and maximum noise datasets.

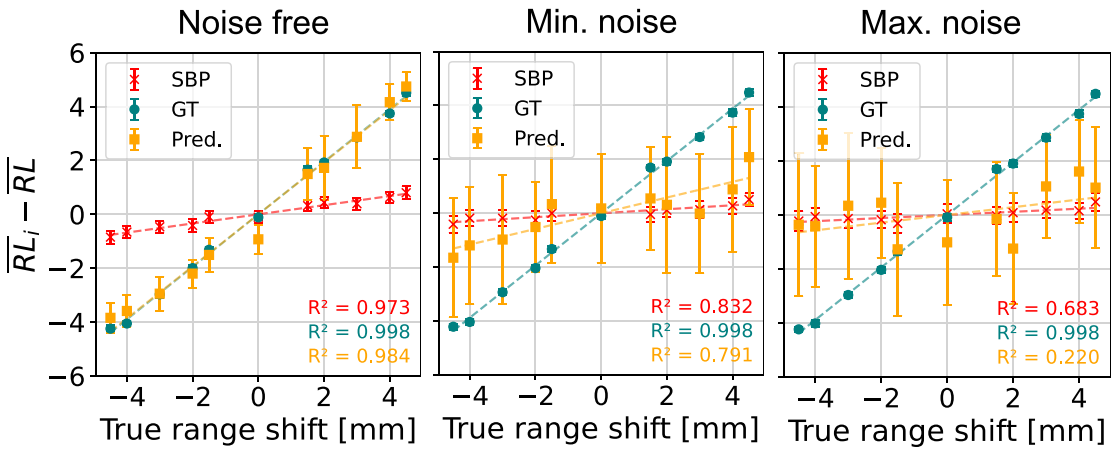

Figure 6. The figures illustrate the linear correlation of $\overline{RL_i} - \overline{RL}$, i.e. the difference in the average range landmark (RL) within a range scenario $\overline{RL_i}$ and the average RL over all range scenarios $\overline{RL}$, and true range shift magnitudes induced during simulations for each image type. Image types include simple back-projection (SBP), ground truth (GT) and pix2pix predicted (Pred.).

The evaluation of individual range shifts also indicated a slight positive bias toward specific range scenarios, where the predicted 2D FN distributions more closely resembled the GT. This is reflected in a lower average MSE and a higher average SSIM for these scenarios compared to the overall averages across all range shift scenarios. For the noise-free dataset, this bias was relatively small, with most range scenarios having mean MSE and SSIM values similar to the overall mean. For the minimum noise dataset, a positive bias was observed toward range shifts of 0.0 and 1.5 mm, while for the maximum noise dataset, positive biases appeared at −1.5, 0.0, 1.5, 3.0, and 4.0 mm.

Preservation of range information in the transformation from SBP to the predicted FN distribution is critical for proton beam range verification. This was assessed using the coefficient of determination (R^2) from a linear regression model fitted to $\overline{RL_i} - \overline{RL}$ as a function of the true range shift magnitudes induced in simulations, where $\overline{RL_i}$ represents the average RL within a range scenario and $\overline{RL}$ the average RL over all range scenarios, as shown in figure 6. The R^2 associated with the SBP images worsened with increased level of detector resolution effects. For the noise-free dataset, a strong correlation was observed, with $R^2 = 0.98$, for the data associated with pix2pix predicted images. As data quality decreased due to increasing detector resolution effects, the correlation weakened, with $R^2 = 0.79$ for the minimum noise dataset and $R^2 = 0.22$

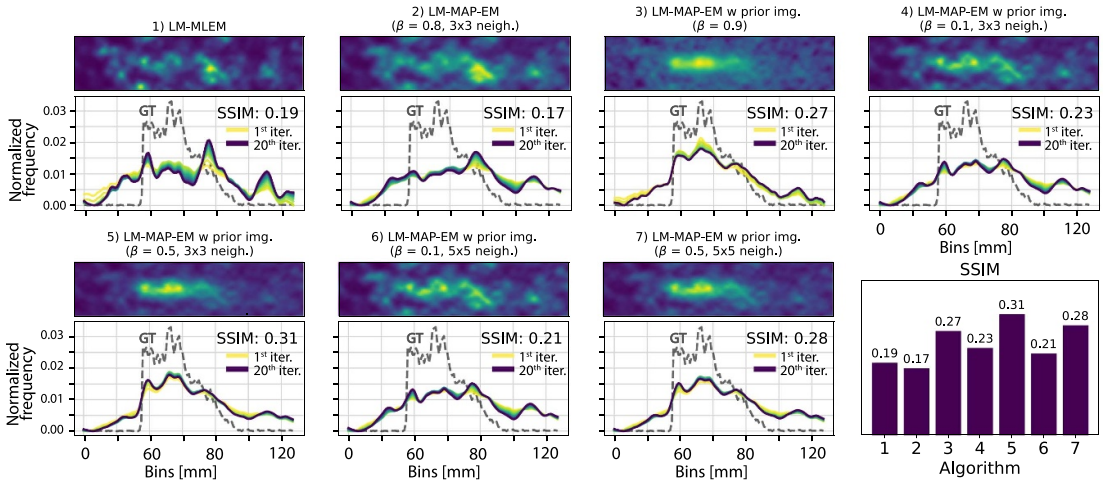

Figure 7. Reconstruction result after 20 iterations for LM-MLEM (1), LM-MAP-EM with a quadratic energy function (2), and LM-MAP-EM with a prior image to guide reconstruction (3–7). The latter group demonstrates higher SSIM values for the longitudinal profiles when compared to the ground truth (GT) than the conventional algorithms where no prior image is used. Longitudinal profiles are normalized so that minimum value is 0 and the sum of weights equals 1 in order to emphasize the distribution shape and compare to the GT profile. Profiles of the 1st though 20th iteration is shown along with the final 20th iteration image, cropped to a 128 × 30 pixel region encompassing peak FN signal. Pixel and bin sizes are 1×1 mm^2 and 1 mm, respectively.

for the maximum noise dataset. Additionally, the variation in computed RLs, and consequently $\overline{RL}_i - \overline{RL}$ values, increased, making it more challenging to distinguish between different range shift scenarios.

Seven different approaches to LM-MAP-EM reconstruction were performed, including a naïve LM-MLEM (non-regularized), conventional LM-MAP-EM with a quadratic energy function over a second-order neighborhood, and five variations of LM-MAP-EM in which a mean estimate of the prior were used to guide reconstruction (respectively labeled with ID 1,2 and 3–7 in table 1). For the latter approaches, the mean estimate of the prior and the reconstructed image were evaluated either over pixel neighborhoods or on a pixel-by-pixel basis, through the energy functions defined in equations (14) and (15). The approaches were demonstrated on an arbitrarily chosen instance of the −3.0 mm range shift scenario (from the test subset of the minimum noise dataset). Using the corresponding SBP image of the −3.0 mm range shift instance, a mean estimate of the prior was predicted by the pix2pix model that was trained on the minimum noise dataset.

The reconstructed images resulting from the different reconstruction approaches are shown in figure 7. As expected, the image reconstructed by naïve LM-MLEM (figure 7.1) exhibited noise-enhancing features, resulting in a speckled image reminiscent of a 'night sky'. Introducing regularization via a quadratic energy function resulted in a more defined FN distribution in the signal region (figure 7.2), with greater smoothing observed for higher β values and larger neighborhoods. Incorporating a mean estimate of the prior image into reconstruction further enhanced the FN production distribution structure. However, the pixel-by-pixel comparison introduced noticeable pixel noise in the background (figure 7.3), which was substantially reduced when using the energy function evaluating the mean estimate prior and reconstructed image over pixel neighborhoods (figures 7.4–7.7). These reconstructed images (corresponding to approach ID 3–7) demonstrated greater SSIM values ($\text{SSIM} \in [0.21, 0.31]$) with the underlying GT distribution than images resulting from naïve LM-MLEM ($\text{SSIM} = 0.19$) and conventional LM-MAP-EM with a quadratic energy function evaluated over a second-order neighborhood ($\text{SSIM} = 0.17$).

4. Discussion

Our study demonstrates that the proposed method performs well on noise-free data, but its effectiveness diminishes as detector resolution degrades data quality (i.e. for the minimum and maximum noise datasets). The minimum noise scenario considers the unavoidable spatial uncertainty inherent in the detector design. The maximum noise scenario models what we could realistically expect from experimental measurements by accounting for all the relevant detector resolutions, that is, the energy, timing and spatial resolution of the detector. For both the minimum and maximum noise datasets, the lower data quality makes it increasingly challenging for pix2pix to capture beam range-related features in the SBP image. In the maximum noise scenario, while the network generates a seemingly plausible GT image, there is significant variance in its alignment with the target GT label, and range correlation of predicted RLs is virtually non-existent. The

correlation appears stronger in the minimum noise case; however, the deviation in RLs remains substantial, making it difficult to distinguish between range scenarios due to nearly completely overlapping Gaussian distributed RLs. The pix2pix model performs arguably best on the noise-free case, exhibiting a strong correlation of RLs to range shift labels and relatively small variations in RLs of predicted images compared to the noisy datasets.

The beam intensity affects the counting statistics and the achievable accuracy. In the present work, the simulated beam intensity was 10^9 protons per spot, and the number of conic sections subsampled (section 2.2.4) corresponded to a beam intensity of 10^8 protons. Subsampling of conic sections corresponding to $5 \cdot 10^8$ protons did not improve network performance, indicating that detector resolution effects may be a greater challenge than statistical noise for training the pix2pix network.

A single case study from the minimum noise dataset demonstrates that a LM-MAP-EM approach, guided by a mean estimate of the prior over pixel neighborhoods, yields the highest visual agreement with corresponding simulated GTs. As expected, we observe that naïve, non-regularized, MLEM amplifies high-frequency components and statistical noise. While enforcing smoothness in reconstruction with a quadratic energy function (without a mean estimate of the prior image) creates more meaningful images, they remain severely affected by noise.

Our study has demonstrated the potential of using deep learning techniques for range verification in proton therapy. However, several areas warrant further investigation to enhance the robustness and clinical applicability of this approach. The choice of hyperparameters, particularly the constant β, plays a crucial role in determining the balance between the influence of the prior image and the data fidelity term corresponding to the Poisson likelihood. Further research is needed to optimize these parameters for optimal reconstruction. Moreover, accurate predictions and the alignment of the mean estimate of the prior with the reconstructed image is essential for accurate results, as misalignment could erroneously guide reconstruction.

In the generation of SBP images, a single cone projection results in a curve on the image plane that spans multiple pixels, which are often widely spaced. Furthermore, there is a one-to-many relationship between pixel correspondences in the SBP and GT images: a single event cone projection in the SBP image increases intensity in several pixels, whereas in the GT image, only one pixel corresponds to the event's true origin. Thus, the spatial correspondence between GT and reconstructed SBP images is not straightforward. Convolutional networks are designed to use small kernels, such as 3×3, to capture meaningful features from spatially adjacent pixels. However, adjacent pixels in the SBP image may not necessarily correspond to neighboring pixels in the GT image. Future work could explore more sophisticated network architectures. One potential candidate is AUTOMAP, a non-GAN convolutional neural network proposed by Zhu *et al* (2018). AUTOMAP incorporates both fully connected (FC) and convolutional layers, where the FC layers are designed to learn mathematical transformations that require spatially non-local manipulation. Moreover, we acknowledge that a comprehensive comparison with other architectures would be valuable, however, the goal of this work is to establish a proof-of-concept for applying deep learning to proton therapy range verification, not to perform an exhaustive search for the optimal network. Network optimization, including architecture design, loss functions, and hyperparameter tuning is left for future work.

A significant challenge lies in generating a diverse and representative training dataset. In present work, generating realistic FN data through MC simulations proved computationally intensive, both in terms of processing time and storage requirements. Due to these constraints, we limited this proof-of-concept study to a single patient anatomy and a single beam energy (85 MeV). This allowed us to focus on evaluating the feasibility of incorporating learned priors in the reconstruction process. We acknowledge this limitation and consider the extension to additional anatomical sites, beam energies, and range shift scenarios an important direction for future work. While data augmentation techniques can help expand the dataset by synthetically generating new data samples, it is crucial to ensure that the fundamental relationship between the SBP and GT images is preserved. Developing methods to synthetically generate realistic training data could be a valuable area of research for cone-projection based imaging systems, not only for NOVCoDA but also systems such as CC and neutron scatter cameras.

5. Conclusion

In conclusion, our study demonstrates the potential of utilizing deep learning techniques to enhance image quality in iterative reconstruction for range verification with FNs in proton therapy. However, we found that the effectiveness of the proposed method is highly dependent on data quality, with significant performance degradation in high-noise scenarios. A potential avenue for future work includes exploring alternative network architectures and developing a custom model tailored to specific challenges, particularly to improve robustness against noise.

Data availability statement

The data cannot be made publicly available upon publication because the cost of preparing, depositing and hosting the data would be prohibitive within the terms of this research project. The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

The NOVO project has received funding from the European Innovation Council (EIC) under Grant Agreement No. 101130979. The EIC receives support from the European Union's Horizon Europe research and innovation program. Partners from The University of Manchester has received funding from UK Research and Innovation under Grant Agreement No. 10102118. The authors acknowledge financial support by the Research Council of Norway under the NANO2021 program (Grant No. 301459).

ORCID iDs

Lena M Setterdahl https://orcid.org/0000-0001-7765-7308
Kyrre Skjerdal https://orcid.org/0000-0002-3164-3100
Hunter N Ratliff https://orcid.org/0000-0003-3761-5415
Kristian Smeland Ytre-Hauge https://orcid.org/0000-0002-3597-1611
William R B Lionheart https://orcid.org/0000-0003-0971-4678
Sean Holman https://orcid.org/0000-0001-8050-2585
Helge E S Pettersen https://orcid.org/0000-0003-4879-771X
Francesco Blangiardi https://orcid.org/0009-0006-7725-6751
Danny Lathouwers https://orcid.org/0000-0003-3810-1926
Ilker Meric https://orcid.org/0000-0001-5659-3907

References

Abadi M *et al* 2015 TensorFlow: large-scale machine learning on heterogeneous systems (v2.13.0) (available at: www.tensorflow.org/)
Agostinelli S *et al* 2003 GEANT4—a simulation toolkit *Nucl. Instrum. Methods Phys. Res. 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 Phys. Res. A* **835** 186–225
Barret H H, White T and Parra L C 1997 List-mode likelihood *J. Opt. Soc. Am. A* **14** 2914–23
Bom V, Joulaeizadeh L and Beekman F 2012 Real-time prompt gamma monitoring in spot-scanning proton therapy using imaging through a knife-edge-shaped slit *Phys. Med. Biol.* **57** 297–308
Clark K *et al* 2013 The cancer imaging archive (TCIA): maintaining and operating a public information repository *J. Digit. Imaging* **26** 1045–57
De Pierro A R 1995 A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography *IEEE Trans. Med. Imaging* **14** 132–7
Draeger E, Mackin D, Peterson S, Chen H, Avery S, Beddar S and Polf J C 2018 3D prompt gamma imaging for proton beam range verification *Phys. Med. Biol.* **63** 035019
Efron B 1979 Bootstrap methods: another look at the jackknife *Ann. Stat.* **7** 1–26
Enghardt W, Crespo P, Fiedler F, Hinz R, Parodi K, Pawelke J and Pönisch F 2004 Charged hadron tumour therapy monitoring by means of PET *Nucl. Instrum. Methods Phys. Res. A* **525** 284–8
Fard A S, Reutens D C, Ramsay S C, Goodman S J, Ghosh S and Vegh V 2024 Image synthesis of interictal SPECT from MRI and PET using machine learning *Front. Neurol.* **15** 1383773
Feng Y, Etxebeste A, Létang J, Sarrut D and Maxim V 2018 Total variation and point spread function priors for MLEM reconstruction in Compton camera imaging *2018 IEEE Nuclear Science Symp. and Medical Imaging Conf. Proc. (NSS/MIC) (Sydney, NSW, Australia)* pp 1–3
Geman S and Geman D 1984 Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-6** 721–41
Golnik C *et al* 2014 Range assessment in particle therapy based on prompt gamma-ray timing measurements *Phys. Med. Biol.* **59** 5399–422
Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y 2020 Generative adversarial networks *Commun. ACM* **63** 139–44
Green P 1990 Bayesian reconstructions from emission tomography data using a modified EM algorithm *IEEE Trans. Med. Imaging* **9** 84–93
Hu Z, Li G, Zhang X, Ye K, Lu J and Peng H 2020 A machine learning framework with anatomical prior for online dose verification using positron emitters and PET in proton therapy *Phys. Med. Biol.* **65** 185003
Hueso-González F, Rabe M, Ruggieri T A, Bortfeld T and Verburg J M 2018 A full-scale clinical prototype for proton range verification using prompt gamma-ray spectroscopy *Phys. Med. Biol.* **63** 185019
Isola P, Zhu J Y, Zhou T and Efros A A 2017 Image-to-image translation with conditional adversarial networks *2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* pp 5967–76
Jiang Z, Polf J, Barajas C, Gobbert M and Ren L 2023 A feasibility study of enhanced prompt gamma imaging for range verification in proton therapy using deep learning *Phys. Med. Biol.* **68** 075001

Kim S, Seo H, Park J, Kim C, Lee C, Lee S, Lee D and Lee J 2013 Resolution recovery reconstruction for a Compton camera *Phys. Med. Biol.* **58** 2823–40
Kingma D P and Ba J 2017 Adam: a method for stochastic optimization (arXiv:1412.6980)
Knopf A-C and Lomax A 2013 *In vivo* proton range verification: a review *Phys. Med. Biol.* **58** R131
Kohlhase N, Stille M, Bolke A, Zvolský M and Rafecas M 2020a Compton camera image reconstruction with a-priori information from a beam tagging hodoscope *2020 IEEE Nuclear Science Symp. and Medical Imaging Conf. (NSS/MIC) (Boston, MA, USA)* pp 1–4
Kohlhase N, Wegener T, Schaar M, Bolke A, Etxebeste A, Sarrut D and Rafecas M 2020b Capability of MLEM and OE to detect range shifts with a Compton camera in particle therapy *IEEE Trans. Radiat. Plasma Med. Sci.* **4** 233–42
Lerendegui-Marco J, Balibrea-Correa J, Babiano V, Caballero L, Domingo-Pardo C and Ladarescu I 2022 Simultaneous neutron and gamma imaging system for real-time range and dose monitoring in hadron therapy and nuclear security applications *EPJ Web Conf.* **261** 05001
Levitan E and Herman G T 1987 A maximum a posteriori probability expectation maximization algorithm for image reconstruction in emission tomography *IEEE Trans. Med. Imaging* **6** 185–92
Ma Q, Mu D, Zhang R, Gao M, Xiao P, Chen X, Lu H M and Xie Q 2022 Feasibility study of range verification in proton therapy based on neutron information *2022 IEEE NSS MIC* pp 1–3
Marafini M, Gasparini L, Mirabelli R, Pinci D, Patera V, Sciubba A, Spiriti E, Stoppa D, Traini G and Sarti A 2017 MONDO: a neutron tracker for particle therapy secondary emission characterisation *Phys. Med. Biol.* **62** 3299–312
Meric I *et al* 2023 A hybrid multi-particle approach to range assessment-based treatment verification in particle therapy *Sci. Rep.* **13** 6709
Moteabbed M, Espana S and Paganetti H 2011 Monte Carlo patient study on the comparison of prompt gamma and PET imaging for range verification in proton therapy *Phys. Med. Biol.* **56** 1063–82
Nakanishi K, Yamamoto S and Watabe T 2021 Prediction of CT images from PET images using deep learning approach for small animal systems *2021 IEEE Nuclear Science Symp. and Medical Imaging Conf. (NSS/MIC)* pp 1–3
Nuyts J and Matej S 2014 Maximum-likelihood expectation-maximization *Nuclear Medicine Physics: A Handbook for Teachers and Students* ed D L Bailey, J L Humm, A Todd-Pokropek and A Van Aswegen (International Atomic Energy Agency) ch 13.3.3., pp 479–85
Paganetti H 2018 *Proton Therapy Physics* 2nd edn (CRC Press) (https://doi.org/10.1201/b22053)
Pandey S *et al* 2024 Prediction of radiologic outcome-optimized dose plans and post-treatment magnetic resonance images: a proof-of-concept study in breast cancer brain metastases treated with stereotactic radiosurgery *Phys. Imaging Radiat. Oncol.* **31** 100602
Parra L and Barrett H 1998 List-mode likelihood: EM algorithm and image quality estimation demonstrated on 2-D PET *IEEE Trans. Med. Imaging* **17** 228–35
Pausch G, Berthold J, Enghardt W, Römer K, Straessner A, Wagner A, Werner T and Kögler T 2020 Detection systems for range monitoring in proton therapy: needs and challenges *Nucl. Instrum. Methods Phys. Res. A* **954** 161227
Platscher M, Zopes J and Federau C 2022 Image translation for medical image generation—ischemic stroke lesions *Biomed. Signal Process. Control.* **72** 103283
Polf J, Peterson S, Ciangaru G, Gillin M and Beddar S 2009 Prompt gamma-ray emission from biological tissues during proton irradiation: a preliminary study *Phys. Med. Biol.* **54** 731–43
Sarrut D *et al* 2014 A review of the use and potential of the GATE Monte Carlo simulation code for radiation therapy and dosimetry applications *Med. Phys.* **41** 064301
Scarselli F and Chung Tsoi A 1998 universal approximation using feedforward neural networks: a survey of some existing methods and some new results *Neural Netw.* **11** 15–37
Setterdahl L M, Lionheart W R B, Holman S, Skjerdal K, Ratliff H N, Ytre-Hauge K S, Lathouwers D and Meric I 2024 Image reconstruction for proton therapy range verification via U-NETs *Medical Image Understanding and Analysis* ed M H Yap, C Kendrick, A Behera, T Cootes and R Zwiggelaar (Springer) pp 232–44
Shepp L A and Vardi Y 1982 Maximum likelihood reconstruction for emission tomography *IEEE Trans. Med. Imaging* **1** 113–22
Smeets J *et al* 2012 Prompt gamma imaging with a slit camera for real-time range control in proton therapy *Phys. Med. Biol.* **57** 3371–405
Stichelbaut F and Jongen Y 2003 Verification of the proton beam position in the patient by the detection of prompt gamma-rays emission *39th Meeting of the Particle Therapy Co-Operative Group*
Sun J, Du Y, Li C, Wu T-H, Yang B and Mok G S P 2022 Pix2Pix generative adversarial network for low dose myocardial perfusion SPECT denoising *Quant. Imaging Med. Surg.* **12** 3539–55
Sweany M, Galindo-Tellez A, Brown J, Brubaker E, Dorrill R, Druetzler A, Kaneshige N, Learned J, Nishimura K and Bae W 2019 Interaction position, time and energy resolution in organic scintillator bars with dual-ended readout *Nucl. Instrum. Methods Phys. Res. A* **927** 451–62
Turko J *et al* 2024 Characterization of organic glass scintillator bars and their potential for a hybrid neutron/gamma ray imaging system for proton radiotherapy range verification *J. Instrum.* **19** P01008
Wang C-C, Wu P-H, Lin G, Huang Y-L, Lin Y-C, Chang Y-P and Weng J-C 2022 Magnetic resonance-based synthetic computed tomography using generative adversarial networks for intracranial tumor radiotherapy treatment planning *J. Pers. Med.* **12** 361
Wang Z, Bovik A, Sheikh H and Simoncelli E 2004 Image quality assessment: from error visibility to structural similarity *IEEE Trans. Image Process.* **13** 600–12
Wilderman S, Clinthorne N, Fessler J and Rogers W 1998a List-mode maximum likelihood reconstruction of Compton scatter camera images in nuclear medicine *1998 IEEE Nuclear Science Symp. Conf. Record. 1998 IEEE Nuclear Science Symp. and Medical Imaging Conf. (Cat. No.98CH36255)* pp 1716–20
Wilderman S, Rogers W, Knoll G and Engdahl J 1998b Fast algorithm for list mode back-projection of Compton scatter camera data *IEEE Trans. Nucl. Sci.* **45** 957–62
Xiao F *et al* 2024 Prompt gamma emission prediction using a long short-term memory network *Phys. Med. Biol.* **69** 235003
Ytre-Hauge K S, Skjerdal K, Mattingly J and Meric I 2019 A Monte Carlo feasibility study for neutron based real-time range verification in proton therapy *Sci. Rep.* **9** 2011
Zhu B, Liu J Z, Cauley S F, Rosen B R and Rosen M S 2018 Image reconstruction by domain-transform manifold learning *Nature* **555** 487–92