Received: 26 July 2022 | Revised: 10 September 2022 | Accepted: 4 October 2022

DOI: 10.1002/mp.16066

RESEARCH ARTICLE

MEDICAL PHYSICS

Small beams, fast predictions: a comparison of machine learning dose prediction models for proton minibeam therapy

F. Mentzel[1] | K. Kröninger[1] | M. Lerch[2] | O. Nackenhorst[1] | A. Rosenfeld[2] | A. C. Tsoi[3] | J. Weingarten[1] | M. Hagenbuchner[3] | S. Guatelli[2]

[1]Department of Physics, TU Dortmund University, Dortmund, Germany

[2]Centre for Medical Radiation Physics, University of Wollongong, Wollongong, New South Wales, Australia

[3]School of Computing and Information Technology, University of Wollongong, Wollongong, New South Wales, Australia

Correspondence
F. Mentzel, Department of Physics, TU Dortmund University, Dortmund, Germany.
Email: florian.mentzel@tu-dortmund.de

Funding information
Deutsche Forschungsgemeinschaft (DFG), Grant/Award Number: 271512359; Department of Health | National Health and Medical Research Council (NHMRC)

Abstract

Background: Dose calculations for novel radiotherapy cancer treatments such as proton minibeam radiation therapy is often done using full Monte Carlo (MC) simulations. As MC simulations can be very time consuming for this kind of application, deep learning models have been considered to accelerate dose estimation in cancer patients.

Purpose: This work systematically evaluates the dose prediction accuracy, speed and generalization performance of three selected state-of-the-art deep learning models for dose prediction applied to the proton minibeam therapy. The strengths and weaknesses of those models are thoroughly investigated, helping other researchers to decide on a viable algorithm for their own application.

Methods: The following recently published models are compared: first, a 3D U-Net model trained as a regression network, second, a 3D U-Net trained as a generator of a generative adversarial network (GAN) and third, a dose transformer model which interprets the dose prediction as a sequence translation task. These models are trained to emulate the result of MC simulations. The dose depositions of a proton minibeam with a diameter of 800μm and an energy of 20–100 MeV inside a simple head phantom calculated by full Geant4 MC simulations are used as a case study for this comparison. The spatial resolution is 0.5 mm. Special attention is put on the evaluation of the generalization performance of the investigated models.

Results: Dose predictions with all models are produced in the order of a second on a GPU, the 3D U-Net models being fastest with an average of 130 ms. An investigated 3D U-Net regression model is found to show the strongest performance with overall 61.0%$\pm$0.5% of all voxels exhibiting a deviation in energy deposition prediction of less than 3% compared to full MC simulations with no spatial deviation allowed. The 3D U-Net models are observed to show better generalization performance for target geometry variations, while the transformer-based model shows better generalization with regard to the proton energy.

Conclusions: This paper reveals that (1) all studied deep learning models are significantly faster than non-machine learning approaches predicting the dose in the order of seconds compared to hours for MC, (2) all models provide reasonable accuracy, and (3) the regression-trained 3D U-Net provides the most accurate predictions.

KEYWORDS
deep learning, dose prediction, proton minibeam therapy

1 | INTRODUCTION

The relatively novel proton minibeam therapy[1] utilizes grids of sub-millimeter proton beams in order to achieve increased tissue sparing while maintaining high tumor control.[2–6] Dose calculations for research on proton minibeams are typically performed using Monte Carlo (MC) simulations with Geant4[7] or software tools based on it.[8] While general purpose MC simulation codes such as Geant4 are usually adopted as gold standard for dosimetric calculations, they have long execution times.[9] In particular, the combination of high dose gradients caused by the Bragg peak and the high requirements on the spatial resolutions make MC simulation for accurate dose predictions of proton mini-beams very time consuming.

In recent years, an increasing number of publications showcase successfully the application of deep learning for dose prediction in radiotherapy.[10–12] With focus on novel and highly conformal treatments, a recent publication[13] focuses on fast and accurate dose predictions for synchrotron minibeams[14] by training a 3D U-Net-based model[15] to mimic a full MC simulation. This approach mitigates the need for approximate analytical algorithms and allows to include all relevant physical effects directly into the training data of the model, combining high dosimetric accuracy with fast execution times.

This study presents the transferability of the recently published approach[13] to proton minibeam therapy and compares it to two additional machine learning (ML) models. First, the 3D U-Net architecture which is trained as a generator as part of a generative adversarial network (GAN),[16] is also trained as a regression model in this work. Second, a separate novel dose prediction model, *Dose Transformer* (DoTA),[17] which is based on the attention mechanism in modern transformer models[18] and which was presented to achieve highly accurate predictions for the proton pencil beam therapy,[19] is compared to the two 3D U-Net-based models. Instead of predicting the dose deposition in the whole target volume at once, as it is the case for the 3D U-Net models, the DoTA model predicts the dose deposition slice by slice along the depth of the beam. This approach may be more flexible than a volumetric prediction as stacking of different tissue slices is a very general and potentially easier transferable approach to novel geometries than predicting the entire 3D phantom geometry at once.[19] This presents, however, a trade-off between the complexity of the model, whose increase usually requires more training data, and the ease with which a volume is being rendered. In the case of the 3D U-Net model, the structure is relatively simple, as it contains a series of downsampling convolutions followed by a number of upsampling convolutions, which are connected at the same level with skip connections. This allows to predict the entire dose volume with a model of relatively few parameters than can be optimized with relatively small training samples/datasets. This is advantageous especially for novel and pre-clinical treatments, for which it can be very difficult to acquire large training datasets.

This paper is structured as follows. In Section 2, the simulated data are described and the three ML models including different variations are explained. Section 3 presents the obtained dose predictions using the different models and compares them systematically. The results and the limitations of the findings are discussed in Section 4 before conclusions are drawn in Section 5.

2 | METHODS

In this section, the digital phantom used for the MC simulation, the resulting simulated datasets, and finally the three ML models are presented.

2.1 | Monte Carlo data simulation

The interaction of the proton minibeam with the simplified head phantom shown in Figure 1a is simulated using Geant4, version 10.6.p02.[7] The resulting energy depositions are used to train and evaluate the performance of the respective ML models. The simplified head comprises multiple material layers: in its center, an ellipsoid of *G4_BRAIN_ICRP*[20] material represents the brain, shown in Figure 1 as blue ellipse. The brain is covered with a layer of skull (*G4_BONE_COMPACT_ICRU*)[20] (black). The bone layer is slightly thinner at the modeled forehead (minimum thickness of 3 mm) and thicker at the back of the head (minimum thickness of 5 mm). The skull is covered with a 2.5 mm layer of *G4_WATER*[20] modeling the skin (red). The head model is surrounded by *G4_AIR*[20] material.

All energy depositions are scored using a $140 \times 18 \times 18$ voxel grid width an edge length of 1 mm. Tissue information of realistic patient or phantom data for use in pre-clinical or even clinical studies is usually obtained using voxelized CT scans. In the case of the digital phantom used in this work, a voxelized tissue density matrix is obtained during the simulation instead which serves as conditional information of the ML models.

The studied single proton minibeam is modeled as a mono-energetic circular beam of radius $r = 0.4$,mm without divergence. This beam size is well within the suggested range for proton minibeams ($r \leq 1$,mm)

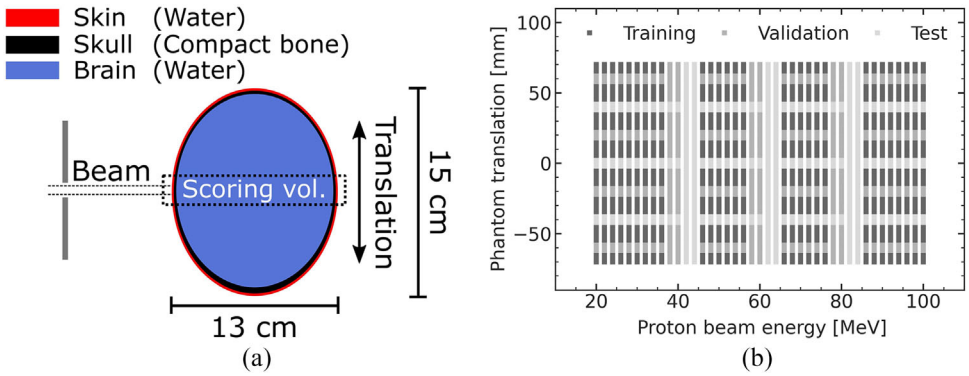

FIGURE 1 (a) Simple head phantom used for data generation[13] (reproduced figure). (b) Visualization of the distribution of the training samples (dark grey), validation samples (medium grey), and test data (light grey) with respect to the beam energy and the phantom translation

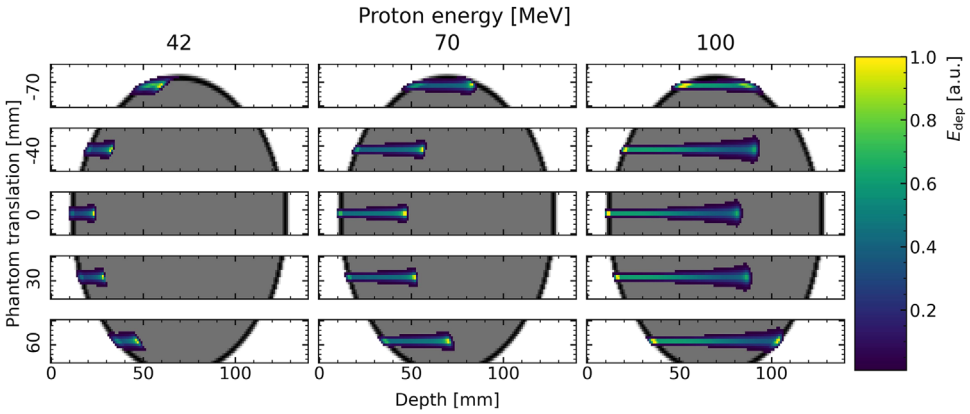

FIGURE 2 Exemplary data samples showing the density matrix in grey-scale (white: air, grey: water, black: bone) and the energy deposition normalized to its maximum for different phantom translations and proton energies

and is around the size of proton minibeams (r = 0.35,mm) used in an early proof-of-concept study.[1] To generate varying prediction geometries, the digital phantom is translated between $\Delta = -70$,mm and $\Delta = +70$,mm in front of a static beam. In addition, the proton beam energy is varied in the range from E = 20,MeV to E = 100,MeV. The pre-built Geant4 physics list *QGSP_BIC_HP* is used as it has been benchmarked for proton therapy.[21] Exemplary simulation results are shown in Figure 2. The change in range of the proton beam, that is, location of the Bragg peak, in the phantom with regard to the proton energy is clearly visible. Wherever the proton beam is incident on parts of the skull which are not perpendicular to the beam, the resulting asymmetry in the shape of the Bragg peak is also observable. For high energies and large phantom translations (e.g., 100 MeV at 70 mm), the proton beam completely traverses the phantom and exists on the distal side.

In total, n = 2911 dose predictions are simulated, split into n_{train} = 1450 training samples, n_{val} = 720 validation samples, and n_{test} = 741 test samples. The split into training, validation, and test data is done by using predetermined cuts on the beam energy and the phantom translation which are shown in Figure 1b: training data shown in dark grey are used to adapt the weights of the neural network (training), the validation data shown in medium grey are used to find the best model configurations (hyperparameter optimization), and the test data in light grey are used to evaluate the performance of the respective best models in a parameter space, which was not previously used during the network optimization (generalization). The exclusive use of certain phantom translations and proton energy ranges for the validation and test data ensures an unbiased estimate of performance during hyperparameter optimization as well as during the generalization assessment and final performance evaluation.

2.2 | Machine learning models

In the following, the different ML models used in this study are presented. The first two models comprise the same dose prediction network architecture presented previously[13] which is based on the 3D U-Net architecture.[15] The two 3D U-Net models differ in the way they are trained. The third model comprises a

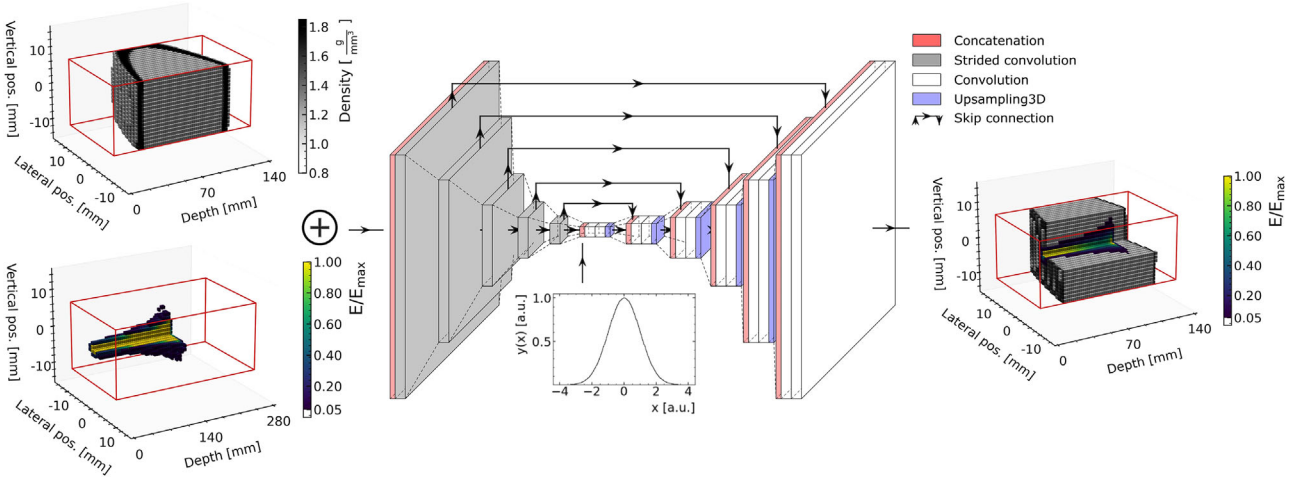

FIGURE 3 Schematic of the 3D U-Net-based model including the model inputs on the left, the layer structure in the center and the model output on the right. Adapted from a previous publication[13]

transformer architecture recently proposed for dose deposition predictions in the proton therapy[19] based on the *self-attention mechanism*.[18]

2.2.1 | 3D U-Net-based model

The 3D U-Net architecture was introduced in 2016 for the purpose of medical image segmentation.[15] In recent studies, it has been shown that 3D U-Nets are also capable of accurate dose/energy deposition predictions in radiotherapy.[13] A schematic of the 3D U-Net-based model developed in the given reference, referred to hereafter as *dose generator network*, is shown in Figure 3. The model is conditioned on two inputs: a 3D tissue density matrix of the region of interest in the head phantom and either the scalar energy of the proton beam or the energy deposition matrix simulated in a homogeneous water phantom. The latter introduces the beam characteristics including energy spectrum and beam size into the model, gives additional information on how the beam would interact with water and does not depend on the target geometry. Such a simulation is performed for every proton energy used in the data set. In the central convolutional block, the *bottleneck*, a Gaussian noise vector of length 100 is concatenated with the data to allow for statistical variations and more robustness. The output of the model comprises the energy deposition matrix in the respective target density matrix.

The use of the energy deposition in homogeneous water as means of introducing the beam characteristics into the phantom, in the following referred to as *water-only condition*[13] proved to be successful for the case of a synchrotron X-ray radiation. This finding needs to be re-evaluated for the application on proton beam dose prediction. For X-ray radiation, phantom heterogeneities mainly influence the magnitude of the energy deposition in the path of the beam as more dense materials like bone receive more energy deposition compared to less dense tissue. In the case of a proton or heavier ion beam, the impact of heterogeneities on the resulting beam spread and attenuation due to the nature of the densely ionizing interactions of protons and heavier ions in a medium is a lot more prominent than in the case of X-rays. The change in Bragg peak location is clearly visible in the exemplary data shown in Figure 2. To investigate the potential benefits of the previously described water-only condition for the proton therapy, it will be compared to a *scalar energy condition*, in which the proton energy E_{proton} will be passed to the model normalized to the maximum used energy $E_{\text{max}} = 100,\text{MeV}$ ($E' = E_{\text{proton}}/E_{\text{max}}$). The resulting scalar value is used to fill a matrix of the same shape as the density matrix ($140 \times 18 \times 18$ voxels) which is then concatenated to the density matrix along a new, fourth axis. This results in the same input data format of $140 \times 18 \times 18$ as in the water-only condition case.

The 3D U-Net dose generator network is trained and compared using two different modes, (1) as part of a GAN model and (2) as a regression model, which are explained in the following two subsections.

2.2.2 | Generative Adversarial Network

It was shown that training the *dose generator network* as part of a GAN yields high-accuracy energy deposition predictions.[13] The idea of a GAN is to train two competing networks. The generator network predicts the energy depositions given a beam characteristic and a density matrix, while the critic network evaluates how likely a given energy deposition matrix is originating from the generator of the MC simulation. The critic receives as conditional information the respective density matrix in addition to the water-only condition matrix and an energy deposition matrix. The difference in the critic responses to either MC or ML generated samples is interpreted as *Wasserstein distance*.[22] This

distance is used as loss function for the generator network, it is effectively trained to produce energy deposition predictions which are indistinguishable from Geant4 simulations.

Two different versions of the GAN model are studied, a default version using the water-only condition and one with the scalar energy condition, which are referred to as model GAN-W (water-only) and model GAN-S (scalar energy).

2.2.3 | Regression model

Interpreting the learning task as a *regression problem* is a more conventional method of optimizing the weights of a neural network toward a desired model output. In this case, a loss function is computed from the comparison of the predicted energy deposition matrix of the trained neural network with the respective Geant4 simulation result. In this study, two commonly used loss functions for regression models, the mean squared error (MSE) and the mean absolute error (MAE) are compared. The network is trained using either of these loss functions with a learning rate between 1×10^{-2} and 1×10^{-5} and a batch size of 32, which is limited by the memory of the used computing hardware (Nvidia GeForce 1080 Ti, 11 GB).

In the results section of this paper, the regression models are named by the used loss function and the exponent of the learning rate, for example, MAE-3 = MAE with learning rate 1×10^{-3}. A following *W* indicates the use of the water-only condition, while models without *W* were trained using the scalar energy condition. One model, denoted with an additional *D*, is trained with a higher dropout ratio[23] than the originally proposed model.

2.2.4 | Transformer-based model

It was recently shown that the transformer architecture is very suitable for dose deposition predictions in the case of the proton beam therapy.[17,19] Transformer models rely on the prediction of sequences to sequences utilizing the so-called *attention mechanism*.[18] In the case of the model of interest, named *DoTA*,[17,19] the source sequence comprises *density matrix tokens*, which were obtained from encoding the phantom tissue density matrix slice by slice using a convolutional encoder. The target sequence comprises *energy deposition tokens* which are decoded via a convolutional decoder into energy deposition slices. The translation of the tissue slice sequence into the dose deposition slice sequence is performed using a so-called transformer encoder.[17,18] The slice-based approach of DoTA has been shown to allow a flexible and accurate prediction in a variety of phantoms. This may render this approach superior to a volume-based one such as a 3D U-Net. A thorough introduction of the DoTA model is available in the respective published article.[19]

Two DoTA-based models are trained for comparison: the original model as taken from a publicly available online repository,[25] and a second using the MAE as loss together with the Adam optimizer[24] instead, referred to as DoTA-O (original) and DoTA-A (adapted), respectively. The same model architecture as published[19] is used for both DoTA models, however, the weights are optimized using the simulated data of this study as the energy ranges and the scoring resolution differ.

2.2.5 | Training stop criterion and performance evaluation

For training and evaluation of the models, the energy deposition is used instead of the dose. This was found to produce more robust training data especially out-of-field in air as the dose values in those areas are suspect to large absolute variations due to the division by a very low density.

All models are trained until the relative frequency of voxels with a relative deviation $\Delta E_{\text{rel}} = (E_{\text{ML}} - E_{\text{Geant4}})/E_{\text{Geant4}}$ of less than 1% does not increase anymore for 100 epochs. In this, E_{ML} and E_{Geant4} are the respective energy deposition matrices obtained from the ML model or the MC simulation, respectively. The frequency of voxels exhibiting a lower deviation than 3% or 1% is being referred to as the respective passing rates. The model with the highest passing rate on the validation data is chosen as the best model, which is evaluated in more detail using the test data. In addition to the 1% passing rate, the results are compared also with respect to the passing rate based on a 3% deviation.

In order to get additional insight into the accuracy of the energy predictions, 2D depth-dependent energy deposition-plots are used to investigate the differences between the applied models.

Both the 3D U-Net and the DoTA approach have been demonstrated to allow for high-accuracy predictions.[13,19] The used metric in this study is chosen to be more strict than for example the commonly used gamma index passing rate,[26] as it does not allow for spatial deviations at all. While it may not be directly comparable to measures for clinic applications, it allows for better differentiation in the scope of this study.

3 | RESULTS

3.1 | Hyperparameter optimization

Out of the investigated models, the 3D U-Net regression model trained with MAE as loss and a learning rate of 1×10^{-3} (MAE-3) performs best. An overview of the training and validation data performance for different

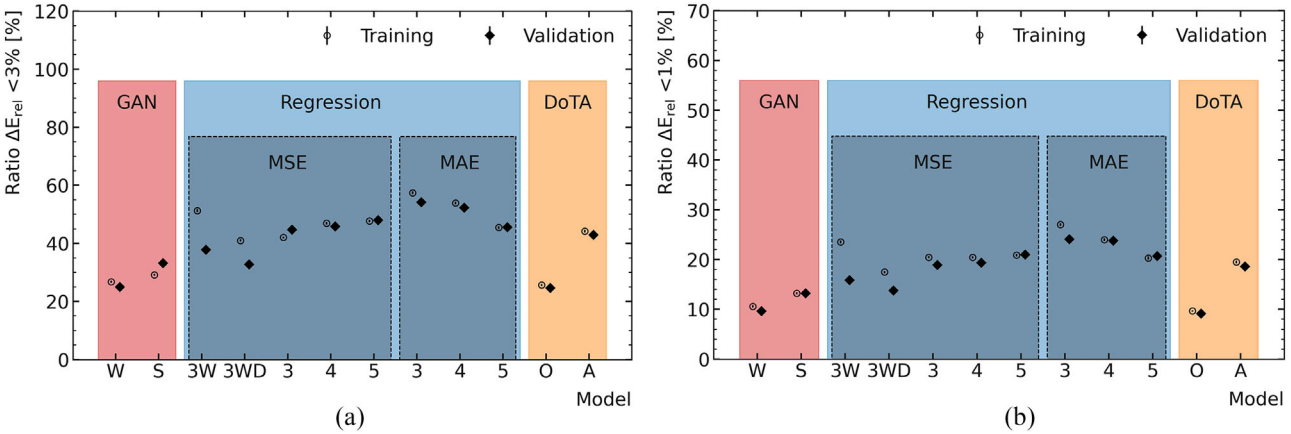

FIGURE 4 Rates of voxels with a relative deviations of less than 3% (a) and 1% (b), respectively for the generative adversarial network (GAN)-based and regression-based 3D U-Net as well as for the Dose Transformer (DoTA)-based model. The areas denoted with mean absolute error (MAE) and mean squared error (MSE) point at the respective loss functions used for training. The IDs shown on the *x*-axis indicate the individual modifications to aspects such as the model input, learning rate and dropout rate and are explained in detail in the text of Section 2.2

tested model configurations, measured using the 1% and 3% passing rates introduced in Section (2.2.5), is shown in Figure 4. The performance of the GAN model is below both the regression and the DoTA-based models. The use of the scalar energy condition (GAN-S) increases the performance relative to the use of the water-only condition (GAN-W). In the case of the regression models, using the water-only condition (MAE-3W) degrades the validation performance and increases the generalization gap, which indicates overtraining. Applying stronger dropout (MAE-3WD) could only partially mitigate the overtraining, while resulting also in a further loss in performance on the validation data. As a consequence, the scalar proton energy is used for the best model, which does not contain any information about the beam shape.

The adapted DoTA-based model (DoTA-A) performs better on this dataset than the original model (DoTA-O), but worse than the best regression model.

3.2 | Performance evaluation

For the final performance evaluation, the respective best GAN, regression and DoTA model is used to predict the energy depositions for the test data set. To inspect the generalization ability of the models, the performance on the test data is compared to the performance on a subset of the training data set in Table 1. The reason for using only a subset of the training dataset is explained in the following.

The performances at the extremes of the phantom translation are the lowest overall. This is a result of the high amount of bone in the beam at the front and the back of the head and also of the fact that high-energy proton beams are not completely stopped inside the phantom. Because most of these extreme cases are assigned to be training data, the average performance on the whole training data is reduced significantly. Whenever the beam hits the phantom in the central part (±50,mm), the performance is quite stable. Therefore, to allow for a more fair comparison between training and test performance in order to examine how well the models are able to generalize to unknown data, samples with a phantom translation of more than ±50,mm are not considered.

The MAE and passing rates confirm the regression-trained 3D U-Net as the most accurate model. While all three models exhibit similar performances on training and test data, the agreement is best for the regression model, indicating that the model generalizes best to unknown data.

The test data performance of the three models with respect to the phantom translation is shown in Figure 5a. For each shown phantom translation, the results are averaged over all proton energies. The regression 3D U-Net outperforms the other models throughout the whole range of phantom translations. For all models, the performance decreases toward larger absolute translations, where the curvature of the head phantom increases which leads to more skull material being located in the path of the beam. The drop in performance toward the maximum translations is not symmetrical. This is a result of the asymmetric head phantom and the fact that the performance is worse whenever more bone material irradiated.

Figure 5b shows the performance in dependence of the proton energies, averaged over all phantom translations. With respect to the energy dependence of the predictions, shown in Figure 5b, the performance of all models increases almost linearly up to approximately 50–60 MeV. This is directly caused by the fact

TABLE 1 MAE and relative deviation passing rates for voxels based on different criteria for the validation dataset

Model	Dataset	MAE (1×10^{-4})	$\Delta E_{rel} < 1\%$ (%)	$\Delta E_{rel} < 3\%$ (%)
3D U-Net	Training	11.64 ± 0.19	10.50 ± 0.19	30.6 ± 0.5
(GAN)	Test	12.98 ± 0.23	11.00 ± 0.18	33.1 ± 0.4
3D U-Net	Training	4.38 ± 0.01	25.87 ± 0.28	61.2 ± 0.5
(Regression)	Test	4.74 ± 0.03	25.62 ± 0.29	61.0 ± 0.5
DoTA	Training	5.27 ± 0.02	21.99 ± 0.28	48.6 ± 0.5
	Test	6.25 ± 0.06	20.45 ± 0.34	46.1 ± 0.6

Abbreviations: DoTA, Dose Transformer; GAN, generative adversarial network; MAE, mean absolute error.

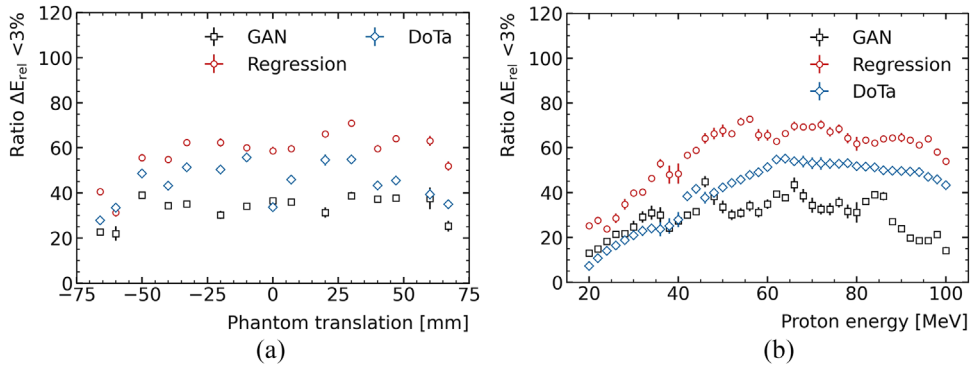

FIGURE 5 Model performances with respect to the phantom translation, averaged over all proton energies (a) and with respect to the proton energies, averaged over all phantom translations (b)

that the lower the beam energies, the higher the relative dose gradients are in very few voxels at the entrance of the phantom. For energies higher than 50 MeV, the performances decrease by a varying degree. While the regression 3D U-Net exhibits the best test data performance, the DoTA model exhibits the lowest variance in prediction performance with respect to the proton energy.

3.3 | Generalization assessment

While Figure 5 allowed for a first assessment of trends in the test data performance, a more detailed analysis of the network performance on all training, validation, and test data reveals additional insight about the generalization, strengths, and weaknesses of the three trained models. The mean value of the absolute errors for each data sample predicted by the three models is shown in Figure 6. The background colors indicate whether samples belong to the training (white), validation (light grey), or test data (dark grey). The data split is the same as in Figure 1b and indicated here for easier optical inspection.

The GAN model shows the overall lower performance than the two other models. With regard to generalization, especially the predictions for energies not part of the training data seems to lead to lower performances, indicating some overtraining. Although the performance along the phantom translation axis is not as good as with the other models, there are less signs of overtraining along that parameter which is indicated by less impact on the performance whether data samples are from the training, the validation, or the test dataset.

The regression model shows an average deviation of less than 5% for most of the parameter space, the deviations for low energies, large phantom translations (especially toward the negative direction) tend to exhibit larger deviations, mostly up to 10%. While good agreement overall between training and test performance is described in Table 1, a lower performance is seen for the validation and test energies around 60 and 80 MeV around the center of the phantom. A similar trend can be seen for the GAN model which suggests that the interpolation along the energy axis might be generally more difficult to achieve with a 3D U-Net model. Especially for negative translations of around −60 mm (part of validation dataset) the deviations are slightly higher than for the surrounding training data samples. This indicates some overtraining in that parameter space.

The DoTA model predictions exhibit interesting features as well. The performance on fringe parameter

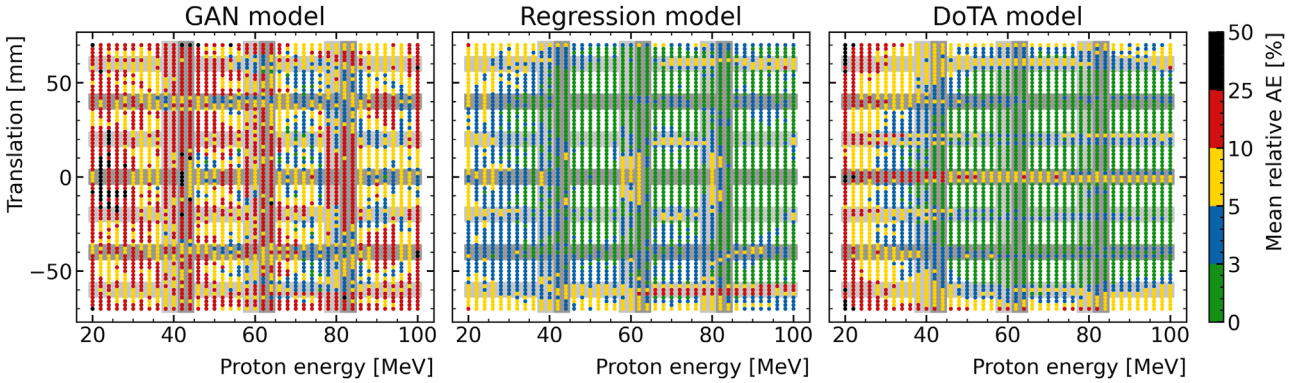

FIGURE 6 Mean relative absolute error (AE) on training (white background), validation (light grey background), and test data (dark grey background) of the three machine learning (ML) models in dependence of both proton energy and the phantom translation

configurations is significantly lower than those surrounded by more training data samples. The performance on validation or test data is generally worse compared to the training data closest in the parameter space, which indicates overtraining. This suggests that while the model exhibits strong interpolation capabilities especially along the energy axis, the model is limited by the amount of training data that densely samples the parameter space.

3.4 | Analysis of exemplary prediction results

In addition to the previously used performance measures and investigations, it is instructive to exemplary inspect individual predicted energy depositions of the ML models and their deviation from the MC simulation and how they differ depending on the phantom configuration and the proton energy. The energy deposition predictions and relative deviations for four exemplary settings are shown in Figure 7. Only voxels with an energy deposition of at least 1% of the maximum energy deposition are shown on top of the visualization of the density matrix to allow for a better visual differentiation. The parameters for the shown examples are chosen to provide a representative overview of the prediction results of apparently challenging configurations: ($\Delta = 0$,mm | $E = 42$,MeV, left column) shows one of the worst examples using the DoTA model, both ($\Delta = 60$,mm|$E = 42$,MeV, second from left column), and ($\Delta = 0$,mm | $E = 62$,MeV, second from right column), are predicted with higher accuracy but appear to be still challenging configurations to both the regression and the DoTA model, while ($\Delta = -60$,mm | $E = 62$ MeV, right column), is among the worst examples for the regression model. Examples with less than 30 MeV were not considered as the results are difficult to inspect visually due to the very low penetration depth of the resulting proton beam. The top row shows the deposited energies of the Geant4 simulation for the given phantom translation and proton energy configurations. Below that, the predictions and their relative deviations from the Geant4 simulation are shown for the GAN model, the regression model, and the DoTA model, respectively.

Although performing worst among the tested models in terms of the passing rates, the GAN model is able to reproduce the overall shape of the energy deposition distribution of the proton beam. However, the range of the proton beam is slightly overestimated by around 1 mm for the first three shown examples, which is visible by an overestimation of more than 10% of the deposited energy toward the end of the range. For the high-energy example in the right column of Figure 7, an underestimation of the deposited energy at the right distal end is observed. This indicates a mismodeling of the Bragg peak shape. While the sloped phantom surface should result in an asymmetric Bragg peak, the GAN model predicts a rather flat distal edge not reproducing this characteristic of the energy deposition pattern.

Although it is shown in Figure 5b that the low-energy regime is the weakest range for dose prediction of the models, the relative deviations of the regression model is mostly smaller than 10% with respect to the Geant4 simulation. For the intermediate energies example, a slight underestimation of about 1 mm of the proton range is seen. For high energies and large negative phantom translations, the lowest prediction performance of the regression model is seen. In Figure 7, it can be seen that this is caused by an underestimation in range of the proton beam for this parameter configuration. The shape of the Bragg peak is accurately predicted despite that. Overall, the predictions are accurate enough to be used in, for example, preliminary treatment plan optimization tasks for which for which underestimations of up to 10% are acceptable.

The DoTA-based model shows similar prediction accuracy as the regression model. In this case though, the model systematically underestimates the range of the beam by about 1 mm. The lowest performance is

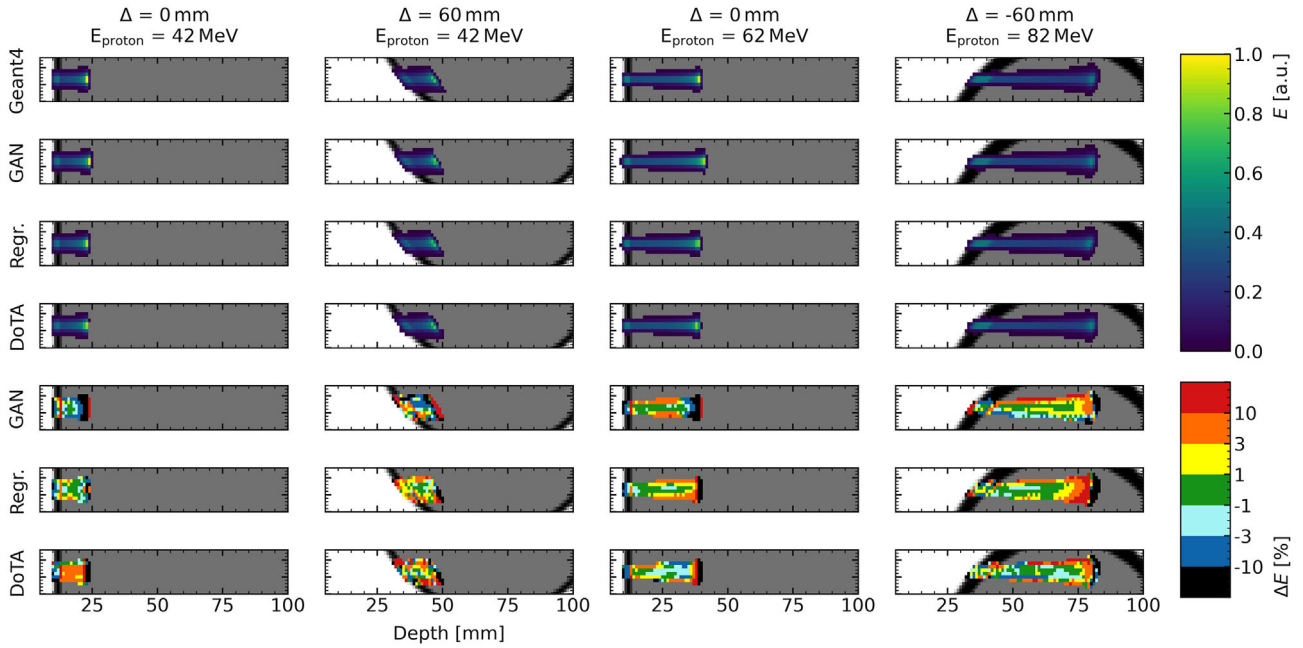

FIGURE 7 2D energy deposition slices obtained from Geant4 simulations (top row) and the respective predictions and relative deviations as achieved by the generative adversarial network (GAN)-trained 3D U-Net (rows 2 and 5), the regression-trained 3D U-Net (rows 3 and 6) and the Dose Transformer (DoTA)-based model (rows 4 and 7) at the vertical center of the field of view for phantom translations and proton energies of 0 mm and 42 MeV (left), 60 mm and 42 MeV (second from left), 0 mm and 62 MeV (second from right), and −60 mm and 82 MeV (right)

achieved in the low-energy range as expected from the results shown in Figure 5b. While the energy deposition in the skull is underestimated, the deposition inside the brain tissue is overestimated. The shape of the distal edge is predicted very accurately which is seen by only small left–right-asymmetry in the deviation from the Geant4 simulations. The shape of the Bragg peak is predicted relatively well even in the case of asymmetric Bragg peaks.

4 | DISCUSSION

While the GAN 3D U-Net model is outperformed by the regression 3D U-Net and the DoTA-based model, the latter both produce reasonably accurate energy deposition predictions for the case of proton minibeams. Overall, the 3D U-Net-based model is found to result in slightly more accurate predictions compared to the DoTA model in this study.

Using a common CPU workstation configuration (Intel Xeon E5-2630 v4 @ 2.20GHz, 10 CPU cores, 20 threads), the Geant4 MC simulation takes approximately 1 h for each sample when being distributed over the available 20 threads. While this time could be reduced by using more CPUs, for example, by accessing a computing cluster, all ML models allow for significantly faster dose predictions, which are all in the order of seconds (see Table 2). On the same CPU architecture, the 3D U-Net takes about 0.65 s, while the DoTA model takes about 2.2 s for a single prediction. Using a consumer-level GPU (Nvidia GeForce 1080 Ti, 11 GB), the 3D U-Net is found to be the fastest with around 0.13 s/prediction, while the DoTA model is found to take around 1.1 s/prediction, which makes the 3D U-Net even around 10 times faster than the DoTA model on a GPU.

Loading the density matrices into memory is part of the assessed prediction time, generating the density matrices from the geometry data is performed in a preprocessing step. The ML predictions are performed using a batch size of 1, meaning that only one sample is predicted at a time. While the sequential prediction of dose distributions is an important use case for treatment plan optimization, significantly higher prediction throughput and thereby average time per prediction can be achieved by predicting multiple samples in the one step. The deviation from faster prediction times reported for the DoTA model[17] might be due to the use of a standard-level GPU in this study in contrast to a

TABLE 2 Prediction times of the 3D U-Net and Dose Transformer (DoTA) model on CPU (Intel Xeon E5-2630 v4 @ 2.20GHz) and GPU (Nvidia GeForce 1080 Ti, 11GB) hardware, compared to the Geant4 Monte Carlo (MC) simulation

	CPU/GPU	Prediction time (s)	Rel. speed
Geant4	1 CPU	≈72 000	Reference
3D U-Net	1 CPU	0.65 ± 0.01	$\approx 1.1 \times 10^6$
	1 GPU	0.13 ± 0.07	$\approx 5.5 \times 10^6$
DoTA	1 CPU	2.2 ± 0.1	$\approx 3.2 \times 10^5$
	1 GPU	1.05 ± 0.04	$\approx 6.8 \times 10^5$

high-end GPU used by the original authors. Depending on the available hardware in future studies, the achieved prediction times will therefore have to be re-evaluated.

An important limitation of this finding is that the hyperparameters of the DoTA-based model were not fine-tuned to the problem at hand but were taken from the publication introducing the model,[17] more specifically the online available code.[25] With further optimization to the studied problem, the DoTA model may improve in performance. While the 3D U-Net model was not fine-tuned in terms of hyperparameters compared to the publication presenting the model[13] either, it was previously optimized on the same phantom used in this study although the previous study was performed using photon beams.

Transformer-based models are known to rely on larger training datasets than other algorithms.[27] This indicates that providing the dose transformer model with a larger database might improve its performance further. Especially for novel and pre-clinical treatments as the proton minibeam therapy discussed in this work, creating vast datasets can be difficult or even prohibitive. Models like the 3D U-Net which can be trained on fewer datasets can be a strong alternative in these cases as shown in this work and other published articles.[13] Aside from this, the observed overtraining in some parts of the parameter space suggests the general need for a more densely sampled parameter space for training data.

While globally no clear sign of overtraining is observed, a closer investigation of the performance in dependence of the proton energy and the phantom translation reveals some local differences indicating overtraining in some parts of the parameter space for all models. By conducting a systematic investigation of performance and differences among different ML approaches, our study shows the merit in evaluating different ML models in a simplified and reduced setup such as the used simplified head phantom. Comparative studies like this can help choosing the right model for users trying to bring ML models to their own pre-clinical and clinical scenarios. Future work should be directed at investigating systematic ways to compare ML algorithms on more complex models potentially suitable for application in pre-clinical or even clinical settings.

5 | CONCLUSION

In this study, three model architectures, two based on the 3D U-Net and one based on the novel DoTA model, were compared in prediction accuracy and speed for dose prediction in the proton minibeam therapy. While overall both the 3D U-Net regression and the DoTA-based model produce accurate results and are promising candidates for fast dose prediction engines in the proton minibeam therapy, the 3D U-Net was found to be more accurate and faster in execution. Especially for applications with limited training data or a sparsely sampled parameter space, the findings implicate that 3D U-Net-based models are most suitable for dose prediction learning tasks.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the computing time provided on the Linux HPC cluster at Technical University Dortmund (LiDO3), partially funded in the course of the Large-Scale Equipment Initiative by the German Research Foundation (DFG) as project 271512359. The authors acknowledge the contribution of the University of Wollongong with NHMRC Near Missed funding.

FUNDING

the Large-Scale Equipment Initiative by the German Research Foundation (DFG), project no.: 271512359

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

REFERENCES

1. Prezado Y, Fois GR. Proton-minibeam radiation therapy: a proof of concept. *Med Phys*. 2013;40:031712.
2. Zlobinskaya O, Girst S, Greubel C, et al. Reduced side effects by proton microchannel radiotherapy: study in a human skin model. *Radiat Environ Biophys*. 2013;52:123-133.
3. Girst S, Greubel C, Reindl J, et al. Proton minibeam radiation therapy reduces side effects in an in vivo mouse ear model. *Int J Radiat Oncol Biol Phys*. 2016;95:234-241.
4. Prezado Y, Jouvion G, Guardiola C, et al. Tumor control in RG2 glioma-bearing rats: a comparison between proton minibeam therapy and standard proton therapy. *Int J Radiat Oncol Biol Phys*. 2019;104:266-271.
5. Lansonneur P, Mammar H, Nauraye C, et al. First proton minibeam radiation therapy treatment plan evaluation. *Sci Rep*. 2020;10:1-8.
6. Lamirault C, Brisebard E, Patriarca A, et al. Spatially modulated proton minibeams results in the same increase of lifespan as a uniform target dose coverage in F98-glioma-bearing rats. *Radiat Res*. 2020;194:715-723.
7. Agostinelli S, Allison J, Amako K, et al. GEANT4—a simulation toolkit. *Nucl Instrum Methods Phys Res A*. 2003;506:250-303.
8. Guardiola C, De Marzi, L, Prezado Y. Verification of a Monte Carlo dose calculation engine in proton minibeam radiotherapy in a passive scattering beamline for preclinical trials. *Br J Radiol*. 2020;93:20190578.
9. Jabbari K. Review of fast Monte Carlo codes for dose calculation in radiation therapy treatment planning. *J Med Signals Sens*. 2011;1:73-86.
10. Kearney V, Chan JW, Haaf S, Descovich M, Solberg TD. DoseNet: a volumetric dose prediction algorithm using 3D fully-convolutional neural networks. *Phys Med Biol*. 2018;63:235022.
11. Kontaxis C, Bol GH, Lagendijk JJ, Raaymakers BW. DeepDose: towards a fast dose calculation engine for radiation therapy using deep learning. *Phys Med Biol*. 2020;65:075013.
12. Kearney V, Chan JW, Wang T, et al. DoseGAN: a generative adversarial network for synthetic dose prediction using attention-gated discrimination and generation. *Sci Rep*. 2020;10:11073.
13. Mentzel F, Kröninger K, Lerch M, et al. Fast and accurate dose predictions for novel radiotherapy treatments in heterogeneous

phantoms using conditional 3D-UNet generative adversarial networks. *Med Phys*. 2022;49:3389-3404.
14. Bräuer-Krisch, E,et al. Medical physics aspects of the synchrotron radiation therapies: Microbeam radiation therapy (MRT) and synchrotron stereotactic radiotherapy (SSRT). *Physica Med*. 2015;31:568-583.
15. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. Medical Image Computing and Computer-Assisted Intervention (MICCAI). *LNCS*. 2016;9901:424-432.
16. Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. *Adv Neural Inf Process Syst*. 2014;27:2672-2680.
17. Pastor-Serrano O, Perkó Z. Learning the physics of particle transport via transformers. 2021. arxiv:2109.03951.
18. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Adv Neural Inf Process Syst*. 2017;30:5999–6009.
19. Pastor-Serrano O, Perkó Z. Millisecond speed deep learning based proton dose calculation with Monte Carlo accuracy. *Phys Med Biol*. 2022;67:105006.
20. Geant4 Collaboration. *Geant4 Material Database—Book For Application Developers 11.0 documentation*. 2017. Accessed: May 30, 2022. https://geant4-userdoc.web.cern.ch/UsersGuides/ForApplicationDeveloper/html/Appendix/materialNames.html
21. Arce P, et al. Report on G4-Med, a Geant4 benchmarking system for medical physics applications developed by the Geant4 Medical Simulation Benchmarking Group. *Med Phys*. 2021;48:19-56.
22. Arjovsky M, Chintala S, Bottou L. Wasserstein GAN. *arxiv: 1701.07875*, 2017.
23. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. *J Mach Learn Res*. 2014;15:1929-1958.
24. Kingma DP, Ba JL. Adam: a method for stochastic optimization. 2015. arXiv:1412.6980.
25. Pastor-Serrano O. Dose calculation via transformers. GitHub Repository. 2022. Accessed: July 25, 2022. https://github.com/opaserr/dota
26. Low DA, Harms WB, Mutic S, Purdy JA. A technique for the quantitative evaluation of dose distributions. *Med Phys*. 1998;25(5):656-661.
27. Xu P, Kumar D, Yang W, et al. Optimizing deeper transformers on small datasets. in *ACL-IJCNLP 2021—59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*. Association for Computational Linguistics (ACL). 2021. pp. 2089-2102.

How to cite this article: Mentzel F, Kröninger K, Lerch M, Nackenhorst O, Rosenfeld A, Tsoi AC, et al. Small beams, fast predictions: a comparison of machine learning dose prediction models for proton minibeam therapy. *Med Phys*. 2022;49:7791–7801.
https://doi.org/10.1002/mp.16066