Artificial Intelligence in Radiotherapy

Probabilistic Deep Learning for Dose Prediction and Anatomy Modeling

Pastor Serrano, O.

DOI
10.4233/uuid:c0c501d2-7c05-4e95-b8e8-d81aab627bb9

Publication date
2023

Document Version
Final published version

Citation (APA)
Pastor Serrano, O. (2023). *Artificial Intelligence in Radiotherapy: Probabilistic Deep Learning for Dose Prediction and Anatomy Modeling*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:c0c501d2-7c05-4e95-b8e8-d81aab627bb9

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Artificial Intelligence in Radiotherapy

Probabilistic Deep Learning for Dose Prediction and Anatomy Modeling

Artificial Intelligence in Radiotherapy

Probabilistic Deep Learning for Dose Prediction and Anatomy Modeling

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus prof. dr. ir. T.H.J.J. van der Hagen
chair of the Board of Doctorates
to be defended publicly on
Monday 1 May 2023 at 17:30 o'clock

by

Oscar PASTOR SERRANO

Master of Science in Physics,
KTH Royal Institute of Technology, Stockholm, Sweden
born in Valencia, Spain.

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof. dr. M.S. Hoogeman	Erasmus University Medical Center, promotor
Dr. ir. D.R. Schaart	Delft University of Technology, promotor
Dr. Z. Perkó	Delft University of Technology, copromotor
Independent members:	
Prof. dr. ir. J. Sonke	University of Amsterdam
Prof. dr. C. Richter	Dresden University of Technology, Germany
Prof. dr. ir. M. Staring	Leiden University Medical Center
Prof. dr. P.A.N. Bosman	Delft University of Technology
Prof. dr. ir. M.B. van Gijzen	Delft University of Technology, reserve member
Other members:	
Prof. dr. L. Xing	Stanford University, USA

Keywords: Radiation therapy, deep learning, treatment plan robustness, inter-fraction uncertainties, intra-fraction uncertainties, latent variable models, convolutional neural networks, transformer neural networks, autoencoder, radiation dose calculation, breathing interplay effects, anatomy models, generative models

ISBN 978-94-6419-790-7
Printed by: Gildeprint - Enschede.

The research described in this thesis was developed in the Medical Physics and Technology section of the Department of Radiation, Science and Technology of the Delft University of Technology (Delft, Netherlands), as well as in the Department of Radiation Oncology of Stanford University (CA, USA). The work described in this thesis has been financially supported by KWF Kanker Bestrijding, being part of the research project PAREL with grant number 11711.

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

Contents

Summary

Radiotherapy cancer treatments aim at killing tumor cells with radiation. Current proton and photon therapy workflows are based on irradiating the patient over the course of several daily sessions (or fractions, ranging from a few to typically 30) using treatment plans based on a planning computed tomography (CT) scan obtained days before starting the treatment course. Ideally, the treatment would completely irradiate the tumor without damaging the surrounding critical organs, but this is physically impossible due to the presence of errors and uncertainties from several sources. Setup errors arise from the imprecision in positioning the patient at the same exact location every fraction. Intra-fraction anatomical variations cause some organs to move during delivery (e.g., liver or lung due to breathing). Most importantly, due to the time lapse between planning and treatment, inter-fraction anatomical changes can cause even larger differences in anatomy (e.g., changes in rectum filling or tumor shrinking). All such uncertainties affect how the dose is deposited in the tumor and surrounding structures, and can ultimately compromise treatment effectiveness if not accounted for.

Current photon and proton plans try to mitigate the detrimental effect of uncertainties a-priori during treatment plan optimization. Photon treatments use margin extensions on the target, directly aiming to irradiate larger volumes. Proton workflows optimize treatment plans simultaneously over a set of error scenarios and subsequently evaluate treatment plan robustness against possible uncertainties in many different error scenarios, ultimately also resulting in irradiated volumes that are larger than the actual clinical target. Photon margins are able to cover for setup errors and also for anatomical variations to some extent, but current proton robust optimization and robustness evaluation approaches only account for setup and range calculation errors, partially due to the lack of realistic anatomy motion models in the literature.

To maximally reduce the effect of inter-fraction organ and tumor motion, next generation online adaptive workflows aim at verifying — and, if necessary, correcting — treatment plans in a couple of minutes before delivery of each treatment session. Such workflows would mitigate the detrimental effects of uncertainties, by adapting to anatomical variations, thereby allowing reduced margins (for photons) and lower robustness settings in the optimization (for protons), and compromising less healthy tissue for the goal of irradiating the tumor. Such adaptive approaches put even stringent requirements on the speed of the dose calculations than robust optimization and robustness evaluation, requiring (among others) algorithms to predict dose delivery in few seconds, currently lacking in the radiotherapy community.

This thesis aims at solving both the problem of the slow dose prediction speed and the absence of anatomical models with a combination of deep learning and probabilistic modeling concepts. The first half of the thesis presents novel methods to predict photon beam or proton pencil beam doses in few milliseconds, while the second half

addresses the simulation of anatomical variations during and in-between fractions.

The primary challenge of current dose calculation approaches is that neither Monte Carlo (MC), nor analytical pencil beam algorithms (PBA) can meet both the stringent speed and accuracy requirements needed for adaptation. As a potential solution, Chapter 2 presents a deep learning based millisecond speed dose calculation algorithm (DoTA) accurately predicting the dose deposited by proton pencil beams for arbitrary energies and patient geometries. Given the forward-scattering nature of protons, 3D particle transport is framed as modeling a sequence of 2D CT geometries in the beam's eye view. DoTA combines convolutional neural networks extracting spatial features (e.g., tissue and density contrasts) with a transformer self-attention backbone routing information between the sequence of geometry slices and a vector representing the beam's energy, and is trained to predict low noise MC simulations of proton beamlets. Predicting beamlet doses in 5 ± 4.9 ms with very high gamma pass rates of $99.37 \pm 1.17\%$ (3 mm, 1%), DoTA significantly improves upon analytical pencil beam algorithms both in precision and speed. Offering MC accuracy 100 times faster than PBAs and 10,000 times faster than MC, our model calculates full treatment plan doses in 10 s to 15 s depending on the number of beamlets (800-2200 in our plans), achieving a $99.70 \pm 0.14\%$ (2mm, 2%) gamma pass rate across 9 test patients. DoTA represents a new state of the art in deep learning-based dose calculation and can directly compete with the speed of even commercial multi GPU MC approaches.

In Chapter 3, the DoTA architecture is extended to predict broad photon beam dose distributions in few milliseconds. The proposed improved Dose Transformer Algorithm (iDoTA) maps arbitrary patient geometries and beam information (in the form of a 3D projected shape resulting from a simple ray tracing calculation) to their corresponding 3D dose distribution. Treating the 3D CT input and dose output volumes as a sequence of 2D slices along the direction of the photon beam, iDoTA solves the dose prediction task as sequence modeling — similar to DoTA —, using a series of convolutions, residual connections and a transformer backbone. iDoTA predicts individual photon beams in $\approx$ 50 milliseconds with a high gamma pass rate of $97.72 \pm 1.93\%$ (2 mm, 2%). Estimating full VMAT dose distributions in 6-12 seconds, iDoTA achieves state-of-the-art performance with a $99.51 \pm 0.66\%$ (2 mm, 2%) gamma pass rate. The proposed model can reduce calculation times from few minutes to just a few seconds, massively speeding up current photon workflows.

While photon workflows partially mitigate the effect anatomical uncertainties with extra margins, the exact effectiveness of added margins on reducing the effects of organ motion is unknown. Most importantly, current proton robust optimization and robustness evaluation approaches only include setup and range uncertainties, basically completely disregarding the detrimental effects of anatomy changes during the treatment course. Accurate models of internal anatomy motion able to simulate dominant motion patterns could alleviate both issues. They could allow determining individualized margins for photon treatment courses, and including multiple anatomies in robust optimization and robustness evaluation for proton treatments. Traditionally, such anatomy models are based on principal component analysis (PCA) and are either patient-specific (requiring several scans per patient) or population-based, applying the same set of deformations to all patients. Chapter 4 presents a hybrid approach which,

based on population data, allows to predict patient-specific inter-fraction variations for an individual patient. This is achieved by a deep learning probabilistic framework that generates deformation vector fields (DVFs) warping a patient's planning CT into possible patient-specific anatomies. The presented daily anatomy model (DAM) uses few random variables capturing groups of correlated movements. Given a new planning CT, DAM estimates the joint distribution over the variables conditioned on the planning CT, with each sample from the distribution corresponding to a different deformation. Focusing on prostate cancer patients, DAM's performance is evaluated by quantifying the contour overlap between real and generated images (DICE score), and comparing the sampled and "ground truth" distributions of volume and center of mass changes. With DICE scores of 0.86 ± 0.05 and an average distance between prostate contours of 1.09 ± 0.93 mm using as few as 8 latent variables, DAM matches and improves the accuracy of previously published PCA-based models. The overlap between the simulated and ground truth distributions of center of mass and volume changes further indicates that DAM's sampled movements match the range and frequency of clinically observed daily changes on repeat CTs. Conditioned only on planning CT values and organ contours of a new patient without any pre-processing, DAM can accurately predict deformations seen during the treatment course, enabling robust treatment planning and robustness evaluation against inter-fraction anatomical changes.

Turning our attention to breathing anatomical variations occurring during radiation delivery, Chapter 5 presents a probabilistic framework to simultaneously generate and classify breathing signal time series. First, the chapter explores the potential of using the variational autoencoder (VAE) and adversarial autoencoder (AAE) algorithms to model breathing signals from individual patients. Second, an extended semi-supervised AAE algorithm is presented, allowing joint semi-supervised classification and generation of different types of signals within a single framework. To simplify the modeling task, a novel pre-processing and post-processing compressing method transforms the multi-dimensional time series into vectors containing only 8 time and position values per cycle, which are transformed back into high-resolution time series data through an additional neural network. The resulting models are able to generate highly realistic samples of breathing. By incorporating 4% and 12% of the labeled samples during training, the presented model outperforms other purely discriminative networks in classifying breathing baseline shift irregularities from a dataset completely different from the training set, with potential applications to generating class-specific breathing signals to be used for simulation of intra-fraction movements.

Based on the breathing signal models introduced in Chapter 5, Chapter 6 addresses the challenge of simulating breathing interplay effects in Intensity Modulated Proton Therapy (IMPT). Interplay effects arise from the interaction between target motion and the movement of the scanning beam, since breathing motion frequency and beam scan speed are of the same order of magnitude. Assessing the detrimental effect of interplay and the clinical robustness of various mitigation techniques requires statistical evaluation procedures that take into account the variability of breathing during dose delivery. Chapter 6 presents a model of intra-fraction respiratory motion based on breathing signals, while also assessing clinically relevant aspects related to the practical evaluation of interplay in IMPT such as how to model irregular breathing, how small breathing

changes affect the final dose distribution, and what is the statistical power (number of different scenarios) required for trustworthy quantification of interplay effects. First, two data-driven methodologies to generate artificial patient-specific breathing signals are compared: a simple sinusoidal model, and a precise probabilistic deep learning model yielding highly realistic samples of patient breathing. Second, the highly fluctuating relationship between interplay doses and breathing parameters is investigated, showing that small changes in breathing period can result in large local variations in the dose. The results indicate that using a limited number of samples to calculate interplay statistics introduces a bigger error than using simple sinusoidal models based on patient parameters or disregarding breathing hysteresis during the evaluation. Furthermore, the chapter illustrates the power of the presented statistical method by analyzing interplay robustness of 4DCT and Internal Target Volume (ITV) based treatment plans for a 8 lung cancer patients. As opposed to 4DCT plans, even 33 fraction ITV plans systematically fail to fulfill robustness requirements, indicating that the current use of ITV plans may be insufficient.

The work presented in this thesis addresses two of the main issues of radiotherapy workflows: the slow speed of dose calculation algorithms and the lack of accurate enough methods to simulate inter-fraction and intra-fraction anatomical changes. The developed algorithms solve both problems, offering millisecond dose prediction speed for proton pencil beams and full photon beams, and simulating realistic inter-fractional and intra-fractional anatomical variations matching the movements typically observed during the treatment course. As a conclusion, Chapter 7 discusses applications and possible future research indications. Offering the speed and anatomical modeling capabilities needed for robust planning, robustness evaluation and ultimately treatment plan adaptation, future research should focus on coupling the presented algorithms to existing clinical workflows, as well as validating their generalization performance in real clinical scenarios.

Samenvatting

Het doel van radiotherapie is het doden van tumorcellen door middel van straling. Bij de huidige protonen- en fotonentherapie wordt de patiënt bestraald tijdens een aantal dagelijkse sessies (ook wel fracties genoemd, gewoonlijk variërend van enkele tot dertig) met behulp van behandelingsplannen die gebaseerd zijn op een planning computertomografiescan (CT-scan) die enkele dagen voor het begin van de behandeling is gemaakt. Idealiter wordt de tumor volledig bestraald zonder de omringende kritieke organen te beschadigen, maar dit is fysiek onmogelijk door de aanwezigheid van verschillende bronnen van fouten en onzekerheden. Onnauwkeurigheid bij het in dezelfde positie plaatsen van de patiënt bij iedere fractie zorgt voor positioneringsfouten. Intra-fractie anatomische variaties veroorzaken dat sommige organen tijdens de toediening bewegen (bv. lever of long door de ademhaling). En nog belangrijker, door de tijd tussen de planning en de behandeling kunnen interfractie anatomische veranderingen nog grotere verschillen in de anatomie veroorzaken (bv. veranderingen in de vulling van het rectum of het krimpen van de tumor). Al deze onzekerheden beïnvloeden de wijze waarop de dosis in de tumor en de omringende structuren terechtkomt, en kunnen uiteindelijk de doeltreffendheid van de behandeling in gevaar brengen als er geen rekening mee wordt gehouden.

Met de huidige fotonen- en protonenplannen wordt getracht het nadelige effect van onzekerheden a-priori, d.w.z. tijdens de optimalisatie van het behandelplan, te beperken. Fotonbehandelingen maken gebruik van marge-uitbreidingen op het doelwit, waarbij direct wordt getracht grotere volumes te bestralen. Protonenworkflows optimaliseren de behandelplannen gelijktijdig over een reeks foutscenario's en evalueren vervolgens de robuustheid van het behandelplan tegen mogelijke onzekerheden in vele verschillende foutscenario's, wat uiteindelijk ook resulteert in bestraalde volumes die groter zijn dan het werkelijke klinische doel. Fotonmarges kunnen tot op zekere hoogte compenseren voor positioneringsfouten en anatomische variaties, maar de huidige robuuste optimalisatie- en robuustheidsevaluatiebenaderingen houden alleen rekening met positionerings- en proton-range fouten, deels vanwege het gebrek aan realistische anatomische bewegingsmodellen in de literatuur.

Om het effect van interfractie orgaan- en tumorbewegingen maximaal te beperken, zijn de volgende generatie online adaptieve workflows gericht op het verifiëren - en zo nodig corrigeren - van behandelplannen, binnen enkele minuten voorafgaand aan het uitvoeren van elke behandelingssessie. Dergelijke workflows zouden de nadelige effecten van onzekerheden verminderen, door zich aan te passen aan anatomische variaties, waardoor kleinere marges (voor fotonen) en lagere robuustheidsinstellingen in de optimalisatie (voor protonen) mogelijk worden, en minder gezond weefsel in gevaar wordt gebracht tijdens het bestralen van de tumor. Dergelijke adaptieve benaderingen stellen nog strengere eisen aan de snelheid van de dosisberekeningen dan robuuste optimalisatie en robuustheidsevaluatie, en vereisen (onder andere) algoritmen die de

dosistoediening in enkele seconden kunnen voorspellen, welke momenteel nog niet beschikbaar zijn binnen de radiotherapiegemeenschap.

Het doel van dit proefschrift is het oplossen van zowel het probleem van de trage dosisvoorspellingssnelheid als het ontbreken van anatomische modellen, met behulp van een combinatie van deep learning en probabilistische modelleerconcepten. De eerste helft van het proefschrift presenteert nieuwe methoden om de dosis van fotonen- of protonenbundels in enkele milliseconden te voorspellen, terwijl de tweede helft de simulatie van anatomische variaties tijdens en tussen fracties behandelt.

De voornaamste uitdaging van de huidige benaderingen voor dosisberekening is dat noch Monte Carlo (MC), noch analytische pencil-beam algoritmen (PBA) kunnen voldoen aan de strenge eisen van snelheid en nauwkeurigheid die nodig zijn voor aanpassing. Als mogelijke oplossing wordt in Hoofdstuk 2 een op deep learning gebaseerd algoritme voor dosisberekening (DoTA) gepresenteerd dat nauwkeurig de door protonenbundels afgegeven dosis voorspelt voor willekeurige energieën en patiëntgeometrieën. Gezien de voorwaarts verstrooiende aard van protonen wordt het 3D deeltjestransport benaderd als het modelleren van een opeenvolging van 2D CT-geometrieën ("plakken") loodrecht op de bundelrichting. DoTA combineert convolutionele neurale netwerken die ruimtelijke kenmerken extraheren (bv. weefsel- en dichtheidscontrasten) met een transformer self-attention backbone die informatie tussen de opeenvolging van geometrieplakken stuurt en een vector die de energie van de bundel weergeeft, en is getraind om MC-simulaties van protonbundels met weinig ruis te voorspellen. De DoTA voorspelt stralingsdoses in 5.9 ± 4.9 ms met zeer hoge Gamma-test slagingspercentages van $99.37 \pm 1.17\%$ (3 mm, 1%) en verbetert de analytische pencil beam-algoritmen aanzienlijk in precisie en snelheid. Met een MC-nauwkeurigheid die 100 keer sneller is dan PBA's en 10.000 keer sneller dan MC, berekent ons model volledige behandelingsplan-doses in 10-15 seconden, afhankelijk van het aantal bundels (800-2200 in onze plannen), met een Gamma slagingspercentage van $99.70 \pm 0.14\%$ (2 mm, 2%) voor 9 testpatiënten. DoTA is de nieuwste op deep learning-gebaseerde dosisberekeningsmethode die zelfs direct kan concurreren met de snelheid van commerciële multi GPU MC benaderingen.

In Hoofdstuk 3 wordt de DoTA-architectuur uitgebreid om in enkele milliseconden brede dosisverdelingen van fotonenbundels te voorspellen. Het voorgestelde improved Transformer Algorithm (iDoTA) brengt de 3D dosisverdeling in kaart van willekeurige patiëntgeometrieën en bundelinformatie (een 3D geprojecteerde vorm die het resultaat is van een eenvoudige ray tracing berekening). Door de 3D CT input en dosis output volumes te behandelen als een opeenvolging van 2D doorsnedes langs de richting van de fotonenbundel, lost iDoTA de dosisvoorspelling op als sequentiemodellering - vergelijkbaar met DoTA - met behulp van een reeks convoluties, residual connections en een transformator backbone. iDoTA voorspelt individuele fotonenbundels in ≈ 50 milliseconden met een hoog Gamma-test slagingspercentage van $97.72 \pm 1.93\%$ (2 mm, 2%). De iDoTA berekent de volledige VMAT-dosisverdeling in 6-12 seconden en met een state of the art Gamma slagingspercentage van $99.51 \pm 0.66\%$ (2 mm, 2%). Het voorgestelde model kan de berekeningstijden terugbrengen van enkele minuten tot slechts enkele seconden, waardoor de huidige fotonworkflows sterk worden versneld.

Hoewel fotonenworkflows het effect van anatomische onzekerheden gedeeltelijk

mitigeren met extra marges, is de exacte effectiviteit van de extra marges op het verminderen van de effecten van orgaanbeweging onbekend. Het belangrijkste is dat de huidige benaderingen voor robuuste optimalisatie en robuustheidsevaluatie van protonen alleen rekening houden met onzekerheden van de patiëntpositionering en de proton range, waardoor de schadelijke effecten van anatomische veranderingen tijdens de behandeling in feite volledig buiten beschouwing blijven. Nauwkeurige modellen van de interne anatomie die dominante bewegingspatronen kunnen simuleren, zouden beide problemen kunnen verkleinen. Zij zouden het mogelijk maken geïndividualiseerde marges te bepalen voor fotonenbehandelingen en meerdere anatomieën op te nemen in robuuste optimalisatie en robuustheidsevaluatie voor protonbehandelingen. Traditioneel zijn dergelijke anatomiemodellen gebaseerd op principal componentanalysis (PCA) en zijn ze ofwel patiëntspecifiek (waarvoor meerdere scans per patiënt nodig zijn) ofwel populatiegericht, waarbij dezelfde reeks vervormingen op alle patiënten wordt toegepast. In Hoofdstuk 4 wordt een hybride aanpak gepresenteerd waarmee op basis van populatiegegevens patiëntspecifieke interfractievariaties voor een individuele patiënt kunnen worden voorspeld. Dit wordt bereikt door een deep learning probabilistisch kader dat deformatievectorvelden (DVF's) genereert die de plannings-CT van een patiënt vervormen tot mogelijke patiëntspecifieke anatomieën. Het gepresenteerde dagelijkse anatomiemodel (DAM) gebruikt enkele willekeurige variabelen die groepen gecorreleerde bewegingen vastleggen. Gegeven een nieuwe planning CT, schat DAM de gezamenlijke verdeling over de variabelen geconditioneerd op de planning CT, waarbij elk monster van de verdeling overeenkomt met een andere vervorming. Bij prostaatkankerpatiënten worden de prestaties van DAM geëvalueerd door de overlapping van de contouren tussen de echte en de gegenereerde beelden (DICE-score) te kwantificeren en de bemonsterde verdelingen van volume- en massamiddelpuntveranderingen te vergelijken met de "ground truth-verdeling. Met DICE-scores van 0.86 ± 0.05 en een gemiddelde afstand tussen prostaatcontouren van 1.09 ± 0.93 mm met slechts 8 latente variabelen evenaart en verbetert DAM de nauwkeurigheid van eerder gepubliceerde PCA-gebaseerde modellen. De overlap tussen de gesimuleerde en ground truth distributies van veranderingen in het massamiddelpunt en volume geeft verder aan dat de in de steekproef opgenomen bewegingen van DAM overeenkomen met het bereik en de frequentie van klinisch waargenomen dagelijkse veranderingen op herhaalde CTs. Wanneer slechts geconditioneerd op plannings-CT-waarden en orgaancontouren van een nieuwe patiënt zonder enige voorbewerking, kan DAM nauwkeurig vervormingen voorspellen die tijdens het behandeltraject worden waargenomen, waardoor robuuste behandelplanning en robuustheidsevaluatie tegen interfractie anatomische veranderingen mogelijk worden.

Als we onze aandacht richten op anatomische ademhalingsvariaties tijdens bestraling, presenteert Hoofdstuk 5 een probabilistisch kader voor het gelijktijdig genereren en classificeren van tijdreeksen van ademhalingssignalen. Ten eerste onderzoekt het hoofdstuk de mogelijkheden van de variationele autoencoder (VAE) en adversariële autoencoder (AAE) om ademsignalen van individuele patiënten te modelleren. Ten tweede wordt een uitgebreid semi-supervised AAE-algoritme gepresenteerd, dat gezamenlijke semi-supervised classificatie en generatie van verschillende soorten signalen binnen één kader mogelijk maakt. Om de modelleringstaak te vereenvoudigen, wor-

den de multidimensionale tijdreeksen door een nieuwe methode voor voor- en nabewerking omgezet in vectoren met slechts 8 tijd- en positiewaarden per cyclus, die via een aanvullend neuraal netwerk weer worden omgezet in tijdreeksgegevens met hoge resolutie. De resulterende modellen kunnen zeer realistische ademhalingsvoorbeelden genereren. Door tijdens de training 4% en 12% van de gelabelde monsters op te nemen, presteert het gepresenteerde model beter dan andere zuiver discriminerende netwerken bij het classificeren van onregelmatigheden in de basislijnverschuiving van de ademhaling uit een dataset die volledig verschilt van de trainingsset, met mogelijke toepassingen voor het genereren van klassespecifieke ademsignalen die kunnen worden gebruikt voor de simulatie van bewegingen binnen de fractie.

Gebaseerd op de in Hoofdstuk 5 geïntroduceerde ademsignaalmodellen, wordt in Hoofdstuk 6 de uitdaging aangegaan om ademhalingseffecten bij intensiteit gemoduleerde protontherapie (IMPT) te simuleren. Interactie-effecten ontstaan door de interactie tussen de beweging van het doelwit en de beweging van de bundel, aangezien de frequentie van de ademhalingsbeweging en de scansnelheid van de bundel van dezelfde orde van grootte zijn. Om het schadelijke effect van interplay en de klinische robuustheid van verschillende mitigatietechnieken te beoordelen, zijn statistische evaluatieprocedures nodig die rekening houden met de variabiliteit van de ademhaling tijdens de toediening van de dosis. Hoofdstuk 6 presenteert een model van ademhalingsbeweging binnen de fractie op basis van ademsignalen, terwijl ook klinisch relevante aspecten met betrekking tot de praktische evaluatie van interplay in IMPT worden beoordeeld, zoals hoe onregelmatige ademhaling kan worden gemodelleerd, hoe kleine ademhalingsveranderingen de uiteindelijke dosisverdeling beïnvloeden en wat het onderscheidend vermogen (aantal verschillende scenario's) is die nodig is voor een betrouwbare kwantificering van interplay-effecten. Eerst worden twee gegevensgestuurde methodologieën voor het genereren van kunstmatige patiëntspecifieke ademsignalen vergeleken: een eenvoudig sinusoïdaal model en een nauwkeurig probabilistisch deep learning-model dat zeer realistische voorbeelden van de ademhaling van patiënten oplevert. Ten tweede wordt de sterk fluctuerende relatie tussen intervaldoses en ademhalingsparameters onderzocht, waaruit blijkt dat kleine veranderingen in de ademhalingsperiode kunnen leiden tot grote lokale variaties in de dosis. Uit de resultaten blijkt dat het gebruik van een beperkt aantal monsters voor de berekening van de interplay-statistieken een grotere fout oplevert dan het gebruik van eenvoudige sinusoïdale modellen op basis van patiëntparameters of het negeren van ademhalingshysterese tijdens de evaluatie. Verder illustreert het hoofdstuk de kracht van de gepresenteerde statistische methode door de interplay-robuustheid te vergelijken tussen 4DCT en Internal Target Volume (ITV) gebaseerde behandelplannen voor 8 longkankerpatiënten. In tegenstelling tot 4DCT-plannen voldoen zelfs 33-fractie ITV-plannen systematisch niet aan de robuustheidsvereisten, wat erop wijst dat het huidige gebruik van ITV-plannen mogelijk ontoereikend is.

Het in dit proefschrift gepresenteerde werk pakt twee van de belangrijkste problemen van radiotherapie workflows aan: de trage snelheid van dosisberekeningsalgoritmen en het gebrek aan voldoende nauwkeurige methoden om interfractie en intrafractie anatomische veranderingen te simuleren. De ontwikkelde algoritmen lossen beide problemen op en bieden een snelheid van milliseconden voor de voorspelling van de

dosis voor protonen- en fotonenbundels, en simuleren realistische interfractionele en intrafractionele anatomische variaties die overeenkomen met de bewegingen die tijdens de behandeling worden waargenomen. Ter afsluiting worden in Hoofdstuk 7 toepassingen en mogelijke toekomstige onderzoeksindicaties besproken. Met de snelheid en de mogelijkheden voor anatomische modellering die nodig zijn voor robuuste planning, robuustheidsevaluatie en uiteindelijk aanpassing van het behandelplan, moet toekomstig onderzoek zich richten op het koppelen van de gepresenteerde algoritmen aan bestaande klinische workflows, en op het valideren van hun generalisatieprestaties in echte klinische scenario's.

(Translation provided by Marc van den Berg, Mischa Hoogeman and Dennis Schaart.)

1

Introduction

Despite significant research efforts, cancer remains responsible for more than 10 million deaths in 2020 worldwide (Sung et al., 2021). With more than 50% of the patients receiving radiation treatments, radiotherapy is at the forefront of current standard of care, playing an important role in improving societal health. Sophisticated computational methods and particle transport simulations have been key to this success (Bernier et al., 2004), enabling highly personalized treatments.

The goal of radiation therapy is to eradicate cancerous tissue while minimizing damage to surrounding healthy organs and structures. While traveling through the patient, radiation deposits energy in the tissue via atomic and nuclear interactions. This energy transfer is quantified as *dose* and is expressed as energy (Joules) absorbed per unit mass, with units of Gray (Gy). Dose translates into radiation damage, which mainly occurs via single or double strand DNA breaks, causing cell death.

Radiotherapy primarily uses photons or protons to irradiate the target, with most patients receiving photon treatments, but proton therapy spreading quickly due to protons' finite range and significantly better ability to focus dose on tumors (Lundkvist et al., 2005). Photons and protons mainly differ in the way they deposit energy along their path. After a short build-up dose region near beam entrance called the 'skin-sparing effect', the energy deposited by photons decreases exponentially with increasing depth. Since the dose delivered by a photon beam is higher near the patient's surface, photon treatments require many beam angles to create high dose regions inside the patient by overlapping low doses from many different beams. As a consequence of using many angles (referred to as *gantry angles* in this thesis) with different collimator shapes that conform the beam laterally to the tumor shape, photon treatments typically result in large irradiated volumes within the patient. Conversely, protons can deliver high doses to deep tumors using only few irradiation angles, since most of the dose deposited by proton beams occurs at a known depth (*range*) in a region referred to as the *Bragg peak*. The location of this high dose Bragg peak can be estimated with reasonable accuracy given the beam energy and the material composition of the patient. This finite range property is used in proton treatments to deliver most of the dose to the tumor, sparing

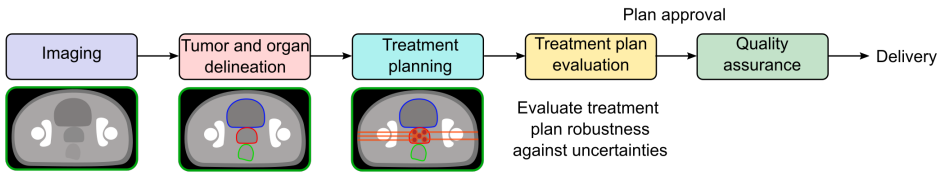


Figure 1.1: **Radiotherapy workflow**. A radiotherapy treatment typically begins with the acquisition of the patient's anatomical information about tissues and structures via imaging (typically in the form of a 3D CT scan), followed by the delineation of the tumor and relevant surrounding organs. Once the treatment irradiation modality has been selected, the image and contours are then used as an input to a treatment planning step, resulting in a list of beam angles, intensities and energies used to irradiate the target. A subsequent treatment plan evaluation step determines if the plan is robust against errors, such as the ones in positioning the patient on the couch, requiring the calculation of the radiation dose delivered in many different evaluation error scenarios.

the surrounding organs and achieving the same tumor dose as in photon treatment with less OAR dose. As a result, an increasingly large number of facilities are adopting intensity modulated proton therapy (IMPT) treatments, using few gantry angles composed of thousands of narrow proton pencil beams with different energies aiming at the target.

1.1. Radiation therapy workflow

As illustrated in Figure 1.1, photon and proton radiotherapy treatments usually follow a 5-step procedure.

1.1.1. Acquiring anatomical information

First, high quality anatomical information is acquired, typically as computed tomography (CT) images (Pereira et al., 2014). CTs are a set of parallel slices showing the voxelized patient geometry (electron density specifically) with units of Hounsfield Unit (HU), representing a 3D image reconstructed from the tomographic projections of image acquisition. Second, target tumors and organs at risk (OARs) to protect are delineated on the anatomy captured by the CT. Other imaging modalities such as positron emission tomography (PET) or magnetic resonance (MR) imaging can also be used for extra assistance.

1.1.2. Treatment planning

Third, a treatment plan is obtained, containing the intensities, energies (in the case of protons), multi-leaf collimator (MLC) shapes (in the case of photons) and angles of the beams used for irradiation. Treatment planning is a complex and computationally expensive task, requiring solving large scale multi-criteria optimization problems with clinical dose constraints and objectives, where the intensities, energies/MLC shapes and angles of all beams are individually tuned until the 'clinically best' total dose distribution is found, satisfying dose constraints on the tumor and surrounding organs (Hussein et al., 2018; Meyer et al., 2018). The most commonly used objective function is the squared difference between clinical prescribed doses and the delivered dose. To

minimize such objectives while meeting the hard dose constraints, treatment planning typically involves exploring the many degrees of freedom available until finding a combination that delivers the required dose to the tumor while minimizing the dose deposited in healthy tissue.

Treatment uncertainties Based on the treatment plan, radiation is delivered during up to 45 different daily sessions (referred to as *fractions*) to allow for healthy tissue to recover from radiation damage, since the healthy cells can repair themselves faster than the cancerous ones. Ideally, a successful treatment would completely eliminate the tumor without damaging the surrounding healthy tissue, but this is physically impossible partially due to errors and uncertainties. Though the same treatment plan is used on each day of the treatment course, the actual delivered doses differ from the planned dose due to these uncertainties, which can degrade treatment effectiveness and the capability to fully eradicate the tumor. Some of these include errors in precisely positioning the patient in the exact same position in every fraction (Liebl et al., 2014; Trofimov et al., 2011), or errors in calibrating the delivery machine. Others include uncertainties in correctly estimating the range of the proton beams — due to the uncertainty in predicting stopping power values (average energy loss of the particle per unit path length) from the HU values in the CT — (Lomax, 2008a, 2008b; Paganetti, 2012), or errors in the delineation of the tumor and structures. The main focus of this work however relates to errors that occur due to the changing anatomy of the patient during or between treatment sessions.

- **Inter-fraction anatomical changes.** During the several weeks that typical radiotherapy treatments last, the internal position of organs and structures continuously changes. As a result, the patient's anatomy captured in the CT scan used for treatment planning can significantly differ from the real anatomy observed for the same patient in a different fraction. Previous studies have demonstrated that such anatomical deformations are one of the main sources of error for certain types of cancer treatments such as prostate (van Herk et al., 2002), where the main anatomical variations include random, daily changes in bladder or rectum fillings, and gradual changes over the treatment course such as tumor shrinkage. Delivering dose based on the original treatment plan to such changing anatomies may result in inaccuracies that could affect treatment success. For example, the high dose regions being partially delivered to healthy tissue instead of entirely covering the target volume could severely under-dose the tumor and increase the chance of complications in the surrounding OARs.

- **Intra-fraction breathing interplay effects.** Real-time tumor and organ movements can also affect the success of treatments. When the tumor is located in an area close to the lung (e.g., lung, esophagus or liver cancer treatments) intra-fraction anatomical variations occur due to breathing during treatment delivery. The continuous movement of internal structures during irradiation affects the final dose distribution, especially in proton treatments using pencil beam scanning. The resulting *breathing interplay effects* — caused by the breathing movement and the scanning of the pencil beam moving at a similar speed — are

detrimental, as during the few minutes in which each fraction is delivered, the continuous movement degrades the final dose distribution (Bert and Durante, 2011; Bert et al., 2008; Lambert et al., 2005), effectively causing some dose to be delivered to healthy structures, or even dose overlap.

Incorporating uncertainties to treatment planning In current clinical practice, the detrimental effect of setup, range (and partially) anatomical uncertainties is a-priori mitigated during treatment planning. In conventional photon radiotherapy, positioning and delivery uncertainties can be accounted for by margins aiming to deliver the required tumor doses at the cost of extra dose deposited in healthy tissues. These margins are applied to extend the clinical target volume (CTV) — a volume containing tumor cells and covering for possible spread — into planning target volume (PTV), supposedly robust against uncertainties. The most known are the Stroom (Stroom and Heijmen, 2002) and van Herk (van Herk et al., 2000) formulas, which calculate the margin based on the standard deviation of random and systematic patient setup errors. While most effective for positioning uncertainties in photon treatments, such PTV extensions also account (to some extent) for intra-fraction and inter-fraction changes. Additionally, to tackle the detrimental effects of breathing, an internal target volume (ITV) can be made combining CTVs at different points of the breathing cycle, e.g., mid-ventilation, exhale and inhale (Shih et al., 2004). However, since proton beams are much more susceptible to uncertainties than photons, margin extensions have been demonstrated to be sub-optimal in proton treatments (Liu et al., 2013), especially given the unaccounted beam range calculation errors (Unkelbach et al., 2018).

Instead of margins, setup and range errors (and in principle other types of delivery uncertainties) can be better taken into account by means of *robust optimization* (Chu et al., 2005; Liu et al., 2012), which results in more resilient plans (Dijk et al., 2016). In particular, proton plans are typically optimized solving a multi-criteria mini-max optimization problem, so that the clinical dose constraints and objectives are always met under a certain set of worst case scenarios. These typically include the nominal scenario (without errors), range errors, and different positioning error scenarios corresponding to rigid shifts along each principal direction. For the elaboration of the treatment plan, the magnitude of such setup and range errors in these scenarios is controlled via the *setup robustness* (SR) setting (in millimeters, most often 6 scenarios corresponding to the positive and negative shifts in the x, y and z directions) and the *range robustness* (RR) setting (in percentage, 2 scenarios for range over/underestimation). As a result, the treatment plan is simultaneously optimized for 9 different scenarios (1 nominal, 6 for SR and 2 for RR), which may lead to either overly conservative or not sufficiently robust plans if used improperly.

Robust optimization can be extended to account for other types of uncertainties apart from setup and range errors, by including the anatomies in different CT scans in the optimization. Such approaches have been proposed to mitigate intra-fraction breathing movements, by including CTs from different phases of the breathing cycle (Bernatowicz et al., 2017; Engelsman et al., 2006). In principle, inter-fraction anatomical variations of organs and target could be similarly incorporated in the optimization if such representative CTs were available beforehand.

1.1.3. Evaluation and quality assurance

Once a treatment plan has been obtained, and typically before approval, a robustness evaluation step allows assessing treatment plan robustness against uncertainties, verifying that the plan meets the clinical constraints. Current proton evaluation protocols only consider a set of worst-case setup and range errors, comparing statistics of the clinical quantities of interest across many simulated dose distributions delivered in the presence such error scenarios, e.g., calculating the minimum dose delivered in 90% of the scenarios. As an example, the Dutch proton therapy group (DUPROTON) protocol in the Netherlands evaluates plan robustness based on 2 artificial dose distributions with the minimum and maximum dose values per voxel across 28 evaluation scenarios with different combinations of geometrical and range errors, which has been shown to lead to overly conservative treatments (Rojo-Santiago et al., 2021).

Last, after approval of the radiation oncologist and medical physicist, a quality assurance step verifies that the treatment plan can be correctly delivered, usually via an additional measurement or a dose calculation independent from the treatment planning system.

1.2. Next steps in improving photon and proton treatments

Overall, margins, robust planning and evaluation approaches are a class of solutions minimally modifying the radiotherapy workflow to produce treatment plans that are inherently more robust against possible uncertainties. The resulting plans typically compromise healthy tissue sparing in favor of delivering the clinically desired doses to the tumor. Current research efforts aim at optimally balancing the successful eradication of the tumor with a lower dose delivered to healthy tissue. In principle, photon margins can be additionally extended on a patient-specific or site-specific basis to cover for anatomy variations, ideally using anatomical information of the patient geometries that are likely to be observed during radiation delivery or the treatment course.

Inspired by the photon margin extension formulas, similar concepts have been proposed for IMPT treatments in the form of robustness recipes (van der Voort et al., 2016). The proposed proton robustness recipes suggest the robustness settings RR and SR to be used during robust optimization for the treatment plan to achieve a specified CTV coverage for a certain percentage of the population. Nevertheless, these recipes have also their limitations, since they cannot sufficiently handle individual patient variations or extreme cases, and they do not directly account for anatomical changes.

Robustness recipes enable to fill in the gap between robust optimization and probabilistic optimization, the latter being one of the main desired milestones in the IMPT field. Instead of jointly optimizing over a small discrete set of equally important error scenarios as in robust optimization, probabilistic optimization minimizes probabilistic formulations of the clinical objectives and constraints. As an example, previous work proposes minimizing the expected value of the clinical objective (Unkelbach et al., 2009), e.g., the expected value of the squared dose difference between delivered and prescribed doses across all sets of error scenarios. Assuming that the infinite number of combinations of errors could be accounted for, the expectation of the objective can be interpreted as a weighted sum of objectives from individual error scenarios, where the

weight corresponds to the relative probability of occurrence of each scenario. Consequently, the main limitation of such probabilistic optimization approaches is the computational power needed to calculate statistics of the clinical quantities of interest, as well as their derivatives with respect to the beam intensities, especially since they have to be continuously estimated in every iteration solving the optimization problem.

Margins and robustness enable counterbalancing the effects of uncertainties by irradiating larger areas of healthy tissue than necessary. The most straightforward way of decreasing these volumes is to mitigate the uncertainties against which margins and robustness are used. Thus, to further reduce the target margins and robustness settings needed during planning, treatment plans should ideally be adapted on a daily basis based on the anatomical differences with respect to the original planning settings.

Online and real-time adaptation Online adaptation aims at adjusting the original treatment plan right before delivery to irradiate the correct dose in the varying anatomy of the patient recorded in that same fraction (Paganetti et al., 2021). Based on the conventional radiotherapy workflow, the modified online adaptive workflow (shown in Figure 1.2) is composed of 5 steps. First, the patient's anatomy is acquired once the patient has been positioned on the treatment couch. Second, the tumor and organ structures are delineated on the resulting CT from the quick imaging step. This can be achieved using automated segmentation software, or transferring the original structures from the planning CT, for which it is necessary to determine the image correspondence via a registration step. Third, the need for adaptation is assessed by quickly estimating the dose delivered by the original treatment plan on the new patient anatomy. Fourth, if needed, the plan has to be optimized on the new geometry, typically based on a re-optimization step. The level of detail of such optimization step can greatly vary, resulting in two variants which differ in the amount of information inherited from the original plan. The first group includes *re-planning* approaches which optimize the initial plan from the start, either with the same objectives and constraints or using new ones adapted to the new geometry (Matter et al., 2019). The second group encapsulates *re-optimization* or *dose restoration* variants that fine-tune the initial plan to achieve the original dose/plan quality in the new anatomy (Bernatowicz et al., 2018; Botas et al., 2018; Jagt et al., 2017, 2018). While the optimization would ideally still be robust or probabilistic to increase robustness of the plans against residual positioning and range uncertainties, daily adaptation allows for a reduction of the RR and SR settings (or the margins, in the case of photons). Finally, the resulting plan follows an automated quality assurance check, confirming that it results in better fraction dose delivery than with the original settings.

While online adaptive workflows circumvent the need to account for inter-fraction anatomical variations, the resulting plans are susceptible to intra-fraction motion, still requiring extra margins or some form of robust planning. To maximally reduce dose to healthy tissue and remove the detrimental effects of uncertainties, real-time adaptation aims at adapting treatments during delivery.

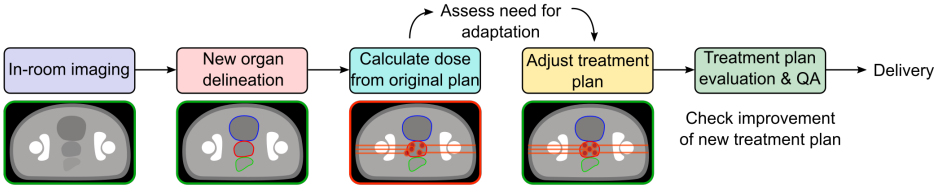


Figure 1.2: **Online adaptive workflow**. After quickly acquiring anatomical information via in-room imaging and automatically delineating the tumor and organ contours, a dose calculation estimates the radiation dose delivered with the original treatment plan (i.e., using the same list of energies, angles and intensities obtained in the first or previous treatment sessions). If the resulting dose distribution does not meet the clinical constraints, the treatment plan can be adjusted, e.g., by quickly re-optimizing the beam intensities. A final evaluation step confirms the superiority of the new treatment plan before delivery.

1.3. Current challenges

Fast and accurate particle transport algorithms are crucial for all the steps of a radiation therapy workflow. CT image reconstruction relies on simulating photon interactions with tissues and detectors; plan optimization requires the spatial dose distribution (typically in more than 1 million voxels) from each available proton or photon beamlet (in the thousands); while for plan evaluation the dose must be calculated for many different geometries (repeatedly recalculating the patient dose from fixed beamlets under uncertainties). Current implementation of robust planning and robustness evaluation methods require simulating the dose distribution in tens of error scenarios, which can be computationally expensive. This is especially the case if including different sources of error with respect to the clinical practice, e.g., extending current protocols to account for anatomical variations, potentially resulting in hundreds of scenarios with different error combinations. With a larger set of uncertainties, the number of scenarios with different combinations of errors grows exponentially, a problem known as the curse of dimensionality. The same applies to probabilistic planning, typically involving thousands of dose calculations for the formulation of the probabilistic metrics, objectives and constraints. With the current dose calculation algorithms and lacking ways to generate scenarios with realistic intra-fraction and inter-fraction variations, extending robustness planning and evaluation approaches to cover anatomical uncertainties remains a challenge.

Algorithmic speed also remains a challenge in daily adaptive workflows. Since the time between imaging and delivery must be reduced as much as possible, online adaptive treatments critically rely on the speed of each of their individual steps, thus requiring very fast dose calculation, image registration, image segmentation, image acquisition and dose optimization. The speed requirements are most acute for next generation real-time adaptive treatments promising ultimate precision with fewest side effects by correcting treatments during irradiation, e.g., to account for anatomical changes due to breathing, coughs or intestinal movements. The optimistic case of real-time adaptation of treatment plans allows to deliver correct doses in-vivo, eliminating residual errors and thus the need of robust treatment planning and evaluation approaches. To finally become reality, such adaptive treatments require algorithms that deliver accurate dose distributions in millisecond speed.

Fast dose calculation algorithms and probabilistic models that quantify anatomical deformations using few random variables are two of the missing pieces for robust treatment planning, robustness evaluation, and ultimately adaptation of treatment plans. Fast methods to predict dose distributions can directly impact clinical practice by allowing to include more scenarios in the optimization and evaluation of the treatment plan. Most critically, a fast dose calculation model can ultimately provide the speed necessary to perform plan adaptation within the ≈ 2 minutes prior to fraction delivery. Alternatively, anatomical models can be used to guarantee a certain coverage given the most likely intra-fraction anatomical variations via extended robustness recipes, probabilistic objectives or robust optimization and evaluation of treatment plans across a larger set of error scenarios.

1.4. Contents of this dissertation

Recently, deep learning algorithms have achieved state-of-the-art performance in many tasks such image processing, natural language processing or sequence modeling, mostly due to learning how to extract highly non-linear and relevant features for the task at hand from large datasets. Optimized for graphics processing unit (GPU) hardware, deep learning algorithms have the potential to massively speed up different steps of the radiotherapy workflow, from dose calculation to image registration. Popular open-source libraries such as Tensorflow (Abadi et al., 2015) or Pytorch (Paszke et al., 2019) enable development and deployment of deep learning models to clinical applications using generic class functions and automated differentiation, providing all the GPU speed benefits without extensive hardware optimization.

This thesis marries the application of deep learning and probabilistic models, providing fast particle transport algorithms and methods to quantify and simulate anatomical movements during and between treatment sessions. The presented tools could in principle be applied to evaluate treatment plan robustness against inter-fraction and intra-fraction anatomical variations, ultimately paving the way for online adaptation and probabilistic treatment planning. The rest of the thesis is organized as follows.

Chapter 2 presents a deep learning-based method to learn particle transport physics from data. Ideally these calculations should be quick and precise, but current analytical pencil beam algorithms (PBA) and stochastic Monte Carlo (MC) dose calculation tools offer a trade-off. PBA yields results without the computational burden of MC engines, but its accuracy is severely compromised in highly heterogeneous or complex geometries, making slow and clinically rarely affordable MC approaches necessary. As a fast and accurate alternative to physics-based models, the chapter presents a deep learning dose transformer algorithm (DoTA), applied to calculate proton pencil beam doses only from CT and beam energy data in few milliseconds. The chapter describes the architecture of the DoTA model and the details of its transformer backbone, together with the dataset, the model training procedure and the evaluation experiments comparing performance to previous deep learning and physics-based algorithms.

Chapter 3 extends the DoTA proton dose calculation algorithm to predict broad photon beam dose distributions. Modern photon delivery modalities such as volumetric modulated arc therapy (VMAT) use hundreds of photon beams conforming the final dose to the tumor as much as possible via continuously changing the MLC shape,

i.e., a dynamic radiation device that can adjust its opening to block undesired parts of the beam. Calculating the dose for such treatments is usually time consuming, given the larger amount of gantry angles considered. To speed up photon dose prediction, an improved DoTA architecture (iDoTA) is presented, which, besides the CT, takes input information about the beam shape and relative position between isocenter and beam source via a 3D projection of the MLC shape. After describing the new architecture, dataset and training procedure, iDoTA is compared to other state-of-the-art deep learning models.

In Chapter 4, we turn our attention to modeling organ movements observed during the course of a radiotherapy treatment. First, a probabilistic daily anatomy model (DAM) is presented, generating deformation fields that warp the planning CT recorded at the beginning of the treatment into plausible repeat CTs. Second, DAM's architecture is described, combining a probabilistic variational framework powered by deep learning models to reduce deformation fields to few random variables with known probability distribution. Third, the training details, dataset and experiments are described. The chapter concludes evaluating DAM's generative capability to generate repeat CTs as observed in the clinic.

Chapter 5 introduces a deep learning-based probabilistic framework to simultaneously classify and generate breathing signals describing tumor motion during radiation delivery. The novelty of this chapter is threefold. First, it explores the use of two previously published models to compress breathing signals into few random variables with known probability distribution, which can be sampled to generate new realistic realizations. Second, a novel joint classification-generative framework is introduced, which allows classifying and subsequently generating signals from a specific class, i.e., with certain traits. Third, a breathing signal pre-processing and post-processing algorithm is presented, which transforms back and forth between 3D motion signals and a vector of position and time stamps. The chapter concludes by demonstrating the ability of the presented probabilistic framework to accurately generate and classify breathing signals with baseline shifts, i.e., upwards or downwards gradual changes in breathing depth.

Chapter 6 applies the results of Chapter 5 to simulate dynamic dose delivery in lung cancer patients. Based on the generated breathing signals and a 4D-CT — a set of 3D CTs capturing anatomical changes over a breathing period composed of different breathing phases — a method to simulate interplay effects is presented, based on delivering pencil beams to their corresponding breathing phase. The presented interplay calculation tool is applied to evaluate the robustness of treatment plans against breathing motion, as well as to determine how to accurately simulate interplay effects in proton pencil beam scanning treatments.

To conclude this thesis, Chapter 7 summarizes the main findings and prospective applications of the presented models, together with some recommendations for future research.

2

Millisecond proton dose calculation with Monte Carlo accuracy

2.1. Introduction

Radiotherapy treatments intimately rely on accurate particle transport calculations. In computed tomography (CT) image acquisition (Pereira et al., 2014) simulations of the interaction between photons, tissues and detectors are used to obtain a detailed 3D image of the patient anatomy, which can be delineated to localize target structures and organs-at-risk. Modern intensity modulated treatments (Hussein et al., 2018; Meyer et al., 2018) require particle transport to compute the spatial distribution of physical dose delivered by thousands of individual electron, photon, proton or other heavy ion beamlets (aimed at the patient from a few different gantry angles), based on which the beamlet intensities can be optimized. Treatment plans – especially sensitive proton and ion treatments – must also be repeatedly evaluated under uncertainties (e.g., setup and range errors, tumor motion or complex anatomical changes) to ensure sufficient plan robustness, requiring recalculating the dose distribution in many different scenarios (Perkó et al., 2016; Rojo-Santiago et al., 2021; van der Voort et al., 2016). With radiotherapy practice steadily moving towards adaptive treatments, accurate, fast and general purpose dose (and particle transport) calculations represent an increasingly pressing, currently unmet need in most clinical settings.

Current physics-based dose calculation tools – by and large falling into 2 categories: analytical pencil beam algorithms (PBAs) (L. Hong et al., 1996; Schaffner et al., 1999) and Monte Carlo (MC) simulations – offer a trade-off between speed and precision. While PBAs yield results without the computational burden of MC engines, their accuracy is severely compromised in highly heterogeneous or complex geometries, making

The contents of this chapter have been published as a journal paper in *Physics in Medicine & Biology* 67 105006 (2022), (Pastor-Serrano and Perkó, 2022a).

slow and clinically often not affordable MC approaches necessary (Grassberger et al., 2014; Saini et al., 2017; Schuemann et al., 2015; Teoh et al., 2020). The problem is most acute for online (and ultimately real-time) adaptive proton therapy aiming at treatment correction prior to (or even during) delivery to account for inter-fractional anatomical changes, motion due to breathing, coughs or intestinal movements. To become reality, such adaptive treatments require algorithms yielding MC accuracy with sub-second speed.

Reducing dose calculation times is an active area of research, with most works focusing on improving existing physics-based algorithms or developing deep learning frameworks. Several studies benefit from the parallelization capabilities of graphics processing units (GPUs) to massively speed up MC simulations, reducing calculations times down to the range of few seconds (Fracchiolla et al., 2021; Wan Chan Tseung et al., 2015) to minutes (Gajewski et al., 2021; J. Ma et al., 2014; Pepin et al., 2018; Qin et al., 2016; Y. Wang et al., 2016), with simulation speeds up to 10^7 protons/s. Deep learning methods have also improved dose calculation times in several steps of the radiotherapy workflow (Meyer et al., 2018), although usually paying the price of limited versatility and generalization capabilities. Some initial studies apply variants of U-net (Ronneberger et al., 2015) and Generative Adversarial Networks (Goodfellow et al., 2014) to aid treatment planning by approximating dose distributions from 'optimal' plans in very specific scenarios based on historical data. As input to these convolutional architectures, most works use organ and tumor masks (Chen et al., 2019; Fan et al., 2019; Kajikawa et al., 2019; Nguyen, Long, et al., 2019), CT images (Kearney et al., 2018) or manually encoded beam information (Barragán-Montero et al., 2019; Nguyen, Jia, et al., 2019) to directly predict full dose distributions, except for few papers predicting the required beam intensities needed to deliver such doses (Lee et al., 2019; W. Wang et al., 2020).

Regarding pure dose calculation, practically all deep learning applications rely on using computationally cheaper physics simulations as additional input apart from CTs. For photons, most works predict low noise MC dose distributions from high noise MC doses (Bai et al., 2021; Neph et al., 2021; Peng, Shan, Liu, Pei, Wang, and Xu, 2019; Peng, Shan, Liu, Pei, Zhou, et al., 2019) or simple analytical particle transport calculations (Dong and Xing, 2020; Xing, Zhang, et al., 2020), with some approaches also utilizing additional manually encoded beam/physics information such as fluence maps (Fan et al., 2020; Kontaxis et al., 2020; Tsekas et al., 2021; Xing, Nguyen, et al., 2020; J. Zhu et al., 2020). For protons, only few works (Javaid et al., 2021; Nomura et al., 2020; C. Wu et al., 2021) compute proton dose distributions via deep learning, using cheap physics models (noisy MC and PBA) or pre-calculated Bragg peak maps as input. While providing significant speed-up compared to pure physics-based algorithms, some even reaching sub-second speeds, all these works depend on secondary physics models to produce their output or are trained to predict only full plan or field doses for specific treatment sites. As a result, these methods do not qualify as generic dose algorithms and do not generalize to other steps of the radiotherapy workflow outside their original scope, e.g., to different plan or field configurations, treatment sites, or applications needing the individual dose distribution from each beamlet separately (such as treatment adaptation).

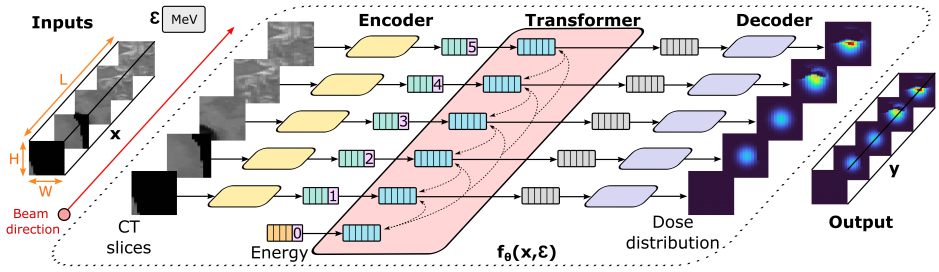


Figure 2.1: **Dose transformer algorithm (DoTA).** A data-driven model learns a mapping $\boldsymbol{y} = f_{\boldsymbol{\theta}}(\boldsymbol{x}, \varepsilon)$ between input CT cubes $\boldsymbol{x}$ and energies ε and output dose distributions $\boldsymbol{y}$. CT and dose distribution 3D volumes are both treated as a sequence of 2D slices in the beam's eye view. An encoder and a decoder individually transform each 2D slice into a feature vector and vice versa, whereas a transformer backbone routes information between different vectors along beam depth.

Instead, this chapter focuses on learning particle transport physics to substitute proton dose engines, providing millisecond speed and high accuracy, and is in principle applicable to all radiotherapy steps requiring dose calculations (e.g., dose-influence matrix calculation, dose accumulation, robustness evaluation). The proposed approach builds upon a previous study (Neishabouri et al., 2021) using long short-term memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) to sequentially calculate proton pencil beam dose distributions from relative stopping power slices in sub-second times, but with the major disadvantage of requiring a separate model per beam energy. As shown in Figure 2.1, proton transport is modeled a sequence of 2D geometry slices in the beam's eye view, introducing an attention-based transformer backbone (Vaswani et al., 2017) that dynamically routes information between elements of the sequence along beam depth. The presented Dose Transformer Algorithm (DoTA) – able to learn the physics of energy dependence in proton transport via a single model – can predict low noise MC proton pencil beam dose distributions purely from beamlet energy and CT data in ≈ 5 ms. Based on the presented experiments and available literature data, in terms of accuracy and overall speed DoTA significantly outperforms pencil beam algorithms and all other deep learning approaches (e.g., LSTM models (Neishabouri et al., 2021) and 'denoising' networks (Javaid et al., 2021; Nomura et al., 2020; C. Wu et al., 2021)), representing the current state-of-the-art in data-driven proton dose calculations and directly competing with (and even improving on) GPU Monte Carlo approaches.

2.2. Dose prediction via transformers

The problem of dose calculation is common to many steps of the radiotherapy workflow and ultimately involves estimating the spatial distribution of physical dose from thousands of pencil beams. A generic deep learning dose engine must be capable of calculating 3D dose distributions for arbitrary patient geometries purely from a list of beam directions and energies for a given beam model, without being conditioned on the type of treatment or task being solved. Therefore, the objective is to accurately predict dose distributions $\boldsymbol{y}$ from individual proton beamlets in sub-second speed, given

patient geometries $\boldsymbol{x}$ and beam energies ε. The proposed DoTA is a parametric model that implicitly captures particle transport physics from data and learns the function $\boldsymbol{y} = f_{\boldsymbol{\theta}}(\boldsymbol{x}, \varepsilon)$ via a series of artificial neural networks with parameters $\boldsymbol{\theta}$.

In particular, DoTA learns a mapping between a 3D CT input voxel grid $\boldsymbol{x} \in \mathbb{R}^{L\times H\times W}$ and output dose distribution $\boldsymbol{y} \in \mathbb{R}^{L\times H\times W}$ conditioned on the energy $\varepsilon \in \mathbb{R}^{+}$, where L is the depth (in the direction of beam propagation), H is the height and W is the width of the grid. While traditional physics-based calculation tools process the entire geometry, DoTA's input CTs are cropped and interpolated to the reduced sub-volume seen by protons as they travel through the patient, with a fixed 2 mm × 2 mm × 2 mm resolution and $L \times H \times W$ size. Framing proton transport as sequence modeling, DoTA processes the input volume as a series of L 2D slices in the forward beam direction. Ideally, the exchange of information between the different elements in the sequence should be dynamic, i.e, the contribution or impact of each 2D slice to the sequence depends on both its position and material composition. Unlike other types of artificial neural networks, the transformer architecture (Vaswani et al., 2017) — and specifically the self-attention mechanism — is notably well suited for this.

2.2.1. Transformer and self-attention

DoTA's backbone is the transformer, based on self-attention (SA) (Vaswani et al., 2017). Though originally introduced for sequential modeling applications in natural language processing such as machine translation, transformers have recently achieved state-of-the-art performance across a wide variety of tasks, with large language (Brown et al., 2020; Devlin et al., 2019) or computer vision (D'Ascoli et al., 2021; Dosovitskiy et al., 2020; Ramachandran et al., 2019; Touvron et al., 2020) models replacing and outperforming recurrent or convolutional architectures. One of the main reasons behind the success of attention-based models is the ability to model interactions between a large sequence of elements without needing an internal memory state. Powered by the SA mechanism, transformers transform each sequence element based on the information it selectively gathers from other members of the sequence based on its content or position.

For modeling the sequentiality in proton transport physics, the advantage of transformers with respect to LSTM frameworks is two-fold. First, every element can directly access information at any point in the sequence without requiring an internal hidden state, which is crucial to include beam energy dependence. The SA routing of information is different for every element, allowing each geometry slice to be independently transformed based on the information it selectively gathers from other slices in the sequence. Second, transformers allow manually encoding the mostly forward scattering nature of proton transport by restricting interaction to only previous slices via causal attention.

Self-attention Given a sequence $\boldsymbol{h} \in \mathbb{R}^{L\times D}$ with L tokens of dimension D, the SA mechanism is based on the interaction between a series of queries $\boldsymbol{Q} \in \mathbb{R}^{L\times D}$, keys $\boldsymbol{K} \in \mathbb{R}^{L\times D}$, and values $\boldsymbol{V} \in \mathbb{R}^{L\times D}$ obtained through a learned linear transformation of the input tokens with weights $\boldsymbol{W}_{QKV} \in \mathbb{R}^{D\times 3D}$ as

$$[\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}] = \boldsymbol{h}\boldsymbol{W}_{QKV}. \tag{2.1}$$

Each token is transformed into a query, key and value vector. Intuitively, for an i^{th} token $\boldsymbol{h}_i \in \mathbb{R}^{1\times D}$, the query $\boldsymbol{q}_i \in \mathbb{R}^{1\times D}$ represents the information to be gathered from other elements of the sequence, while the key $\boldsymbol{k}_i \in \mathbb{R}^{1\times D}$ contains token's information to be shared with other sequence members. The token $\boldsymbol{h}_i$ is then transformed into $\boldsymbol{h}'_i$ via a weighted sum of all values in the sequence $\boldsymbol{v}_j \in \mathbb{R}^{1\times D}$ as

$$\boldsymbol{h}'_i = \sum_{j=1}^{L} w_j \boldsymbol{v}_j, \tag{2.2}$$

where each weight is based on a the similarity between the i^{th} query and the other keys in the sequence, measured as the dot product $w_{ij} = \boldsymbol{q}_i^T \boldsymbol{k}_j$. The output sequence of transformed tokens $\boldsymbol{h} \in \mathbb{R}^{L\times D}$ is the result of the SA operation applied to all sequence elements, defined by the attention matrix containing all weights $\boldsymbol{A} \in \mathbb{R}^{L\times L}$ and the operations

$$\boldsymbol{A} = \text{softmax}\Big(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{D}}\Big), \tag{2.3}$$

$$\boldsymbol{h}' = \text{SA}(\boldsymbol{h}) = \boldsymbol{A}\boldsymbol{V}. \tag{2.4}$$

A variant of SA called multi-head self-attention (MSA) runs N_h parallel SA operations focusing on different features or inter-dependencies of the data. The outputs of the different SA operations, called *heads*, are first concatenated and then linearly projected with learned weights $\boldsymbol{W}_h \in \mathbb{R}^{N_h D\times D}$ as

$$\text{MSA}(\boldsymbol{h}) = \underset{n\in\{N_h\}}{\text{concat}}[\text{SA}_n(\boldsymbol{h})]\boldsymbol{W}_h. \tag{2.5}$$

By definition, every token can attend to all previous and future tokens. Causal SA is a variant of SA applied to sequence modeling tasks restricting access to future information, where all elements above the diagonal in the attention matrix $\boldsymbol{A}$ are masked to 0. Additionally, since SA is invariant to the relative order of elements in the sequence, a fixed (Vaswani et al., 2017) or learned (Dosovitskiy et al., 2020) positional embedding $\boldsymbol{r} \in \mathbb{R}^{L\times D}$ is usually added or concatenated to the input tokens, where is element in the positional embedding sequence contains unique information about its position.

2.3. Model architecture and training

Figure 2.2 shows DoTA's architecture, which first applies the same series of convolutions to each 2D slice of the input sequence $\{\boldsymbol{x}_i | \boldsymbol{x}_i \in \mathbb{R}^{1\times H\times W}, \forall i = 1, ..., L\}$ separately. This convolutional encoder contains two blocks — both with a convolution, a Group Normalization (GN) (Y. Wu and He, 2020) and a pooling layer, followed by a Rectified Linear Unit (ReLU) activation — which extract important features from the input, e.g., material contrasts and tissue boundaries. After the second block, the outputs of a final convolution with K filters are flattened into a vector of embedding dimension

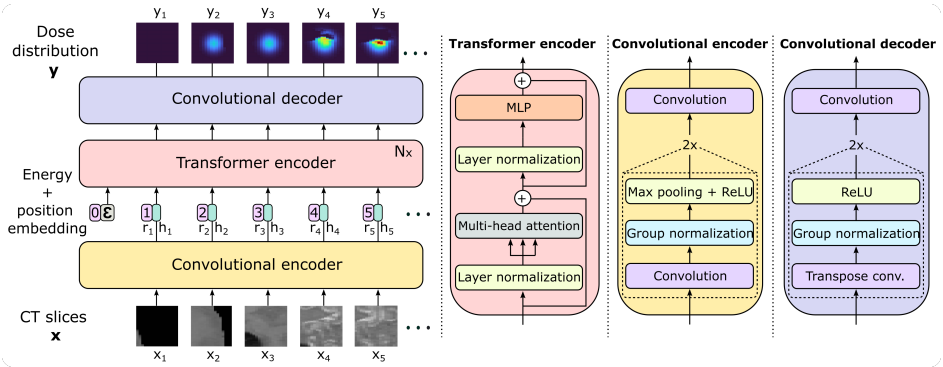


Figure 2.2: **DoTA architecture**. The input and output 3D volumes are treated as a sequence of 2D slices. A convolutional encoder extracts important geometrical from each slice into a feature vector. The particle energy is added at the beginning of the resulting sequence. A transformer encoder with causal self-attention subsequently combines information from the different elements of the sequence. Finally, a convolutional decoder individually transforms the low-dimensional vectors into output 2D dose slices.

$D = H' \times W' \times K$, where H' and W' are the reduced height and width of the images after the pooling operations. The convolutional encoder applies the same operation to every element $\boldsymbol{x}_i$, resulting in a sequence of L vectors $\{\boldsymbol{h}_i | \boldsymbol{h}_i \in \mathbb{R}^D, \forall i = 1, ..., L\}$ referred to as tokens in the remainder of the thesis.

A transformer encoder models the interaction between tokens $\boldsymbol{h}_i$ via causal MSA, resulting in an output sequence $\boldsymbol{h}' \in \mathbb{R}^D$. Since transformers operate on sets and by default do not account for the relative position of the slices in the sequence, a learnable positional encoding $\boldsymbol{r}_i \in \mathbb{R}^D$ is added to each token $\boldsymbol{h}_i$, e.g., $\boldsymbol{r}_1$ is always added to the token $\boldsymbol{h}_1$ from the first slice seen by the proton beam. The energy dependence is included via a 0^{th} token $\boldsymbol{h}_0 = \boldsymbol{W}_0 \varepsilon \in \mathbb{R}^D$ at the beginning of the sequence, where $\boldsymbol{W}_0 \in \mathbb{R}^{D \times 1}$ is a learned linear projection of the beam energy ε. Therefore, the transformer encoder blocks computes the operations

$$\boldsymbol{h} = [\boldsymbol{h}_e; \boldsymbol{h}] + \boldsymbol{r}, \tag{2.6}$$

$$\boldsymbol{h}_{int} = \boldsymbol{h} + \text{MSA}(\text{LN}(\boldsymbol{h})), \tag{2.7}$$

$$\boldsymbol{h}' = \boldsymbol{h}_{int} + \text{MLP}(\text{LN}(\boldsymbol{h}_{int})), \tag{2.8}$$

where MLP denotes a two layer feed-forward network with Dropout (Srivastava et al., 2014) and Gaussian Error Linear Unit (GELU) activations (Hendrycks and Gimpel, 2016). DoTA is based on the standard pre-Layer Normalization (LN) (Ba et al., 2016) transformer block (Xiong et al., 2020), alternating LN and residual connections with a self-attention operation and a MLP block.

Finally, a convolutional decoder independently transforms every output token to a 2D slice of the same size as the input $\{\boldsymbol{y}_i | \boldsymbol{y}_i \in \mathbb{R}^{1 \times H \times W}, \forall i = 1, ..., L\}$. The decoder's structure is identical to that of its encoder counterpart, but substituting the down-sampling convolution + pooling operation with an up-sampling convolutional transpose layer.

Dataset DoTA is trained to predict low noise MC dose distributions calculated with MCsquare (Souris et al., 2016), obtained using a set of 30 CT scans from prostate, lung and head and neck (H&N) cancer patients (Aerts et al., 2014, 2015; Clark et al., 2013) with 2 mm isotropic grid resolution. Given that proton beams have approximately 25 mm diameter and travel up to 300 mm through a small sub-volume of the CT, DoTA's input blocks $\boldsymbol{x} \in \mathbb{R}^{150\times24\times24}$ cover a volume of approximately $48 \times 48 \times 300$ mm^3. From each patient CT, $\approx$ 2,500 of such blocks are obtained — corresponding to beamlets being shot at different angles and positions — by effectively rotating, linearly interpolating and cropping the CT scan in steps of 10° and by applying 10 mm lateral shifts.

For each block, 2 different dose distributions are calculated using 10^7 primary particles to ensure MC noise values around 0.3% and always below 0.5%, zeroing out dose values below noise levels. Both dose distributions correspond to a randomly sampled beam energy between 70 and 220 MeV, with a 140 MeV cap in lung and H&N geometries given the potential to overshoot the patient. As a result, $\approx$ 80,000 individual CT block–dose distribution input–output pairs are obtained. This amount is further quadrupled by rotating the CT and dose blocks in steps of 90° around the beam direction axis, yielding a final training dataset consisting of $\approx$ 320,000 samples, 10% of which are used as a validation set to prevent overfitting.

An independent test set of 18 additional patients unseen during training is used for evaluation purposes, equally split into prostate, H&N and lung. Half of these patients (3 prostate, 3 H&N and 3 lung) are used to obtain 3,888 test beamlet dose distributions (1,386 lung, 1,512 H&N and 990 prostate samples), with the other half serving to evaluate DoTA's performance in full plans.

Training details The model is trained end-to-end using Tensorflow (Abadi et al., 2015), with the LAMB optimizer (You et al., 2019) and 8 samples per mini-batch, limited by the maximum internal memory of the Nvidia Tesla T4® GPU used during model training. The loss function is the mean squared error, with a scheduled learning rate starting at 10^{-3} that is halved every 4 epochs, with a restart after 28 epochs. In total, DoTA is trained for 56 epochs, saving the weights resulting in the lowest validation mean squared error. The best performing model consists of one transformer block with 16 heads and 12 convolutional filters in the last encoder layer, as obtained from a hyperparameter grid search evaluating the lowest validation loss across all possible combinations of transformer layers $N \in \{1,2,4\}$, convolutional filters $K \in \{8,10,12,16\}$ and attention heads $N_h \in \{8,12,16\}$. Given the two down-sampling pooling operations, the transformer processes tokens of dimension $D = H/4 \times W/4 \times K$, which in for the specific values of height $H = 24$, width $W = 24$, and $K = 12$ kernels results in $D = 432$.

2.4. Model evaluation

Using the ground truth MC dose distributions in the test set, DoTA' performance is compared to that of several data-driven dose engines, including LSTM (Neishabouri et al., 2021), and deep learning frameworks using noisy MC (Javaid et al., 2021) and PBA (C. Wu et al., 2021) doses as additional input. Since PBA is the analytical dose calculation method commonly used in the clinic and one of DoTA's competitors in terms of

speed and accuracy, the results include a PBA baseline from the open-source treatment planning software matRad (Wieser et al., 2017) [1].

Test set accuracy metrics The main mechanism used to compare predictions to the ground truth 3D dose distributions from the test set is the gamma analysis (Low et al., 1998). The gamma analysis is based on the notion that doses delivered in neighboring voxels have similar biological effects. Intuitively, for a set reference points — the voxel centers in the ground truth 3D volume — and their corresponding dose values, this method searches for similar predicted doses within small spheres around each point. The sphere's radius is referred to as distance-to-agreement criterion, while the dose similarity is usually quantified as a percentage of the reference dose, e.g., dose values are accepted similar if within 1% of the reference dose. Each voxel with coordinates $\boldsymbol{p} \in \mathbb{R}^3$ in the reference grid is compared to points $\boldsymbol{p}'$ of the predicted dose grid and assigned a gamma value $\gamma(\boldsymbol{p})$ according to

$$\gamma(\boldsymbol{p}) = \min_{\boldsymbol{p}'} \{\Gamma_{\boldsymbol{p},\boldsymbol{p}'}(d_{ta}, d_d)\}, \tag{2.9}$$

$$\Gamma_{\boldsymbol{p},\boldsymbol{p}'}(d_{ta}, d_d) = \sqrt{\frac{\left|\boldsymbol{p} - \boldsymbol{p}'\right|^2}{d_{ta}^2} + \frac{\left|\hat{y}_{\boldsymbol{p}} - y_{\boldsymbol{p}'}\right|^2}{d_d^2}}, \tag{2.10}$$

where $\hat{y}_{\boldsymbol{p}}$ is the reference dose at point $\boldsymbol{p}$, d_{ta} is the distance-to-agreement, and d_d is the dose difference criterion. A voxel passes the gamma analysis if $\gamma(\boldsymbol{p}) < 1$. The reported gamma pass rates — calculated as the fraction of passed voxels over the total number of voxels — further reduce the gamma evaluation to a single number per sample. All calculations are based on the PyMedPhys gamma evaluation functions [2].

Additionally, the average relative error ρ is used to explicitly compare dose differences between two beamlet dose distributions. Given the predicted output $\boldsymbol{y}$ and the ground truth dose distribution $\hat{\boldsymbol{y}}$ with $n_v = L \times H \times W$ voxels, the average relative error is calculated as

$$\rho = \frac{1}{n_v} \frac{\left\|\hat{\boldsymbol{y}} - \boldsymbol{y}\right\|_{L_1}}{\max \hat{\boldsymbol{y}}} \times 100. \tag{2.11}$$

Both the mean squared error (MSE) cost function used during training, and compute the root mean squared error (RMSE) between ground truth and predicted beamlet dose distributions are additionally reported. The RMSE is defined as

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{n_v} \sum_{i=1}^{n_v} .(\hat{y}_i - y_i)^2}. \tag{2.12}$$

Finally, an alternative metric to the gamma pass rate for full dose distribution comparison is the relative distribution error (RDE) (Nomura et al., 2020) between the ground truth and predicted D_{95}, D_{90}, D_{50} and D_{20} values, where D_v is the dose received by v% of the tumor volume. The RDE is computed relative to the planned dose D_{pr} as

[1] Publicly available at https://e0404.github.io/matRad/
[2] Publicly available at https://docs.pymedphys.com

$$\mathrm{RDE}(D_v, \hat{D}_v) = \frac{D_v - \hat{D}_v}{D_{pr}} \times 100. \tag{2.13}$$

Experiments A generic data-driven dose engine must yield accurate predictions for both single beamlet and full plan dose distributions. To ensure DoTA's suitability for replacing conventional particle transport tools in dose prediction tasks, its performance is assessed in two different settings:

- Individual beamlets. First, DoTA's speed and accuracy in predicting single beamlet doses is evaluated for 9 patients in the test set, comparing gamma pass rate distributions and inference times of DoTA, the LSTM models and the PBA baseline. Given the 2 mm × 2 mm × 2 mm grid resolution, a gamma evaluation $\Gamma(3\,\mathrm{mm}, 1\%)$ using a distance-to-agreement criterion $d_{ta} = 3$ mm ensures a neighborhood search of at least one voxel, while a dose criterion $d_d = 1\%$ disregards any uncertainty due to MC noise. Since DoTA's outputs are hardly ever 0 due to numerical inaccuracies of the last convolutional linear layer, and to disregard voxels not receiving any dose, voxels with doses below 0.1% of the maximum dose are excluded from the gamma pass rate calculations, resulting in a stricter metric (as the many voxels with near 0 dose could artificially increase the passing rate). Other reported results include the relative error ρ and RMSE between PBA/DoTA predictions and MC dose distributions, and where their error ρ and the gamma pass rate probability densities across all test samples.

- Full plans. A treatment plan with 2 fields is obtained for the remaining 9 test set patients using matRad. Given the list of beam intensities and energies in the plan, all dose distributions are recalculated using PBA, MCsquare (Souris et al., 2016) and DoTA, and their performance is evaluated via the gamma pass rate, masking voxels receiving a dose lower than 10% of the maximum dose. For each for each angle in the treatment plan, the CT is rotated prior to calculating the dose from each beamlet, while the resulting dose is rotated back to its original angle for dose accumulation. To allow for a fair comparison with other data-driven models — referred to as baselines B1 (Javaid et al., 2021) and B2 (C. Wu et al., 2021) — three gamma evaluations $\Gamma(1\,\mathrm{mm}, 1\%)$, $\Gamma(2\,\mathrm{mm}, 2\%)$ and $\Gamma(3\,\mathrm{mm}, 3\%)$ are computed, comparing the pass rate results to the available values in these baseline studies. DoTA is compared to the third baseline B3 via the RDEs, since the original B3 study (Nomura et al., 2020) does not report gamma pass rates. For more information about the experiments, Table 2.1 contains a description of the metrics and evaluation settings.

2.5. Results

In this section, DoTA's performance and speed is compared to state-of-the-art models and clinically used methods. The analysis is three-fold: assessing the accuracy in predicting beamlet dose distributions and full dose distributions from treatment plans, as well as DoTAs' potential as a fast dose engine by evaluating its calculation runtimes.

Table 2.1: **Overview of experiments.** Summary of the experiments, metrics and baselines used to evaluate DoTA's accuracy. D_{max} refers to the maximum dose value in a dose distribution and only voxels receiving dose above the cutoff level are included in the Γ calculations.

Experiment	Test data	Metric	Dose cutoff (Gy)	Baseline
Beamlets	3,888 beamlets (1,386 lung, 990 prostate, 1,512 H&N)	$\Gamma(3\text{ mm}, 1\%)$	0	LSTM
			0.1% of D_{max}	PBA
		Error ρ	0	PBA
		RMSE	0	PBA
Full plans	9 treatment plans	$\Gamma(1\text{ mm}, 1\%)$	10% of D_{max}	PBA, B2
		$\Gamma(2\text{ mm}, 2\%)$	10% of D_{max}	B1
		$\text{RDE}_{v\in\{20,50,90,95\}}$	Tumor doses	B3

2.5.1. Individual beamlets

For each individual beamlet in the test set, DoTA's predictions are compared to MC ground truth dose distributions using a $\Gamma(3\text{ mm}, 1\%)$ gamma analysis. In Table 2.2, the average, standard deviation, minimum and maximum of the distribution of gamma pass rates across test samples are reported. By disregarding voxels whose dose is below 0.1% of the maximum dose, the reported gamma pass rates are stricter than those of previous state-of-the-art studies (Neishabouri et al., 2021), where only voxels with a gamma value of 0 — which typically correspond to voxels not receiving any dose — are excluded from the pass rate calculation. Even with the stricter setting and including energy dependence, DoTA outperforms both the LSTM and PBA dose engines in all aspects: the average pass rates are higher, the standard deviation is lower, and the minimum is at least 5.5% higher. Similar results are observed for stricter gamma evaluation settings in Table 2.3. The left plot in Figure 2.3 further demonstrates DoTA's superiority, showing a gamma pass rate distribution that is more concentrated towards higher values. Each beam dose distribution is subsequently divided into 4 fragments of equal size between the entrance and the Bragg peak, where each fragment is referred to as *beam section* in the remainder of the chapter. The right plot in Figure 2.3 shows the proportion of voxels failing the gamma evaluation in each beam section, out of the total number of failed voxels, indicating for both PBA and DoTA that most of the failing voxels belong to the 4^{th} section, i.e., the high dose region around the Bragg peak where the effect of tissue heterogeneity is most evident.

As an additional measure of model performance, Table 2.4 shows the mean and standard deviation of the relative error ρ and RMSE between predictions and ground truth MC dose distributions in the test set. The results confirm DoTA's improvement, with mean, maximum error and standard deviation less than half of PBA's. The left plot in Figure 2.4 displays the distribution of ρ across all test samples, showing that values are smaller and closer to 0 for DoTA. As with the gamma pass rate, the beam is divided in 4 sections from entrance (1^{st}) to the Bragg peak (4^{th}), and the average relative error per section is shown in the right plot in Figure 2.4. Although both models show a similar trend with errors increasing towards the beam's end, DoTA is on average twice better than PBA.

Table 2.2: **Gamma pass rate of beamlet dose distributions.** Gamma analysis $\Gamma(3\text{mm}, 1\%)$ for the presented DoTA, the pencil beam algorithm (PBA) from matRad (Wieser et al., 2017) and the LSTM models are listed. Gamma pass rates are calculated using all test samples, with LSTM rates directly obtained from (Neishabouri et al., 2021). The reported values include the mean, standard deviation (Std), minimum (Min) and maximum (Max) across the test set for different sites, and 'Multi-site' refers to computing statistics using all sites.

Model	Site	Energy (MeV)	Mean (%)	Std (%)	Min (%)	Max (%)
LSTM	Lung	67.85	98.56	1.3	95.35	99.79
		104.25	97.74	1.48	92.57	99.74
		134.68	94.51	2.99	85.37	99.02
DoTA (ours)	Lung	[70, 140]	99.46	0.81	93.19	100
	H&N	[70, 140]	99.21	1.23	93.49	100
	Prostate	[70, 220]	99.51	1.46	94.06	100
DoTA (ours)	Multi-site	[70, 220]	99.37	1.17	93.19	100
PBA	Multi-site	[70, 220]	98.68	3.14	87.53	100

Table 2.3: **Additional gamma pass rate of beamlet dose distributions.** Gamma analysis $\Gamma(1\text{mm}, 1\%)$ and $\Gamma(2\text{mm}, 1\%)$ for DoTA and the pencil beam algorithm (PBA) from matRad (Wieser et al., 2017) are listed. The reported values include the mean, standard deviation (Std), minimum (Min) and maximum (Max) across all test samples.

Model	Energy (MeV)	Γ settings	Mean (%)	Std (%)	Min (%)	Max (%)
DoTA (ours)	[70, 220]	1mm, 1%	96.58	3.83	82.31	100
		2mm, 1%	98.67	2.04	89.69	100
PBA	[70, 220]	1mm, 1%	92.54	6.07	65.21	99.41
		2mm, 1%	97.20	4.27	76.49	100

Table 2.4: **Error of beamlet dose distributions.** The reported values include the mean, standard deviation (Std), minimum (Min) and maximum (Max) values of the relative error ρ and root mean squared error (RMSE) between 3,888 test predictions and reference MC dose distributions, for both the pencil beam algorithm (PBA) from matRad (Wieser et al., 2017) and DoTA.

Model	Relative error ρ (%)				RMSE (Gy)			
	Mean	Std	Min	Max	Mean	Std	Min	Max
DoTA (ours)	0.126	0.109	0.025	1.258	0.083	0.041	0.024	0.277
PBA (matRad)	0.306	0.309	0.059	4.077	0.294	0.126	0.057	1.293

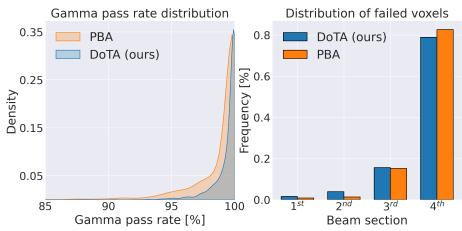


Figure 2.3: **Gamma pass rate distribution.** (Left) Distribution of the gamma pass rates $\Gamma(3\,\text{mm}, 1\%)$ of the test samples for the pencil beam algorithm (PBA) and the presented DoTA model. (Right) Distribution of the failed voxels along the beam, where each bin is an equally-sized fragment (referred to as section) of the beam from dose entrance (1^{st}) to Bragg Peak and dose falloff (4^{th}). Each bin shows the ratio of the number of test set voxels that fail the gamma evaluation within a section divided by the total number of failed voxels.

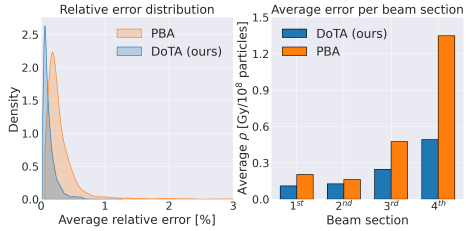


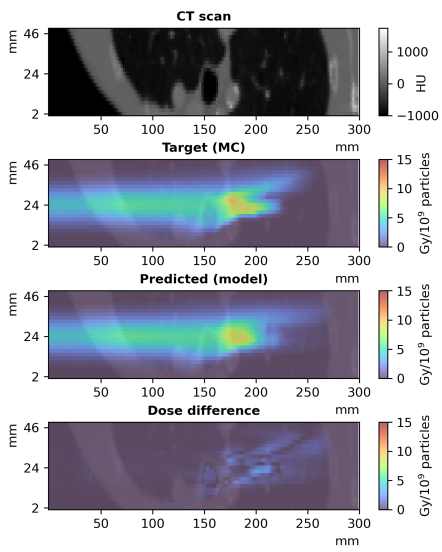
Figure 2.4: **Average relative error ρ distribution.** (Left) Distribution of the average relative error across the test samples for the pencil beam algorithm (PBA) and the presented DoTA model. (Right) Average relative error per beam section, where each bin is a section (4 equally-sized fragments) of the beam from dose entrance (1^{st}) to Bragg Peak and dose falloff (4^{th}). Each bin shows the average of the relative error values recorded within a section of the beam.

Finally, Figure 2.5b shows DoTA's test sample with the lowest gamma pass rate, together with PBA's prediction of the same sample (Figure 2.5a). Likewise, Figure 2.5d and Figure 2.5c show the predictions of the worst PBA sample from both models. In both cases, PBA results in errors as high as 80% of the maximum dose, severely overdosing parts of the geometry, while for DoTA errors are below 20% of the maximum dose.

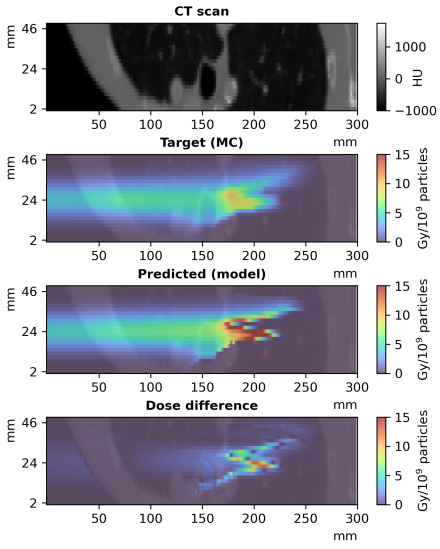
2.5.2. Full dose recalculation

To assess the feasibility of using DoTA as a dose engine in real clinical settings, DoTA's recalculated full dose distributions are compared to the reference MC doses via 3 different gamma analysis: $\Gamma(1\,\text{mm}, 1\%)$, $\Gamma(2\,\text{mm}, 2\%)$ and $\Gamma(3\,\text{mm}, 3\%)$, in decreasing order of strictness. The resulting gamma pass rates for each of the 9 test patients are shown in Table 2.5, showing values that are consistently high and similar across treatment sites, always at least 10% higher than PBA. DoTA is additionally compared to recently published state-of-the-art deep learning approaches: a MC-denoising U-net (B1) (Javaid et al., 2021) , and a U-net correcting PBA (B2) (C. Wu et al., 2021). Except for the prostate plans, DoTA outperforms both approaches, even without requiring the additional physics-based input.

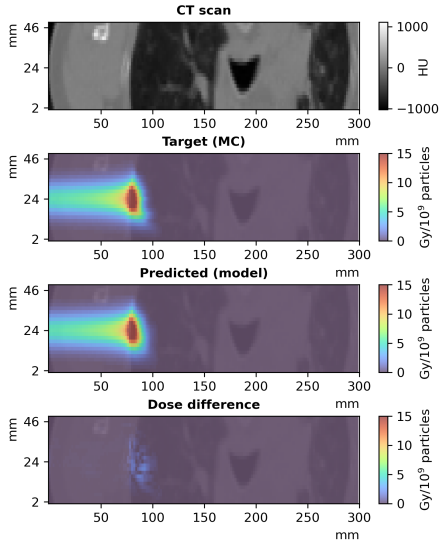
Figure 2.6 shows the RDE of DoTA and the B3 baseline (a convolutional neural network predicting dose distributions from Bragg peak position maps). B3 results are taken directly from the paper (Nomura et al., 2020), while DoTA values are computed using all test set dose distributions. With a significantly lower spread and values much closer to 0%, the results further confirm DoTA's superiority and accuracy gains.



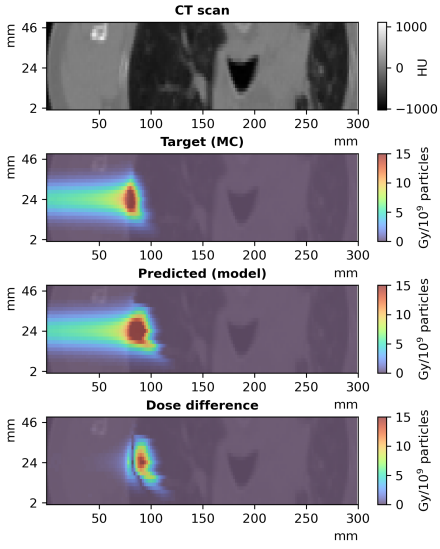
(a) DoTA's worst prediction.



(b) Pencil beam algorithm prediction of the worst DoTA sample.



(c) DoTA's prediction of the worst pencil beam algorithm sample.

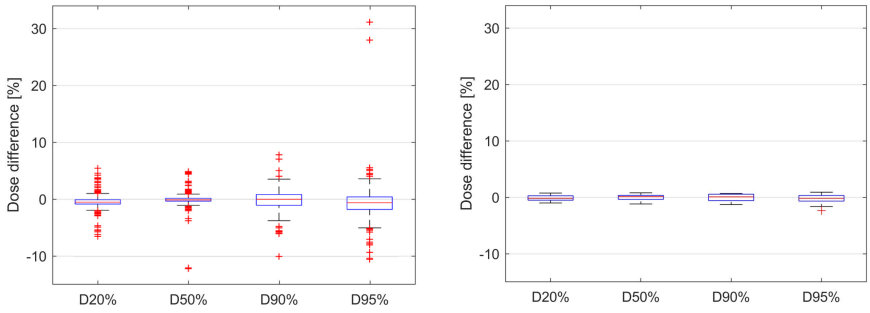


(d) Pencil beam algorithm worst prediction.

Figure 2.5: **Worst performing DoTA and PBA test sample.** (a) Worst performing test sample in the gamma evaluation for DoTA, with gamma pass rate of 93.19%, and (b) the pencil beam algorithm (PBA) prediction for the same sample. (d) Worst performing prediction in the gamma evaluation across the test set for PBA, with gamma pass rate of 87.53%, and (c) DoTA's prediction of the same sample. In descending order, all 4 subplots show: the central slice of the 3D input CT grid, the MC ground truth dose distribution, the model's prediction and the dose difference between the predicted and MC beams.

Table 2.5: **Gamma pass rate of planned dose distributions.** Treatment plans of 9 test patients are recalculated using the presented DoTA model, and compared to ground truth MC dose distributions via 3 different gamma analysis: $\Gamma(1\,\text{mm}, 1\%)$, $\Gamma(2\,\text{mm}, 2\%)$ and $\Gamma(3\,\text{mm}, 3\%)$. The $\Gamma(1\,\text{mm}, 1\%)$ pass rate for dose distributions recalculated with the pencil beam algorithm (PBA) from matRad (Wieser et al., 2017) is also reported. The baseline B1 corresponds to a MC-denoising U-net (Javaid et al., 2021), while B2 is a U-net correcting PBA (C. Wu et al., 2021), whose values are directly taken for their corresponding papers.

Site		**Number of spots**	**DoTA (ours)** $\Gamma(1,1)$	$\Gamma(2,2)$	$\Gamma(3,3)$	**PBA** $\Gamma(1,1)$	**B1** $\Gamma(2,2)$	**B2** $\Gamma(1,1)$
Lung	1	954	95.86	99.73	99.99	80.38	84.1	89.7±3.8
	2	2245	96.31	99.72	99.98	79.83		
	3	1646	95.63	99.64	99.97	78.92		
HN	4	1554	95.02	99.39	99.81	68.32	76.5	92.8±2.9
	5	1064	94.71	99.62	99.97	76.63		
	6	708	96.93	99.88	99.99	83.02		
Prostate	7	1598	96.38	99.81	99.99	87.34	-	99.6±0.3
	8	2281	95.78	99.82	99.99	77.12		
	9	1518	96.18	99.71	99.98	83.64		



(a) B3 baseline.

(b) DoTA.

Figure 2.6: **Relative dose errors.** Error between the ground truth and predicted D_{95}, D_{90}, D_{50} and D_{20} for (a) the B3 baseline (Nomura et al., 2020) and (b) the proposed DoTA model, relative to planned doses. Red crosses are outliers, red lines represent the median, and box boundaries denote the 25th and 75th percentiles.

Table 2.6: **Beamlet prediction runtime**. The reported values include the mean inference time and standard deviation (Std) taken by each model to predict individual beamlet dose distributions. Both the DoTA and LSTM models run on GPU hardware, while the pencil beam algorithm (PBA) (Wieser et al., 2017) and Monte Carlo (MC) dose engine use CPUs with multiple threads. LSTM inference times are taken directly from (Neishabouri et al., 2021).

Model	Mean (ms)	Std (ms)
LSTM[a]	6.0	1.5
DoTA[b] (ours)	5.0	4.9
PBA[c] (matRad)	728.3	30.9
MC[c]	43,636.9	12,291.6

[a] Nvidia® Quadro RTX 6000 64 Gb RAM
[b] Debian 10 4 vCPUs - Nvidia® A100 40 Gb RAM
[c] CentOS 7 8 CPUs intel Xeon® E5-2620 16Gb RAM

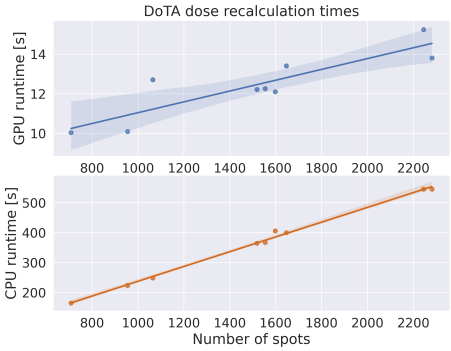


Figure 2.7: **Full dose recalculation runtime**. Time needed to recalculate planned dose distributions with DoTA using (top) a Nvidia® A100 GPU or (bottom) an intel Xeon® CPU. Estimates include time for loading CT and beam weights from plan data, for dose inference by DoTA and for the necessary CT and dose interpolations. Shaded areas denote the 95% confidence interval.

2.5.3. Prediction times

Apart from high prediction accuracy, fast inference is critically important for clinical applications. Table 2.6 displays the mean and standard deviation runtime taken by each model to predict a single beamlet. Being particularly well-suited for GPUs, DoTA is on average faster than LSTM and physics-based engines, offering more than 100 times speed-up with respect to PBA. Additionally, although dependent on hardware, DoTA approximates doses four orders of magnitude faster than MC, providing millisecond dose calculation times without requiring any extra computations for real-time adaptive treatments.

Regarding full dose recalculation from treatment plans, Figure 2.7 shows total runtimes for DoTA using both GPU and CPU hardware, including all steps from loading CT and beamlet weights from plan data files, necessary CT rotations and interpolations, DoTA dose inference time and reverse rotations and interpolation to assign dose on the original CT grid. Being optimized for GPU acceleration, DoTA is the fastest alternative, needing less than 15 seconds to calculate full dose distributions. For the baselines in this chapter, PBA runtimes oscillate between 100 and 150 seconds, while B1 and B2 report needing only few seconds to correct/denoise their inputs, but must add the runtime necessary to generate their respective PBA (123 s to 303 s in (C. Wu et al., 2021)) or MC (≈ 10 s in (Javaid et al., 2021)) input doses, as well as data transfer times between the physics engine and the deep learning framework. Furthermore, B2 is a per beam network, hence its runtime scales linearly with the number of beams, in practice meaning 2-4 times higher total calculation times.

2.6. Discussion

The presented DoTA model builds upon previous work learning proton transport as sequence modeling task via LSTM networks (Neishabouri et al., 2021), by introducing energy dependence and significantly improving its performance in a varied set of treatment sites. DoTA greatly outperforms analytical physics-based PBA algorithms in predicting dose distributions from individual proton pencil beams, achieving high accuracy even in the most heterogeneous patient geometries, demonstrated by the 6% improvement in the minimum gamma pass rate. With millisecond inference times, DoTA provides at least a factor 100 reduction in calculation time compared to the clinically still predominant analytical PBAs.

The drastic reduction in spot dose prediction times translates into the ability to calculate full dose distributions in 12 s on average and less than 15 s even for the plan with more than 2200 pencil beams, which times include the required time for all steps from loading CT and pencil beam weights from plan data (≈ 1 s on average), CT interpolation and beamlet geometry extraction (≈ 1 s), DoTA model and weights loading (≈ 2 s), dose inference by DoTA (≈ 7.5 s) and interpolating the final dose distribution back to the original CT grid (≈ 1 s). Although publicly available deep learning frameworks are optimized for GPU architectures and may offer an advantage with respect to adapting MC and PBA to GPU hardware, DoTA achieves this 10 s to 15 s speed on a single GPU card, even without any optimization of GPU settings for inference, which can reportedly yield up to 9 times speed-ups depending on the task [3]. Without sacrificing accuracy, DoTA represents at least a factor 10 speed-up with respect to PBAs and a 33% speed-up (and ≈ 80% considering the difference in MC noise levels) with respect to the fastest GPU MC competitor available in the literature — clinically used GPU MC software Raystation® (Fracchiolla et al., 2021), typically running in clusters or workstations with multiple GPUs and CPU cores. Moreover, DoTA offers a 10-25% increase in the $\Gamma(1\,\text{mm}, 1\%)$ gamma pass rate compared to PBA, and with a $\Gamma(2\,\text{mm}, 2\%)$ gamma pass rate >99% it matches (Y. Wang et al., 2016) or outperforms (Qin et al., 2016; Wan Chan Tseung et al., 2015) the accuracy of GPU MC approaches. DoTA's accuracy is also on par with the agreement between commercial MC engines (Raystation®) and experimental measurements (Schreuder et al., 2019a, 2019b). While the GPU-based PBA algorithm reported in (Silva et al., 2015) calculates a full distribution in 0.22 s and is faster than DoTA, it was tested only on a single patient showing worse accuracy with a 3% lower $\Gamma(2\,\text{mm}, 2\%)$ pass rate.

The proposed DoTA is also substantially superior to the only 3 published deep learning approaches for proton full plan dose calculations (Javaid et al., 2021; Nomura et al., 2020; C. Wu et al., 2021). DoTA achieves 15% and 25% higher $\Gamma(2\,\text{mm}, 2\%)$ pass rates compared to the MC-denoising U-net of (Javaid et al., 2021), and 6% and 2% higher $\Gamma(1\,\text{mm}, 1\%)$ pass rates compared to the PBA correcting U-net of (C. Wu et al., 2021) in lung and H&N patients, respectively. With lower RDE values much more concentrated around 0, DoTA also improves upon the dose prediction U-net based on Bragg peak position maps (Nomura et al., 2020). DoTA shows a slight inferiority in prostate patients, with a ≈ 3% lower $\Gamma(1\,\text{mm}, 1\%)$ pass rates than (C. Wu et

[3] Discussed in the non-peer-reviewed study in https://huggingface.co/transformers/v2.10.0/benchmarks.html

al., 2021). However, this direct comparison is somewhat unfair to DoTA. The Intensity Modulated Proton Therapy plans available in this work have small, 3 mm to 5 mm spot sizes, while in (C. Wu et al., 2021) double scattering proton therapy plans were used, which in general are less conformal and smoother, and therefore are expected to be easier to predict with data-driven approaches. The input and output voxel resolution of 2 mm × 2 mm × 2 mm is also finer compared to the 2 mm × 2 mm × 2.5 mm used in (C. Wu et al., 2021). Furthermore, (C. Wu et al., 2021) also reports site specific fine-tuning of their deep learning approach, unlike for DoTA. Last, (C. Wu et al., 2021) has the further disadvantage of using per beam PBA calculations as input, thus the reported 2 s to 3 s dose correction times easily translate to full treatment plan calculation times in the 5 min to 10 min range depending on the number of beams (taking into account the >2 min PBA run times), even without accounting for the additional time for the necessary CT rotations and interpolations.

DoTA's accuracy may further be increased by training with larger datasets, as demonstrated by the improvement achieved when increasing training data from 4 lung patients (Pastor-Serrano and Perkó, 2022b) to 30 patients with varied anatomies in the current study. Using dose distributions with lower MC noise could further improve performance. Convincingly outperforming all recent works learning corrections for 'cheap' physics-based predictions (Javaid et al., 2021; C. Wu et al., 2021) both in terms of accuracy and speed, DoTA has the flexibility to be used in a great variety of treatment sites and clinical settings.

Limitations The current version of DoTA is trained to predict MC ground truth dose distributions from a specific machine with unique settings and beam profiles, necessitating a specific model per machine. Likewise, range shifters — which are often dependent on treatment location and site — affect the dose delivered by some spots while inserted, thereby modifying the final dose distribution. Both problems could in principle be addressed by constructing a model that takes extra shape and range shifter specifications as input in the form of tokens at the beginning of the sequence, similar to how DoTA currently handles the energy dependence.

DoTA is trained for a specific voxel grid resolution, requiring either an individual model per resolution level or an additional interpolation step that will likely negatively interfere with the gamma pass rate results, especially for gamma evaluations $\Gamma(1,1\%)$ with a distance-to-agreement criterion lower than the voxel resolution level. While DoTA also works for finer nominal CT grids (Pastor-Serrano and Perkó, 2022b), an additional study testing the dose recalculation performance with more patients and finer grid resolution should confirm its suitability for direct clinical application needing such resolutions. MC noise may also affect the results of the gamma evaluation, as demonstrated in previous work (Cohilis et al., 2020) showing that even 1% MC noise levels introduce significant under-estimation in the gamma pass rate. Such detrimental effect is limited in the reported experiments given the lower noise levels of 0.3% in the ground truth MC doses (which level is considered as reference "denoised" in (Cohilis et al., 2020)).

One of the main problems of deep learning algorithms is their limited generalization or extrapolation capability outside the domain of the used training dataset. Us-

ing an independent test set of patients with varied geometries unseen during training, DoTA is clearly superior to all other methods in all evaluated scenarios, showing strong evidence of high level of generalization. Nevertheless, just like any deep learning approach, DoTA may also yield unrealistic predictions for data that vastly differs from the training data (e.g., in the presence of metallic implants), contrarily to MC engines, which – when using enough particles – are certain to provide valid results. Whether or not "more physics-based" PBAs perform better than DoTA in such cases is less straightforward. First, PBA clearly performed worse than DoTA in all tests, and in particular showed worse performance in the examples of Figure 2.5 exhibiting high heterogeneity (in Figures 2.5a-2.5b) and the Bragg peak position coinciding with a sharp change in density (in Figures 2.5c-2.5d). Second, the impact of approximations inherent to PBA approaches on the predicted dose in cases of unusual geometries (e.g., implants) is not easy to foresee without detailed analysis. The same holds for the error due to DoTA's potential generalization limitations in such cases. Although not supported by direct evidence, physics-based approaches (even approximative ones) may maintain a higher level of accuracy when going far beyond the training dataset domain. For the specific case of radiotherapy however, to a large extent these problems could be mitigated by including geometries with metallic implants in the training data set and teaching DoTA to accurately predict dose distributions in such scenarios too and by limiting use to (the vast majority of) patients who do not have implants until such improved model is available.

2.7. Summary

In this chapter, DoTA is presented: a generic, fast and accurate dose engine that implicitly learns proton particle transport physics and can be applied to speed up several steps of the radiotherapy workflow. Framing particle transport as sequence modeling of 2D geometry slices in the proton's beam travel direction, DoTA uses the power of transformers to predict individual beamlets with millisecond speed and close to MC precision. The presented results show that DoTA has the right attributes to potentially replace the proton dose calculation tools currently used in the clinics for applications that critically depend on runtime. Predicting dose distributions from single pencil beams in milliseconds, DoTA offers 100 times faster inference times than widely used PBAs, yielding close to MC accuracy as indicated by the very high gamma pass rate $\Gamma(3\,\text{mm}, 1\%)$ of $99.37 \pm 1.17\%$, thus has the potential to enable next generation online and real-time adaptive radiotherapy cancer treatments. The presented model predicts MC quality full plan dose distributions with at least a 10% improvement in gamma pass rate $\Gamma(1\,\text{mm}, 1\%)$ with respect to current analytical approaches and reduces dose calculation times of planned doses to less than 15 seconds, representing a tool that can directly benefit current clinical practice too.

3

Sub-second speed photon beam dose prediction

3.1. Introduction

Modern radiotherapy techniques such as intensity modulated radiation therapy (IMRT) or volumetric modulated arc therapy (VMAT) critically rely on accurate and fast calculations of the radiation dose delivered within the patient by photon beams, typically shaped by multi-leaf collimators (MLC) (Hussein et al., 2018). With modern workflows moving towards online or real time adaptation, fast dose calculations are critical for quick plan evaluation, re-optimization and finally being able to account for motion due to breathing or anatomical changes.

Commercial treatment planning systems mainly use pencil beam (PB) (Mohan et al., 1986), collapsed cone (CC) (Ahnesjö, 1989; Boyer and Mok, 1985), or Monte Carlo (MC) dose engines. While both PB and CC algorithms are usually faster than MC, the assumptions and approximations they use to solve photon particle transport result in less accurate results. Conversely, MC methods — the gold standard in dose calculation — simulate individual stochastic particle trajectories abiding the physical laws of nuclear interactions and track the deposited dose along these paths. By averaging results from enough particles (typically several millions), MC methods achieve very high accuracy even in the most complex patient geometries, at the cost of high computation times. Current commercial treatment planning systems mainly use improved PB or CC variations yielding close-to-MC accuracy, e.g., the anisotropic analytical algorithm (AAA) (Sievinen et al., 2013; Ulmer et al., 2005) based on the PB convolution (Mohan et al., 1986) in Eclipse (Varian Medical Systems) or the CC convolution algorithm in Pinnacle (Philips) (Boyer and Mok, 1985). Some recent MC implementations also use the parallelization capabilities of graphics processing units (GPUs) to reduce dose calculation times from several hours to minutes (Hissoiny et al., 2011; Jahnke et al., 2012;

The contents of this chapter have been accepted for publication as journal paper in *Medical Physics*.

Jia et al., 2011). Despite these advances, the need for accurate and fast dose calculation algorithms is still unmet in most clinical workflows, as neither PB nor MC are fast enough for real time treatment plan correction.

Recently, deep learning models have been applied to several steps of the radiotherapy workflow (Meyer et al., 2018), mainly as U-net convolutional architectures (Ronneberger et al., 2015) or Generative Adversarial Networks (Goodfellow et al., 2014). Most works aim to aid treatment planning by predicting clinically optimal doses based on historical data. As a result, they are constrained to a specific site, clinical optimum choice, and often fixed beam configurations, limiting their generalization capabilities. These models typically directly predict the full dose distribution using computed tomography (CT) images (Kearney et al., 2018), organ masks (Chen et al., 2019; Fan et al., 2019; Kajikawa et al., 2019; M. Ma et al., 2019; Nguyen, Long, et al., 2019), or additional information about the photon beam configuration (Nguyen, Jia, et al., 2019) as input. To further aid treatment planning, few studies additionally provide the beam intensities needed to deliver the predicted dose distribution (Lee et al., 2019; W. Wang et al., 2020).

Aiming at predicting dose distributions in generic setups, several subsequent studies present dose calculation models that estimate beam or full dose distributions from CTs and additional physics input such as high noise MC (Bai et al., 2021; Neph et al., 2021; Peng, Shan, Liu, Pei, Wang, and Xu, 2019) or PB doses (Dong and Xing, 2020; Xing, Zhang, et al., 2020); fluence maps, e.g., resulting from simple ray tracing calculations (Fan et al., 2020; Xing, Nguyen, et al., 2020); energy released per unit mass (J. Zhu et al., 2020); or a combination of the previous with additional beam information (Kontaxis et al., 2020; Tsekas et al., 2021). The reason for their success are the convolutional layers that excel at capturing local features and are heavily optimized for GPU hardware, but are less appropriate for modeling long-range dependencies, e.g., changes along the beam direction through the patient.

Although some of the most recent models can quickly predict dose distributions in most cases with good accuracy (Kontaxis et al., 2020; Tsekas et al., 2021; Tsekas et al., 2022), there is room for improvement with newer architectures that require less input information and can model distant features in the data. Recent transformer architectures (Vaswani et al., 2017) are particularly well-suited to process local and distant features, yielding excellent results in a wide range of sequence modeling tasks (Brown et al., 2020; Devlin et al., 2019; Dosovitskiy et al., 2020). For smaller datasets, transformers perform particularly well when combined with convolutional layers (D'Ascoli et al., 2021). Based on these synergies between convolutions and transformers, a recent study presented a transformer-based algorithm predicting proton beamlet 3D dose distributions as a sequence of 2D slices in the beam depth, with state-of-the-art performance and speed (Chapter 2, Pastor-Serrano and Perkó, 2022a, 2022b).

This chapter presents a deep learning model that can predict dose distributions in few milliseconds with clinically acceptable accuracy. Based on the previous transformer-based proton dose calculation model in Chapter 2, the model harnesses the power of hybrid transformer and 3D convolutional architectures to predict the dose of much bigger photon broad beams, as in concurrent work (F. Xiao et al., 2022). As shown in Figure 3.1, the proposed improved Dose Transformer Algorithm (iDoTA) combines

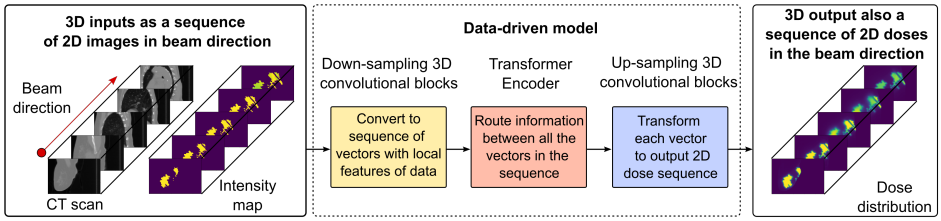


Figure 3.1: **Model overview.** A deep learning data-driven model learns the mapping $\boldsymbol{y} = f_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{r})$ between input 3D CT $\boldsymbol{x}$ and projected shape $\boldsymbol{r}$ volumes, and the corresponding output 3D dose distributions $\boldsymbol{y}$. The problem is formulated as a sequence prediction task, where all input and output cubes are treated as a sequence of 2D slices in the beam's eye view. Each 2D slice is mapped into a vector via a series of down-sampling convolutional blocks. A transformer backbone routes information between all elements of the resulting sequence. Finally, a several convolutional operations up-sample and transform each vector into a 2D dose distribution map.

a series of 3D convolutional layers modeling local dose and tissue variations, with a transformer backbone routing information along the depth of the entire photon beam. The model treats input 3D CT and projected shape volumes (containing beam geometrical information) as a sequence of 2D slices in the direction of the beam, framing dose calculation as sequence modeling to produce a sequence of 2D dose slices forming the 3D dose distribution. After comparison to the best-performing data-driven models, iDoTA shows superior speed and accuracy for photon dose calculation tasks, being capable to speed up beam prediction times down to few milliseconds and reducing treatment plan computation times to few seconds.

3.2. Model architecture and training

This section presents the problem setup and architecture of the iDoTA model, used to predict photon beam doses from 3D CT and projected shape inputs. Additionally, the dataset and training procedure used to optimize the model parameters are described, as well as the evaluation metrics used to assess iDoTA's performance as a generic photon dose calculation engine.

Proposed framework Photon dose calculation involves estimating the radiation dose delivered in the patient geometry. If the machine parameters do not change, the predicted dose distribution mainly depends on the irradiated geometry and the beam geometrical information such as the MLC aperture shape, the beam angle and the relative position of the isocenter. All the necessary beam shape information is captured in a 3D projected shape $\boldsymbol{\kappa} \in \mathbb{R}^{L\times H\times W}$ of depth L, height H and width W, containing the result of a simple ray tracing operation propagating the photon beam shape through the patient geometry CT scan $\boldsymbol{x} \in \mathbb{R}^{L\times H\times W}$. The outcome of the dose calculation operation predicted by our model is another grid $\boldsymbol{y} \in \mathbb{R}^{L\times H\times W}$ with the 3D distribution of dose per monitor unit (MU).

As shown in Figure 2.2, the patient CT $\boldsymbol{x}$ and the 3D projected shape $\boldsymbol{\kappa}$ are inputs to iDoTA, which during training implicitly learns the mapping $\boldsymbol{y} = f_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\kappa})$ via a cascade

of neural networks layers with parameters $\boldsymbol{\theta}$. Framing the dose prediction task as modeling a sequence of D elements in the direction of the photon beam, iDoTA combines the strengths of both convolutional and transformer architectures into a single model. The input geometry $\boldsymbol{x}$ can be expressed a sequence of L images in the direction of the beam $\{\boldsymbol{x}_i | \boldsymbol{x}_i \in \mathbb{R}^{1\times H\times W}, \forall i = 1,...,L\}$, while the projected shape 3D input $\boldsymbol{\kappa}$ is similarly viewed as a sequence 2D slices $\{\boldsymbol{\kappa}_i | \boldsymbol{\kappa}_i \in \mathbb{R}^{1\times H\times W}, \forall i = 1,...,L\}$ containing beam information. Likewise, the final dose volume $\boldsymbol{y}$ is also expressed as the sequence of 2D dose slices $\{\boldsymbol{y}_i | \boldsymbol{y}_i \in \mathbb{R}^{1\times H\times W}, \forall i = 1,...,L\}$).

Model architecture As seen in Figure 3.2, the proposed architecture combines a series of convolutional blocks modeling local features with a transformer backbone that processes information along the entire beam depth.

- First, a series of down-sampling convolutional blocks extract local features of the data into a sequence of vectors $\{\boldsymbol{h}_i | \boldsymbol{h}_i \in \mathbb{R}^D, \forall i = 1,...,L\}$ — referred to as tokens in the remainder of the paper — of size D. Each block contains a 3D convolutional layer with kernel size equal to 3, modeling local features from the immediately preceding and succeeding elements in the sequence, followed by a layer normalization (Ba et al., 2016), a rectified linear unit (ReLU) activation function and a max-pooling operation. All such operations in the block are applied in parallel to every element of the input sequence. Due to the max-pooling operation, the height H and width W of the slices are halved after each block. A total of lev blocks result in lev resolution levels. After the last block, a final convolution with K filters and flattening operation transforms the resulting features into tokens of dimension $D = \left(\frac{H}{2}\right)^{lev} \times \left(\frac{W}{2}\right)^{lev} \times K$. The result is a sequence of L tokens containing local features about the corresponding input slices, e.g., the third token $\boldsymbol{h}_3$ represents local features from the inputs $\boldsymbol{x}_3$, $\boldsymbol{\kappa}_3$ and their neighboring slices.

- A transformer backbone routes information between the extracted features along the depth of the entire volume, with the self-attention mechanism (Vaswani et al., 2017) making the information exchange dynamic, i.e., each token $\boldsymbol{h}_i$ is independently transformed based on its content and information selectively gathered from other sequence elements. To account for the relative distance between tokens, a learnable positional embedding $\boldsymbol{r}_i \in \mathbb{R}^D$ is added to each token $\boldsymbol{h}_i$, i.e., a sequence of vectors $\{\boldsymbol{r}_i | \boldsymbol{r}_i \in \mathbb{R}^D, \forall i = 1,...,L\}$ is learned and added to the token sequence before the first operation in the transformer. The transformer block follows a pre-Layer Normalization architecture (Xiong et al., 2020), which consists of a Layer Normalization (LN) (Ba et al., 2016) operation, followed by a self-attention operation (Vaswani et al., 2017), and two fully-connected layers with Dropout (Srivastava et al., 2014) and a Gaussian Error Linear Unit (GeLU) activation (Hendrycks and Gimpel, 2016).

- Finally, a series of lev up-sampling convolutional blocks convert the token sequence into the output dose volume. For each level, the sequence previously obtained from the same level down-sampling convolutional block is appended along the feature dimension, similar to U-net architectures. The up-sampling

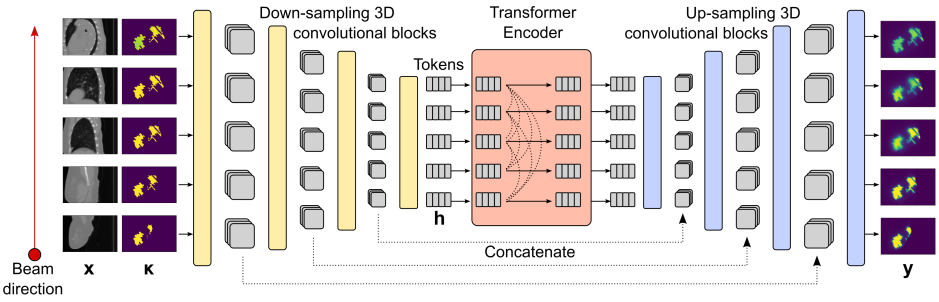


Figure 3.2: **Model architecture**. The proposed model solves the dose prediction task as sequence modeling, mapping two input sequences of 2D CT slices $\boldsymbol{x}$ and projected shapes $\boldsymbol{\kappa}$ with beam shape information into a sequence of 2D dose distributions $\boldsymbol{y}$. First, a series of down-sampling convolutional blocks merges and compresses the two sequences from the data into a sequence of feature vectors $\boldsymbol{h}$ (referred to as tokens). A transformer encoder with causal self-attention routes long-range dependencies along the beam direction. Finally, a series of up-sampling convolutional blocks transform the output tokens into a sequence of 2D dose distributions. In each block, the exact same 3D convolution operation is applied to all sequence elements, extracting local features from the preceding and following element in the sequence.

block's architecture is identical to that of its down-sampling counterpart, except for the use of a nearest-neighbor up-sampling interpolation operation instead of the max-pooling.

Projected shape and dose calculation Apart from the values in the CT, the additional 3D projected shape input $\boldsymbol{\kappa}$ encodes beam information such as the MLC aperture shape, the angle or relative distance between the isocenter and the source, including basic material information with a simple correction based on tissue densities. Such projected shape is generated via an algorithm that estimates the dose at each voxel through the percentage depth dose (PDD), corrected by an off-axis factor. The PDD is measured at 100 cm source-to-surface distance (SSD) with a 10 cm × 10 cm field size, adjusting for different SSDs using the Mayneord factor. The depth for determining the percentage dose is the water equivalent distance, calculated via ray tracing for all voxels. The off-axis correction factor is calculated by sampling from a diagonal beam profile for a 40 cm × 40 cm field size at 10 cm depth, projecting it to different depths using the lateral distance of the voxel to the center beam axis and the longitudinal distance from the voxel to the source. This ray tracing calculation estimates the dose using the commissioning data and is optimized for speed over accuracy, taking around 0.1 ms per beam in a GPU. The corresponding ground truth dose distributions (to be predicted by the model) are obtained via the AcurosXB V15.6.05 algorithm in the Varian Eclipse TPS system (with the option of calculating dose to medium). Both the dose and the projected shapes have similar ranges from 0 to ≈ 3, with units cGy/MU.

Dataset iDoTA is trained to predict individual photon beams using a training dataset of 17 clinical patient CTs with disease sites of brain, head neck, lung, abdomen and pelvis. All CTs were recorded using a General Electrics LightSpeed CT scanner with

2.5 mm × 2.5 mm × 2.5 mm resolution. For each patient, 100 different co-planar photon beams were computed, using, for each beam, a random gantry angle and an isocenter location randomly selected within the patient, and an aperture shape that was generated by randomly sampling leaf positions, keeping the couch angle fixed. After calculating the dose per MU and projecting the aperture shape, the input 3D CT $\boldsymbol{x} \in \mathbb{R}^{96\times96\times64}$, projected shape $\boldsymbol{\kappa} \in \mathbb{R}^{96\times96\times64}$ and output dose $\boldsymbol{y} \in \mathbb{R}^{96\times96\times64}$ blocks are obtained, covering a volume of approximately $240 \times 240 \times 160$ mm^3, so that the beam always travels in the same direction along the first dimension $L = 96$ with angles between -45° and 45°. All 1700 input CT volumes are normalized to the range [0,1] dividing by using the maximum value of 3,071 observed across the entire dataset. Likewise, both projected shapes and dose distributions are normalized using the maximum dose value of 3.075 cGy/MU in the dataset. During training, 10% of the samples are set aside for validation purposes, i.e., finding the best model configuration.

The best model's performance is evaluated using an independent test dataset of 584 beam dose distributions corresponding to a prostate and a lung patient unseen during training. Additionally, to assess iDoTA's performance in predicting full dose distributions composed of many photon beams, the model is tested on 11 additional clinical VMAT treatment plans with 2 arcs and 99-178 control points per arc, corresponding to 1 brain, 3 HN, 3 lung, and 4 prostate cancer patients.

Training details iDoTA is trained using the mean squared error as a loss function, with mini-batches of 4 samples and the layer-adaptive LAMB optimizer (You et al., 2019), finding the combination of a low batch size and the LAMB optimizer to be critical for convergence. During training, the dataset size is augmented via rotations (in steps of 90 °, perpendicular to the direction of the beam) and random shifts along the beam direction (shifting the entire volume up to 15 positions along the first dimension). Training consists of 10 cycles with 120 epochs/cycle, where the learning rate is set to 10^{-3} at the beginning of each cycle, and halved every 15 epochs.

After hyper-parameter tuning using the validation data, the best-performing model has 4 transformer heads, $lev = 4$ levels with $K = 10$ filters in the last encoder convolution. The four down-sampling operations in the encoder transform the input slices with dimensions $H = 96$ and $W = 64$ into tokens of size $D = H/2^4 \times W/2^4 \times K = 240$. All training and experiments are run in a Nvidia A40® GPU using Tensorflow (Abadi et al., 2015).

3.3. Model evaluation

For evaluation purposes, iDoTA's predictions are compared to corresponding ground truth dose distributions in the independent test set of patients unseen during training. The main method to assess dosimetric differences is the gamma analysis (Low et al., 1998), based on the intuition that two neighboring voxels with a similar dose result in equivalent biological effects. Intuitively, a voxel in the predicted dose distribution passes the gamma evaluation $\Gamma(d_{ta}\,\text{mm}, d_d\%)$ if another voxel with a similar value — deviating less than d_d% of the maximum dose — is found within a sphere of radius d_{ta} mm in the ground truth dose grid. Three gamma evaluations $\Gamma(1\,\text{mm}, 1\%)$,

$\Gamma(2\,\text{mm}, 2\%)$ and $\Gamma(3\,\text{mm}, 3\%)$ are computed, calculating the gamma passing rate by dividing the number of passed voxels by the total amount of eligible voxels, i.e., voxels with values within 10% and 100% of the maximum dose.

The average relative error ρ serves as an additional metric to measure explicit voxel dose differences, expressed as a percentage of the maximum dose in the grid. As for the gamma pass rate, the average relative error is calculated only for voxels with values within 10% and 100% of the maximum dose. For model predictions $\boldsymbol{y}$, and corresponding ground truth 3D dose distributions $\hat{\boldsymbol{y}}$ (both with $M = L \times H \times W$ voxels), the average relative error is calculated using the L_1-norm as

$$\rho = \frac{1}{M} \frac{\left\| \boldsymbol{y} - \hat{\boldsymbol{y}} \right\|_{L_1}}{\max \hat{\boldsymbol{y}}} \times 100. \tag{3.1}$$

3.4. Results

To assess iDoTA's suitability as a generic photon dose calculation tool and determine its improvements with respect to other data-driven algorithms, different evaluation metrics are computed using the independent test data. In particular, iDoTA's accuracy and speed are compared to previous approaches for prediction of both individual photon beam prediction and full dose distributions from clinical VMAT plans.

3.4.1. Individual beams

Table 3.1 reports the $\Gamma(1\,\text{mm}, 1\%)$, $\Gamma(2\,\text{mm}, 2\%)$ and $\Gamma(3\,\text{mm}, 3\%)$ gamma pass rate for the 584 beams in the test dataset, comparing their mean, standard deviation and minimum to those reported in previous studies achieving state-of-the-art performance, i.e, the B1 and B2 convolutional architectures for photon dose prediction in standard linear accelerator (Linac) (Kontaxis et al., 2020) and MR-Linac settings (Tsekas et al., 2021), respectively. Table 3.1 additionally includes the pass rates of a purely convolutional version of iDoTA without the transformer encoder, referred to as iDoTA-conv. The overall lower pass rates achieved by iDoTA-conv demonstrate the added benefit of combining transformers and convolutions. In general, iDoTA achieves better pass rates than previous convolutional models, with higher means and smaller standard deviations. Most importantly, the minimum gamma pass rate across all test samples is >20% higher than that of the 3D-U-net based architectures.

iDoTA can better predict photon beams in pelvic anatomies than in lung scans, which is likely caused by the more heterogeneous nature of lung geometries (i.e., the contrasts between bony structures, air, and water-like tissues). Figure 3.3 further confirms iDoTA's superiority for the pelvic cases over lung, showing gamma pass rates, and ρ distributions with lower lung pass rates and higher errors. Figure 3.4 visually compares the target and predicted beam dose distributions for the worst-performing lung and pelvic samples, and an average-performing pelvic beam. The overall errors are low and mostly occur at the beam lateral falloff, which may be caused by the coarse resolution of the input projected shapes. Since the average relative error in test data of 2.18% is similar to the final error in validation data of $2.19 \pm 1.08\%$, and relatively close to the error for training data of $1.54 \pm 0.64\%$, we can conclude that the model generalizes well.

Table 3.1: **Model accuracy for individual broad beams.** Gamma pass rates for photon beams are computed using 3 different criteria in the gamma evaluation. The reported values, which include the mean, standard deviation (std), and minimum across all test samples from pelvic and lung cancer patients, are compared to other state-of-the-art 3D U-net deep learning models as reported in their respective studies, referred to as B1 (Kontaxis et al., 2020) and B2 (Tsekas et al., 2021). To determine the added benefit of using transformers, a purely convolutional variant of iDoTA — without the transformer encoder, denoted as iDoTA-conv — is trained and evaluated using the same training procedure and dataset.

Site	Model	$\Gamma(1,1)$ [%]		$\Gamma(2,2)$ [%]		$\Gamma(3,3)$ [%]	
		Mean±std	Min	Mean±std	Min	Mean±std	Min
Pelvic	3D U-net (B1)	**89.9±5.1**	44.5	97.8±3.0	55.2	99.4±2.5	62.5
	3D U-net (B2)	87.6±8.3	47.5	97.9±2.6	68.2	99.5±1.0	77.5
	iDoTA - conv	85.8±8.6	32.46	97.0±4.6	52.8	99.2±2.1	76.2
	iDoTA (ours)	89.0±5.4	**66.9**	**98.1±1.7**	**87.7**	**99.6±0.5**	**94.7**
Lung	iDoTA - conv	84.3±4.1	65.5	95.6±2.0	86.9	98.8±0.8	94.0
	iDoTA (ours)	84.1±4.7	**68.9**	**96.9±2.0**	**90.1**	**99.2±0.8**	**94.2**

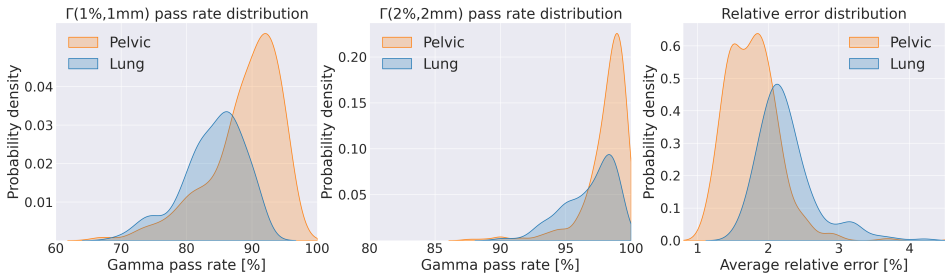


Figure 3.3: **Accuracy metrics distribution.** (Left)$\Gamma(1\,\text{mm}, 1\%)$. pass rate, (middle) $\Gamma(2\,\text{mm}, 2\%)$ pass rate and (right) average relative error distributions across all beams in the test dataset. The lower errors and higher pass rate values in orange correspond to beams in the pelvic area, while blue distributions are from lung samples.

3.4.2. Full dose distributions

For 11 additional patients outside the training dataset with clinical VMAT plans available, Table 3.2 compares the $\Gamma(1\,\text{mm}, 1\%)$, $\Gamma(2\,\text{mm}, 2\%)$ and $\Gamma(3\,\text{mm}, 3\%)$ gamma pass rate to the values reported in previous studies. In particular, iDoTA's accuracy are compared to those of: convolutional U-net architectures predicting each beam in the plan individually B1 (Kontaxis et al., 2020), B2 (Tsekas et al., 2021), B3 (Tsekas et al., 2022); convolutional models de-noising MC dose distributions B4 (Neph et al., 2021), B5 (Bai et al., 2021); and the concurrent TransDose transformer model for MR-Linac dose prediction (F. Xiao et al., 2022).

Table 3.2 shows the mean and standard deviation of the gamma pass rates separately for pelvic, lung and HN patients, comparing them to other models. With a $99.51 \pm 0.66\%$ (2 mm, 2%) pass rate, an average relative dose error of $0.75 \pm 0.36\%$ across all patients, and higher pass rates in all treatment sites, iDoTA outperforms all previous approaches. Additionally, the average error ρ in HN, lung and pelvic plans is

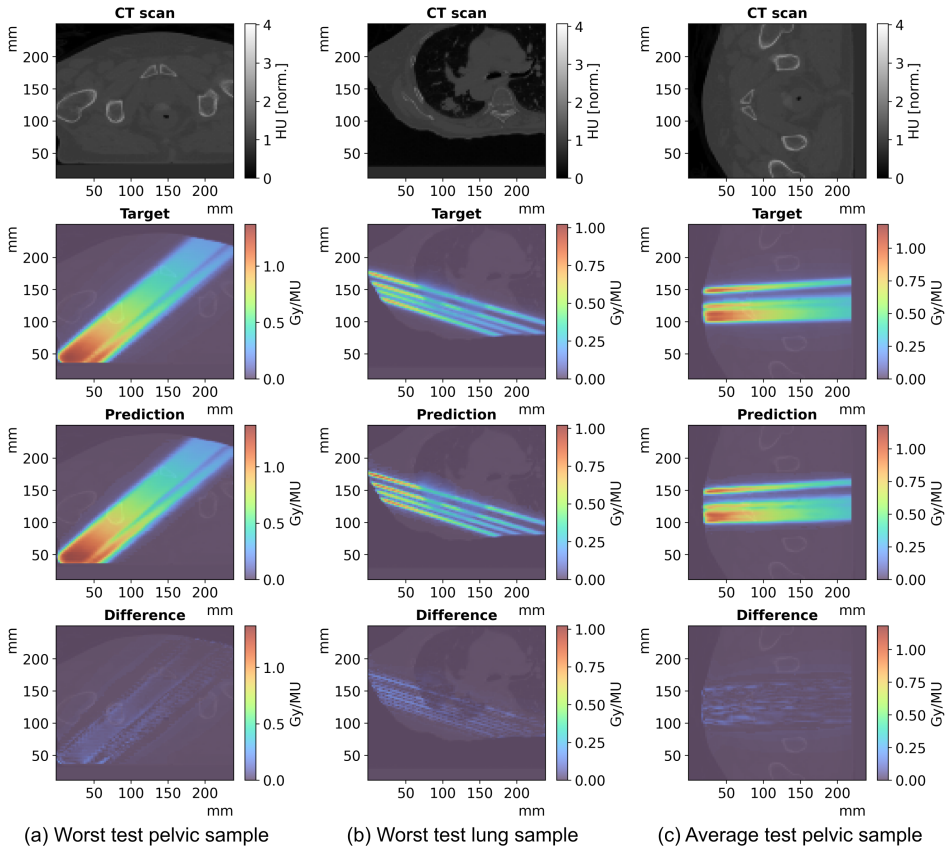


Figure 3.4: **Individual beam test samples.** (a) Worst performing pelvic test sample in the gamma evaluation, with $\Gamma(2\,\text{mm}, 2\%)$ gamma pass rate of 87.7%; (b) worst performing prediction in the gamma evaluation across the lung test samples, with $\Gamma(2\,\text{mm}, 2\%)$ gamma pass rate of 90.1%, and (c) average performing sample. Given the $96 \times 96 \times 64$ 3D volumes — a sequence of 96 2D slices of dimension 96×64 — all plots show the central slice along the beam direction, i.e, slice 32 out of 64. From top to bottom rows, the subplots show the 3D input CT grid, the reference dose distribution, the model's prediction and the dose difference between the predicted and reference beams.

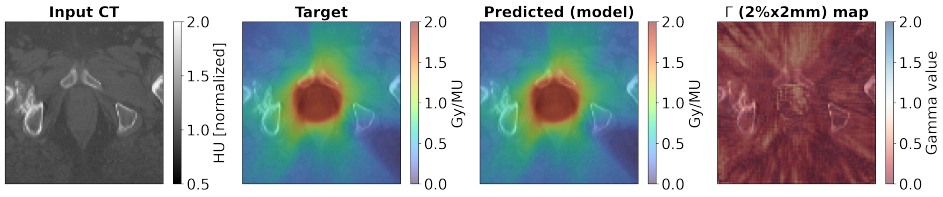
Table 3.2: **Model accuracy for full clinical dose distributions.** For different treatment sites, the gamma pass rates of full photon dose distributions is calculated using 3 different criteria. The values from few of the best-performing models are also included as reported in their respective studies. In particular, iDoTA is compared to the 3D U-net models B1 (Kontaxis et al., 2020), B2 (Tsekas et al., 2021), B3 (Tsekas et al., 2022); the de-noising 3D U-net models B4 (Neph et al., 2021), B5 (Bai et al., 2021); and the concurrent TransDose transformer model (F. Xiao et al., 2022). All pass rates include the average and standard deviation across all available dose distributions.

Treatment site	Model	$\Gamma(1,1)$ [%]	$\Gamma(2,2)$ [%]	$\Gamma(3,3)$ [%]
Head & Neck	B4	70.9±2.9	89.4±3.7	-
	TransDose	-	96.7±2.3	-
	iDoTA (ours)	**80.5±8.6**	**98.9±0.9**	**99.9±0.1**
Pelvic	B1	89.9±3.3	99.5±0.7	99.9±0.3
	B2	82.2±9.7	96.1±3.1	99.4±0.6
	B3	84.2±2.9	99.0±0.4	99.9±0.1
	B5	-	95.4±1.6	-
	TransDose	-	97.9±0.4	-
	iDoTA (ours)	**95.8±3.1**	**99.8±0.2**	**99.9±0.0**
Lung	TransDose	-	96.7±1.4	-
	iDoTA (ours)	**94.3±1.5**	**99.8±0.2**	**99.8±0.1**

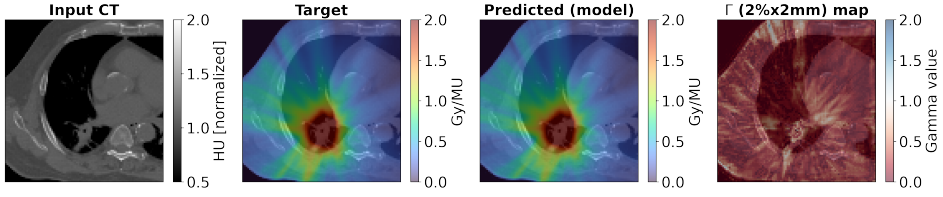
1.11%, 0.64%, and 0.45%, respectively. For the remaining patient with a brain tumor, a $\Gamma(1\,\text{mm}, 1\%)$, $\Gamma(2\,\text{mm}, 2\%)$ and $\Gamma(3\,\text{mm}, 3\%)$ gamma pass rate of 93.5, 99.7, and 99.9, respectively. As seen in the individual beams, iDoTA is more accurate in pelvic cases and less precise in HN anatomies, which is also likely due to the bone, water and air (cavities) heterogeneities. Nevertheless, the overall pass rate is still significantly higher than other approaches. Finally, Figure 3.5 shows very similar reference and predicted dose distributions for a prostate (Figure 3.5a) and lung (Figure 3.5b) VMAT plan, along with the corresponding $\Gamma(2\,\text{mm}, 2\%)$ map with mostly all voxels passing the gamma evaluation. To further evaluate the similarity between ground truth and predicted doses, Figure 3.6 shows dose volume histograms (DVHs) from three test patients. The almost perfectly overlapping DVH lines indicate that iDoTA's predictions are practically identical to the reference data.

3.4.3. Prediction times

Computation speed is critically important in adaptive workflows. In Table 3.3, iDoTA's total time needed to predict individual beams and full plans is compared to the reported values from models in previous studies. All prediction times for all models include the time needed to generate and prepare the inputs, predict the output and (for full dose distributions) accumulate beam doses. For individual beam prediction, iDoTA is significantly faster than any other competitor, being 30-60x faster than the 3D U-net models and 6x faster than the concurrent transformer model TransDose (F. Xiao et al., 2022). Likewise, iDoTA predicts full dose distribution from VMAT plans (with 194-354 beams per plan) on average in 8 seconds, representing a 10-80x speed-up com-



(a) Prostate patient.



(b) Lung patient.

Figure 3.5: **Dose distributions from VMAT plans.** From left to right, the input CT, target and predicted dose distributions and $\Gamma(2\,\text{mm}, 2\%)$ gamma map are shown for two clinical VMAT plans from a (a) prostate and (b) lung cancer patient. To show details of the high dose region, the images display crops around the target volume.

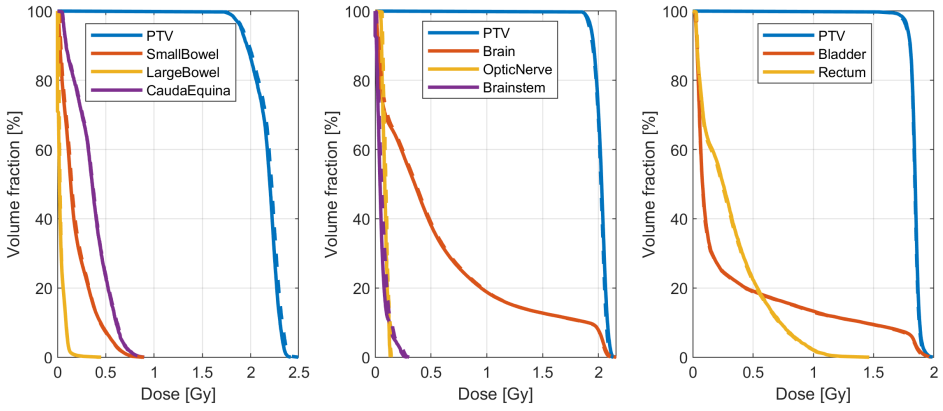


Figure 3.6: **Dose volume histograms from 3 VMAT plans.** Three dose volume histograms from a (left) pelvic, (mid) brain and (right) prostate test patients are shown, indicating the dose received by a specific fraction of the volume of an organ. All plots include the planning target volume (PTV) and few of the surrounding organs at risk. Solid lines represent iDoTA's predictions, while dotted lines indicate ground truth values.

Table 3.3: **Average prediction time.** iDoTA's computing speed is compared to the fastest models in literature via the average computing time needed to predict a photon beam or full dose distribution. The reported baselines include the 3D U-net models B1 (Kontaxis et al., 2020), B2 (Tsekas et al., 2021), B3 (Tsekas et al., 2022); the de-noising 3D U-net models B4 (Neph et al., 2021), B5 (Bai et al., 2021); and the concurrent TransDose transformer model (F. Xiao et al., 2022). The reported values include the time needed to generate and process the model inputs. iDoTA's CPU prediction times are also included for comparison, as well as the average number of beams in the evaluated treatment plans.

	Model	**Hardware**	**Average time [ms]**
Photon beams	B1	GPU	1500
	B2	GPU	3000
	B3	GPU	7000
	TransDose	GPU	310
	iDoTA (ours)	CPU	1480
	iDoTA (ours)	GPU	**50**
	Model	**Hardware**	**Average time [s]**
Full plans	B1 (< 20 beams)	GPU	60
	B4 (< 20 beams)	GPU	660
	B5 (< 20 beams)	GPU	150
	iDoTA (ours, 194 – 354 beams)	CPU	450
	iDoTA (ours, 194 – 354 beams)	GPU	**8**

pared to the IMRT (with ≈ 10 beams) U-net models. With CPU settings (intel® Core™ i7-8550U 1.8 GHz), iDoTA still remains competitive with previous GPU-based models, predicting beam doses in 1.48 ± 0.13 seconds and full plans in 300 to 600 seconds, depending on the number of beams.

3.5. Discussion

Framing photon dose calculation as sequence modeling, iDoTA is able to predict beam doses with high accuracy and speed, achieving an overall $97.72 \pm 1.93\%$ $\Gamma(2\,\text{mm}, 2\%)$ pass rate in lung and pelvic geometries. This per-beam prediction precision translates into a very high $\Gamma(2\,\text{mm}, 2\%)$ pass rate of $99.51 \pm 0.66\%$ in dose distributions from clinical VMAT plans, which also outperforms all previous models. Compared to the best-performing convolutional models B1 and B2 (Kontaxis et al., 2020; Tsekas et al., 2021) (and the recently published VMAT model B3 (Tsekas et al., 2022)), iDoTA offers more than 30x faster beam dose prediction even in the most heterogeneous geometries, achieving better gamma pass rates on average with lower spread, and 20% higher pass rates in the most difficult samples. Furthermore, iDoTA only uses the 3D CT and beam intensity to predict doses, in contrast to the 5 different input volumes containing physics information required by the 3D U-nets, allowing for lower input generation times and faster calculation times overall. iDoTA also convincingly outperforms MC de-noising models B4 and B5 (Bai et al., 2021; Neph et al., 2021), with a 5-10% increase in gamma pass rates and a 20-80x speed-up, partially caused by the time needed to generate the high-noise MC dose inputs. In general, iDoTA achieves higher gamma

pass rates than all previous convolutional models, also compared to the purely convolutional iDoTA-conv variant trained with identical dataset, training procedure and architecture (except for the transformer encoder). As in Chapter 2 and the concurrent TransDose (F. Xiao et al., 2022), these findings demonstrate that the addition of the transformer — being able to capture relationships between distant features, as opposed to convolutions — seems to be beneficial for dose prediction tasks.

Moreover, our method outperforms the concurrent TransDose transformer model in both accuracy and speed. Although TransDose is trained to predict photon beams under magnetic fields for MR-Linac applications — which could be a more difficult task to learn — part of iDoTA's success may be due to differences in the model, i.e., that the data-demanding transformer architecture in iDoTA routes information only between each of the 96 2D slices, instead of the 5000 voxels that are input to the transformer in TransDose. As a result, iDoTA's transformer has less parameters, which can be favorable with smaller datasets and accelerates inference.

With higher accuracy and lower computing times than any other previously introduced deep learning model, the proposed iDoTA represents a new state of the art in data-driven photon dose calculation. iDoTA can predict full dose distributions in 6-10 seconds, including CT cropping and rotation time (≈ 25 ms per beam), ray tracing input calculation (≈ 0.1 ms per beam) loading the model and weights (≈ 2 s), inferring the beam dose distribution (≈ 20 ms per beam) and accumulating the doses in the final grid (≈ 5 ms per beam). As a result, iDoTA is an order of magnitude faster than clinically used algorithms or MC approaches adapted to GPU hardware (Hissoiny et al., 2011; Jahnke et al., 2012; Jia et al., 2011). While such MC-GPU implementations are several orders of magnitude faster and almost as accurate as their CPU counterparts, their total calculation times are still in the order of minutes. Furthermore, iDoTA is 20x and 60x faster than the Eclipse Acuros XB and AAA algorithms (Varian Medical Systems) used in ≈ 80% of the clinics, which predict VMAT doses in 2-3 and ≈ 10 minutes, respectively (Fogliata et al., 2012; Yan et al., 2017). Most importantly, the photon beams can be predicted in parallel in several batches depending on the number of GPUs and their internal memory, practically allowing for further reduction in total calculation times.

Limitations Like all other data-driven algorithms, iDoTA is trained to emulate dose distributions from a specific machine and settings. Deep learning algorithms have limited extrapolation capabilities outside the training domain, which would require a different model each time the machine configuration is changed (or even the CT scanner, unless different CT machines are included in the training dataset). In such cases, fine-tuning iDoTA starting from the provided weights using a smaller dataset can save time without significantly degrading performance.

Ideally, all machine characteristics would be given to the model as separate inputs. Alternatively, to account for geometrical information and machine characteristics, iDoTA requires the additional input projected shape, necessitating ray-tracing pre-calculations. As for the machine parameters, such beam information could be included in the input as separate tokens, e.g., the aperture shape could be given as 2D binary mask at the beginning of the input sequence.

iDoTA is trained using a certain resolution and grid dimensions, which must be fixed for both training and inference. For dose prediction in finer grid resolutions, iDoTA can be coupled to neural representation models capable of accurate super resolution (Vasudevan et al., 2022). Regarding grid size, predicting dose distributions from treatment plans or beams through anatomies larger than the predetermined voxel grid must be done in steps, obtaining several input volumes and accumulating the outputs along the beam depth. Conversely, all doses can be predicted for the same fixed grid covering the part of the anatomy containing the structures of interests, which neglects the (usually) low doses near patient entrance. As observed in proton dose prediction (chapter 2), iDoTA is expcted to perform equally well for different grid settings, with calculation times going up for larger grids and finer resolutions, but still within sub-second speed.

Finally, iDoTA is trained and evaluated on a dataset that differs from the ones used in previous models, which can affect the final evaluation metrics. Likewise, the high-end GPU used in our experiments may affect iDoTA's reported prediction times. Nevertheless, our GPU is not expected to offer significant speed improvements with respect to the one from previous studies, especially if compared to the fastest alternative (F. Xiao et al., 2022) using modern, similar GPU hardware. iDoTA's intrinsic speed is further confirmed by its competitive prediction times even when using a CPU, as shown in Table 2.6, which is partly due to using less parameters and a faster input generation.

3.6. Summary

Combining the convolutional layers extracting local features with a Transformer backbone routing distant information, the presented iDoTA model outperforms any previous deep learning model in photon dose calculation. iDoTA can predict beam dose distributions in 50 milliseconds with high accuracy, achieving an average $\Gamma(2,2)$ pass rate of 97.72%. The per-beam prediction speed translates into estimating full VMAT dose distributions in less than 10 seconds with a $\Gamma(2,2)$ pass rate of 99.51% on average, instead of the several minutes required by clinical algorithms or previous data-driven models. Given its speed and versatility, iDoTA can accelerate several steps of the radiotherapy workflow: from treatment planning and quality assurance to real-time adaption.

4

Modeling inter-fraction daily anatomical variations

4.1. Introduction

Modern radiotherapy techniques such as intensity modulated proton therapy (IMPT) have the potential to deliver highly conformal doses to tumors while maximally sparing organs at risk (OARs). Although offering dosimetric advantages with respect to conventional modalities, such treatments are particularly sensitive to geometrical uncertainties arising from setup errors before delivery or range errors caused by organ movements between or during treatment sessions. In the presence of uncertainties, planned doses are delivered to anatomies different from the 3D computed tomography (CT) scan used during treatment planning, which may translate into shifting high dose regions away from clinical target volumes (CTVs) into critical OARs. Being one of the main sources of error in, e.g., prostate cancer treatments (van Herk et al., 2002), the magnitude of the deformations and their effect on the final dose distribution must be quantified to ensure robust delivery. Ideally, treatments could be real-time adapted via image guidance, or alternatively adjusted before each treatment session (Jagt et al., 2017, 2018), but such adaptive workflows are constrained by the speed of the CT acquisition, delineation, dose calculation and treatment re-optimization processes in practice.

An efficient alternative currently used in the clinic consists of including setup and range uncertainties during treatment planning optimization to design robust treatment plans that withstand positioning and range errors (Rojo-Santiago et al., 2021; Unkelbach and Paganetti, 2018; van der Voort et al., 2016). Similarly, inter-fractional movement information could be incorporated during treatment planning or treatment evaluation to make treatment plans robust against complex geometrical variations. To account for such anatomical changes, some published works propose computing ex-

The contents of this chapter have been accepted for publication as a journal paper in *Physics in Medicine & Biology*.

pected dose distributions using weighted scenarios, where each scenario corresponds to the dose deposited in a patient geometry generated by an anatomy model. Typically, such models extract the main eigenmodes of organ deformation — groups of correlated movements — via principal component analysis (PCA) (Budiarto et al., 2011; Jeong et al., 2010; Söhn et al., 2005; Szeto et al., 2017). During the last decades, linear PCA models have been successfully employed to quantify and understand the effect of organ deformations in different treatment sites and modalities (Magallon-Baro et al., 2019; Rios et al., 2017; Thörnqvist, Hysing, Zolnay, Söhn, Hoogeman, Muren, and Heijmen, 2013); to extend clinical volumes with extra margins and compensate for anatomical changes (Bondar et al., 2014; Thörnqvist, Hysing, Zolnay, Söhn, Hoogeman, Muren, Bentzen, and Heijmen, 2013); to characterize respiratory deformations (Badawi et al., 2010; Q. Zhang et al., 2007); and to simulate dosimetric outcomes of delivery in the presence of geometrical uncertainties (Nie et al., 2012; Söhn et al., 2012; Tilly et al., 2017; Xu et al., 2014). Focusing on conventional photon-based modalities, most of these studies are based only on organ contours without including CT intensity values, and require time-consuming image registrations as pre-processing to find corresponding points across a population of patients before being usable for learning generic deformations. Furthermore, all previously introduced models are either patient-specific (requiring several CTs per patient) or population-based (applying the same set of deformations to all patients), which limits their accuracy and applicability. For widespread adoption of anatomically robust treatment planning, accurate probabilistic models quickly generating patient-specific treatment anatomies are required.

All published PCA models learn correlated organ movements from a dataset of 3D deformation vector fields (DVFs), where each vector indicates the magnitude and direction of displacement for each point in a voxelized volume. Such DVFs can be obtained via image registration algorithms finding a non-linear correspondence between, e.g., two CT scans (Ashburner, 2007; Bruveris and Holm, 2015; Vásquez Osorio et al., 2009). While traditional not data-driven algorithms require minutes to solve a registration task, recent deep learning based methods reduce computing times down to few seconds and additionally increase registration accuracy (Balakrishnan et al., 2019; de Vos et al., 2017), typically using 2D (Ronneberger et al., 2015) or 3D (Çiçek et al., 2016) U-net convolutional architectures in combination with spatial transformer networks (Jaderberg et al., 2015). Several architectures generating DVFs and warping pairs of images have been proposed and applied to radiotherapy problems such as 4D image registration of moving images due to breathing (Lei et al., 2020; Romaguera et al., 2020) or automated contour propagation in adaptive workflows (Liang et al., 2021).

Our objective, however, is to generate a set of DVFs to warp a single planning CT into different repeat CTs that are likely to be observed during the course of a radiotherapy treatment. Ideally, a suitable model would be able to implicitly capture the relative likelihood of correlated groups of movements depending on the input patient geometry. Probabilistic frameworks based on variational inference (Blei et al., 2017; Kingma and Welling, 2014; Rezende et al., 2014) have been successfully applied to model uncertainty in organ segmentation tasks (Baumgartner et al., 2019; Hu et al., 2019; Kohl et al., 2019, 2018), making use of auxiliary latent variables that represent the main factors of variation behind the model's predictions. Similar probabilistic U-net based architec-

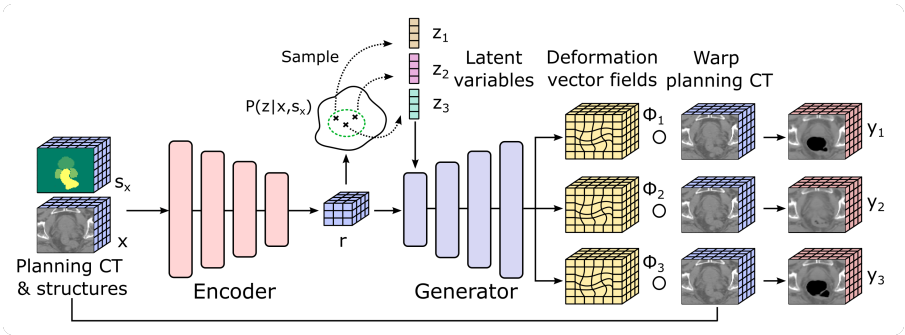


Figure 4.1: **Proposed generative framework.** The probabilistic models are embedded within a U-net, where the down-sampling path is referred to as Encoder, and the up-sampling path is the Generator. The Encoder takes the planning CT and structures and outputs both a compressed representation of the input $\boldsymbol{r}$ as well as a distribution $P(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{s}_x)$ over the region of the latent space containing variables corresponding to plausible patient-specific movements. Given $\boldsymbol{r}$ and any sample $\boldsymbol{z}$ from the latent space distribution, the Generator outputs a deformation vector field that is used to warp the planning CT into an artificial repeat CTs.

tures have also been proposed for pure image registration tasks (Dalca, Balakrishnan, et al., 2019; Krebs et al., 2019), with applications to unsupervised contouring problems (Dalca, Yu, et al., 2019) and breathing movement prediction based on motion surrogates (Romaguera et al., 2021).

Extending on these recent architectures, this chapter presents a probabilistic deep learning framework that represents common anatomical movements and deformations in a population of patients using few latent variables. The proposed daily anatomy model (DAM) first generates DVFs conditioned on an input planning CT scan and latent variables, where each combination of latent variables corresponds to a different group of movements; and subsequently warps the planning CT with the generated DVFs into a set of artificial repeat scans. The model is trained using a dataset containing planning and repeat CTs recorded at different stages of prostate cancer treatments in three different institutions, evaluating whether DAM is able to learn realistic movements with two external patients. Compared to previous methods, DAM does not require any pre-processing registration step and can in principle be applied to quickly simulate patient anatomies for treatment adaptation and robustness evaluation purposes.

4.2. Model architecture and training

This section describes the fundamentals of diffeomorphic transformations and the variational framework used to capture anatomical variations, including the different parametric models and the procedure used to tune their parameters. Subsequently, the model architecture is described in detail, together with the data and the evaluation metrics used in each experiment.

4.2.1. Proposed framework

During the course of a radiotherapy treatment, the internal structures and organs of the patient change between fractions/days. As a result, the anatomy captured in the planning image $\boldsymbol{x} \in \mathbb{R}^M$ and organ structures $\boldsymbol{s}_x \in \mathbb{R}^M$ (both represented as 3D matrices) can significantly differ from the repeat images $\boldsymbol{y} \in \mathbb{R}^M$ and structures $\boldsymbol{s}_y \in \mathbb{R}^M$ taken during following treatment sessions. M voxels comprise the entire volume, where the voxels in $\boldsymbol{x}$ and $\boldsymbol{y}$ represent image intensity values, and the voxels in $\boldsymbol{s}_x$ and $\boldsymbol{s}_y$ contain an integer corresponding to the organ present in the voxel.

As demonstrated in previous studies (Budiarto et al., 2011) for treatment sites like prostate, common anatomical variations such as volume and contour changes are observed across an entire population. Based on the existence of such generic movements it is assumed that, given a planning image $\boldsymbol{x}$ and structures $\boldsymbol{s}_x$, there is an unknown patient-specific generative distribution $P^*(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{s}_x)$ of repeat scans that can be approximated via a probabilistic model with learned parameters. Given a planning image from a new patient, the resulting model distribution $P_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{s}_x)$ parametrized by $\boldsymbol{\theta}$ can be sampled to generate a set of artificial anatomies observed at future treatment stages.

In this case, $\boldsymbol{\theta}$ corresponds to the parameters of the U-net neural network that is used to compute a DVF $\Phi : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ mapping coordinates between images. The presented model predicts a diffeomorphic transformation Φ, which is invertible and practically allows obtaining the forward and inverse transformations in a very simple manner. Based on the seminal works in (Ashburner, 2007; Dalca, Balakrishnan, et al., 2019), the selected diffeomorphic transformation is represented via the ordinary differential equation

$$\frac{\partial \Phi^{(t)}}{\partial t} = v(\Phi^{(t)}) \tag{4.1}$$

describing the evolution of the deformation over time, where $t \in [0, 1]$ is time, $\Phi^{(0)}$ is the identity transformation and $v : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ is a stationary velocity field. Generating a DVF starts from the identity transformation $\Phi^{(0)}$, integrating over time to obtain $\Phi^{(1)}$. The integration is done via scaling and squaring (Arsigny et al., 2006; Moler and Van Loan, 2003), which involves recursively updating the DVF in T successive small time steps

$$\Phi^{(1/2^T)} = \boldsymbol{p} + v(\boldsymbol{p})/2^T \tag{4.2}$$

$$\Phi^{(1/2^{t-1})} = \Phi^{(1/2^t)} \circ \Phi^{(1/2^t)} \tag{4.3}$$

$$\Phi^{(1)} = \Phi^{(1/2)} \circ \Phi^{(1/2)} \tag{4.4}$$

where $\boldsymbol{p}$ are spatial locations. Typically, T is chosen so that $v(\boldsymbol{p})/2^T$ is small, with higher T leading to more accurate solutions.

As for the inputs, in the presented DAM the velocity field $\boldsymbol{v} \in \mathbb{R}^{M\times 3}$ and DVF $\boldsymbol{\Phi} \in \mathbb{R}^{M\times 3}$ are discretized into M voxels, using $\boldsymbol{\Phi}(\boldsymbol{p})$ to denote the displacement applied to the voxel centered at location $\boldsymbol{p} \in \mathbb{R}^3$. Following previous work (Dalca, Balakrishnan, et al., 2019), the U-net predicts $\boldsymbol{v}$, which is exponentiated via scaling and squaring using

a spatial transformer network (Jaderberg et al., 2015) to obtain the final DVF $\boldsymbol{\Phi}$ used to warp planning images into artificial repeats $\boldsymbol{y} = \boldsymbol{\Phi} \circ \boldsymbol{x}$.

Generative model DAM is probabilistic model that conditions the generated DVFs (and thus also the repeat images) on N unobserved latent variables $\mathbf{z} \in \mathbb{R}^N$ capturing the main factors of variation in the data, i.e., the main groups of anatomical deformations. The latent variables distribute following a multivariate Gaussian prior probability distribution that depends on the input planning anatomy

$$P(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{s}_x) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{s}_x), \boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{s}_x)), \tag{4.5}$$

where the mean $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ and diagonal covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ are deterministic functions calculated by a neural network referred to as *Encoder* (Figure 4.1), which corresponds to the down-sampling path of a U-net. The prior dependence on the input results in a different distribution over latent variables per patient, which allows the model to select the groups of movements that are likely to be observed for each specific input image. The Encoder additionally outputs a volume $\boldsymbol{r} = g_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{s}_x)$, which is the results of several deterministic convolution operations containing features from the input. Since $\boldsymbol{r}$ is a deterministic function of the input, any conditioning on $\boldsymbol{r}$ is substituted with $\boldsymbol{x}$ and $\boldsymbol{s}_x$ in the remainder of the chapter.

The relationship between the input planning image and latent variables and the output warped repeat images is computed in the up-sampling path of the U-net, which takes sampled latent variables and the low-dimensional features $\boldsymbol{r}$ to generate a velocity field $\boldsymbol{v}_{\boldsymbol{z},\boldsymbol{\theta}} = f_{\boldsymbol{\theta}}(\boldsymbol{z}, \boldsymbol{x}, \boldsymbol{s}_x)$, where the subscripts denote the deterministic dependence to $\boldsymbol{z}$ and $\boldsymbol{\theta}$. After integrating $\boldsymbol{v}_{\boldsymbol{z},\boldsymbol{\theta}}$ to obtain the DVF $\boldsymbol{\Phi}_{\boldsymbol{z},\boldsymbol{\theta}}$, the output repeat image $\boldsymbol{y} \in \mathbb{R}^M$ is obtained by warping the input as $\boldsymbol{y} = \boldsymbol{\Phi}_{\boldsymbol{z},\boldsymbol{\theta}} \circ \boldsymbol{x}$.

Different latent variable samples $\boldsymbol{z}$ result in different repeat images given the same input planning scan, and the modeled distribution of repeat images can be recovered as a function of the prior $P(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{s}_x)$ and a likelihood $P_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{x}, \boldsymbol{s}_x)$ distributions as

$$P_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{s}_x) = \int P_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{x}, \boldsymbol{s}_x) P(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{s}_x) d\boldsymbol{z}. \tag{4.6}$$

The choice of the likelihood distribution affects the final loss function. Based on previous work (Krebs et al., 2019), the likelihood distribution is a function of the *cross-correlation* (CC) between predicted $\boldsymbol{y}$ and ground-truth $\hat{\boldsymbol{y}}$ images, scaled by a constant λ as

$$P_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{x}, \boldsymbol{s}_x) \propto \exp(\lambda \mathrm{CC}(\hat{\boldsymbol{y}}, \boldsymbol{y} = \boldsymbol{\Phi}_{\boldsymbol{z},\boldsymbol{\theta}} \circ \boldsymbol{x})). \tag{4.7}$$

The CC has been empirically found to yield better similarity than other metrics such as the mean squared error (Balakrishnan et al., 2019), with larger CC values corresponding to more alike images. Let $y(\boldsymbol{p})$ and $\hat{y}(\boldsymbol{p})$ denote the intensity values for each voxel at position $\boldsymbol{p}$ in the predicted and ground-truth images, respectively. If $w(\boldsymbol{p})$ and $\hat{w}(\boldsymbol{p})$ are images where each voxel is the local mean of the n^3 neighboring voxels, e.g., $w(\boldsymbol{p}) = \frac{1}{n^3} \sum_{j=1}^{n^3} y(\boldsymbol{p}_j))$ and $\hat{w}(\boldsymbol{p}) = \frac{1}{n^3} \sum_{j=1}^{n^3} \hat{y}(\boldsymbol{p}_j))$, the CC is defined as

$$\mathrm{CC}(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \sum_{\boldsymbol{p}\in\Omega} \frac{\left[\sum_{i=1}^{n^3}(\hat{y}(\boldsymbol{p}_i) - \hat{w}(\boldsymbol{p}))(y(\boldsymbol{p}_i) - w(\boldsymbol{p}))\right]^2}{\left[\sum_{i=1}^{n^3}(\hat{y}(\boldsymbol{p}_i) - \hat{w}(\boldsymbol{p}))\right]\left[\sum_{i=1}^{n^3} y(\boldsymbol{p}_i) - \hat{w}(\boldsymbol{p}))\right]}. \tag{4.8}$$

As in previous work (Krebs et al., 2019), DAM always uses the mode of the distribution $\boldsymbol{\Phi}_{\boldsymbol{z},\boldsymbol{\theta}} \circ \boldsymbol{x}$, instead of sampling the likelihood $P_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{x}, \boldsymbol{s}_x)$ each time during inference to generate anatomies.

Learning With the presented probabilistic formulation, the goal is to maximize Equation 4.6 by learning the parameters $\boldsymbol{\theta}$ from a dataset containing planning $\boldsymbol{x}$ and repeat $\boldsymbol{y}$ pairs. However, estimating the integral over the latent space would require sampling a large number of latent variables, being intractable in practice. A variational framework is used instead, defining an *approximate posterior* distribution $Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{s}_x, \boldsymbol{y}, \boldsymbol{s}_y)$, parametrized by an *Inference Network* with parameters $\boldsymbol{\phi}$. During training, the Inference Network has access to the real repeat scans and predicts the parameters of Gaussian distribution covering a small region of the latent space containing variables that are likely to explain the deformation between $\boldsymbol{x}$ and $\boldsymbol{y}$ scans. Thus, the predicted Gaussian is

$$Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{s_x}, \boldsymbol{y}, \boldsymbol{s_y}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{s}_x, \boldsymbol{y}, \boldsymbol{s}_y), \boldsymbol{\Sigma}_{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{s}_x, \boldsymbol{y}, \boldsymbol{s}_y)), \tag{4.9}$$

with deterministic mappings $\boldsymbol{\mu}_{\boldsymbol{\phi}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\phi}}$ computed by the Inference neural network. Our formulation allows estimating the model parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ by minimizing the negative *evidence lower bound* as

$$\begin{aligned}\log\left(P_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{s}_x)\right) \leq &-\mathbb{E}_{\boldsymbol{z}\sim Q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{s}_x,\boldsymbol{y},\boldsymbol{s}_y)}[\log(P_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{x}, \boldsymbol{s}_x))] \\ &+D_{KL}(Q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{s}_x, \boldsymbol{y}, \boldsymbol{s}_y)||P_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{s}_x)).\end{aligned} \tag{4.10}$$

The lower bound balances two terms: the $D_{KL}(\cdot||\cdot)$ term — Kullback - Leibler (KL) divergence — forces the approximated posterior to be close to the prior distribution, while the first term corresponds to maximizing the CC, encouraging similarity between real and generated images. Further details about deriving the lower bound are included in Appendix A.

Explicit regularization terms The current form of the likelihood enforces image similarity regardless of structure overlap or DVF quality. To enforce realistic predicted anatomies, the lower bound is modified by adding two extra regularization terms. To encourage smooth and realistic DVFs, a *spatial* regularization term penalizes large unrealistic spatial gradients $\nabla\boldsymbol{\Phi}_{\boldsymbol{z},\boldsymbol{\theta}}(\boldsymbol{p}) = \left(\frac{\partial\boldsymbol{\Phi}_{\boldsymbol{z},\boldsymbol{\theta}}(\boldsymbol{p})}{\partial x}, \frac{\partial\boldsymbol{\Phi}_{\boldsymbol{z},\boldsymbol{\theta}}(\boldsymbol{p})}{\partial y}, \frac{\partial\boldsymbol{\Phi}_{\boldsymbol{z},\boldsymbol{\theta}}(\boldsymbol{p})}{\partial z}\right)$ of the DVF $\boldsymbol{\Phi}_{\boldsymbol{z},\boldsymbol{\theta}}$, which is multiplied by a constant κ_r as

$$R(\boldsymbol{\Phi}_{\boldsymbol{z},\boldsymbol{\theta}}) = -\kappa_r \sum_{\boldsymbol{p}\in\Omega} \left\|\nabla\boldsymbol{\Phi}_{\boldsymbol{z},\boldsymbol{\theta}}(\boldsymbol{p})\right\|_2. \tag{4.11}$$

A *segmentation* regularization term is added to improve the overlap between propagated and ground-truth structures, using the DICE score (defined between 0 and 1,

where 1 denotes perfect overlap). For K structures, let $\hat{\boldsymbol{s}}_y^k$ be the voxels in the ground-truth scan with structure number $k \in [1, K]$, $\boldsymbol{s}_y^k = \boldsymbol{\Phi}_{\boldsymbol{z},\boldsymbol{\theta}} \circ \boldsymbol{s}_x^k$ the predicted voxels with structure number k, and $|\hat{\boldsymbol{s}}_y^k|$ the cardinality of structure $\hat{\boldsymbol{s}}_y^k$, i.e, the number of elements in $\hat{\boldsymbol{s}}_y^k$. The DICE score is defined as

$$\text{DICE}(\hat{\boldsymbol{s}}_y^k, \boldsymbol{s}_y^k) = 2\frac{\left|\hat{\boldsymbol{s}}_y^k \cap \boldsymbol{s}_y^k\right|}{\left|\hat{\boldsymbol{s}}_y^k\right| + \left|\boldsymbol{s}_y^k\right|}. \tag{4.12}$$

With these two terms multiplying the likelihood in the lower bound of Equation 4.10, the final optimization problem becomes

$$\begin{aligned}\boldsymbol{\theta}^*, \boldsymbol{\phi}^* = \underset{\boldsymbol{\theta},\boldsymbol{\phi}}{\text{argmin}}\, \mathbb{E}_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{s}_x,\boldsymbol{s}_y \sim P_D(\boldsymbol{x},\boldsymbol{y},\boldsymbol{s}_x,\boldsymbol{s}_y)} \Big[\mathbb{E}_{\boldsymbol{z}\sim Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{s}_x,\boldsymbol{y},\boldsymbol{s}_y)} \Big[-\lambda \text{CC}(\hat{\boldsymbol{y}}, \boldsymbol{y}) \\ -\frac{1}{K}\sum_{k=1}^{K} \text{DICE}(\hat{\boldsymbol{s}}_y^k, \boldsymbol{\Phi}_{\boldsymbol{z},\boldsymbol{\theta}} \circ \boldsymbol{s}_x^k)) + \kappa_r \sum_{\boldsymbol{p}\in\Omega} \left\|\nabla \boldsymbol{\Phi}_{\boldsymbol{z},\boldsymbol{\theta}}(\boldsymbol{p})\right\|_2 \Big] \\ + D_{KL}(Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{s}_x,\boldsymbol{y},\boldsymbol{s}_y)||P_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{s}_x)) \Big],\end{aligned} \tag{4.13}$$

with $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{s}_x$ and $\boldsymbol{s}_y$ sampled from the real data distribution $P_D(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{s}_x, \boldsymbol{s}_y)$.

4.2.2. Dataset

To learn the model parameters in a training stage, a dataset with 369 CTs from 40 prostate cancer patients is used, including prostate, seminal vesicles, bladder and rectum delineations with no overlap. For each of the patients, 3-11 repeat CTs were recorded at different points during their treatment at 3 different institutions: Erasmus University Medical Center (Rotterdam, Netherlands), Haukeland Medical Center (Bergen, Norway) and the Netherlands Cancer Institute (Amsterdam, Netherlands) (Deurloo et al., 2005; Sharma et al., 2012; Xu et al., 2014). In total, 329 planning-repeat CT pairs are available, 312 of which are used for training and validation, while the remaining 22 CTs — corresponding to 2 independent test patients, as in previous studies (Budiarto et al., 2011) — serve to evaluate performance on unseen geometries. After rigidly aligning each repeat to the planning CT, all volumes are cropped to a region of $64 \times 64 \times 48$ voxels around the prostate with a voxel resolution of 2 mm, resulting in sub-volumes of $128 \times 128 \times 96$ that in all cases covers the prostate, seminal vesicles, rectum and a large portion the bladder. These sub-volumes are $\boldsymbol{x} \in \mathbb{R}^{64\times64\times48}$ and $\boldsymbol{y} \in \mathbb{R}^{64\times64\times48}$ with the original CT intensity values rescaled to the range [0,1], and $\boldsymbol{s}_x \in \mathbb{R}^{64\times64\times48}$ and $\boldsymbol{s}_y \in \mathbb{R}^{64\times64\times48}$ with categorical labels depending on the organ present in each voxel. Given the stochasticity in the density of the rectum fillings, all voxels in the rectum are masked, setting their intensity to -1000 (vacuum).

4.2.3. Model architecture

As shown in Figure 4.2, the proposed variational framework comprises two different models, parametrized by artificial neural networks: the Inference network and the probabilistic U-net with down-sampling and up-sampling paths denoted as Encoder

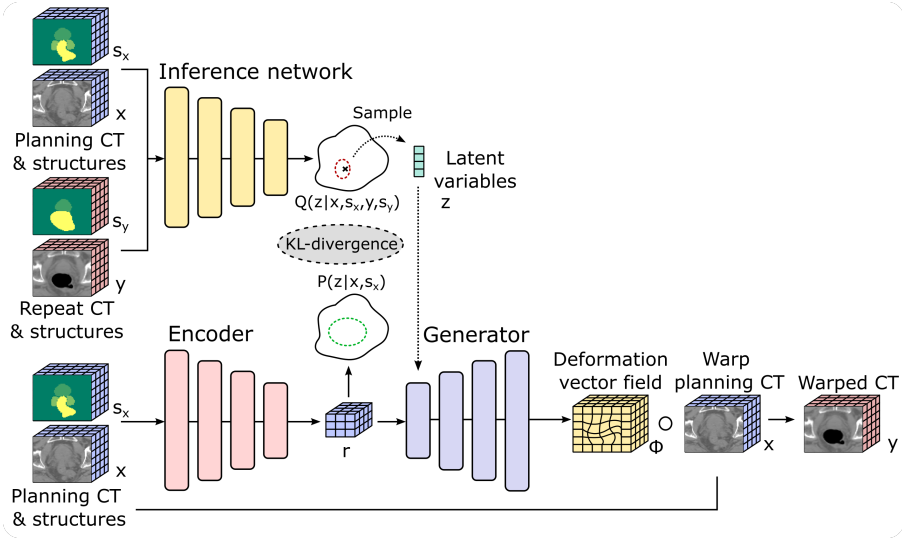


Figure 4.2: **Learning the model parameters**. An additional Inference Network takes a pair of planning and repeat CT and outputs the parameters of a distribution over a smaller region of the latent space that is likely to capture the deformation between the two images. The prior distribution predicted by the Encoder is forced to the distribution produced by the Inference Network via a KL-divergence term in the loss. Additionally, a reconstruction term encourages the resulting artificial CT (obtained after warping the planning scan with the predicted deformation) to be similar to real repeat CT.

and Generator, respectively. Based on the input planning CT and structures, the Encoder computes (i) a low-dimensional volume of input image features $\boldsymbol{r}$, and (ii) the parameters $\boldsymbol{\mu_\theta}$ and $\boldsymbol{\Sigma_\theta}$ of the prior distribution $P_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{x},\boldsymbol{s}_x)$ over a region of the latent space containing movements that are likely to be observed for the patient. The prior depends on the input, thus one of the functions of the Encoder is selecting primary groups of movements for each patient based on planning CT anatomy. The Generator takes the features $\boldsymbol{r}$ and sampled latent variables $z \sim P_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{x},\boldsymbol{s}_x)$ and produces the velocity field $\boldsymbol{v}_{\boldsymbol{z},\boldsymbol{\theta}}$ that is exponentiated to obtain a diffeomorphic transformation $\boldsymbol{\Phi}_{\boldsymbol{z},\boldsymbol{\theta}}$.

During training, the Inference network takes a pair of planning and repeat CTs and outputs the parameters $\boldsymbol{\mu_\phi}$ and $\boldsymbol{\Sigma_\phi}$ of the distribution $Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{s}_x,\boldsymbol{y},\boldsymbol{s}_y)$ over a much smaller region of the latent space containing latent variables that explain the deformation between both images. The DVF resulting from such latent variables is used to warp the planning CT into artificial repeat CTs $\boldsymbol{y}$ and structures $\boldsymbol{\Phi} \circ \boldsymbol{s}_x$. The distributions $Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})$ from the Inference network and $P_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$ from the Encoder are forced to overlap via the KL divergence in Equation 5.5, while the artificial CT and structures are forced to match the ground-truth repeat CTs via the CC and DICE terms in the likelihood.

For the model with the lowest validation loss, the Encoder and Inference network are identical: three consecutive convolutional blocks, where each block contains a 3D convolutional layer with 32 channels and a $3 \times 3 \times 3$ kernel followed by Group Normalization (Y. Wu and He, 2020), a rectified linear (ReLU) activation and a max pooling

down-sampling operation. At the lowest level, an additional 3D convolution with 4 channels results in the low-dimensional feature volume $\boldsymbol{r} \in \mathbb{R}^{4\times8\times8\times6}$, which is mapped to the means and variances of the prior distribution via two different fully-connected layers. Conversely, the Generator first concatenates the latent variables to $\boldsymbol{r}$ as an additional channel, and then applies three up-sampling convolutional blocks with 32 channels. Two additional 3D convolution operations with 16 and 3 channels result in the final velocity field $\boldsymbol{v}_{\boldsymbol{z},\boldsymbol{\theta}}$. All models are trained for 1000 epochs using a learning rate of 0.001, hyper-parameters $\kappa = 0.1$ and $\lambda = 1000$, and the Adam optimizer (Kingma and Ba, 2017) with default parameters.

4.2.4. Experiments

DAM is evaluated in terms of accuracy in both generating feasible groups of deformations and reconstructing the ground-truth repeat scans. Additional experiments aim at exploring the structure of the latent space and the types of movements triggered by different latent variables.

- Reconstruction accuracy. Given a planning and one of its repeat CTs in the test set, the Inference network can be used to obtain the latent variables corresponding to the deformation between both images, which are in turn used to get the DVF and warp the planning scan. For all 22 test planning/repeat pairs, the generated repeat CTs are compared to the ground truth repeats via computing the CC (Equation 4.8) and the DICE score (Equation 4.12). Additionally, after warping points $\boldsymbol{\pi}_i \in \mathbb{R}^3$ on the surface of the planning prostate, their distance to corresponding points $\hat{\boldsymbol{\pi}}_i \in \mathbb{R}^3$ on the surface of the repeat prostates is assessed via the mean surface error as

$$\boldsymbol{e} = \frac{1}{L}\sum_{i=1}^{L} \|\hat{\boldsymbol{\pi}}_i - \boldsymbol{\Phi} \circ \boldsymbol{\pi}_i\|_2 \,. \tag{4.14}$$

  To allow for a fair comparison with PCA-based methods, we can calculate the mean and standard deviation across the same number $L = 5864$ of randomly chosen points as in previous studies (Budiarto et al., 2011). Finally, the effect of the latent space dimensionality is evaluated by comparing all accuracy metrics for different models trained with a varying number of latent variables.

- Generative performance. To finally be applied in clinical settings, the generated movements must match those from the recorded CT scans. Based on a previous study quantifying anatomical changes in prostate patients (Antolak et al., 1998), the volume changes and center of mass shifts between planning and repeat scans are compared to their corresponding 'ground-truth' distributions obtained using real and artificial repeat CTs. To be able to compare to the reference values (Antolak et al., 1998), center of mass shifts are reduced to a single value by computing the average of absolute differences across coordinates.

- Latent space analysis. By individually varying the values of each latent variable while keeping the other fixed, we can numerically and visually assess the volume

changes and center of mass shifts triggered by each variable. Finally, to understand the structure of the latent space, the latent variables from all pairs in the dataset are obtained and classified according to the magnitude of their induced center of mass shifts and volume changes. Ideally, similar latent variables should correspond to similar deformations. which can be verified by plotting a 2D representation of the N latent variables using t-SNE (Maaten and Hinton, 2008) together with their associated label to determine the presence of clusters.

4.3. Results

This section assesses DAM's performance in generating realistic CTs with anatomical changes that match those of the real recorded repeat CTs. First, the reconstruction accuracy of real CTs is assessed, followed by an analysis of the latent space, and the types of deformations captured by the latent variables.

4.3.1. Reconstruction accuracy

Given a planning-repeat pair of CT scans and structures in the test set, a repeat scan can be reconstructed via the same framework as used during training: sampling latent variables with the Inference network that are used by the Generator to generate a DVF. To verify the similarity between DAM's reconstructions and the real repeat CTs, three metrics assessing CT and structure overlap are computed: the CC, DICE score, and surface error $\boldsymbol{e}$. All three metrics in Figure 4.3 are computed for different models trained with a varying number of latent variables, from 1 to 32. The values shown for 0 latent variables correspond to using the planning CTs as a prediction, which is equivalent to disregarding any model. First, the cross correlation between the real and reconstructed repeat CT is shown in the left plot of Figure 4.3, indicating that the model significantly improves when adding the first few variables, whereas no substantial is observed beyond 10 variables. As seen in DICE scores for the prostate and rectum from the middle plot in Figure 4.3, DAM can model prostate deformations with high accuracy even with a single latent variable, while representing rectum movements generally requires a slightly larger latent space with ≈ 8 variables. The relative simplicity in capturing prostate movements is further confirmed from the right plot in Figure 4.3, showing that most surface error (Equation 4.14) reduction results from adding the first latent variable. On average, DAM matches — and even outperforms in the low-dimensional regime — the accuracy of countour-based PCA models (Budiarto et al., 2011). The larger spread in error values is likely caused by the fact that, unlike for the values reported in the PCA study, all surface points are not equidistant but randomly sampled over the surface, increasing the distance between correspondent points in under-sampled areas.

4.3.2. Generative performance

Besides generating realistic CT scans, DAM should produce patient-specific movements whose distribution approximately matches those observed in the clinics, as reported in previous work (Antolak et al., 1998). For the 2 test patients, Figure 4.4 displays the distribution of the anatomical variations seen in the 11 recorded repeat CTs (blue), compared to the deformations seen in 100 randomly sampled CTs (orange). Except for the

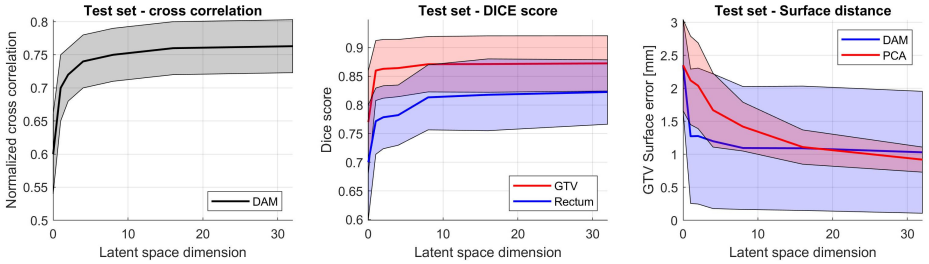
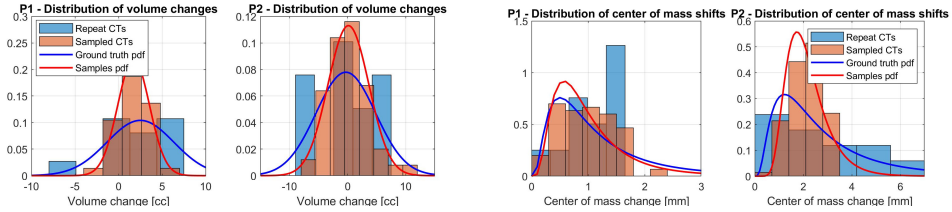


Figure 4.3: **Reconstruction accuracy metrics**. All figures show the mean (solid line) and standard deviation across all test planning-repeat pairs of the different metrics for a different number of latent variables, where 0 latent variables refers to using no model (always using the planning CT as a prediction). The left plot shows the cross-correlation between the real and reconstructed repeat CTs. In the middle plot, the DICE score is shown, measuring overlap between the warped planning structures and the organs delineated in the repeat CTs. Finally, the right figure shows the error between surface points in the prostate, compared to reference PCA values directly taken from (Budiarto et al., 2011).

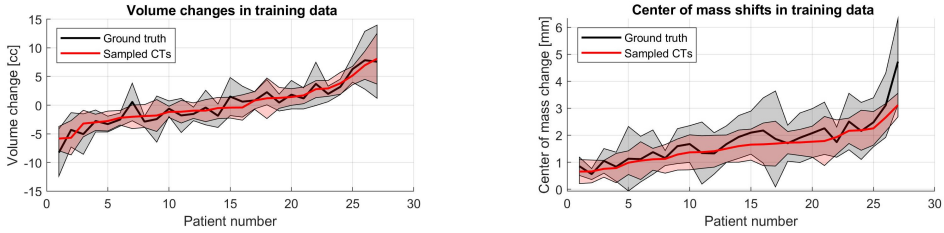


(a) Volume changes.

(b) Center of mass shifts.

Figure 4.4: **Test set histograms of anatomical variations**. For the two independent test patients, histograms of prostate (a) volume changes and (b) center of mass shifts are plotted. Blue histograms correspond to changes between the planning CT and the 11 available repeat CTs, additionally showing their corresponding fitted normal and log-normal distributions in the same colors. Orange histograms are calculated using 100 randomly sampled CTs, obtained from 100 different latent variable combinations.

large center of mass movements seen for the second patient Figure 4.4b, the ranges of values for both volume changes in Figure 4.4a, and center of mass shifts in Figure 4.4b are approximately equal. Similarly, Figure 4.5 shows the center of mass shift and volume changes distributions for all training patients with more than 5 repeat CTs. To compress all the information into one plot, the mean and standard deviation are plotted instead of the full histogram. The good overlap between distributions demonstrates that DAM captures the correct frequency and range of movements. As for the test patients, the biggest differences between both distributions occur for the last patient in Figure 4.5b with large center of mass shifts, which is aggravated by the fact that this patient has three big outliers of >7 mm shift. Finally, Figure 4.6 displays generated and real anatomies for one of the patients, showing high quality images and contours with similar features and shapes.



(a) Volume changes.

(b) Center of mass shifts.

Figure 4.5: **Training set distribution of anatomical variations**. For all the patients in the training set with 5 or more repeat CTs, the mean (solid line) and standard deviation of prostate (a) volume changes and (b) center of mass shifts are plotted. Black lines are computed using the available planning-repeat pairs of CT. The red curves are calculated using 100 randomly sampled CTs, obtained from 100 different latent variable combinations.

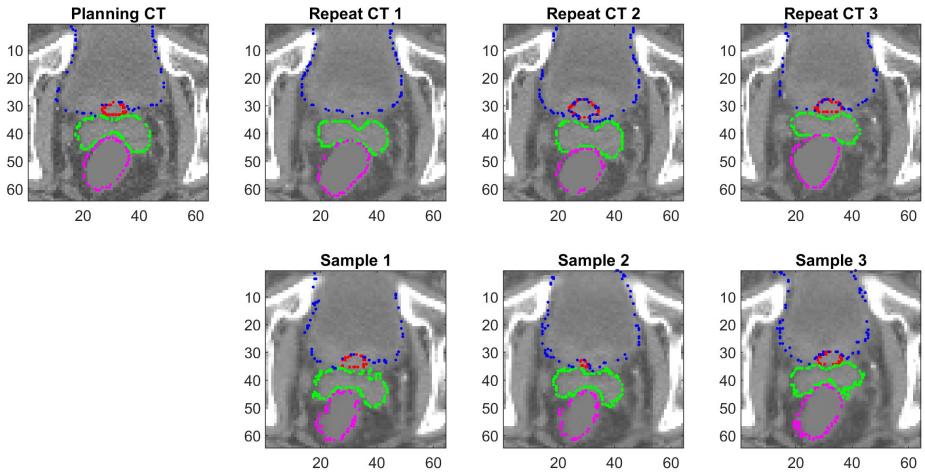


Figure 4.6: **Real vs. sampled anatomies**. Three recorded repeat CTs (top row), and three anatomies generated by the model (bottom row) are shown for one of the planning CTs, including prostate (red), seminal vesicles (green), bladder (blue) and rectum (pink) contours. The images correspond to a perpendicular slice in the cranial-caudal axis, showing the top of the prostate.

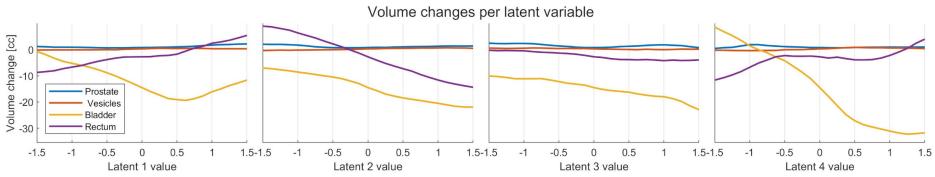
4.3.3. Latent space analysis

To investigate the deformations captured by the latent variables, we can compute the center of mass shifts and volume changes triggered by each variable independently, while keeping the rest fixed. Figure 4.7 displays such changes for 4 randomly picked variables from the model with 8 latent variables, whose value was modified between -1.5 and 1.5 times the standard deviation of the prior distribution. The results show magnitudes and correlations between changes as can be expected: smaller prostate volume changes, and large bladder and rectum variations shifting the center of mass of the prostate and seminal vesicles. To further demonstrate DAM's learned correlated groups of movements, Figure 4.8 shows a grid of structures corresponding to simultaneously varying two latent variables. Individual changes in the horizontal and vertical axis mainly control the bladder and rectum volumes, respectively. Correlated deformations arise: the increase of bladder volume above the seminal vesicles, together with the decrease of rectum filling below the prostate, cause a prostate and vesicles shift and rotation.

The structure of the latent space can be analyzed by determining if similar deformations (shifts and volume changes) or anatomical features (organ volume) result in similar latent variables. Figure 4.10 shows a two-dimensional t-SNE representation of the latent variables, where only samples with the smallest and largest movements or volumes are included, i.e., samples whose with center of mass shifts or volumes that fall above the 90% percentile or below the 10% percentile. Most of the latent space information seems to concern center of mass shifts and bladder/rectum volume changes, since their 2D representations can be clearly separated. Ideally, similar latent variables that are clustered together will correspond to different anatomical deformations, and will not carry information about anatomical features of the patient such as absolute organ volume. Instead, the Encoder is in charge to mapping deformations to anatomical traits observed in the planning CT or structures. Prostate and bladder volume seem to have no effect in how the latent space is organized, since similar latent variables correspond to very different sizes. To some extent, the effect of rectum size is also limited, resulting from the possible correlation between rectum fillings and volume changes.

4.4. Discussion

This chapter presents a probabilistic framework to model patient-specific inter-fraction movements based on population data. The presented DAM captures deformation patterns, generating DVFs only based on the planning CT scan and delineations. Based on the metrics obtained in Figure 4.3 for the 22 scans from two independent test patients, DAM can generate realistic CTs with anatomical variations that resemble those recorded in the clinics using a small number of latent variables. The structure overlap of a model with a single variable, measured as a DICE score of 0.856 ± 0.058, agrees with that of previous state-of-the-art pure segmentation/registration (non-generative) deep learning studies (Elmahdy et al., 2019, 2021; Liang et al., 2021; Yuan et al., 2019). Compared to linear PCA models where each eigenvector captures an independent mode of motion, the non-linearities in DAM allow representing different groups of correlated movements using different values of only one latent variable. Given that a single latent variable practically suffices to capture prostate movements, and that both the



(a) Volume changes.

(b) Center of mass shifts.

Figure 4.7: **Effect of individual latent variables on deformations**. (a) Volume changes and (b) center of mass shifts triggered by independently varying latent variables. For a model with $N = 8$ latent variables, four randomly selected variables are varied between values within -1.5 and 1.5 of their standard deviation, while keeping the remaining seven variables fixed and equal to their mean.

CC and rectum DICE score keep increasing with larger latent spaces, we can conclude that most of the computational effort is dedicated to modeling rectum deformations. Prostate IMPT treatments typically solely involve lateral beams, for which the impact of error due to rectum movement is small. In some cases, models with as little as 4-8 variables may be accurate enough, while 8-16 variables additionally ensure accurate rectum deformations for plans requiring more precision.

For clinical application, it is critical that the model generates realistic shifts and deformations of the volume to be irradiated/treated (in this case, the prostate). Overall, based on the results in Figure 4.4 and Figure 4.5, the center of mass shifts and volume changes produced by DAM show good overlap to the deformations and shifts recorded in the clinic, matching previously reported values (Antolak et al., 1998). One reason why DAM struggles in simulating the most extreme shifts or slides is the regularization term of the loss, which limits large deformations. Despite this limitation, such large anatomical variations are typically taken care of by adapting the treatment plan to the new anatomy, whereas robust treatment planning and evaluation — the main potential applications of DAM — are in principle oriented to incorporating average, frequent deformations into treatment design and evaluation, and DAM is expected to be useful for such purposes.

Comparison to other methods All the previously published approaches are either patient-specific or population models based on PCA. Patient-specific methods (Nie et al., 2012; Söhn et al., 2005; Thörnqvist, Hysing, Zolnay, Söhn, Hoogeman, Muren, Bentzen, and Heijmen, 2013; Thörnqvist, Hysing, Zolnay, Söhn, Hoogeman, Muren, and Heijmen, 2013; Q. Zhang et al., 2007) require at least a few CTs recorded during a patient's treatment, and therefore they are unfeasible for pre-delivery robust treatment planning and evaluation, being restricted to post-treatment analysis. Conversely, pop-

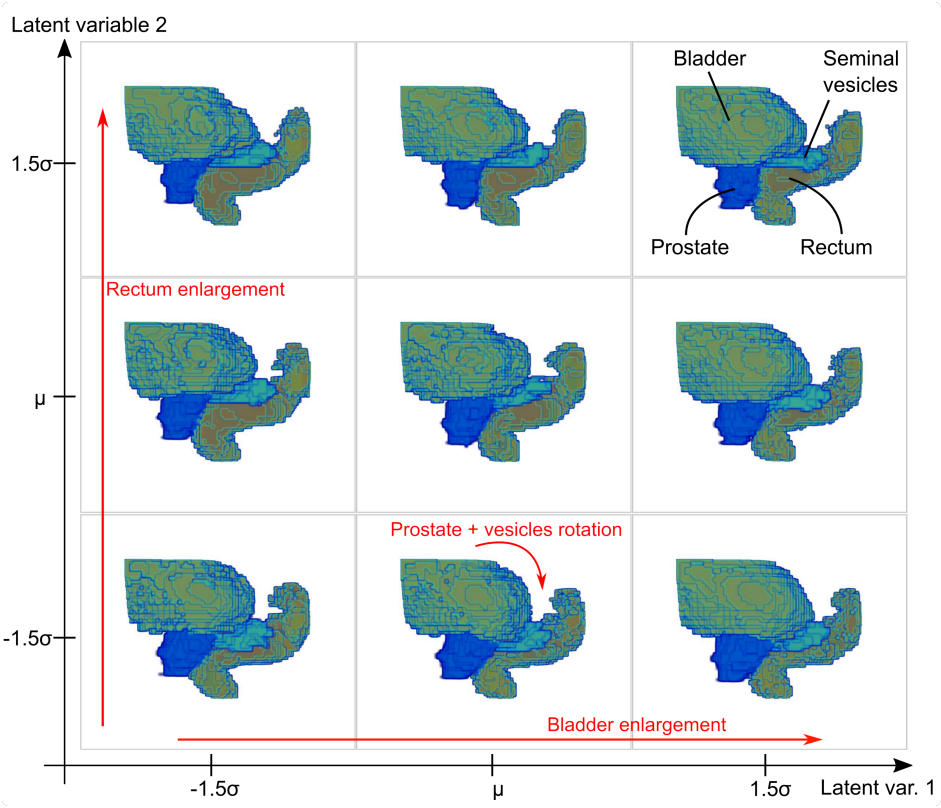
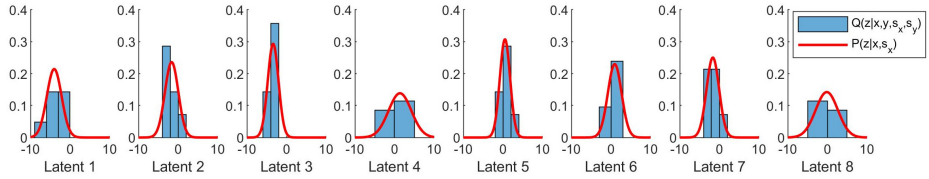
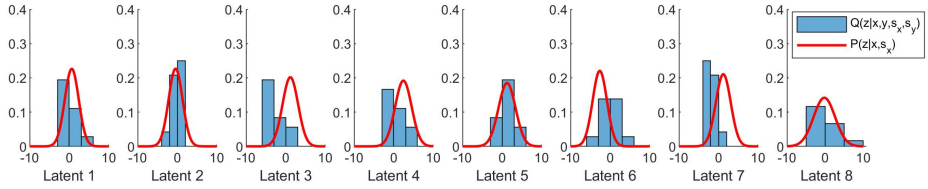


Figure 4.8: **Latent space visualization**. Grid plot of the prostate (blue), seminal vesicles (green), bladder (yellow) and rectum (orange) volumes. Each box corresponds to a different combination of latent variables in a 2D plane of the latent space, where the values for each variable are shown on the axes, with σ being the standard deviation and μ the mean. Changes in the horizontal axis translate into bladder enlargements, while the vertical axis controls rectum volume. Correlated groups of movements are observed, e.g., as prostate rotations triggered by an enlarged bladder and smaller rectum.



(a) Training set patient.



(b) Test set patient.

Figure 4.9: **Prior - posterior distribution overlap**. For (a) a training patient, and (b) a test patient, the prior Gaussian probability density function $P(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{s}_x)$ is compared to a normalized histogram of samples from the posterior distribution $Q(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y},\boldsymbol{s}_x,\boldsymbol{s}_y)$. The parameters of the prior distribution are obtained from the Encoder, given a planning CT and structure volume. Histograms are obtained by sampling once each posterior distribution corresponding to each of the planning-repeat pairs available for both patients.

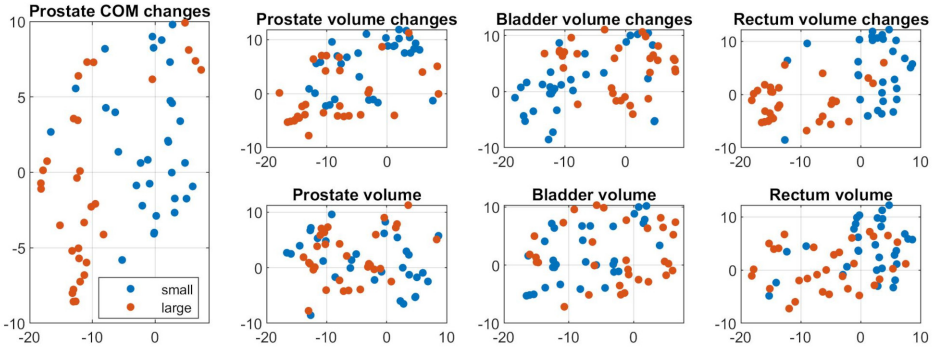


Figure 4.10: **Latent space structure**. Each latent variable is reduced to 2D space t-SNE representation and classified, from left to right, according to whether they correspond to small (blue) or large (orange): prostate center of mass shifts, prostate, bladder and rectum volume changes, or prostate, bladder and rectum sizes. 'Small' samples fall below the 10% percentile of all values, while 'large' samples include all values above the 90% percentile.

ulation models (Budiarto et al., 2011; Magallon-Baro et al., 2019; Rios et al., 2017; Szeto et al., 2017; Tilly et al., 2017) use a set of planning-repeat CT/contour pairs from previous patients, but simulate the same type of deformations for all patients regardless of their anatomy. In contrast, as seen in Figure 4.4 and Figure 4.5, DAM is able to retrieve patient-specific magnitude and frequency of movements from the entire population based on the planning CT anatomy, making the model suitable for a wider range of applications.

Most previous studies (Budiarto et al., 2011; Magallon-Baro et al., 2019; Söhn et al., 2005; Thörnqvist, Hysing, Zolnay, Söhn, Hoogeman, Muren, Bentzen, and Heijmen, 2013) model only the surface of the organs and not the intensities values in the CT. Without CT values the dose distributions are always calculated on the same planning CT with varying contours, which limits its applicability, especially in IMPT given the protons' finite range and tissue sensitivity. Conversely, PCA-based models modeling full DVFs require 7 (Tilly et al., 2017) or up to 100 principal components (Szeto et al., 2017) to capture 90% of the variance in the training data. A large number of components (equivalent to DAM's latent variables) adds more variation, increases the chance of sampling unrealistic deformations and limits their applicability as reduced order models. Most importantly, all previous population-based methods require a time-consuming pre-processing step involving multiple deformable image registration steps between scans and patients to an organ or CT template. The accuracy of such registration calculation degrades the final accuracy and generative performance of the model, with previous studies (Szeto et al., 2017; Tilly et al., 2017) showing surface errors of around 1.5 ± 1.0 mm introduced in their pre-processing step alone that are comparable the DAM's total errors reported in the right plot of Figure 4.3. Given the lack of uniformity in treatment site and evaluation metrics in previous studies – where most focus on evaluating the variance captured by the PCA model components and the errors on the DVFs caused by truncating the number of eigenmodes — DAM is compared to a PCA model of the prostate (Budiarto et al., 2011) in the right plot of Figure 4.3. Even without adding any pre-processing errors, DAM matches the overall performance and is to capture prostate motion with a lower number of modeling parameters. Being trained directly on CT images in an unsupervised manner, DAM bypasses any performance or time losses from any pre-processing step, and can be easily applied to generate new anatomies in few milliseconds, compared to the tens of minutes or hours needed to obtain accurate enough registrations using conventional clinical software.

Like PCA-based models, DAM assigns realistic correlated deformations to different values of the latent variables. Figure 4.7 and Figure 4.8 show that variables control shifts, volume changes and rotations similar to those reported in previous studies (Budiarto et al., 2011; Magallon-Baro et al., 2019). Figure 4.10 demonstrates that the latent variables almost exclusively carry information about deformations, and not about anatomical traits from the patients. Instead, the Encoder is in charge of independently mapping planning anatomies to a subset of latent variables. Furthermore, unlike all previous approaches not evaluating the generative performance of their proposed models, this chapter demonstrates the DAM also generates the adequate range and frequency of deformations for each patient.

Limitations Like PCA-based models, DAM will struggle to generate deformations that are not represented in the training data, for which continuously updating the model (e.g, using cone beam CTs) can be a solution. Likewise, low resolution images with poor contrast can also affect performance by masking small movements of structures, especially in areas with similar organ tissue densities. DAM's implementation in the clinic thus requires a quality assurance protocol that evaluates robustness in predictions e.g., by training several models using different data, and evaluating result similarity on a same test dataset.

As for many other deep learning algorithms, DAM's generalization capabilities depend on the size and variability of the data in the dataset, as well as on the quality and resolution of the CT images. Due to the rather small size of the dataset in this preliminary study — caused by the scarcity of recorded sets of planning and repeat CTs — and based on the initial positive results, further testing appears warranted.

DAM's accuracy in generating reasonable patient-specific movements depends on the extent to which movements can be predicted only from the planning CT and structures. As with other classical and deep learning registration algorithms, DAM would struggle to register rectum structures due to the randomness in their intensity values. Following clinical practice, the rectum voxels are masked with air. As a result, all deformed CTs have air-filled rectum structures, which can affect the accuracy in the dose calculation, especially for beams delivered in the anterior-posterior direction. Possible solutions include adding an additional generative model that generates rectum voxel intensities based on the organ mask shape.

4.5. Summary

This chapter presents DAM, a deep learning-based daily anatomy model to simulate patient-specific deformations that may be observed during the course of a prostate cancer radiotherapy treatment. DAM captures groups of correlated movements via few auxiliary latent variables, where few variables are able to model prostate deformations and shifts with similar accuracy as state-of-the-art models based on principal component analysis. Compared to previous population models, DAM can generate realistic CT images and contours in less than a second without any pre-processing, with volume changes and center of mass shifts that match in frequency and range those reported in the clinics and in previous studies. Given its simplicity and speed to generate CTs based on a single planning scan and delineations, DAM can be tested in treatment planning and evaluation to design treatment plans that are robust against inter-fraction variations.

5

Modeling and classifying intra-fraction breathing variations

5.1. Introduction

Biomedical data is the driving force behind most modern advances in medicine. The use of biomedical records is associated however with a series of problems such as the lack of reliable models capable of simulating data with clinical precision, the absence of personalized models for diagnosis, or the lack of labeled samples since the labels containing personal features that compromise privacy or simply are not recorded (Neal and Kerckhoffs, 2010). Some of the initial efforts to model biomedical data include analytical approaches: e.g., an electrocardiogram (ECG) model based on three coupled ordinary differential equations (McSharry et al., 2003), or sinusoidal model to represent breathing (George et al., 2005).

Recent advances in deep learning and the introduction of algorithms such as the variational autoencoder (VAE) (Kingma and Welling, 2014; Rezende et al., 2014) and generative adversarial networks (GANs) (Goodfellow et al., 2014) have resulted in a wide variety of methods capable of generating and classifying biomedical signals, most of them having been applied to ECG data. Most published works focus on classification algorithms, using convolutional neural networks (CNN) for computer aided diagnosis based on biomedical signals (Acharya, Fujita, Lih, et al., 2017; Acharya, Fujita, Oh, et al., 2017; Cimr et al., 2020; Fujita and Cimr, 2019; Ö. Yildirim et al., 2018), while some works propose autoencoder compression models using artificial neural networks (ANNs) (O. Yildirim et al., 2018). Recent CNN (Chen et al., 2021) and long short-term memory (LSTM) implementations O. Yildirim et al., 2019 have resulted in minimal classifica-

The contents of this chapter have been published as a journal paper in *Computer Methods and Programs in Biomedicine* 209, 106312 (2021), (Pastor-Serrano, Lathouwers, and Perkó, 2021).

tion error of arrhythmia in ECG signals. With respect to generation, most works propose generative models of realistic ECG signals combining recurrent and convolutional architectures under an adversarial training objective (Delaney et al., 2019; Golany and Radinsky, 2019; F. Zhu et al., 2019), with the exception of an auto-regressive model able that produce longer signals with high variability (Wulan et al., 2020). Practically all previously proposed methods focus either on generation or classification and result in models that depend on large labeled datasets and supervised training; are resource intensive and require significant amounts of computing power; are inaccurate when the dataset is imbalanced (i.e., there are very few labels for some classes of interest), or generate data that lacks variability and has a limited temporal dependence (S. Hong et al., 2020; C. Xiao et al., 2018). Furthermore, most of the approaches are not capable of compressing the data into a low-dimensional compact manifold in which specific regions correspond to similar samples.

In this chapter, we turn our attention to mechanical breathing signals representing the movement of chest markers during respiration. Among their many applications, these type of biomedical signals are of great importance in radiotherapy cancer treatments, where they are used to quantify the impact of respiration and to design robust lung cancer radiotherapy treatments that withstand the detrimental effect of breathing motion during treatment delivery. Among the most important breathing irregularities are baseline shifts, which are gradual or sudden changes in the exhale position and trend of respiration. Baseline shifts negatively affect the outcome of radiotherapy treatments (Takao et al., 2016). To date, there are no previous studies that develop breathing generative models resulting in realistic respiratory traces. Likewise, only one classification autoencoder framework that discriminates between apnea and regular breathing, has been proposed for radiotherapy treatments (Abreu et al., 2020).

This chapter introduces a semi-supervised framework that simultaneously classifies and generates breathing motion with high accuracy using a small subset of labeled data, outperforming purely discriminative models. The main contributions of this research are threefold. First, the suitability of probabilistic generative models for the task of modeling breathing signals is investigated. Second, building upon these breathing models, a modified semi-supervised algorithm is proposed to train a joint generative-discriminative model with a partially-labeled dataset. The proposed model can be used, e.g, to simultaneously generate and classify samples of irregular breathing such as baseline shifts. Third, a method to pre-process and post-process breathing signals is presented, transforming back and forth the breathing signals from their original 3D time series form into a simplified vector form containing pairs of position-time values. This transformation significantly reduces the dimensionality of the inputs and speeds up training.

5.2. Semi-supervised probabilistic models

Consider $\boldsymbol{x} \in \mathbb{R}^M$ to be a random vector over a vector space $\mathscr{X}$, with unknown underlying probability distribution $P^*(\boldsymbol{x})$. Given a dataset $\{\boldsymbol{x}^{(i)}\}_{i=1}^{N_D}$ with N_D independent and identically distributed (i.i.d) data points, the goal is to model a probability distribution $P_{\boldsymbol{\theta}}(\boldsymbol{x})$ that approximates the unknown true probability distribution generating the

data using a probabilistic graphical model with parameters $\boldsymbol{\theta}$. Let this probabilistic model be a latent variable model, which conditions the observed variable $\boldsymbol{x}$ on the unobserved random variable $\boldsymbol{z} \in \mathbb{R}^N$ over the latent space $\mathcal{Z}$ containing N latent variables that are assumed to capture the principal factors of variation in the data. The latent variable model represents the joint distribution of observed and unobserved variables and factorizes as $P_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z}) = P_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})P(\boldsymbol{z})$. The (target) marginal distribution of the observed variables can be recovered as

$$P_{\boldsymbol{\theta}}(\boldsymbol{x}) = \int P_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z} = \int P_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})P(\boldsymbol{z}) d\boldsymbol{z}, \tag{5.1}$$

where $P(\boldsymbol{z})$ is the prior probability distribution over $\mathcal{Z}$ and $P_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$ is a conditional distribution that can be parametrized using neural networks. In principle, the prior could be any function not conditioned on the observations. Point-estimates of the parameters $\boldsymbol{\theta}$ of the latent variable model can be obtained via maximum likelihood estimation, i.e., by maximizing the (log-) marginal distribution of the observed data

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{\boldsymbol{x} \in D} \log\left(P_{\boldsymbol{\theta}}(\boldsymbol{x})\right) \simeq \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathbb{E}_{\boldsymbol{x} \sim P_D(\boldsymbol{x})} \log(P_{\boldsymbol{\theta}}(\boldsymbol{x})), \tag{5.2}$$

where the expected value is computed over the empirical data distribution $P_D(\boldsymbol{x})$. The empirical data distribution is different from the true underlying data generating distribution $P^*(\boldsymbol{x})$ to be approximated that cannot be accessed. $P_D(\boldsymbol{x})$ is defined as a mixture of Dirac delta distributions $\delta(\boldsymbol{x})$ that assigns probability mass $1/N_D$ to each data point in the dataset as

$$P_D(\boldsymbol{x}) = \frac{1}{N_D} \sum_{i=1}^{N_D} \delta(\boldsymbol{x} - \boldsymbol{x}^{(i)}). \tag{5.3}$$

In practice, computing the integral over the space $\mathcal{Z}$ in Equation 5.1 is intractable. Thus, the optimization in Equation 5.2 is simplified by maximizing a lower bound on the marginal distribution.

5.2.1. Variational autoencoder

Using variational inference (Kingma and Welling, 2014; Rezende et al., 2014), the latent variable model parameters can be estimated by maximizing a lower bound on Equation 5.2. The VAE algorithm requires an inference model that approximates the (also) intractable true posterior distribution $P_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$ using a family of probability distributions of the latent variables $Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ conditioned on observed data points, parametrized by an ANN with parameters $\boldsymbol{\phi}$ shared across data points $\boldsymbol{x}$. By including the inference model, the lower bound optimization objective is formulated as

$$\log\left(P_{\boldsymbol{\theta}}(\boldsymbol{x})\right) \geq \mathbb{E}_{\boldsymbol{z} \sim Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}[\log(P_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}))] - D_{KL}(Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||P(\boldsymbol{z})), \tag{5.4}$$

where the second term is the Kullback - Leibler (KL) divergence, denoted $D_{KL}(\cdot||\cdot)$. Essentially, the KL divergence quantifies "the difference" between distributions. Further details about the lower bound and how to compute the KL-divergence are included in Appendix A.

In the VAE framework, the prior is the multivariate Gaussian $P(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{I})$, where $\boldsymbol{I}$ is the identity matrix. The likelihood conditional distribution $P_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$ is represented as a multivariate Gaussian probability distribution with identity covariance $P_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; f_{\boldsymbol{\theta}}(\boldsymbol{z}), \boldsymbol{I})$, where the function $f_{\boldsymbol{\theta}}(\boldsymbol{z}) : \mathcal{Z} \rightarrow \mathbb{R}^M$ is parameterized with an ANN referred to as the probabilistic decoder. With this formulation, $P_{\boldsymbol{\theta}}(\boldsymbol{x})$ is an infinite mixture of Gaussian distributions. In the same way as with the probabilistic decoder, we can parameterize the inference model conditional distribution using a neural network that performs a mapping $g_{\boldsymbol{\phi}}(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X} \rightarrow (\boldsymbol{\mu}(\boldsymbol{x}), \boldsymbol{\sigma}(\boldsymbol{x})) \in \mathbb{R}^{2N}$ and outputs the mean $\boldsymbol{\mu}(\boldsymbol{x})$ and standard deviation $\boldsymbol{\sigma}(\boldsymbol{x})$ of the normal distribution $Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}(\boldsymbol{x}), \text{diag}\, \boldsymbol{\sigma^2}(\boldsymbol{x}))$.

The lower bound balances two terms: the first term encourages the probabilistic decoder to produce samples that resemble the observed data, while the second term forces the approximated posterior distribution obtained from the inference model to be close to the prior distribution. Using the negative lower bound as optimization objective, the minimization problem becomes

$$\boldsymbol{\theta}^*, \boldsymbol{\phi}^* = \underset{\boldsymbol{\theta}, \boldsymbol{\phi}}{\text{argmin}}\, \mathbb{E}_{\boldsymbol{x} \sim P_D(\boldsymbol{x})} \big[-\mathbb{E}_{\boldsymbol{z} \sim Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} [\log(P_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}))] + \lambda_{KL} D_{KL}(Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) || P(\boldsymbol{z})) \big], \quad (5.5)$$

where λ_{KL} is a hyperparameter that can be used to weigh the reconstruction and regularization terms (Higgins et al., 2022). The minimization in Equation 5.5 can be performed using first order stochastic methods such as stochastic gradient descent (SGD). The reparametrization trick is usually employed to propagate the weights' gradients through the encoder (Kingma and Welling, 2014, 2019).

5.2.2. Adversarial autoencoder

In an alternative formulation to the lower bound, adversarial autoencoders (Makhzani et al., 2016) approximate the KL divergence with the optimal value of an adversarial loss forcing the aggregated posterior distribution $Q_{\boldsymbol{\phi}}(\boldsymbol{z})$ to be close to the prior distribution:

$$Q_{\boldsymbol{\phi}}(\boldsymbol{z}) = \int Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) P_D(\boldsymbol{x}) d\boldsymbol{x} \simeq P(\boldsymbol{z}). \quad (5.6)$$

In the original paper, the authors explore the use of both probabilistic encoders and deterministic encoders with $g_{\boldsymbol{\phi}}(\boldsymbol{x})$ as a deterministic mapping. The used probabilistic encoder can in principle learn any arbitrary posterior distribution by employing random noise $\eta \in \mathcal{H} \in \mathbb{R}$ with distribution $P(\eta) = \mathcal{N}(\eta; 0, 1)$. Such encoders take additional random noise values to produce samples $\boldsymbol{z} = g_{\boldsymbol{\phi}}(\boldsymbol{x}, \eta)$, and can use different noise values η to map the same input $\boldsymbol{x}$ to a sub-domain of $\mathcal{Z}$. The aggregated posterior can be computed as

$$Q_{\boldsymbol{\phi}}(\boldsymbol{z}) = \int \int \delta(\boldsymbol{z} - g_{\boldsymbol{\phi}}(\boldsymbol{x}, \eta)) P(\eta) P_D(\boldsymbol{x}) d\eta d\boldsymbol{x}, \quad (5.7)$$

The adversarial loss is based on GANs. Let the encoder network perform a mapping $g_{\boldsymbol{\phi}}(\boldsymbol{x}, \eta) : \mathcal{X} \times \mathcal{H} \rightarrow \mathcal{Z}$, via an ANN with parameters $\boldsymbol{\phi}$. A discriminator model is introduced, modeled also with an ANN with mapping function $d_{\boldsymbol{\xi}}(\boldsymbol{z}) : \mathcal{Z} \rightarrow \mathbb{R}$ that outputs a single scalar. The value $S(d_{\boldsymbol{\xi}}(\boldsymbol{z})) \in [0, 1]$ represents the probability that $\boldsymbol{z}$ is a sample

from the prior distribution $P(\boldsymbol{z})$ (true samples) rather than being a latent space mapping from the encoder (fake samples), where $S(z) := (1+e^{-z})^{-1}$ is the logistic sigmoid function. This translates into a min-max optimization problem

$$\min_{\boldsymbol{\phi}} \max_{\boldsymbol{\xi}} \mathbb{E}_{\boldsymbol{z}\sim P(\boldsymbol{z})}[\log(S(d_{\boldsymbol{\xi}}(\boldsymbol{z})))] + \mathbb{E}_{\boldsymbol{x}\sim P_D(\boldsymbol{x})}\mathbb{E}_{\eta\sim P(\eta)}[\log(1-S(d_{\boldsymbol{\xi}}(g_{\boldsymbol{\phi}}(\boldsymbol{x},\eta))))], \tag{5.8}$$

where first the discriminator is trained to correctly distinguish between real and encoder samples by maximizing the probability of classifying real samples from the prior $\boldsymbol{z}_r$ as real ($S(d_{\boldsymbol{\xi}}(\boldsymbol{z}_r)=1)$) and fake samples from the encoder $\boldsymbol{z}_f$ as false ($S(d_{\boldsymbol{\xi}}(\boldsymbol{z}_f)=0)$). Second, the encoder is trained to minimize the probability $1-S(d_{\boldsymbol{\xi}}(\boldsymbol{z}_f))$ that the discriminator identifies its samples $\boldsymbol{z}_f$ as fake, where $d_{\boldsymbol{\xi}}(\boldsymbol{z_f})=1$ means that the discriminator classifies a fake sample as a true sample. Training the probabilistic decoder $P_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$, the inference model $Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ and the discriminator $d_{\boldsymbol{\xi}}(\boldsymbol{z}_f)$ can be done with SGD in two alternating steps: a reconstruction phase forces the decoder to produce realistic samples by using the $\boldsymbol{z}_f$ variables produced by the inference model, and the regularization phase updating the parameters of the encoder and discriminator. As shown in Appendix B, optimizing the adversarial objective results in an approximation to the lower bound, where the regularization term $\mathbb{E}_{\boldsymbol{x}\sim P_D(\boldsymbol{x})}[D_{KL}(Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||P(\boldsymbol{z}))]$ in Equation 5.4 is replaced by $D_{KL}(Q_{\boldsymbol{\phi}}(\boldsymbol{z})||P(\boldsymbol{z}))$.

5.2.3. Joint generative-discriminative models

One of the advantages of the AAE algorithm is that the standard architecture can be slightly modified in order to additionally perform semi-supervised classification based on few labeled data points. The most notable difference with respect to the standard AAE architecture is the introduction of an extra discrete latent variable $\boldsymbol{c}\in\{0,1\}^C$, which represents the class to which the input belongs over C classes. The class $\boldsymbol{c}$ is practically implemented as a sparse one-hot vector with a 1 entry at the position corresponding to the class. In the case of breathing, the $\boldsymbol{c}$ variable could indicate the presence of irregularities or the patient to which breathing pertains. The encoder now outputs the joint distribution $Q_{\boldsymbol{\phi}}(\boldsymbol{c},\boldsymbol{z}|\boldsymbol{x})$ that factorizes as

$$Q_{\boldsymbol{\phi}}(\boldsymbol{c},\boldsymbol{z}|\boldsymbol{x}) = Q^c_{\boldsymbol{\phi}}(\boldsymbol{c}|\boldsymbol{x})Q^s_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}), \tag{5.9}$$

where $Q^c_{\boldsymbol{\phi}}(\boldsymbol{c}|\boldsymbol{x})$ is a categorical distribution that performs the mapping $S_{mx}(g^c_{\boldsymbol{\phi}}(\boldsymbol{x},\boldsymbol{\eta})) : \mathscr{X}\rightarrow[0,1]^C$ based on the input $\boldsymbol{x}$, with $g^c_{\boldsymbol{\phi}}(\boldsymbol{x},\eta)$ being a deterministic function. The use of the S_{mx}softmax non-linearity and one-hot vectors as a target forces sparsity in $Q^c_{\boldsymbol{\phi}}(\boldsymbol{c}|\boldsymbol{x})$. The Gumbel-softmax reparametrization trick (Jang et al., 2017; Maddison et al., 2017) is used to back-propagate the gradients through the categorical distribution. As in the standard AAE, the approximate posterior $Q^s_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\eta})$ is a probabilistic mapping based on noise η. In order to simultaneously classify and generate new samples given a specific input, the proposed modified AAE architecture uses a single discriminator for both the classification and style heads. In this way, the aggregated approximated posterior is forced to match the mixture prior distribution

$$Q_{\boldsymbol{\phi}}(\boldsymbol{z},\boldsymbol{c}) = \int Q^s_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})Q^c_{\boldsymbol{\phi}}(\boldsymbol{c}|\boldsymbol{x})P_D(\boldsymbol{x})d\boldsymbol{x} \simeq P(\boldsymbol{z},\boldsymbol{c}). \tag{5.10}$$

where the prior distribution factorizes as the mixture

$$P(\boldsymbol{z},\boldsymbol{c}) = P(\boldsymbol{z})P(\boldsymbol{c}) = \mathcal{N}(\boldsymbol{z};\boldsymbol{0},\boldsymbol{I})\text{Cat}(\boldsymbol{c};[1/C]^C),$$

With this setup, each label $\boldsymbol{c}$ is associated with an independent low-dimensional space where $\boldsymbol{z}$ is distributed according to $P(\boldsymbol{z})$. Sampling from each cluster is easy, as opposed to the models presented in (Makhzani et al., 2016) that are specifically trained either for clustering or conditional generation of samples, and where $\boldsymbol{z}$ is jointly distributed according to $P(\boldsymbol{z})$ over all $\boldsymbol{c}$ classes.

Proposed semi-supervised model Let $P_D(\boldsymbol{x}_l,\boldsymbol{c}_l)$ be the joint empirical distribution of labeled data $\boldsymbol{x}_l$ with labels $\boldsymbol{c}_l$. The proposed semi-supervised AAE (SAAE) is trained in 3 stages: a reconstruction and regularization phase that are identical to the ones in the standard AAE, and a supervised classification phase minimizing the cross-entropy $\lambda_c \cdot \mathbb{E}_{\boldsymbol{x}_l,\boldsymbol{c}_l\sim P_D(\boldsymbol{x}_l,\boldsymbol{c}_l)}[-\log Q^c_{\boldsymbol{\phi}}(\boldsymbol{y_l}|\boldsymbol{x_l})]$ using the available labels, where λ_c controls the weight of the classification loss. The optimization problem is defined as

5

$$\text{Regularization: } \max_{\boldsymbol{\xi}} \mathbb{E}_{\boldsymbol{z},\boldsymbol{c}\sim P(\boldsymbol{z},\boldsymbol{c})}[\log(S(d_{\boldsymbol{\xi}}(\boldsymbol{z},\boldsymbol{c})))] + \mathbb{E}_{\boldsymbol{x}\sim P_D(\boldsymbol{x})}\mathbb{E}_{\eta\sim P_{(\eta)}}[\log(1-S(d_{\boldsymbol{\xi}}(g_{\boldsymbol{\phi}}(\boldsymbol{x},\eta))))]. \tag{5.11}$$

$$\text{Classification: } \min_{\boldsymbol{\phi}} \lambda_c \cdot \mathbb{E}_{\boldsymbol{x}_l,\boldsymbol{c}_l\sim P_D(\boldsymbol{x}_l,\boldsymbol{c}_l)}[-\log Q^c_{\boldsymbol{\phi}}(\boldsymbol{c}_l|\boldsymbol{x}_l)]. \tag{5.12}$$

$$\text{Reconstruction: } \max_{\boldsymbol{\theta},\boldsymbol{\phi}} \mathbb{E}_{\boldsymbol{x}\sim P_D(\boldsymbol{x})}\mathbb{E}_{\boldsymbol{z},\boldsymbol{c}\sim Q_{\boldsymbol{\phi}}(\boldsymbol{z},\boldsymbol{c}|\boldsymbol{x})}[\log(P_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{c})) + \mathbb{E}_{\boldsymbol{x}\sim P_D(\boldsymbol{x})}\mathbb{E}_{\eta\sim P_{(\eta)}}[d_{\boldsymbol{\xi}}(g_{\boldsymbol{\phi}}(\boldsymbol{x},\eta))]. \tag{5.13}$$

5.3. Model architecture and training

This chapter describes the patient data used to train the VAE and the AAE models of patient-specific respiratory motion, as well as the pre-processing and post-processing steps. The architecture of the proposed SAAE — a population breathing model capable of simultaneously classifying and generating specific types of breathing — is presented, together with the experiments used to evaluate its performance. The breathing signals used for model training and evaluation are time series representing the position of chest markers in lung cancer patients. Figure 5.1 shows an overview of the workflow, including the models' architecture and the final post-processing time series reconstruction step.

5.3.1. Patient and population data

Different breathing signals were obtained with the radiosurgery system Cyberknife® (Accuray Inc., Sunnyvale CA, US), which tracks breathing movement using correspondence of markers positioned on the patient's chest (Coste-Manière et al., 2005). The

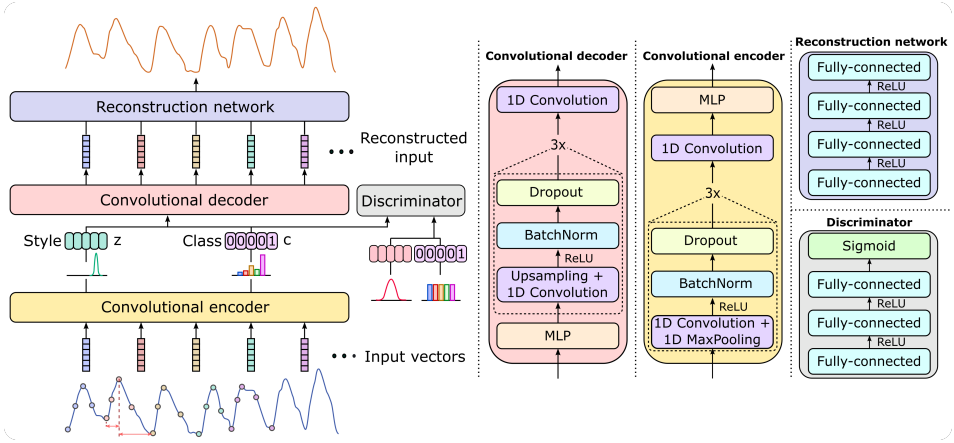
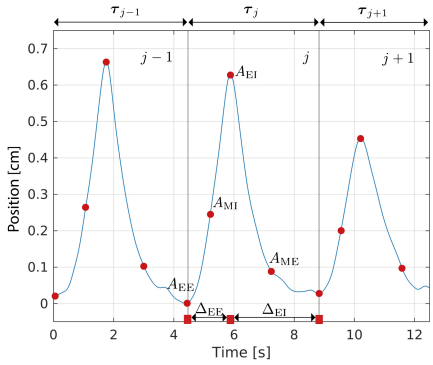


Figure 5.1: **Semi-supervised breathing model architecture**. First, the original time series is pre-processed via principal component analysis (PCA), to obtain the input vectors $\boldsymbol{x}$. Population models are then obtained using a SAAE with one-dimensional convolutional layers. The inference model generates a class label latent variable $\boldsymbol{c}$ in addition to the continuous low-dimensional vector $\boldsymbol{z}$. Labeled data can be leveraged during training in order to learn the classification task in a semi-supervised manner. During generation (red dashed square), the sampled latent variables are transformed into the input vector form. These new vectors $\hat{\boldsymbol{x}}$ are then transformed into a time series with the help of an auxiliary reconstruction neural network.

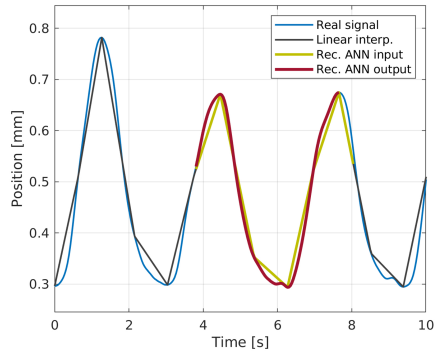
data used in this work consists of long respiratory traces for 21 different patients. The optical device tracks data with a 26 Hz frequency, for a total duration between ten and thirty minutes. The breathing signals for 15 out of the 21 patients were obtained from the open-access database recorded at Georgetown University Hospital (Washington DC, United States) (Ernst, 2011), with breathing amplitudes between 0.5 and 10 mm. The 6 remaining respiratory traces were recorded during treatments at Erasmus MC (Rotterdam, Netherlands) and correspond to 6 patients with smaller amplitudes between 0.5 and 2 mm. The 2 datasets are referred to as the GUH and EMC datasets in the remainder of the chapter.

Input data & pre-processing The first step consists of removing errors related to machine recalibration during signal acquisition. This results in a 3D time series, where each dimension correspond to a physical dimension in the Cartesian coordinate system. The 3D signals are further compressed into a 1D signal projecting them onto the main axis of movement via principal component analysis (PCA) (using the eigenvector with highest eigenvalue). Given the correlation between the three physical dimensions, such projection onto the principal axis retains around 95% of the original variance. The resulting trace is divided into different periods $\boldsymbol{\tau}_j$, each of them corresponding to the time between start of different inhales. Each period j is discretized into 4 points with $A_{s,j}$ denoting position and a $\Delta_{s,j}$ representing the difference in time between consecutive points. Thus, a period is parametrized by the vector

$$\boldsymbol{\tau}_j = (A_{\mathrm{EE},j}, \Delta_{\mathrm{EE},j}, A_{\mathrm{MI},j}, A_{\mathrm{EI},j}, \Delta_{\mathrm{EI},j}, A_{\mathrm{ME},j}), \tag{5.14}$$



(a) Times series discretization.



(b) Time series reconstruction.

Figure 5.2: **Input pre-processing and output reconstruction**. (a) Discretization of a breathing signal into periods and time-position points. In practice, the time series is discretized into a pair of time-position coordinates that are concatenated for a number of periods covering a certain desired time. (b) Transformation of the vector $\boldsymbol{x}$ into a time series. An additional ANN is trained to generate realistic breathing signals from linearly interpolated time series.

5

where s denotes the stage within each breathing period: EE for the end of exhale (or beginning of inhale), EI for the end of inhale (or beginning of exhale), and ME, MI for the 2 intermediate points between EE and EI. For simplicity, the redundant $\Delta_{\mathrm{ME},j}$ and $\Delta_{\mathrm{MI},j}$ time coordinates is omitted, since they are equal to $\Delta_{\mathrm{EI},j}/2$ and $\Delta_{\mathrm{EE},j}/2$, respectively.

Figure 5.2a displays a fragment of the time series and its discretization into time-position points. A breathing sample is obtained by concatenating consecutive periods for the desired length of the signal. Each sample is assumed to be i.i.d. and is characterized by a vector $\boldsymbol{x} = (\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, ..., \boldsymbol{\tau}_{N_T}) \in \mathbb{R}^{N_T \times 6}$ formed by N_T discretized periods. Vectors of length $N_T = 25$ are used to model shorter signals of 1 to 2 minutes, and $N_T = 100$ for longer signals of several minutes corresponding to the typical duration of radiotherapy treatments. This compression step allows reducing the dimensionality of the breathing time series two orders of magnitude.

The pre-processing step results in 36,430 and 4,468 breathing fragments for the GUH and EMC datasets, respectively. Each data sample is assigned a baseline shift label according to the slope of the signal: if the slope of a sample is above a certain threshold value, the breathing sample is labeled as upwards baseline shift. Likewise, if the (negative) slope is below the threshold, the data point is labeled as downwards baseline shift. The threshold values correspond to the 7.5 upper and lower percentile of the distributions of slopes in the GUH dataset.

Time series reconstruction The output vectors $\hat{\boldsymbol{x}}$ from the models have the same structure as the discretized input vector. Therefore, they must be transformed back into a time series by reconstructing the position values between two consecutive points in $\hat{\boldsymbol{x}}$. A first order approximation is a simple linear interpolation between the four position points in each cycle, which requires little time but lacks accuracy.

Alternatively, a realistic breathing time series can be reconstructed using an ad-

ditional feed-forward neural network, denoted as *reconstruction ANN* in the remainder of the chapter. The input is the linearly interpolated series, and the ANN learns a general mapping from the linear time series into realistic shapes. The input for the reconstruction ANN is no longer a vector of dimension $M = 6 \times N_T$, but a fragment of 120 position values (see Figure 5.2b). The number 120 is a hyperparameter that is selected from a set of different candidate lengths. The output of the ANN is the first 100 transformed values of the input series. By adding 20 extra positions, the network achieves higher accuracy without discontinuities during concatenation of consecutive fragments. The ANN is composed of 4 feed-forward layers and is trained with a learning rate of 10^{-4}, decay rate of 10^{-6} per epoch, and batches of 256 samples.

The training data for the reconstruction ANN consists of slices with 120 elements of position values from the recorded breathing signals, and the corresponding linear interpolations. During training, the input and output slices are normalized to the interval [0,1]. Ideally, a single general ANN would reconstruct the time series from any patient in the population and make the process highly scalable. Such reconstruction ANN would be trained using only a subset of the data (either data from a single patient or a subset of data from all the patients). To test this, two reconstruction ANNs are trained using (i) data from one patient (referred to as PatBR model from on) and (ii) a subset of data from the GUH data (referred to as PopBR model), while both models are tested using the EMC dataset. The PatBR is trained using a single patient from the GUH dataset, while the PopBR is trained on 10% of the GUH dataset, instead of on all available samples. This is due to the fact that, unlike with the AAE, VAE and SAAE vector inputs, the training dataset for the reconstruction task consists of few million fragments of the breathing time series (vectors with 120 position values) obtained from linear interpolation of the generated vectors.

5.3.2. Patient-specific models

First, to investigate the potential and limitations of signal modeling with probabilistic autoencoders, the standard VAE and AAE algorithms are applied for modeling breathing signals from individual patients in the dataset separately. Both the AAE and VAE frameworks use an isotropic Gaussian prior distribution $P(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{I})$, with an encoder and a decoder of 4 convolutional layers and an MLP block with 2 fully-connected layers. Batch normalization (Ioffe and Szegedy, 2015) and dropout between layers are empirically found to improves convergence, resulting in better generalization.

For the VAE, the value of the parameter λ_{KL} in Equation 5.5 is determined using the input dimension M and latent dimension N (which vary per model) as $\lambda_{KL} = 0.02(N/M)$. The fixed value of 0.02 is empirically found to yield an optimum balance between a Gaussian latent space that is closer to the prior and good reconstruction performance.

Of all patient data available, 20% of the patient data is equally split into a validation and a test set, while the remaining 80% is used during training, with a batch size of 256 samples and the Adam optimizer (Kingma and Ba, 2017) with learning rate 10^{-4}. After training the models, the input vector $\boldsymbol{x}$ can be reconstructed by sampling the inference model $Q_\phi(\boldsymbol{z}|\boldsymbol{x})$ to obtain $\boldsymbol{z}$, and then sampling the decoder. Artificial breathing signals can be obtained by decoding random samples from the prior $P(\boldsymbol{z})$.

5.3.3. Population models of breathing irregularities

The proposed semi-supervised SAAE framework is applied to model and classify baseline shift breathing irregularities, which are gradual downward or upward shifts of the exhale position. As a first step, two simple experiments using simplified sinusoidal breathing signals investigate whether it is possible to obtain good models that classify and generate signals with upward or downward shift, or no shift at all (regular signals). In the first experiment (S1), only the slope of the signals is varied; while the second experiment (S2) includes signals with randomly sampled slopes, periods and amplitudes.

In the third experiment, the SAAE model is trained using real breathing signals, investigating the number of labeled samples needed to obtain accurate classification, with the GUH dataset as the training set (with 10% as validation data) and testing on the EMC dataset. The models are trained using a batch size of 256 samples with unequal learning rates for the Adam optimizer: 10^{-4} in the reconstruction and supervised classification phase and $2 \cdot 10^{-4}$ for the discriminator. α values of around 5-10 significantly enhance classification when the number of labels is limited, while higher values do not improve and even hinder performance.

5

5.4. Model evaluation

Evaluating patient-specific models A good model can reconstruct unseen signals and generates artificial signals that distribute according to the training data. Several tests asses the generative performance of the patient-specific model:

- Analyzing reconstruction error. The generalization performance of the patient-specific models is evaluated via the reconstruction error of signals from the test set. Based on how varying the dimensionality of the latent space affects reconstruction error of unseen test data (given a fixed decoder and encoder architecture), the proposed experiments also aim at determining an optimal number of latent variables. Additionally, the advantages of using convolutional layers are verified by comparing models purely based on fully-connected layers to the proposed one-dimensional convolutional models in terms of reconstruction performance.

- Assessing the generative performance. To determine if the model captures the data distribution, an external classifier is trained to distinguish between reconstructed and artificial samples from the model. As in (Razavi et al., 2019), the external classifier takes the reconstructed vectors as ground-truth input instead of the original input vectors, since the compression through the latent space usually removes high-frequency noise in the original data that can be easily used by the classifier to distinguish samples. The classifier performance is also evaluated for different latent space sizes.

- Investigating the structure of the latent space. The presence of "empty" regions in the latent space where no encodings $\boldsymbol{z}$ data are observed often results in low quality and variability of training samples. Computing the distribution of the distance between neighboring $\boldsymbol{z}$ from the dataset can show the presence of empty

regions in the latent space. Furthermore, a possible mismatch between the aggregated posterior and prior distribution can be assessed by comparing the distribution of the L2 norm of the encodings of the training samples and the samples from the prior.

Evaluating breathing irregularity models The evaluation of the joint models is based on the F1-score (van Rijsbergen, 1979). For a given class, the precision is the proportion of correctly predicted samples over the total number of examples labeled as such class, while the recall is the fraction of correctly predicted samples over the total number of true samples for the given class. For multi-label classification, the macro F1-score (mF1) can be used, which is the average of F1-scores for the different classes. The baseline shift breathing irregularity models are tested with regards to both their classification and generative performance.

- Assessing classification performance. The discriminative performance (i.e., the ability to label signals having upward, downward or no baseline shift) is evaluated by comparing the classification accuracy of SAAE models to other neural network models purely optimized for classification. Specifically, convolutional neural network and fully-connected neural network discriminators are trained using a labeled subset of the training data. This additional convolutional classifier is similar to the encoder and inspired by state-of-the-art one-dimensional convolutional ECG models in (Acharya, Fujita, Lih, et al., 2017; Acharya, Fujita, Oh, et al., 2017; Fujita and Cimr, 2019). The effect of the number of labeled examples used during the supervised phase of training on the classification accuracy of the SAAE is evaluated by comparing its mF1-score to that of pure classifier networks.

- Evaluating generative performance. Inspired by (Ravuri and Vinyals, 2019) and (Razavi et al., 2019), Classification Accuracy Score (CAS) is used to evaluate the model's generative performance (i.e., whether the model generates realistic and varied samples). The CAS is obtained by training a discriminative model on data generated by the model, and evaluating the mF1-score on the real data test set.

- Analyzing the reconstruction error. Additionally, the reconstruction error on GUH and EMC test data is reported for two SAAE models using 15 and 30 latent variables.

5.5. Results

5.5.1. Patient-specific models

The results of the evaluation of the AAE and VAE patient specific models in terms of reconstruction and generative performance are shown in Figure 5.3 for 2 randomly selected patients. The models for the first and second random patient were trained using 1890 and 2653 samples, respectively. Figure 5.3a displays the reconstruction error on unseen test set data for different latent space dimensionalities, for both models based on one-dimensional convolutional architectures and models purely based on

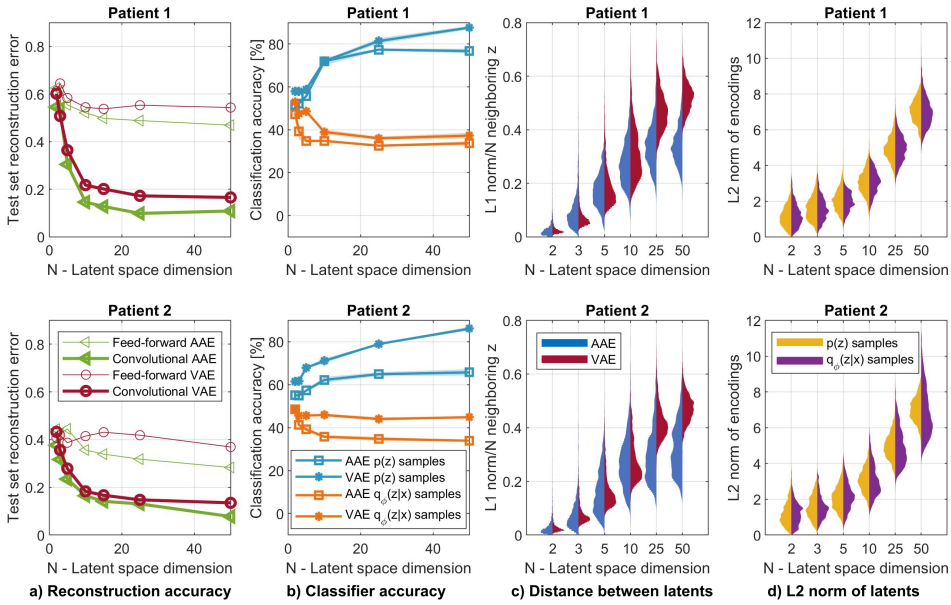


Figure 5.3: **Patient-specific model evaluation**. (a) Reconstruction error on the test set for different latent space dimensionalities N. (b) Performance of an additional classifier trained to distinguish samples from the dataset from artificial samples from the model. Shaded regions represent the standard deviation around the mean (solid). (c) Distribution of the distance between neighboring encodings, for the AAE (blue) and VAE (red). The L1 norm distance is normalized by dividing by the latent space dimensionality. (d) Distribution of the L2 norm of the real data encodings $\boldsymbol{z}$ (yellow) and sampled encodings from the prior distribution (purple).

fully connected layers. The error values are re-scaled to the interval [0,1] to facilitate comparison, where 1 corresponds to the maximum error achieved at weight initialization. Although the error always decreases with increasing latent dimension N, the convolutional architectures notably increase the accuracy in the reconstruction. For qualitative evaluation, Figure 5.4 shows reconstructions of the original inputs using a convolutional model with a 5-dimensional latent space ($N = 5$).

The generative performance is shown in Figure 5.3b, depicting the accuracy of a CNN classifier trained to distinguish reconstructed data points from artificial samples generated by models with varying latent dimensionality. The reported average, maximum and minimum values correspond to 3 different classifiers trained on distinct artificial data. The data is generated either by sampling the prior $P(\boldsymbol{z})$, or by taking $\boldsymbol{z}$ encodings in the vicinity of $Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$, where the latter cover a much smaller domain of the latent space. The auxiliary classifier performs worse when distinguishing real and AAE samples, hinting that these better capture the distribution of the data. Note that the binary cross entropy loss values are almost always above 1 for the $P(\boldsymbol{z})$ classifier, which indicates the presence of uncertainty and significant miss classification errors.

To study the structure of the latent space Figure 5.3c shows the distribution of the distance between neighboring encodings. Since the L^n-norm distance metric always

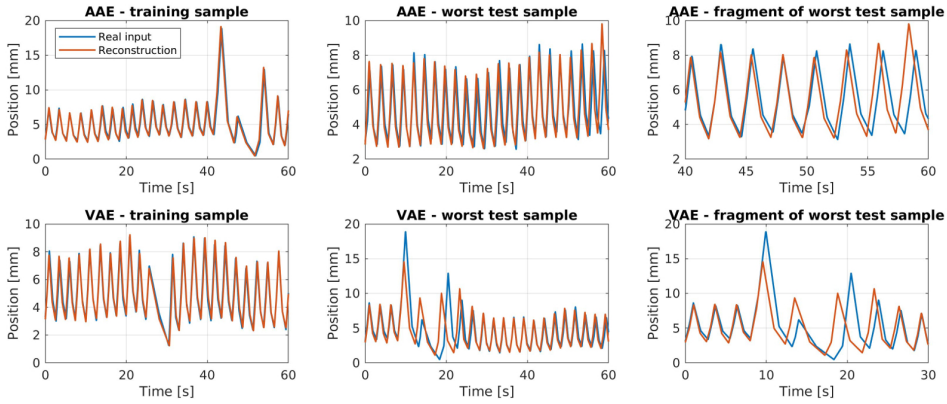


Figure 5.4: **Breathing signal input reconstruction**.(Top row) Reconstruction of breathing signals from AAE patient-specific models and (bottom row) VAE-based models for (left) a sample from the training set, (middle) the worst performing sample from the GUH test set, and (right) and a fragment of the worst reconstructed GUH test sample, with the highest reconstruction error. The discretized reconstructed signals are linearly interpolated and transformed back into a time series.

5

increases with the number of latent dimensions N, the reported L1 norm between nearby $\boldsymbol{z}$ is devided by the latent space dimensionality. The plotted distributions indicate that the AAE encodings are more evenly distributed. This, together with the fact that the classifier in Figure 5.3b struggles to distinguish real signals from samples in the vicinity of $Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ hints that the latent space is more compact in the AAE-based models. On top of that, the AAE algorithm seems to be a more effective latent space regularizer, whose models have a latent space that closely resembles the prior distribution. This is deduced from Figure 5.3d, where the distribution of the L2 norm of the encodings is compared to the distribution of the L2 norm of samples from the prior. The results suggest a possible relationship between more compact and similar to the prior AAE latent space and the lower classifier performance for AAE samples in Figure 5.3b.

5.5.2. Baseline shift population models

First, the effect of slope, period and amplitude variations on the classification accuracy is evaluated using an artificial dataset based on sinusoidal signals. The SAAE models achieve a mF1-score of 100% in S1 by using as little as 300 labeled examples during the supervised classification phase. Adding period and amplitude variability to the sinusoidal signals in S2 results in additional difficulty, and the models need 1500 labeled examples (around 4% of the training data points) in order to achieve null classification error.

Based on these results, a baseline shift model is trained using real data. The performance and added benefits of jointly classifying and modeling breathing signals are evaluated by assessing the classification accuracy, generation variability and the reconstruction error. The classification performance is assessed by comparing the SAAE models to purely discriminative models trained to only classify baseline shifts using a

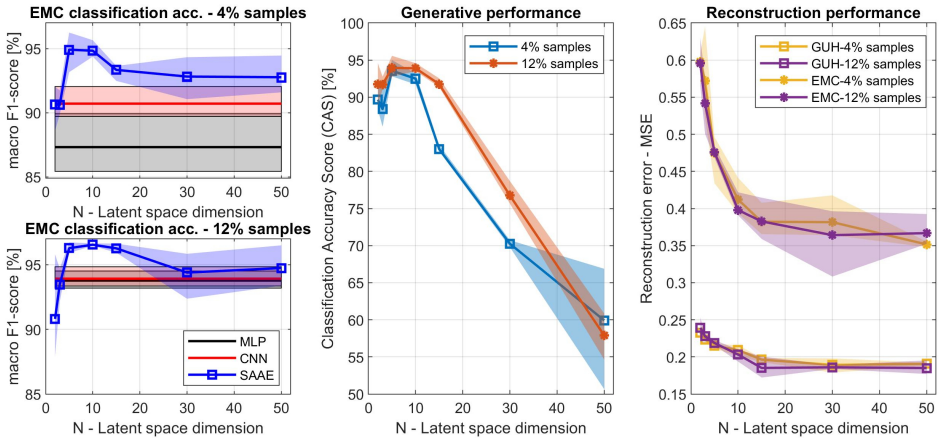


Figure 5.5: **Joint semi-supervised model results**.Classification, generation and reconstruction performance of the SAAE semi-supervised models, for varying latent space dimension. The models use 4% or 12% of the training data during the supervised classification phase, which corresponds to 1500 and 6000 data points, respectively. The reported mean, maximum and minimum values correspond to 3 independent models with different training-test dataset splits and weight initialization. The relative reconstruction error is expressed as a percentage, where 100% corresponds to the maximum error corresponding to a model with randomly initialized weights.

subset of the available labels. Specifically, a feed-forward (MLP) classifier and a convolutional (CNN) classifier were trained using 4% and 12% of the GUH training labeled data. Figure 5.5 shows that the SAAE model with 5 to 15 latent variables outperforms both architectures, achieving a mean mF1-score of 94.91% and 96.54% on the unseen test EMC dataset when trained with 4% and 12% of the labels, respectively.

The generative performance and sample variability are evaluated with the CAS mF1-score. A CNN classifier is trained using 36430 randomly generated samples from the SAAE model, which allows a fair comparison with the model trained using the real GUH data. The classifier is then evaluated on EMC data, achieving a remarkable 93.90% mF1-score for the model with 10 latent variables trained with 12% of the labels, which is on par with the performance of the feed-forward and CNN classifiers trained with real data observed in Figure 5.5 (MLP and CNN in the two left plots). As with the patient-specific models, the generative performance significantly degrades for higher latent space dimensionality.

Finally, the reconstruction error on test set data is shown in Figure 5.5. As with the patient-specific models, the error is expressed relative to the maximum error corresponding to predictions from a randomly initialized model. The models perform similarly when using more than 10 latent variables. Higher latent space dimensionality seems to beneficial in the complicated task of reconstructing EMC samples that follow a different distribution, where the models achieve similar reconstruction performance to the feed-forward patient-specific models in Figure 5.3a.

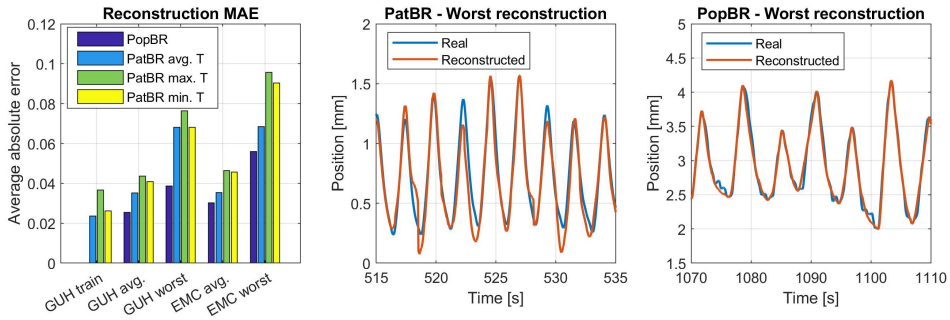


Figure 5.6: **Output signal reconstruction performance**. (Left) Average absolute error achieved by the PatBR and PopBR models in the reconstruction of breathing time series. The error is shown for the training patient(s), the worst-performing patient and the entire set of patients present in each of the GUH and EMC datasets. (Middle) Reconstruction of the EMC signal fragment with highest error, using the PatBR model trained with data from the patient with maximum amplitude. (Right) Worst-performing reconstruction over all the EMC dataset using the PopBR model.

5

5.5.3. Time series reconstruction

Three different PatBR models are trained using the data from three patients: the patients with the largest and lowest breathing period in the dataset, and one of the patients with an average period. From these PatBR models, the former (largest period) achieves the largest error, precisely on signals of the patient with the lowest period. For each of the PatBR models and the PopBR model, a comparison of the average absolute error (average L1-norm) on the training set, the test set and the worst performing patient from the test set is shown in the left plot of Figure 5.6. The average absolute error is calculated as the average L1-norm $|\boldsymbol{w}_{\text{real}} - \boldsymbol{w}_{\text{rec}}|$ between all position points in the recorded and reconstructed time series vectors $\boldsymbol{w}$. The middle and right plots in Figure 5.6 show the worst EMC test sample reconstruction from the PatBR and PopBR models, respectively.

The PopBR reconstruction ANN consistently outperforms the single patient PatBR networks and opens the door to using a single model to reconstruct breathing signals for any patient. PatBR models fail to reconstruct time series from other patients, especially when they are evaluated on patients whose period significantly differs from that of the samples used for training, as seen in the left plot of Figure 5.6. The generalization error of the PopBR model is very low and it provides accurate reconstructions for patients whose breathing signal was recorded in a different location and machine. The error could in principle be further decreased by training a specific PatBr for each specific patient, at the expense of slightly longer computation time.

5.6. Discussion

The standard VAE, standard AAE and SAAE architectures result in breathing models that capture the variability of respiration through few latent variables, as opposed to approaches that use implicit adversarial models (Golany and Radinsky, 2019; Wulan et al., 2020). The models are easy to sample and the decoders generate realistic breath-

ing samples. The convolutional layers result in 25% reduction of the reconstruction error on test data. AAEs outperform standard VAE models in reconstruction, generalization and generative performance. Much of the AAE success seems to be related to their more compact latent space: their aggregated posterior distributions are closer to the prior, and their encodings are more evenly spaced, as seen in Figure 5.3c and Figure 5.3d. The problem of aggregated posterior-prior mismatch in VAEs is not new, and the results in this work are in line with previous findings (Dai and Wipf, 2019; Rezende and Viola, 2018; Rosca et al., 2019).

For the set of all possible models, the reconstruction performance is in theory independent of the latent dimension. Very powerful autoencoders with deep encoders and decoders could perfectly reconstruct the input using as few as one latent dimension, but this is not observed in practice. In general, the performance can be practically improved by adding more latent variables or increasing the capacity of the model. However, it has been observed that very powerful decoder architectures tend to ignore the information encoded in $\boldsymbol{z}$ (Bowman et al., 2016; Chen et al., 2017; Zhao et al., 2018). In concordance with Figure 5.3, adding dimensions helps, especially in low-dimensional latent spaces. Nevertheless, there is a certain latent space dimensionality beyond which adding more latent units seems to add little information. For the VAE, this may manifest as "inactive latent variables", where some latent units remain equal to the prior distribution during the whole training process (Burda et al., 2016; Sønderby et al., 2016). For the specific case of breathing and given the presented encoder and decoder convolutional architectures, the limit seems to be around 10 latent variables. This is supported by the fact that the test reconstruction error and classifier performance plateau around $N = 10$ in Figure 5.3.

Even though SAAE models are mainly trained to reconstruct breathing signals, they outperform pure discriminative architectures based on state-of-the-art one dimensional convolutional models (Acharya, Fujita, Lih, et al., 2017; Acharya, Fujita, Oh, et al., 2017). The fact that a single model can (better) classify and selectively sample types of signals is a novelty with respect to previous architectures that specialize in only one of such tasks (Wulan et al., 2020; F. Zhu et al., 2019). One interesting remark is the fact that there seems to be a latent dimension range between 5 and 15 where SAAE models are superior in the classification task. In general, increasing the number of latent variables means that each variable carries less information about the input. A plausible reason behind the loss of accuracy for increasing latent space dimensions is that some of the information encoded in $\boldsymbol{c}$ may leak into the style variables $\boldsymbol{z}$. Models with a large enough number of latent variables would not benefit from the joint discriminative-generative modeling task, since they could completely encode the input using $\boldsymbol{z}$ and simply learn the label $\boldsymbol{c}$ separately. However, this should be confirmed in future research.

The generative performance of the SAAE models degrades with increasing latent dimensionality. As in the patient-specific models, a possible reason is the "emptier" latent space with larger distance between encodings. This directly follows from the increasing volume of the multi-variate Gaussian latent space and the fixed number of samples used to cover such volume during training. Additionally, SAAE models perform similarly to the patient-specific models in terms of reconstruction and general-

ization on test samples from the same distribution, as indicated by the reconstruction error on GUH test samples. Although the reconstruction accuracy significantly decreases, the SAAE models also perform reasonably well in the much more complicated task of generalizing to test samples from the EMC dataset with different distribution, and their reconstruction error is on par with feed-forward patient-specific models (Figure 5.3a). As in the patient-specific models, the SAAE models seem to benefit little from adding extra latent variables for latent space dimensionalities above 10. Since the classification and generative performance attain their maximum between 5 and 10 latent variables, we can conclude that the optimum latent space dimension lays around 10.

The presented models can be applied to a wide range of tasks involving signal generation and classification. Regarding generation, the models can be used to capture the variability in breathing of a patient and generate artificial samples. For the specific application of proton therapy — where a very narrow (1-3 mm) proton beam is used to actively scan the tumor — the movement of the beam and the breathing motion are on comparable time scales, leading to the so-called "interplay effect" degrading therapeutic effectiveness. The presented generative framework presents significant advantages in addressing this problem compared to the commonly used simple sinusoidal artificial signals that fail to capture irregular motion and the true variability of the breathing. The realistic generated samples can be incorporated into treatment design in order to make treatments less sensitive to breathing motion during dose delivery. Since each generated breathing sample results in a different virtual delivered dose, repeated sampling allows deriving the distribution of plausible treatment outcomes, which can subsequently be used to assess treatment plan robustness before actual delivery or to directly optimize treatment plans to be robust against breathing movements. As a result, the desired clinical outcomes can be better ensured or the likelihood that a patient will present a certain type of breathing can be estimated - tasks that are infeasible with currently available methods.

An important advantage of the presented methodology is the fact that it achieves feasible compute times. Training times are massively reduced by using Graphics Processing Units (GPUs), which are needed to train the presented convolutional architectures due to the requirements of the latest version of the Tensorflow package (Abadi et al., 2015). Most of the training was done with an NVIDIA® Tesla® K80, and the training times vary around 10 minutes for the VAE and AAE patient-specific models, 30 minutes for the reconstruction PopBR and PatBR models, and 20 minutes for the SAAE models. Generating and classifying breathing samples is almost instantaneous.

Limitations One drawback of the proposed method is the uninformative prior $P(\boldsymbol{c})$ in the semi-supervised model, which assumes no previous knowledge about the proportion between different classes. For cases when there is class imbalance, i.e., many more samples of regular breathing compared to irregular breathing, using such uninformative prior may result in the model miss-classifying some samples in order to match the uniform prior. The solution to this problem is dataset-dependent approach and involves determining the naturally occurring proportion of classes.

5.7. Summary

In this chapter, a semi-supervised algorithm based on the AAE is presented, allowing simultaneous classification and generation of biomedical signals within a single framework. The resulting models classify signals with greater accuracy than discriminative models specifically trained for classification using only few labeled data points; are easy to sample, and compress the data into a reduced latent space with few independent parameters with known probability distributions. In view of the results, 10 of such latent variables are able to capture most of the variation in the data and achieve excellent reconstruction and generation of samples. For the particular case of breathing, the adversarial objective used in AAEs is a better regularizer of the latent space and overcomes some of the previously studied problems of the VAE framework.

Given the length of the input time series, all models are trained on compressed input vectors containing information about the period and amplitude of the biomedical signal. The compressed output vectors produced by the model are transformed back into a time series with the help of an additional reconstruction network. Reconstruction ANN models trained with the data of a single patient (PatBR) do not achieve good generalization when evaluated on other patients, and are outperformed by a population-based reconstruction models (PopBR) trained with a subset of the data of a population of patients. The population model is trained only once and achieves great accuracy when applied to new unseen data. Even though this work is based on mechanical breathing signals, the framework shows potential applicability to simulation and diagnostic purposes using any other biomedical signal with a quasi-periodic structure.

6

Simulating interplay effects in proton therapy

6.1. Introduction

In Intensity Modulated Proton Therapy (IMPT), breathing interplay effects arise from the interaction between the scanning beam and moving organs during treatment delivery. This is detrimental, as during the few minutes in which each fraction is delivered, the continuous movement of the target due to breathing degrades the final dose distribution (Bert and Durante, 2011; Bert et al., 2008; Lambert et al., 2005). Given the adoption of IMPT in treating moving tumors, there is a growing need for computational methods that allow sound statistical evaluation of interplay effects, where the error introduced by modeling approximations (e.g., using few breathing realizations of sinusoidal breathing) is known and justified.

Several techniques aimed at minimizing the detrimental effect of breathing during delivery include beam gating, rescanning, beam tracking, breath-hold and compression. During beam gating the patient breathes freely and the dose delivery is constrained to a specific part of the breathing cycle (e.g., end of exhale) (Bert et al., 2009; Ohara et al., 1989). Beam tracking consists of adjusting the treatment delivery system to real-time predicted target movement (Bert et al., 2009; Y. Zhang et al., 2014). In rescanning or repainting the target is irradiated several times during the same fraction, which helps smooth the final dose distribution (Phillips et al., 1992; Seco et al., 2009). Finally, breath-hold and compression methods aim at immobilizing the target during delivery (Boda-Heggemann et al., 2016; Péguret et al., 2016).

From a treatment planning perspective, different approaches are used to account for target motion by including information about different breathing phases (e.g., exhale, inhale, mid-ventilation) into the optimization. Internal Target Volume (ITV) planning aims at irradiating an ITV volume in the reference phase, which is defined as the

The contents of this chapter have been published as a journal paper in *Physics in Medicine & Biology* 66 (23), 235003 (2021), (Pastor-Serrano, Habraken, et al., 2021).

union of all Clinical Target Volume (CTV) contours of the different breathing phases (Shih et al., 2004). With the help of surrogate models that generate artificial motion, ITVs can be extended to probabilistic ITVs that capture breathing variability (Krieger et al., 2020). 4DCT planning is based on optimizing the dose distribution using min-imax robust optimization (Pflugfelder et al., 2008), including multiple Computerized Tomography (CTs) from different breathing phases so that the dose prescriptions are met in all the included breathing phases (Bernatowicz et al., 2017; Engelsman et al., 2006; Heath et al., 2009). Some 4DCT approaches also account for beam tracking (Eley et al., 2014) and temporal structure (Engwall, Fredriksson, and Glimelius, 2018) during optimization.

Previous work shows that fractionation effectively limits interplay dose degradation in Intensity Modulated Radiation Therapy (IMRT) delivery techniques with moving parts such as multi-leaf collimators (MLCs) (Bortfeld et al., 2002; Jiang et al., 2003), but may be insufficient to tackle negative biological effects in treatments with many segments of few monitor units (MUs) (Seco et al., 2007). Several studies further investigate the effects of regular breathing motion and collimator speed on the outcome of MLC treatments (L. Court et al., 2010; L. E. Court et al., 2008), showing that non-negligible interplay effects increase with target magnitude, plan complexity and breathing period.

While the problem of interplay is common for all dynamically delivered treatments, its nature differs between IMPT and IMRT: proton pencil beams are narrower than photon beams, deliver the dose more locally, and their irradiation times are usually an order of magnitude smaller. Several studies quantify the negative effect of interplay in IMPT and evaluate the effectiveness of different mitigation techniques such as re-painting in lung and liver patients (Engwall, Glimelius, and Hynning, 2018; Li et al., 2014; Seco et al., 2009; Y. Zhang et al., 2016), breath-hold (Emert et al., 2021; Yu et al., 2017) or a comparison between different mitigation techniques used in liver treatments (Y. Zhang et al., 2018), showing that neither rescanning nor gating alone can mitigate interplay effects. Regarding the effect of motion parameters, large breathing amplitudes are known to produce significant local under- and overdosing (Jakobi et al., 2018; Kardar et al., 2014; Kraus et al., 2011).

Evaluating interplay is usually time consuming and requires many dose distributions corresponding to different realizations of breathing during treatment delivery. While alternative, more realistic and computationally demanding approaches use simulated 4DCT scans with dynamic dose delivery (Boye et al., 2013) or motion surrogates (den Boer et al., 2021), most of the interplay evaluation studies are based on a single 4DCT scan and many breathing signals to simulate different breathing scenarios. Obtaining enough of such signals involves either taking fragments from the recorded respiratory signal — which is often short and does not offer much variability — or using a sinusoidal approximation, oversimplifying breathing and failing to capture typical irregularities such as baseline shifts and amplitude changes. Furthermore it is not known how realistic and irregular these signals need to be, how small breathing variations affect the final dose, and how many different breathing samples are needed to accurately capture the statistical variation of interplay. Except for one published paper hinting the possible systematic error in IMRT interplay evaluation caused by the use of a limited number of motion samples in both planning and evaluation (Evans et al., 2005),

no previous study has investigated the statistical significance of evaluating interplay effects using few samples and simplified breathing models disregarding any breathing cycle hysteresis.

Building on previous IMRT studies (Kissick et al., 2005; Seco et al., 2007), this work investigates the interplay dependence on breathing uncertainties for proton treatments with many pencil beams — where the order of magnitude of beam delivery times is a factor 100 lower that the period of breathing motion — and specifically the relationship between dose and breathing parameters such as period and amplitude changes. The proposed method for evaluation of interplay is based on a 4DCT scan representing the different anatomies of the patient in a breathing cycle, and breathing signals that capture how these alternate during the course of a treatment fraction. Extending on previous work (L. Court et al., 2010; L. E. Court et al., 2008), such breathing signals have both constant and variable breathing periods. This chapter covers the following topics related to the simulation and evaluation of breathing interplay effects:

- A method to statistically assess interplay effects in lung IMPT is presented and applied to evaluate robustness, comparing 4DCT and ITV planning approaches and the impact of fractionation for 8 stage III lung cancer patients.

- The proposed approach is used to evaluate error introduced in the interplay evaluation caused by (i) using simplistic sinusoidal breathing approximations, (ii) using a limited set of scenarios, or (iii) disregarding breathing hysteresis.

  (i) Two methods to generate patient-specific breathing signals which differ in accuracy and computational complexity are compared, referred to as *breathing models*. Specifically, given the popularity of sinusoidal models, this chapter addresses the dosimetric impact of evaluating motion using simple sinusoidal breathing patterns, which are the most commonly used approach when lacking a sufficiently long recorded breathing signal with enough variation.

  (ii) Interplay statistical analyses lacking statistical power (i.e., only consider a limited number of breathing scenarios) can result in errors. The proposed interplay evaluation tool is used to determine whether evaluating interplay with a small number of such breathing scenarios — as observed in most of the previous studies — leads to significant errors.

  (iii) The dosimetric impact of disregarding hysteresis in the breathing cycle is investigated by considering symmetrical inhale and exhale during evaluation.

- The dependence of IMPT interplay dose distributions on breathing parameters such as amplitude, period or starting phase is investigated. More specifically, evaluating how the dose and Dose Volume Histogram (DVH) values vary with small changes in the breathing signal, this chapter aims at determining which parameters (e.g.,breathing amplitude, period) have the biggest effect on dose.

6.2. Interplay effect simulation

This section describes the patient data, the proposed methodology to simulate breathing interplay effects, and the design choices in treatment planning and delivery simulation.

6.2.1. Patient data and treatment plans

Different breathing signals are obtained with the stereotactic body radiation surgery (SBRT) system Cyberknife® (Accuray Inc., Sunnyvale CA, US), which tracks targets that move with respiration using a correlation model that relates the internal target position with external markers taped to the chest of the patient (Coste-Manière et al., 2005; Hoogeman et al., 2009). The long respiratory traces represent tumor movement during treatment for 8 different lung cancer patients. Each signal is matched to a 4DCT scan from a stage III lung cancer patient (having been treated with IMRT and recorded with a Siemens Sensation Open® CT scanner using phase binning) and subsequently rescaled to the maximum 4DCT amplitude. The 4DCT scans are discretized into 8 phases in the breathing cycle: 0%, 25%, 50%, 75%, 100% inhale, and 75%, 50% and 25% exhale. The structures of interest are clinically delineated in all scans, with the exception of the ITV, which is obtained by combining in the mid-ventilation 50% exhale reference phase the CTV volumes from all the breathing phases.

Table 6.1 describes the motion and tumor sizes of the patients in the dataset. Two treatment plans are obtained per patient: ITV plans targeting the ITV in the reference phase, and 4DCT robust plans targeting CTV contours from three phases: the reference 50% exhale phase, and the two extreme 0% and 100% phases. Both ITV and 4DCT robust IMPT plans use a 5 mm setup robustness setting, a 5% range robustness setting, and a 2 mm extra margin around the target(s), based on current clinical practice at Holland PTC (Delft, Netherlands). The treatment is divided into 33 fractions of 2 Gray (Gy), with plans made using Erasmus-iCycle, an in-house Treatment Planning System (TPS) which uses automated multi-criteria prioritized optimization and a pencil beam dose algorithm to calculate the dose delivered per spot (Breedveld et al., 2012; Water et al., 2013), including range shifters and filtering of low-weight beams. No breathing uncertainty mitigation technique is applied during planning or delivery, except for one experiment where volumetric repainting is applied per gantry angle.

6.2.2. Interplay dose calculation

The proposed model calculates an interplay dose distribution based on the treatment plan, the machine parameters, a 4DCT scan and a breathing signal that can be either a fragment of the real recorded signal or an artificial signal from one of the breathing models discussed below. The number of spots — regions irradiated by a single monoenergetic pencil beam — and the order in which they are delivered can be obtained from the treatment plan. Spots are ordered in descending order according to pencil beam energies, on a per gantry angle basis. The machine parameters determine a *spot-timeline*, which is a list ordering the spots in time using information such as the elapsed time between two consecutive spots or the time needed to change layers and beams. The irradiation time for each spot is directly obtained from the optimized plans with beam data corresponding to standard Varian ProBeam® settings, resulting in a fixed

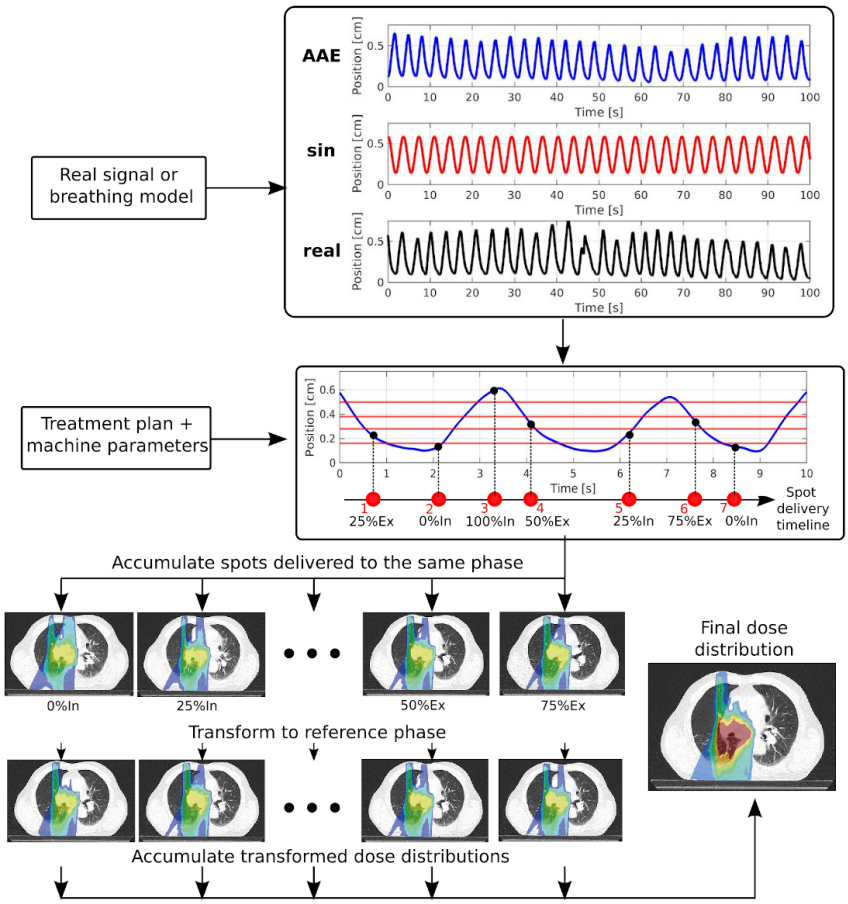


Figure 6.1: **Interplay calculation workflow**. The input breathing signal, treatment plan and machine parameters are used to distribute the spots over the breathing phases included in the 4DCT scan and determine in which phase each spot is delivered. Breathing phase dose distributions are first obtained by accumulating the dose delivered by all the spots in the same phase, and are subsequently transformed to the reference phase. The final interplay dose distribution is the result of adding the transformed doses.

Table 6.1: **Dataset description and treatment delivery times**. The reported values include the breathing amplitude along the lateral, anterior-posterior (A-P) and cranial-caudal (C-C) axes, and the combined volume of the CTV including lymph nodes. Treatment delivery time per gantry angle for both the 4DCT and ITV plans are also reported.

Patient	Breathing amplitude (mm)			Target size (cm^3)	Delivery time (s)					
					Beam 1		Beam 2		Beam 3	
	Lateral	A-P	C-C	CTV & nodes	4DCT	ITV	4DCT	ITV	4DCT	ITV
1	3.2	1.8	4.1	39.1	46.8	46.1	58.2	57.9	28.3	29.1
2	2.1	2.4	5.9	130.9	38.4	33.0	67.9	60.2	39.4	34.8
3	1.5	3.6	9.4	211.5	33.6	29.7	66.2	61.3	35.6	28.5
4	1	0.7	8.7	489.7	68.0	62.2	101.0	96.0	76.5	56.8
5	0.8	0.4	2.4	400.6	34.0	31.9	98.3	76.5	50.4	42.3
6	0.8	2.4	1.7	286.7	53.3	52.6	87.9	65.5	49.7	42.8
7	1.2	1.9	5.4	404.5	61.5	61.7	76.0	76.5	40.2	39.6
8	1.2	0.3	2.3	162.1	40.6	38.5	90.5	69.9	65.5	44.2

current and variable local dose rates between 10 and 54 Gy/s. Beam data measurements are based on integral dose depth (IDD) curves, lateral spot profiles and absolute dosimetry (MU calibration) under reference conditions in a water phantom. For the machine parameters, 10 ms off-beam time are added after delivery of each spot, as well as an average of 0.7 seconds to change energy layer. Range shifter fixed insertion times equal 16 seconds, while the variable time needed to change the gantry angle depends on a linearly increasing, bounded angular acceleration.

Figure 6.1 illustrates the process of simulating an interplay dose distribution. After a breathing signal is generated and a treatment starting phase is sampled, the signal is binned into between full inhale and exhale, according to the maximum 4DCT amplitude (5 bins delimited by 4 red horizontal boundary lines in Figure 6.1). The breathing signal indicates the phase in which each spot is delivered, with all the points of the signal that fall between consequent binning boundaries being considered to be the part of the same being phase. For fractions where the patient presents shallow breathing with low amplitude, the dose will be deposited in only a subset of phases. For baseline shifts, the dose delivery will gradually shift from inhale (75%In, 100%, 75%Ex) to exhale phases (25%Ex, 0%, 25%In) as the treatment proceeds. The number of phases used during interplay evaluation may differ from the number of phases used to optimize the 4DCT plan. In this study, 4DCT plans are made using 3 phases, whereas all the 8 available phases are used for the evaluation. After binning, each point of the signal corresponds to phase of the 4DCT, resulting in the *CT-timeline* containing the different phases ordered in time. Pairing the CT- and spot-timelines results in each spot being assigned to a certain phase. Dose distributions per phase are obtained by adding the doses from individual spots in the same phase, which are later transformed (via a non-rigid thin-plate spline registration deformation field) to the reference phase before being added to form the final dose distribution.

6.2.3. Breathing models

Breathing signals are used to represent respiratory motion during a treatment fraction, and each of them ultimately results in a different dose distribution. The statistical evaluation of interplay effects requires statistics of the dosimetric quantities of interest using many different dose distributions, requiring a large set of respiratory traces. Except for this study using breathing signals that were deliberately recorded during a long time, the available signals from regular patients are usually short and do not contain enough variability, thus requiring commonly used artificial sinusoidal approximations that potentially introduce errors. Two different types of data-driven breathing models that capture uncertainty and variability in respiratory motion are compared. The first model is based on simple sinusoidal waves (denoted as 'sin' in the remainder of the chapter), while the second model is based on the Adversarial Autoencoder (AAE) algorithm described in Chapter 5.

1. **Sinusoidal model**. In the sinusoidal model the respiratory time series is generated by using a sinusoidal function $\sin^{2n}$ as $x(t) = A_0 + A \cdot \sin^{2n}(\pi t/T + \psi)$, where $x(t)$ is the time dependent position of the tumor, A_0 is the position at the beginning of inhale (in centimeters), T is the breathing period (time between two consecutive inhales, in seconds), and A is the amplitude (distance from inhale to exhale, in centimeters). The parameter ψ represents offset in phase, and effectively symbolizes the moment when the treatment starts within the first cycle. This chapter considers the simplest sinusoidals $\sin^{2n}$ with $n = 1$ (Lujan et al., 2003), with constant amplitude and period. Each signal has a different period and amplitude sampled from Gaussian distributions fitted to both the periods $\mathcal{N}(\mu_T, \sigma_T)$ and amplitudes $\mathcal{N}(\mu_A, \sigma_A)$ present in the recorded breathing signal. The parameter A_0 is often fixed and calculated by the average across breathing cycles in previous studies (George et al., 2005; Lujan et al., 2003). In this study A_0 is considered an independent parameter in order to provide the model with extra variability, and its distribution is also considered to be normal fitted to the breathing data $\mathcal{N}(\mu_{A_0}, \sigma_{A_0})$.

2. **AAE model**. The AAE breathing models are based on artificial neural networks. First, an encoder computes a few *latent* parameters (a low dimensional embedding) that uniquely characterize each high-dimensional breathing signal. The number of low-dimensional latent variables is optimally configured. A decoder reconstructs the original breathing signal using the latent variables from the encoder. A training process using a large set of samples ensures that the decoder accurately reconstructs breathing signals and that each of the latent parameters is approximately distributed according to the Gaussian distribution $\mathcal{N}(0,1)$. As previously shown in Chapter 5, using as few as 5 parameters, the AAE breathing models can generate patient-specific realistic breathing signals with high accuracy and variability in period and amplitude, as opposed to the sinusoidal model always yielding regular sinusoidal samples.

Once the models are obtained, artificial breathing signals are generated by sampling model parameters from their distributions, with each parameter combination

resulting in a unique signal. A sinusoidal signal is thereby obtained by sampling a period, amplitude and inhale position from $\mathcal{N}(\mu_T, \sigma_T)$, $\mathcal{N}(\mu_A, \sigma_A)$ and $\mathcal{N}(\mu_{A_0}, \sigma_{A_0})$. For the AAE breathing models, different signals are obtained by sampling the 5 latent parameters from a Gaussian distribution $\mathcal{N}(0,1)$. For both models, the starting phase ψ — the starting point of delivery within the first breathing cycle — is sampled from a uniform distribution $\mathcal{U}(0, 2\pi)$.

6.3. Statistical evaluation of interplay

Testing robustness against interplay effects involves a statistical evaluation using a set of N_b different dose distributions and DVHs corresponding to N_b different breathing samples. A DVH(D) is a function obtained for a given structure of interest that indicates the fraction of volume V that receives a dose greater than or equal to D. The quantity $V_f = \text{DVH}(fD_p)$ indicates the fraction of the target volume that receives at least a certain percentage f of the prescribed dose D_p. Alternatively, the value $D_f = \text{DVH}^{-1}(fV)$ represents the lowest dose received by at least a fraction f of the volume. Typical values for these quantities are used to assess the adequacy of treatment plans, e.g., the D_{98} or dose that 98% of the volume receives. Additionally, the *homogeneity index* (HI) is determines target dose homogeneity, defined as $HI = (D_2 - D_{98})/D_p$, where D_{98} and D_2 are the dose received by the 98% and 2% of the volume. HIs quantify how uniformly the majority of the target volume is irradiated, with lower values indicating smaller differences between the dose delivered to different parts of the target. The $V_{107/95}$ indicates the fraction of the target volume that receives a dose outside the usually clinically accepted interval $(0.95D_p, 1.07D_p)$, and it is calculated as $V_{107/95} = V_{107} + (1 - V_{95})$.

The proposed interplay evaluation is based on comparing the distributions of D_2, D_{98}, HI and $V_{107/95}$, referred to as quantities of interest (Λ) in the remainder of this chapter. Such distributions are approximated using a collection of n_i percentiles Λ_i obtained from the N_b available computed Λ values, which are compiled into a *percentile vector* $\boldsymbol{\delta}_\Lambda = \{\Lambda_i\}_{i=1}^{n_i}$. Subsequently, the similarity between the results of different statistical analyses is assessed by comparing distributions of each quantity of interest Λ via the percentile vectors $\boldsymbol{\delta}_\Lambda$. If different statistical evaluations yield similar distributions, the analysis of interplay and conclusions drawn regarding the quality of the plan will approximately be the same.

Overview After obtaining a treatment plan that satisfies the planning constraints and objectives, the interplay simulation proceeds as follows:

1. N_b different breathing signals are obtained either by randomly sampling the parameters of the breathing models, or by cropping N_b random fragments from the original recorded signal, where the width of the slicing window is equal to the treatment length.

2. Using the N_b signals, treatment plan information and machine parameters, N_b interplay dose distributions are calculated. Each dose distribution results in a DVH from which the HI, $V_{107/95}$, D_2 and D_{98} are calculated. For each patient, the final robustness evaluation is based on first calculating 1000 interplay dose

distributions using fragments of the recorded signal, and subsequently analyzing the difference in $\boldsymbol{\delta}_\Lambda$ between 4DCT and ITV plans.

3. Distributions are numerically compared using the relative distribution error (RDE). For a quantity of interest and its corresponding vector $\boldsymbol{\delta}_\Lambda$, the RDE quantifies the difference between two distributions as

$$\text{RDE}(\boldsymbol{\delta}^{N_b}_{\Lambda,1}, \boldsymbol{\delta}^{N_b}_{\Lambda,2}) = \frac{1}{n_i} \sum_{i=1}^{n_i} \frac{\left|\Lambda_{i,1} - \Lambda_{i,2}\right|}{\Lambda_{ref}} \times 100. \tag{6.1}$$

where $n_i = 3$ (median, 2 and 98 percentiles), and the reference value for the quantity of interest Λ_{ref} is used to compute the relative error and is obtained from a single interplay dose distribution corresponding to a sinusoidal with average period, amplitude and initial inhale position.

4. A series of experiments evaluate how using a limited number of samples, using artificial signals or ignoring breathing hysteresis compromises evaluation accuracy:

   i) Distributions over the quantities of interest are computed for a different number of breathing samples $N_b = \{20, 50, 100, 500, 1000\}$ of the recorded signal. Two independent statistical analyses for each number of breathing samples N_b using a different subset of N_b interplay dose distributions, result in two different vectors $\boldsymbol{\delta}^{N_b}_{\Lambda,1}$ and $\boldsymbol{\delta}^{N_b}_{\Lambda,2}$ that are compared via the RDE.

   ii) The effect of using artificial breathing signals from the sin or AAE models is determined by computing the RDE between their corresponding $\boldsymbol{\delta}^{1000}_{\Lambda,sin}$ or $\boldsymbol{\delta}^{1000}_{\Lambda,AAE}$ and the reference $\boldsymbol{\delta}^{1000}_{\Lambda,real}$ from the real recorded signals, where all the statistics are calculated using 1000 samples.

   iii) Finally, the dosimetric impact of disregarding motion hysteresis is assessed via the RDE between the results of two different interplay evaluations with 1000 samples: one including 8 breathing phases, and the other only 5 phases identical during inhale and exhale.

6.4. Results

6.4.1. Robustness of 4DCT and ITV plans against interplay

Reliable statistical analyses allow direct assessment of the robustness of treatment plans, as well as comparison between different planning approaches. To illustrate this, Figure 6.2 shows the distribution of D_{98}, HI and $V_{107/95}$ corresponding to each plan and patient combination, for both single fractions and a fully fractionated treatment. As seen in the top row, the 4DCT plans result in higher D_{98} values regardless of fractionation, tumor size and breathing amplitude, while ITV plans systematically fail to meet the clinical constraints. Likewise, the HI and $V_{107/95}$ (middle and bottom rows) are consistently lower in 4DCT treatments, indicating that the delivered dose distribution is more homogeneous and the target receives a dose within the clinically acceptable limits in most of the scenarios.

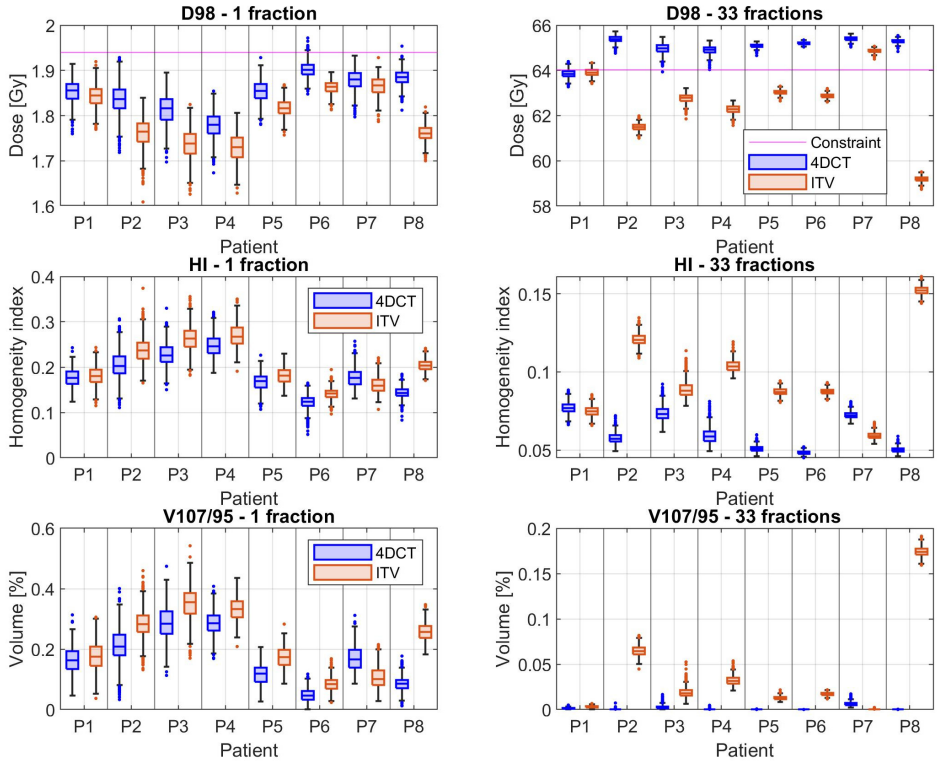
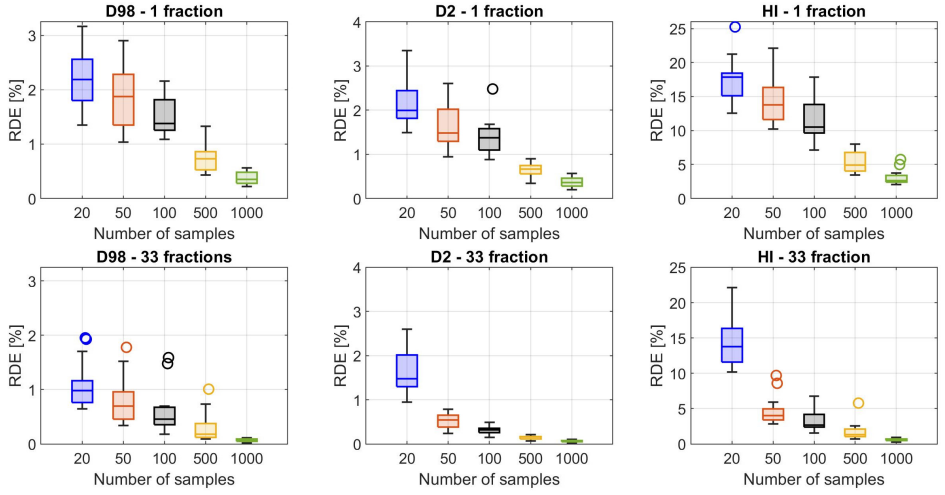


Figure 6.2: **Interplay evaluation results.** The distribution of 1000 different D_{98}, HI and $V_{107/95}$ CTV values is shown for every patient and plan, and for (left) individual fractions and (right) fully fractionated treatments, using the real recorded signal. The pink line in the top row denotes the clinical near-minimum CTV dose constraint. For each box, the centered line represents the median, while the boundaries correspond to the upper and lower quartiles (25th and 75th percentiles), and the individual points outside the whiskers are outliers. Higher HI and $V_{107/95}$ correspond to more heterogeneous dose distributions with hot and cold spots.

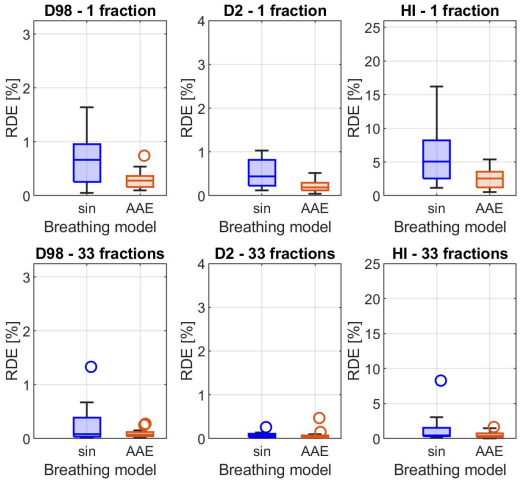
6.4.2. Influence of sample size, breathing models and hysteresis.

A relevant question is how many different interplay dose distributions are necessary in order to perform a statistical analysis that yields reliable results. For this reason, independent statistical analyses are performed using different sub-sample sizes selected according to published results (Engwall, Glimelius, and Hynning, 2018; Engwall, Fredriksson, and Glimelius, 2018; Jakobi et al., 2018; Seco et al., 2009). Figure 6.3a shows a reduction in RDE as more breathing samples are used to calculate the statistics, confirming that the distributions gradually converge. The RDE illustrates how much the results from two statistical analysis could vary for a given sample size simply due to chance, being higher for single fraction analyses using <100 samples.

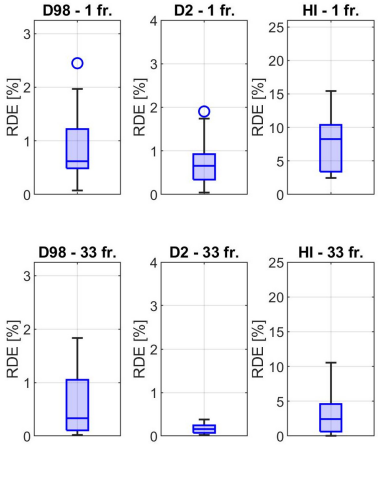
Figure 6.3b shows a comparison of the error introduced by using artificial signals from the sin and AAE models instead of the recorded signals from the patient, showing that for single fractions AAE model slightly outperforms the sin model, but the differ-



(a) Impact of using different sample sizes to calculate interplay statistics during evaluation.



(b) Dosimetric effect of using breathing signals from a model instead of the real recorded signals.



(c) Impact of ignoring breathing hysteresis during the evaluation.

Figure 6.3: **Effect of the evaluation parameters on the interplay statistics.** The reported RDE represents the difference between two different distributions of a quantity of interest, in this case the D_{98}, D_2 and HI, and can be used to determine whether two independent interplay evaluations yield the same results. Each box includes all RDE values across patients and planning approaches, showing the dosimetric impact of varying one of the following evaluation parameters, while keeping the rest fixed: (a) the number of samples used to compute the statistics, (b) the breathing signal model, and (c) the absence of respiratory motion hysteresis, with identical inhale and exhale. Each variation results in an independent distribution, which is compared to either (a) a duplicate distribution obtained using the same settings, or (b,c) a reference distribution obtained from a statistical analysis using 1000 samples from the recorded signal and considering breathing hysteresis. Each box contains the median in the center and the upper and lower quartiles (25th and 75th percentiles) as box boundaries, with outliers represented as individual points outside the whiskers.

ences between models are minimal in fully fractionated treatments. Figure 6.3c shows the effect of disregarding breathing hysteresis. Although using a model of respiratory motion results in non-negligible errors in single fraction doses, its dosimetric impact is much lower than using few samples in the evaluation.

6.4.3. Interplay dose dependence on breathing parameters.

In order to investigate the relationship between small changes in the breathing parameters and interplay doses, Figure 6.4 shows the dose and D_{98} for different amplitudes, periods and starting phases of a sinusoidal breathing signal. Each of the parameters is varied independently, one at a time, leaving the rest fixed. Amplitude changes have a lower and less fluctuating effect on the dose compared to changes in period or starting phase. The latter affect the time structure of treatment delivery, and as a result, small variations can effectively shift the breathing phases in which subsequent spots are delivered, with a great local impact on voxel doses. On the other hand, changes in amplitude are responsible for shifting only few spots to neighboring phases, hence inducing smaller changes in the delivered dose. Repainting contributes to better target coverage and reducing the magnitude of interplay effects, as indicated by the lower spread of voxel doses around the target 2 Gy fraction dose, and the higher D_{98} values. For fractions delivered without repainting, period changes can result in up to 50% variations over the target dose and 4 Gy differences in D_{98}, as seen in the top row of Figure 6.4.

6.5. Discussion

The results indicate that 4DCT plans outperform ITV plans in terms of dose coverage and homogeneity, regardless tumor size and breathing amplitude. Using a fully fractionated robust 4DCT treatment planning approach with the exhale, inhale and mid-ventilation phases may be sufficient to compensate the detrimental effect of breathing motion, as indicated by the high D_{98} values and lower HI and $V_{107/95}$ shown in Figure 6.2. Contrariwise, robust ITV plans seem to fail to meet the required dose constraints in IMPT lung cancer treatments, and may require the use of additional margins or motion mitigation techniques, or increased robustness settings. Finally, as seen in the left plots of Figure 6.2 by the lower D_{98} and higher HI and $V_{107/95}$ in single fraction doses, interplay effects seem to be aggravated by larger amplitudes and tumor sizes (P3 and P4).

Accuracy of the interplay evaluation Among all possible simplifications (i.e., using few breathing samples or ignoring hysteresis), the error of using artificial signals seems to be the lowest, where the AAE breathing model clearly outperforms the sin model at a considerably higher computational cost and more patient-specific data. Based on the results, a simple sinusoidal model may be sufficiently accurate in fully fractionated treatments as long as the parameter distribution is patient specific. Disregarding hysteresis, however, introduces errors that can be as high as 2.5% of the D_{98} of the delivered dose in some cases, even when considering the smoothing effect of fractionation (Figure 6.3c).

Using a few realizations (<100) of interplay dose distributions in order to evaluate

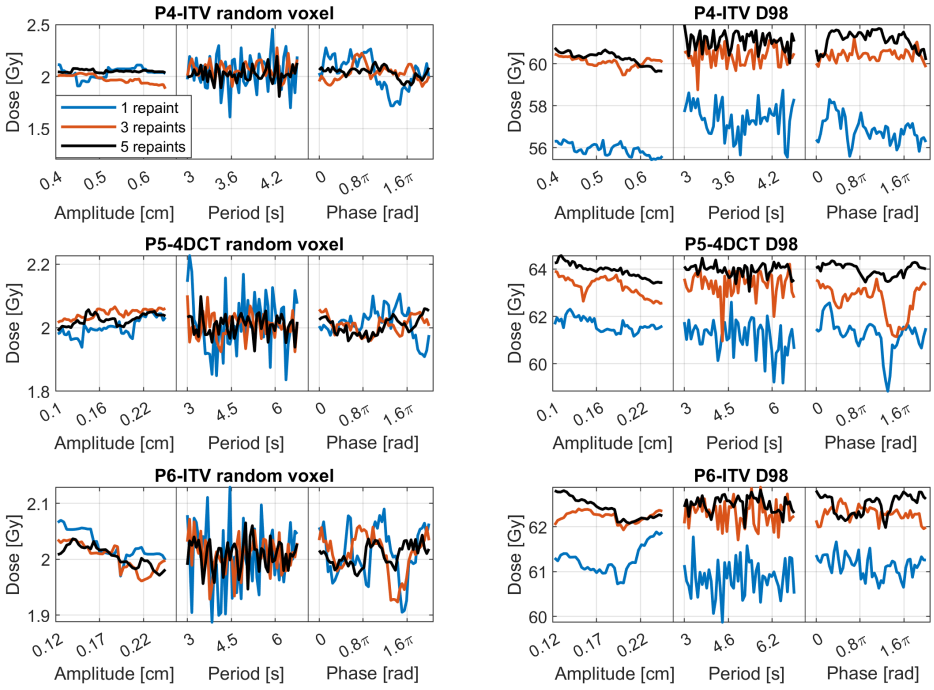


Figure 6.4: **Effect of breathing parameters on the final dose distribution.** (Left) Fraction dose in a random CTV voxel and dependence on the parameters of a sinusoidal breathing signal, for a different number of repaintings and 3 patients (from top to bottom, patient 4, patient 5 and patient 6). Blue lines correspond to dose distributions without repainting, whereas read and black lines indicate 3 and 5 repaints. Repainting smooths out interplay effects in the local fraction doses and reduces the fluctuations around the 2 Gy target dose. As a result Similar results are obtained for other randomly selected voxels. (Right) D_{98} dependence on small breathing variations. Repainted dose distributions fluctuate less and result in better target coverage, as indicated by the higher D_{98} values.

interplay effects lacks statistical power. The presented results from lung cancer patients show that at least 500 different interplay dose distributions are needed to achieve the same level of error as the one introduced by other simplifications such as using sinusoidal breathing or no hysteresis, also for fractionated delivery. Only when >500 samples are used the differences are generally below 1% of the reference dose and 5% of the HI values, which can be limiting with computationally expensive interplay dose calculation models. Most of the previous studies are short on samples: ranging from 300 different simulated treatments (Seco et al., 2009) to as few as 10 samples (Engwall, Glimelius, and Hynning, 2018), 20 samples (Engwall, Fredriksson, and Glimelius, 2018) or 64 samples (Jakobi et al., 2018). Other published works do not explicitly reference this number but use few realizations with different starting phases (Kardar et al., 2014; Li et al., 2014), or are based on 30 dose distributions weighted by their probability of occurrence (Kraus et al., 2011).

The combined smoothing benefits of repainting and fractionation in lung cancer

treatments has been previously investigated (Engwall, Glimelius, and Hynning, 2018; Li et al., 2014; Seco et al., 2009) and is further exemplified in Figure 6.4. We can therefore assume that the worst-case scenario in the interplay dose degradation occurs for single fraction dose distributions with no motion mitigation, which explains the fact that the errors in the statistical evaluation diminish as fractionation increases. As a result, the relative errors between distributions may be minimal if repainting or other mitigation technique are applied, requiring fewer samples to obtain reliable results and thus compensating for the longer calculation times needed to simulate repainting.

This chapter focuses on IMPT lung cancer patients, that represent a worst-case scenario for breathing motion. Other treatment modalities such as SBRT or hypofractionated IMPT treatments deliver the dose more intensely using less fractions. The considerably higher dose per fraction could exacerbate interplay effects (and in particular may cause bigger inhomogeneities in the dose), especially in terms of biological dose. For such cases, evaluating the dose degradation due to motion using only few samples could lead to even larger inaccuracies.

Dose dependence on breathing parameters The results in Figure 6.4 demonstrate the beneficial effect of repainting in both smoothing out great local dose variations and improving target coverage, as seen in the reduced fluctuations around the 2 Gy target dose that translate into higher D_{98} values. However, rescanning alone does not fully mitigate interplay effects, in concordance with previous results (Y. Zhang et al., 2018), resulting in local doses that may vary up to 10% of the target dose and D_{98} values that are always below the constraint. Delivery without repainting results in dose fluctuations amounting up to 50% of the target fraction dose. This effect may be caused by the fact that small period and starting phase changes can simultaneously shift a significant number of subsequent spots, the effect being more dramatic for the parts of the tumor that receive dose only from few individual pencil beams, or spots delivered later within a fraction. The results are consistent with previous findings for IMRT dynamic delivery (Kissick et al., 2005) that demonstrate the detrimental effect of intra-fraction random changes of the breathing parameters. We can further hypothesize that these results are independent of the 4DCT resolution: adding more 4DCT phases during evaluation results in some spots shifting to consequent phases with similar anatomy, and thus the effect may not as dramatic as with period or phase changes, where small variations may cause the delivery of a later spot to shift from full inhale to exhale.

The degrading effects of time changes can also impact currently applied clinical protocols. Most of the treatment centers establish their criteria for interplay mitigation in terms of breathing amplitude (e.g., no mitigation is considered if the breathing amplitude for a given patient is lower than 5 mm). The results show that not only does period influence the fluctuating behavior but it also highly affects the degree of degradation of the dose. Thus, more research is needed to determine whether making planning or clinical decisions purely based on amplitude criteria suffices, and whether strategies that weigh both period and amplitude changes offer additional benefits.

Limitations The most limiting design choice is the use of a single 4DCT, under the assumption that it captures the variations in patient anatomy from full inhale (maximum

amplitude) to full exhale and breathing hysteresis, as well as the mismatch between 4DCT and the signal motion surrogate. While this assumption to speeds up and simplifies the interplay dose distribution calculation, some irregularities may not be captured in the 4DCT, for which a bio-mechanical model could be used to simulate hiccups or coughs as in (Boye et al., 2013). Similarly, the temporal resolution of the 4DCT is significantly lower than that of the spot delivery. Although using a coarser resolution is not expected to be as significant as disregarding hysteresis, the most detailed interplay simulations should be based on variable time dependent 4DCT data with finer temporal resolution. Finally, the accuracy and calculation times of the presented interplay dose calculation method ultimately depend on that of the dose engine and the registration algorithm. Traditional (usually slow) image registration methods have been been recently outperformed by data-driven approaches (Balakrishnan et al., 2018, 2019; Dalca, Balakrishnan, et al., 2019; Dalca, Yu, et al., 2019). Similarly, recent deep learning based dose engines (Chapter 2 & 3, Pastor-Serrano and Perkó, 2022b; C. Wu et al., 2021) have been shown to overcome the speed limitations of Monte Carlo methods, while offering better performance than the pencil beam algorithms commonly used in the clinics.

6.6. Summary

This chapter presents a practical method to simulate dose delivery under motion interplay effects and assess treatment robustness based on hundreds of (sampled) breathing signals. The proposed statistical evaluation shows that ITV plans systematically fall behind their computationally more expensive 4DCT robust counterpart, regardless of tumor size and breathing amplitude. After analyzing the error introduced by simplifications such as neglecting motion hysteresis or using few interplay scenarios and sinusoidal breathing signals, we can conclude that the statistical analysis of fully fractionated treatments requires at least 500 different dose distributions corresponding to 500 different samples of regular sinusoidal breathing (based on patient-specific parameter distributions) with hysteresis to yield acceptable precision. This chapter further demonstrates that small breathing period variations have a highly non-linear effect on local dose deposition and can cause local doses to fluctuate up to 50% of the target fraction dose.

7

Conclusion

This thesis presents methods to simulate typical anatomical uncertainties (e.g., breathing motion or relative organ movements) that occur during the delivery of a photon and proton radiation therapy treatment, and to subsequently quickly calculate the dose delivered under such uncertainty. The main motivation is to enable robustness evaluation of treatment plans against intra-fraction and inter-fraction anatomical changes in clinically feasible times by (i) quantifying and simulating the errors for a given patient, and (ii) quickly estimating the dose delivered in such scenarios.

7.1. Outcomes of this dissertation

The main contribution of this thesis is in the fields of fast dose prediction algorithms and probabilistic models to generate anatomical uncertainties, which are missing pieces towards clinical, fast robust treatment planning, robustness evaluation and online adaptation. The most important findings can be summarized as follows:

- As a solution to slow physics-based dose calculation algorithms, a deep learning Dose Transformer Algorithm (DoTA) was developed, predicting proton pencil beam doses in few milliseconds with accuracy close to Monte Carlo (MC) based dose calculation. As demonstrated by the fast prediction times and high gamma pass rates, DoTA outperforms existing approaches and represents a new state-of-the-art from which current clinical practice could benefit in numerous aspects. The small number of potential geometries currently used to clinically evaluate treatment plan robustness — which is primarily limited by the speed of the dose calculation algorithm — can be extended with many additional samples, capturing a more diverse and realistic set of uncertainties (e.g., inter-fraction and intra-fraction geometrical variations). Moreover, DoTA's millisecond speed further allows calculating probabilistic metrics to be used during probabilistic optimization. DoTA's capability to quickly and accurately estimate fraction dose distributions based on pre-treatment daily computed tomography (CT) images could transform dosimetric quality assurance protocols, enabling a fast, independent

dose calculation based on the machine parameters stored in the delivery system, and allowing to directly compare the planned and estimated doses, which is a necessary prerequisite for online adaptation of plans (Albertini et al., 2020; Jagt et al., 2017, 2018). Most crucially, by pre-computing the input volumes and updating their CT values in real time, the millisecond speed for individual pencil beam dose calculation makes DoTA well-suited for real-time dose prediction during radiation delivery, which could in the future enable real-time adaptation if coupled to fast imaging and re-optimization algorithms.

- The original DoTA model was extended to predict dose distribution from broad photon beams, extending the its speed benefits to photon treatments. Conditioned only on the beam shape projection and the input CT scan, the new improved DoTA (iDoTA) also outperforms other deep learning dose calculation approaches, representing a state of the art in photon dose prediction, especially when calculating full dose distributions from volumetric modulated arc therapy (VMAT) treatments. Like its proton DoTA counterpart, iDoTA is a versatile algorithm that can drastically reduce computing times in any application involving repeated calculation of dose distributions, e.g., checking plan robustness by quickly predicting the dose in each of the many possible error scenarios or anatomical variations of the patient (Tilly et al., 2017). However, the most straightforward application of iDoTA is reducing computation times in VMAT plan dose calculation, offering 10x faster predictions than clinical software, potentially massively speeding up planning and evaluation in VMAT treatments consisting of many control points, such as pediatric total body irradiation treatments.

- To enable margin personalization and robust optimization and evaluation protocols that take into account inter-fraction anatomical changes, a method to quantify inter-fraction deformations is required. For this, a daily anatomy model (DAM) was introduced, being able to simulate the organ movements and shifts seen in prostate patients. The model generates patient-specific deformations by selectively querying and sampling deformation fields based on the ones seen for patients with similar anatomies within a population. DAM's main application in robust treatment planning and robustness evaluation involves generating patient anatomies on which corresponding dose distributions will be calculated. With prediction times of few milliseconds per generated anatomy, DAM offers huge speed-up possibilities for plan evaluation when coupled to fast dose calculation algorithms such as DoTA or iDoTA. Furthermore, few (3-5) representative scenarios corresponding to points around mean of the posterior distribution can be sampled to be used for scenario-based anatomically robust optimization of proton treatment plans, or for patient-specific target margin individualization or optimization in photon plans, which may translate into a dosimetric advantage. In principle, the same generic modeling framework could be applied to any treatment site (e.g., pancreatic tumors), provided that a dataset with planning and repeated imaging is available, e.g., CT or magnetic resonance (MR). Other applications involve formulating anatomical robustness recipes (van der Voort

et al., 2016) that jointly cover range, setup and anatomical uncertainties.

- Focusing on intra-fraction breathing variations, a semi-supervised probabilistic framework was applied to model breathing signals. When coupled to the presented pre-processing and post-processing steps, the novel semi-supervised adversarial autoencoder algorithm can accurately generate realistic, artificial breathing signals. Besides applications to simulate breathing during the delivery of radiation therapy treatments, the SAAE framework can in principle be applied for computer aided diagnosis of breathing abnormalities, as well as for dataset augmentation when the available data for a patient is scarce. An example is classifying breathing irregularities and generating additional samples that present the identified irregularity. One of the advantages of training the proposed framework in a semi-supervised way is the possibility to build such models requiring only a small subset of labeled data. The proposed model can in principle be applied to any other kind of biomedical data that shows a repetitive or periodic structure, like electrocardiogram (ECG) signals composed of well-defined intervals randomly varying in amplitude and length. The added advantage of the proposed generative approach with respect to other models in the literature that do not explicitly model the data distribution (Delaney et al., 2019; F. Zhu et al., 2019) is the possibility to map the data samples to specific regions or classes in the latent space, enabling classification and generation of class-specific data by sampling latent variables from the desired regions.

- To demonstrate the application of the breathing signal models in proton therapy treatments, an interplay effect quantification tool was introduced, calculating the dose dynamically delivered in moving lung tumors. The method can be used to evaluate robustness of treatment plans against intra-fraction breathing movements by sampling many different breathing signals (much more realistic than currently used sinusoidal approximations) and estimating the dose delivered for each scenario. The interplay calculation tool was applied to compare robustness of 4DCT and internal target volume (ITV) based treatment plans, showing the clear superiority of 4DCT robust treatment planning. Even though only 8 patients were used for such comparison, the results show that ITV plans consistently fail to meet the prescribed clinical objectives, indicating that the current ITV planning methodology may under-perform and need further scrutiny.

7.2. Recommendations

The presented models allow fast anatomy change modeling and corresponding dose calculation, thus solving a challenge in robust planning, robustness evaluation and possibly online adaptation. Future research should focus on extending the capabilities of the dose calculation algorithms, while further validating the intra-fraction and inter-fraction models to enable their implementation in the clinic.

To be a truly generic dose calculation tool such as physics-based algorithms that can be applied in any clinic with minimal effort, a single DoTA model should be able to process machine and beam characteristics, such as the angle or shape of the beam, which could be added as additional input tokens in the sequence, as it is done with

the beam energy. DoTA's spectrum of applications can be extended to predicting additional quantities, e.g., particle flux, estimating radiobiological weighted dose — typically significantly slower to simulate with MC methods than pure physical dose calculation — and potentially even speeding up DNA damage MC simulation tools (Faddegon et al., 2020; Perl et al., 2012). A clinically highly relevant follow-up study is to include geometries with metallic implants in the training dataset and ensuring prediction accuracy in such challenging geometries too. DoTA also offers great potential to speed up dose calculation times in heavy ion treatments with particles such as carbon and helium. Such heavy ions share similar, mostly forward scatter physics, with MC dose calculations that often take much longer to simulate given the larger amount of secondary particles generated as the beam travels through the patient.

Similar to DoTA, future work to increase iDoTA's generalization capabilities could focus on including 2D aperture shapes, machine or beam characteristics via tokens into the sequence, removing the dependence on the current 3D input beam intensity shape, and thus potentially reducing computation times. For MR image-guided treatments, the magnetic field strength could even be added as an additional token in the sequence, similar to the energy token in previous transformer-based proton dose prediction models. Compared to previous works (Tsekas et al., 2021; F. Xiao et al., 2022), such deep learning model would be the first to predict high accuracy dose distributions in few milliseconds given an input image and the magnetic field strength.

Several modifications can also improve and extend DAM capabilities. Focusing on performance, using multiple resolution levels (Kohl et al., 2019; Krebs et al., 2019; Sønderby et al., 2016) could increase the quality of the modeled deformations, by capturing different types of the deformations per level, e.g., the coarser resolution level modeling global deformations, with finer levels focusing on more specific, local movements. Further extensions include adding temporal dependence for treatments where patients' anatomies change following a clear pattern during, e.g., simulating anatomies changing gradually during the fraction due to periodical breathing or over the different fractions of the treatment as a result of progressive changes such as tumor shrinkage. Compared to 4DCT images, such time-dependent model would offer continuous anatomy changes through a breathing cycle, which could be coupled to interplay effect simulation tools in Chapter 6 and the breathing signal models in Chapter 5 to obtain treatment plans that mitigate the detrimental effect of movement during delivery. Furthermore, to solve the problem of working with masked rectum structures, a similar probabilistic generative model based on a variational autoencoder could fill the deformed contours with realistic gray values, enabling DAM's application to photon workflows using beams traversing the rectum.

For fast robustness evaluation or robust optimization of treatment plans against inter-fraction motion, DAM must be coupled to a fast dose calculation such as DoTA. Alternatively, polynomial chaos expansion (PCE) methods can be used to establish a dependence between latent variables in DAM and dose values, enabling quick prediction of the dose for each latent value combination. Such PCE methods have been previously successfully applied to model the dependence between range and positioning errors and the corresponding dose (Perkó et al., 2016). PCE methods can simulate thousands of dose distributions for many different error scenarios in seconds, but they

are built using few tens to few hundred ground-truth scenarios calculated with the reference dose calculation algorithm. As a result, DoTA may be better suited for time-sensitive applications requiring few representative error scenarios such as robust optimization, while PCE may be more advantageous in robustness evaluation protocols necessitating detailed statistics. Extending the PCE methodology to include intra-fraction movements could further result in robustness recipes producing treatment plans that are jointly robust against most position, range and organ motion errors. Most importantly, to finally implement DAM in the clinic, a thorough quality assurance protocol that evaluates prediction robustness (different from treatment plan robustness) is required e.g., by training several models using different datasets from different institutions and machines, and evaluating result similarity on a common external test dataset.

One of the main bottlenecks preventing the presented interplay calculation tool to be applied in the clinics, is the speed of the image registration. For testing plan robustness against inter-fraction anatomical uncertainties together with positioning and range errors, the dose needs to be recalculated and deformed in each scenario, implying long computation times of tens of minutes due to the required image registration steps between the reference phase and each of the breathing phases. Recently, deep learning models have achieved state-of-the-art accuracy and prediction times in registration tasks too. Thus, substituting the original pencil beam dose and cubic spline registration algorithms in the presented interplay calculation tool by deep learning models such as DoTA or Voxelmorph (Balakrishnan et al., 2018, 2019; Dalca, Balakrishnan, et al., 2019; Dalca, Yu, et al., 2019) can reduce the simulation of a dynamically delivered dose from the current $\approx$ 8 minutes to several seconds, providing the needed speed to evaluate interplay robustness in clinical times.

Offering fast dose calculation and anatomical change modeling, the presented models are some of the missing pieces needed for clinical automated adaptation of treatment plans. Within the online adaptation workflow, DoTA and iDoTA can be used to quickly calculate the dose delivered in the new anatomy using the original plan and subsequently re-optimize plans or restore doses, while the interplay calculation tool can be used to quickly evaluate robustness. To be finally applied in the clinic, these models should be coupled to automatic segmentation and registration models and fast optimization solvers, while being continuously monitored and retrained using the available clinical data.

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., … Research, G. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems* (tech. rep.).

Abreu, M., Fred, A., Valente, J., Wang, C., & Plácido da Silva, H. (2020). Morphological autoencoders for apnea detection in respiratory gating radiotherapy. *Computer Methods and Programs in Biomedicine*, *195*, 105675.

Acharya, U. R., Fujita, H., Lih, O. S., Hagiwara, Y., Tan, J. H., & Adam, M. (2017). Automated detection of arrhythmias using different intervals of tachycardia ECG segments with convolutional neural network. *Information Sciences*, *405*, 81–90.

Acharya, U. R., Fujita, H., Oh, S. L., Hagiwara, Y., Tan, J. H., & Adam, M. (2017). Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals. *Information Sciences*, *415-416*, 190–198.

Aerts, H. J. W. L., Rios Velazquez, E., Leijenaar, R. T. H., Parmar, C., Grossmann, P., Carvalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., Rietveld, D., Hoebers, F., Rietbergen, M. M., Leemans, C. R., Dekker, A., Quackenbush, J., Gillies, R. J., & Lambin, P. (2015). Data From NSCLC-Radiomics-Genomics [Version Number: 1 Type: dataset].

Aerts, H. J. W. L., Velazquez, E. R., Leijenaar, R. T. H., Parmar, C., Grossmann, P., Carvalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., Rietveld, D., Hoebers, F., Rietbergen, M. M., Leemans, C. R., Dekker, A., Quackenbush, J., Gillies, R. J., & Lambin, P. (2014). Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*, *5*(1), 4006.

Ahnesjö, A. (1989). Collapsed cone convolution of radiant energy for photon dose calculation in heterogeneous media. *Medical Physics*, *16*(4), 577–592.

Albertini, F., Matter, M., Nenoff, L., Zhang, Y., & Lomax, A. (2020). Online daily adaptive proton therapy. *The British Journal of Radiology*, *93*(1107), 20190594.

Antolak, J. A., Rosen, I. I., Childress, C. H., Zagars, G. K., & Pollack, A. (1998). Prostate target volume variations during a course of radiotherapy. *International Journal of Radiation Oncology, Biology, Physics*, *42*(3), 661–672.

Arsigny, V., Commowick, O., Pennec, X., & Ayache, N. (2006). A Log-Euclidean Framework for Statistics on Diffeomorphisms. In R. Larsen, M. Nielsen, & J. Sporring (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006* (pp. 924–931). Springer.

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *NeuroImage*, *38*(1), 95–113.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer Normalization.

Badawi, A. M., Weiss, E., Sleeman IV, W. C., Yan, C., & Hugo, G. D. (2010). Optimizing principal component models for representing interfraction variation in lung cancer radiotherapy. *Medical Physics*, *37*(9), 5080–5091.

Bai, T., Wang, B., Nguyen, D., & Jiang, S. (2021). Deep dose plugin: Towards real-time Monte Carlo dose calculation through a deep learning-based denoising algorithm. *Machine Learning: Science and Technology*, *2*(2), 25033–25033.

Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J., & Dalca, A. V. (2018). An Unsupervised Learning Model for Deformable Medical Image Registration, 9252–9260.

Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J., & Dalca, A. V. (2019). VoxelMorph: A Learning Framework for Deformable Medical Image Registration. *IEEE Transactions on Medical Imaging*, *38*(8), 1788–1800.

Barragán-Montero, A. M., Nguyen, D., Lu, W., Lin, M.-H., Norouzi-Kandalan, R., Geets, X., Sterpin, E., & Jiang, S. (2019). Three-dimensional dose prediction for lung IMRT patients with deep neural networks: Robust learning from heterogeneous beam configurations. *Medical Physics*, *46*(8), 3679–3691.

Baumgartner, C. F., Tezcan, K. C., Chaitanya, K., Hötker, A. M., Muehlematter, U. J., Schawkat, K., Becker, A. S., Donati, O., & Konukoglu, E. (2019). PHiSeg: Capturing Uncertainty in Medical Image Segmentation. *arXiv:1906.04045 [cs, eess, stat]*.

Bernatowicz, K., Geets, X., Barragan, A., Janssens, G., Souris, K., & Sterpin, E. (2018). Feasibility of online IMPT adaptation using fast, automatic and robust dose restoration. *Physics in Medicine and Biology*, *63*(8), 085018.

Bernatowicz, K., Zhang, Y., Perrin, R., Weber, D. C., & Lomax, A. J. (2017). Advanced treatment planning using direct 4D optimisation for pencil-beam scanned particle therapy. *Physics in Medicine and Biology*, *62*(16), 6595–6609.

Bernier, J., Hall, E. J., & Giaccia, A. (2004). Radiation oncology: A century of achievements. *Nature Reviews Cancer*, *4*(9), 737–747.

Bert, C., & Durante, M. (2011). Motion in radiotherapy: Particle therapy. *Physics in Medicine and Biology*, *56*(16), R113–R144.

Bert, C., Gemmel, A., Saito, N., & Rietzel, E. (2009). Gated Irradiation With Scanned Particle Beams. *International Journal of Radiation Oncology Biology Physics*, *73*(4), 1270–1275.

Bert, C., Grözinger, S., & Rietzel, E. (2008). Quantification of interplay effects of scanned particle beams and moving targets. *Physics in Medicine and Biology*, *53*(9), 2253–2265.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877.

Boda-Heggemann, J., Knopf, A.-C., Simeonova-Chergou, A., Wertz, H., Stieler, F., Jahnke, A., Jahnke, L., Fleckenstein, J., Vogel, L., Arns, A., Blessing, M., Wenz, F., & Lohr, F. (2016). Deep Inspiration Breath Hold—Based Radiation Therapy: A Clinical Review. *International Journal of Radiation Oncology Biology Physics*, *94*(3), 478–492.

Bondar, L., Intven, M., Burbach, J. P., Budiarto, E., Kleijnen, J. P., Philippens, M., Van Asselen, B., Seravalli, E., Reerink, O., & Raaymakers, B. (2014). Statistical modeling of CTV motion and deformation for IMRT of early-stage rectal cancer. *International Journal of Radiation Oncology Biology Physics*, *90*(3), 664–672.

Bortfeld, T., Jokivarsi, K., Goitein, M., Kung, J., & Jiang, S. B. (2002). Effects of intra-fraction motion on IMRT dose delivery: Statistical analysis and simulation. *Physics in Medicine and Biology*, *47*(13), 2203–2220.

Botas, P., Kim, J., Winey, B., & Paganetti, H. (2018). Online adaption approaches for intensity modulated proton therapy for head and neck patients based on cone beam CTs and Monte Carlo simulations. *Physics in Medicine and Biology*, *64*(1), 015004.

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Jozefowicz, R., & Bengio, S. (2016). Generating Sentences from a Continuous Space. *arXiv:1511.06349 [cs.LG]*.

Boye, D., Lomax, T., & Knopf, A. (2013). Mapping motion from 4D-MRI to 3D-CT for use in 4D dose calculations: A technical feasibility study. *Medical Physics*, *40*, 061702.

Boyer, A., & Mok, E. (1985). A photon dose distribution model employing convolution calculations. *Medical Physics*, *12*(2), 169–177.

Breedveld, S., Storchi, P. R. M., Voet, P. W. J., & Heijmen, B. J. M. (2012). iCycle: Integrated, multicriterial beam angle, and profile optimization for generation of coplanar and noncoplanar IMRT plans. *Medical Physics*, *39*(2), 951–963.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 2020-Decem*.

Bruveris, M., & Holm, D. D. (2015). *Geometry of image registration: The diffeomorphism group and momentum maps* (Vol. 73).

Budiarto, E., Keijzer, M., Storchi, P. R., Hoogeman, M. S., Bondar, L., Mutanga, T. F., De Boer, H. C., & Heemink, A. W. (2011). A population-based model to describe geometrical uncertainties in radiotherapy: Applied to prostate cases. *Physics in Medicine & Biology*, *56*(4), 1045–1061.

Burda, Y., Grosse, R., & Salakhutdinov, R. (2016). Importance Weighted Autoencoders. *arXiv:1509.00519 [cs.LG]*.

Chen, X., Kingma, D. P., Salimans, T., Duan, Y., Dhariwal, P., Schulman, J., Sutskever, I., & Abbeel, P. (2017). Variational lossy autoencoder. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 1–17.

Chen, X., Cheng, Z., Wang, S., Lu, G., Xv, G., Liu, Q., & Zhu, X. (2021). Atrial fibrillation detection based on multi-feature extraction and convolutional neural network for processing ECG signals. *Computer Methods and Programs in Biomedicine*, *202*, 106009.

Chen, X., Men, K., Li, Y., Yi, J., & Dai, J. (2019). A feasibility study on an automated method to generate patient-specific dose distributions for radiotherapy using deep learning. *Medical Physics*, *46*(1), 56–64.

Chu, M., Zinchenko, Y., Henderson, S. G., & Sharpe, M. B. (2005). Robust optimization for intensity modulated radiation therapy treatment planning under uncertainty. *Physics in Medicine and Biology*, *50*(23), 5463.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3D U-net: Learning dense volumetric segmentation from sparse annotation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9901 LNCS*, 424–432.

Cimr, D., Studnicka, F., Fujita, H., Tomaskova, H., Cimler, R., Kuhnova, J., & Slegr, J. (2020). Computer aided detection of breathing disorder from ballistocardiography signal using convolutional neural network. *Information Sciences*, *541*, 207–217.

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L., & Prior, F. (2013). The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *Journal of Digital Imaging*, *26*(6), 1045–1057.

Cohilis, M., Sterpin, E., Lee, J. A., & Souris, K. (2020). A noise correction of the gamma index method for Monte Carlo dose distribution comparison. *Medical Physics*, *47*(2), 681–692.

Coste-Manière, È., Olender, D., Kilby, W., & Schulz, R. A. (2005). Robotic whole body stereotactic radiosurgery: Clinical advantages of the Cyberknife® integrated system. *The International Journal of Medical Robotics and Computer Assisted Surgery*, *1*(2), 28–39.

Court, L., Wagar, M., Berbeco, R., Reisner, A., Winey, B., Schofield, D., Ionascu, D., Allen, A. M., Popple, R., & Lingos, T. (2010). Evaluation of the interplay effect when using RapidArc to treat targets moving in the craniocaudal or right-left direction. *Medical Physics*, *37*(1), 4–11.

Court, L. E., Wagar, M., Ionascu, D., Berbeco, R., & Chin, L. (2008). Management of the interplay effect when using dynamic MLC sequences to treat moving targets. *Medical Physics*, *35*(5), 1926–1931.

Dai, B., & Wipf, D. (2019). Diagnosing and Enhancing VAE Models. *arXiv:1903.05789 [cs.LG]*.

Dalca, A. V., Balakrishnan, G., Guttag, J., & Sabuncu, M. R. (2019). Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Medical Image Analysis*, *57*, 226–236.

Dalca, A. V., Yu, E., Golland, P., Fischl, B., Sabuncu, M. R., & Iglesias, J. E. (2019). Unsupervised Deep Learning for Bayesian Brain MRI Segmentation. *arXiv preprint, arXiv:1904.11319 [cs, eess]*.

D'Ascoli, S., Touvron, H., Leavitt, M., Morcos, A., Biroli, G., & Sagun, L. (2021). ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases.

Delaney, A. M., Brophy, E., & Ward, T. E. (2019). Synthesis of Realistic ECG using Generative Adversarial Networks. *arXiv:1909.09150 [eess.SP]*.

den Boer, E., Wulff, J., Mäder, U., Engwall, E., Bäumer, C., Perko, Z., & Timmermann, B. (2021). Technical Note: Investigating interplay effects in pencil beam scanning proton therapy with a 4D XCAT phantom within the RayStation treatment planning system. *Medical Physics*, *48*(3), 1448–1455.

Deurloo, K. E. I., Steenbakkers, R. J. H. M., Zijp, L. J., de Bois, J. A., Nowak, P. J. C. M., Rasch, C. R. N., & van Herk, M. (2005). Quantification of shape variation of prostate and seminal vesicles during external beam radiotherapy. *International Journal of Radiation Oncology*Biology*Physics*, *61*(1), 228–238.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (1950), 4171–4186.

de Vos, B. D., Berendsen, F. F., Viergever, M. A., Staring, M., & Išgum, I. (2017). End-to-End Unsupervised Deformable Image Registration with a Convolutional Neural Network. In M. J. Cardoso, T. Arbel, G. Carneiro, T. Syeda-Mahmood, J. M. R. Tavares, M. Moradi, A. Bradley, H. Greenspan, J. P. Papa, A. Madabhushi, J. C. Nascimento, J. S. Cardoso, V. Belagiannis, & Z. Lu (Eds.), *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (pp. 204–212). Springer International Publishing.

Dijk, L. V. v., Steenbakkers, R. J. H. M., Haken, B. t., Laan, H. P. v. d., Veld, A. A. v. t., Langendijk, J. A., & Korevaar, E. W. (2016). Robust Intensity Modulated Proton Therapy (IMPT) Increases Estimated Clinical Benefit in Head and Neck Cancer Patients. *PLOS ONE*, *11*(3), e0152477.

Dong, P., & Xing, L. (2020). Deep DoseNet: A deep neural network for accurate dosimetric transformation between different spatial resolutions and/or different dose calculation algorithms for precision radiation therapy. *Physics in Medicine & Biology*, *65*(3), 35010–35010.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.

Eley, J. G., Newhauser, W. D., Lüchtenborg, R., Graeff, C., & Bert, C. (2014). 4D optimization of scanned ion beam tracking therapy for moving tumors. *Physics in Medicine and Biology*, *59*(13), 3431–3452.

Elmahdy, M. S., Beljaards, L., Yousefi, S., Sokooti, H., Verbeek, F., Van Der Heide, U. A., & Staring, M. (2021). Joint Registration and Segmentation via Multi-Task Learning for Adaptive Radiotherapy of Prostate Cancer. *IEEE Access*, *9*, 95551–95568.

Elmahdy, M. S., Jagt, T., Zinkstok, R. T., Qiao, Y., Shahzad, R., Sokooti, H., Yousefi, S., Incrocci, L., Marijnen, C., Hoogeman, M., & Staring, M. (2019). Robust contour propagation using deep learning and image registration for online adaptive proton therapy of prostate cancer. *Medical Physics*, *46*(8), 3329–3343.

Emert, F., Missimer, J., Eichenberger, P. A., Walser, M., Gmür, C., Lomax, A. J., Weber, D. C., & Spengler, C. M. (2021). Enhanced Deep-Inspiration Breath Hold Superior to High-Frequency Percussive Ventilation for Respiratory Motion Mitigation: A Physiology-Driven, MRI-Guided Assessment Toward Optimized Lung Cancer Treatment With Proton Therapy. *Frontiers in Oncology*, *11*.

Engelsman, M., Rietzel, E., & Kooy, H. M. (2006). Four-dimensional proton treatment planning for lung tumors. *International Journal of Radiation Oncology Biology Physics*, *64*(5), 1589–1595.

Engwall, E., Glimelius, L., & Hynning, E. (2018). Effectiveness of different rescanning techniques for scanned proton radiotherapy in lung cancer patients. *Physics in Medicine and Biology*, *63*(9), 095006.

Engwall, E., Fredriksson, A., & Glimelius, L. (2018). 4D robust optimization including uncertainties in time structures can reduce the interplay effect in proton pencil beam scanning radiation therapy. *Medical Physics*, *45*(9), 4020–4029.

Ernst, F. (2011). *Compensating for Quasi-periodic Motion in Robotic Radiosurgery*. Springer Science & Business Media.

Evans, P. M., Coolens, C., & Nioutsikou, E. (2005). Effects of averaging over motion and the resulting systematic errors in radiation therapy. *Physics in Medicine and Biology*, *51*(1), N1–N7.

Faddegon, B., Ramos-Méndez, J., Schuemann, J., McNamara, A., Shin, J., Perl, J., & Paganetti, H. (2020). The TOPAS tool for particle simulation, a Monte Carlo simulation tool for physics, biology and clinical research. *Physica Medica*, *72*, 114–121.

Fan, J., Wang, J., Chen, Z., Hu, C., Zhang, Z., & Hu, W. (2019). Automatic treatment planning based on three-dimensional dose distribution predicted from deep learning technique. *Medical Physics*, *46*(1), 370–381.

Fan, J., Xing, L., Dong, P., Wang, J., Hu, W., & Yang, Y. (2020). Data-driven dose calculation algorithm based on deep U-Net. *Physics in Medicine and Biology*, *65*(24), 245035–245035.

Fogliata, A., Nicolini, G., Clivio, A., Vanetti, E., & Cozzi, L. (2012). Critical Appraisal of Acuros XB and Anisotropic Analytic Algorithm Dose Calculation in Advanced Non-Small-Cell Lung Cancer Treatments. *International Journal of Radiation Oncology Biology Physics*, *83*(5), 1587–1595.

Fracchiolla, F., Engwall, E., Janson, M., Tamm, F., Lorentini, S., Fellin, F., Bertolini, M., Algranati, C., Righetto, R., Farace, P., Amichetti, M., & Schwarz, M. (2021). Clinical validation of a GPU-based Monte Carlo dose engine of a commercial treatment planning system for pencil beam scanning proton therapy. *Physica Medica*, *88*, 226–234.

Fujita, H., & Cimr, D. (2019). Computer Aided detection for fibrillations and flutters using deep convolutional neural network. *Information Sciences*, *486*, 231–239.

Gajewski, J., Garbacz, M., Chang, C. W., Czerska, K., Durante, M., Krah, N., Krzempek, K., Kopeć, R., Lin, L., Mojżeszek, N., Patera, V., Pawlik-Niedzwiecka, M., Rinaldi, I., Rydygier, M., Pluta, E., Scifoni, E., Skrzypek, A., Tommasino, F., Schiavi, A., & Rucinski, A. (2021). Commissioning of GPU–Accelerated Monte Carlo Code FRED for Clinical Applications in Proton Therapy. *Frontiers in Physics*, *8*.

George, R., Vedam, S. S., Chung, T. D., Ramakrishnan, V., & Keall, P. J. (2005). The application of the sinusoidal model to lung cancer patient respiratory motion. *Medical Physics*, *32*(9), 2850–2861.

Golany, T., & Radinsky, K. (2019). PGANs: Personalized Generative Adversarial Networks for ECG Synthesis to Improve Patient-Specific Deep ECG Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 557–564.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. *27*.

Grassberger, C., Daartz, J., Dowdell, S., Ruggieri, T., Sharp, G., & Paganetti, H. (2014). Quantification of proton dose calculation accuracy in the lung. *International Journal of Radiation Oncology Biology Physics*, *89*(2), 424–430.

Heath, E., Unkelbach, J., & Oelfke, U. (2009). Incorporating uncertainties in respiratory motion into 4D treatment plan optimization. *Medical Physics*, *36*(7), 3059–3071.

Hendrycks, D., & Gimpel, K. (2016). Gaussian Error Linear Units (GELUs), 1–9.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., & Lerchner, A. (2022). Beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework.

Hissoiny, S., Raaijmakers, A. J. E., Ozell, B., Després, P., & Raaymakers, B. W. (2011). Fast dose calculation in magnetic fields with GPUMCD. *Physics in Medicine & Biology*, *56*(16), 5119–5129.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780.

Hong, L., Goitein, M., Bucciolini, M., Comiskey, R., Gottschalk, B., Rosenthal, S., Serago, C., & Urie, M. (1996). A pencil beam algorithm for proton dose calculations. *Physics in Medicine and Biology*, *41*(8), 1305–1330.

Hong, S., Zhou, Y., Shang, J., Xiao, C., & Sun, J. (2020). Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review. *Computers in Biology and Medicine*, *122*, 103801.

Hoogeman, M., Prévost, J.-B., Nuyttens, J., Pöll, J., Levendag, P., & Heijmen, B. (2009). Clinical Accuracy of the Respiratory Tumor Tracking System of the CyberKnife: Assessment by Analysis of Log Files. *International Journal of Radiation Oncology Biology Physics*, *74*(1), 297–303.

Hu, S., Worrall, D., Knegt, S., Veeling, B., Huisman, H., & Welling, M. (2019). Supervised Uncertainty Quantification for Segmentation with Multiple Annotations. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11765 LNCS*, 137–145.

Hussein, M., Heijmen, B. J. M., Verellen, D., & Nisbet, A. (2018). Automation in intensity modulated radiotherapy treatment planning—a review of recent innovations. *The British Journal of Radiology*, *91*(1092), 20180270–20180270.

Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *Proceedings of the 32nd International Conference on Machine Learning*, 448–456.

Jaderberg, M., Simonyan, K., Zisserman, A., & kavukcuoglu koray, k. (2015). Spatial Transformer Networks. *Advances in Neural Information Processing Systems*, *28*.

Jagt, T., Breedveld, S., van de Water, S., Heijmen, B., & Hoogeman, M. (2017). Near real-time automated dose restoration in IMPT to compensate for daily tissue density variations in prostate cancer. *Physics in Medicine and Biology*, *62*(11), 4254–4272.

Jagt, T., Breedveld, S., van Haveren, R., Heijmen, B., & Hoogeman, M. (2018). An automated planning strategy for near real-time adaptive proton therapy in prostate cancer. *Physics in Medicine and Biology*, *63*(13), 135017.

Jahnke, L., Fleckenstein, J., Wenz, F., & Hesser, J. (2012). GMC: A GPU implementation of a Monte Carlo dose calculation based on Geant4. *Physics in Medicine and Biology*, *57*(5), 1217–1229.

Jakobi, A., Perrin, R., Knopf, A., & Richter, C. (2018). Feasibility of proton pencil beam scanning treatment of free-breathing lung cancer patients. *Acta Oncologica*, *57*, 203–210.

Jang, E., Gu, S., & Poole, B. (2017). Categorical reparameterization with gumbel-softmax. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 1–13.

Javaid, U., Souris, K., Huang, S., & Lee, J. A. (2021). Denoising proton therapy Monte Carlo dose distributions in multiple tumor sites: A comparative neural networks architecture study. *Physica Medica*, *89*, 93–103.

Jeong, Y., Radke, R. J., & Lovelock, D. M. (2010). Bilinear models for inter-and intra-patient variation of the prostate. *Physics in Medicine & Biology*, *55*(13), 3725–3739.

Jia, X., Gu, X., Graves, Y. J., Folkerts, M., & Jiang, S. B. (2011). GPU-based fast Monte Carlo simulation for radiotherapy dose calculation. *Physics in Medicine & Biology*, *56*(22), 7017–7031.

Jiang, S. B., Pope, C., Jarrah, K. M. A., Kung, J. H., Bortfeld, T., & Chen, G. T. Y. (2003). An experimental investigation on intra-fractional organ motion effects in lung IMRT treatments. *Physics in Medicine and Biology*, *48*(12), 1773–1784.

Kajikawa, T., Kadoya, N., Ito, K., Takayama, Y., Chiba, T., Tomori, S., Nemoto, H., Dobashi, S., Takeda, K., & Jingu, K. (2019). A convolutional neural network approach for IMRT dose distribution prediction in prostate cancer patients. *Journal of Radiation Research*, *60*(5), 685–693.

Kardar, L., Li, Y., Li, X., Li, H., Cao, W., Chang, J. Y., Liao, L., Zhu, R. X., Sahoo, N., Gillin, M., Liao, Z., Komaki, R., Cox, J. D., Lim, G., & Zhang, X. (2014). Evaluation and mitigation of the interplay effects of intensity modulated proton therapy for lung cancer in a clinical setting. *Practical Radiation Oncology*, *4*(6), e259–e268.

Kearney, V., Chan, J. W., Haaf, S., Descovich, M., & Solberg, T. D. (2018). DoseNet: A volumetric dose prediction algorithm using 3D fully-convolutional neural networks. *Physics in Medicine and Biology*, *63*(23), 235022–235022.

Kingma, D. P., & Ba, J. (2017). Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*.

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, (1050), 1–14.

Kingma, D. P., & Welling, M. (2019). An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, *12*(4), 307–392.

Kissick, M. W., Boswell, S. A., Jeraj, R., & Mackie, T. R. (2005). Confirmation, refinement, and extension of a study in intrafraction motion interplay with sliding jaw motion. *Medical Physics*, *32*(7Part1), 2346–2350.

Kohl, S. A. A., Romera-Paredes, B., Maier-Hein, K. H., Rezende, D. J., Eslami, S. M. A., Kohli, P., Zisserman, A., & Ronneberger, O. (2019). A Hierarchical Probabilistic U-Net for Modeling Multi-Scale Ambiguities.

Kohl, S. A., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J. R., Maier-Hein, K. H., Ali Eslami, S. M., Rezende, D. J., & Ronneberger, O. (2018). A probabilistic U-net for segmentation of ambiguous images. *Advances in Neural Information Processing Systems*, *2018-Decem*(NeurIPS), 6965–6975.

Kontaxis, C., Bol, G. H., Lagendijk, J. J. W., & Raaymakers, B. W. (2020). DeepDose: Towards a fast dose calculation engine for radiation therapy using deep learning. *Physics in Medicine and Biology*, *65*(7), 75013–75013.

Kraus, K. M., Heath, E., & Oelfke, U. (2011). Dosimetric consequences of tumour motion due to respiration for a scanned proton beam. *Physics in Medicine and Biology*, *56*(20), 6563–6581.

Krebs, J., Delingette, H., Mailhe, B., Ayache, N., & Mansi, T. (2019). Learning a Probabilistic Model for Diffeomorphic Registration. *IEEE transactions on medical imaging*, *38*(9), 2165–2176.

Krieger, M., Giger, A., Salomir, R., Bieri, O., Celicanin, Z., Cattin, P. C., Lomax, A. J., Weber, D. C., & Zhang, Y. (2020). Impact of internal target volume definition for pencil beam scanned proton treatment planning in the presence of respiratory motion variability for lung cancer: A proof of concept. *Radiotherapy and Oncology*, *145*, 154–161.

Lambert, J., Suchowerska, N., McKenzie, D. R., & Jackson, M. (2005). Intrafractional motion during proton beam scanning. *Physics in Medicine and Biology*, *50*(20), 4853–4862.

Lee, H., Kim, H., Kwak, J., Kim, Y. S., Lee, S. W., Cho, S., & Cho, B. (2019). Fluence-map generation for prostate intensity-modulated radiotherapy planning using a deep neural-network. *Scientific Reports*, *9*(1), 15671–15671.

Lei, Y., Fu, Y., Wang, T., Liu, Y., Patel, P., Curran, W. J., Liu, T., & Yang, X. (2020). 4D-CT deformable image registration using multiscale unsupervised deep learning. *Physics in Medicine & Biology*, *65*(8).

Li, Y., Kardar, L., Li, X., Li, H., Cao, W., Chang, J. Y., Liao, L., Zhu, R. X., Sahoo, N., Gillin, M., Liao, Z., Komaki, R., Cox, J. D., Lim, G., & Zhang, X. (2014). On the interplay effects with proton scanning beams in stage III lung cancer. *Medical Physics*, *41*(2), 021721.

Liang, X., Bibault, J. E., Leroy, T., Escande, A., Zhao, W., Chen, Y., Buyyounouski, M. K., Hancock, S. L., Bagshaw, H., & Xing, L. (2021). Automated contour propagation of the prostate from pCT to CBCT images via deep unsupervised learning. *Medical Physics*, *48*(4), 1764–1770.

Liebl, J., Paganetti, H., Zhu, M., & Winey, B. A. (2014). The influence of patient positioning uncertainties in proton radiotherapy on proton range and dose distributions. *Medical Physics*, *41*(9), 091711.

Liu, W., Frank, S. J., Li, X., Li, Y., Zhu, R. X., & Mohan, R. (2013). PTV-based IMPT optimization incorporating planning risk volumes vs robust optimization. *Medical Physics*, *40*(2), 021709.

Liu, W., Zhang, X., Li, Y., & Mohan, R. (2012). Robust optimization of intensity modulated proton therapy. *Medical Physics*, *39*(2), 1079–1091.

Lomax, A. J. (2008a). Intensity modulated proton therapy and its sensitivity to treatment uncertainties 1: The potential effects of calculational uncertainties. *Physics in Medicine and Biology*, *53*(4), 1027.

Lomax, A. J. (2008b). Intensity modulated proton therapy and its sensitivity to treatment uncertainties 2: The potential effects of inter-fraction and inter-field motions. *Physics in Medicine & Biology*, *53*(4), 1043.

Low, D. A., Harms, W. B., Mutic, S., & Purdy, J. A. (1998). A technique for the quantitative evaluation of dose distributions. *Medical Physics*, *25*(5), 656–661.

Lujan, A. E., Balter, J. M., & Ten Haken, R. K. (2003). A method for incorporating organ motion due to breathing into 3D dose calculations in the liver: Sensitivity to variations in motion. *Medical Physics*, *30*(10), 2643–2649.

Lundkvist, J., Ekman, M., Ericsson, S. R., Jönsson, B., & Glimelius, B. (2005). Proton therapy of cancer: Potential clinical advantages and cost-effectiveness. *Acta Oncologica*, *44*(8), 850–861.

Ma, J., Beltran, C., Seum Wan Chan Tseung, H., & Herman, M. G. (2014). A GPU accelerated and Monte Carlo-based intensity modulated proton therapy optimization system. *Medical Physics*, *41*(12).

Ma, M., K. Buyyounouski, M., Vasudevan, V., Xing, L., & Yang, Y. (2019). Dose distribution prediction in isodose feature-preserving voxelization domain using deep convolutional neural network. *Medical Physics*, *46*(7), 2978–2987.

Maaten, L. v. d., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605.

Maddison, C. J., Mnih, A., & Teh, Y. W. (2017). The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. *arXiv:1611.00712 [cs.LG]*.

Magallon-Baro, A., Loi, M., Milder, M. T., Granton, P. V., Zolnay, A. G., Nuyttens, J. J., & Hoogeman, M. S. (2019). Modeling daily changes in organ-at-risk anatomy in a cohort of pancreatic cancer patients. *Radiotherapy and Oncology*, *134*, 127–134.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., & Frey, B. (2016). Adversarial Autoencoders. *arXiv:1511.05644 [cs.LG]*.

Matter, M., Nenoff, L., Meier, G., Weber, D. C., Lomax, A. J., & Albertini, F. (2019). Intensity modulated proton therapy plan generation in under ten seconds. *Acta Oncologica*, *58*(10), 1435–1439.

McSharry, P., Clifford, G., Tarassenko, L., & Smith, L. (2003). A dynamical model for generating synthetic electrocardiogram signals. *IEEE Transactions on Biomedical Engineering*, *50*(3), 289–294.

Mescheder, L., Nowozin, S., & Geiger, A. (2017). Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. *34th International Conference on Machine Learning, ICML 2017*, *5*, 3694–3707.

Meyer, P., Noblet, V., Mazzara, C., & Lallement, A. (2018). Survey on deep learning for radiotherapy. *Computers in Biology and Medicine*, *98*(May), 126–146.

Mohan, R., Chui, C., & Lidofsky, L. (1986). Differential pencil beam dose computation model for photons. *Medical Physics*, *13*(1), 64–73.

Moler, C., & Van Loan, C. (2003). Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-Five Years Later. *SIAM Review*, *45*(1), 3–49.

Neal, M. L., & Kerckhoffs, R. (2010). Current progress in patient-specific modeling. *Briefings in Bioinformatics*, *11*(1), 111–126.

Neishabouri, A., Wahl, N., Mairani, A., Köthe, U., & Bangert, M. (2021). Long short-term memory networks for proton dose calculation in highly heterogeneous tissues. *Medical Physics*, *48*(4), 1893–1908.

Neph, R., Lyu, Q., Huang, Y., Yang, Y. M., & Sheng, K. (2021). DeepMC: A deep learning method for efficient Monte Carlo beamlet dose calculation by predictive denoising in magnetic resonance-guided radiotherapy. *Physics in Medicine & Biology*, *66*(3), 35022–35022.

Nguyen, D., Jia, X., Sher, D., Lin, M.-H., Iqbal, Z., Liu, H., & Jiang, S. (2019). 3D radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected U-net deep learning architecture. *Physics in Medicine & Biology*, *64*(6), 65020.

Nguyen, D., Long, T., Jia, X., Lu, W., Gu, X., Iqbal, Z., & Jiang, S. (2019). A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning. *Scientific Reports*, *9*(1), 1076.

Nie, X., Liang, J., & Yan, D. (2012). Organ sample generator for expected treatment dose construction and adaptive inverse planning optimization. *Medical Physics*, *39*(12), 7329–7337.

Nomura, Y., Wang, J., Shirato, H., Shimizu, S., & Xing, L. (2020). Fast spot-scanning proton dose calculation method with uncertainty quantification using a three dimensional convolutional neural network. *Physics in Medicine and Biology*, *65*(21), 215007.

Ohara, K., Okumura, T., Akisada, M., Inada, T., Mori, T., Yokota, H., & Calaguas, M. J. B. (1989). Irradiation synchronized with respiration gate. *International Journal of Radiation Oncology Biology Physics*, *17*(4), 853–857.

Paganetti, H. (2012). Range uncertainties in proton therapy and the role of Monte Carlo simulations [Publisher: IOP Publishing]. *Physics in Medicine and Biology*, *57*(11), R99.

Paganetti, H., Botas, P., Sharp, G. C., & Winey, B. (2021). Adaptive proton therapy. *Physics in Medicine and Biology*, *66*(22), 22TR01.

Pastor-Serrano, O., Habraken, S., Lathouwers, D., Hoogeman, M., Schaart, D., & Perkó, Z. (2021). How should we model and evaluate breathing interplay effects in IMPT? *Physics in Medicine and Biology*, *66*(23), 235003–235003.

Pastor-Serrano, O., Lathouwers, D., & Perkó, Z. (2021). A semi-supervised autoencoder framework for joint generation and classification of breathing. *Computer Methods and Programs in Biomedicine*, *209*, 106312–106312.

Pastor-Serrano, O., & Perkó, Z. (2022a). Millisecond speed deep learning based proton dose calculation with Monte Carlo accuracy. *Physics in Medicine and Biology*, *67*(10), 105006.

Pastor-Serrano, O., & Perkó, Z. (2022b). Learning the Physics of Particle Transport via Transformers. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(11), 12071–12079.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv: 1912.01703 [cs, stat]*.

Péguret, N., Ozsahin, M., Zeverino, M., Belmondo, B., Durham, A.-D., Lovis, A., Simons, J., Long, O., Duclos, F., Prior, J., Denys, A., Beigelman, C., Sozzi, W. J., Grant, K., Gautier-Dechaud, V., Peters, S., Vienne, M., Moeckli, R., & Bourhis, J. (2016). Apnea-like suppression of respiratory motion: First evaluation in radiotherapy. *Radiotherapy and Oncology*, *118*(2), 220–226.

Peng, Z., Shan, H., Liu, T., Pei, X., Wang, G., & Xu, X. G. (2019). MCDNet – A Denoising Convolutional Neural Network to Accelerate Monte Carlo Radiation Transport Simulations: A Proof of Principle With Patient Dose From X-Ray CT Imaging. *IEEE Access*, *7*, 76680–76689.

Peng, Z., Shan, H., Liu, T., Pei, X., Zhou, J., Wang, G., & Xu, X. G. (2019). Deep learning for accelerating Monte Carlo radiation transport simulation in intensity-modulated radiation therapy, 1–8.

Pepin, M. D., Tryggestad, E., Wan Chan Tseung, H. S., Johnson, J. E., Herman, M. G., & Beltran, C. (2018). A Monte-Carlo-based and GPU-accelerated 4D-dose calculator for a pencil beam scanning proton therapy system. *Medical Physics*, *45*(11), 5293–5304.

Pereira, G. C., Traughber, M., & Muzic, R. F. (2014). The Role of Imaging in Radiation Therapy Planning: Past, Present, and Future. *BioMed Research International*, *2014*(2), 1–9.

Perkó, Z., Van Der Voort, S. R., Van De Water, S., Hartman, C. M., Hoogeman, M., & Lathouwers, D. (2016). Fast and accurate sensitivity analysis of IMPT treatment plans using Polynomial Chaos Expansion. *Physics in Medicine and Biology*, *61*(12), 4646–4664.

Perl, J., Shin, J., Schümann, J., Faddegon, B., & Paganetti, H. (2012). TOPAS: An innovative proton Monte Carlo platform for research and clinical applications. *Medical Physics*, *39*(11), 6818–6837.

Pflugfelder, D., Wilkens, J. J., & Oelfke, U. (2008). Worst case optimization: A method to account for uncertainties in the optimization of intensity modulated proton therapy. *Physics in Medicine and Biology*, *53*(6), 1689–1700.

Phillips, M. H., Pedroni, E., Blattmann, H., Boehringer, T., Coray, A., & Scheib, S. (1992). Effects of respiratory motion on dose uniformity with a charged particle scanning method. *Physics in Medicine and Biology*, *37*(1), 223–233.

Qin, N., Botas, P., Giantsoudi, D., Schuemann, J., Tian, Z., Jiang, S. B., Paganetti, H., & Jia, X. (2016). Recent developments and comprehensive evaluations of a GPU-based Monte Carlo package for proton therapy. *Physics in Medicine and Biology*, *61*(20), 7347–7362.

Ramachandran, P., Bello, I., Parmar, N., Levskaya, A., Vaswani, A., & Shlens, J. (2019). Stand-alone self-attention in vision models. *Advances in Neural Information Processing Systems*, *32*.

Ravuri, S., & Vinyals, O. (2019). Classification Accuracy Score for Conditional Generative Models. *Advances in Neural Information Processing Systems, 32*.

Razavi, A., van den Oord, A., & Vinyals, O. (2019). Generating Diverse High-Fidelity Images with VQ-VAE-2. *Advances in Neural Information Processing Systems, 32*.

Rezende, D. J., Mohamed, S., & Wierstra, D. (2014). Stochastic Backpropagation and Approximate Inference in Deep Generative Models. *arXiv:1401.4082 [cs, stat]*.

Rezende, D. J., & Viola, F. (2018). Taming VAEs. *arXiv:1810.00597 [stat.ML]*.

Rios, R., De Crevoisier, R., Ospina, J. D., Commandeur, F., Lafond, C., Simon, A., Haigron, P., Espinosa, J., & Acosta, O. (2017). Population model of bladder motion and deformation based on dominant eigenmodes and mixed-effects models in prostate cancer radiotherapy. *Medical Image Analysis, 38*, 133–149.

Rojo-Santiago, J., Habraken, S. J. M., Lathouwers, D., Romero, A. M., Perkó, Z., & Hoogeman, M. S. (2021). Accurate assessment of a Dutch practical robustness evaluation protocol in clinical PT with pencil beam scanning for neurological tumors. *Radiotherapy and Oncology, 163*, 121–127.

Romaguera, L. V., Mezheritsky, T., Mansour, R., Carrier, J.-F., & Kadoury, S. (2021). Probabilistic 4D predictive model from in-room surrogates using conditional generative networks for image-guided radiotherapy. *Medical Image Analysis, 74*, 102250–102250.

Romaguera, L. V., Plantefève, R., Romero, F. P., Hébert, F., Carrier, J. F., & Kadoury, S. (2020). Prediction of in-plane organ deformation during free-breathing radiotherapy via discriminative spatial transformer networks. *Medical Image Analysis, 64*.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.). Springer International Publishing.

Rosca, M., Lakshminarayanan, B., & Mohamed, S. (2019). Distribution Matching in Variational Inference. *arXiv:1802.06847 [stat.ML]*.

Saini, J., Maes, D., Egan, A., Bowen, S. R., St James, S., Janson, M., Wong, T., & Bloch, C. (2017). Dosimetric evaluation of a commercial proton spot scanning Monte-Carlo dose algorithm: Comparisons against measurements and simulations. *Physics in Medicine and Biology, 62*(19), 7659–7681.

Schaffner, B., Pedroni, E., & Lomax, A. (1999). Dose calculation models for proton treatment planning using a dynamic beam delivery system: An attempt to include density heterogeneity effects in the analytical dose calculation. *Physics in Medicine and Biology, 44*(1), 27–41.

Schreuder, A. N., Bridges, D. S., Rigsby, L., Blakey, M., Janson, M., Hedrick, S. G., & Wilkinson, J. B. (2019a). Validation of the RayStation Monte Carlo dose calculation algorithm using a realistic lung phantom. *Journal of Applied Clinical Medical Physics, 20*(12), 127–137.

Schreuder, A. N., Bridges, D. S., Rigsby, L., Blakey, M., Janson, M., Hedrick, S. G., & Wilkinson, J. B. (2019b). Validation of the RayStation Monte Carlo dose calculation algorithm using realistic animal tissue phantoms. *Journal of Applied Clinical Medical Physics, 20*(10), 160–171.

Schuemann, J., Giantsoudi, D., Grassberger, C., Moteabbed, M., Min, C. H., & Paganetti, H. (2015). Assessing the Clinical Impact of Approximations in Analytical Dose Calculations for Proton Therapy. *International Journal of Radiation Oncology Biology Physics*, *92*(5), 1157–1164.

Seco, J., Sharp, G. C., Turcotte, J., Gierga, D., Bortfeld, T., & Paganetti, H. (2007). Effects of organ motion on IMRT treatments with segments of few monitor units. *Medical Physics*, *34*(3), 923–934.

Seco, J., Robertson, D., Trofimov, A., & Paganetti, H. (2009). Breathing interplay effects during proton beam scanning: Simulation and statistical analysis. *Physics in Medicine and Biology*, *54*(14), N283–N294.

Sharma, M., Weiss, E., & Siebers, J. V. (2012). Dose deformation-invariance in adaptive prostate radiation therapy: Implication for treatment simulations. *Radiotherapy and Oncology*, *105*(2), 207–213.

Shih, H. A., Jiang, S. B., Aljarrah, K. M., Doppke, K. P., & Choi, N. C. (2004). Internal target volume determined with expansion margins beyond composite gross tumor volume in three-dimensional conformal radiotherapy for lung cancer. *International Journal of Radiation Oncology Biology Physics*, *60*(2), 613–622.

Sievinen, J., Ulmer, W., & Kaissl, W. (2013). AAA Photon Dose Calculation Model in Eclipse.

Silva, J. d., Ansorge, R., & Jena, R. (2015). Sub-second pencil beam dose calculation on GPU for adaptive proton therapy. *Physics in Medicine and Biology*, *60*(12), 4777–4795.

Söhn, M., Birkner, M., Yan, D., & Alber, M. (2005). Modelling individual geometric variation based on dominant eigenmodes of organ deformation: Implementation and evaluation. *Physics in Medicine and Biology*, *50*(24), 5893–5908.

Söhn, M., Sobotta, B., & Alber, M. (2012). Dosimetric treatment course simulation based on a statistical model of deformable organ motion [Publisher: IOP Publishing]. *Physics in Medicine and Biology*, *57*(12), 3693–3709.

Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K., & Winther, O. (2016). Ladder variational autoencoders. *Advances in Neural Information Processing Systems*, 3745–3753.

Souris, K., Lee, J. A., & Sterpin, E. (2016). Fast multipurpose Monte Carlo simulation for proton therapy using multi- and many-core CPU architectures. *Medical Physics*, *43*(4), 1700–1712.

Srivastava, N., Hinton, G., Krizhevsky, A., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, *15*, 1929–1958.

Stroom, J. C., & Heijmen, B. J. M. (2002). Geometrical uncertainties, radiotherapy planning margins, and the ICRU-62 report. *Radiotherapy and Oncology*, *64*(1), 75–83.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, *71*(3), 209–249.

Szeto, Y. Z., Witte, M. G., van Herk, M., & Sonke, J. J. (2017). A population based statistical model for daily geometric variations in the thorax. *Radiotherapy and Oncology*, *123*(1), 99–105.

Takao, S., Miyamoto, N., Matsuura, T., Onimaru, R., Katoh, N., Inoue, T., Sutherland, K. L., Suzuki, R., Shirato, H., & Shimizu, S. (2016). Intrafractional Baseline Shift or Drift of Lung Tumor Motion During Gated Radiation Therapy With a Real-Time Tumor-Tracking System. *International Journal of Radiation Oncology Biology Physics*, *94*(1), 172–180.

Teoh, S., Fiorini, F., George, B., Vallis, K. A., & Van den Heuvel, F. (2020). Is an analytical dose engine sufficient for intensity modulated proton therapy in lung cancer? *British Journal of Radiology*, *93*(1107).

Thörnqvist, S., Hysing, L. B., Zolnay, A. G., Söhn, M., Hoogeman, M. S., Muren, L. P., Bentzen, L., & Heijmen, B. J. M. (2013). Treatment simulations with a statistical deformable motion model to evaluate margins for multiple targets in radiotherapy for high-risk prostate cancer. *Radiotherapy and Oncology*, *109*(3), 344–349.

Thörnqvist, S., Hysing, L. B., Zolnay, A. G., Söhn, M., Hoogeman, M. S., Muren, L. P., & Heijmen, B. J. M. (2013). Adaptive radiotherapy in locally advanced prostate cancer using a statistical deformable motion model. *Acta Oncologica*, *52*(7), 1423–1429.

Tilly, D., Van De Schoot, A. J., Grusell, E., Bel, A., & Ahnesjö, A. (2017). Dose coverage calculation using a statistical shape model - Applied to cervical cancer radiotherapy. *Physics in Medicine and Biology*, *62*(10), 4140–4159.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2020). Training data-efficient image transformers & distillation through attention, 1–22.

Trofimov, A., Nguyen, P. L., Efstathiou, J. A., Wang, Y., Lu, H.-M., Engelsman, M., Merrick, S., Cheng, C.-W., Wong, J. R., & Zietman, A. L. (2011). Interfractional Variations in the Setup of Pelvic Bony Anatomy and Soft Tissue, and Their Implications on the Delivery of Proton Therapy for Localized Prostate Cancer. *International Journal of Radiation Oncology Biology Physics*, *80*(3), 928–937.

Tsekas, G., Bol, G. H., Raaymakers, B. W., & Kontaxis, C. (2021). DeepDose: A robust deep learning-based dose engine for abdominal tumours in a 1.5 T MRI radiotherapy system. *Physics in Medicine and Biology*, *66*(6), 65017–65017.

Tsekas, G., Bol, G. H., & Raaymakers, B. W. (2022). Robust deep learning-based forward dose calculations for VMAT on the 1.5T MR-Linac. *Physics in Medicine and Biology*.

Ulmer, W., Pyyry, J., & Kaissl, W. (2005). A 3D photon superposition/convolution algorithm and its foundation on results of Monte Carlo calculations. *Physics in Medicine and Biology*, *50*(8), 1767–1790.

Unkelbach, J., Alber, M., Bangert, M., Bokrantz, R., Chan, T. C. Y., Deasy, J. O., Fredriksson, A., Gorissen, B. L., Herk, M. v., Liu, W., Mahmoudzadeh, H., Nohadani, O., Siebers, J. V., Witte, M., & Xu, H. (2018). Robust radiotherapy planning. *Physics in Medicine and Biology*, *63*(22), 22TR02.

Unkelbach, J., Bortfeld, T., Martin, B. C., & Soukup, M. (2009). Reducing the sensitivity of IMPT treatment plans to setup errors and range uncertainties via probabilistic treatment planning. *Medical Physics*, *36*(1), 149–163.

Unkelbach, J., & Paganetti, H. (2018). Robust Proton Treatment Planning: Physical and Biological Optimization. *Seminars in radiation oncology*, *28*(2), 88–96.

van der Voort, S., van de Water, S., Perkó, Z., Heijmen, B., Lathouwers, D., & Hoogeman, M. (2016). Robustness Recipes for Minimax Robust Optimization in Intensity Modulated Proton Therapy for Oropharyngeal Cancer Patients. *International Journal of Radiation Oncology Biology Physics*, *95*(1), 163–170.

van Herk, M., Remeijer, P., & Lebesque, J. V. (2002). Inclusion of geometric uncertainties in treatment plan evaluation. *International Journal of Radiation Oncology Biology Physics*, *52*(5), 1407–1422.

van Herk, M., Remeijer, P., Rasch, C., & Lebesque, J. V. (2000). The probability of correct target dosage: Dose-population histograms for deriving treatment margins in radiotherapy. *International Journal of Radiation Oncology Biology Physics*, *47*(4), 1121–1135.

van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworths.

Vásquez Osorio, E. M., Hoogeman, M. S., Bondar, L., Levendag, P. C., & Heijmen, B. J. M. (2009). A novel flexible framework with automatic feature correspondence optimization for nonrigid registration in radiotherapy. *Medical Physics*, *36*(7), 2848–2859.

Vasudevan, V., Shen, L., Huang, C., Chuang, C., Islam, M. T., Ren, H., Yang, Y., Dong, P., & Xing, L. (2022). Implicit neural representation for radiation therapy dose distribution. *Physics in Medicine and Biology*, *67*(12), 125014.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *2017-Decem*, 5999–6009.

Wan Chan Tseung, H., Ma, J., & Beltran, C. (2015). A fast GPU-based Monte Carlo simulation of proton transport with detailed modeling of nonelastic interactions. *Medical Physics*, *42*(6), 2967–2978.

Wang, W., Sheng, Y., Wang, C., Zhang, J., Li, X., Palta, M., Czito, B., Willett, C. G., Wu, Q., Ge, Y., Yin, F.-F., & Wu, Q. J. (2020). Fluence Map Prediction Using Deep Learning Models – Direct Plan Generation for Pancreas Stereotactic Body Radiation Therapy. *Frontiers in Artificial Intelligence*, *3*(September), 1–10.

Wang, Y., Mazur, T. R., Green, O., Hu, Y., Li, H., Rodriguez, V., Wooten, H. O., Yang, D., Zhao, T., Mutic, S., & Li, H. H. (2016). A GPU-accelerated Monte Carlo dose calculation platform and its application toward validating an MRI-guided radiation therapy beam model. *Medical Physics*, *43*(7), 4040–4052.

Water, S. v. d., Kraan, A. C., Breedveld, S., Schillemans, W., Teguh, D. N., Kooy, H. M., Madden, T. M., Heijmen, B. J. M., & Hoogeman, M. S. (2013). Improved efficiency of multi-criteria IMPT treatment planning using iterative resampling of randomly placed pencil beams. *Physics in Medicine and Biology*, *58*(19), 6969–6983.

Wieser, H. P., Cisternas, E., Wahl, N., Ulrich, S., Stadler, A., Mescher, H., Muller, L. R., Klinge, T., Gabrys, H., Burigo, L., Mairani, A., Ecker, S., Ackermann, B., Ellerbrock, M., Parodi, K., Jakel, O., & Bangert, M. (2017). Development of the open-

source dose calculation and optimization toolkit matRad. *Medical Physics*, *44*(6), 2556–2568.

Wu, C., Nguyen, D., Xing, Y., Montero, A. B., Schuemann, J., Shang, H., Pu, Y., & Jiang, S. (2021). Improving proton dose calculation accuracy by using deep learning. *Machine Learning: Science and Technology*, *2*(1), 15017–15017.

Wu, Y., & He, K. (2020). Group Normalization. *International Journal of Computer Vision*, *128*(3), 742–755.

Wulan, N., Wang, W., Sun, P., Wang, K., Xia, Y., & Zhang, H. (2020). Generating electrocardiogram signals by deep learning. *Neurocomputing*, *404*, 122–136.

Xiao, C., Choi, E., & Sun, J. (2018). Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review. *Journal of the American Medical Informatics Association*, *25*(10), 1419–1428.

Xiao, F., Cai, J., Zhou, X., Zhou, L., Song, T., & Li, Y. (2022). TransDose: A transformer-based UNet model for fast and accurate dose calculation for MR-LINACs. *Physics in Medicine and Biology*, *67*(12), 125013.

Xing, Y., Nguyen, D., Lu, W., Yang, M., & Jiang, S. (2020). Technical Note: A feasibility study on deep learning-based radiotherapy dose calculation. *Medical Physics*, *47*(2), 753–758.

Xing, Y., Zhang, Y., Nguyen, D., Lin, M.-H., Lu, W., & Jiang, S. (2020). Boosting radiotherapy dose calculation accuracy with deep learning. *Journal of Applied Clinical Medical Physics*, *21*(8), 149–159.

Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., & Liu, T. Y. (2020). On layer normalization in the transformer architecture. *37th International Conference on Machine Learning, ICML 2020*, 10455–10464.

Xu, H., Vile, D. J., Sharma, M., Gordon, J. J., & Siebers, J. V. (2014). Coverage-based treatment planning to accommodate deformable organ variations in prostate cancer treatment. *Medical Physics*, *41*(10).

Yan, C., Combine, A. G., Bednarz, G., Lalonde, R. J., Hu, B., Dickens, K., Wynn, R., Pavord, D. C., & Saiful Huq, M. (2017). Clinical implementation and evaluation of the Acuros dose calculation algorithm. *Journal of Applied Clinical Medical Physics*, *18*(5), 195–209.

Yildirim, O., Baloglu, U. B., Tan, R.-S., Ciaccio, E. J., & Acharya, U. R. (2019). A new approach for arrhythmia classification using deep coded features and LSTM networks. *Computer Methods and Programs in Biomedicine*, *176*, 121–133.

Yildirim, O., Tan, R. S., & Acharya, U. R. (2018). An efficient compression of ECG signals using deep convolutional autoencoders. *Cognitive Systems Research*, *52*, 198–211.

Yildirim, Ö., Pławiak, P., Tan, R.-S., & Acharya, U. R. (2018). Arrhythmia detection using deep convolutional neural network with long duration ECG signals. *Computers in Biology and Medicine*, *102*, 411–420.

You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., Song, X., Demmel, J., Keutzer, K., & Hsieh, C.-J. (2019). Large Batch Optimization for Deep Learning: Training BERT in 76 minutes.

Yu, J., Park, S. S., Herman, M. G., Langen, K., Mehta, M., & Feigenberg, S. J. (2017). Free Breathing versus Breath-Hold Scanning Beam Proton Therapy and Car-

diac Sparing in Breast Cancer. *International Journal of Particle Therapy*, *3*(3), 407–413.

Yuan, Y., Qin, W., Guo, X., Buyyounouski, M., Hancock, S., Han, B., & Xing, L. (2019). Prostate segmentation with encoder-decoder densely connected convolutional network (ed-densenet). *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 434–437.

Zhang, Q., Pevsner, A., Hertanto, A., Hu, Y.-C., Rosenzweig, K. E., Ling, C. C., & Mageras, G. S. (2007). A patient-specific respiratory model of anatomical motion for radiation treatment planning. *Medical Physics*, *34*(12), 4772–4781.

Zhang, Y., Huth, I., Weber, D. C., & Lomax, A. J. (2018). A statistical comparison of motion mitigation performances and robustness of various pencil beam scanned proton systems for liver tumour treatments. *Radiotherapy and Oncology*, *128*(1), 182–188.

Zhang, Y., Huth, I., Wegner, M., Weber, D. C., & Lomax, A. J. (2016). An evaluation of rescanning technique for liver tumour treatments using a commercial PBS proton therapy system. *Radiotherapy and Oncology*, *121*(2), 281–287.

Zhang, Y., Knopf, A., Tanner, C., & Lomax, A. J. (2014). Online image guided tumour tracking with scanned proton beams: A comprehensive simulation study. *Physics in Medicine and Biology*, *59*(24), 7793–7817.

Zhao, S., Song, J., & Ermon, S. (2018). InfoVAE: Information Maximizing Variational Autoencoders. *arXiv:1706.02262 [cs.LG]*.

Zhu, F., Ye, F., Fu, Y., Liu, Q., & Shen, B. (2019). Electrocardiogram generation with a bidirectional LSTM-CNN generative adversarial network. *Scientific Reports*, *9*, 6734.

Zhu, J., Liu, X., & Chen, L. (2020). A preliminary study of a photon dose calculation algorithm using a convolutional neural network. *Physics in Medicine and Biology*, *65*(20).

A

Lower bound derivation

Even though there are different ways to obtain the ELBO, the most common derivation is based on Jensen's inequality. For a concave function such as the natural logarithm the Jensen inequality states that

$$\log\left(\mathbb{E}[a]\right) \geq \mathbb{E}\left[\log(a)\right].$$

A.1. Lower bound of breathing models

Starting from the marginal likelihood of the probabilistic model, the expression of the lower bound can be obtained as

$$\log\left(P_{\boldsymbol{\theta}}(\boldsymbol{x})\right) = \log \int P_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{z})\,d\boldsymbol{z} \tag{A.1}$$

$$= \log \int P_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{z})\frac{Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}{Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\,d\boldsymbol{z} \tag{A.2}$$

$$= \log \mathbb{E}_{\boldsymbol{z}\sim Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\frac{P_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{z})}{Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\right] \tag{A.3}$$

$$\geq \mathbb{E}_{\boldsymbol{z}\sim Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log\left(\frac{P_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{z})}{Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\right)\right] \tag{A.4}$$

$$= \mathbb{E}_{\boldsymbol{z}\sim Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log\left(\frac{P_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\,P(\boldsymbol{z}))}{Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\right)\right] \tag{A.5}$$

$$= \mathbb{E}_{\boldsymbol{z}\sim Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log P_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right] - D_{KL}(Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||P(\boldsymbol{z})), \tag{A.6}$$

where the KL-divergence D_{KL} is defined as

$$D_{KL}(P(x)||Q(x)) = \int \log\left(\frac{P(x)}{Q(x)}\right)P(x)\,dx = \mathbb{E}_{\mathrm{x}\sim P(x)}\log\left(\frac{P(x)}{Q(x)}\right). \tag{A.7}$$

The output of the probabilistic decoder is the likelihood conditional distribution $P_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$. This distribution is represented as a multivariate Gaussian probability distribution with identity covariance matrix $P_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; f_{\boldsymbol{\theta}}(\boldsymbol{z}), \boldsymbol{I})$, where the function $f_{\boldsymbol{\theta}}(z) : \mathcal{Z} \rightarrow \mathbb{R}^M$ is parametrized with an ANN and represents the mean. The log-likelihood is formulated as

$$\log(P_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})) = \log\left(\frac{1}{\sqrt{(2\pi)^M|\boldsymbol{I}|}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-f_{\boldsymbol{\theta}}(\boldsymbol{z}))^T\boldsymbol{I}^{-1}(\boldsymbol{x}-f_{\boldsymbol{\theta}}(\boldsymbol{z}))\right)\right) \propto \frac{1}{2}\|\boldsymbol{x}-f_{\boldsymbol{\theta}}(\boldsymbol{z})\|_2^2, \tag{A.8}$$

This result has the same form as the squared error (SE), which is computed for the model output $\hat{\boldsymbol{x}}$ approximating the true output $\boldsymbol{x}$ as

$$\text{SE} = \|\boldsymbol{x}-\hat{\boldsymbol{x}}\|_2^2. \tag{A.9}$$

Thus, minimizing the log-likelihood with respect to the parameters $\boldsymbol{\theta}$ (which is done by approximating the expectation $\mathbb{E}_{\boldsymbol{z}\sim Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\log(P_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}))$ by taking Monte Carlo samples for $\boldsymbol{z} \sim Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$) yields the same result as minimizing the SE. On the other hand, when p and q are both Gaussian distributions, the KL-divergence can be computed in closed form. In our case the prior is $P(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{I})$ and the encoder distribution is $Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}(\boldsymbol{x}), \text{diag}\,\boldsymbol{\sigma}(\boldsymbol{x})^2)$. For an N-dimensional latent space, the KL-divergence can be analytically computed as:

$$D_{KL}(Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||P(\boldsymbol{z})) = \frac{1}{2}\left(-\sum_i^N(\log\sigma(\boldsymbol{x})_i^2+1)+\sum_i^N\sigma(\boldsymbol{x})_i^2+\sum_i^N\mu(\boldsymbol{x})_i^2\right). \tag{A.10}$$

Note that the contribution of the KL-divergence to the lower bound scales linearly with the latent dimensionality, so an increase in the lower bound caused by an increase of the latent space dimensionality could in theory be compensated by increasing the variance of the approximated posterior $Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ (lower KL-divergence per latent dimension).

A.2. Lower bound of organ models

Starting from the marginal likelihood of the probabilistic model in Equation 4.6, the lower bound is obtained as

$$\log\left(P_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{s}_x)\right) = \log\int P_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{x},\boldsymbol{s}_x)P(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{s}_x)d\boldsymbol{z} \tag{A.11}$$

$$= \log\mathbb{E}_{\boldsymbol{z}\sim Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{s}_x,\boldsymbol{y},\boldsymbol{s}_y)}\left[\frac{P_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{x},\boldsymbol{s}_x)P(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{s}_x)}{Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{s}_x,\boldsymbol{y},\boldsymbol{s}_y)}\right] \tag{A.12}$$

$$\geq \mathbb{E}_{\boldsymbol{z}\sim Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{s}_x,\boldsymbol{y},\boldsymbol{s}_y)}\left[\log\left(\frac{P_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{x},\boldsymbol{s}_x)P(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{s}_x)}{Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{s}_x,\boldsymbol{y},\boldsymbol{s}_y)}\right)\right] \tag{A.13}$$

$$= \mathbb{E}_{\boldsymbol{z}\sim Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{s}_x,\boldsymbol{y},\boldsymbol{s}_y)}[\log P_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{x},\boldsymbol{s}_x)] - D_{KL}(Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{s}_x,\boldsymbol{y},\boldsymbol{s}_y)||P(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{s}_x)). \tag{A.14}$$

B

Adversarial variational objective

AAEs do not optimize the exact variational lower bound, but an approximation. This section describes the approximated variational objective in AAEs. Let $\boldsymbol{x}$ be the data underlying data generating distribution $P_D(\boldsymbol{x})$ that we want to approximate, $\boldsymbol{z}$ be the corresponding latent variables with prior distribution $P(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \mathbf{0}, \boldsymbol{I})$, and η be random noise with distribution $p(\eta) = \mathcal{N}(\eta; 0, 1)$. In (Makhzani et al., 2016), the authors propose to regularize the latent space by introducing a discriminator model, modeled also with an ANN with mapping function $d_{\boldsymbol{\xi}}(\boldsymbol{z}) : \mathcal{Z} \rightarrow \mathbb{R}$ that outputs a single scalar logit. The discriminator is assumed to be capable of approximating any function. Given the encoder mapping $g_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, \eta) : \mathcal{X} \times \mathcal{H} \rightarrow \mathcal{Z}$, and the approximated posterior distribution $Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) = \int \delta(\boldsymbol{z} - g_{\boldsymbol{\phi}}(\boldsymbol{x}, \eta)) P(\eta) d\eta$, the adversarial regularization objective maximization can be formulated as

$$\max_{\boldsymbol{\xi}} \mathbb{E}_{\mathbf{z} \sim P(\boldsymbol{z})}[\log(S(d_{\boldsymbol{\xi}}(\boldsymbol{z})))] + \mathbb{E}_{\mathbf{x} \sim P_D(\boldsymbol{x})} \mathbb{E}_{\mathbf{z} \sim Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}[\log(1 - S(d_{\boldsymbol{\xi}}(\boldsymbol{z})))] \tag{B.1}$$

$$= \max_{\boldsymbol{\xi}} \int p(\boldsymbol{z}) \log(S(d_{\boldsymbol{\xi}}(\boldsymbol{z}))) d\boldsymbol{z} + \int \int P_D(\boldsymbol{x}) Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \log(1 - S(d_{\boldsymbol{\xi}}(\boldsymbol{z}))) d\boldsymbol{z} d\boldsymbol{x} \tag{B.2}$$

$$= \max_{\boldsymbol{\xi}} \int \Big[P(\boldsymbol{z}) \log(S(d_{\boldsymbol{\xi}}(\boldsymbol{z}))) + \int P_D(\boldsymbol{x}) Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \log(1 - S(d_{\boldsymbol{\xi}}(\boldsymbol{z}))) d\boldsymbol{x} \Big] d\boldsymbol{z}. \tag{B.3}$$

In the last step, we applied Fubini's theorem to change the order in the integration. As in (Goodfellow et al., 2014) and (Mescheder et al., 2017), it can be shown that the discriminator achieves its optimum value at

$$d^*_{\boldsymbol{\xi}}(\boldsymbol{z}) = \log(P(\boldsymbol{z})) - \log\Big(\int Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) P_D(\boldsymbol{x}) d\boldsymbol{x} \Big) = \log(P(\boldsymbol{z})) - \log(Q_{\boldsymbol{\phi}}(\boldsymbol{z})). \tag{B.4}$$

This follows from the fact that for any $(a, b) \in \mathbb{R}^2 \setminus [0, 0]$, a function that has the form $f(h) = a \log h + b \log(1 - h)$ attains it maximum in $[0, 1]$ at $h = a/(a + b)$. Thus, the optimum value of Equation B.3 is

$$S(d^*_{\boldsymbol{\xi}}(\boldsymbol{z})) = \frac{P(\boldsymbol{z})}{P(\boldsymbol{z}) + \int Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) P_D(\boldsymbol{x}) d\boldsymbol{x}}, \tag{B.5}$$

which is equivalent to Equation B.4. The lower bound in Equation 5.5 can be reformulated based on the definition of the KL divergence in Equation A.7 as

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}\sim P_D(\boldsymbol{x})}[\log(P_{\boldsymbol{\theta}}(\boldsymbol{x}))] \geq \mathbb{E}_{\mathbf{x}\sim P_D(\boldsymbol{x})}\mathbb{E}_{\mathbf{z}\sim Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}[\log(P_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}))] - & \quad \text{(B.6)} \\
\mathbb{E}_{\mathbf{x}\sim P_D(\boldsymbol{x})}[D_{KL}(Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||P(\boldsymbol{z}))] & \quad \text{(B.7)} \\
= \mathbb{E}_{\mathbf{x}\sim P_D(\boldsymbol{x})}\mathbb{E}_{\mathbf{z}\sim Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}[\log(P_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}))] + & \\
\mathbb{E}_{\mathbf{x}\sim P_D(\boldsymbol{x})}\mathbb{E}_{\mathbf{z}\sim Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}[\log(P(\boldsymbol{z})) - \log(Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}))]. & \quad \text{(B.8)}
\end{aligned}$$

As described in (Makhzani et al., 2016), the AAE algorithm replaces the last term in Equation B.8 (regularization term, equivalent to the KL term) with "an adversarial procedure that encourages $Q_{\boldsymbol{\phi}}(\boldsymbol{z})$) to match to the whole distribution of $P(\boldsymbol{z})$". Mathematically, this translates into replacing the KL term with $\mathbb{E}_{\mathbf{x}\sim P_D(\boldsymbol{x})}\mathbb{E}_{\mathbf{z}\sim Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}[d^*_{\boldsymbol{\xi}}(\boldsymbol{z})]$, effectively approximating the variational bound as

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}\sim P_D(\boldsymbol{x})}\log\left(P_{\boldsymbol{\theta}}(\boldsymbol{x})\right) \geq \mathbb{E}_{\mathbf{x}\sim P_D(\boldsymbol{x})}\mathbb{E}_{\mathbf{z}\sim Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}[\log(P_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}))] + \mathbb{E}_{\mathbf{x}\sim P_D(\boldsymbol{x})}\mathbb{E}_{\mathbf{z}\sim Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}[d^*_{\boldsymbol{\xi}}(\boldsymbol{z})] & \quad \text{(B.9)} \\
= \mathbb{E}_{\mathbf{x}\sim P_D(\boldsymbol{x})}\mathbb{E}_{\mathbf{z}\sim Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}[\log(P_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}))] - D_{KL}(Q_{\boldsymbol{\phi}}(\boldsymbol{z})||P(\boldsymbol{z})), & \quad \text{(B.10)}
\end{aligned}$$

where, compared to the bound in Equation B.6, the term $\mathbb{E}_{\mathbf{x}\sim P_D(\boldsymbol{x})}[D_{KL}(Q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||P(\boldsymbol{z}))]$ is approximated with $D_{KL}(Q_{\boldsymbol{\phi}}(\boldsymbol{z})||P(\boldsymbol{z}))$. As a result, the AAE translates into a modified variational objective that does not preserve the original formulation.

Nomenclature

Math symbols	Description
$\boldsymbol{x}$	Model input variables.
$\boldsymbol{y}$	Model output variables.
$\hat{\boldsymbol{y}}$	Ground-truth outputs in dataset.
$\boldsymbol{z}$	Continuous latent variables.
$\boldsymbol{c}$	Discrete latent variables.
M	Input dimensionality, typically a 3D voxel grid of $M = H \times W \times L$ of height H, width W and depth L.
N	Latent space dimensionality.
D	Token dimensionality (for transformers).
$\boldsymbol{\theta}$	Model parameters.
$\boldsymbol{\phi}$	Parameters of the variational inference network.
$\boldsymbol{\xi}$	Parameters of a discriminator neural network.
$\boldsymbol{h}$	Token sequence, input of a transformer neural network.
$\boldsymbol{r}$	Positional embedding, input to a transformer neural network.
$\boldsymbol{W}$	Matrix of weights of a linear transformation.
$\boldsymbol{q}$	Queries in a self attention operation.
$\boldsymbol{k}$	Keys in a self attention operation.
$\boldsymbol{v}$	Values in a self attention operation.
$\boldsymbol{A}$	Attention matrix with dynamic weights.
$\boldsymbol{\kappa}$	Additional model input variable with beam shape information.
$\boldsymbol{s}_x$	Organ structure masks from image $\boldsymbol{x}$.
$\hat{\boldsymbol{s}}_x$	Ground-truth organ structure masks from image $\boldsymbol{x}$.
$\boldsymbol{p}$	3D Cartesian coordinates of a given point.
$\boldsymbol{\Phi}$	Diffeomorphic deformation vector field.
$\boldsymbol{u}$	Stationary velocity field characterizing $\boldsymbol{\Phi}$.
λ_i	Hyper-parameter multiplying the i^{th} loss term.
$P(x)$	Probability distribution function of a random variable x.
$Q(x)$	Posterior probability distribution function of a random variable x
$P^*(x)$	Underlying unknown ground truth probability distribution function (to be approximated).
$P_D(x)$	Empirical distribution of data points x in dataset.

$f_{\boldsymbol{\theta}}, g_{\boldsymbol{\phi}}, d_{\boldsymbol{\xi}}$	Functions computed by neural network with parameters $\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\xi}$, respectively.
$D_{KL}(P\|Q)$	Kullback-Leibler divergence between probability distributions $P(x)$ and $Q(x)$.
$\mathcal{N}(x;\mu,\sigma)$	Gaussian probability distribution of variable x with mean μ and standard deviation σ.
$\text{Cat}(x;\boldsymbol{\zeta})$	Categorical distribution of variable x with vector of probabilities for each class $\boldsymbol{\zeta}$.
$\boldsymbol{\tau}$	Vector of time and position stamps of a breathing signal.
T	Breathing period of a breathing signal.
A	Breathing amplitude of a breathing signal.
δ^n_Λ	n^{th} - percentile of a quantity of interest Λ.
$\|\boldsymbol{a}\|_1$	L1 norm of vector $\boldsymbol{a}$, calculated as the sum of the absolute value of its elements.
$\|\boldsymbol{a}\|_2$	L2 norm of vector $\boldsymbol{a}$, calculated as the square root of the sum of the squares of its elements.

Metrics	**Description**	**Units**
$\gamma(\boldsymbol{p})$	Gamma value for a voxel centered at point $\boldsymbol{p}$.	-
$\Gamma(d_{ta}, d_d)$	Gamma analysis with distance-to-agreement d_{ta} and dose difference d_d criteria.	-
ρ	Average absolute error, relative to the maximum value.	%
D_p	Prescribed dose.	Gy
D_v	Minimum dose received by v% of the volume.	Gy
V_f	Fraction of the volume receiving at least a percentage f of the prescribed dose.	%
HI	Homogeneity index, calculated as $(D_2 - D_{98})/D_p$.	-
$V_{107/95}$	Fraction of the volume receiving a dose between the 95% and 107% of the prescribed dose, calculated as $V_{107} + (1 - V_{95})$.	%
$\text{RDE}_{\Lambda_1,\Lambda_2}$	Error between two quantities of interest Λ_1 and Λ_2, relative to a reference value (e.g., the prescribed dose).	%
MSE	Mean squared error between predictions $\boldsymbol{y}$ and ground-truth values $\hat{\boldsymbol{y}}$.	(Same as $\boldsymbol{y}$)^2
RMSE	Root mean squared error between predictions $\boldsymbol{y}$ and ground-truth values $\hat{\boldsymbol{y}}$.	Same as $\boldsymbol{y}$
CC	Cross correlation between between predictions $\boldsymbol{y}$ and ground-truth values $\hat{\boldsymbol{y}}$.	-

Radiotherapy abbreviations	**Description**
RT	Radiation therapy.

CT	Computed tomography.
MR	Magnetic resonance.
DVH	Dose volume histogram.
HU	Hounsfield unit.
PBA	Pencil beam algorithm.
MC	Monte Carlo.
DVF	Deformation vector field.
IMRT	Intensity modulated radiation therapy.
IMPT	Intensity modulated proton therapy.
VMAT	Volumetric modulated arc therapy.
4DCT	Four-dimensional computed tomography.
ITV	Internal target volume.
GTV	Gross tumor volume.
CTV	Clinical target volume.
PTV	Planning target volume.

Deep learning abbreviations	**Description**
ANN	Artificial neural network.
GAN	Generative adversarial network.
VAE	Variational autoencoder.
AAE	Adversarial autoencoder.
SA	Self-attention.
MSA	Multi-head self attention.
ReLU	Rectified linear unit activation.
MLP	Multi layer perceptron.
SGD	Stochastic gradient descent.

Other	**Description**
PCA	Principal component analysis.
GPU	Graphics processing unit.
MeV	Mega electron-volt (beam energy unit).
H&N	Head and neck.

Acknowledgements

The way of the PhD student is one of a kind. I am grateful to all the wonderful people that have provided support and encouragement along the way, helping me grow both personally and professionally.

First and foremost, I would like to thank my daily supervisor Zoltán Perkó. During these last 4 years at TU Delft, Zoltán has shown me the light side of the force, always leading by example in how to be a critical thinker and creative researcher. Playing a very important role in my professional development, Zoltán has always supported me and cared about my personal goals, and for that I will be eternally grateful. I would also like to thank my promotors Mischa Hoogeman and Dennis Schaart for their constant feedback and interesting discussions, which have undoubtedly been of great importance for my work. By asking the right questions, Dennis and Mischa have constantly made sure that the research ideas we developed were always addressing clinical problems. Last, but not least, I am grateful to Danny Lathouwers and Steven Habraken for their support, from theoretical discussions to providing clinical insights and data. I feel lucky to have worked with such an outstanding supervisory team.

My period as a visiting researcher in Stanford University has been an immensely enriching experience. I am extremely thankful to Lei Xing for giving me the opportunity to join his group and for being an excellent host. I would also like to thank my collaborators Peng, Charles, Varun and Yusuke, as well as the rest of the members in Xing's lab. Apart from the very interesting scientific discussions, they all welcomed me with open arms and showed me what is like to do research in such a high-performing environment. I would also like to thank all the great people I met in sunny California, the German gang and especially my dearest Sri, Zissy, Kathy and Marius. All these fellow researchers and good friends made my stay at Stanford one of the most fulfilling of my life.

My stay in the Reactor Institute at TU Delft has been full of positive experiences. First, I would like to thank my fellow PhD colleagues Jaen, Thomas, Marc, Tibi, Celebrity, Bouke, Andries, Anand and Daniel for the fun times inside and outside the office, from Fridays at 't Koepeltje to home made pizzas (extra shout-out to Marc for helping translate the Summary). Likewise, I am also grateful to the great people in the Reactor Physics and Nuclear Materials group, who hosted me in their group even though my research had little to do with nuclear reactors. To Fahad, Marco, and the old batch of PhDs I owe my gratitude for welcoming me to the group and their advice, and I wish Mikolaj, Ana and Nick good luck with their PhD endeavor. Especial shout-out to Aldo and Matteo, who also became good friends despite our failed attempts to organize ski trips. During my PhD, I also had the chance to supervise 11 brilliant Master and Bachelor students, all of whom have contributed to the research presented in this thesis. I wish all of them the best of luck in their future career. Finally, as a side project, I also had the pleasure of working with Sarwan, Dilnoza, Theresa, and the THRIVE Institute

team in building a more circular healthcare system, to whom I am grateful for the work and dinner sessions that helped to cheer up the boring lock-down days.

I would like to conclude thanking my paranymphs (and most importantly, my close friends) Marius and Jesus. Their friendship is one of the most important things I take from these four years, and I cannot think of better companions for sharing the stand during my defense ceremony. I want to thank my close friend Fran (who I expect to keep chasing around the world), and my dear Conchita and Albert, all of which brought their Spanish warmth to make the cold and rainy autumns and winters more bearable. Para acabar, gracias a todas las personas en España, mi familia y amigos que me han acompañado durante todos estos años. Y, sobre todo, gracias a Oscar, Carmen, Dani, Isa, y Paloma, mi paranymph vitalicia. Gran parte de este gran paso es gracias a vosotros.

About the author

Oscar Pastor Serrano was born in 1994 in Valencia, Spain. He attended Salesianos San Antonio Abad school in Valencia from kindergarten to high school. In 2012, he enrolled in the Bachelor Degree of Energy Engineering at Valencia Polytechnic University (Universidad Politecnica de Valencia) in Valencia, Spain. He spent the last semester of his Bachelor studies as a visiting student at the University of Wisconsin-Madison in Wisconsin, USA, learning about nuclear physics and Monte Carlo transport methods. After graduating in 2016, he enrolled in the Master's degree in Nuclear Engineering at the Royal Institute of Technology (KTH Kungliga Tekniska Högskolan) in Stockholm, Sweden, where he specialized in physics-based simulations for nuclear reactor and medical physics problems. During his master thesis internship at the radiotherapy company Elekta AB in Stockholm, he developed Monte Carlo particle transport software simulating the interaction of charged particles (up to carbon ions) in media, applied to radiation dose calculations. In 2019, he joined the Radiation Science and Technology section of the Delft University of Technology in Delft, Netherlands, as a PhD candidate. While pursuing his PhD studies, he had the opportunity to spend 6 months as a guest researcher at the Laboratory of Artificial Intelligence and Medicine and Biomedical Physics of Stanford University in California, USA. The results of his 4 years of research at TU Delft and Stanford University are summarized in this thesis.

List of Publications

Journal publications

6. **O. Pastor-Serrano**, S. Habraken, M. Hoogeman, D. Schaart, D. Lathouwers, Y. Nomura, L. Xing, Z. Perkó, *A probabilistic deep learning model of inter-fraction anatomical variations in radiotherapy*, Accepted for publication in Physics in Medicine and Biology (2023).

5. **O. Pastor-Serrano**, P. Dong, C. Huang, L. Xing, Z. Perkó, *Sub-second photon dose prediction via transformer neural networks*, Accepted for publication in Medical Physics (2023).

4. C. Huang, V. Vasudevan, **O. Pastor-Serrano**, M.T. Islam, Y. Nomura, P. Dubrowski, J. Wang, J. Schulz, Y. Yang, L. Xing, *Learning image representations for content based image retrieval of radiotherapy treatment plans*, Under review in Physics in Medicine and Biology (2022).

3. **O. Pastor-Serrano**, Z. Perkó, *Millisecond speed deep learning based proton dose calculation with Monte Carlo accuracy*, Physics in Medicine and Biology 67 (10), 105006 (2022).

2. **O. Pastor-Serrano**, S. Habraken, D. Lathouwers, M. Hoogeman, D. Schaart, Z. Perkó, *How should we model and evaluate breathing interplay effects in IMPT?*, Physics in Medicine and Biology 66 (23), 235003 (2021).

1. **O. Pastor-Serrano**, D. Lathouwers, Z. Perkó, *A semi-supervised autoencoder framework for joint generation and classification of breathing*, Computer Methods and Programs in Biomedicine 209, 106312 (2021).

Conference publications

5. **O. Pastor-Serrano**, Z. Perkó, *Learning the physics of particle transport via transformers*, Proceedings of the AAAI Conference on Artificial Intelligence, 36(11), 12071-12079 (2022).

4. V. Vasudevan, **O. Pastor-Serrano**, C. Huang, C. Chuang, Z. Perkó, P. Dong, L. Xing, *Three-dimensional dose super-resolution using implicit neural representation*, Medical Physics 49 (6), E711-E712.

3. H. van der Wind, **O. Pastor-Serrano**, S. Habraken, D. Schaart, D. Lathouwers, M. Hoogeman, Z. Perkó, *Evaluating the effectiveness of interplay mitigation techniques in proton therapy*, Radiotherapy and Oncology 170, S1499-S1500 (2022).

2. **O. Pastor-Serrano**, Z. Perkó, *Sub-second speed proton dose calculation with Monte Carlo accuracy using deep learning*, Radiotherapy and Oncology 170, S10-S11 (2022).

1. **O. Pastor-Serrano**, S. Habraken, D. Lathouwers, M. Hoogeman, D. Schaart, Z. Perkó, *Breathing motion robustness of 4D-CT and ITV based treatment plans in lung cancer IMPT*, Radiotherapy and Oncology 152, S158-S159 (2020).