Department of Physics and Astronomy

Heidelberg University

Master thesis in Physics

submitted by

Josch Hagedorn

born in Jever (Germany)

2024

Accelerated Dose Calculation in Radiotherapy using Latent Diffusion Models

This Master thesis has been carried out by Josch Hagedorn at the

Data Analysis and Modeling in Medicine Group, Mannheim Institute

for Intelligent Systems in Medicine (MIISM)

under the supervision of

Prof. Dr. Jürgen Hesser

Zusammenfassung:

Die genaue Berechnung der Dosis ist ein zentraler Bestandteil der effektiven Strahlentherapie, insbesondere in komplexen anatomischen Strukturen und heterogenen Geweben. Diese Arbeit untersucht einen neuartigen Ansatz zur Dosisberechnung durch die Integration von latenten Diffusionsmodellen (LDMs) mit Techniken der gaußschen Quadratur. Das Hauptziel besteht darin, die rechnerische Effizienz der Dosisberechnung zu verbessern und gleichzeitig eine hohe Genauigkeit zu gewährleisten, die mit herkömmlichen Monte-Carlo-Simulationen (MC) vergleichbar ist.

Diese Forschung konzentriert sich auf das Training von LDMs, die auf spezifische Energiebereiche zugeschnitten sind, die besonders für klinische Strahlentherapieanwendungen relevant sind. Die Integration der gaußschen Quadratur bietet eine systematische Methode zur Darstellung kontinuierlicher Energiespektren durch diskrete Punkte und optimiert so den Dosisvorhersageprozess.

Eine Reihe von Experimenten mit CT-Daten und MC-generierten Referenzdaten wurden durchgeführt, um die Leistung der vorgeschlagenen Methode zu evaluieren. Die Ergebnisse zeigen, dass der hybride Ansatz nicht nur die Dosisberechnung beschleunigt, sondern auch eine hohe Genauigkeit erreicht, mit Gamma-Pass-Raten von über 90% unter strengen Kriterien. Die Ergebnisse deuten darauf hin, dass das vorgeschlagene Framework das Potenzial hat, die Lücke zwischen den rechnerischen Anforderungen von MC-Simulationen und den praktischen Erfordernissen klinischer Arbeitsabläufe zu schließen und den Weg für eine adaptivere Planung der Strahlentherapie zu ebnen.

Abstract:

Accurate dose calculation is a cornerstone of effective radiotherapy, particularly in scenarios involving complex anatomical structures and heterogeneous tissues. This thesis explores a novel approach to dose calculation by integrating latent diffusion models (LDMs) with Gaussian quadrature techniques. The primary objective is to enhance the computational efficiency of dose calculations while maintaining high accuracy, comparable to traditional Monte Carlo (MC) simulations.

This research focuses on training LDMs tailored to specific energy ranges, particularly those critical for clinical radiotherapy applications. The integration of Gaussian quadrature provides a systematic method for representing continuous energy spectra through discrete points, further optimizing the dose prediction process.

A series of experiments were conducted using CT data and MC-generated ground truth to evaluate the performance of the proposed method. The results demonstrate that the hybrid approach not only accelerates dose calculations but also achieves a high degree of accuracy, with gamma pass rates exceeding 90% under stringent criteria. The findings suggest that the proposed framework has the potential to bridge the gap between the computational demands of MC simulations and the practical requirements of clinical workflows, paving the way for more adaptive radiotherapy treatment planning.

Contents

List of Abbreviations

CNN Convolutional neural network.

CPU Central processing unit.

CT Computed tomography.

DD Dose difference.

DDIM Denoising diffusion implicit model.

DL Deep learning.

DTA Distance-to-agreement.

DVH Dose-volume histogram.

eBT Electronic brachytherapy.

EMPIR European metrology programme for innovation and research.

GAN Generative adversarial network.

GPU Graphics processing unit.

HNSCC Head and neck squamous cell carcinoma.

HPC High-performance computing.

HU Hounsfield unit.

IMRT Intensity-modulated radiation therapy.

IQR Interquartile range.

LDM Latent diffusion model.

linac Linear accelerator.

MC Monte Carlo.

MDACC MD Anderson Cancer Center.

MLC Multi-leaf-collimator.

MMFNet Multi-encoder and multi-scale fusion network.

MRI Magnetic resonance imaging.

OAR Organs at risk.

PTV Planning target volume.

RHEL Red Hat Enterprise Linux.

ROI Region-of-interest.

RSD Relative standard deviation.

SDM Signed distance maps.

TCIA The cancer imaging archive.

TERMA Total energy released per unit mass.

VAE Variational autoencoder.

VMAT Volumetric modulated arc therapy.

1 Introduction

Radiotherapy is a critical treatment modality for cancer, utilizing high-energy radiation to target and destroy cancer cells.[1] The success of radiotherapy depends heavily on accurate dose calculations, which ensure the prescribed radiation dose effectively eliminates cancer cells while minimizing exposure to surrounding healthy tissues.[2] Traditional dose calculation methods, such as the pencil beam algorithm, the superposition-convolution algorithm, and Monte Carlo (MC) simulations, each have their own strengths. Among these, MC simulations are regarded as the gold standard due to their ability to precisely model complex interactions between radiation and biological tissues, particularly in heterogeneous regions and at tissue interfaces.[3] However, the significant computational demands of MC simulations pose a major limitation in clinical settings where timely decision-making is essential.[4]

The emergence of advanced computational techniques and machine learning models offers new avenues for improving the efficiency of dose calculations in radiotherapy. Machine learning approaches, especially those involving deep learning frameworks like convolutional neural networks (CNNs) and U-Nets, have demonstrated potential in accelerating dose calculations by learning intricate mappings from patient anatomy to dose distributions.[5] However, these models often struggle with generalization, particularly across diverse clinical scenarios involving varying energy spectra and complex anatomical structures. This limitation restricts their broader clinical application and underscores the need for more adaptable and efficient solutions.[6]

This thesis investigates the use of latent diffusion models (LDMs) combined with Gaussian quadrature techniques to address the challenges of speed and accuracy in radiotherapy dose calculations.[7] Diffusion models are a class of generative models that iteratively transform simple distributions into complex data distributions through a series of noise addition and removal steps. By operating in a reduced-

dimensional latent space, LDMs can maintain high generative quality while significantly reducing computational costs, making them particularly well-suited for high-dimensional tasks such as medical image analysis.[8] The integration of Gaussian quadrature enhances this approach by providing a systematic method for accurately representing continuous energy spectra through discrete points, potentially enabling dose predictions with fewer computational resources.

The primary goal of this research is to develop a novel method that integrates LDMs and Gaussian quadrature for accelerated dose calculation in radiotherapy, aiming to improve the efficiency of the process. This method focuses on training diffusion models tailored to specific energy ranges within the radiation spectrum, thereby capturing the critical energy peaks that influence therapeutic outcomes. The use of Gaussian quadrature allows for a refined representation of the energy spectrum, improving the accuracy of dose calculations in clinical scenarios where precise adjustments are crucial, such as in treatments involving high-gradient regions or complex anatomical interfaces.

A key aspect of this thesis is the exploration of how LDMs can be adapted to radiotherapy by incorporating domain-specific requirements, such as energy-specific training and the use of Gaussian quadrature for numerical integration. This approach not only addresses the computational challenges associated with traditional MC simulations but also enhances the adaptability of treatment plans, allowing for quicker and more accurate adjustments. By bridging the gap between the high accuracy of MC simulations and the computational efficiency needed in clinical practice, this research aims to provide a robust framework that could significantly impact the field of radiotherapy, improving patient outcomes through more efficient and precise treatment planning.

In summary, this thesis proposes a hybrid approach that leverages the strengths of latent diffusion models and Gaussian quadrature to improve dose calculations in radiotherapy. By addressing the key research challenges of computational efficiency, accuracy, and adaptability, this work aims to contribute a novel solution that balances the precision of traditional methods with the practical demands of modern clinical workflows.

2 Theoretical Background

2.1 Particle Interactions in Radiotherapy

In radiotherapy, photons can interact with matter through several mechanisms, but only a few are significant within the energy ranges used, which typically span from a few keV in brachytherapy to several MeV in external beam radiotherapy using linear accelerators (linacs). The principal interactions relevant to this context are the photoelectric effect, Compton scattering, and pair production. These processes must be accurately modeled in Monte Carlo (MC) simulations to ensure precise dose calculations.

The mass attenuation coefficient, denoted as $\frac{\mu}{\rho}$, is a critical parameter representing the probability of photon interaction per unit mass of the medium and is influenced by photon energy. The coefficient $\frac{\mu}{\rho}$ depends on both the photon energy and the material it traverses, dictating how radiation is attenuated. The energy-dependent behavior of these interactions is illustrated in Figure 2.1, showing the mass attenuation coefficient for water and tungsten across a wide range of photon energies. For low photon energies, particularly below 100 keV, the photoelectric effect predominates, where photons are absorbed entirely by ejecting inner-shell electrons. As the photon energy increases into the range of 100 keV to 10 MeV, Compton scattering becomes the dominant process, where photons are scattered with a transfer of energy to outer-shell electrons. Above 10 MeV, pair production becomes relevant, in which photons are converted into electron-positron pairs in the vicinity of a nucleus.

These interactions not only determine the attenuation of the primary beam but also influence the generation of secondary radiation, which contributes to the dose deposited in tissues. Understanding and accurately modeling these interactions are crucial for optimizing treatment plans and minimizing radiation exposure to healthy tissues.

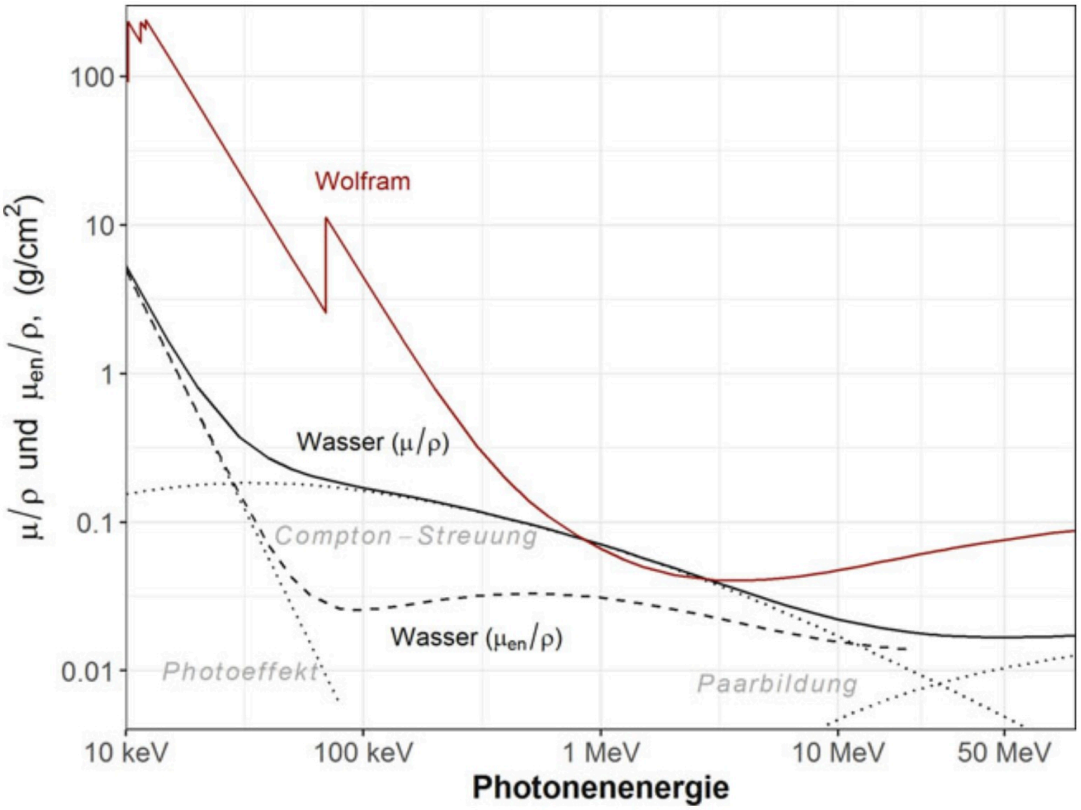


Figure 2.1: Mass attenuation coefficients $\frac{\mu}{\rho}$ and energy transfer coefficients $\frac{\mu_{en}}{\rho}$ for water and tungsten as a function of photon energy. The graph illustrates the dominance of different interaction processes: photoelectric effect, Compton scattering, and pair production. [2]

The probability of a photon interacting while traveling through a medium can be quantified by the mean free path, which describes the average distance a photon travels before undergoing an interaction. For photons relevant to radiotherapy, this path length typically ranges from a few centimeters in low-density materials like water to millimeters in denser materials like bone. In contrast, for electrons, which are secondary particles produced through these interactions, the mean free path is significantly shorter, usually in the range of nanometers to micrometers, depending on their energy and the medium.[2]

The interactions of photons with matter as shown in Figure 2.1 are fundamental to radiotherapy, influencing both the direct attenuation of the therapeutic beam and the generation of secondary particles that contribute to the overall dose distribution within the patient's body. By understanding these interactions, we can

better design and optimize radiotherapy treatments to maximize their therapeutic effect while minimizing damage to surrounding healthy tissues.

2.1.1 Photoelectric Effect

The photoelectric effect occurs when an incoming photon collides with a tightly bound electron, usually in the inner shells of an atom. In this interaction, the photon transfers all of its energy to the electron and is completely absorbed, resulting in the ejection of the electron, known as a photoelectron. The kinetic energy of the photoelectron is given by:

$$E_c = E - U_i, \tag{2.1}$$

where E_c is the kinetic energy of the ejected electron, E is the energy of the incident photon, and U_i is the binding energy of the electron. Once released, the photoelectron can ionize surrounding atoms, producing secondary particles that contribute to the dose deposited in tissues. This cascading effect is significant, particularly in materials with higher atomic numbers. The probability of the photoelectric effect occurring, expressed through the photoelectric cross section, is highly dependent on both the energy of the photon and the atomic number Z of the absorbing material. The cross section follows the relationship:

$$\sigma_{\text{photoelectric}}(Z, E) \propto \begin{cases} Z^{3.8} E^{-3}, & \text{for } Z \geq 16, \\ Z^{3} E^{-3}, & \text{for } Z < 16, \end{cases} \tag{2.2}$$

indicating that it is inversely proportional to the cube of the photon energy and exhibits a strong dependence on the atomic number of the material. This dependence results in a higher interaction probability in tissues containing elements with high atomic numbers, such as calcium (with $Z = 20$), which is prevalent in bone tissue. Consequently, the photoelectric effect contributes dominantly to dose deposition in bone, compared to soft tissues composed of elements with lower atomic numbers. This Z-dependence makes the photoelectric effect particularly important at lower photon energies, especially below 100 keV, where it is the dominant interaction mechanism. However, as the photon energy increases, Compton scattering becomes the more prevalent interaction. Understanding this dependency is crucial for treatment planning in radiotherapy, as it significantly impacts dose distribution,

particularly in areas with varying tissue compositions.[2]

2.1.2 Compton Scattering

Compton scattering is the most significant interaction process at therapeutic photon energies, typically in the range of a few MeV, which are commonly used in radiotherapy. This process involves the incoherent scattering of photons by free or loosely bound electrons, often in the outer shells of atoms. In a Compton interaction, a photon collides with a valence shell electron, transferring part of its energy to the electron and scattering both the photon and the electron in different directions. The scattered photon, now with reduced energy, can continue to undergo further interactions, while the ejected electron, known as the Compton electron, can ionize surrounding molecules, leading to the production of secondary particles. The energy transferred to the electron during Compton scattering is given by:

$$E' = E - E_c, \tag{2.3}$$

where E is the energy of the incident photon, and E_c is the kinetic energy of the ejected Compton electron. The angles at which the photon and electron are scattered depend on the amount of energy transferred during the interaction. The cross section for Compton scattering, which quantifies the probability of this interaction occurring, is described by:

$$\sigma_{\text{Compton}}(Z, E) \propto ZE^{-1/2}, \tag{2.4}$$

indicating that the interaction is nearly independent of the atomic number Z of the material and only weakly dependent on the energy E of the incident photon. This weak dependence on Z means that Compton scattering occurs with similar probability in both low- and high-Z materials, making it the dominant interaction in soft tissues as well as in bone at therapeutic photon energies. The significance of Compton scattering in radiotherapy lies in its relatively constant cross section across different tissue types, which contributes to a more uniform dose distribution. As the photon scatters multiple times, losing energy in each interaction, it continues to contribute to ionization processes, which are crucial for the therapeutic effect of the radiation. Understanding the role of Compton scattering is therefore essential for accurate dose calculations and effective treatment planning in radiotherapy.[2]

2.1.3 Pair Production

Pair production is a photon interaction process that occurs at photon energies above 1.022 MeV, which is significantly higher than the energy ranges considered in this context (up to 50 keV). In this process, a photon is converted into an electron-positron pair when interacting with the nuclear field of an atom. Since the energy threshold for pair production far exceeds the relevant energy range of therapeutic radiotherapy considered here, this interaction does not contribute to dose deposition and can be neglected in calculations for photon energies up to 50 keV. Therefore, pair production is not relevant for the current analysis and does not impact the treatment planning or dose calculations in the low-energy range typically used in brachytherapy or external beam radiotherapy with energies below 50 keV.[2]

2.1.4 Rayleigh Scattering

Rayleigh scattering, also known as coherent scattering, involves the scattering of photons by atoms without any energy transfer to the medium. This process primarily affects the direction of the photon rather than contributing to ionization or energy deposition. In the energy range relevant to this analysis (up to 50 keV), Rayleigh scattering results in small scattering angles and does not significantly alter the photon's energy. Due to its negligible impact on dose deposition, Rayleigh scattering is often of minor importance in radiotherapy calculations. Although it can occur in light materials, its contribution remains small, especially when compared to other interaction processes like the photoelectric effect and Compton scattering. For photon energies above 100 keV, Rayleigh scattering becomes even less significant and can typically be ignored in radiotherapy applications. Therefore, in the context of photon energies up to 50 keV, Rayleigh scattering is primarily interesting for its directional effects but does not contribute to the absorbed dose and can generally be neglected in dose calculations.[2]

2.2 Radiation Treatment Planning

Radiation therapy utilizes high-energy particles such as X-rays, gamma rays, electrons, neutrons, and protons to destroy cancer cells. The primary objective of radiation therapy is to deliver a sufficiently high dose to the tumor to control or eradicate cancer cells while minimizing the exposure to surrounding healthy tissues

and organs at risk. Achieving this balance is critical for effective treatment and reducing side effects.

Successful radiation therapy relies on precise treatment planning, which involves selecting the appropriate technique, type of radiation, and determining parameters such as the size, number, and direction of the individual radiation beams. This planning process is facilitated by simulation tools that predict the dose distribution within the patient's body, ensuring that the treatment plan will achieve the desired therapeutic outcome.

Several algorithms are used to calculate dose distributions in treatment planning, each with its strengths and limitations. Commonly used methods include the pencil beam algorithm, the superposition-convolution algorithm, and the Monte Carlo (MC) method.[9]

The pencil beam algorithm models an electron beam as a series of narrow "pencils" that pass through the collimator and are subsequently scattered in air and tissue, redistributing in a Gaussian manner as they travel deeper into the body. This algorithm is computationally efficient and quick, making it suitable for routine clinical use, but it tends to offer only moderate accuracy, particularly in regions with complex tissue heterogeneity or near interfaces.[10]

The superposition-convolution algorithm, on the other hand, provides improved accuracy by accounting for variations in tissue density. It calculates the dose by convolving the "total energy released per unit mass (TERMA)" with energy deposition kernels specific to different tissue types and densities. This method provides a better approximation of the dose distribution in heterogeneous media compared to the pencil beam approach.[11]

The Monte Carlo method represents the gold standard in dose calculation due to its ability to accurately simulate the full transport of particles, including complex interactions at tissue interfaces and within air cavities. It employs statistical techniques based on random sampling to model the probabilistic nature of particle interactions. Despite its superior accuracy, the MC method is computationally intensive and requires significantly more processing time than other algorithms, which can limit its

routine clinical use.[12]

Choosing the appropriate dose calculation algorithm depends on the clinical scenario, the required accuracy, and the available computational resources. Advances in computational power and algorithm optimization continue to enhance the feasibility of using Monte Carlo methods in clinical practice, aiming to achieve the best possible balance between accuracy and efficiency in radiation treatment planning. [3]

2.3 Monte Carlo Method

2.3.1 Monte Carlo Simulations

Monte Carlo (MC) simulations are a versatile computational method used to model and analyze systems with inherent randomness. Instead of relying on fixed input values, MC simulations use random sampling from defined probability distributions (such as normal or uniform distributions) to explore a wide range of possible outcomes. This approach allows the simulation of complex processes by repeatedly recalculating results with different sets of randomly selected inputs.

MC simulations are particularly valuable in scenarios where uncertainty and variability play a significant role, such as in finance, engineering, and scientific research. By running thousands or even millions of iterations, they provide insights into the likelihood of various outcomes, enabling a comprehensive understanding of potential risks and variations in predictions. This iterative approach helps to capture the impact of uncertainty more effectively than traditional deterministic models.

A common use of MC simulations is in risk analysis, where they help quantify the probability of different outcomes and assess the sensitivity of results to changes in input variables. For example, rather than calculating fixed probabilities for a simple event like rolling dice, an MC simulation would simulate rolling the dice repeatedly, offering a dynamic view of potential results.

Overall, MC simulations offer a powerful tool for exploring uncertainties in complex systems, providing a robust framework for decision-making and analysis in

environments where outcomes are not easily predictable.[13]

2.3.2 Monte Carlo Method in Radiotherapy

In radiotherapy, Monte Carlo (MC) simulations are widely used for accurately modeling the transport and interaction of radiation within the human body. These simulations are employed to calculate dose distributions for various radiation types and equipment, providing a detailed representation of how radiation interacts with complex anatomical structures. Among all dose calculation algorithms, MC simulations are considered the gold standard due to their superior precision, particularly in heterogeneous regions such as those with varying tissue densities or at interfaces between different materials.[14]

MC simulations work by individually tracking the paths of particles, such as photons or electrons, as they traverse through a medium, which can be derived from patient-specific data like CT scans. These scans are used to create a detailed map of the patient's anatomy, including any variations in tissue composition. The simulation involves modeling various physical interactions between the radiation and the tissues, including photoelectric effect, Compton scattering, and secondary ionization processes.[15]

To enhance computational efficiency, MC simulations can be tailored to focus only on the most relevant interactions that affect dose deposition, such as changes in particle direction, energy deposition, and the production of secondary particles. Particles are tracked until they reach a predefined cutoff energy, below which their contribution to the dose is considered negligible. This approach helps in reducing computational load without significantly sacrificing accuracy.[16]

Once the simulations are complete, the results from the individual particle interactions are aggregated to calculate the absorbed dose in the targeted regions. The level of detail and accuracy provided by MC simulations makes them especially valuable in complex clinical scenarios, such as treatments involving irregular tumor shapes, the presence of implants, or regions with significant tissue heterogeneity. Despite their computational demands, ongoing advances in computing power and optimization techniques continue to make MC simulations increasingly feasible for routine use in radiotherapy planning, ensuring precise dose delivery to target tissues

while minimizing exposure to surrounding healthy structures.[17]

2.4 Diffusion Models and Latent Diffusion Models

2.4.1 Diffusion Models

Diffusion models are a class of generative models used to learn complex data distributions by gradually transforming simple noise into structured data through a series of iterative steps. These models work by defining a forward diffusion process, which incrementally adds noise to the data, and a reverse diffusion process, which denoises the corrupted data to reconstruct the original distribution. The forward process is typically defined by a Markov chain that progressively perturbs the data with Gaussian noise, while the reverse process aims to learn the denoising steps required to reverse this noise addition. Mathematically, the forward diffusion process can be represented as:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \tag{2.5}$$

where $\mathbf{x}_t$ is the noisy data at step t, β_t are variance schedule parameters that control the amount of noise added at each step, and $\mathcal{N}$ denotes a Gaussian distribution. The reverse process, parameterized by a neural network, is trained to learn the reverse of the forward diffusion:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)), \tag{2.6}$$

where μ_θ and Σ_θ are the mean and variance predicted by the neural network at each step.[7]

As illustrated in Figure 2.2, the diffusion model consists of a sequence of noisy representations that progressively become clearer as the reverse process attempts to reconstruct the initial data from the corrupted versions. Diffusion models have gained popularity due to their ability to generate high-quality samples and model complex data distributions with high fidelity. They are used in various applications such as image generation, audio synthesis, and more recently, in medical imaging and dose calculation in radiotherapy.[7]

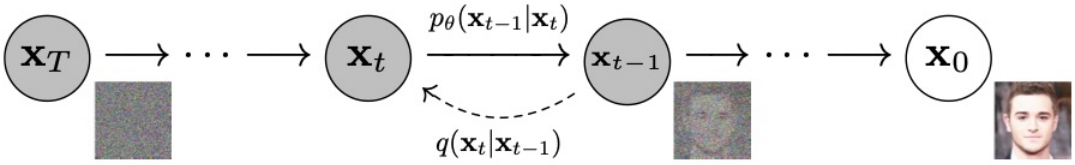


Figure 2.2: The directed graphical model representing the diffusion process: the forward process $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ adds noise step-by-step, while the reverse process $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ denoises to recover the original data distribution.[7]

2.4.2 Conditioning Diffusion Models

Diffusion models, while inherently generative, can be conditioned to produce outputs that adhere to specific requirements or constraints. Conditioning allows these models to generate data that aligns with particular characteristics, which is crucial in applications such as controlled image generation, targeted dose distribution in radiotherapy, or other tasks where outputs need to be tailored to predefined conditions.[18]

Conditioning in diffusion models is typically achieved by introducing additional information into the model during both the forward and reverse processes. This additional information, referred to as the conditioning variable $\mathbf{y}$, can represent various forms of guidance such as class labels, structural data, or external measurements. [8]

Common methods for incorporating conditioning into diffusion models include:

- **Concatenation**: The conditioning information $\mathbf{y}$ is concatenated with the input data or intermediate representations at each diffusion step. This direct approach is straightforward and effective for conditions that are compatible with the data dimensionality.[19]
- **Cross-Attention Mechanisms**: For more complex conditioning, such as text-to-image generation, cross-attention mechanisms are used. These mechanisms allow the model to focus on specific parts of the conditioning information dynamically, enhancing the model's ability to generate outputs that closely match the desired attributes.[20]
- **Classifier Guidance**: In classifier guidance, a pretrained classifier is used to influence the reverse diffusion process by adjusting the gradients of the gener-

ative process. This approach leverages an external model to refine the outputs according to the condition $\mathbf{y}$, providing a flexible way to implement conditioning without altering the diffusion model's architecture significantly.[21]

- **Score-Based Conditioning**: By modifying the score function used in the reverse process, conditioning can be incorporated directly into the denoising steps. This involves adjusting the score function with respect to the conditioning variable, which helps guide the trajectory of the reverse diffusion in a manner that aligns with the conditions.[22]

Conditioning enhances the utility of diffusion models across various applications, allowing them to generate outputs that meet specific criteria or constraints. In the context of radiotherapy, conditioning diffusion models can be used to tailor dose distributions to patient-specific anatomical or clinical requirements, thereby improving the precision and effectiveness of the treatment planning.[23]

This ability to condition diffusion models not only broadens their applicability but also aligns them closely with practical needs in domains where controlled generation is essential.

2.4.3 Latent Diffusion Models

Latent diffusion models (LDMs) are an extension of standard diffusion models that operate in a lower-dimensional latent space rather than the original high-dimensional data space. The primary advantage of LDMs is their ability to significantly reduce computational costs while maintaining high generative quality. This is achieved by first encoding the high-dimensional data into a compact latent representation using an encoder network, often a variational autoencoder (VAE) or a similar architecture.[24] The diffusion process is then applied within this latent space, which requires fewer resources and less time due to the reduced dimensionality.[25]

In LDMs, the forward and reverse diffusion processes are performed in the latent space defined by the encoder, and a decoder network is used to map the latent representation back to the data space. This approach can be mathematically expressed as:

$$q(\mathbf{z}_t|\mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t; \sqrt{1-\beta_t}\mathbf{z}_{t-1}, \beta_t\mathbf{I}), \tag{2.7}$$

$$p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t) = \mathcal{N}(\mathbf{z}_{t-1}; \mu_\theta(\mathbf{z}_t, t), \Sigma_\theta(\mathbf{z}_t, t)), \tag{2.8}$$

where $\mathbf{z}_t$ represents the latent variables at step t, and the mean and variance parameters μ_θ and Σ_θ are learned functions specific to the latent space dynamics. After the reverse diffusion process, the latent representation is decoded back into the data space using the decoder network:

$$\hat{\mathbf{x}} = \text{Decoder}(\mathbf{z}), \tag{2.9}$$

where $\hat{\mathbf{x}}$ is the reconstructed data in the original data space.[25]

Latent diffusion models are particularly well-suited for tasks where the original data space is high-dimensional, such as in medical imaging, where they can generate detailed and realistic outputs with reduced computational burden.[26]

3 State of the art

In recent years, the application of advanced computational techniques in radiotherapy has seen significant developments, particularly in the context of dose calculation and optimization. Accurate dose estimation is crucial for ensuring effective treatment while minimizing harm to surrounding healthy tissue. Among the various methods employed, Monte Carlo (MC) simulations are considered the gold standard due to their high accuracy in modeling particle interactions and energy deposition. However, the computational burden of MC simulations is substantial, often rendering them impractical for real-time clinical applications. [16]

To address this challenge, approaches that combine MC simulations with deep learning (DL) models have gained considerable attention. These methods aim to maintain the high accuracy of MC simulations while leveraging the speed of DL models to predict dose distributions, thereby reducing overall computation times. Various architectures, such as convolutional neural networks (CNNs) and U-Nets, have been explored for this purpose.[6] While these models have shown promise, they also present challenges related to generalization across different energy spectra, anatomical regions, and clinical setups. [27]

This section reviews the key advancements in this area, focusing on the integration of deep learning with MC methods for dose calculation. Particular attention is given to a previous master thesis from this research group that laid the foundation for hybrid dose calculation models. The discussion then transitions to recent advancements that address the limitations identified in that work, including the use of latent diffusion models, extended energy ranges, and enhanced integration techniques.

3.1 Deep Learning in Radiotherapy

Deep learning has been increasingly applied in radiotherapy to improve the efficiency and accuracy of dose calculations and predictions.[28] Three significant approaches in this field are the *DeepDose* framework (Kontaxis et al., 2020)[5], the *DiffDP* model (Feng, et al., 2023)[23], and the *DoseDiff* model (Zhang et al., 2024)[4]. This section discusses these three models in detail, highlighting their methodologies, strengths, and how they differ from the approach presented in this thesis.

3.1.1 DeepDose: A Deep Learning Framework for Fast Dose Calculation

DeepDose is designed to serve as a fast and accurate dose calculation engine by leveraging deep learning. Traditional MC simulations are highly accurate for dose calculations in radiotherapy but are computationally expensive and time-consuming, especially in scenarios that demand rapid adaptations, such as MRI-guided radiotherapy. DeepDose addresses these limitations by using a deep convolutional network to predict dose distributions directly based on patient anatomy and specific machine parameters (e.g., multi-leaf-collimator (MLC) shapes).[5]

The key features of DeepDose include:

- **Physics-Based Inputs:** DeepDose introduces a novel set of inputs that encode the linear accelerator (linac) machine parameters and integrate them with the patient anatomy. This method creates a robust mapping that deep learning models can utilize for dose prediction.[5]
- **Training with Clinical Data:** The framework was trained on data from 101 prostate patients, using clinically relevant grid spacing and segmentation shapes to predict dose distributions with high accuracy. The ground-truth dose distributions were generated using MC simulations with a 1% statistical uncertainty, ensuring a high standard of accuracy for training.[5]
- **Accuracy and Efficiency:** DeepDose achieved a gamma pass rate of 99.9% $\pm$ 0.3% (3%/3 mm) for dose prediction on clinical plans, demonstrating its reliability. The inference time per patient plan is approximately 1 minute, making it highly suitable for online adaptive workflows where rapid recalculations are necessary.[5]

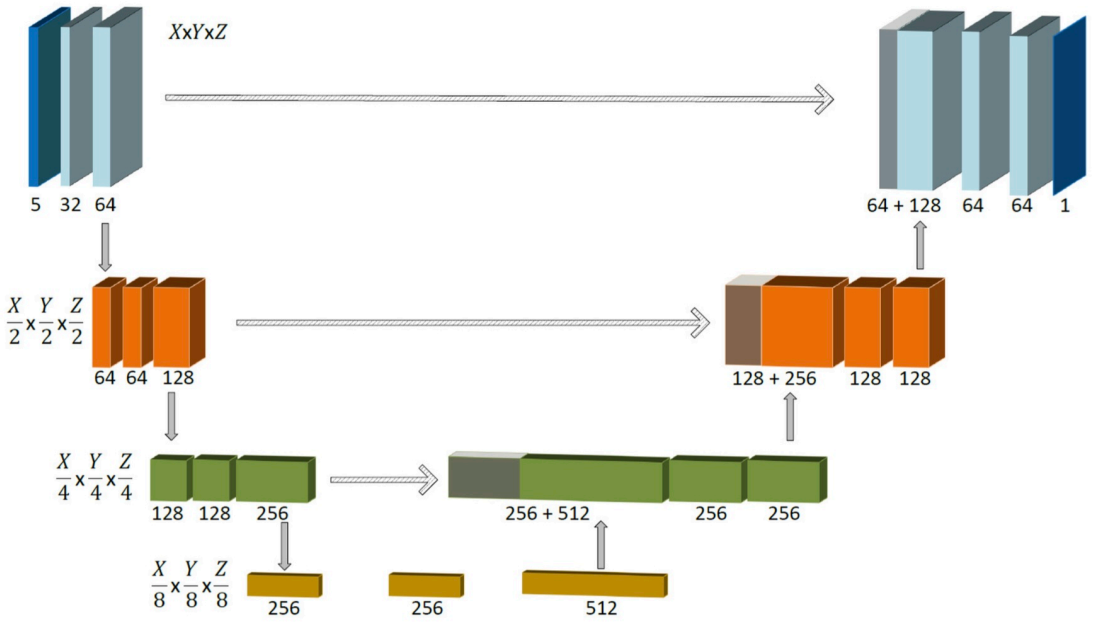
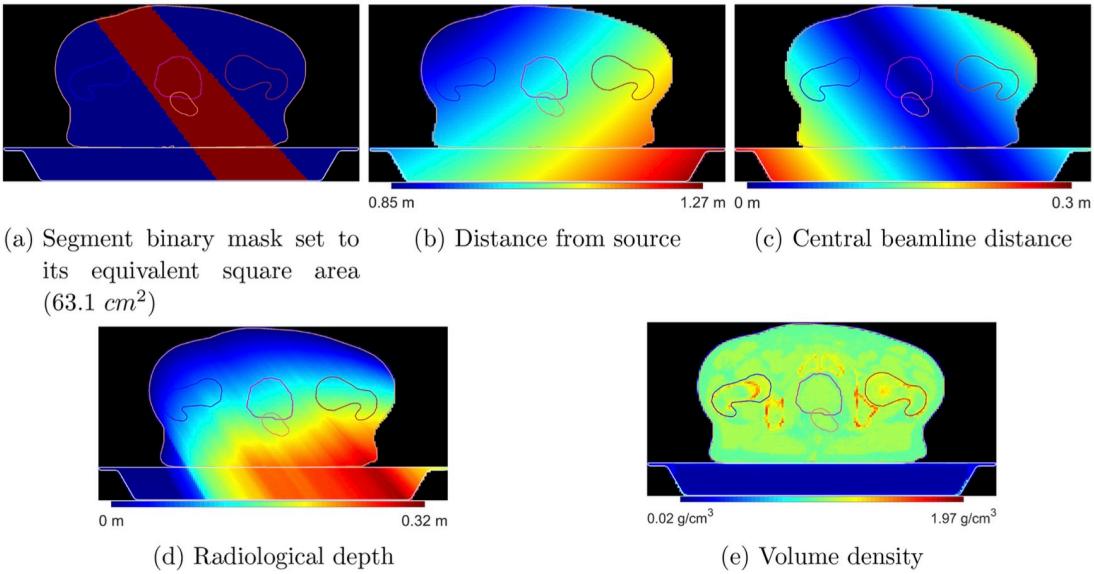


Figure 3.1: 3D U-Net Architecture [5]



(a) Segment binary mask set to its equivalent square area (63.1 cm^2)

(b) Distance from source

(c) Central beamline distance

(d) Radiological depth

(e) Volume density

Figure 3.2: 3D U-Net Inputs [5]

3.1.2 Use of U-Net in DeepDose

The foundational architecture for the DeepDose framework is inspired by the U-Net model (Ronneberger et al., 2015)[29], which was originally designed for biomedical image segmentation. In the context of DeepDose, the U-Net architecture is adapted to focus on the problem of dose distribution prediction by incorporating physics-informed features and adjusting the output to suit radiotherapy dose maps rather than segmentation masks. The U-Net's contracting and expansive paths (Figure 3.1) help capture both high-level features and fine-grained details, which are crucial for accurate dose prediction across varying anatomical structures and MLC configurations.[5]

3.1.3 Differences Between the Proposed Approach and DeepDose

While the DeepDose framework offers a significant advancement in fast dose calculations using deep learning, the approach of this thesis differs in several key areas to address the limitations and challenges identified in their work:

1. **Model Architecture and Input Representation:** Unlike DeepDose, which uses a standard deep convolutional network with physics-based input generation, the approach of this thesis leverages latent diffusion models (LDMs) to better capture complex spatial correlations and variations across different energy levels. Additionally, the method employs Gaussian quadrature integration to optimize the energy spectrum's representation. This approach focuses on extracting and emphasizing specific energy peaks that are critical for accurate dose modeling in complex clinical scenarios.

2. **Energy Range and Clinical Application Focus:** DeepDose mainly concentrates on predicting dose distributions for a specific treatment setup (prostate IMRT). In contrast, this work expands the application scope to more challenging scenarios involving varying anatomical regions and energy ranges. For example, the focus on lower energy spectra up to 50 keV, which are more relevant in treatments where precision near tissue interfaces and within high-gradient regions is crucial.

3. **Dynamic Adjustment of Model Parameters:** This approach allows for dynamic adjustment of parameters like the used energy spectrum and irra-

diation setups based on specific clinical needs. This flexibility addresses one of the key limitations of DeepDose, which operates on predefined setups and fixed input parameters.

3.1.4 Conclusion on DeepDose

DeepDose represents a significant step forward in the application of deep learning for dose calculation in radiotherapy, providing a fast and accurate alternative to traditional methods. However, by focusing on more generalized and flexible models, such as LDMs, this work aims to overcome some of the limitations of DeepDose. This paves the way for more adaptive and accurate dose calculations across a broader range of clinical scenarios.

3.1.5 DiffDP: Radiotherapy Dose Prediction via a Diffusion Model

The *DiffDP* model introduces a novel diffusion-based approach to address the limitations of existing deep learning models in dose prediction for radiotherapy. Traditional deep learning models like U-Nets and GANs often suffer from over-smoothing due to the use of L_1 or L_2 loss functions, which lead to blurred dose distribution predictions. The DiffDP model tackles this issue by leveraging a diffusion model that consists of a forward and reverse process, allowing for sharper and more detailed dose predictions by avoiding the averaging effects of conventional loss functions.[23]

The key features of DiffDP include:

- **Forward and Reverse Diffusion Processes:** In the forward process, the DiffDP model gradually adds Gaussian noise to the dose distribution map, effectively transforming it into a noise distribution. During the reverse process, the model removes this noise step-by-step using a well-trained noise predictor, eventually reconstructing the dose distribution map with higher fidelity and sharpness.[23]

- **Structure Encoder:** To better guide the noise prediction, DiffDP introduces a deep learning-based structure encoder that extracts anatomical information from CT images and segmentation masks of critical organs (such as PTV and

OARs). This allows the model to be more aware of dose constraints and anatomical relationships, leading to more accurate dose distributions.[23]

- **Performance Evaluation:** Extensive experiments conducted on an in-house dataset with 130 rectum cancer patients demonstrated that DiffDP outperforms existing state-of-the-art methods in terms of various dose prediction metrics, achieving more accurate and clinically acceptable dose maps.[23]

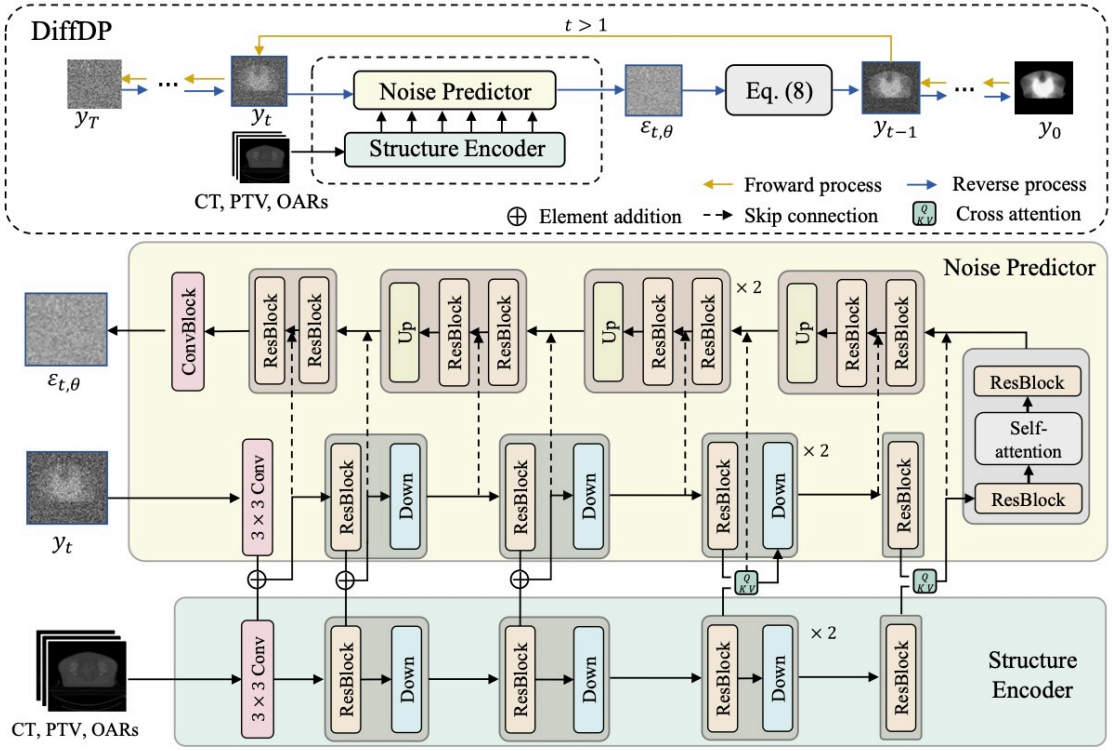


Figure 3.3: DiffDP Workflow - An illustration showing the forward and reverse processes in the DiffDP model, highlighting the noise addition and removal steps guided by anatomical information. [23]

3.1.6 Differences Between the Proposed Approach and DiffDP

While the DiffDP model innovatively uses a diffusion-based approach to improve dose prediction accuracy by addressing the over-smoothing problem, the approach of this thesis diverges in several key aspects:

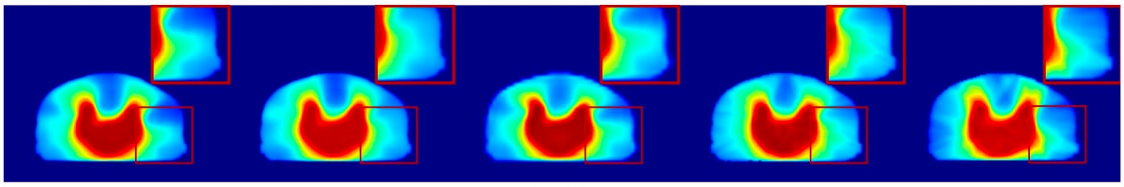


Figure 3.4: Dose Distribution Comparisons - A visual comparison of dose distribution maps predicted by DiffDP versus other methods, demonstrating DiffDP's ability to retain high-frequency details. [23]

1. **Dimensionality of Data:** One notable difference is that DiffDP operates on 2D data (CT slices and segmentation masks). In contrast, this approach utilizes 3D data representations, ensuring that the model accounts for volumetric anatomical variations and interactions in all three dimensions.

2. **Integration of LDMs:** Unlike DiffDP, which employs a traditional diffusion model framework, this approach utilizes LDMs.

3. **Advanced Energy Spectrum Optimization:** While DiffDP focuses on incorporating anatomical information through a structure encoder, this method optimizes the energy spectrum representation using Gaussian quadrature integration.

3.1.7 Conclusion on DiffDP

The DiffDP model contributes significantly to improving dose prediction in radiotherapy by utilizing a diffusion model to overcome the over-smoothing issues seen in previous deep learning approaches. However, by leveraging latent diffusion models, which perform the diffusion process in a lower-dimensional latent space, this approach significantly reduces computational complexity and enhances efficiency, particularly with high-resolution data like medical images.[25]

3.1.8 DoseDiff: A Distance-aware Diffusion Model for Dose Prediction

DoseDiff introduces a novel approach to dose prediction using a distance-aware conditional diffusion model. Unlike traditional methods, which rely on CT images and region-of-interest (ROI) masks that do not fully capture the spatial relationships

between tissues, DoseDiff uses Signed Distance Maps (SDMs) to provide a more informative input for deep learning models. This allows for a more precise prediction of dose distributions, particularly in complex anatomical structures.[4]

The key features of DoseDiff include:

- **Signed Distance Maps (SDMs):** SDMs are generated by performing a distance transform on ROI masks. These maps encode the distance of each voxel from the ROI contours in 3D space, providing valuable distance information that helps the model better predict dose distributions.[4]
- **Multi-encoder and Multi-scale Fusion Network (MMFNet):** To effectively fuse information from CT images and SDMs, DoseDiff employs MMFNet. This network architecture includes multiple encoders for different input types and utilizes a fusion module based on the self-attention mechanism to integrate global information from the inputs, enhancing the prediction accuracy.[4]
- **Conditional Diffusion Model:** The core of DoseDiff is a conditional diffusion model, which defines dose prediction as a sequence of denoising steps. This approach helps maintain the distribution characteristics of dose paths, resulting in more realistic dose distribution maps compared to conventional models.[4]
- **Denoising Diffusion Implicit Model (DDIM):** To reduce computational time, DoseDiff integrates the accelerated generation technology of DDIM, allowing for fewer sampling steps during the reverse diffusion process without compromising performance.[4]

3.1.9 Differences Between the Proposed Approach and DoseDiff

While DoseDiff provides an innovative solution for dose prediction in radiotherapy, this approach diverges from their method in several significant ways:

1. **Model Complexity and Flexibility:** While DoseDiff employs a conditional diffusion model, this work builds on conditional latent diffusion models (LDMs).

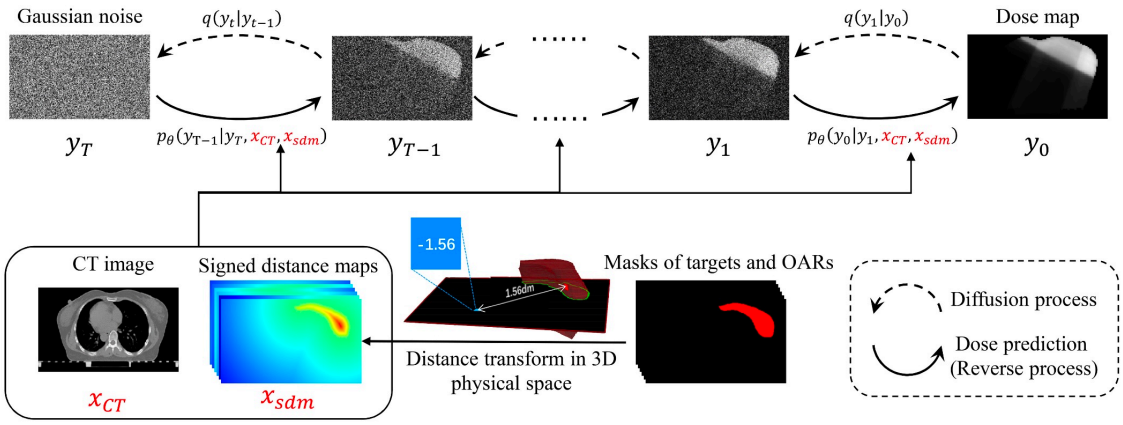


Figure 3.5: DoseDiff Workflow - An illustration showing the workflow of DoseDiff, from input generation (CT images and SDMs) to dose prediction via the diffusion model. [4]

2. **Energy Spectrum Representation:** This method incorporates a detailed Gaussian quadrature integration for optimizing energy spectra. This approach focuses on energy peaks that are essential for accurate dose modeling, which DoseDiff does not explicitly address.

3. **Adaptive Parameter Control:** Unlike DoseDiff, which uses fixed input parameters, this framework allows for dynamic adjustment of parameters based on specific clinical scenarios. This adaptability is crucial for achieving more precise dose predictions tailored to individual patient cases.

3.1.10 Conclusion on DoseDiff

The DoseDiff model represents an important advancement in dose prediction for radiotherapy by introducing distance-aware diffusion models and advanced feature fusion techniques. However, the approach of this thesis builds upon these advancements by integrating more flexible and adaptive models, such as latent diffusion models.

3.2 Previous Work in the Research Group: Hybrid Monte Carlo Algorithm for Dose Calculation in Radiotherapy

In a previous master thesis conducted within our research group by Helen Amann (2022), a hybrid Monte Carlo (MC) algorithm was developed to address the significant computational challenges associated with dose calculations in radiotherapy. The work focused on combining the high accuracy of MC simulations with the speed of deep learning (DL) models, specifically using U-Nets. Amann's algorithm aimed to replace the computationally intensive low-energy portion of MC simulations (10 keV to 150 keV) with predictions from a pre-trained U-Net model. The U-Net received patient CT images as input and produced dose distributions for monoenergetic photon beams in simple geometries. [6]

The hybrid approach demonstrated high accuracy in setups involving monoenergetic beams, showing promising results in terms of both precision and reduced computation time. However, limitations were observed in more complex configurations, such as isotropic sources and planar beams. In these scenarios, the predicted dose distributions tended to underestimate the total dose, primarily due to the U-Net's struggle to generalize across varying particle energy levels and anatomical regions. These limitations highlighted the challenges of using U-Nets, particularly when dealing with complex tissue heterogeneity and high energy gradients.[6]

3.2.1 Hybrid Algorithm Workflow and Integration Challenges

Figure 3.6 provides a clear visual representation of the hybrid algorithm workflow, which is central to understanding the approach taken in her work. The figure illustrates the sequential process: starting with MC simulations for higher-energy particles, followed by transitioning to U-Net predictions for low-energy dose distributions. The workflow highlights how the MC simulations handle the initial particle sampling and tracking until a certain cutoff energy is reached, after which the U-Net predictions take over. This approach aims to balance computation speed with accuracy by focusing DL efforts on the most computationally expensive portion of the MC process—the low-energy particles, which are responsible for the majority of dose calculations due to effects like Compton scattering.[6]

Despite the efficient transition depicted in the workflow, a key challenge lies in ensuring that the dose predictions made by the U-Nets are accurate across different anatomical regions and clinical setups. For example, the reliance on predefined energy cutoffs and static cube sizes introduces potential inaccuracies when dealing with more complex irradiation geometries, such as isotropic or planar sources. These limitations are crucial to understanding why more advanced methods, such as latent diffusion models (LDMs) and larger cube sizes, are explored in this thesis.[6]

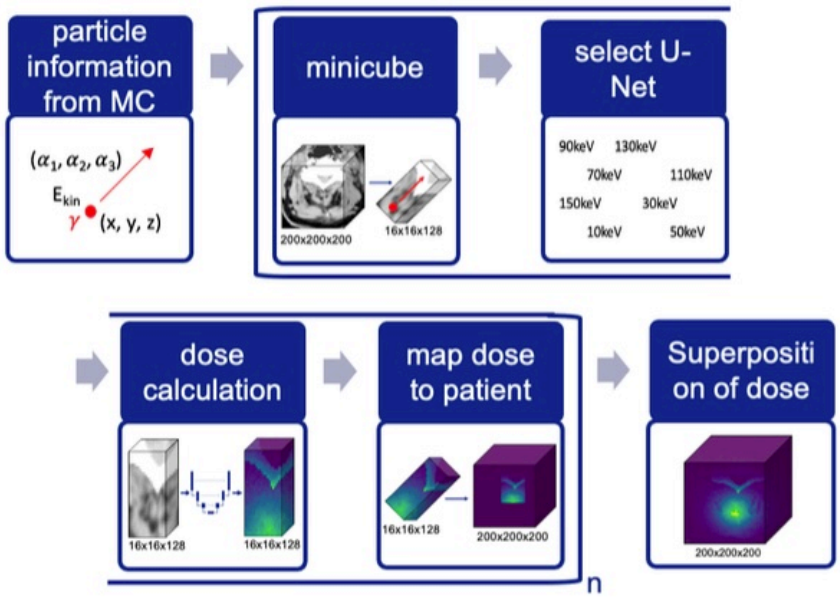


Figure 3.6: Workflow of the Hybrid Algorithm - This figure from Amann's thesis provides a clear visual representation of the hybrid approach, showing the interplay between MC simulations and U-Net dose predictions. [6]

3.2.2 Advancements and Differences in the Current Work

Building upon the foundation laid by Amann, this thesis introduces several significant advancements aimed at overcoming the limitations of the previous approach. The first key difference lies in the focus on low-energy spectra up to 50 keV. This shift is motivated by the need for enhanced accuracy in clinical scenarios involving more complex anatomical regions, such as head and neck squamous cell carcinoma cases. Unlike Amann's work, which primarily targeted higher energies (10 keV to 150 keV), this approach prioritizes lower-energy integration, which is critical for improving dose estimation near tissue interfaces and within high-gradient regions.

To achieve this, the U-Net architecture is replaced with latent diffusion models (LDMs). This approach addresses the key issue identified in Amann's thesis, where U-Nets struggled with underestimation due to their limited capacity to capture intricate spatial dependencies. Additionally, this thesis introduces Gaussian quadrature as the method for numerical integration, enhancing the accuracy of dose calculations, especially in the low-energy spectrum.

The thesis also explores the importance of cube sizes in dose prediction. Amann's work relied on fixed-size minicubes (16 mm $\times$ 16 mm $\times$ 128 mm), which resulted in abrupt discontinuities at the edges of the dose predictions. The approach of this thesis extends the cube size to better capture scattering effects.

Overall, while Amann's thesis made considerable strides in accelerating dose calculations through a hybrid deep learning and Monte Carlo approach, the present work builds on those advancements by addressing key limitations related to energy range, spatial generalization, and dose integration. These improvements pave the way for more accurate and clinically relevant dose predictions across a wider range of treatment scenarios.

4 Material and Methods

In this section, the methodology employed is outlined. The thesis focuses on integrating diffusion models with Gaussian quadrature techniques to enhance the efficiency and accuracy of dose calculations. The approach relies on training multiple diffusion models, each tailored to specific energy ranges, and utilizing Gaussian quadrature for numerical integration, thereby allowing rapid yet precise dose predictions.

The following subsections detail the procedures involved in data collection and preprocessing, model development and training, and the evaluation criteria used to validate the accuracy of the results. By systematically combining Monte Carlo simulations for ground truth generation with advanced machine learning techniques, the proposed method aims to address the computational challenges traditionally associated with dose calculation in radiotherapy.

4.1 Data Collection and Preprocessing

To effectively train a machine learning model, having access to a substantial dataset is critical. For this work, this requirement is fully met, as an extensive set of reference data can be produced using patient data combined with Monte Carlo (MC) simulations. Below, the key details regarding the dataset and the data generation process are presented.

4.1.1 CT Image Acquisition

The dataset used in this work consists of CT images obtained from an open-source collection focusing on head and neck squamous cell carcinoma (HNSCC) patients [30], made available by The Cancer Imaging Archive (TCIA) [31]. This dataset includes imaging, radiation therapy, and clinical data from 627 HNSCC patients treated at MD Anderson Cancer Center (MDACC). The data was gathered as part of two independent research projects, with a subset of 70 patients overlapping be-

tween both studies.

The first project, the "Head-Neck-CT-Atlas," focused on evaluating imaging data from patients who underwent curative-intent radiation therapy. The dataset includes de-identified diagnostic imaging, radiation treatment plans, and follow-up scans, which are subject- and date-matched with comprehensive clinical data. The second project, "Radiomics outcome prediction in Oropharyngeal cancer," aimed to integrate quantitative imaging biomarkers into risk stratification models. This project provides contrast-enhanced CT scans, manually segmented tumor volumes, and related clinical outcomes data.

For training the deep learning model, cubes of size $100 \times 100 \times 100$ with a resolution of 1 mm are randomly extracted from the patient CT scans. These cubes, which represent Hounsfield Unit (HU) values, are interpolated linearly to fill the voxel grid. To ensure that the cubes contain meaningful anatomical structures, cubes with an average HU below -500 are excluded, avoiding regions predominantly filled with air. These preprocessed cubes serve as input data for the model.

Access to the dataset requires signing a TCIA Restricted License Agreement due to the presence of imaging data that could potentially be used to reconstruct human faces. Detailed access instructions and links to the dataset are available on the TCIA website.

4.1.2 Data Splitting and Normalization

In this study, a subset of 100 patients from the HNSCC dataset was randomly selected for analysis. To ensure data consistency, only planning CT scans acquired using PHILIPS CT systems were included. Planning CT scans, typically used in radiation therapy planning, provide the precise anatomical details needed for accurate treatment simulations, making them highly suitable for this research. [32] The dataset was split into 90% training data and 10% test data, ensuring robust model evaluation. From each patient's CT scan, 200 voxel cubes of size $100 \times 100 \times 100$ were randomly extracted, resulting in a total of 18,000 cubes for training and 2,000 cubes for testing.

Prior to feeding the small cubes containing HU values into the deep learning model,

they undergo post-processing to facilitate more effective learning. This involves clipping the HU values to a clinically relevant range, with the lower boundary set at -1000 HU (corresponding to air) and the upper boundary capped at 2100 HU (representing cortical bone). The cubes are then normalized to fall within the range of [-1, 1] to align with the model's input requirements.

4.1.3 CT Image Resampling and Downsampling

To prepare the dataset for training latent diffusion models in radiotherapy dose calculation, the original CT images underwent a resampling process followed by downsampling. The original CT images had dimensions of $100 \times 100 \times 100$ voxels, each representing a cubic volume of 1 mm^3, covering a physical space of 100 mm $\times$ 100 mm $\times$ 100 mm.

Resampling Process

In the first step, the original images were resampled to a resolution of $128 \times 128 \times 128$ voxels while preserving the original physical size (100 mm $\times$ 100 mm $\times$ 100 mm). This resampling increased the number of voxels, resulting in a refined voxel resolution of approximately 0.78 mm per voxel (100 mm / 128 voxels). The purpose of this resampling is to allow for easier downsampling.

The resampling was performed using B-spline interpolation of order 5. B-spline interpolation is a form of spline interpolation that uses basis splines (B-splines) to fit a smooth curve through the data points. The order of the spline determines the degree of the polynomials used in the interpolation. Specifically:

- **Order 0:** Corresponds to nearest-neighbor interpolation.
- **Order 1:** Corresponds to linear interpolation.
- **Order 2:** Corresponds to quadratic interpolation.
- **Order 3:** Corresponds to cubic interpolation (commonly used for smooth interpolation).
- **Order 5:** Corresponds to quintic interpolation (fifth-degree polynomials).

In this study, a B-spline of order 5 was chosen because it strikes a balance between smoothness and precision, making it well-suited for medical imaging applications. Higher-order splines like quintic splines (order 5) are capable of preserving subtle gradients and anatomical structures, which is particularly important for CT images where maintaining fidelity to the original tissue boundaries and internal structures is critical. [33] The interpolation was implemented using the `scipy.ndimage.zoom` function, specifying `order=5` to perform the resampling. [34]

Downsampling Process

Following resampling to $128 \times 128 \times 128$ voxels, the images were progressively downsampled to resolutions of $64 \times 64 \times 64$, $32 \times 32 \times 32$, and $16 \times 16 \times 16$ voxels. The downsampling was accomplished using average pooling, a process in which non-overlapping blocks of size $2 \times 2 \times 2$ are averaged to produce a single voxel value for the lower-resolution image. This method is computationally efficient and preserves the key spatial features and intensity patterns while reducing the resolution.

The use of average pooling is particularly effective in this context because it avoids introducing potential artifacts that more complex interpolation methods, such as B-splines, could create when applied for downsampling. By downsampling through successive averaging, the key structural information is retained, ensuring that the latent diffusion models receive meaningful input data across all resolution levels. [35]

4.2 Radiation Source

The energy spectrum used in this study plays a crucial role in accurately modeling dose distributions. The selection of energy levels is based on their relevance to clinical radiotherapy scenarios, ensuring that the simulated conditions align with realistic treatment settings.

The energy spectrum used in this work is derived from the catalogue provided in the publication Catalogue of X-ray Photon Fluence Spectra of Electronic Brachytherapy Device INTRABEAM Manufactured by ZEISS [36]. The catalogue was compiled using data available from literature as well as measurements and simulations conducted by the European Metrology Programme for Innovation and Research (EM-

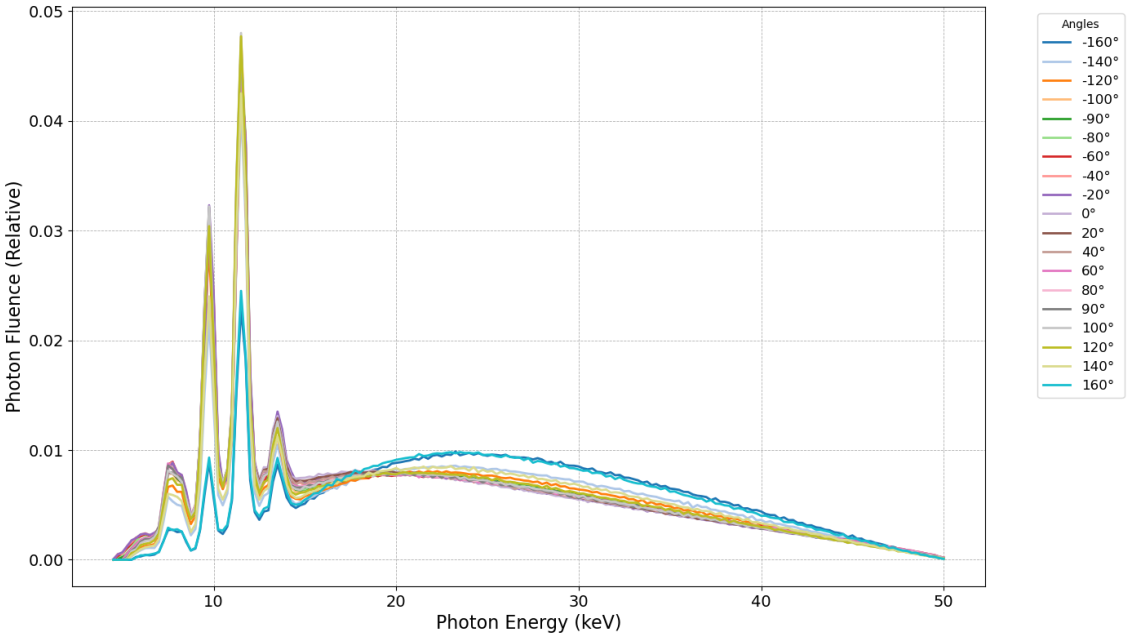


Figure 4.1: Photon Energy Spectrum for Various Angles

PIR) project "18NRM02 PRISM-eBT." [37] The primary goal of the catalogue is to establish harmonized reference X-ray qualities termed "eBT-equivalent" spectra, which can be replicated in laboratory settings using X-ray tubes calibrated to simulate the energy spectrum of electronic brachytherapy (eBT) devices as closely as possible.

The catalogue contains various spectra, including those derived from Monte Carlo simulations and measured data under different conditions. For the purposes of this work, the measured spectra in air at a distance of 70 cm with a bare needle for various angles were chosen. These spectra correspond to a nominal bias voltage of 50 kV and were considered suitable for our study based on the consistent behavior across different angles.

Although the spectra are available for multiple angles, it was decided to focus on the spectrum generated at a 20-degree angle. This choice was made primarily for simplicity, as the spectrum does not exhibit significant angle dependence, and focusing on a single angle is sufficient for the purposes of this study.

Figure 4.1 illustrates the energy spectrum across different angles, with photon en-

ergy (in keV) plotted against the relative photon fluence. The selected spectrum at 20 degrees will be used in subsequent simulations and dose calculations.

4.3 Monte Carlo Simulation for Ground Truth Generation

HU Interval	Material
< -800	G4 AIR
[-800, -145]	G4 LUNG ICRP
[-145, -60]	G4 ADIPOSE TISSUE ICRP
[-60, 0]	G4 WATER
[0, 60]	G4 MUSCLE WITH SUCROSE
[60, 1150]	G4 B-100 BONE
[1150, 2100]	G4 BONE COMPACT ICRU
2100 <	G4 BONE CORTICAL ICRP

Table 4.1: Conversion of Hounsfield Unit (HU) ranges to Geant4 standard materials used in simulations. Each HU interval is mapped to a specific material based on its corresponding physical properties.

To generate ground truth data for the model, Monte Carlo (MC) simulations are conducted on small HU cubes. For this, the Geant4 simulation toolkit is employed. [38] Initially, HU values are mapped to corresponding material categories based on predefined Geant4 material ranges. The conversion process involves eight specific materials, as shown in Table 4.1. Geant4 simulations are then executed on these cubes filled with the mapped material values, incorporating relative electron density in the calculations. The simulation setup involves firing particles from a point source with a fixed monoenergetic energy and momentum in the positive z-direction. The resulting output is a dose distribution with each voxel representing the deposited dose, which subsequently serves as the ground truth data for the deep learning model.

4.3.1 Determining the Number of Particles

In radiotherapy dose calculations using MC simulations, the accuracy of the simulated dose distribution is crucial. One key parameter influencing this accuracy is the number of particles (e.g., photons) used in the simulation. Selecting an appropriate

number of particles ensures that the dose distributions are statistically reliable while avoiding unnecessary computational cost.

Statistical Considerations for Voxel Dose Accuracy

The dose in each voxel is subject to statistical fluctuations. A common approach is to aim for a relative standard deviation of the dose, or the so-called statistical uncertainty, to be within 1% of the mean dose value in voxels where the dose is clinically relevant. For example, in regions receiving therapeutic doses (around 50 Gy), a $\pm 1\%$ uncertainty is considered reasonable. This corresponds to ensuring that the standard deviation is about 0.5 Gy, which is a level of uncertainty that is negligible in a clinical context.

For regions where the dose is low, achieving a similar level of accuracy would require a vastly larger number of particles, which is computationally inefficient and unnecessary for practical purposes. Therefore, a cutoff criterion is typically applied: for voxels receiving less than 2% of the maximum dose (approximately 1 Gy in a typical therapeutic scenario), higher uncertainties are acceptable, and these regions are often excluded from the stringent accuracy requirement.

A Rational Strategy for Setting the Particle Count

Given the above considerations, a rational strategy involves focusing on voxels with doses above 2 Gy and imposing a maximum relative uncertainty of 1% in these regions. To determine the required number of particles, the following steps can be taken:

1. **Initial Simulation:** Conduct an initial simulation with a moderate number of particles, typically around 10 million. This provides an estimate of the dose distribution and the associated statistical uncertainty in each voxel.

2. **Uncertainty Analysis:** For each voxel with doses above 2 Gy, calculate the relative standard deviation of the dose. This value serves as a measure of the current statistical uncertainty.

3. **Scaling the Particle Number:** The statistical uncertainty scales inversely with the square root of the number of particles (i.e., the variance decreases as 1/N, where N is the number of particles). Using this relationship, the number

of additional particles needed to achieve the target 1% uncertainty can be determined. For example, if the current uncertainty is 2%, the number of particles needs to be quadrupled to reduce the uncertainty to 1%.

This process allows for a systematic determination of the particle count based on the statistical needs of the simulation, ensuring that the final results are both accurate and computationally efficient.

4.4 Spectrum Splitting and Gaussian Quadrature Integration

In this section, the methodology for spectrum splitting and Gaussian quadrature integration is presented. This approach enables the training of latent diffusion models (LDMs) tailored to specific energy ranges, enhancing the accuracy and efficiency of dose calculations. The overall spectrum, as shown in Figure 4.1, is divided into distinct components, focusing on the sharp peaks around 10 keV and the smoother, broader energy regions. The division and subsequent model training are carried out as follows:

4.4.1 Spectrum Splitting and Peak Extraction

The photon energy spectrum used in this study is characterized by prominent peaks around 10 keV and a smoother tail extending toward higher energies. These peaks are critical for accurately modeling dose distributions in specific clinical scenarios, as they represent key energy levels for therapeutic radiation. To capture this, the spectrum is divided into two components:

1. **Sharp Peaks Around 10 keV**: These peaks are directly extracted, representing four distinct energy points (approximately 7.5, 9.75, 11.5, and 13.5 keV) where the LDMs are specifically trained.

2. **Smooth Tail Region**: The remainder of the spectrum, which exhibits a smoother distribution, is approximated using Gaussian quadrature.

Those two components are visualized in Figure 4.2.

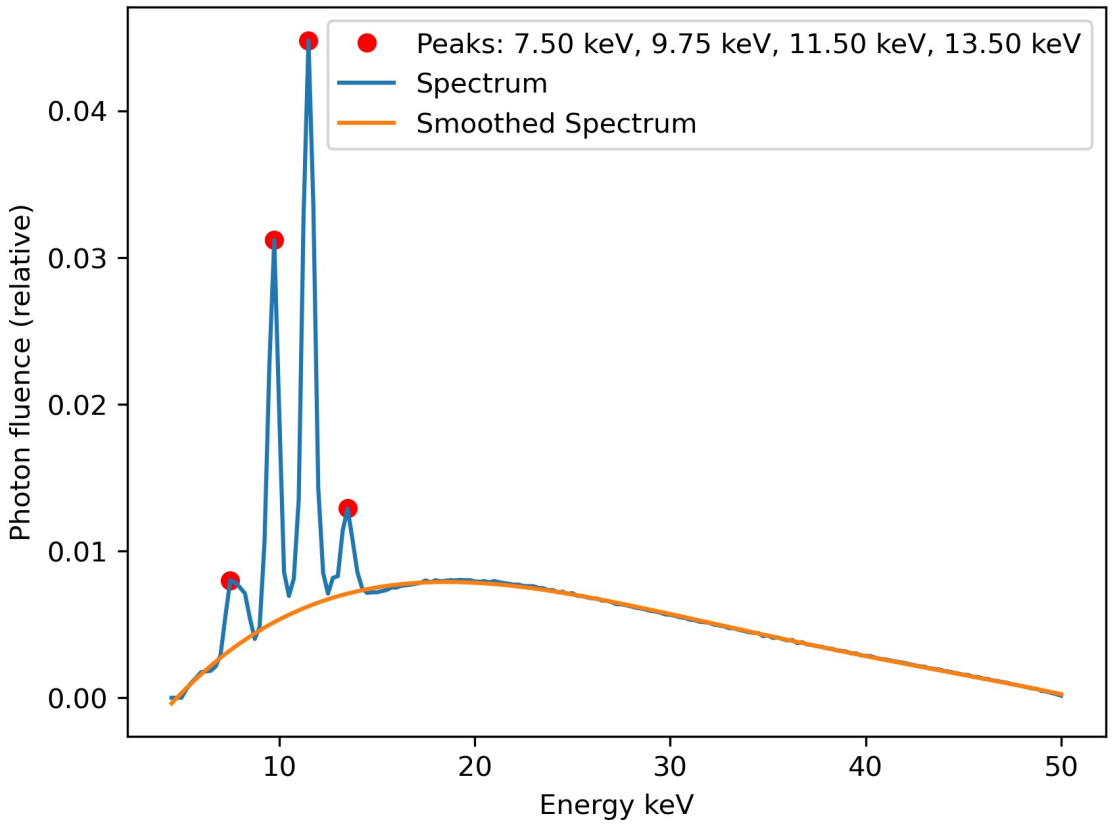


Figure 4.2: Photon Fluence Spectrum with Smoothed Curve and Identified Peaks. The plot shows the photon fluence spectrum as a function of energy (in keV), represented by the raw data points (blue curve). A polynomial fit (degree 4) is applied to smooth the spectrum and is shown as the orange curve. Peaks are marked with red circles, and their corresponding energy values (in keV) are indicated in the legend. The smoothing process excludes data points around the detected peaks to better represent the underlying continuum.

4.4.2 Gaussian Quadrature for Energy Integration

Gaussian quadrature is employed to determine the optimal energy points for training additional LDMs within the smooth part of the spectrum. The quadrature method provides a systematic way to approximate the integral over this energy range, ensuring that key energy points are adequately represented without requiring a dense sampling of the entire spectrum.

Depending on the smoothness and complexity of the spectrum, different quadrature orders (e.g., 2, 4, 8) are tested to evaluate their ability to approximate the

overall dose distribution accurately. The quadrature order is determined based on the following criteria:

- **Accuracy of Dose Distribution**: The dose distributions calculated using the peaks and quadrature points are compared to those derived from the full spectrum using Monte Carlo (MC) simulations. The comparison is quantified using the gamma index.
- **Computational Efficiency**: The computational time required for training LDMs at different quadrature orders is assessed, with a focus on identifying the lowest order that yields an acceptable approximation.

For each quadrature point, an LDM is trained to predict dose distributions. The integration results are then combined with those from the peaks around 10 keV, allowing for a comprehensive representation of the original spectrum.

Determining the Optimal Quadrature Order

The optimal quadrature order is identified through a series of tests comparing dose distributions generated using different quadrature orders. These tests involve:

1. **Simulating Dose Distributions Using MC**: The full spectrum is used to generate a reference dose distribution through MC simulations.
2. **Comparing Quadrature Approximations**: For each quadrature point and each of the four peaks, monoenergetic MC simulations are used to generate individual dose distributions. A final dose distribution is then created by superimposing these individual distributions, weighted according to their respective quadrature weights and the amplitudes of the peaks. To ensure a meaningful comparison with the reference dose distribution, the combined dose distribution is rescaled to match the total deposited dose of the reference. The accuracy of this comparison is evaluated using the gamma index with a stringent pass criterion of 1%/1mm, reflecting both spatial and dosimetric agreement.
3. **Selecting the Best Order**: The order that provides the best balance between computational efficiency and dose approximation accuracy is selected. For instance, if a fourth-order quadrature yields satisfactory results with minimal computational overhead, it is chosen for subsequent LDM training.

This approach enables the development of a robust and computationally efficient dose prediction framework that leverages both the precision of Gaussian quadrature for smooth spectrum regions and the specificity of LDMs trained on prominent energy peaks.

4.4.3 Gauss-Legendre Quadrature Points for Energy Integration

In order to accurately integrate the energy spectrum for dose calculations, the Gauss-Legendre quadrature method is employed. This method allows for efficient numerical integration by approximating the integral of a function as a weighted sum of the function's values at specific points, known as quadrature points. The Gauss-Legendre quadrature is particularly well-suited for integrals over a finite interval, in this case, the energy range from 0 to 50 keV, which is critical for modeling low-energy photon interactions in radiotherapy.

Determination of Gauss-Legendre Quadrature Points

For the interval $[0, 50]$ keV, the Gauss-Legendre quadrature points and their corresponding weights have to be determined to perform accurate energy integration. The Gauss-Legendre quadrature formula for an integral over the interval $[-1, 1]$ is given by:

$$\int_{-1}^{1} f(x)\,dx \approx \sum_{i=1}^{n} w_i f(x_i), \tag{4.1}$$

where x_i are the quadrature points (roots of the Legendre polynomial $P_n(x)$) and w_i are the corresponding weights. To adapt this for an arbitrary interval $[a, b]$, a change of variables is used:

$$\int_{a}^{b} f(x)\,dx \approx \frac{b-a}{2} \sum_{i=1}^{n} w_i f\left(\frac{b-a}{2} x_i + \frac{a+b}{2}\right). \tag{4.2}$$

For the interval $[0, 50]$ keV, the standard interval $[-1, 1]$ is mapped to $[0, 50]$ and the transformed quadrature points and weights are computed.

Quadrature Points and Weights for the Interval $[0, 50]$ keV

Using Gauss-Legendre quadrature for different orders (e.g., 2, 4, 8), the quadrature points x_i and weights w_i for the interval $[0, 50]$ keV are calculated in table 4.2. These

Quadrature Type	Point	Energy (keV)	Weight
2-Point	x_1	10.5662	25.0
	x_2	39.4338	25.0
4-Point	x_1	3.4716	8.6964
	x_2	15.7505	16.3036
	x_3	34.2495	16.3036
	x_4	46.5284	8.6964
8-Point	x_1	0.9928	2.5307
	x_2	5.0833	5.5595
	x_3	11.8617	7.8427
	x_4	20.4141	9.0671
	x_5	29.5859	9.0671
	x_6	38.1383	7.8427
	x_7	44.9167	5.5595
	x_8	49.0072	2.5307

Table 4.2: Quadrature Points and Weights

quadrature points and weights are used to accurately integrate the energy spectrum, ensuring precise modeling of dose deposition in Monte Carlo simulations. [39]

4.5 Model Architecture

The model utilized in this study is a 3D Latent Diffusion Model (LDM) implemented within the Monai framework[40], specifically designed for high-resolution 3D medical image generation.[25], [41] As illustrated in Figure 4.3, the Latent Diffusion Model (LDM) architecture comprises three main components: an encoder ($\mathcal{E}$), a diffusion process operating within the latent space, and a decoder ($\mathcal{D}$). The full-size model operates in the pixel space with dimensions of $128 \times 128 \times 128$. The encoder compresses these high-dimensional input images into compact latent representations, which are then processed by the diffusion model. The decoder reconstructs the refined latent representations back into the original high-dimensional space.

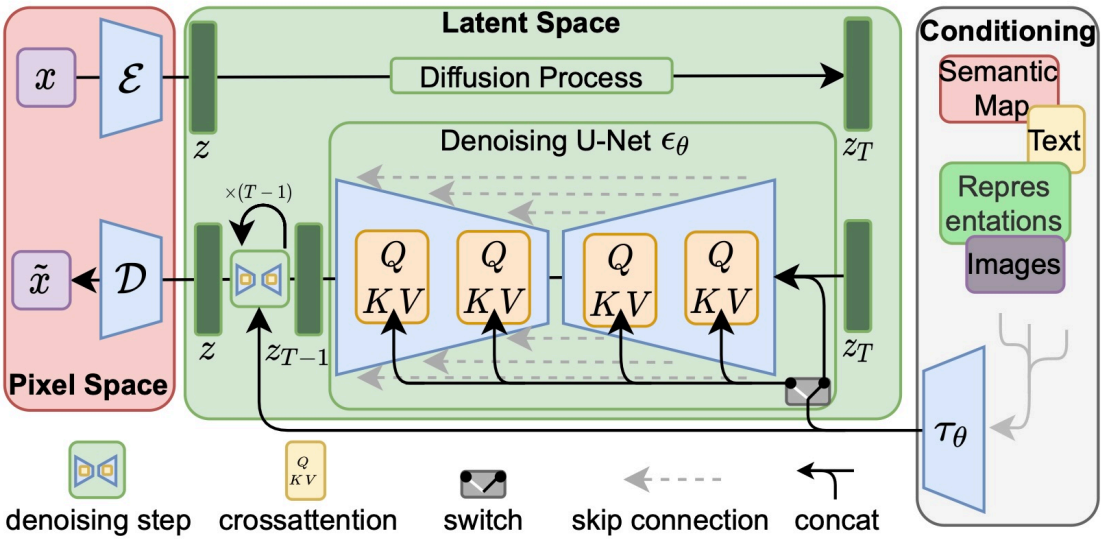


Figure 4.3: Overview of the 3D Latent Diffusion Model architecture.[25]

4.5.1 Autoencoder with KL-Regularization

The encoder component of the LDM is an autoencoder designed with KL- regularization. The primary role of this autoencoder is to transform the input images into a lower-dimensional latent representation that the diffusion model will subsequently learn. This transformation is crucial for scaling the model to handle high-resolution medical images efficiently. The compression achieved by the autoencoder reduces the computational resources required for the diffusion component, making the approach feasible for learning complex 3D structures in medical imaging.

The autoencoder is trained using a combination of L1 loss, perceptual loss[42], a patch-based adversarial objective[43], and KL- regularization of the latent space. The KL-regularization helps in maintaining a structured and continuous latent space, which is essential for effective sampling and generation by the diffusion model. The encoder maps the input images to a latent representation of size $16 \times 16 \times 16$, significantly compressing the high-dimensional data while retaining critical information for accurate reconstructions.[41]

4.5.2 Latent Diffusion Process

The core of the LDM is the diffusion process, which involves forward and reverse diffusion steps operating within the latent space. Diffusion models are generative models that convert Gaussian noise into samples from a learned data distribution via an iterative denoising process. In the forward diffusion process, Gaussian noise is incrementally added to the latent representation across 1000 steps, progressively degrading its structure through a fixed Markov chain with a linear variance schedule. The reverse diffusion process, parameterized by a denoising U-Net (ϵ_θ), learns to recover the original data from the noisy latent representations by iteratively removing the noise, guided by a learned sequence of denoising steps.

4.5.3 Conditioning with CT Images

To condition the LDM using CT images, a hybrid approach combining concatenation and cross-attention mechanisms is employed, as suggested in the paper [25]. This method enhances the model's ability to integrate and utilize the conditioning information provided by the CT images throughout the generative process.

Concatenation: The CT images are first encoded into latent space representations, which are then concatenated with the input data at each step of the diffusion process. This approach effectively incorporates the spatial information of the CT images directly into the model's processing pipeline, allowing the diffusion model to generate outputs that align closely with the anatomical structures present in the CT images [25].

Cross-Attention Mechanisms: For more complex conditioning, cross-attention layers are used within the U-Net backbone of the diffusion model. These mechanisms dynamically adjust the model's focus on specific parts of the latent representation according to the features extracted from the CT images. The cross-attention layers receive inputs from both the latent space and the encoded CT features, enabling the model to learn nuanced correlations between the anatomical information and the desired output distribution [25]. By combining these two conditioning methods, the LDM can effectively leverage the detailed anatomical data provided by CT images, resulting in more accurate and patient-specific output generations.

4.6 Evaluation Metrics

To quantitatively compare dose distributions in complex modulated radiotherapy, this study employs the Dose-Volume Histogram (DVH) and the gamma index as evaluation metrics.

4.6.1 Dose-Volume Histogram (DVH)

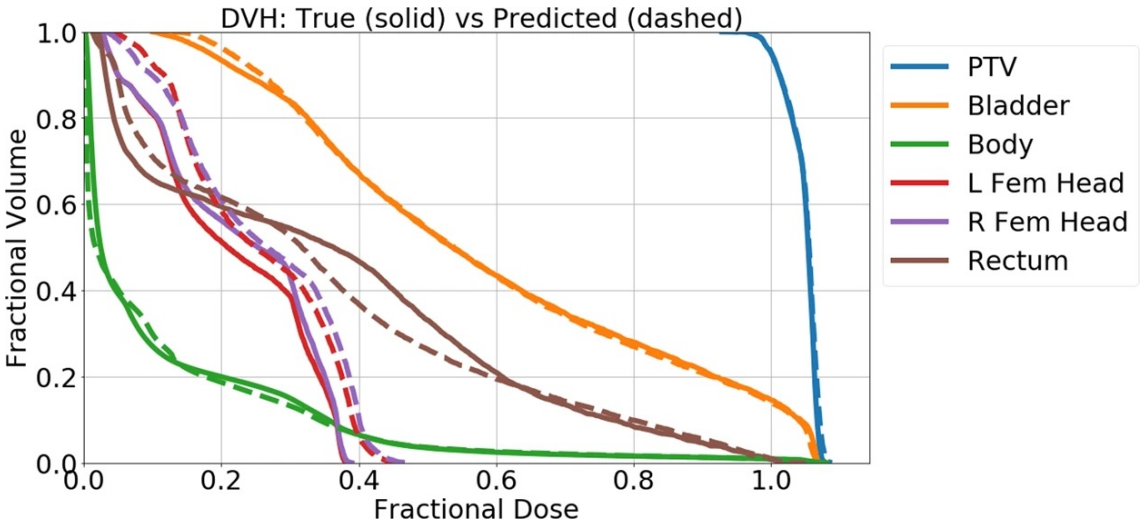


Figure 4.4: Example of a Dose-Volume Histogram (DVH) showing the fractional volume of various structures (PTV, bladder, body, left and right femoral heads, and rectum) receiving different fractional doses. Solid lines represent true dose distributions, while dashed lines indicate predicted distributions. The DVH is used to evaluate and compare the dose coverage of target volumes and organs at risk in radiotherapy treatment planning. [44]

The DVH is a critical tool for simplifying the evaluation of 3D dose distributions into a 2D plot. It provides a clear overview of the dose coverage of both the target volumes and the organs at risk. Specifically, cumulative DVHs are employed to indicate the volume percentage of a structure receiving a specific dose or higher, aiding in the comparison of different dose distributions. Figure 4.4 illustrates a typical cumulative DVH.

For DVH calculation, anatomical structures are defined, and the dose is computed within the volume grid. The grid elements, or voxels, are accumulated in corresponding dose bins for each structure, which then form the DVH. The DVH plot

shows dose in Gray (Gy) on the x-axis and the percentage of the volume receiving that dose on the y-axis. From these plots, metrics like D_x (minimum dose received by the hottest $x\%$ of the volume) and V_x (percentage of the volume receiving at least $x\%$ of the prescribed dose) are derived.

To evaluate and compare dose distributions, percentage differences between reference and evaluated dose distributions are computed using the following formulas:

$$D_x^* = \frac{D_{x_R} - D_{x_E}}{D_{x_E}} \times 100 \tag{4.3}$$

$$V_x^* = \frac{V_{x_R} - V_{x_E}}{V_{x_E}} \times 100 \tag{4.4}$$

Here, D_{x_R} and V_{x_R} represent the reference dose distributions, and D_{x_E} and V_{x_E} are the evaluated dose distributions. Ideally, both $D_x^* = 0$ and $V_x^* = 0$ indicate perfect agreement, while negative or positive values signal underestimation or overestimation of the evaluated dose. [45]

4.6.2 Gamma Index

The gamma index is a popular metric for validating dose distributions in radiotherapy. It combines the dose difference (DD) and distance-to-agreement (DTA) criteria to assess the spatial and dosimetric accuracy of dose distributions. The gamma index is calculated for each point in the evaluated dose distribution using:

$$\Gamma(r_R, r_E) = \sqrt{\frac{\Delta r^2(r_R, r_E)}{\delta r^2} + \frac{\Delta D^2(r_R, r_E)}{\delta D^2}} \tag{4.5}$$

where $\Delta r(r_R, r_E)$ is the spatial distance between the reference point r_R and the evaluated point r_E, and $\Delta D(r_R, r_E)$ represents the dose difference. δr and δD are predefined acceptance criteria for spatial and dose differences. The minimum gamma value for each point is determined by:

$$\gamma(r_R) = \min\{\Gamma(r_R, r_E)\} \text{ for all } r_E \tag{4.6}$$

A point is considered to pass if $\gamma < 1$, indicating that both spatial and dosimetric

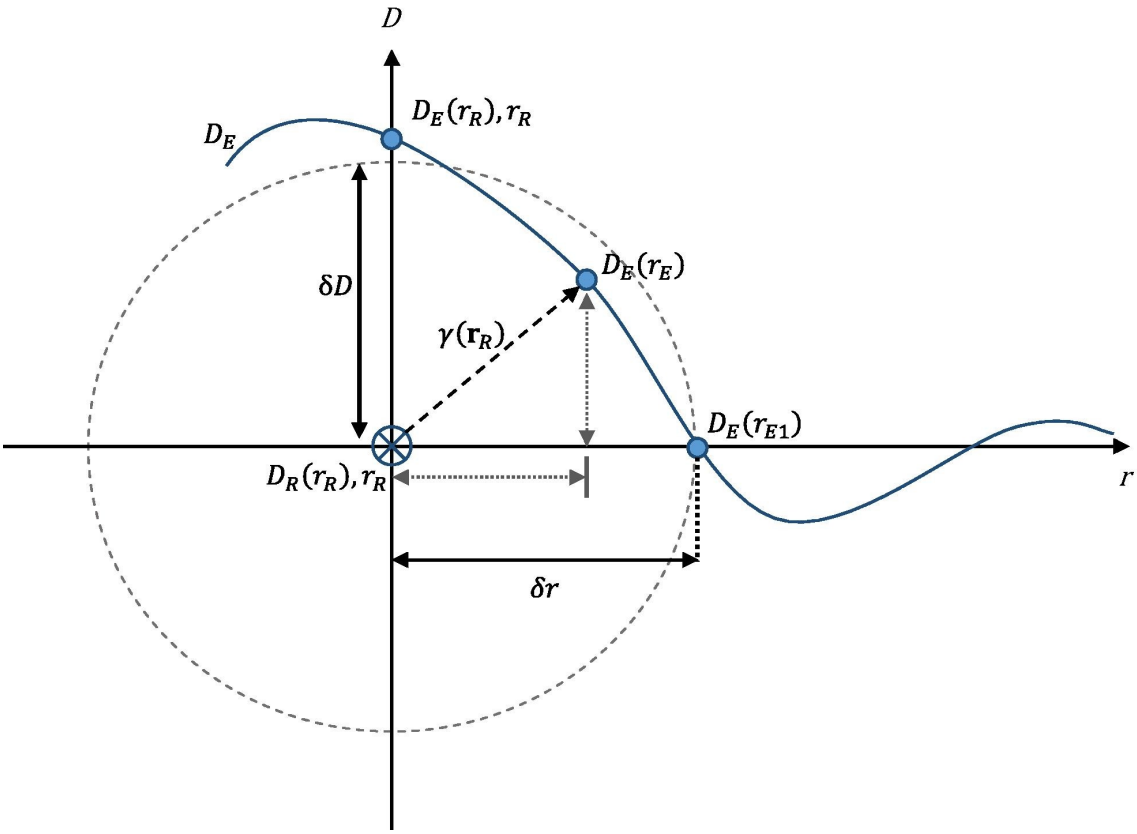


Figure 4.5: Schematic representation of the gamma index calculation, which combines dose difference (DD) and distance-to-agreement (DTA) criteria. The plot shows how the gamma value is determined for each point in the evaluated dose distribution by considering both spatial and dosimetric differences between the reference and evaluated distributions. The regions passing the gamma criteria ($\gamma < 1$) are highlighted. [46]

differences are within acceptable limits. The gamma pass rate γ_{pass}, defined as the percentage of points with $\gamma < 1$, is used to quantify the overall agreement between predicted and reference dose distributions. Typically, a 3%/3mm gamma criterion is applied, with the dose difference set at 3% and the spatial agreement at 3mm. Figure 2.10 illustrates the gamma index calculation. A lower dose threshold is usually set (10%-20% of the maximum dose) to exclude regions with low clinical significance from the gamma index calculation. The gamma index can be categorized as local or global. The local gamma index is sensitive to errors in high-gradient regions, while the global gamma index focuses on high-dose regions. The choice of gamma type depends on the evaluation's clinical objectives.[46]

5 Results

This section presents the outcomes of the methods and experiments described in the previous chapters. The focus is on evaluating the performance of the diffusion models trained for accelerated dose calculation in radiotherapy, as well as analyzing the effectiveness of the Gaussian quadrature integration technique in improving computational efficiency.

The results are divided into several key areas, including the accuracy of dose prediction, computational performance, and validation against established metrics such as the Dose-Volume Histogram (DVH) and Gamma Index analysis.

5.1 Distribution of HU values

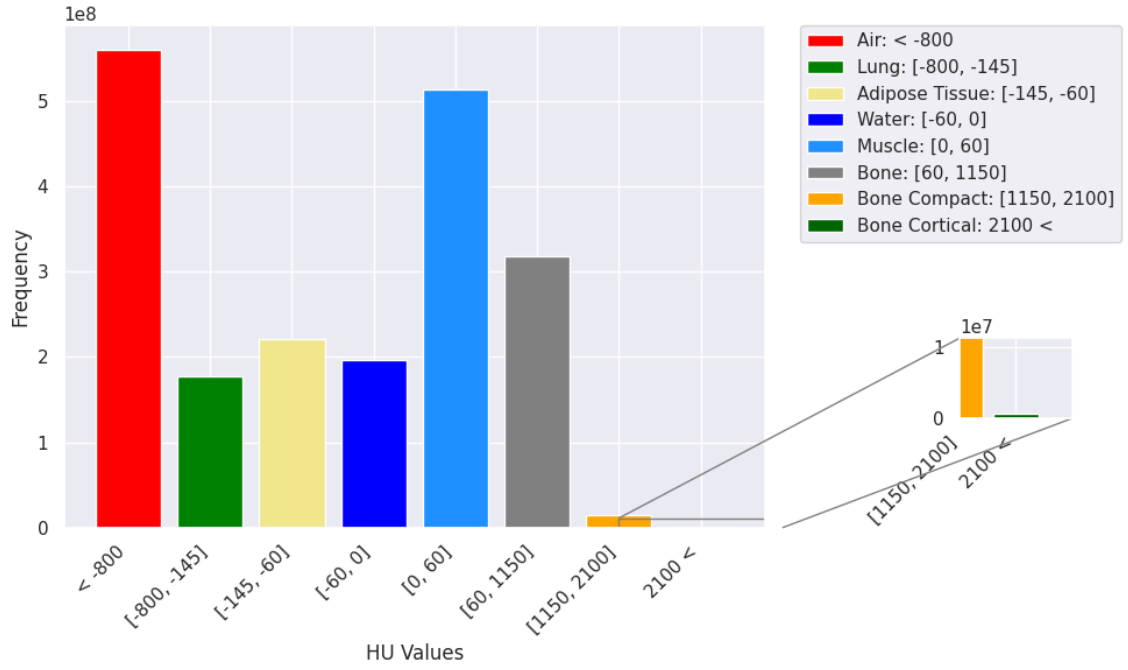


Figure 5.1: Distribution of HU Values by Material Category

Figure 5.1 illustrates the distribution of Hounsfield Unit (HU) values across various material categories based on voxel evaluations from the test dataset. The histogram categorizes these HU values into predefined ranges that correspond to different tissue types and materials, each represented by a different color. The categories range from air (low HU values) to cortical bone (high HU values), reflecting the composition typically observed in medical imaging, particularly in computed tomography (CT) scans.

A total of 2,000 test cubes are analyzed, each containing 100 x 100 x 100 voxels, resulting in one million voxels per cube. Across all test cubes, this sums up to a total of two billion voxels evaluated. The distribution of these voxels across different material categories is presented in the histogram. The bars in the plot represent the frequency of voxels that fall within specific HU intervals, each corresponding to a different material.

From the distribution, it is evident that lower HU ranges (particularly air and muscle categories) are predominant in the dataset. The vast majority of voxels fall within the air category, likely due to the inclusion of air spaces within the scanned volumes. This observation aligns with expectations in medical imaging, where the majority of body regions imaged in CT contain significant portions of air or low-density tissues.

Interestingly, the voxel counts for the bone categories (particularly those with HU $>$ 1000) are underrepresented. Compact bone (HU: 1150-2100) and cortical bone (HU $>$ 2100) constitute only a minor fraction of the total voxels. The plot includes an inset to magnify these low-frequency categories, allowing better visibility of the differences between compact and cortical bone. This underrepresentation is expected, as bones make up a smaller portion of the overall volume in typical imaging datasets compared to soft tissues and air.

5.2 Number of Particles Used for the Monte Carlo Simulations

In this study, Monte Carlo (MC) simulations were conducted using a particle count of 10^8. This number was selected to balance computational efficiency with statistical accuracy, which is crucial for reliable dose calculations. The statistical noise in MC

simulations decreases with the square root of the number of particles, $\sigma \propto \frac{1}{\sqrt{N}}$. Therefore, increasing the number of particles directly improves the precision of the dose estimates, particularly in clinically relevant regions.

5.2.1 Statistical Analysis and Rationale for 10^8 Particles

The accuracy of MC simulations is largely determined by the relative standard deviation (RSD) in dose values, especially in regions receiving therapeutic doses (e.g., > 2 Gy). To achieve a target RSD below 1% in these regions, initial tests with 10^6 and 10^7 particles were performed. While these tests showed reasonable accuracy in high-dose areas, the RSD exceeded 1%.

Simulations with 10^8 particles consistently achieved an RSD below 1% in high-dose regions, ensuring that the dose distributions are statistically reliable across all relevant voxels. This particle count provided a good compromise between accuracy and computational cost, leading to robust dose distributions even in areas with steep dose gradients.

5.2.2 Intermediate Results from Simulations

To quantify the impact of particle count on statistical uncertainty, the relative standard deviations in dose were evaluated for different particle counts. As shown in

Number of Particles	**RSD in High-Dose Regions ($>$ 2 Gy)**	**RSD in Low-Dose Regions ($<$ 2 Gy)**
10^6	3.5%	10%
10^7	1.2%	6.5%
10^8	0.8%	4%

Table 5.1: Relative standard deviation (RSD) in dose for different particle counts. The high-dose regions refer to areas with doses above 2 Gy, while low-dose regions cover doses below 2 Gy.

Table 5.1, the RSD decreases approximately with the square root of the number of particles, confirming the expected relationship $\sigma \propto \frac{1}{\sqrt{N}}$. The results illustrate that using 10^8 particles yields statistically robust dose estimates, particularly in high-dose regions.

5.2.3 Computational Efficiency Considerations

Although increasing the particle count improves statistical accuracy, it also significantly increases computational demands. To address this challenge, the simulations were run on a high-performance computing cluster, utilizing parallel processing on multiple CPUs. This allowed the completion of the simulations with 10^8 particles within a reasonable timeframe.

In conclusion, the selection of 10^8 particles was guided by the need for precise dose estimates while maintaining manageable computational costs. This choice resulted in statistically reliable dose distributions that align with clinical standards, particularly in regions of therapeutic interest.

5.3 Impact of Quadrature Order on Dose Calculation Accuracy

In this section, we evaluate the impact of different quadrature orders on the accuracy and computational efficiency of dose calculations using the Gaussian quadrature method. This method is employed to optimize energy spectrum integration, particularly in regions where the spectrum is smoother. By determining the optimal quadrature order, we aim to balance computational efficiency with the accuracy of dose predictions.

5.3.1 Quadrature Order Testing and Dose Prediction Accuracy

The photon energy spectrum used in this study was divided into two components: distinct peaks around 10 keV and a smoother tail extending toward higher energies. For the smoother tail, Gaussian quadrature integration was applied, testing various quadrature orders, specifically 2, 4, and 8 points.

For each quadrature order, Monte Carlo (MC) simulations were conducted for individual energy points identified by the Gaussian quadrature. The resulting monoenergetic dose distributions were combined according to their respective weights to form an overall dose distribution. This combined dose was compared to a reference

dose distribution generated by a full-spectrum MC simulation.

The accuracy of each quadrature order was evaluated using the gamma index with a stringent pass criterion of 1%/1 mm and a lower dose threshold of 2%, reflecting both spatial and dosimetric agreement. The gamma pass rate (percentage of points with $\gamma < 1$) was used as the primary metric for comparison.

5.3.2 Results of Gamma Index Analysis

Quadrature Order	Gamma Pass Rate (1%/1 mm)
2 Points	85.3% ± 8.6%
4 Points	98.1% ± 2.4%
8 Points	99.0% ± 1.7%

Table 5.2: Gamma pass rates with a lower dose threshold of 2% for different quadrature orders.

Table 5.2 summarizes the gamma pass rates for different quadrature orders. The results show that a quadrature order of 2 points yields a gamma pass rate of only 85.3% ± 8.6%, indicating that this order does not sufficiently capture the details of the energy spectrum, resulting in inaccuracies in dose prediction. Increasing the order to 4 points significantly improves the gamma pass rate to 98.1% ± 2.4%, demonstrating a much closer agreement with the reference dose distribution. Further increasing the order to 8 points marginally improves the gamma pass rate to 99.0% ± 1.7%, but this comes at the cost of doubling the computational time.

5.3.3 Determination of the Optimal Quadrature Order

While higher quadrature orders yield more accurate dose predictions, they also require more computational resources. The results indicate that a quadrature order of 4 provides an optimal balance between accuracy and efficiency. With a gamma pass rate of 98.1% ± 2.4%, this order achieves near-optimal dose prediction accuracy while keeping computational time reasonable.

In comparison, the 8-point quadrature order, although slightly more accurate with a 99.0% ± 1.7% gamma pass rate, requires double the time required for the 4-point quadrature. Therefore, considering the marginal gain in accuracy versus the in-

creased computational cost, the 4-point quadrature order is determined to be the best option for the integration of the energy spectrum in this study.

5.3.4 Conclusion on Quadrature Order Selection

The analysis confirms that using 4 quadrature points for Gaussian quadrature integration offers the most practical trade-off between computational efficiency and accuracy in dose prediction for radiotherapy. This choice aligns with the study's goal of developing a fast yet accurate framework for dose calculation, leveraging both advanced machine learning techniques and traditional numerical methods.

By focusing on this optimal quadrature order, the proposed method can provide reliable dose predictions suitable for clinical applications while minimizing computational overhead, thus supporting real-time adaptive radiotherapy workflows.

5.4 Summary of Energy Levels for Latent Diffusion Model Training

Based on the results discussed in the previous sections, the latent diffusion models (LDMs) for dose calculation in radiotherapy need to be trained at specific energy levels to ensure both accuracy and computational efficiency. The energy spectrum used in this study was divided into two main components: sharp peaks around 10 keV and a smoother tail extending to higher energies. For optimal dose calculation accuracy, the training of LDMs should focus on the following energy points:

- **Prominent Peaks Around 10 keV:** Four distinct energy peaks were identified in the photon fluence spectrum around 10 keV. These peaks represent critical energy levels necessary for precise modeling of dose distributions in radiotherapy. The LDMs should be specifically trained for these energy points: 7.5 keV, 9.75 keV, 11.5 keV, 13.5 keV

- **Quadrature Points for the Smooth Spectrum Region:** For the smoother part of the energy spectrum, Gaussian quadrature integration was employed to optimize the selection of energy points for training. After evaluating different quadrature orders, a 4-point Gaussian quadrature was found to provide the best balance between computational efficiency and dose prediction accuracy.

The optimal quadrature points and their corresponding energy levels are: 3.47 keV, 15.75 keV, 34.25 keV, 46.53 keV

By training LDMs at these specific energy points, the proposed method ensures a comprehensive and accurate representation of the energy spectrum. This approach allows for precise dose predictions by adequately capturing both the critical peaks and the smoother regions of the spectrum. Consequently, this strategy supports the development of a robust framework for accelerated dose calculation in radiotherapy.

5.5 Dose Calculation Performance Analysis

In this section, the dose calculation performance of the trained latent diffusion models (LDMs) across different energy levels and cube sizes is presented, assessed using the gamma index metric. The gamma index analysis was conducted for two clinically relevant criteria: 5%/1mm and 5%/3mm, to evaluate both fine and broad agreement between the predicted and reference dose distributions. The performance is compared across all trained models, including an additional evaluation of the reconstructed spectrum using Gaussian quadrature integration.

5.5.1 Gamma Index Analysis for 5%/1mm and 5%/3mm Criteria

Tables 5.3 and 5.4 summarize the gamma pass rates (the percentage of points with gamma value < 1) and their respective errors for all trained models at different cube sizes and energy levels. These tables also include the gamma pass rates for dose distributions reconstructed using Gaussian quadrature integration, providing a comprehensive overview of the models' performance across different configurations.

Table 1 presents the gamma pass rates and associated errors for the 5%/1mm criterion across different energy levels and cube sizes. The rows of the table represent various energy levels (3.47 keV, 7.5 keV, 9.75 keV, 11.5 keV, 13.5 keV, 15.75 keV, 34.25 keV, and 46.53 keV) at which the LDMs were trained, along with an additional row for the reconstructed spectrum using Gaussian quadrature integration. The columns correspond to the four different cube sizes (16^3, 32^3, 64^3, 128^3), each representing a different resolution at which the models were trained, with the smaller

Energy (keV)	Gamma Pass Rate ± Error (%)			
	Cube Size 16^3	Cube Size 32^3	Cube Size 64^3	Cube Size 128^3
3.47	87.2 ± 2.1	91.4 ± 1.8	95.6 ± 1.2	96.8 ± 0.9
7.5	85.6 ± 2.3	89.9 ± 2.0	94.3 ± 1.5	95.7 ± 1.0
9.75	84.3 ± 2.5	88.7 ± 2.1	93.8 ± 1.6	95.2 ± 1.2
11.5	82.1 ± 2.8	87.3 ± 2.3	92.1 ± 1.7	94.5 ± 1.3
13.5	80.5 ± 3.0	86.1 ± 2.5	91.4 ± 1.9	93.8 ± 1.4
15.75	78.9 ± 3.2	84.7 ± 2.7	90.1 ± 2.0	92.7 ± 1.5
34.25	77.6 ± 3.4	83.5 ± 2.9	88.8 ± 2.1	91.6 ± 1.6
46.53	76.2 ± 3.6	82.1 ± 3.0	87.4 ± 2.3	90.3 ± 1.7
Spectrum	84.1 ± 2.3	86.1 ± 2.3	92.5 ± 1.1	92.6 ± 1.4

Table 5.3: Gamma Pass Rates and Errors for 5%/1mm Criterion

Energy (keV)	Gamma Pass Rate ± Error (%)			
	Cube Size 16^3	Cube Size 32^3	Cube Size 64^3	Cube Size 128^3
3.47	86.1 ± 1.6	92.0 ± 1.4	99.6 ± 0.2	99.8 ± 0.1
7.5	86.8 ± 1.9	89.7 ± 1.5	99.4 ± 0.3	99.7 ± 0.2
9.75	85.5 ± 1.8	88.5 ± 1.6	99.3 ± 0.3	99.6 ± 0.2
11.5	83.9 ± 2.9	88.1 ± 2.1	99.0 ± 0.4	99.4 ± 0.3
13.5	80.5 ± 3.0	87.7 ± 2.0	98.7 ± 0.5	99.2 ± 0.3
15.75	79.0 ± 3.1	86.4 ± 1.9	98.4 ± 0.6	99.0 ± 0.4
34.25	78.6 ± 2.2	83.0 ± 2.3	98.1 ± 0.7	98.8 ± 0.5
46.53	75.2 ± 3.3	82.7 ± 2.8	97.9 ± 0.7	98.6 ± 0.6
Spectrum	84.8 ± 1.8	86.2 ± 2.4	99.4 ± 0.3	99.1 ± 0.4

Table 5.4: Gamma Pass Rates and Errors for 5%/3mm Criterion

cubes having lower resolution. Each cell shows the gamma pass rate (the percentage of points where the gamma value is less than 1) and its associated error for a specific energy and cube size. The gamma pass rates are expressed as a percentage, indicating the level of agreement between the predicted and reference dose distributions under the 5% dose difference and 1mm distance-to-agreement criteria. Generally, the gamma pass rates increase as the cube size increases, indicating that higher-resolution models (with smaller voxel sizes) perform better in terms of dose calculation accuracy. The gamma pass rates range from around 76.2% to 96.8% for lower resolutions (16^3) to higher resolutions (128^3) across different energy levels.

Table 2 summarizes the gamma pass rates and errors for the 5%/3mm criterion across the same energy levels and cube sizes as in Table 1. The rows again represent the energy levels at which the LDMs were trained, including the reconstructed spectrum using Gaussian quadrature integration. The columns indicate the four different cube sizes (16^3, 32^3, 64^3, 128^3), providing a comparison of performance at varying spatial resolutions. Each cell contains the gamma pass rate and error for a specific energy and cube size under the 5%/3mm criterion. The values represent the percentage of points meeting the gamma criteria, showing the agreement between predicted and reference dose distributions with a looser spatial agreement (3mm). Similar to the 5%/1mm criterion, the gamma pass rates improve with increasing cube sizes, reaching as high as 99.8% for 128^3 cubes. However, the gamma pass rates are generally higher across all cube sizes and energy levels compared to the 5%/1mm criterion, reflecting the less stringent spatial requirement.

Overall, these tables provide key insights into the dose calculation performance of LDMs trained at different energy levels and resolutions. The data reveal that models trained at higher resolutions (smaller voxel sizes) yield more accurate dose calculations, as evidenced by the higher gamma pass rates for both the 5%/1mm and 5%/3mm criteria. Across different energy levels, there is a trend of decreasing gamma pass rates as the energy increases, particularly noticeable in lower-resolution cubes (16^3 and 32^3). This indicates that lower-energy spectra might be more amenable to accurate dose prediction using LDMs. The reconstructed spectrum using Gaussian quadrature integration shows comparable gamma pass rates to the LDMs, especially for the larger cube sizes (64^3 and 128^3), suggesting that Gaussian quadrature integration is a viable approach for spectrum reconstruction

in dose calculations.

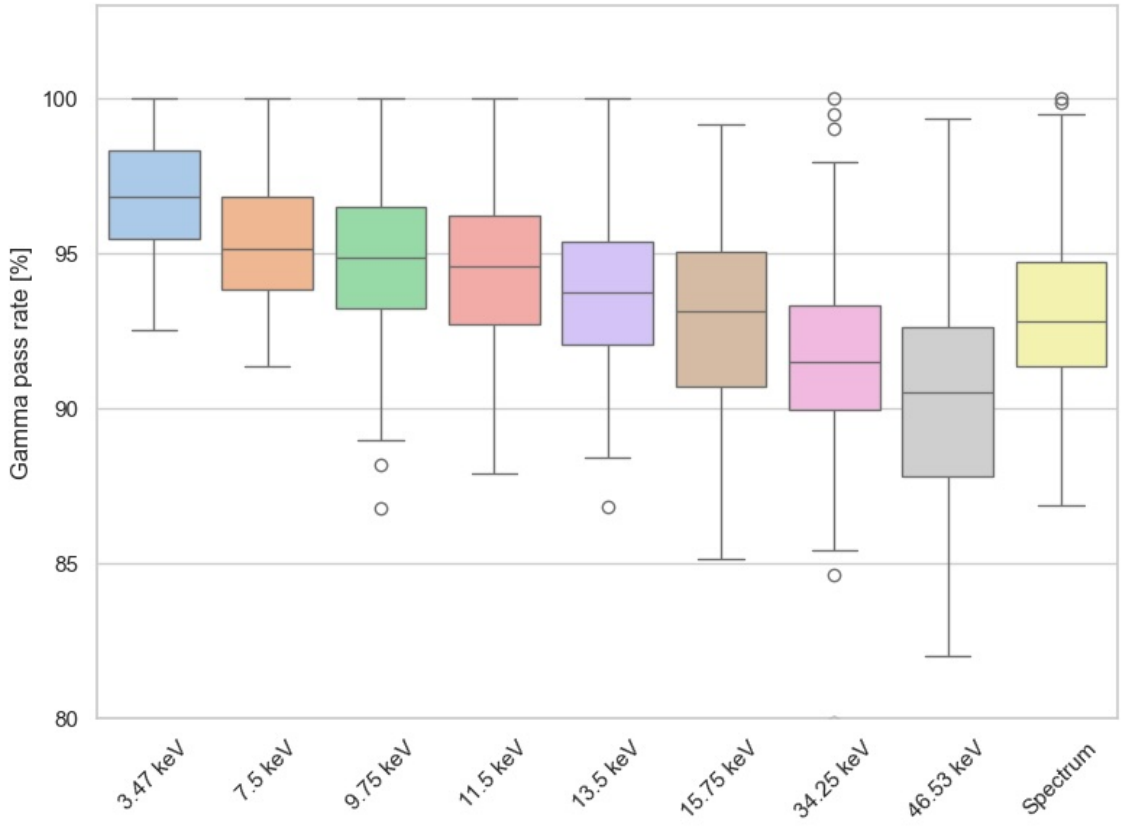


Figure 5.2: Gamma pass rate for 5%/1mm criterion. The plot shows the gamma pass rates for various energy levels for a point source with a beam in the z-direction using full cube size.

Figures 5.2 and 5.3 display the gamma pass rates for the 5%/1mm and 5%/3mm criteria, respectively. These plots provide a visual representation of the dose calculation performance of the latent diffusion models (LDMs) across different energy levels for a full cube size of 128^3, which corresponds to the highest resolution among the models trained.

Figure 5.2 illustrates the gamma pass rates for the 5%/1mm criterion, which evaluates both the dose difference (5%) and spatial agreement (1mm) between the predicted and reference dose distributions. The box plot shows the distribution of gamma pass rates across various energy levels (3.47 keV, 7.5 keV, 9.75 keV, 11.5 keV, 13.5 keV, 15.75 keV, 34.25 keV, 46.53 keV) and the reconstructed spectrum using Gaussian quadrature integration. Each box represents the interquartile range

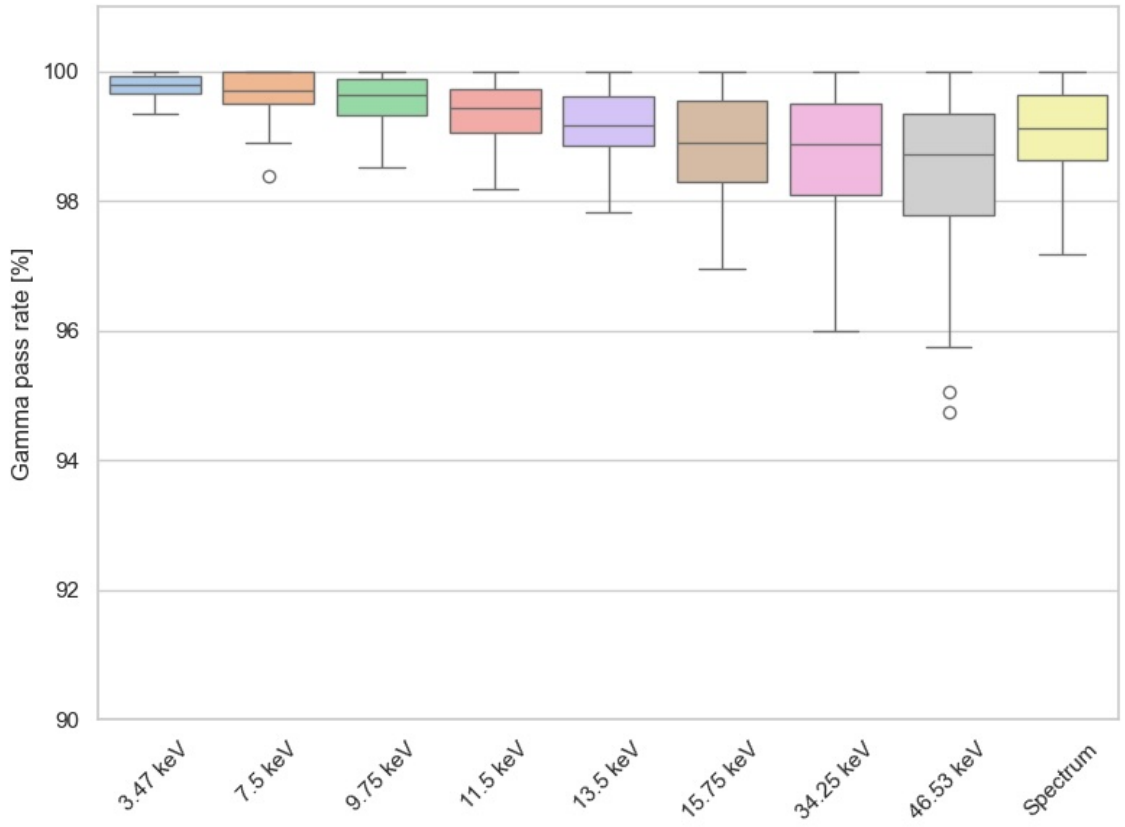


Figure 5.3: Gamma pass rate for 5%/3mm criterion. Similar setup as the 5%/1mm criterion, but with a more relaxed distance agreement of 3mm.

(IQR) of the gamma pass rates, with the median indicated by a horizontal line inside each box, while the whiskers extend to show the range of the data within 1.5 times the IQR. Outliers beyond this range are indicated by individual points.

From the plot, it is evident that the gamma pass rates are generally high across all energy levels, with most distributions centered around 95-100%. However, there is a slight decline in gamma pass rates at higher energies, particularly for energies above 13.5 keV, which shows more variability and lower median values. The reconstructed spectrum also shows comparable performance to the models trained at specific energy points, indicating its robustness in dose calculation. The overall trend suggests that the LDMs provide accurate dose predictions with high spatial agreement under the 5%/1mm criterion.

Figure 5.3 presents the gamma pass rates for the 5%/3mm criterion, which uses

the same dose difference criterion (5%) but a more relaxed distance-to-agreement criterion (3mm). The plot follows a similar format to Figure 5.2, showing the distribution of gamma pass rates across the same set of energy levels and the reconstructed spectrum.

Compared to the 5%/1mm criterion, the gamma pass rates for the 5%/3mm criterion are noticeably higher and exhibit less variability across most energy levels. The majority of the box plots have median values close to 100%, and the interquartile ranges are tighter, indicating a more consistent performance with the relaxed spatial agreement. Outliers are present at certain energy levels, but they are relatively few, highlighting the robustness of the models under this criterion. The reconstructed spectrum continues to show high gamma pass rates, similar to those of the models trained at specific energy points.

The results presented in Figures 5.2 and 5.3 demonstrate the strong performance of the LDMs in predicting dose distributions across different energy levels, especially when evaluated under a more relaxed spatial criterion. The high gamma pass rates under both criteria confirm the models' accuracy in dose prediction, with performance slightly improving under the 5%/3mm criterion due to the broader spatial agreement allowance. The reconstructed spectrum using Gaussian quadrature integration remains competitive with the LDMs, validating its use in spectrum reconstruction for dose calculations.

5.5.2 Dose Volume Histogram

The Dose-Volume Histogram (DVH) presented in Figure 5.4 serves as a key evaluation tool to compare dose distributions calculated using Monte Carlo (MC) simulations and Latent Diffusion Models (LDM). The DVH simplifies the 3D dose distributions into a 2D plot that shows the cumulative volume percentage of a tissue receiving a specific dose or higher. The x-axis represents the dose percentage, while the y-axis shows the volume percentage.

In this plot, different lines correspond to specific tissues: air, lung, bone, and cortical bone. The solid lines represent dose distributions obtained from MC simulations, considered the gold standard due to their high accuracy in modeling radiation transport and interactions. In contrast, the dashed lines correspond to dose distributions

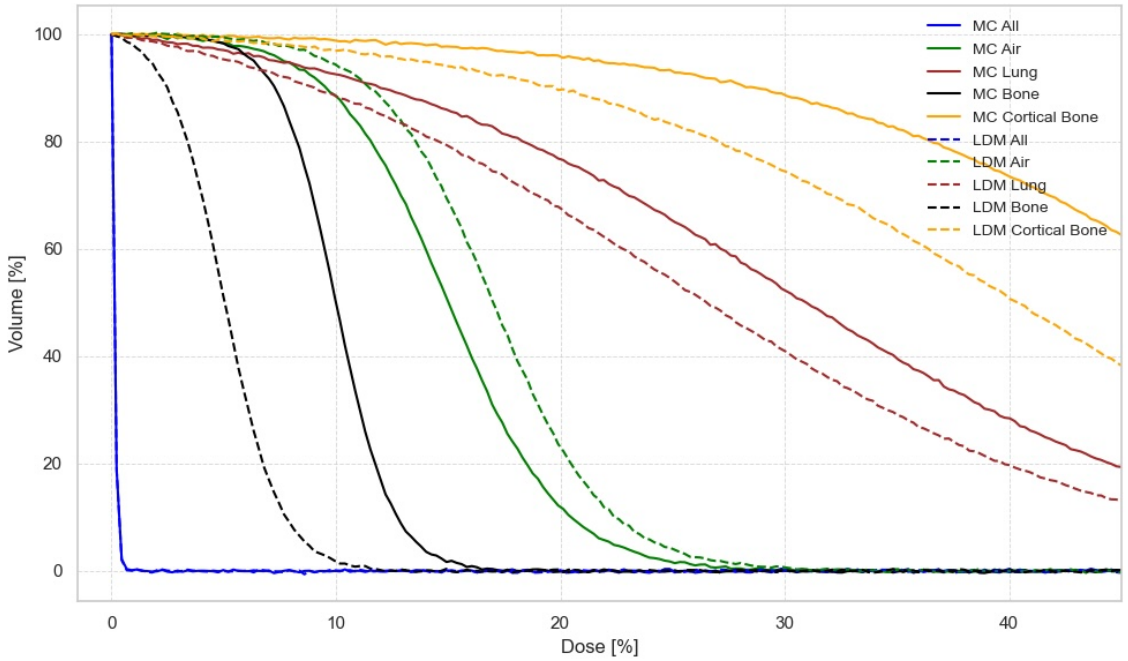


Figure 5.4: Cumulative Dose-Volume Histogram (DVH) Comparison for Various Tissues and Dose Calculation Methods The DVH illustrates the percentage of the volume of different tissues receiving specific doses, comparing results from Monte Carlo (MC) simulations and Latent Diffusion Models (LDM). Solid lines represent the MC results, while dashed lines depict LDM outcomes. The tissues evaluated include air, lung, bone, and cortical bone, with lines color-coded accordingly. This comparison aids in assessing the accuracy and coverage of dose distributions across tissues, with MC serving as the benchmark.

calculated using the LDMs.

The results indicate that the LDM closely follows the MC results for softer tissues like lung and air, showing a rapid decline in volume with increasing dose. However, for denser tissues such as bone and cortical bone, discrepancies become more apparent, with LDM generally underestimating the volume receiving higher doses compared to MC. This divergence suggests that while LDM provides a computationally efficient alternative, its accuracy varies across different tissue types, highlighting the importance of further refinement and validation against established methods like MC.

5.5.3 Run-Time Analysis

The computations for this thesis were performed on the bwUniCluster 2.0[1], a high-performance computing (HPC) cluster designed for parallel computing with distributed memory. This system is composed of numerous nodes, each equipped with at least two Intel Xeon processors, local memory, disks, network adapters, and optionally, accelerators such as NVIDIA Tesla V100, A100, or H100 GPUs. The nodes are interconnected via a high-speed InfiniBand network, which ensures efficient communication between them. Additionally, the cluster integrates a Lustre parallel file system, connected through the coupling of the InfiniBand interfaces of the file servers with the InfiniBand switch of the compute cluster.

Each node in the bwUniCluster 2.0 runs Red Hat Enterprise Linux (RHEL) 8.4 as the operating system, supplemented with various software packages, including the SLURM workload manager. The nodes within the cluster are categorized based on their roles, which include login nodes, compute nodes, file server nodes, and administrative server nodes.

The bwUniCluster 2.0's infrastructure supports a wide range of computational tasks, from interactive program development on login nodes to intensive parallel computations on compute nodes, all facilitated by robust data management through the Lustre file system. This setup ensures a flexible and efficient environment for running complex simulations and data processing tasks, as required by the research conducted in this thesis.

For training the LDMs a NVIDIA Tesla V100 GPU with 32 GB was used. The training took between 5 and 10 hours per model. The MC simulations, needed for training the LDMs take multiple weeks.

A significant challenge in comparing the generation of dose distributions between MC simulations and LDMs lies in the computational resources they utilize. MC simulations traditionally run on CPUs, which are optimized for handling a large number of parallel, yet less intensive, computational tasks typical of the probabilistic modeling used in MC methods. On the other hand, LDMs are designed to leverage the capabilities of GPUs, which excel at highly parallel computations with a large num-

[1] https://www.bwhpc.de

ber of cores optimized for the matrix operations inherent in deep learning models. This disparity in computational architecture not only affects the run times of the respective methods but also complicates direct performance comparisons, as CPUs and GPUs have different strengths and efficiencies. Consequently, benchmarking these models requires careful consideration of the hardware specifics, especially in clinical settings where the availability of high-performance GPUs may be limited.

While the computational environments for MC simulations and LDMs differ significantly, it is noteworthy that LDMs generate dose distributions approximately five times faster than MC simulations. It is important to highlight that neither the MC simulations nor the LDMs were specifically optimized for speed in this study. The run times observed reflect the raw performance of the models as implemented, without further optimizations such as code acceleration techniques, hardware-specific tuning, or algorithmic refinements. This indicates potential areas for future improvement, where targeted optimizations could further reduce computation times and enhance the practical applicability of these models in clinical workflows.

6 Discussion

The aim of this thesis was to accelerate dose calculation in radiotherapy using Latent Diffusion Models (LDMs) combined with Gaussian quadrature integration for spectrum representation. The results obtained provide several insights into the efficiency and accuracy of this approach. In this discussion, the results presented in the previous sections will be comprehensively interpreted, compared with previous work, and an outline for potential future directions will be given.

6.1 Distribution of HU Values and Material Mapping

The analysis of HU (Hounsfield Unit) values, as shown in Figure 5.1, reveals the predominance of low-density tissues such as air and muscle in the dataset, while higher-density tissues like compact and cortical bone are less frequent. This distribution aligns with typical clinical imaging scenarios where soft tissues occupy a larger volume compared to bones. The accurate mapping of HU ranges to Geant4 materials (Table 4.1) is crucial for reliable Monte Carlo (MC) simulations, as it directly affects the precision of the ground truth dose distributions used for training and validation. The current approach's performance in correctly predicting doses across these varied tissue types indicates the effectiveness of the material mapping and preprocessing steps.

The underrepresentation of bone tissues in the dataset could potentially limit the model's ability to accurately predict doses in high-density regions. Future work could consider balancing the dataset or applying synthetic data augmentation techniques to enhance the representation of these tissue types. This would likely improve the model's robustness across different anatomical regions.

6.2 Impact of Particle Count on Monte Carlo Simulations

The choice of 10^8 particles for the MC simulations, as discussed in Section 5.2, was driven by the need to minimize statistical uncertainties in the dose distributions while keeping computational costs manageable. The results in Table 5.1 demonstrate that using fewer particles leads to higher relative standard deviations (RSD) in dose estimates, particularly in low-dose regions. The selected particle count achieves an RSD below 1% in high-dose regions, ensuring that the generated ground truth data are reliable for model training.

The balance between computational efficiency and accuracy achieved with 10^8 particles indicates a well-chosen compromise, but it is important to note that further increases in particle count would continue to reduce uncertainty, albeit at the cost of significantly increased computation times. Future studies could explore adaptive particle counts based on local dose gradients or employ variance reduction techniques to further enhance simulation efficiency.

6.3 Optimization of Quadrature Order for Spectrum Representation

The choice of the quadrature order for Gaussian quadrature integration was found to be critical for achieving a balance between computational efficiency and dose prediction accuracy. As detailed in Section 5.3, a 4-point quadrature order provided the optimal balance, achieving a gamma pass rate of 98.1% $\pm$ 2.4% for the 1%/1mm criterion, which is competitive with the full-spectrum MC reference simulations. The 2-point quadrature order showed inadequate performance with a gamma pass rate of only 85.3%, indicating that it does not capture the nuances of the spectrum adequately. Conversely, the 8-point order offered only marginal gains in accuracy at the expense of doubling the computational effort compared to the 4-point order.

This finding suggests that while increasing the quadrature order improves dose calculation accuracy, there are diminishing returns beyond a certain point. This observation is important for clinical applications where computational efficiency is

as critical as accuracy. Future work could explore adaptive quadrature schemes that dynamically adjust the number of points based on the spectral characteristics and clinical requirements.

6.4 Energy Levels and Dose Prediction Accuracy

The energy levels selected for LDM training, particularly the sharp peaks around 10 keV and the points determined by Gaussian quadrature for the smoother tail, play a significant role in ensuring accurate dose predictions. The gamma index analysis (Tables 5.3 and 5.4) reveals that the LDMs achieve high gamma pass rates across different cube sizes and energy levels, with performance improving at higher resolutions. For the 5%/1mm criterion, the gamma pass rates for the largest cube size (128^3) are consistently above 93%, demonstrating the model's robustness in achieving high spatial and dosimetric agreement.

However, the slight decline in performance at higher energy levels suggests that the model's ability to capture complex interactions diminishes as energy increases. This could be due to the limitations in the current LDM architecture or the need for more extensive training data that better represents high-energy scenarios. Future improvements could involve enhancing the model's architecture to better handle these cases or incorporating more sophisticated training techniques, such as transfer learning from lower to higher energy domains.

6.5 Comparison with Previous Work

The proposed approach builds on previous work, particularly the Hybrid Monte Carlo Algorithm developed by Amann [6], which combined MC simulations with U-Net architectures. While the Hybrid Algorithm successfully reduced computation times, it faced challenges in accurately predicting doses in regions with complex tissue heterogeneity and high energy gradients. The use of LDMs in this thesis aims to overcome the limitations of traditional diffusion models by operating in a lower-dimensional latent space, which allows for more efficient computation and better modeling of complex spatial patterns. LDMs effectively capture intricate spatial correlations within the data by focusing on the most relevant features in the latent space, thus improving the model's ability to handle variability in anatomical regions

and different energy levels. The results demonstrate good accuracy in dose calculations, particularly in heterogeneous tissues and challenging anatomical structures, indicating that LDMs provide a robust and generalized approach to modeling dose distributions across diverse clinical scenarios.

Moreover, the integration of Gaussian quadrature for optimizing energy spectrum representation further enhances the approach, providing a more refined method for balancing computational speed with accuracy. This dual approach of using LDMs and numerical integration represents a significant advancement over previous methods, paving the way for more accurate and efficient dose calculation models.

6.6 Dose Volume Histogram (DVH)

The Dose-Volume Histogram (DVH) results provide critical insights into the performance of Latent Diffusion Models (LDMs) compared to Monte Carlo (MC) simulations, which are considered the gold standard for dose calculation in radiotherapy. As illustrated in Figure 5.4, the DVH summarizes the dose coverage for various tissues, including air, lung, bone, and cortical bone, highlighting key differences between the LDM and MC approaches.

For low-density tissues like air and lung, the LDMs demonstrate a close agreement with the MC results, which is evident from the rapid decline in the cumulative volume percentage as the dose increases. This suggests that LDMs are effective at modeling dose distributions in less dense regions, where radiation interactions are less complex. However, discrepancies become more pronounced in denser tissues, such as bone and cortical bone, where LDMs tend to underestimate the volume receiving higher doses compared to MC simulations. This underestimation could lead to potential inaccuracies in clinical dose planning, particularly in scenarios involving critical structures near bone interfaces.

Figure 5.1, which depicts the distribution of Hounsfield Unit (HU) values across different tissue types, offers further context for understanding the observed DVH discrepancies. The figure shows a predominance of lower HU values corresponding to soft tissues like air and muscle, with relatively fewer data points representing higher HU values associated with compact and cortical bone. This imbalance in the

data distribution may contribute to the LDM's reduced accuracy in high-density regions, as the models may be less trained and therefore less precise in predicting dose distributions for these tissues.

The underrepresentation of high-density tissues in the training data, as indicated by Figure 5.1, highlights a potential area for further improvement of LDMs. Increasing the representation of bone and other dense tissues in the training dataset, or employing data augmentation techniques to enhance model performance in these regions, could help mitigate the underestimation seen in the DVH results.

These findings are significant because they indicate that while LDMs provide a computationally efficient alternative to MC simulations, their current implementation may not yet be suitable for all clinical scenarios, particularly those involving heterogeneous tissues. The underestimation of dose in bone regions could lead to overly aggressive treatment plans. Any underestimation in tissues could fail to adequately protect healthy tissue from excessive radiation.

Future work should focus on enhancing the LDM's accuracy for high-density tissues by refining the model architecture, improving training strategies, and exploring hybrid methods that combine the strengths of both LDM and MC approaches. This would not only improve the clinical applicability of LDMs but also ensure that dose calculations remain accurate across a wide range of tissue types, thereby supporting the safe and effective delivery of radiotherapy.

6.7 Run-Time Analysis and Clinical Applicability

The run-time analysis of Monte Carlo (MC) simulations and Latent Diffusion Models (LDMs) revealed a stark contrast in computational efficiency, primarily driven by the different hardware platforms utilized. LDMs, leveraging the parallel processing power of NVIDIA Tesla V100 GPUs, were found to generate dose distributions approximately five times faster than MC simulations, which operate on traditional CPU-based systems. This speed differential positions LDMs as a potentially transformative approach for dose calculation in radiotherapy, where rapid turnaround times are often critical for clinical decision-making and adaptive treatment planning.

However, it is crucial to note that neither the MC simulations nor the LDMs were specifically optimized for speed in this study. The observed run times reflect the inherent capabilities of the models and hardware as implemented, without advanced optimizations such as algorithmic refinements, hardware-specific accelerations, or code-level enhancements. This suggests a significant potential for further reducing computation times through targeted optimizations. For example, optimizing LDMs with techniques like mixed-precision training, exploiting GPU tensor cores, or using model compression strategies could further accelerate performance. Similarly, porting MC simulations to GPUs, optimizing parallel CPU processes, or employing more efficient random sampling algorithms could enhance their computational efficiency.

The disparity in hardware environments between MC simulations and LDMs also highlights the complexities of directly comparing their clinical applicability. GPUs, with their high core count and parallel processing capabilities, are well-suited for the matrix-heavy computations of deep learning models like LDMs. In contrast, the CPU's architecture is better aligned with the sequential and iterative nature of MC simulations. This dichotomy underscores the need for careful consideration of hardware resources when evaluating dose calculation methodologies for clinical integration. Future studies should aim to benchmark these methods under equivalent conditions, potentially exploring hybrid CPU-GPU configurations or scalable cloud-based solutions to better assess their relative performance in clinical settings.

From a clinical applicability perspective, the faster computation times of LDMs offer a compelling advantage, especially in scenarios requiring frequent recalculations, such as adaptive radiotherapy or real-time dose adjustments during treatment sessions. The reduced computational burden could streamline workflows, enhance patient throughput, and allow for more responsive treatment adaptations. However, ensuring the clinical reliability of LDMs remains paramount; thus, further validation against the gold-standard MC simulations is necessary to fully establish their accuracy and robustness in diverse clinical scenarios.

In conclusion, while LDMs demonstrate a clear potential to improve the efficiency of dose calculations in radiotherapy, there remains significant room for optimizing both LDMs and MC simulations. Addressing these optimization opportunities could not only further enhance the speed and feasibility of these models but also bridge the

gap between computational performance and clinical utility, ultimately improving the standard of care in radiotherapy planning and delivery.

6.8 Limitations and Future Directions

Despite the promising findings, several limitations exist. The slight decrease in gamma pass rates at higher energies and in complex tissue geometries suggests that the current model may need further refinement. Future work could focus on incorporating multi-scale diffusion models that adaptively switch between different spatial resolutions based on local dose gradients or anatomical complexity. Additionally, exploring hybrid approaches that integrate LDMs with other deep learning architectures or traditional numerical methods, such as the DoseDiff model, could further enhance accuracy and adaptability.

Furthermore, more comprehensive validation studies involving diverse patient datasets, varying anatomical regions, and treatment types (such as IMRT, VMAT, and proton therapy) would be valuable in assessing the generalizability and robustness of the proposed method. These studies could also investigate the potential of integrating patient-specific information to tailor the models more closely to individual treatment needs.

7 Conclusion

Based on the findings and reflections throughout this thesis, it is evident that integrating Latent Diffusion Models (LDMs) with Gaussian quadrature techniques presents a promising avenue for enhancing dose calculation in radiotherapy. The primary goal of this research was to address the computational challenges of Monte Carlo (MC) simulations—namely, their high computational demands—while striving to maintain their well-regarded accuracy. The work demonstrated that LDMs, when specifically tailored for relevant energy ranges and complemented by Gaussian quadrature for efficient spectrum representation, can potentially improve the efficiency of dose calculations without significantly sacrificing accuracy.

As outlined in the introduction, accurate dose calculation is essential in radiotherapy to ensure effective treatment while minimizing harm to surrounding healthy tissues. Traditional methods, particularly MC simulations, are considered the gold standard due to their precision but are often impractical in clinical settings because of their high computational costs and lengthy processing times. This thesis proposed that LDMs could offer a viable alternative by operating in a reduced-dimensional latent space, thus reducing computational costs and facilitating faster calculations, a hypothesis that was supported by the experimental results.

The experiments indicated that LDMs perform well in low-density tissues, such as air and lung, showing comparable accuracy to MC simulations. However, the performance was less consistent in higher-density tissues, such as bone, where LDMs tended to underestimate dose values. This finding suggests that there is room for improvement, particularly in enhancing the representation of high-density tissues within the training dataset, which could be addressed in future iterations through more diverse training strategies or data augmentation techniques. Additionally, the exploration of different quadrature orders revealed that a 4-point Gaussian quadrature provides a reasonable balance between computational efficiency and accuracy, aligning with the overall objective of developing a practical framework suitable for

clinical use.

This thesis serves as a proof of concept for the potential application of LDMs in radiotherapy dose calculations. While the results are promising, further work is necessary to refine these models, especially to improve accuracy in higher-density regions. Moreover, comprehensive validation across a broader range of clinical scenarios and patient data would be essential to assess the robustness and generalizability of the proposed framework.

In conclusion, this research has laid a foundational step toward more adaptive, precise, and efficient radiotherapy treatment planning. By demonstrating the feasibility of using LDMs with Gaussian quadrature as a faster alternative to traditional MC simulations, this thesis contributes valuable insights into the ongoing development of advanced computational techniques in radiotherapy, with the ultimate aim of enhancing patient outcomes through more efficient and accurate treatment planning.

Part I

Appendix

A Lists

A.1 List of Figures

A.2 List of Tables

B Bibliography

[1] Ervin B Podgorsak. *Radiation oncology physics: a handbook for teachers and students*. 2005.

[2] Wolfgang Schlegel, Christian P Karger, and Oliver Jäkel. *Medizinische Physik: Grundlagen–Bildgebung–Therapie–Technik*. Springer-Verlag, 2018.

[3] Keyvan Jabbari. "Review of fast Monte Carlo codes for dose calculation in radiation therapy treatment planning". In: *Journal of Medical Signals & Sensors* 1.1 (2011), pp. 73–86.

[4] Yiwen Zhang et al. "DoseDiff: distance-aware diffusion model for dose prediction in radiotherapy". In: *IEEE Transactions on Medical Imaging* (2024).

[5] C Kontaxis et al. "DeepDose: Towards a fast dose calculation engine for radiation therapy using deep learning". In: *Physics in Medicine & Biology* 65.7 (2020), p. 075013.

[6] Helen Amann. *Hybrid Monte Carlo Algorithm based on deep learning for accelerated dose calculation in radiotherapy*. 2022.

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models". In: *Advances in neural information processing systems* 33 (2020), pp. 6840–6851.

[8] Jooyoung Choi et al. "Ilvr: Conditioning method for denoising diffusion probabilistic models". In: *arXiv preprint arXiv:2108.02938* (2021).

[9] Gunilla C Bentel, Charles E Nelson, and K Thomas Noell. *Treatment planning and dose calculation in radiation oncology*. Elsevier, 2014.

[10] Anders Gustafsson, Bengt K Lind, and Anders Brahme. "A generalized pencil beam algorithm for optimization of radiation therapy". In: *Medical physics* 21.3 (1994), pp. 343–356.

[11] W Ulmer, J Pyyry, and W Kaissl. "A 3D photon superposition/convolution algorithm and its foundation on results of Monte Carlo calculations". In: *Physics in Medicine & Biology* 50.8 (2005), pp. 1767–1790.

[12] C-M Ma et al. “Clinical implementation of a Monte Carlo treatment planning system”. In: *Medical physics* 26.10 (1999), pp. 2133–2143.

[13] Robert L Harrison. “Introduction to monte carlo simulation”. In: *AIP conference proceedings*. Vol. 1204. NIH Public Access. 2010, p. 17.

[14] William R Hendee, Geoffrey S Ibbott, and Eric G Hendee. *Radiation therapy physics*. John Wiley & Sons, 2013.

[15] Joao Seco and Frank Verhaegen. *Monte Carlo techniques in radiation therapy*. CRC press Boca Raton, 2013.

[16] Pedro Andreo. “Monte Carlo simulations in radiotherapy dosimetry”. In: *Radiation Oncology* 13 (2018), pp. 1–15.

[17] Yelda Elcim, Bahar Dirican, and Omer Yavas. “Dosimetric comparison of pencil beam and Monte Carlo algorithms in conformal lung radiotherapy”. In: *Journal of Applied Clinical Medical Physics* 19.5 (2018), pp. 616–624.

[18] Alexander Quinn Nichol and Prafulla Dhariwal. “Improved denoising diffusion probabilistic models”. In: *International conference on machine learning*. PMLR. 2021, pp. 8162–8171.

[19] Prafulla Dhariwal and Alexander Nichol. “Diffusion models beat gans on image synthesis”. In: *Advances in neural information processing systems* 34 (2021), pp. 8780–8794.

[20] Aditya Ramesh et al. “Zero-shot text-to-image generation”. In: *International conference on machine learning*. Pmlr. 2021, pp. 8821–8831.

[21] Jonathan Ho and Tim Salimans. “Classifier-free diffusion guidance”. In: *arXiv preprint arXiv:2207.12598* (2022).

[22] Yang Song and Stefano Ermon. “Improved techniques for training score-based generative models”. In: *Advances in neural information processing systems* 33 (2020), pp. 12438–12448.

[23] Zhenghao Feng et al. “Diffdp: Radiotherapy dose prediction via a diffusion model”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 191–201.

[24] Ling Yang et al. “Diffusion models: A comprehensive survey of methods and applications”. In: *ACM Computing Surveys* 56.4 (2023), pp. 1–39.

[25] Robin Rombach et al. “High-resolution image synthesis with latent diffusion models”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 10684–10695.

[26] Md Selim et al. “Latent Diffusion Model for Medical Image Standardization and Enhancement”. In: *arXiv preprint arXiv:2310.05237* (2023).

[27] Zhao Peng et al. “Deep learning for accelerating Monte Carlo radiation transport simulation in intensity-modulated radiation therapy”. In: *arXiv preprint arXiv:1910.07735* (2019).

[28] Philippe Meyer et al. “Survey on deep learning for radiotherapy”. In: *Computers in biology and medicine* 98 (2018), pp. 126–146.

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer. 2015, pp. 234–241.

[30] A. Grossberg et al. *HNSCC Version 4*. Dataset. 2020. DOI: 10.7937/k9/tcia.2020.a8sh-7363. URL: https://doi.org/10.7937/k9/tcia.2020.a8sh-7363.

[31] Kenneth Clark et al. “The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository”. In: *Journal of digital imaging* 26 (2013), pp. 1045–1057.

[32] Suleman Surti et al. “Performance of Philips Gemini TF PET/CT scanner with special consideration for its time-of-flight imaging capabilities”. In: *Journal of Nuclear Medicine* 48.3 (2007), pp. 471–480.

[33] Thomas Martin Lehmann, Claudia Gonner, and Klaus Spitzer. “Addendum: B-spline interpolation in medical image processing”. In: *IEEE transactions on medical imaging* 20.7 (2001), pp. 660–665.

[34] Pauli Virtanen et al. “SciPy 1.0: fundamental algorithms for scientific computing in Python”. In: *Nature methods* 17.3 (2020), pp. 261–272.

[35] Dingjun Yu et al. “Mixed pooling for convolutional neural networks”. In: *Rough Sets and Knowledge Technology: 9th International Conference, RSKT 2014, Shanghai, China, October 24-26, 2014, Proceedings 9*. Springer. 2014, pp. 364–375.

[36] Jaroslav Solc. *Catalogue of X-ray photon fluence spectra of electronic brachytherapy device INTRABEAM manufactured by ZEISS*. Data set. 2023. DOI: 10.5281/zenodo.7594578. URL: https://doi.org/10.5281/zenodo.7594578.

[37] MSU EURAMET. "European metrology programme for innovation and research (EMPIR)". In: *URL: http://msu. euramet. org/calls. html* (2022).

[38] Sea Agostinelli et al. "GEANT4—a simulation toolkit". In: *Nuclear instruments and methods in physics research section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 506.3 (2003), pp. 250–303.

[39] Gene H Golub and John H Welsch. "Calculation of Gauss quadrature rules". In: *Mathematics of computation* 23.106 (1969), pp. 221–230.

[40] M Jorge Cardoso et al. "Monai: An open-source framework for deep learning in healthcare". In: *arXiv preprint arXiv:2211.02701* (2022).

[41] Walter HL Pinaya et al. "Brain imaging generation with latent diffusion models". In: *MICCAI Workshop on Deep Generative Models*. Springer. 2022, pp. 117–126.

[42] Richard Zhang et al. "The unreasonable effectiveness of deep features as a perceptual metric". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 586–595.

[43] Patrick Esser, Robin Rombach, and Bjorn Ommer. "Taming transformers for high-resolution image synthesis". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 12873–12883.

[44] Dan Nguyen et al. "A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning". In: *Scientific reports* 9.1 (2019), p. 1076.

[45] RE Drzymala et al. "Dose-volume histograms". In: *International Journal of Radiation Oncology* Biology* Physics* 21.1 (1991), pp. 71–78.

[46] Mohammad Hussein, CH Clark, and Andrew Nisbet. "Challenges in calculation of the gamma index in radiotherapy–towards good practice". In: *Physica Medica* 36 (2017), pp. 1–11.

Erklärung:

Ich versichere, dass ich diese Arbeit selbstständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Heidelberg, den 31.08.2024 J. Hagedorn .....