IMPERIAL

Project Code: PART-Maier-4

Applying Self-Attention and Sequence Learning to Simulated LDM Detection in Superfluid Helium-4

CID: 02025332

Supervisor: Benedikt Maier

Assessor: Alexander Tapper

Word Count: 9997

Layperson's Summary

Finding the Invisible: Using AI to Help Detect Dark Matter

What is the universe made of? Surprisingly, everything that can be seen – stars, planets, and people – accounts for just a tiny fraction of what exists. Most of the universe is believed to be composed of something invisible called **dark matter**. Although dark matter does not shine or absorb light, its gravitational effects shape galaxies and influence the expansion of the universe. Discovering the true nature of dark matter remains one of the most profound challenges in modern physics.

Some theories suggest that dark matter could be made up of extremely light particles. When these particles interact with ordinary matter, they would leave behind incredibly faint signals. The **DELight experiment** has been designed to detect such tiny interactions using **superfluid helium** – a special, frictionless form of liquid helium that is remarkably sensitive to small energy deposits. A dark matter particle colliding with a helium atom is expected to create tiny splashes of light, heat, and quantum excitations.

Because these signals are so faint, and often obscured by random detector noise, advanced **machine learning** techniques were applied to simulated data to help pick out real signals. Three models were created: one that could reconstruct where an interaction took place, one that could also classify the type of interaction, and a third that could simultaneously determine the interaction's position, type, and energy. This progressive development allowed an investigation into whether solving multiple tasks at once could improve model performance.

Each model combined two cutting-edge AI methods: **LSTMs**, which are good at understanding time sequences, and **Transformers**, which excel at spotting patterns across many inputs at once and form the basis of modern language models like ChatGPT. These networks were trained on thousands of simulated interactions across a wide range of energies, both with and without realistic noise added.

The results were very promising. At higher energies, the best-performing model could locate events to within a few millimetres – about the width of a grain of rice – and classify event types with near-perfect accuracy. It also estimated the energy deposited with only small errors. Notably, it was found that the model trained to solve all three tasks simultaneously was more robust to noisy data than the simpler models, suggesting that **multi-task learning** made it better at focusing on important features.

However, challenges remained at the lowest energies, where the signals were often so weak that even a superfluid helium detector struggled to record them. In such cases, the model sometimes defaulted to guessing that the event happened near the centre of the detector – a reasonable assumption when little information is available.

This project shows that by using advanced AI techniques, the ability to detect tiny signals from dark matter interactions can be significantly improved. Such approaches could play a vital role in helping answer one of the deepest questions in science: what is the universe really made of?

Abstract

Light dark matter (LDM) detection demands new experimental strategies due to the small energy depositions expected. The new Direct search Experiment for LDM (DELight) employs superfluid ^{4}He to exploit multiple detection channels, enabling sensitive searches for LDM-nucleon interactions. A novel hybrid model, combining a Long Short-Term Memory network (LSTM) with a Transformer network, was developed to simultaneously perform three key tasks: position reconstruction, recoil type classification (electronic vs nuclear), and energy regression. Using Monte Carlo simulations of DELight's multi-channel detector signals, three model variants were trained and evaluated under both idealised and noise-injected conditions. Results demonstrated sub-centimetre position reconstruction at high energies, consistent near-perfect classification for energies ≥ 100 eV, and energy regression with biases below 15%. The multi-task framework was found to act as an effective regulariser, improving model robustness. Low-energy nuclear recoils remained challenging due to intrinsic signal sparsity. The models showed strong potential for future experimental deployment, with proposed improvements including repeated training runs for uncertainty quantification, loss balancing, and data augmentation.

Contents

1 Introduction

1.1 Dark Matter

The standard cosmological model, based on Einstein's general relativity, explains much of the universe's structure but fails to account for observations such as the anomalously high rotational velocities of stars at galactic edges [1]. Newtonian mechanics predicts declining orbital speeds with distance, yet observed flat rotation curves imply unseen "dark matter". This discrepancy is further supported by gravitational lensing observations revealing excess mass [2].

Modified Newtonian Dynamics (MOND) proposes modified gravitational behaviour at low accelerations as an alternative to dark matter [3]. While successfully explaining galactic rotation curves, MOND fails to describe larger-scale phenomena like galaxy clusters and contradicts cosmic microwave background observations, while remaining incompatible with general relativity.

The MACHO hypothesis alternatively attributes missing mass to compact baryonic objects like brown dwarfs or black holes [4]. However, microlensing surveys demonstrate these cannot constitute most dark matter [5], with baryon abundance from nucleosynthesis and CMB data further excluding this possibility [6].

The prevailing ΛCDM model incorporates cold dark matter (CDM) and dark energy (Λ), where weakly interacting CDM particles provide gravitational scaffolding for structure formation [7]. This framework explains both galactic dynamics and large-scale structure while predicting CMB anisotropies and lensing observations [8, 9]. As the most comprehensive model accommodating diverse cosmological phenomena, ΛCDM remains dominant despite the unresolved nature of dark matter particles.

The landscape of dark matter detection has undergone a significant evolution as traditional weakly interacting massive particle (WIMP) searches continue to yield null results. Liquid xenon experiments like XENONnT [10], LUX-ZEPLIN [11], and PandaX [12] have pushed WIMP-nucleon cross-section sensitivities to unprecedented levels, now probing below 10^{-48} cm^2 for masses above 10 GeV. Similarly, cryogenic detectors such as SuperCDMS [13] and CRESST-III [14] have extended sensitivity to lower WIMP masses in the GeV range, while indirect detection efforts through gamma-ray (Fermi-LAT [15]) and neutrino (IceCube [16]) observations have constrained annihilation cross-sections across multiple orders of magnitude. Despite these remarkable achievements, the persistent absence of a definitive signal has necessitated a fundamental re-evaluation of detection strategies.

One compelling explanation for dark matter's elusiveness is that its mass could lie far below the GeV-scale predictions of traditional WIMP models. This Light Dark Matter (LDM) scenario, characterised by sub-GeV masses, has gained traction as experiments continue to exclude canonical WIMP parameter space. While LDM interactions with nuclei would produce detectable signals in principle, their extremely low recoil energies challenge conventional detectors, which lack the sensitivity to resolve such faint signatures. Recent theoretical advances, however, suggest that novel detection techniques could bridge this gap. The Direct search Experiment for LDM (DELight) exemplifies this effort, employing superfluid helium to probe LDM-nucleon interactions at previously inaccessible energy thresholds [17]. Figure 1 illustrates the current landscape of spin-independent scattering limits, highlighting DELight's projected sensitivity in this unexplored regime.

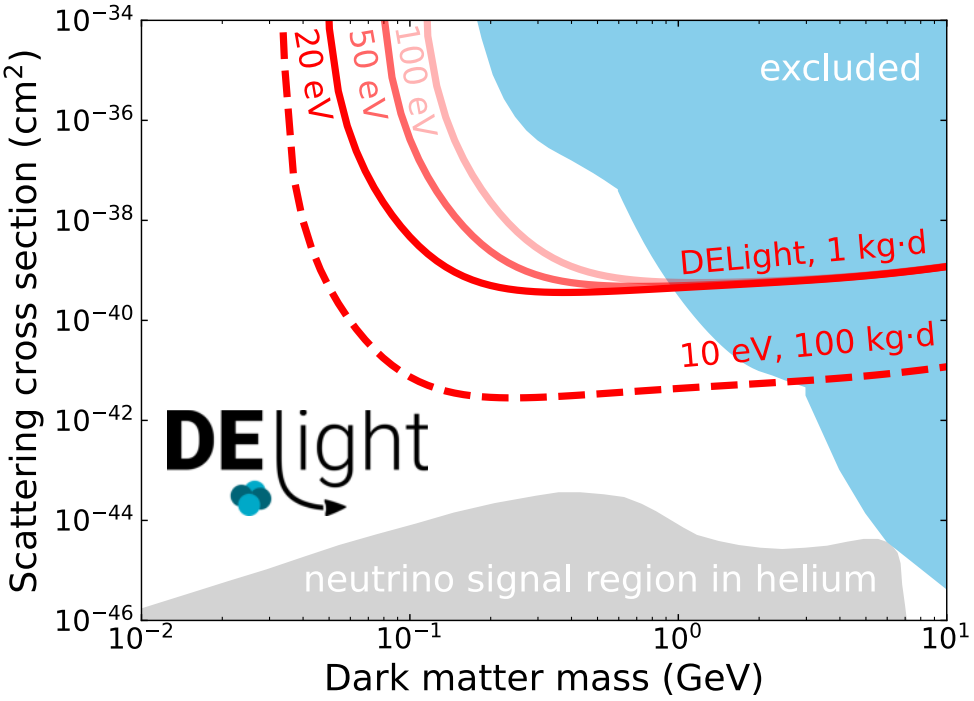


Figure 1: Projected sensitivity of the DELight experiment to spin-independent dark matter-nucleon scattering, highlighting expected limits for various phases and thresholds, alongside existing exlusions from the parameter space (in blue) from CRESST-III [14], DarkSide [18], and XENONnT [10]. (DELight internal)

1.2 Superfluid ^{4}He and DELight

Superfluid ^{4}He (LHe) has emerged as a particularly promising target medium for light dark matter (LDM) detection due to its unique quantum properties and multiple measurable signal channels. The superfluid phase of ^{4}He, occurring below 2.17 K, enables the production and propagation of quasiparticle excitations (phonons and rotons) that carry energy from particle interactions with exceptionally long mean free paths [19, 20]. This property, combined with ^{4}He's low nuclear mass (providing better kinematic matching for low-mass dark matter particles), intrinsic radiopurity (impurities freeze out at mK temperatures), and absence of long-lived radioactive isotopes, makes it ideally suited for probing the sub-GeV mass range inaccessible to conventional dark matter detectors.

When energy is deposited in LHe through particle interactions, it is partitioned into four distinct channels: quasiparticles (phonons and rotons), ultraviolet (UV) photons from singlet excimer decays, infrared (IR) photons from higher excited states, and long-lived triplet excimers. Figure 2 illustrates the response to incoming particles. Figure 3 displays the distributions obtained from Monte Carlo simulations of the theory [21]. The relative energy distribution among these channels depends critically on whether the interaction is caused by a nuclear recoil (NR) or electronic recoil (ER) event. NR events, such as those expected from dark matter scattering, produce dense ionisation tracks that enhance Penning quenching —- a process where excited helium atoms (He and He_2) are deactivated through collisions with nearby species, reducing the yield of scintillation photons while leaving quasiparticle production largely unaffected. ER events, typically from background electron or gamma interactions, produce more sparse ionisation with less quenching, resulting in higher UV and triplet excimer yields [21]. This differential partitioning enables powerful background discrimination.

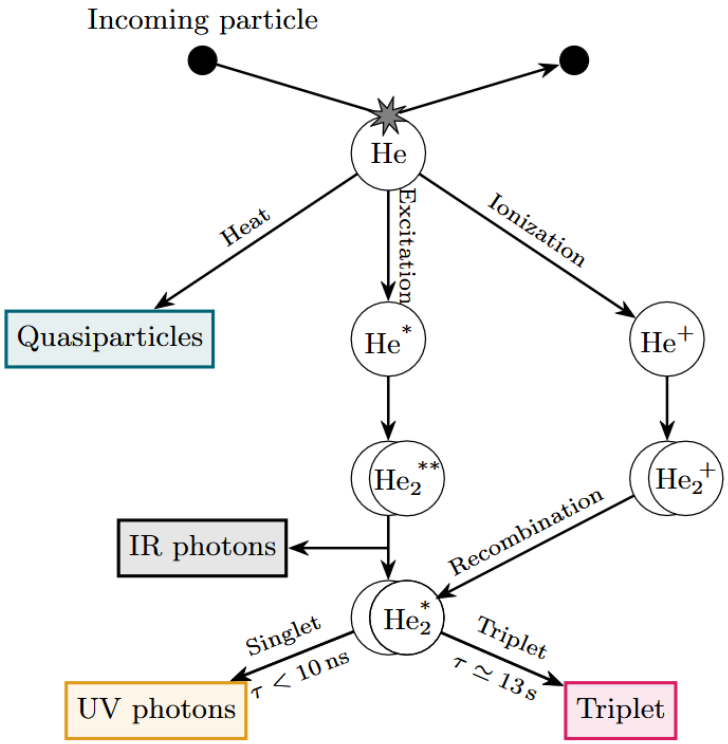


Figure 2: Schematic of LHe response to an incoming particle. Energy deposition leads to heat (quasiparticles), excitation (He), and ionisation (He^+). Subsequent processes include the formation of excimers (He_2*, He_2**), emission of IR and UV photons, and generation of long-lived triplet states. Adapted from [21].

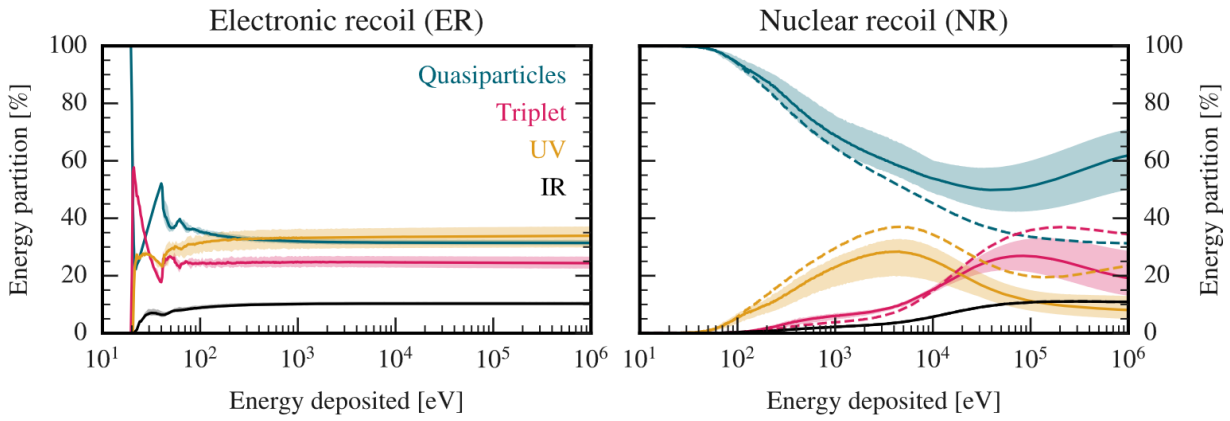


Figure 3: Energy partitioning in superfluid helium for electronic (ER, left) and nuclear recoils (NR, right) as a function of deposited energy. Adapted from [21].

The choice of ^{4}He over ^{3}He is motivated by several key factors. While both isotopes become superfluid at sufficiently low temperatures, ^{4}He's zero nuclear spin eliminates spin-dependent backgrounds that would complicate signal interpretation in ^{3}He. Additionally, ^{4}He's larger atomic mass provides better kinematic matching for the expected mass range of light dark matter candidates. The superfluid properties of ^{4}He also allow for longer quasiparticle propagation lengths compared to ^{3}He, enhancing detection efficiency. Crucially, ^{4}He's first nuclear excited state at ~ 21 MeV means all energy depositions below this threshold produce only the four aforementioned signal channels without nuclear excitation backgrounds [22].

DELight has been specifically designed to exploit LHe properties for LDM detection, employing magnetic microcalorimeters (MMCs) in a dual configuration where 20% of sensors above

the surface detect evaporated helium atoms and UV photons, while 80% measure submerged scintillation and triplet excimers [17], enabling 3D event reconstruction via timing correlations (Figure 4). Unlike TES-based approaches like HeRALD [23], MMCs provide superior energy resolution (sub-eV thresholds for X-rays) through paramagnetic sensors with SQUID readouts [24], optimised for surface detection of evaporated atoms. Operating below 20 mK to minimise noise, each MMC couples a particle absorber to a paramagnetic sensor in a weak magnetic field, translating energy depositions into magnetisation changes [25].

The quasiparticle channel provides particularly crucial information for energy reconstruction. When quasiparticles with energies ≥ 0.8 meV reach the liquid surface, they evaporate helium atoms with $\sim 30\%$ efficiency [20, 26]. These atoms are then detected by the above-surface MMCs through their adsorption energy on silicon substrates (~ 10 meV/atom), providing natural signal amplification. Each keV of recoil energy produces approximately 100 evaporated atoms, enabling sensitivity to very low energy depositions. The time delay between prompt scintillation (singlet excimers, $\tau \approx 1$ ns) and the arrival of evaporated atoms (determined by quasiparticle propagation velocities of 150-200 m/s for rotons and 238 m/s for phonons [20, 26]) constrains the interaction depth within the detector.

The UV and triplet excimer channels provide complementary information for particle identification. Singlet excimers decay promptly, emitting 15-16 eV UV photons that are efficiently detected due to LHe's transparency to its own scintillation light [27]. Triplet excimers, with their much longer lifetime ($\tau \approx 13$ s), propagate ballistically through the superfluid before decaying at surfaces or being detected directly by submerged sensors. The ratio of UV to triplet signals differs characteristically between NR and ER events due to the aforementioned Penning quenching effects in dense NR tracks [21].

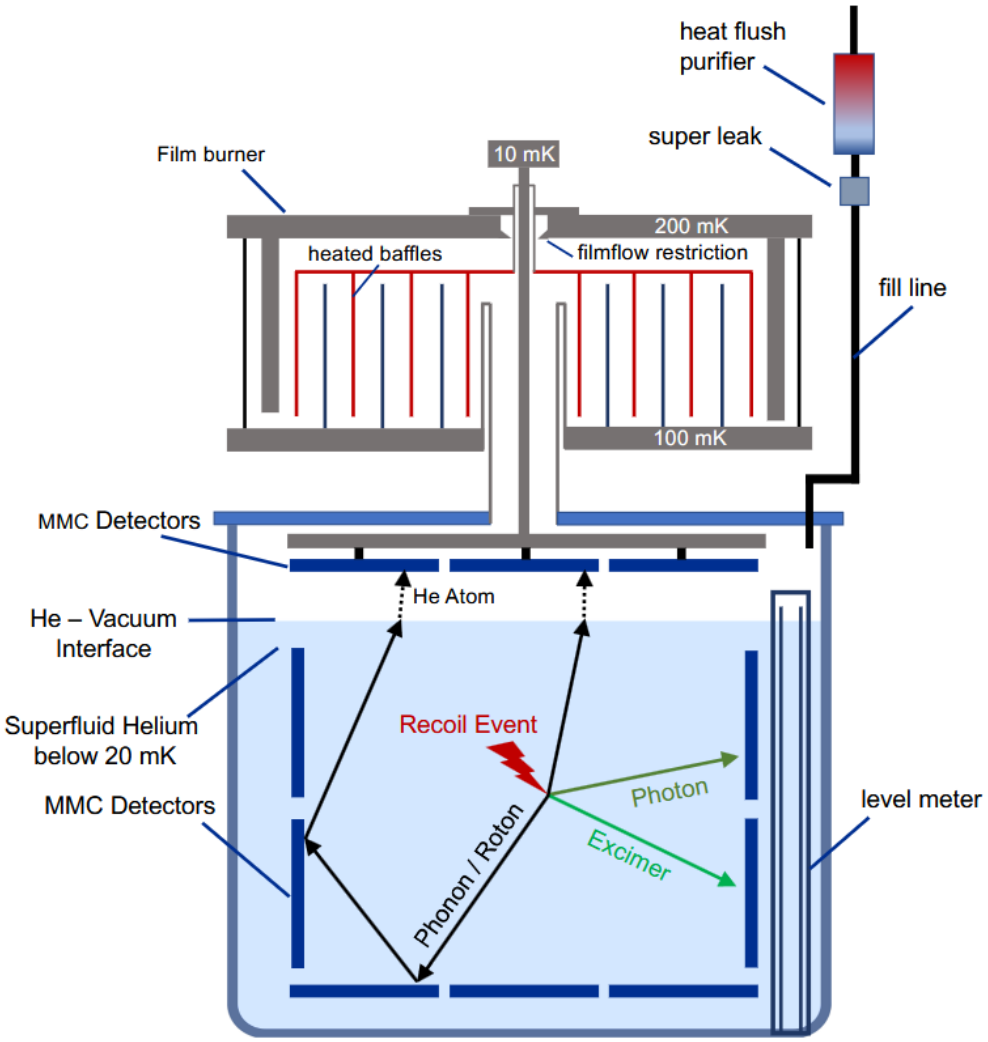


Figure 4: Preliminary design of the DELight LHe cell. The primary signal pathways are also depicted. Adapted from [17].

The expected dominant background in DELight comes from electron recoils caused by Compton scattering of environmental gamma rays. However, the multi-channel detection scheme provides several handles for background suppression [17]. The different ionisation densities of NR and ER events lead to characteristic differences in the ratio of quasiparticle to photon signals, the singlet-to-triplet excimer ratio, and the temporal development of signals. Surface backgrounds can be rejected through position reconstruction using the timing differences between prompt scintillation and delayed evaporation signals. Additionally, the pure LHe target inherently suppresses many radioactive backgrounds through its self-cleaning properties at cryogenic temperatures.

Monte Carlo simulations of signal partitioning in superfluid helium-4 reveal important features for LDM detection. Below the first excitation threshold of 19.82 eV, all energy goes into quasiparticles, but this is not a significant limitation since no gamma-induced ER events occur at these low energies. Between 20-200 eV, NR events show a gradual decrease in quasiparticle fraction from 100% to $\sim$ 90%, while ER events exhibit pronounced peaks in UV and triplet yields corresponding to discrete atomic excitation levels. Above 200 eV, Penning quenching in NRs suppresses UV and triplet excimer signals by promoting non-radiative de-excitation, thereby increasing the quasiparticle yield. This effect grows with energy as higher excimer densities enhance quenching, redirecting energy from photons to quasiparticles, particularly above 5 keV, where ionisation processes dominate [21]. This is illustrated in Figure 3.

The energy-dependent differences in signal partitioning between NR and ER events allow DELight to maintain high sensitivity while rejecting backgrounds. For a 10 keV energy deposition, typical NR events might distribute energy as $\sim 60\%$ quasiparticles, $\sim 20\%$ UV, and $\sim 20\%$ triplets, while ER events would partition more evenly with $\sim 30\%$ quasiparticles, $\sim 30\%$ UV, and $\sim 30\%$ triplets (the remainder going to IR in both cases) [21]. These distinct "fingerprints" enable statistical separation of potential dark matter signals from backgrounds. The eventual data from DELight's operational phase will provide a critical test of these theoretical predictions, requiring sophisticated analysis to correlate observed signal partitions with energy depositions and interaction types.

1.3 Motivation

The analysis of these multi-channel signals aims to achieve two primary objectives: energy regression and ER/NR discrimination. Accurate energy regression will help constrain the energy range of potential dark matter interactions, while ER/NR discrimination is critical for identifying NR events, which are the primary signature of dark matter-nucleus scattering. In the absence of a dark matter detection, this analysis will still provide stringent limits on dark matter interactions by distinguishing background ER events from potential NR signals. Additionally, position reconstruction of events is valuable for background discrimination and detector characterisation. By mapping the spatial distribution of events, it can help distinguish between neutron-induced nuclear recoils (which may cluster near detector walls or shielding due to external sources) and potential dark matter signals (which would be uniformly distributed). Furthermore, position information can identify localised background sources, such as radioactive contaminants or surface interactions, enabling targeted mitigation strategies.

This project explores the use of machine learning (ML) to develop a unified model for the DELight experiment, capable of performing position reconstruction, energy regression, and ER/NR classification in an end-to-end framework. By training on multi-channel detector responses, leveraging timing correlations for position, energy partitioning for particle identification, and combined channel amplitudes for energy reconstruction, the model can simultaneously optimize all three tasks. This approach is critical for distinguishing NRs (with higher quasiparticle yields) from ERs (dominated by UV/IR signals), as their distinct multi-dimensional signatures are inherently captured by the ML architecture. A key advantage of this method is its adaptability to detector-specific variations, such as noise or non-uniform channel responses, without requiring manual recalibration. Since real DELight data is not yet available, the model is developed and validated using Monte Carlo simulations of expected waveforms, ensuring readiness for future experimental deployment.

While previous experiments have utilised convolutional neural networks (CNNs) [28] and deep neural networks (DNNs) [29], the models developed in this project employ a Transformer network for its self-attention mechanism, which has proven useful in large language models and in modelling physical systems [30]. In the context of this project, the Transformer is used to establish connections and patterns between the different detectors. Additionally, a Long Short-Term Memory network (LSTM) is also used for the extraction of timing relationships between the channels. This project presents a novel solution to this multi-channel analysis problem – an LSTM-Transformer architecture. While this architecture has found use in other fields [31, 32], it has not yet been applied to the particular field of dark matter experiments. The process of building and testing this LSTM-Transformer architecture is described in detail in the following section.

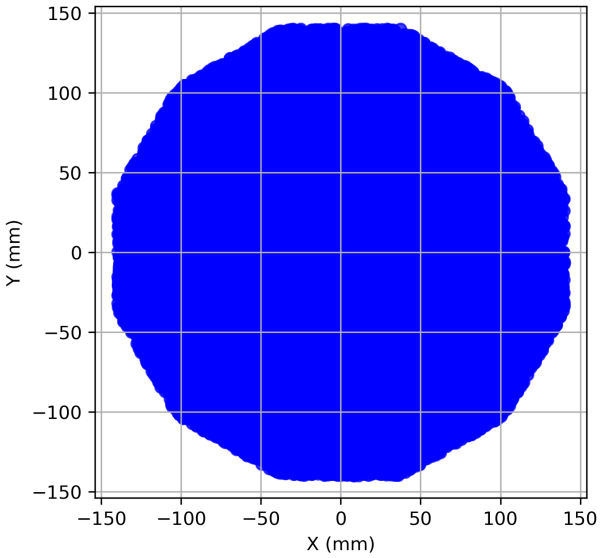
2 Computational Methods

2.1 Data

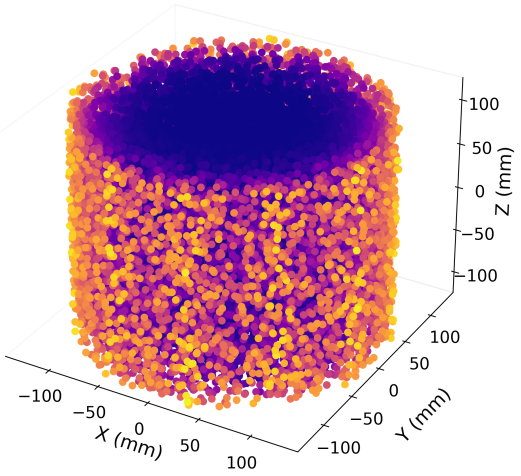
The data used consists of simulated events represented as 54 parallel detector readouts, each sampled over a 64 millisecond time window at a frequency of 256 Hz. This gives 16384 time steps per channel or a sequence length of 16384. The 54 channels correspond to individual MMCs, of which 9 are positioned above the superfluid surface. These upper detectors are the only ones that can detect the quasiparticle channel via helium atom evaporation. The remaining 45 detectors are submerged and primarily capture UV scintillation and triplet excimer signals. Because of this geometric configuration, and the differing propagation times and coupling mechanisms of these signal channels, the waveforms contain distinct spatial and temporal signatures. These signals are of either NR or ER events, covering a range of discrete energies –- 50 eV, 100 eV, 500 eV, 1000 eV, 10000 eV and 100000 eV. Each file contains 128 such events of each type and has the types specified for each event, so that the data is arranged into tensors of shape (128, 54, 16384) per event type (256 events per file). During loading, the events were assigned class values of 0 and 1 for ER and NR respectively.

The files also contain the positions and energies of the events. The event positions are confined to a volumetric cell defined by an x-y range of ± 140 mm (280 mm total span) and a z range of -1953 to -1713 mm (240 mm total span). The xy cross-section is geometrically constrained by a dodecagonal prismatic structure, reflecting the theoretical arrangement of detectors – 36 positioned along the sides of the cell, with 9 distributed across the bottom of the cell and 9 above the LHe. This configuration results in a non-cylindrical, polygonal volume that extends uniformly along the z-axis. A plot of the xy cross-section of the positions and a 3D plot of the positions can be seen in Figure 5.

Two main datasets were used: noiseless (idealised) and noised (realistic), with the latter generated by adding Gaussian white noise to the waveforms to simulate detector backgrounds. Figure 6 displays example waveforms. Preliminary tests revealed that Fourier-transformations of the data offered no significant improvement in training efficiency or model accuracy. Thus, the models used raw data with only two preprocessing steps: centring the z-coordinate at zero and log-normalising energies to handle their wide range spanning multiple orders of magnitude.



(a) XY cross-section of simulated event positions.



(b) 3D plot of simulated event positions, with radial distance encoded by a skewed colour map (scaling as r^7). Yellow indicates maximal distance from the centre, while purple corresponds to points near the origin. The non-linear colour map accentuates the edges of the volume.

Figure 5: Visualisation of the spatial distribution of simulated events.

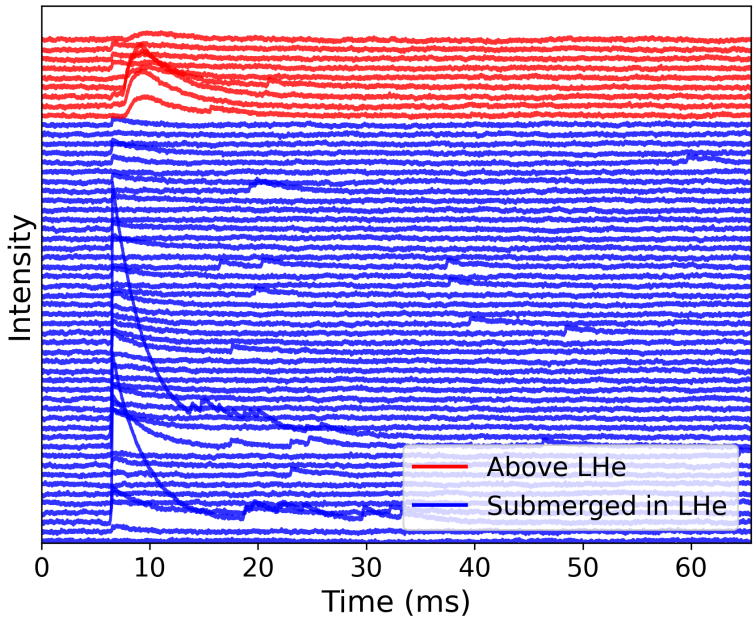


Figure 6: Plot displaying simulated waveforms from the 54-detector array for a 10 keV NR event, including noise contributions. The traces are colour-coded by detector location: red represents the 9 detectors positioned above the LHe volume, while blue corresponds to the 45 submerged detectors. The early waveform double peaks in the above-LHe detectors originate from UV/IR and quasiparticle detection, with the more rounded peak resulting from quasiparticle detection. Subsequent peaks beyond 20 ms are from triplet detection.

2.2 Model Architecture

Three models were constructed: PR (Position Reconstruction), PRC (Position Reconstruction and Classification), and PRCE (Position Reconstruction, Classification and Energy regression). This progressive approach first validated the most challenging timing-based position reconstruction, then tested classification capability, and finally integrated energy estimation to examine how multi-task learning improves performance while isolating each component's behaviour.

To learn from this waveform data, neural networks were employed to create the models. A neural network is a computational model made up of interconnected layers of artificial "neurons." Each neuron performs a simple mathematical operation – it takes one or more inputs, applies a set of learned weights to them, sums the result, adds a bias term, and passes it through a non-linear activation function [33]. Mathematically, the operation performed by a single neuron is expressed as

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right)$$

where x_i are the inputs, w_i are the weights, b is the bias, and is f the activation function. The purpose of the activation function is to introduce non-linearity, allowing the network to model complex patterns [33]. In the models made in this project, the ReLU (Rectified Linear Unit) and GELU (Gaussian Error Linear Unit) activations were the main ones used. The ReLU activation function outputs the input directly if it is positive; otherwise, it outputs zero [34]. It is expressed as

$$ReLU(x) = max(0, x).$$

It introduces non-linearity in neural networks, helping them learn complex patterns while being computationally efficient. This can cause issues when trying to predict negative values such as for position reconstruction. For this reason, it was only implemented for energy regression. The GELU activation function weights inputs by their percentile under a Gaussian distribution, smoothly approximating ReLU [34]. It is expressed as

$$GELU(x) = x\Phi(x),$$

where $\Phi(x)$ is the cumulative Gaussian distribution. The GELU activation function was chosen for both position reconstruction and classification tasks because, unlike ReLU, it permits negative values – an important feature for modeling signed spatial coordinates. For the binary classification task specifically, GELU's non-zero gradient across all inputs helps prevent dying neurons during training, which is particularly valuable when learning from binary targets (0/1 labels for ER and NR respectively).

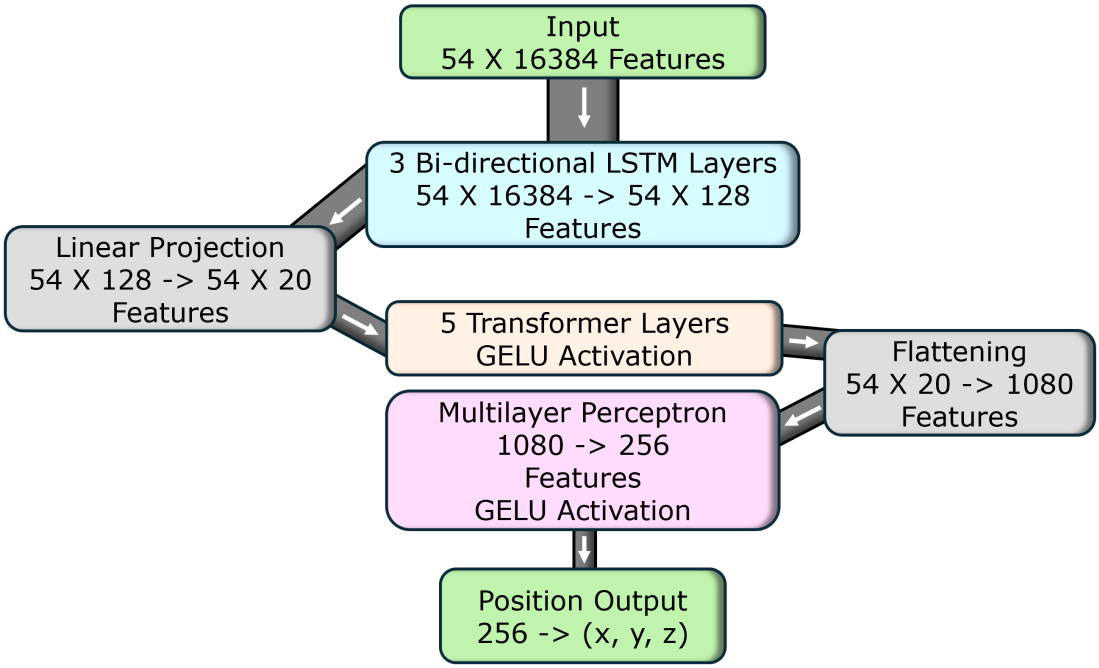


Figure 7: Schematic of the PR model. It shows the flow of a single input (one event). 54 is the number of detectors, 16384 is the sample length (features).

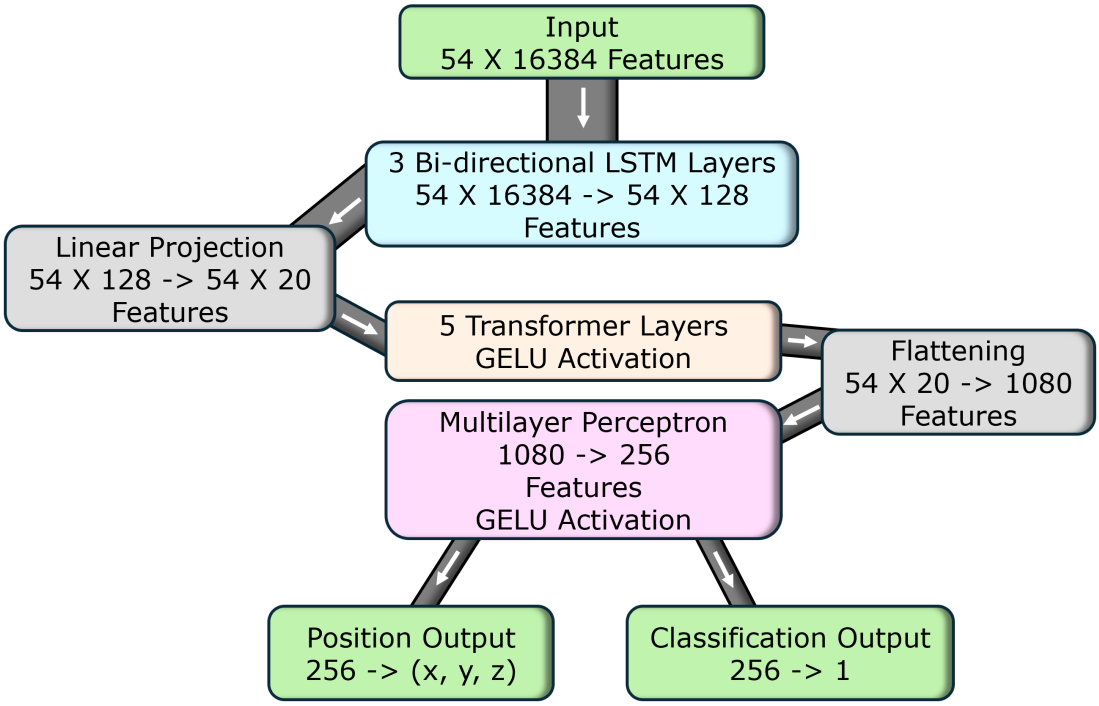


Figure 8: Schematic of the PRC model.

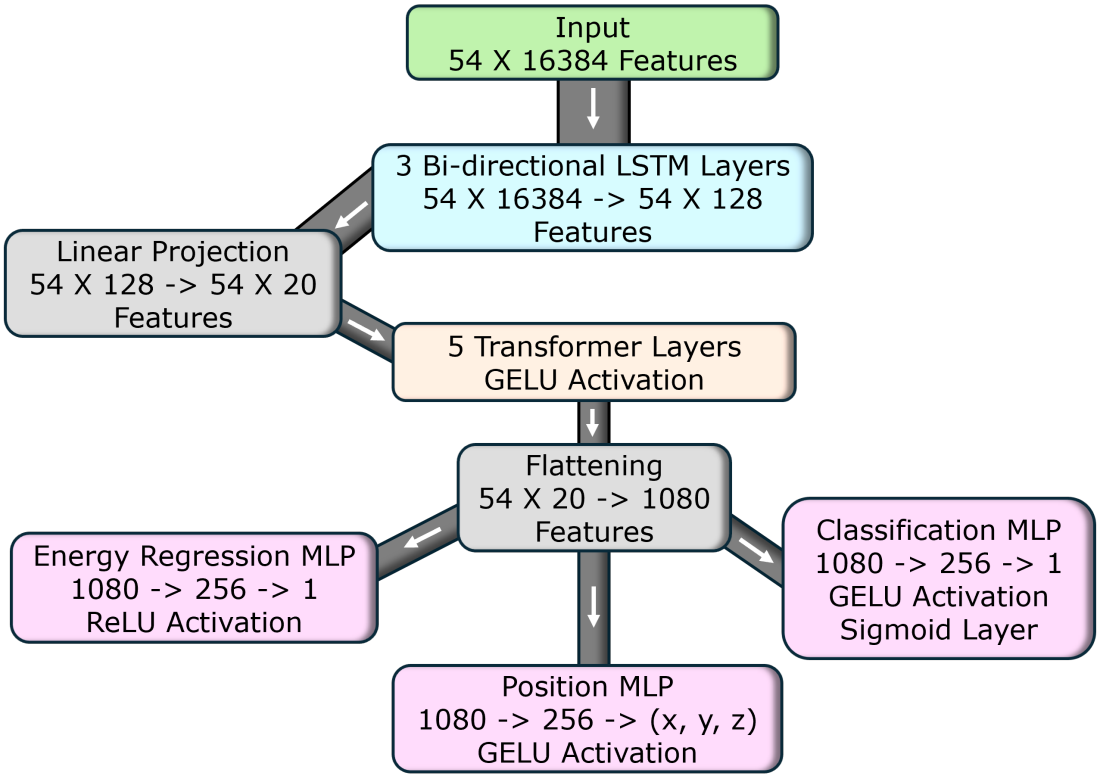


Figure 9: Schematic of the PRCE model.

All models were implemented in Python [35] using the PyTorch framework [36]. The architecture used across all tasks shares a consistent backbone combining recurrent and attention-based components – specifically, a bidirectional Long Short-Term Memory (LSTM) network followed by a Transformer block. Diagrams of each model's architecture can be seen in Figures 7, 8 and 9.

Each model begins with a three-layer bidirectional LSTM. LSTMs are recognised as a specialised type of recurrent neural network which have been designed to model sequential data effectively through their distinctive gating architecture [37]. In contrast to standard RNNs, the vanishing gradient problem is addressed by LSTMs via their cell state mechanism and three varieties of gates: input gates by which new information is regulated, forget gates through which unnecessary information is discarded, and output gates that determine what is passed to the subsequent time step [37]. This architecture renders them particularly suitable for processing the detector's high-resolution waveforms, which extend across 16,384 time steps.

The bidirectional implementation is characterised by each sequence being processed in both temporal directions – one pass moving forwards through time and another moving backwards, enabling contextual information from the complete sequence to be incorporated when any given time point is analysed. Each LSTM layer is composed of 64 units, though the bidirectional architecture results in 128 features per detector being produced, as the forwards and backwards passes generate separate outputs that are concatenated. These features are designed to capture the most significant temporal patterns from the raw waveforms while preserving information about long-range dependencies.

To prevent overfitting during training, dropout regularisation is applied with a rate of 0.3. Dropout operates by randomly deactivating a fraction of neurons during each training iteration, which compels the network to develop redundant representations and prevents excessive dependence on any individual neuron [38]. Finally, the LSTM outputs are passed through a linear projection layer where the dimensionality is reduced to 20 features per detector, thereby matching the required input size for the subsequent Transformer block.

A Transformer is a powerful architecture for sequential data that leverages self-attention mechanisms to dynamically assess relationships across the entire input sequence. Unlike traditional recurrent networks, Transformers process all sequence elements in parallel, with self-attention weights determining the relevance of each element to others—enabling the model to focus on the most informative features while preserving global context [39].

The Transformer block used in the models is composed of five stacked encoder layers. Each encoder layer consists of multi-head attention (using three parallel attention heads), followed by a position-wise feed-forward network with 128 hidden dimensions and GELU activation. Residual connections and layer normalisation stabilise training across the stacked layers, while dropout (applied at 0.2 rate) regularises activations to prevent overfitting.

The Transformer serves as a relational reasoning module, processing the LSTM-extracted waveforms from all 54 detectors. Through self-attention, it dynamically models inter-detector relationships, weighing their feature-based importance to uncover spatial patterns. Unlike isolated processing, this explicit linking enables position inference via "triangulation" of collective waveform features. The Transformer's relational reasoning can be also leveraged for energy regression, as inter-detector correlations are weighted according to their predictive relevance for deposition magnitude. The attention mechanism can highlight detector-level patterns diagnostic of recoil type (e.g., UV-dominated vs. quasiparticle-dominated signals), allowing ER/NR discrimination.

The quadratic complexity of self-attention necessitates sequence length reduction before the Transformer layer [39], hence the reduction after the LSTM. This computational optimisation improves efficiency and scalability while focusing on salient features, enhancing generalisation. Inputs are permuted to match the Transformer's expected (sequence length, batch size, detector dimension) format, then restored for downstream processing. Batch size corresponds to the number of events (in the context of this work) processed simultaneously.

The output of the Transformer is then dimensionally flattened and passed to one or more Multilayer Perceptrons (MLPs) depending on the task. An MLP is a class of feedforward artificial neural networks composed of multiple layers of interconnected neurons that use non-linear activation functions to learn complex patterns in data [40]. Each layer transforms the input through weighted connections and activations, enabling the network to approximate virtually any continuous function given sufficient hidden units and proper training.

In the PR model, the output is sent through a two-layer MLP with 256 hidden units and GELU activation, projecting to a final output of three values corresponding to the (x, y, z) coordinates of the interaction. The PRC model has a shared MLP with 256 GELU-activated units with two outputs: one for spatial coordinates and another that produces a single scalar classification logit. The PRCE model uses 3 separate MLPs (heads) – position, classification and energy regression. The position head follows the same GELU-activated structure as above. The classification head also uses GELU, followed by a sigmoid activation to convert the output into a probability. The energy regression head uses a ReLU activation to ensure that the predicted energy remains non-negative, followed by a linear projection to a scalar output. During initial training, the PRCE model would return NaN (Not a Number) loss values, which was likely caused by division by zero. Having a separate MLP for classification with a final sigmoid activation layer alleviated this issue.

2.3 Training and Evaluation

Training, validation, and testing are the three phases of evaluating a machine learning model's performance. Datasets follow typical splits of 70% for training, 15% for validation, and 15% for testing, though ratios may vary based on data size. The model learns patterns during training, the validation tunes hyperparameters and detects overfitting, while the testing provides a final,

unbiased performance assessment [41]. Hyperparameters are predefined configuration settings that control the learning process of a model (e.g., learning rate, batch size, or number of layers), which are set before training and influence how the model learns from the data. Overfitting occurs when a model learns the noise or specific details of the training data too closely, harming its ability to generalise to unseen data.

Validation and testing employ identical technical procedures [41]. The uniform energy distributions and theoretically constrained signal patterns of the datasets can allow validation metrics to reliably approximate test performance without requiring separate evaluation. Preliminary checks showed negligible validation-test performance differences, which motivated the choice of a final data split of 80% training and 20% validation without a dedicated test set. Cross-energy generalisation (e.g., training on 50 eV and testing on 100 eV) was not attempted, as energy-specific training aligns with the intended use case of matching experimental energy bins.

The PR model was trained and evaluated exclusively on NR data, with separate runs performed for each energy (50–100000 eV). In contrast, the PRC and PRCE models were trained on both ER and NR data to enable classification. While PRC was evaluated per energy, PRCE was trained jointly across all energies to optimize cross-energy regression performance. A preliminary PRCE evaluation restricted to energies $\geq$ 500 eV was conducted following the PR model's poor low-energy performance (Section 3.2). This revealed significant classification differences despite unchanged energy and position reconstruction results (Section 3.3). All models were first evaluated on noiseless data, then on noised data. This allowed a direct comparison between idealistic and realistic conditions.

All models were trained using supervised learning, a paradigm where models learn from input-output pairs to approximate the mapping between them [42]. During training, the models iteratively minimised their prediction errors through optimisation of a loss function – a mathematical measure of the discrepancy between predicted and true values [41]. This optimisation process systematically adjusted model parameters to improve performance on the three target tasks. A mean squared error (MSE) loss was applied to the predicted spatial coordinates (x, y, z) to quantify position reconstruction error. It is calculated using the formula

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{r}_i^{\mathrm{true}} - \mathbf{r}_i^{\mathrm{pred}}\|^2,$$

where $\mathbf{r}_i = (x_i, y_i, z_i)$ denotes spatial coordinates and n is the number of events in each batch. For classification, a binary cross-entropy (BCE) loss with logits was used to evaluate the PRC and PRCE models' ER/NR classification performance. This formulation allowed for numerical stability by operating directly on the model's raw logits rather than probability scores. The numerically stable formulation used in PyTorch is

$$\mathcal{L}_{\mathrm{BCE}} = \frac{1}{n}\sum_{i=1}^{n} \left[\max(\hat{y}_i, 0) - \hat{y}_i y_i + \log(1 + e^{-|\hat{y}_i|})\right],$$

where $\hat{y}_i$ are the model's raw logits, $y_i \in \{0, 1\}$ are labels. Energy regression was also supervised using MSE loss, applied to the predicted energy values. For the PRCE model these three losses were computed independently during each training iteration and then summed without explicit weighting, reflecting an assumption of equal task importance and simplifying the training objective. The PR model only used MSE loss, while the PRC model used the sum of both MSE and BCE loss.

The network weights were initialised randomly and optimised through backpropagation, computing gradients via backward loss propagation. The data structure (128 events/file) naturally aligned with batch processing – the PR and PRC models used single-file batches (128)

while the PRCE model employed double-file batches (256) for enhanced parallel computation on GPU hardware. This batch sizing balanced computational efficiency with gradient accuracy, as larger batches leverage parallel processing but may oversmooth gradients, while smaller batches preserve finer data patterns through more frequent updates [43]. Preliminary testing confirmed these sizes optimally negotiated the trade-off between efficient hardware utilisation (favoured by larger batches) and precise gradient estimation (better with smaller batches).

The calculated gradients were used by an optimizer to adjust the weights in a direction that reduced the loss. For this purpose, the AdamW optimizer was employed. AdamW combines adaptive learning rate adjustments with momentum-based updates and a decoupled weight decay mechanism, which improves generalisation and mitigates overfitting [44]. The learning rate determines the size of the steps the optimizer takes when updating the model's weights during training, where a larger rate speeds up convergence but risks overshooting optimal solutions, while a smaller rate ensures stability but may slow progress. Weight decay is a regularisation technique that discourages overly large weights in a neural network by adding a penalty term (proportional to the squared magnitude of the weights) to the loss function, helping to prevent overfitting and improve generalisation.

A relatively high weight decay coefficient of 0.04 was applied to penalise large weights and promote sparsity in parameter values. To further refine the learning process, a cosine annealing scheduler with warm restarts was applied to the learning rate to modulate it in a cyclic manner, decreasing it according to a cosine schedule over 30 epochs before resetting it to its initial value [45]. The initial value was chosen to be 3×10^{-4} after preliminary testing. The minimum learning rate was constrained to 1×10^{-8}. This approach introduced periodic variability in the learning rate, which encouraged exploration of different regions of the loss landscape and helped the optimizer to avoid shallow local minima.

To mitigate the issue of exploding gradients, a gradient clipping threshold of 1.0 was introduced. This restricted the norm of the gradients to remain below the specified value, thereby ensuring stable updates, especially in architectures containing recurrent elements. LSTM layers are particularly susceptible to unstable gradients due to their recurrent structure. Even modest clipping like the one used here protects against instability when backpropagating through long sequences (16,384 samples) or cross-detector attention, without neutering the model's ability to learn [46].

Mixed precision training is a technique that uses both 16-bit and 32-bit floating-point numbers during neural network training (16-bit for faster computation and memory savings), while critical operations (like weight updates) retain 32-bit precision to maintain stability and accuracy [47]. It was implemented using automatic casting of operations to 16-bit floating point where appropriate, combined with dynamic gradient scaling to prevent numerical underflow during backpropagation, reduce memory usage and improve computational speed. A gradient scaler was used to manage this process, scaling the loss before computing gradients and unscaling them prior to gradient clipping and optimizer updates.

At the conclusion of each training epoch, the models were switched to evaluation mode. In this mode, dropout and other stochastic layers were disabled to ensure deterministic outputs. Validation was then conducted using a separate portion of the dataset, and losses were computed without computing gradients. Validation predictions for position, classification, and energy were recorded and saved alongside the corresponding true values for subsequent analysis. One training and one validation epoch together form a full epoch. Due to the time constraints and limited availability of GPU resources the PR model was trained for 600 epochs while the PRC and PRCE models were trained for 300 epochs. If a previous training session had been interrupted, the script was designed to resume from the last completed epoch by reloading model weights, optimizer states, and previously logged loss data. Table 1 summarises the training differences between the models.

The average training and validation losses were recorded per epoch in a structured CSV file. Total losses were recorded for the three models. However, MSE and BCE losses were only recorded for the PRC model's trainings on noised data. Section 3 explains how task performances were assessed in detail.

Model	Batch Size	Loss Funtions	Epochs	NR/ER	Energies
PR	128	MSE	600	NR	One at a time
PRC	128	MSE + BCE	300	Both	One at a time
PRCE	256	MSE + BCE + MSE	300	Both	All at once

Table 1: Training differences between the models.

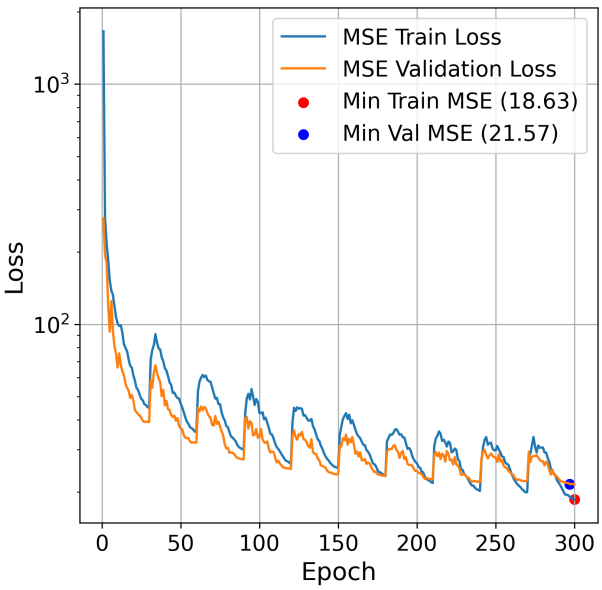
3 Results

The models' performances were evaluated using performance metrics, which measure how well a model generalises, balancing errors (e.g., false positives/negatives) to avoid overfitting, or in other words they asses predictive reliability and model confidence. Commonly used ones are accuracy, precision, and ROC curves. Sections 3.2, 3.3 and 3.4 include more information on the specific metrics used and task performances, while Section 3.1 focuses on examining the loss plots, which were made as a preliminary study on the models' performances. The position reconstruction and energy regression results from the PRCE limited energy range training on noiseless data were unremarkable and as such are omitted.

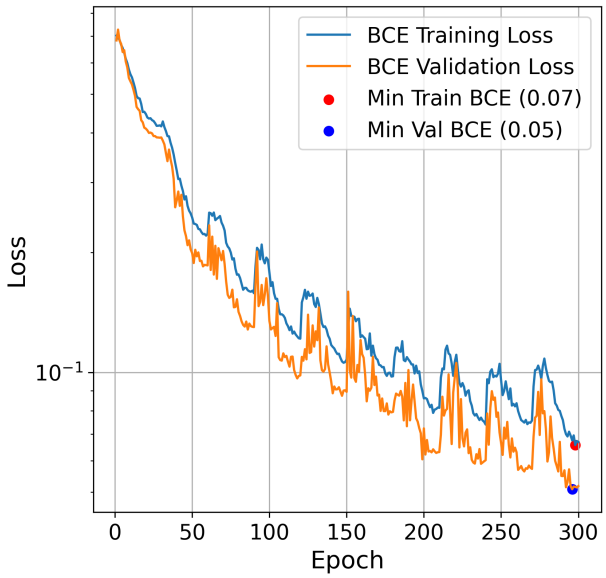
3.1 Loss

The losses observed in the figures shown in this section are MSE for position reconstruction, BCE for classification and total losses (sum of two or more losses depending on the model). Minimum values for training and validation losses are shown in red and blue respectively. Loss plots for the PR model were analysed but are omitted here to avoid redundancy, as the PRC model's performance inherently encompasses position reconstruction.

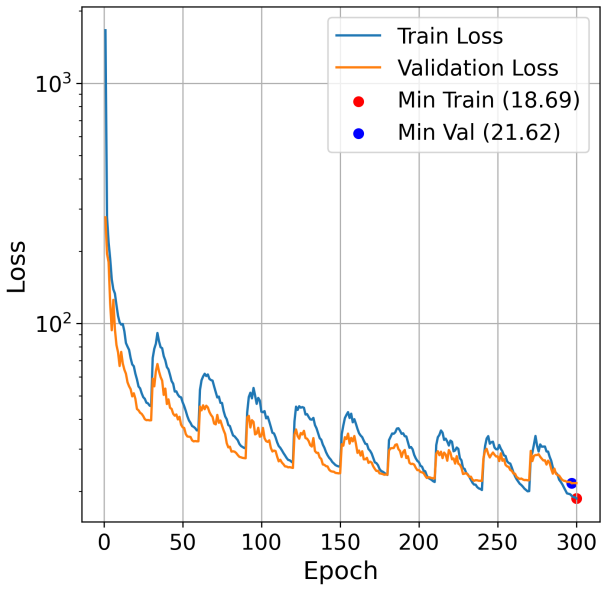
In all figures presented here, minimum values for training and validation losses are shown in red and blue respectively. Figure 10 shows that the unbalanced weighing of the separate loss functions allowed position loss to dominate. Significant overfitting was observed during training, particularly at low energies (50 and 100 eV), with even more pronounced effects in noised data (Figures 11 and 12). Notably, while position loss stagnated for noised 50 eV data, classification loss continued to decrease. Performance improved markedly at $\geq$ 500 eV as seen in Figures 13 and 14, where regularisation effects increased training loss, resulting in a lower validation loss than training loss, particularly at higher energies. Model performance generally scaled with energy. Figure 15 displays the PRCE model's loss trajectories across noiseless and noised data.



(a) MSE loss plot.

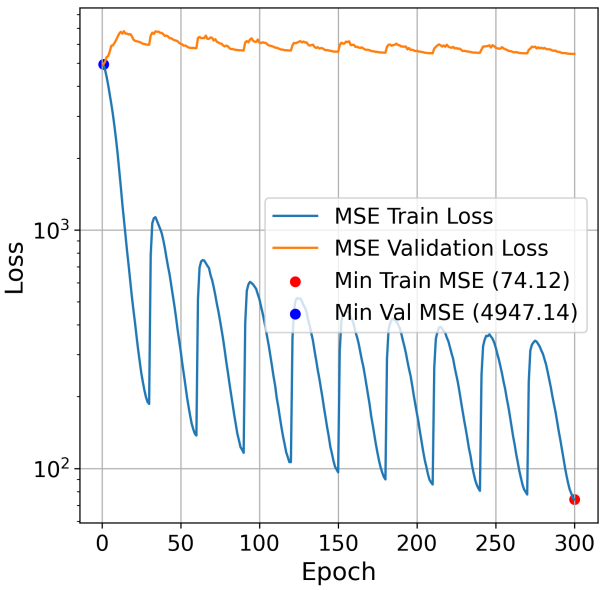


(b) BCE loss plot.

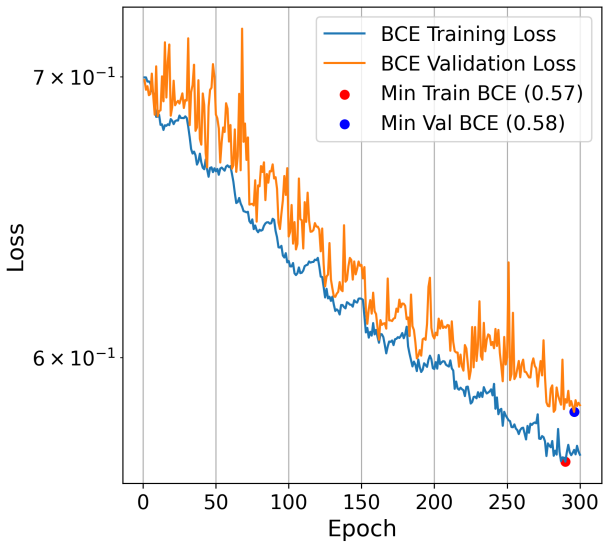


(c) Total loss plot.

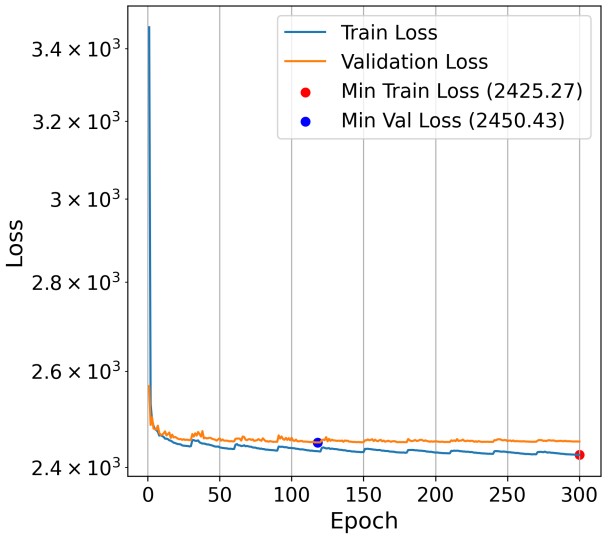
Figure 10: PRC model comparison of noiseless and noised 10 keV data performance. Note that position loss dominated. Also note that regularisation effects inflated the training losses.



(a) MSE loss plot, noised data.

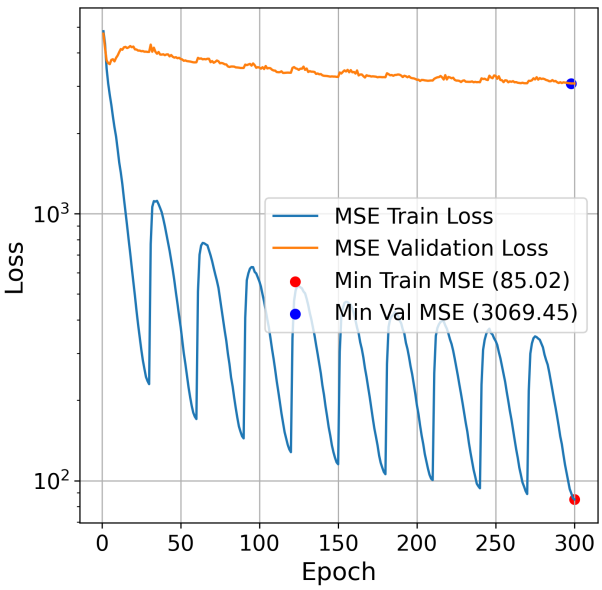


(b) BCE loss plot, noised data.

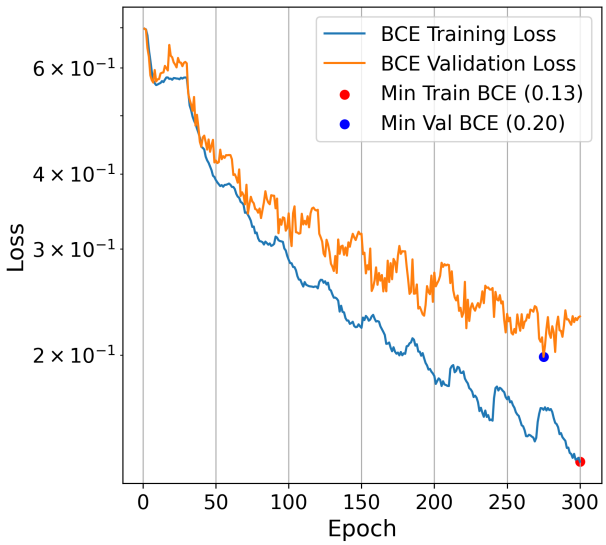


(c) Total loss plot, noiseless data.

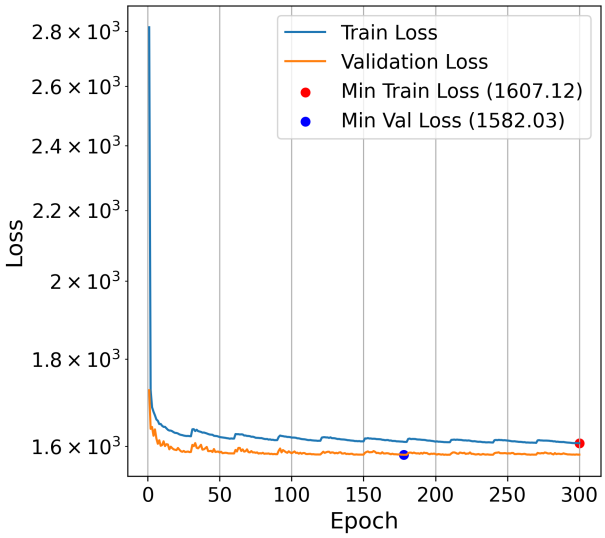
Figure 11: PRC model comparison of noiseless and noised 50 eV data performance. Severe overfitting can be observed in (a), while (b) shows minor overfitting, increasing with epoch number. Some overfitting can be observed in (c). Note the degradation in position validation loss, the lack of convergence for classification losses, and the early convergence for total loss.



(a) MSE loss plot, noised data.

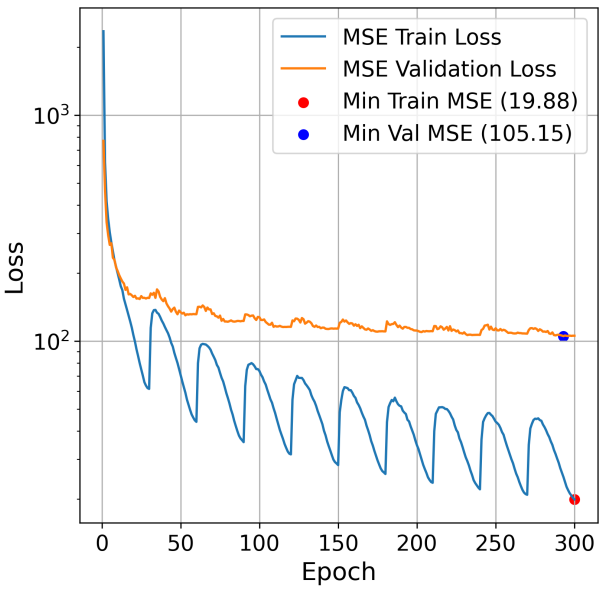


(b) BCE loss plot, noised data.

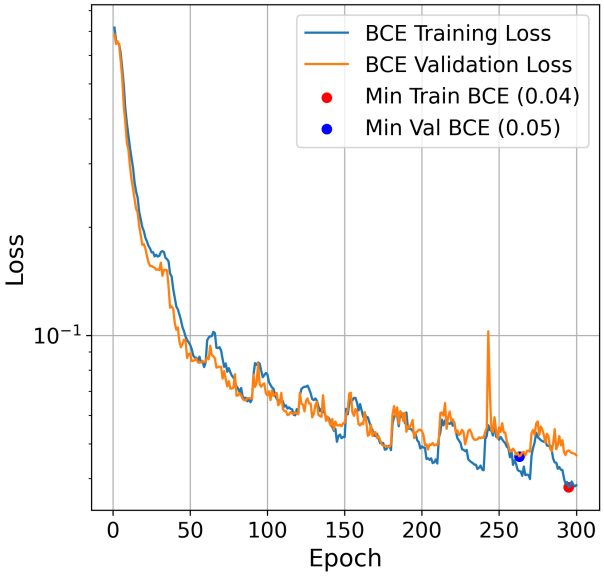


(c) Total loss plot, noiseless data.

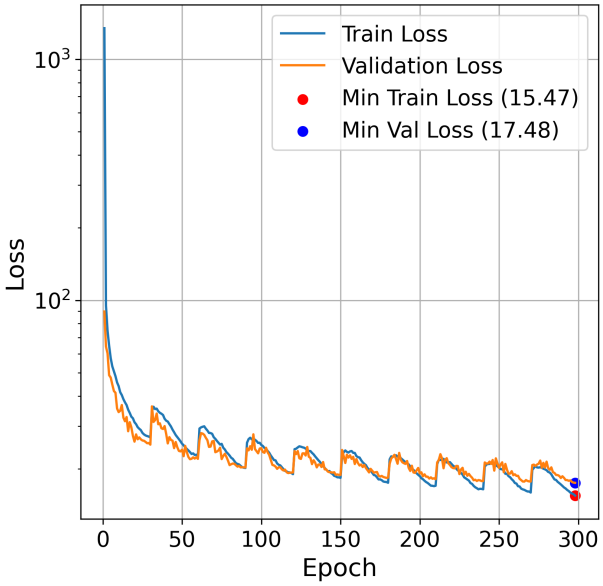
Figure 12: PRC model comparison of noiseless and noised 100 eV data performance. Severe overfitting can be observed in (a), while (b) shows an increase in overfitting with epoch number. Training losses in (c) were inflated due to regularisation. Note the overall better performance compared to 50 eV.



(a) MSE loss plot, noised data.

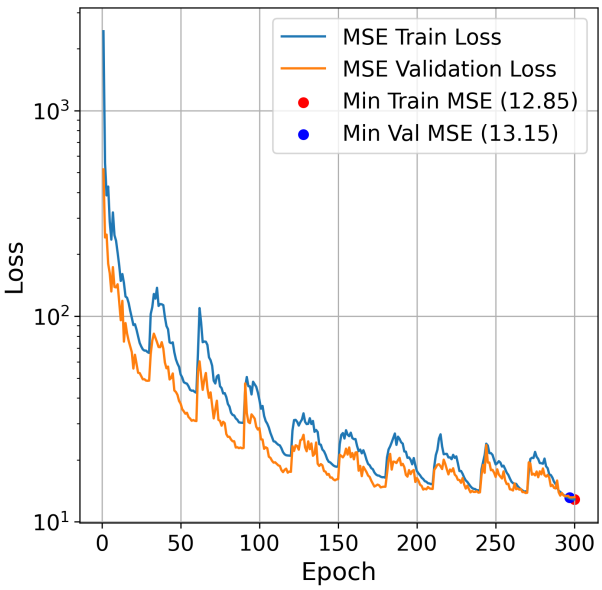


(b) BCE loss plot, noised data.

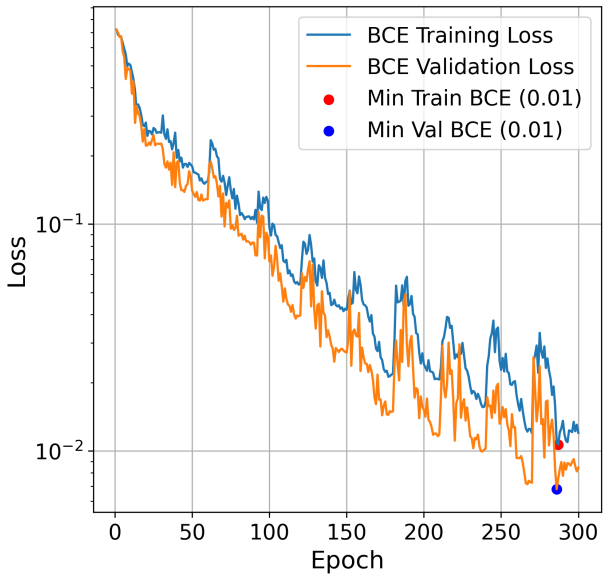


(c) Total loss plot, noiseless data.

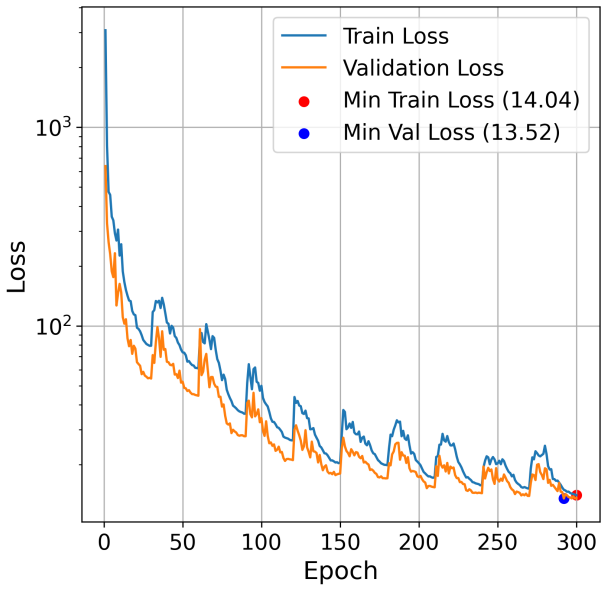
Figure 13: PRC model comparison of noiseless and noised 500 eV data performance. Less overfitting compared to lower energies can be observed for position reconstruction (a). Limited overfitting can be observed for noiseless data (c) and classification on noised data (b). Note the improved performance compared to lower energies.



(a) MSE loss plot, noised data.

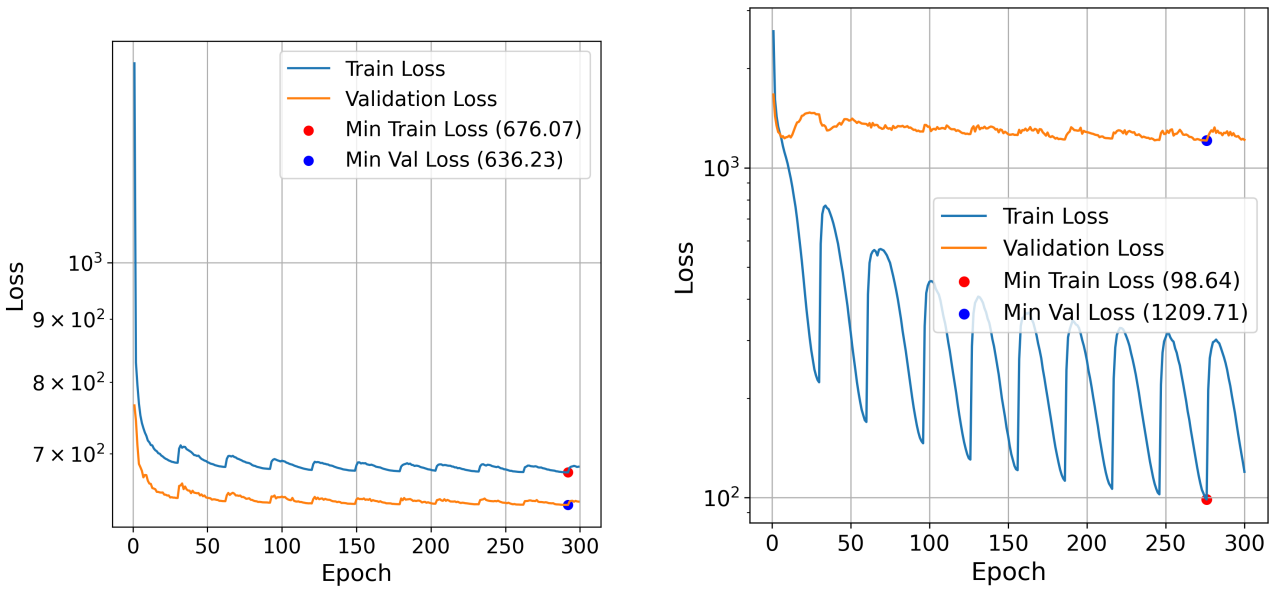


(b) BCE loss plot, noised data.



(c) Total loss plot, noiseless data.

Figure 14: PRC model comparison of noiseless and noised 100 keV data performance. Regularisation inflated training losses. Note the overall improvement in performance over lower energies.



(a) Total loss plot, noiseless data.

(b) Total loss plot, noised data.

Figure 15: PRCE model comparison of noiseless and noised data performance. Lower energies (50 and 100 eV) inflated the losses. Regularisation inflated training losses in (a), whereas severe overfitting can be observed in (b).

3.2 Position Reconstruction

The performance metrics used for position reconstruction assessment were MSE loss and Mean Euclidian Distance (MED). The Mean Euclidean Distance (MED) quantifies position reconstruction accuracy and is defined as

$$\text{MED} = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{r}_i^{\text{pred}} - \mathbf{r}_i^{\text{true}}\|_2,$$

where $\mathbf{r}_i = (x_i, y_i, z_i)$ denotes 3D coordinates for event i, and $\|\cdot\|_2$ is the Euclidean norm. The random guess distributions in the figures shown in this section represent the MEDs between true positions and points randomly generated within the LHe volume. This comparison baseline helps quantify the improvement over spatial randomness, as any meaningful reconstruction should significantly outperform random chance. The radial distances in the heat maps shown in the figures are calculated as straight-line displacements in the xy-plane between true and reconstructed positions. Bins without events in the heat maps are coloured black. The z errors are quantified as the magnitude of the difference between true and predicted positions.

Table 2 displays the lowest (best) MSE losses and MEDs by energy for both the PR and PRC models using noiseless data. The corresponding metrics for noised data are shown in Table 3. The PR model results are strictly for NRs. The performance of the PR model on energies 50, 500 and 100000 eV is visualised in Figures 16, 17 and 18, using histograms and heat maps.

Table 4 presents in detail the performance of the PRCE and PRC models, broken down per noise conditions (noiseless or noised), recoil type and energy. Figures 19 and 20 display the ER and NR performance discrepancies at 50 and 100 eV. Figure 21 visualises the position reconstruction performance of the PRCE model.

The position reconstruction performance was evaluated across energy ranges and recoil types for the PRC, and PRCE models. The PR model was evaluated across energy ranges, only for NRs. For noiseless data (Table 2), significantly higher MSE losses and MEDs were observed at

low energies (50 and 100 eV), with values decreasing sharply above 500 eV. The PRC model was found to outperform the PR model at low energies (e.g., 62.69 vs 114.72 mm MED at 50 eV), while comparable performance was achieved above 500 eV. The PRCE model (Table 4) showed marginal improvements over PRC for both ER and NR events at most energies (e.g., 5.94 vs 5.52 mm NR MED at 500 eV noiseless).

Under noised conditions (Table 3), degraded performance was observed across all energies, though the trends were smoother. The PRC model's advantage was reduced, with superior performance only seen at select energies (e.g., 80.23 vs 95.44 mm MED at 100 eV). The PRCE model demonstrated more consistent improvements over PRC under noise, particularly for ER events (e.g., 85.54 vs 116.68 mm at 50 eV), while NR performance remained similar between models.

When examined by recoil type (Table 4), ER events were reconstructed more accurately than NRs at low energies by all models. The PRCE model showed the strongest ER/NR performance gap reduction at intermediate energies (500 and 1000 eV) under noise conditions (e.g., 17.65 vs 14.88 mm MED at 500 eV). At high energies ($\geq$ 10 keV), all models achieved similar performance, with PRCE showing slight but consistent MED reductions compared to PRC.

	PR Model		**PRC Model**	
Energy (eV)	**MSE Loss**	**MED (mm)**	**MSE Loss**	**MED (mm)**
50	4822.10	114.72	2450.43	62.69
100	3137.49	77.35	1582.03	42.58
500	22.47	5.78	17.48	5.72
1000	13.64	5.79	12.97	5.68
10000	19.06	6.83	17.70	6.54
100000	13.12	5.74	13.52	5.81

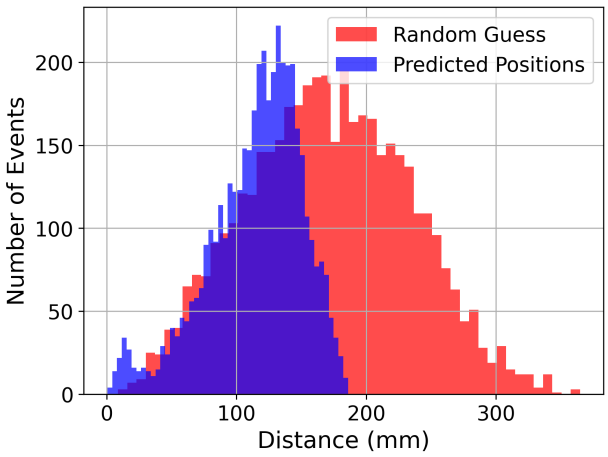
Table 2: Position reconstruction performance on noiseless data. Note the sharp increase in performance at energies $\geq$ 500 eV and the slight anomaly at 10 keV.

	PR Model		**PRC Model**	
Energy (eV)	**MSE Loss**	**MED (mm)**	**MSE Loss**	**MED (mm)**
50	4952.95	117.48	4947.14	117.30
100	4043.09	95.44	3069.45	80.23
500	83.79	13.41	105.15	15.12
1000	57.59	11.80	66.79	12.58
10000	25.83	7.89	21.57	7.27
100000	12.35	5.60	13.15	5.74

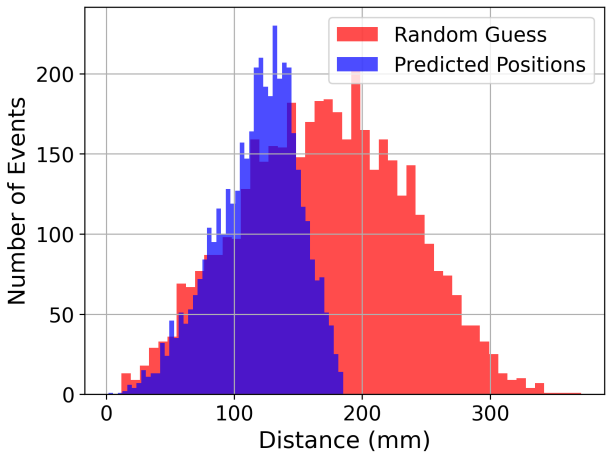
Table 3: Position reconstruction performance on noised data. Note the smoother trend in performance. Also note that the PRC model outperforms the PR model only at select energies, despite poorer performance elsewhere, suggesting an energy-dependent anomaly.

	Noiseless MED (mm)				Noised MED (mm)			
	ER		NR		ER		NR	
Energy (eV)	**PRCE**	**PRC**	**PRCE**	**PRC**	**PRCE**	**PRC**	**PRCE**	**PRC**
50	10.73	10.82	114.35	114.56	85.54	116.68	129.82	117.91
100	8.19	8.11	77.51	77.06	53.46	60.68	97.16	100.98
500	6.53	5.91	5.94	5.52	17.65	16.97	14.88	13.27
1000	6.16	6.01	5.43	5.34	13.58	13.80	12.16	11.36
10000	5.88	6.68	5.34	6.41	8.07	7.46	7.28	7.08
100000	5.28	5.86	4.98	5.75	6.10	5.77	5.94	5.70

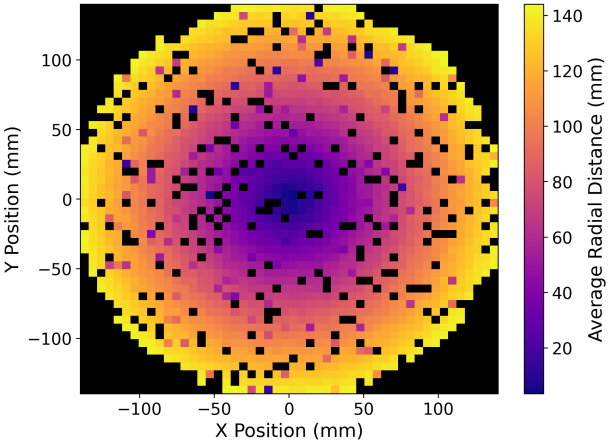
Table 4: PRCE and PRC model performance across energies and recoil types for noiseless and noised data. Note the minimum MED of 4.98 mm (PRCE noiseless NR). Note the gap in performance between ER and NR MEDs for the PRCE model.



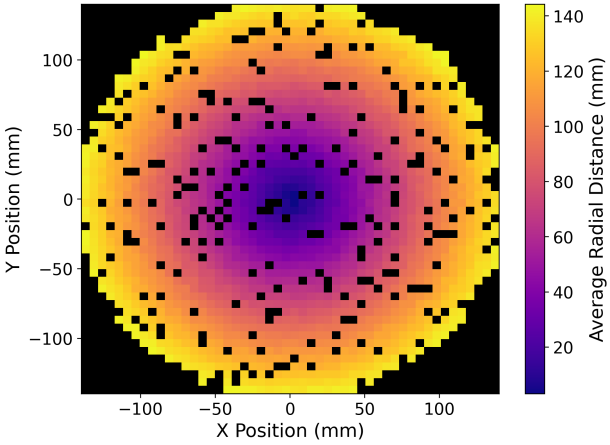
(a) Distribution of MEDs, noiseless data.



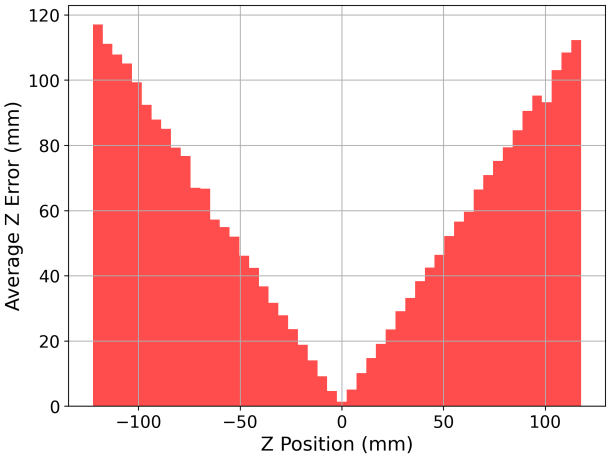
(b) Distribution of MEDs, noised data.



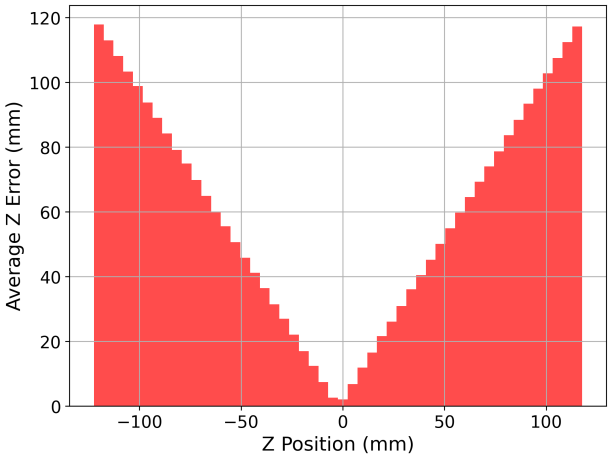
(c) Distribution of radial distances, noiseless data.



(d) Distribution of radial distances, noised data.

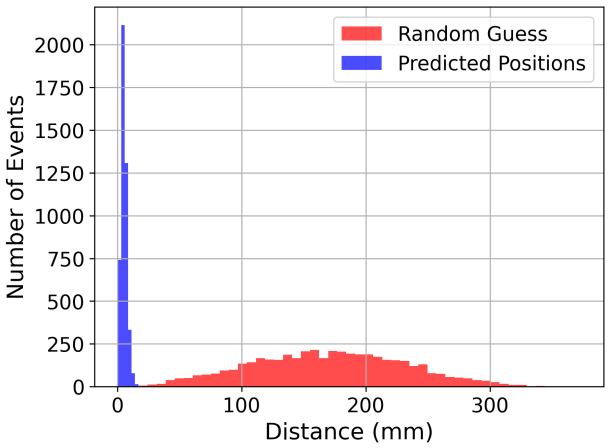


(e) Distribution of z errors, noiseless data.

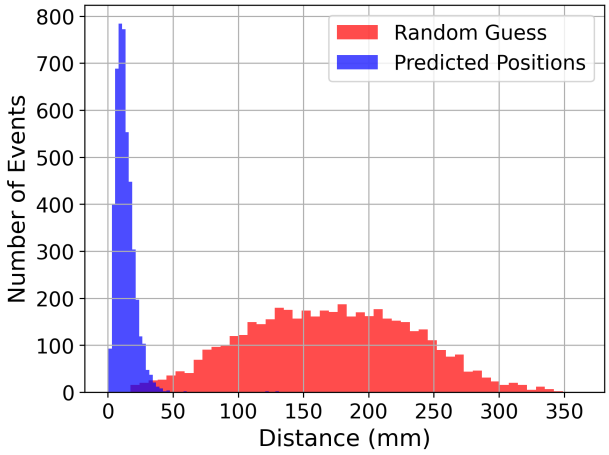


(f) Distribution of z errors, noised data.

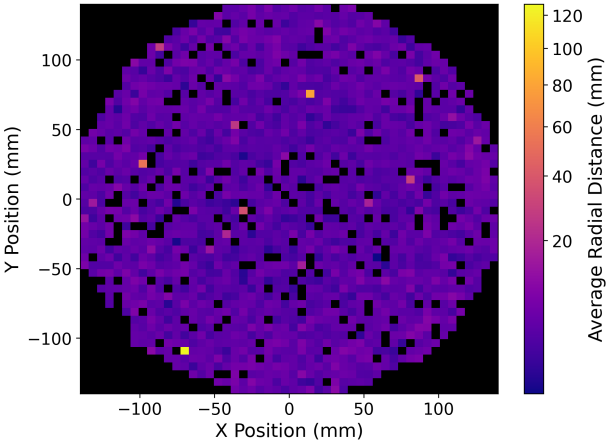
Figure 16: Comparison of PR model performance for position reconstruction on 50 eV NR data. Note the small peak near 0 in (a), the scattered darker bins in (c), and the uniformity of (f) compared to (e).



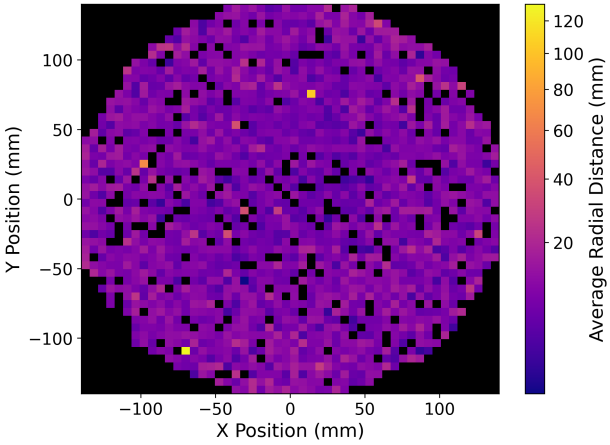
(a) Distribution of MEDs, noiseless data.



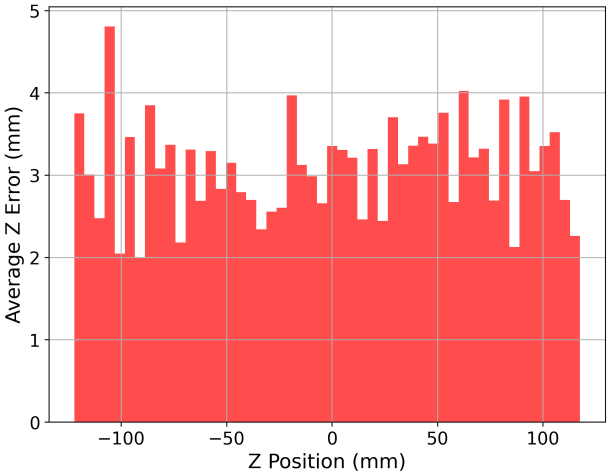
(b) Distribution of MEDs, noised data.



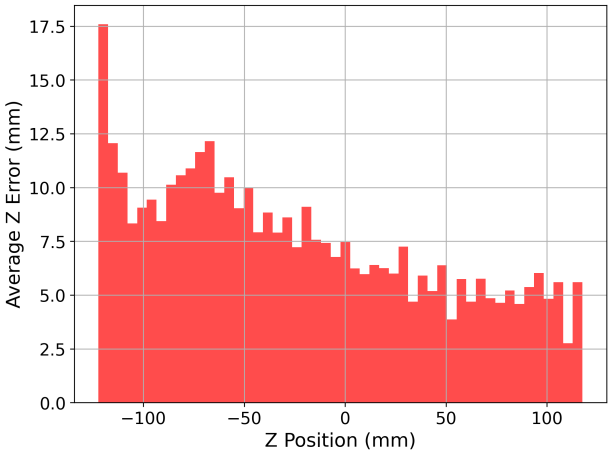
(c) Distribution of radial distances, noiseless data.



(d) Distribution of radial distances, noised data.

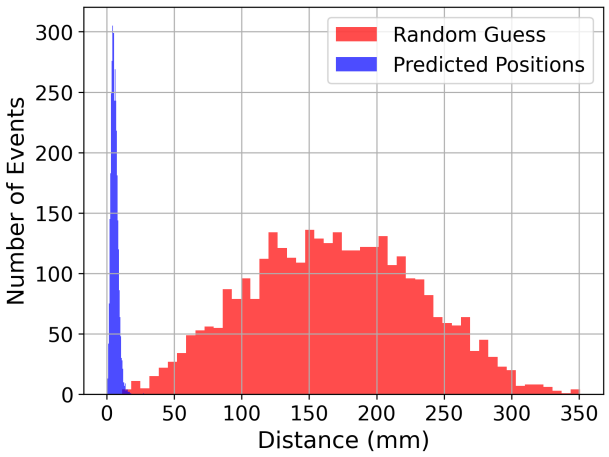


(e) Distribution of z errors, noiseless data.

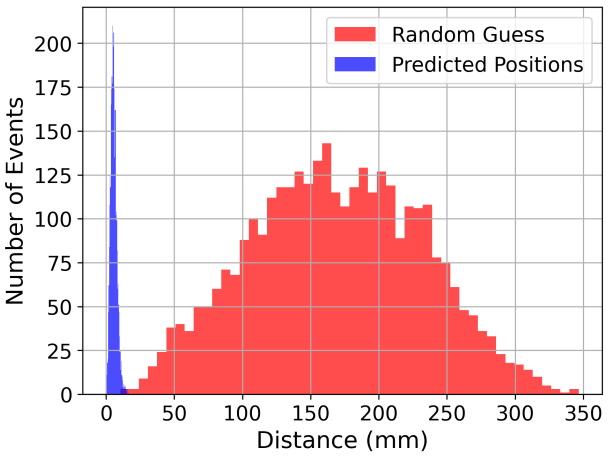


(f) Distribution of z errors, noised data.

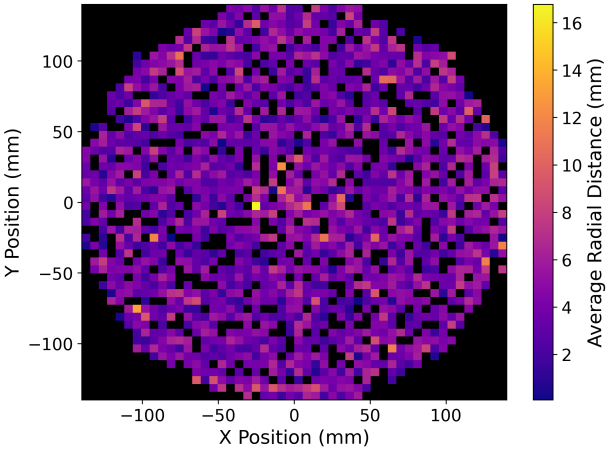
Figure 17: Comparison of PR model performance for position reconstruction on 500 eV NR data. The colour maps in (c) and (d) is scaled to highlight smaller distances, while accommodating rare outliers. Note the width of the predicted peak in (b) compared to (a), the overall colour in (d) compared to (c), and the scales of (e) and (f).



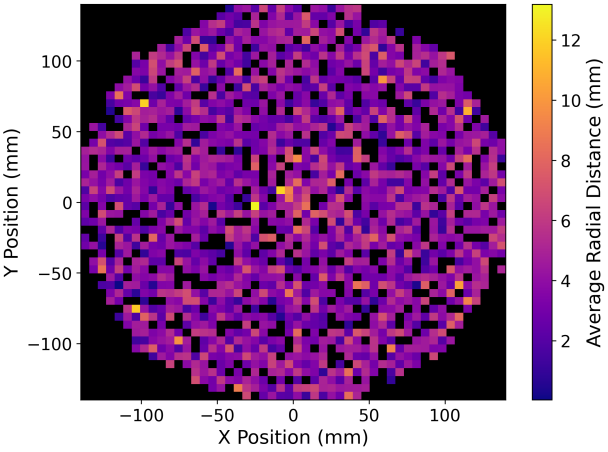
(a) Distribution of MEDs, noiseless data.



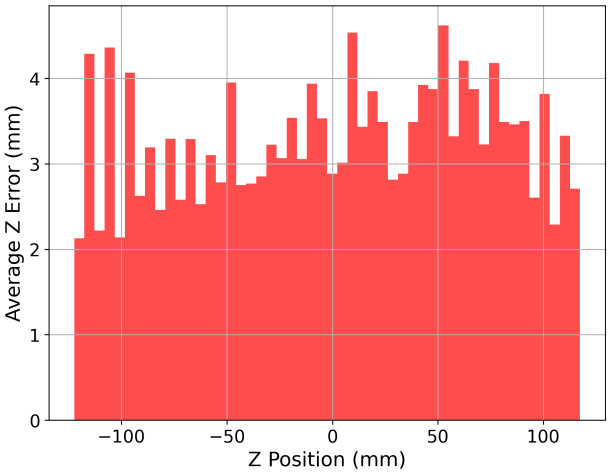
(b) Distribution of MEDs, noised data.



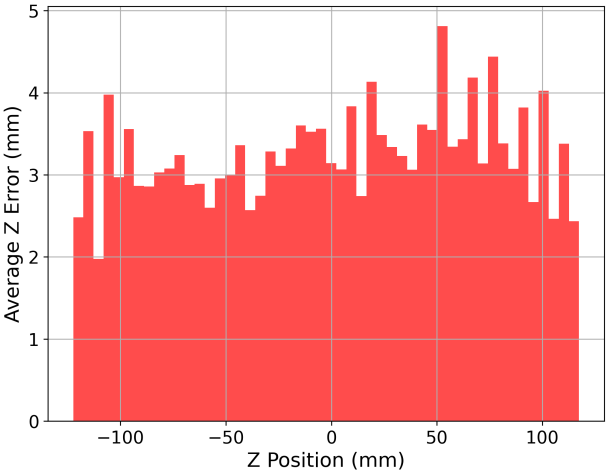
(c) Distribution of radial distances, noiseless data.



(d) Distribution of radial distances, noised data.

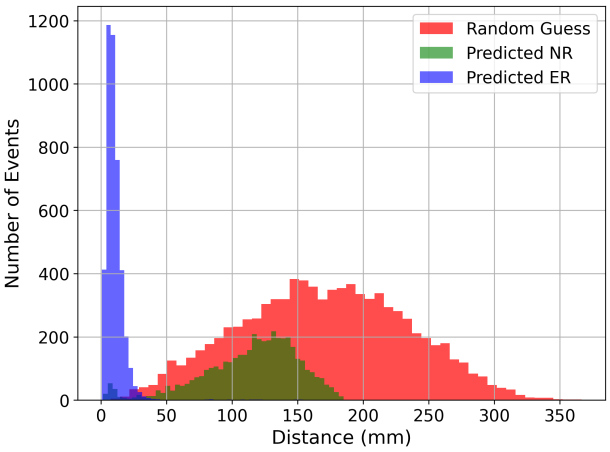


(e) Distribution of z errors, noiseless data.

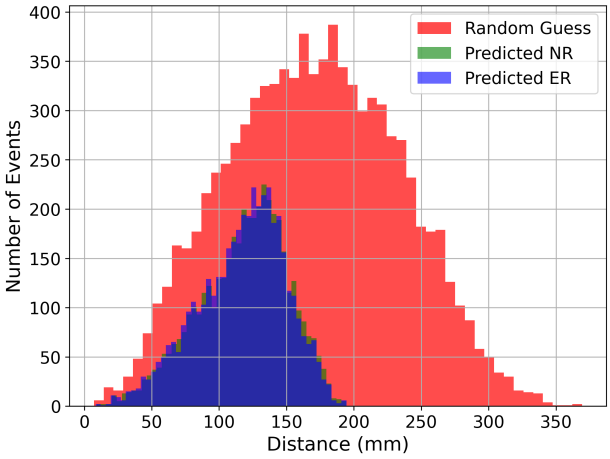


(f) Distribution of z errors, noised data.

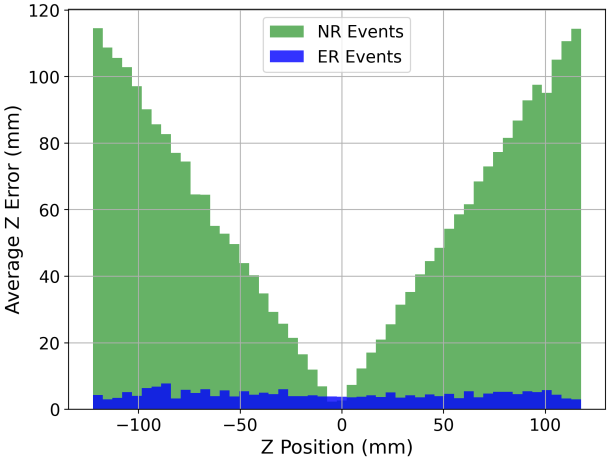
Figure 18: Comparison of PR model performance for position reconstruction on 100 keV NR data. The colour map in (c) and (d) is scaled to highlight smaller distances, while accommodating rare outliers. Note the bin widths in (a) and (b), the overall colour in (d) compared to (c), and the scales of (e) and (f).



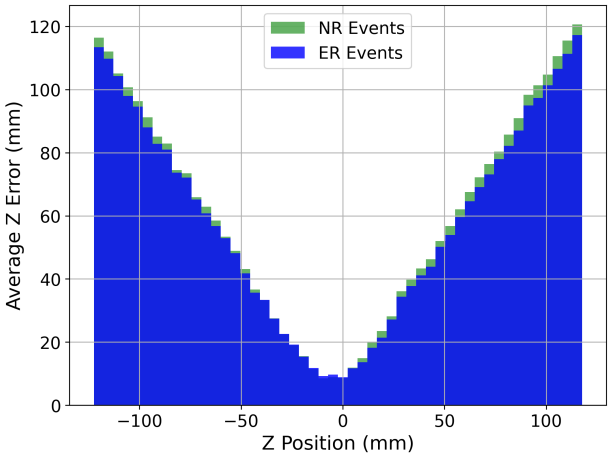
(a) Distribution of MEDs, noiseless data.



(b) Distribution of MEDs, noised data.

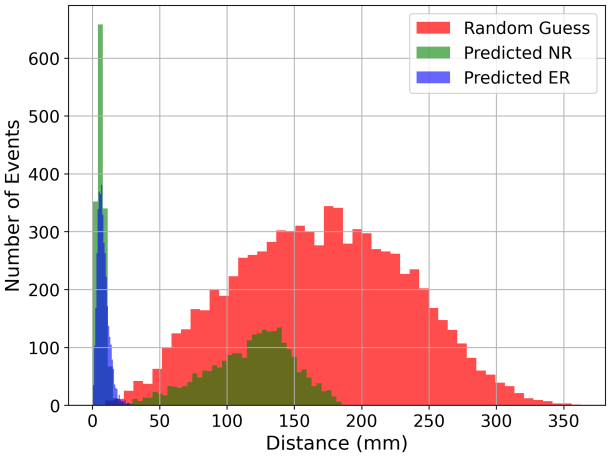


(c) Distribution of z errors, noiseless data.

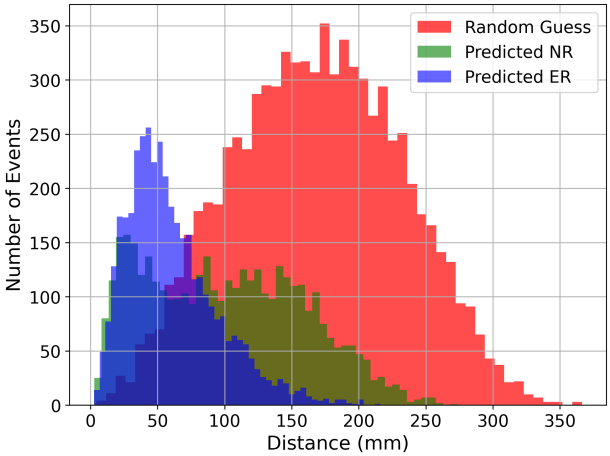


(d) Distribution of z errors, noised data.

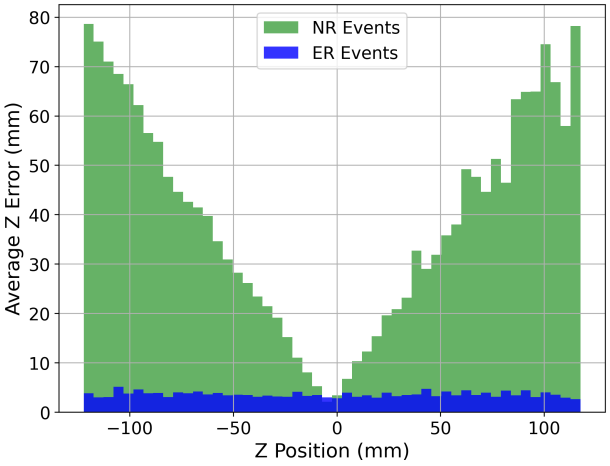
Figure 19: Discrepancies in PRC model performance on 50 eV ER and NR data. Note the short NR peak near 0 in (a) and the narrower ER and NR bins in (b) (same number as Random Guess bins). Note the overall poor performance on noised data and the large discrepancy in performance for recoil types seen in (c).



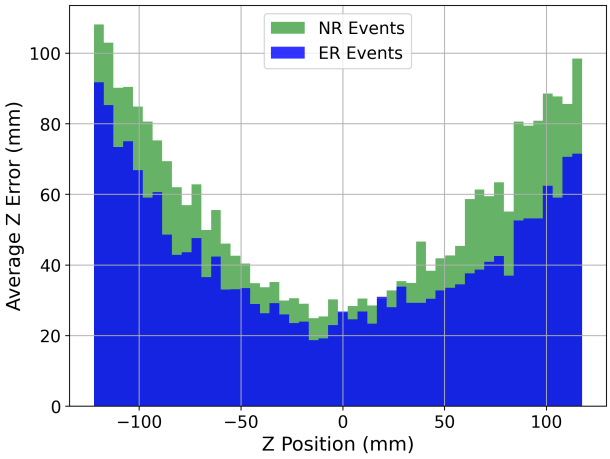
(a) Distribution of MEDs, noiseless data. Note the taller NR peak near 0. There is an equal number of bins for each distribution but with different widths.



(b) Distribution of MEDs, noised data. Note that ER and NR bins are narrower than Random Guess bins.

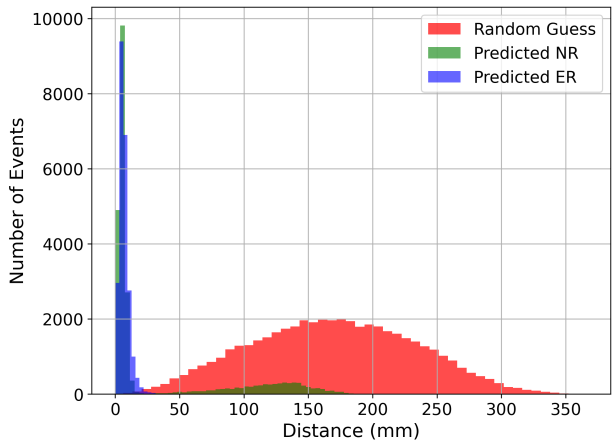


(c) Distribution of z errors, noiseless data.

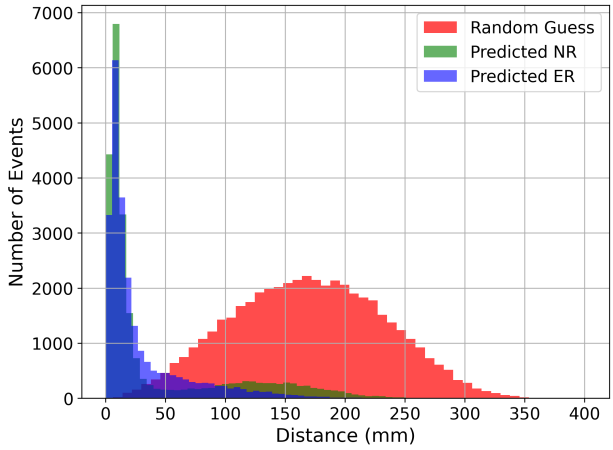


(d) Distribution of z errors, noised data.

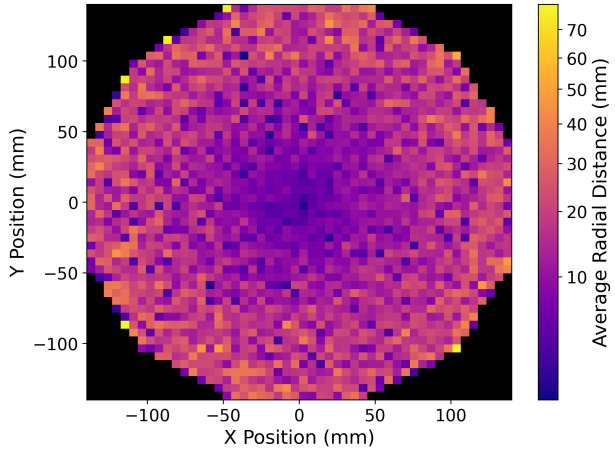
Figure 20: Discrepancies in PRC model performance on 100 eV ER and NR data. Note the bin widths in (a) and (b) and the lower maximum z error in (b) compared to 50 eV. Also note the poorer central position predictions for noised data, despite the better ER reconstruction performance.



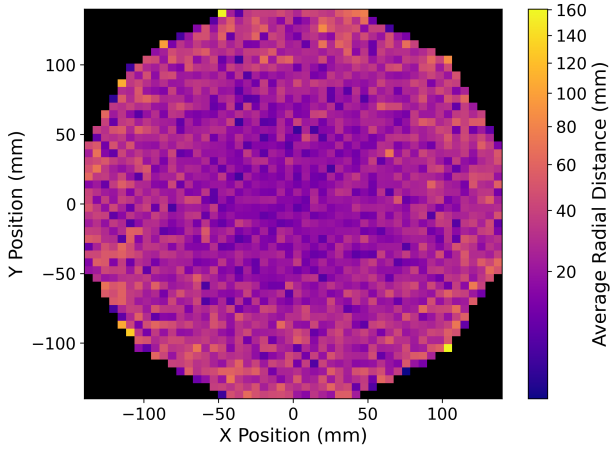
(a) Distribution of MEDs, noiseless data.



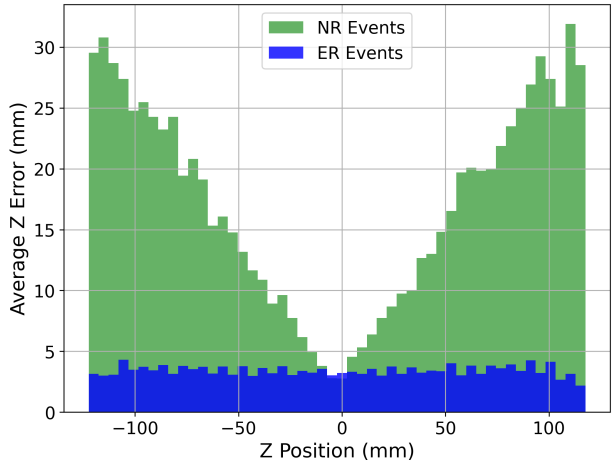
(b) Distribution of MEDs, noised data.



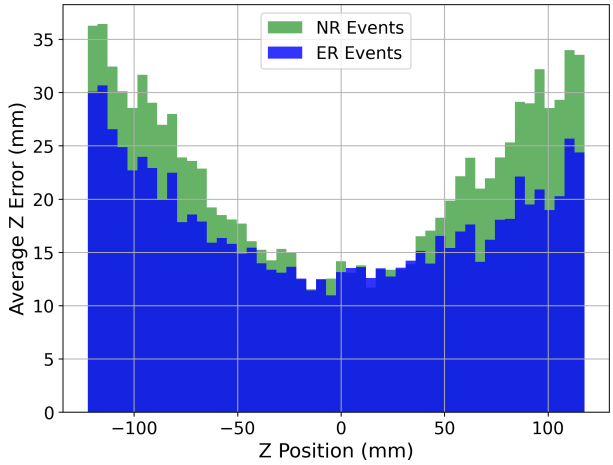
(c) Distribution of radial distances, noiseless data.



(d) Distribution of radial distances, noised data.



(e) Distribution of z errors, noiseless data.



(f) Distribution of z errors, noised data.

Figure 21: PRCE model position reconstruction performance on all energies. Note the smoother and more stretched out ER distribution in (b) compared to (a), the ranges of the colour maps in (c) and (d) and the stark contrast in ER reconstruction between (e) and (f). Also note the overall poor reconstruction of central events for noised data.

3.3 Classification

The classification performance of the PRC and PRCE models was evaluated using the ROC-AUC score (Receiver Operating Characteristic-Area Under Curve), a metric that quantifies the ability of a binary classifier to distinguish between classes (ER and NR events in this case). The ROC curve is constructed by plotting the true positive rate against the false positive rate across all classification thresholds, with the AUC score providing a summary of the classifier's overall performance [48]. A perfect classifier is represented by an AUC of 1.0, corresponding to complete separation between classes, while an AUC of 0.5 indicates performance no better than random guessing. Values below 0.5 suggest systematic misclassification.

For evaluation, a script was implemented to calculate AUC scores per epoch. For the PRC model (trained separately per energy), the epoch with the highest energy-specific AUC was selected for each energy. For the PRCE model (trained on all energies simultaneously), the epoch with the highest average AUC across all energies was chosen.

Table 5 shows an AUC score comparison between the PRC and PRCE models for both noiseless and noised data, while Table 6 shows a performance comparison between training on full vs. partial energy ranges for the PRCE model (noiseless data). Figure 22 displays the ROC curves for both models. Figure 23 shows confusion matrices for the PRC model's performance on 50 eV, visualising the differences in performance on noiseless and noised data.

The PRC model's classification performance maintained near-perfect levels (AUC $\geq$ 0.994) across all energies for noiseless data, while noised conditions showed moderate degradation at low energies (AUC 0.761 at 50 eV) before recovering to excellence (AUC $\geq$ 0.981) at 100 eV (Table 5). In contrast, the PRCE model demonstrated severe performance degradation above 100 eV for noiseless data (AUC 0.481-0.526), but showed remarkable improvement under noise, particularly at energies 100-100000 eV, with the exception of 50 eV – AUC 0.982 (noiseless) vs. 0.950 (noised).

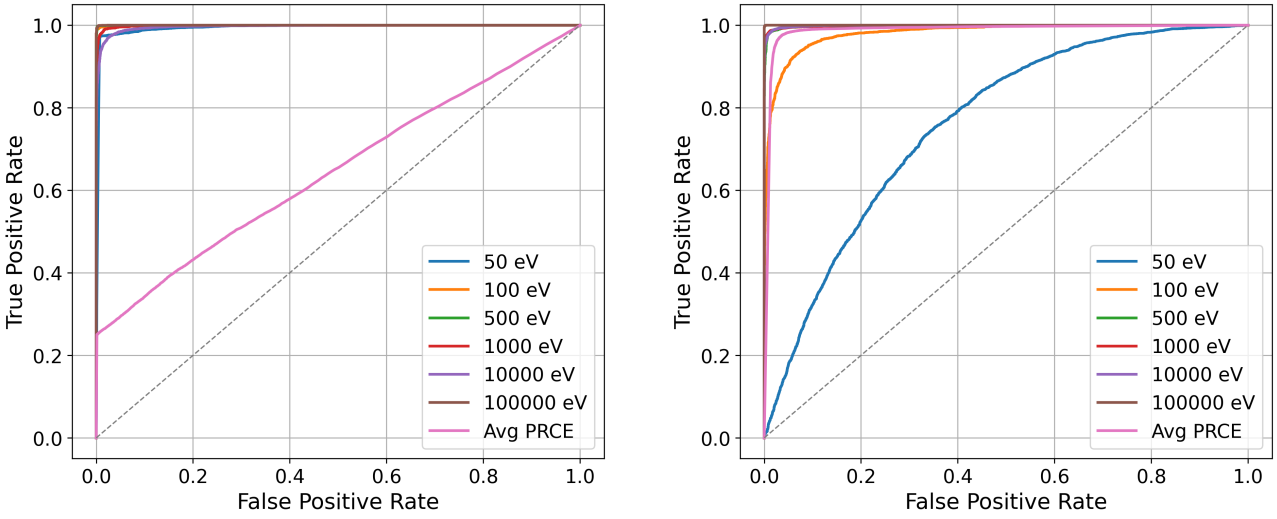
When trained on partial high-energy ranges (500 eV and above), the PRCE model achieved consistently excellent classification (AUC $\geq$ 0.984), while full-range training yielded substantially worse high-energy performance (AUC 0.481-0.526) (Table 6).

	Noiseless AUC Scores		**Noised AUC Scores**	
Energy (eV)	**PRC**	**PRCE**	**PRC**	**PRCE**
50	0.994	0.982	0.761	0.950
100	1.000	0.841	0.981	0.987
500	1.000	0.526	0.998	0.998
1000	0.999	0.489	0.999	0.999
10000	0.996	0.481	0.999	0.999
100000	1.000	0.513	1.000	1.000

Table 5: Classification performance of the PRC and PRCE models per energy. Note the decrease in noiseless PRCE performance with increase in energy compared to the increase for noised data. Also note that all AUC scores shown as 1.000 represent values such as 0.9995 or higher in the unrounded calculation.

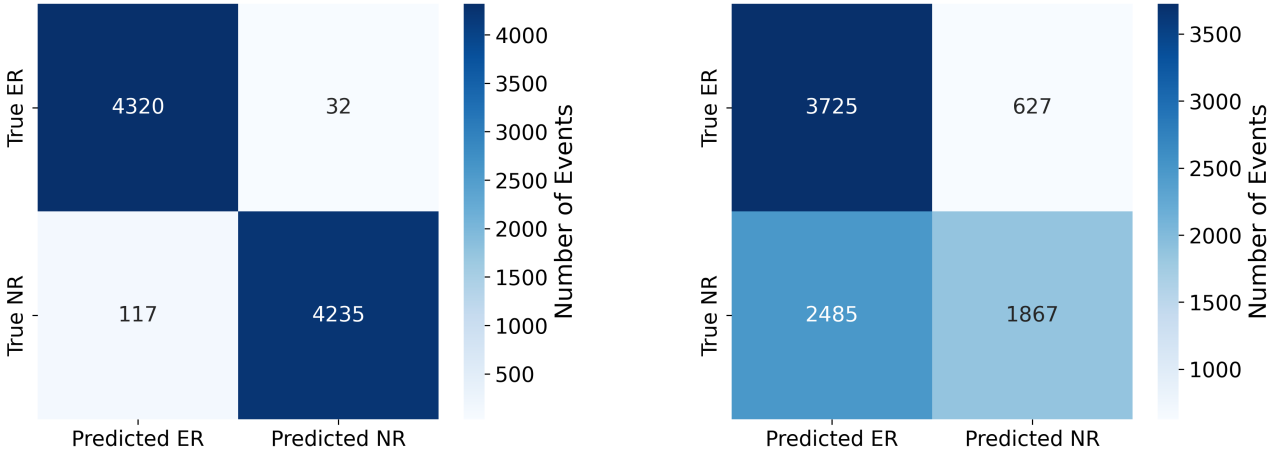
Energy (eV)	Partial Range AUC	Full Range AUC
500	0.994	0.526
1000	0.986	0.489
10000	0.984	0.481
100000	0.999	0.513

Table 6: AUC score comparison for PRCE trained on partial vs. full energy ranges (noiseless). Note the large discrepancy in performance.



(a) (b)

Figure 22: ROC curves with AUC scores comparing PRC and PRCE models on noiseless (a) and noised (b) data. The dashed line shows random guessing performance (AUC=0.5) for reference. Note the kinks in 50 eV and the average PRCE curve in (a).



(a) (b)

Figure 23: Confusion matrices for noiseless (a) and noised (b) 50 eV data, showing stronger classification performance for ERs than NRs.

3.4 Energy Regression

In evaluating the performance of the energy regression model, two key metrics were used to quantify accuracy and precision: bias and resolution. The bias, which serves as a measure of accuracy, is defined as the relative difference between the mean predicted energy and the true energy for each unique energy level. The bias is calculated using the formula

$$\text{Bias} = \frac{\langle E_{\text{pred}} \rangle - E_{\text{true}}}{E_{\text{true}}},$$

where $\langle E_{\text{pred}} \rangle$ is the mean of the predicted energies at a given true energy level, and E_{true} is the true energy. A low bias (e.g., $< 10\%$) is indicative of high accuracy in the energy predictions. The resolution, representing the model's precision, is calculated as the relative standard deviation of the predicted energies. It is given by

$$\text{Resolution} = \frac{\sigma(E_{\text{pred}})}{\langle E_{\text{pred}} \rangle},$$

where $\sigma(E_{\text{pred}})$ is the standard deviation of the predicted energies, and $\langle E_{\text{pred}} \rangle$ is the mean of the predicted energies. Lower values of resolution (e.g., $< 30\%$) correspond to higher consistency in the predictions.

To robustly determine the best-performing training epoch to be used for performance evaluation, both metrics are considered through a combined scoring scheme. The combined score for evaluating the best-performing epoch is defined as

$$\text{Combined Score} = 0.7 \times \text{Bias} + 0.3 \times \text{Resolution},$$

weighting accuracy more heavily in recognition of its greater importance in reconstructing true energy distributions reliably. The chosen weights reflect the priority of minimising systematic deviations from true energies, which is particularly critical when dealing with physical quantities where accurate mean estimation is paramount. While precision is still valued for ensuring consistent predictions, a moderate resolution is acceptable if it is accompanied by sufficiently low bias.

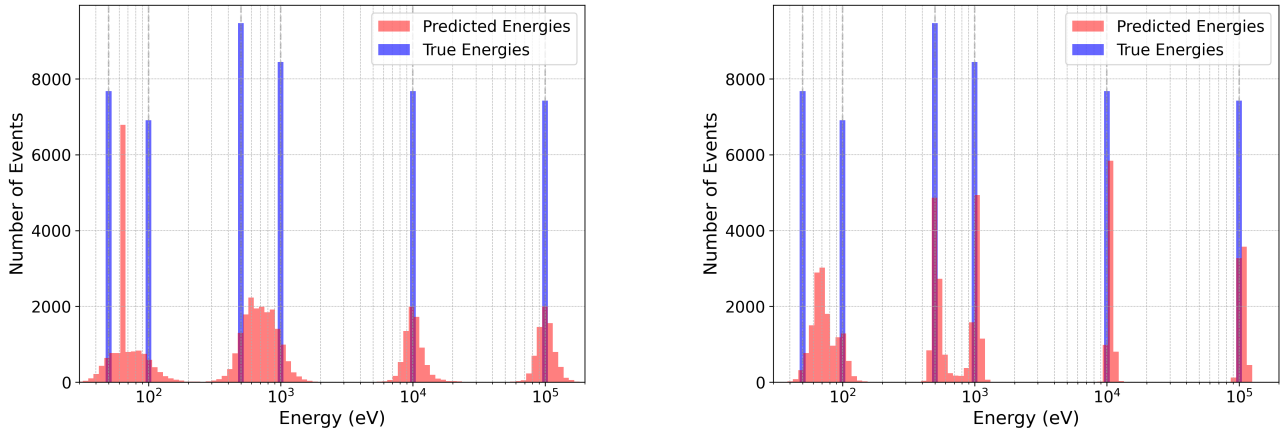
A Python script was created and used to iterate over the saved predicted and true energies for all epochs and calculate the combined score (averaged over all energies) for each epoch. The epoch corresponding to the lowest combined score was selected as optimal, balancing the trade-off between accuracy and precision. Table 7 displays the calculated biases and resolutions per energy along with their averages for the best performing epoch for both noiseless and noised data. Figure 24 visualises the distributions for each energy.

The PRCE model's energy regression performance was characterised by consistent improvements in resolution under noised conditions, with an average reduction from 21.37% (noiseless) to 11.39%. Notable resolution enhancements were observed at higher energies (4.50% at 10 keV and 5.28% at 100 keV with noise), while biases remained stable across conditions (average absolute bias: 11.66% noiseless vs. 10.91% noised). Energy-dependent bias patterns were identified, including systematic under-prediction at 100 eV (both noiseless and noised) and minor overestimation at 50 eV and 500 eV in noiseless data. The strongest performance is achieved at 100 keV with noise, where resolution below 5.5% and biases are confined to $< 6.5\%$.

Resolutions of 4.50–5.28% (10–100 keV, noised) meet typical sub-10% targets for particle detectors, while biases $\leq 6.5\%$ at these energies are acceptable. Low-energy biases (15–33%) exceed desirable limits but align with theoretical expectations for faint signals.

	Noiseless		Noised	
Energy (eV)	**Bias (%)**	**Resolution (%)**	**Bias (%)**	**Resolution (%)**
50	20.65	15.85	33.03	15.06
100	-15.67	29.92	-15.38	21.40
500	16.64	19.20	4.01	11.55
1000	-10.28	21.90	1.58	10.01
10000	4.77	25.74	6.37	4.50
100000	1.94	15.59	5.12	5.28
Average	11.66	21.37	10.91	11.39

Table 7: Energy regression performance of the PRCE model. Bias averages are calculated from absolute bias values. Note the trends in biases for noiseless and noised data. Negative bias values indicate the mean is predicted at a lower energy than the respective true energy, with positive biases showing the opposite.



(a) (b)

Figure 24: Distributions of predicted energies for every unique true energy for noiseless (a) and noised (b) data. Note the tall peak between 60 eV and 70 eV in (a) and the narrower energy spreads in (b).

3.5 Post-Hoc Reconstruction Analysis

The observed gap in low-energy performance between ER and NR data (Table 4, Figures 19 and 20) was investigated using the noiseless dataset, where signals lack background noise and thus provide unambiguous ground truth. A Python script quantified events with at least one signal, revealing 99.3% detection for 50 eV ERs versus 2.7% for NRs, increasing to 100% versus 35.5% at 100 eV. By 500 eV, both types approached full detection (100% ER vs 99.9% NR).

An additional comparison was performed by replacing the random distribution of points within the cell volume with a fixed set of points at (0, 0, 0) (maintaining the same array length), as prompted by closer inspection of Figures 16 and 19. Figure 25 illustrates this comparison.

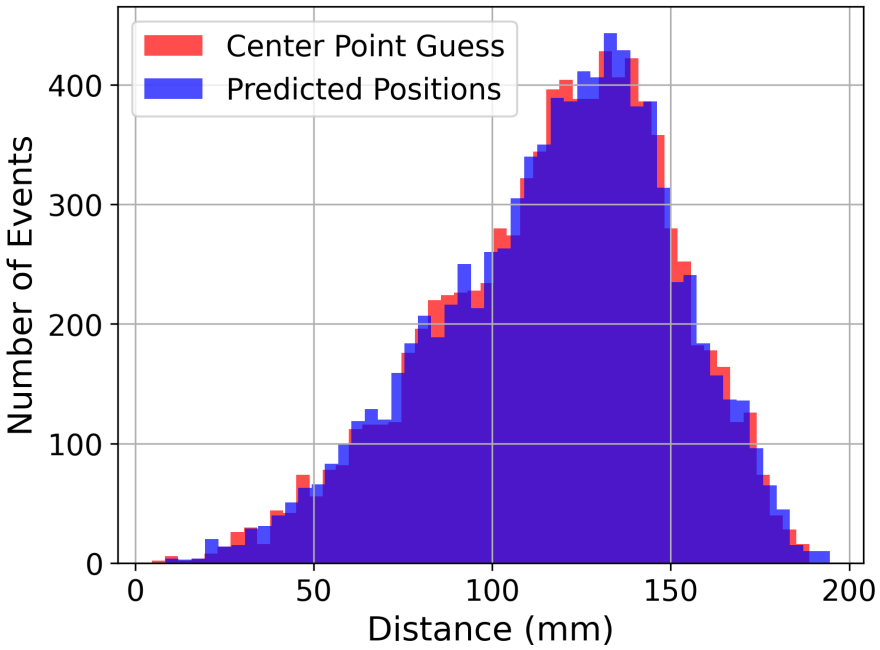


Figure 25: Distribution of MEDs for noised 50 eV PRC performance compared with predicting every event to be at (0, 0, 0) – Centre Point Guess. Note that predicted position MEDs are not split by recoil type. Also note how closely the distributions match.

4 Discussion

Training on noised data resulted in visible overfitting, as reflected in the validation loss plots discussed in Section 3.1. This is likely due to the model learning specific noise patterns rather than generalisable features. Although not explored here, data augmentation techniques, such as introducing physically motivated signal distortions, could help mitigate this issue by diversifying the training distribution. Another limitation arose from the unbalanced total loss function. Without scaling, the MSE loss for position reconstruction dominated due to its larger absolute values compared to the BCE loss for classification and the MSE loss for energy regression. Although all loss components were necessary, future work could apply normalisation or explicit weighting to balance training across tasks.

Despite these constraints, the models performed well overall, with energy regression performance (Section 3.4) falling within acceptable ranges for low-energy detectors (bias $< 15\%$, resolution $< 20\%$). Interestingly, the PRCE model showed consistent improvement under noised conditions, particularly in energy regression and classification. This likely stems from the injected white noise acting as a regulariser: it disrupts overfitting to idealised simulations while introducing variability akin to experimental signal uncertainties. The improved resolution with stable bias for noised data (Table 7) suggests the model developed more robust pattern recognition capabilities.

The observed bias and resolution trends across energies align with the signal partitioning physics described in [21]. At low energies (50 and 100 eV), the elevated but stable biases reflect inherent reconstruction challenges for quasiparticle-dominated events. The resolution improvements at intermediate energies (500 and 1000 eV) demonstrate noise's beneficial role in modelling transitional UV/triplet channels, while high-energy performance (10000 and 100000 eV) confirms effective learning of ionisation processes. The architecture's ability to maintain precision under noise-induced fluctuations suggests strong potential for real-world deployment, though the persistent 50 eV bias may indicate fundamental limitations for pure-quasiparticle

reconstruction in experimental-like conditions.

The position reconstruction results reveal trends in how noise and model structure affect performance (Section 3.2). The modest and inconsistent across energies differences between the PR and PRC models shown in Table 3 are unlikely to reflect systematic architectural advantages, considering the inherent stochasticity of deep learning training. In contrast, Table 4 shows that the PRCE model consistently outperformed the PRC model under noised conditions at low energies, suggesting that the multi-task setup of the PRCE model, incorporating energy regression alongside classification and position reconstruction, helped the model learn more stable features under noise.

Interpreting these trends through the theoretical framework described in [21], the degradation in performance at lower energies (50 and 100 eV) is consistent with the increasingly discrete nature of signal production in LHe. At energies below a few hundred eV, the creation of excitations such as phonons and rotons becomes subject to quantum statistical fluctuations, and signal amplitudes exhibit greater variability. This leads to an intrinsic limit on the amount of information that can be extracted from event traces. The energy dependence of the position reconstruction performance therefore reflects not only model limitations but also fundamental physical noise inherent to the detection medium.

The addition of white noise further degraded model performance at 50 and 100 eV by obscuring low-amplitude signals that were already limited by quantum production thresholds. As anticipated from signal propagation considerations in superfluid helium [21], random background noise particularly hampers the detection of small quasiparticle and photon populations, which are critical for accurate position and energy estimation at low event energies. The observed smoothing of model performance at higher energies, where signal amplitudes are larger and more stable, aligns with the expectation that macroscopic signal production dominates in this regime, mitigating the relative impact of added noise.

The stronger performance of the PRCE model at low energies suggests that multi-task learning, particularly the energy regression head, helped disambiguate faint signals by encouraging the network to preserve amplitude information, offering a partial countermeasure against the degradation induced by white noise and fundamental physical limitations.

Figure 25 revealed that the models defaulted to predicting events near the central coordinates (0,0,0) for noised ER and NR events at 50 eV. This behaviour can be attributed to noise masking critical signal features, particularly for low-energy events, as evidenced by the performance degradation shown in Figures 19d, 20d, 21f, and 21d. Notably, central positions were reconstructed more accurately at 50 eV than at 100 eV, despite overall performance away from the centre being superior at 100 eV. The exact cause of this discrepancy remains unclear, though it may be linked to the sparsity of central training points, which could lead to systematic mispredictions at the centre.

Future work could explore the precise energy threshold where noised and noiseless NR position reconstruction performance converges, with theoretical predictions suggesting this transition occurs above 200 eV due to Penning quenching effects [21]. Additional investigation of the anomalous central position reconstruction accuracy at 50 eV compared to 100 eV would be valuable, potentially through targeted simulations with increased sampling near the cell centre. The multi-task architecture's success suggests further development of task-specific loss balancing could optimise performance across all energy regimes.

Classification performance (Section 3.3) was generally excellent across energy levels and noise conditions for the PRC model, which achieved near-perfect AUC scores (typically > 0.98). However, two anomalies were observed at 50 eV. First, the ROC curve for the PRC model on noiseless data displayed a visible kink (Figure 22a), likely caused by sharp transitions in classification around a small subset of difficult events despite overall good separability. This was likely affected by the general lack of NR event signals at 50 eV as explained above. A

similar kink was observed for the PRCE model (note that the ROC curve shown in Figure 22a is an average) and is attributed to the same low-energy intricacies. These kinks can still appear even when the AUC score is high, because the AUC measures the overall ranking ability and is not sensitive to localised sharp changes in the curve.

Second, when noise was added, the PRC model's classification performance dropped substantially for 50 eV (AUC score of 0.761 compared to 0.994 for noiseless), indicating poor generalisation to noisy data at the lowest energy, likely caused by the model's reliance on fragile, low-level features that were disrupted by noise. Interestingly, the PRCE model performed better under these conditions, even compared to the limited-range noiseless training (Tables 5 and 6). When considered together, these discrepancies imply that the random white noise only worked as a regulariser for the PRCE model, while negatively affecting the PRC model's performance at 50 and 100 eV.

The multi-task framework of the PRCE model also explains its superior performance on noised data over the PRC model. The energy regression task's regularisation of shared features, particularly its suppression of noise-corrupted quasiparticle signals in 50 eV NRs, indirectly stabilises classification. This cross-task benefit clearly illustrates one of the strengths of multi-task architectures: performance in one task can indirectly stabilise another by influencing the shared feature space.

The overall poor classification performance of the PRCE model on noiseless data (Table 5) can be attributed to the model overfitting to the simpler low-energy signals, which dominated gradient updates and led to the forgetting of features important for higher-energy events, a phenomenon known as catastrophic forgetting [49]. The lack of energy-aware loss balancing and the noiseless nature of the training data further encouraged the model to prioritise easy low-energy patterns at the expense of broader generalisation. Notably, this forgetting was fully mitigated under noised conditions, where injected white noise prevented over-specialisation to low-energy patterns by artificially diversifying the training distribution, forcing the model to maintain robust features across all energies.

The stark contrast in performance between training on the restricted range and the full range (Table 6) implies that limiting the energy range removed the model's ability to overfit to low-energy events and forced the development of more discriminative high-energy features. This suggests broad energy range training, especially on noiseless data, may require strategies such as loss weighting, staged curriculum training, or model regularisation to maintain balanced performance across all signal domains.

The severe NR signal depletion observed at energies of 50 and 100 eV (Section 3.5) can be explained by the quasiparticle channel's low efficiency (main signal for NRs at low energies as seen in 3), caused by quantum evaporation's high Kapitza resistance (Section II of [21]). Crucially, this implies that overall performance limitations on noiseless 50 eV NR data stemmed from the underlying physical model used in simulations, rather than deficiencies in the analysis framework.

One major limitation of this study was the lack of repeated training runs. Each model was trained only once (with the exception of the PRCE model) due to time constraints, meaning variability due to random weight initialisation, shuffling, or batch selection was not captured. As such, some of the performance differences could be statistically insignificant. Future work could address this via training with multiple seeds, Monte Carlo dropout during inference, or bootstrap resampling of the test data to estimate predictive uncertainty. Such steps would not only provide confidence intervals for key metrics like AUC or resolution, but also offer insight into model stability, which is critical for any future deployment in an experimental setting. Performance could be also improved through the implementation of more straightforward measures, such as training on larger datasets and for more epochs, since the position reconstruction performance floor likely scales with dataset size.

The successful application of the LSTM-Transformer architecture to the DELight simulations demonstrates the potential of machine learning to revolutionise dark matter detection. By achieving robust classification (AUC > 0.98) and energy regression (bias $< 15\%$), specifically for noised data, this work paves the way for real-time, multi-task analysis in next-generation detectors. Future research should focus on integrating real experimental data to refine noise resilience and extending the model to continuous energy spectra beyond discrete simulations. The architecture's adaptability suggests broader utility in other rare-event searches, such as neutrino experiments or quantum materials research, where multi-channel signal discrimination is critical. Additionally, the observed limitations in low-energy position reconstruction highlight the need for improved detector designs or hybrid approaches combining ML with physical models.

5 Conclusion

This project successfully developed a multi-task machine learning framework to model the DELight experiment's simulated detector response to light dark matter interactions. An LSTM-Transformer architecture was designed to leverage temporal and inter-detector spatial correlations, enabling simultaneous position reconstruction, recoil type classification, and energy regression.

The models achieved strong performance across most tasks. High-energy (≥ 500 eV) position reconstruction reached sub-centimetre accuracy, while classification attained near-perfect AUC scores even under noise. Energy regression reached biases below 15% and resolutions under 20%, meeting the stringent demands of low-energy event reconstruction. Noise was shown to regularise training, leading to better generalisation and stability only for energy regression compared to noiseless training. The multi-task PRCE model outperformed simpler models under noisy conditions, highlighting the benefits of jointly optimising multiple physical targets.

Nonetheless, challenges were observed, particularly for 50 eV nuclear recoils, where sparse input signals limited performance. For both recoils at 50 eV events, reconstruction frequently defaulted to central positions as noise completely dominated the faint signals, while spatial distributions at 100 eV were degraded but non-random, consistent with partial signal preservation. Future improvements could include better loss balancing across tasks, uncertainty quantification through repeated training runs, and data augmentation to simulate a broader range of noise and detector artefacts.

Importantly, this work demonstrates that modern machine learning architectures can robustly process complex multi-channel detector data, offering a path toward real-time analysis in next-generation dark matter searches. Integrating these models into DELight's experimental pipeline could significantly enhance signal reconstruction and background rejection capabilities. Further extensions could include adapting the architecture to continuous energy spectra, fine-tuning models on real experimental data, and exploring hybrid approaches that combine machine learning with physically motivated signal models.

Overall, the results pave the way for a new class of machine learning-assisted rare event searches, where deep models trained on simulated or real data augment the sensitivity of low-threshold detectors across multiple physical observables.

Acknowledgements

I sincerely thank my supervisor for his invaluable guidance and for the insightful discussions about the theory. I am grateful to my project partner for his continued support throughout this work. Special thanks to my brother for his feedback on the machine learning framework, and to

Dr. Francesco Toschi for providing the Monte Carlo simulation tools essential to this research. This work was made possible through access to Imperial College London's High Energy Physics computing cluster.

Bibliography

[1] V. Trimble. "Existence and nature of dark matter in the universe". In: *Annual review of astronomy and astrophysics* 25.1 (1987), pp. 425–472. DOI: 10.1146/annurev.aa.25.090187.002233.

[2] R. Massey, T. Kitching, and J. Richard. "The dark matter of gravitational lensing". In: *Reports on Progress in Physics* 73.8 (2010), p. 086901. DOI: 10.1088/0034-4885/73/8/086901.

[3] J. D. Bekenstein. "The modified Newtonian dynamics—MOND and its implications for new physics". In: *Contemporary Physics* 47.6 (2006), pp. 387–403. DOI: 10.1080/00107510701244055.

[4] G. F. Chapline and P. H. Frampton. "A new direction for dark matter research: intermediate-mass compact halo objects". In: *Journal of Cosmology and Astroparticle Physics* 2016.11 (2016), p. 042. DOI: 10.1088/1475-7516/2016/11/042.

[5] C. Alcock, R. Allsman, D. R. Alves, T. Axelrod, A. C. Becker, D. Bennett, K. H. Cook, N. Dalal, A. J. Drake, K. Freeman, et al. "The MACHO Project: Microlensing Results from 5.7 Years of LargeMagellanic Cloud Observations". In: *The Astrophysical Journal* 542.1 (2000), p. 281. DOI: 10.1086/309512.

[6] J. Yoo, J. Chaname, and A. Gould. "The end of the MACHO era: limits on halo dark matter from stellar halo wide binaries". In: *The Astrophysical Journal* 601.1 (2004), p. 311. DOI: 10.1086/380562.

[7] P. J. E. Peebles. "Large-scale background temperature and mass fluctuations due to scale-invariant primeval perturbations". In: *The Astrophysical Journal* 263 (1982), pp. L1–L5. DOI: 10.1086/183911.

[8] V. Sahni and A. A. Sen. "A new recipe for ΛCDM". In: *The European Physical Journal C* 77.4 (2017), pp. 1–8. DOI: 10.1140/epjc/s10052-017-4796-7.

[9] M. S. Turner. "ΛCDM: Much more than we expected, but now less than what we want". In: *Foundations of Physics* 48.10 (2018), pp. 1261–1278. DOI: 10.1007/s10701-018-0178-8.

[10] E. Aprile, K. Abe, F. Agostini, S. Ahmed Maouloud, L. Althueser, B. Andrieu, E. Angelino, J. Angevaare, V. Antochi, D. Antón Martin, et al. "First dark matter search with nuclear recoils from the XENONnT experiment". In: *Physical Review Letters* 131.4 (2023), p. 041003. DOI: 10.1103/PhysRevLett.131.041003.

[11] D. S. Akerib. "LUX, ZEPLIN and LUX-ZEPLIN: Developments in liquid xenon detectors and the search for WIMP dark matter". In: *Nuclear Physics B* 1003 (2024), p. 116437. DOI: 10.1016/j.nuclphysb.2024.116437.

[12] X. Cui, A. Abdukerim, W. Chen, X. Chen, Y. Chen, B. Dong, D. Fang, C. Fu, K. Giboni, F. Giuliani, et al. "Dark matter results from 54-ton-day exposure of PandaX-II experiment". In: *Physical review letters* 119.18 (2017), p. 181302. DOI: 10.1103/PhysRevLett.119.181302.

[13] R. Agnese, A. J. Anderson, M. Asai, D. Balakishiyeva, R. Basu Thakur, D. Bauer, J. Beaty, J. Billard, A. Borgland, M. Bowles, et al. "Search for low-mass weakly interacting massive particles with SuperCDMS". In: *Physical review letters* 112.24 (2014), p. 241302. DOI: 10.1103/PhysRevLett.112.241302.

[14] G. Angloher, S. Banik, G. Benato, A. Bento, A. Bertolini, R. Breier, C. Bucci, J. Burkhart, L. Canonica, A. D'Addabbo, et al. "Results on sub-GeV dark matter from a 10 eV threshold CRESST-III silicon detector". In: *Physical Review D* 107.12 (2023), p. 122003. DOI: 10.1103/PhysRevD.107.122003.

[15] S. Abdollahi, M. Ajello, L. Baldini, J. Ballet, D. Bastieri, J. B. Gonzalez, R. Bellazzini, A. Berretta, E. Bissaldi, R. Bonino, et al. "The fermi-LAT lightcurve repository". In: *The Astrophysical Journal Supplement Series* 265.2 (2023), p. 31. DOI: 10.3847/1538-4365/acbb6a.

[16] M. Ahlers, K. Helbing, and C. Pérez de los Heros. "Probing particle physics with IceCube". In: *The European Physical Journal C* 78.11 (2018), p. 924. DOI: 10.1140/epjc/s10052-018-6369-9.

[17] B. von Krosigk, K. Eitel, C. Enss, T. Ferber, L. Gastaldo, F. Kahlhoefer, S. Kempf, M. Klute, S. Lindemann, M. Schumann, et al. "DELight: A Direct search Experiment for Light dark matter with superfluid helium". In: *SciPost Physics Proceedings* 12 (2023), p. 016. DOI: 10.21468/SciPostPhysProc.12.016.

[18] P. Agnes, I. F. d. M. Albuquerque, T. Alexander, A. Alton, G. Araujo, D. M. Asner, M. Ave, H. O. Back, B. Baldin, G. Batignani, et al. "Low-mass dark matter search with the DarkSide-50 experiment". In: *Physical review letters* 121.8 (2018), p. 081307. DOI: 10.1103/PhysRevLett.121.081307.

[19] A. Wyatt. "Evaporation of liquid 4He; A quantum process". In: *Physica B+ C* 126.1-3 (1984), pp. 392–399. DOI: 10.1016/0378-4363(84)90193-1.

[20] S. Bandler, C. Enss, G. Goldhaber, R. Lanou, H. Maris, T. More, F. Porter, and G. Seidel. "The evaporation signal from α particles stopped in superfluid helium". In: *Journal of Low Temperature Physics* 93 (1993), pp. 715–720. DOI: 10.1007/BF00693501.

[21] F. Toschi, A. Brunold, L. Burmeister, K. Eitel, C. Enss, E. Fascione, T. Ferber, R. Gabriel, L. Hauswald, F. Kahlhoefer, et al. "Signal partitioning in superfluid He 4: A Monte Carlo approach". In: *Physical Review D* 111.3 (2025), p. 032013. DOI: 10.1103/PhysRevD.111.032013.

[22] A. Csótó and G. M. Hale. "Nature of the first excited state of 4 He". In: *Physical Review C* 55.5 (1997), p. 2366. DOI: 10.1103/PhysRevC.55.2366.

[23] S. A. Hertel, A. Biekert, J. Lin, V. Velan, and D. McKinsey. "Direct detection of sub-GeV dark matter using a superfluid He 4 target". In: *Physical Review D* 100.9 (2019), p. 092007. DOI: 10.1103/PhysRevD.100.092007.

[24] A. Fleischmann, C. Enss, and G. Seidel. "Metallic magnetic calorimeters". In: *Cryogenic particle detection* (2005), pp. 151–216. DOI: 10.1007/10933596_4.

[25] S. Kempf, A. Fleischmann, L. Gastaldo, and C. Enss. "Physics and applications of metallic magnetic calorimeters". In: *Journal of Low Temperature Physics* 193 (2018), pp. 365–379. DOI: 10.1007/s10909-018-1891-6.

[26] C. Enss, S. Bandler, R. Lanou, H. Maris, T. More, F. Porter, and G. Seidel. "Quantum evaporation of4He: Angular dependence and efficiency". In: *Physica B: Condensed Matter* 194 (1994), pp. 515–516. DOI: 10.1016/0921-4526(94)90587-8.

[27] F. Carter, S. Hertel, M. Rooks, P. McClintock, D. McKinsey, and D. Prober. "Calorimetric Observation of Single He _2^* He 2* Excimers in a 100-mK He Bath". In: *Journal of low temperature physics* 186 (2017), pp. 183–196. DOI: 10.1007/s10909-016-1666-x.

[28] A. Grobov, A. Ilyasov, et al. "Convolutional neural network approach to event position reconstruction in DarkSide-50 experiment". In: *Journal of Physics: Conference Series*. Vol. 1690. 1. IOP Publishing. 2020, p. 012013. DOI: 10.1088/1742-6596/1690/1/012013.

[29] S. Delaquis, M. Jewell, I. Ostrovskiy, M. Weber, T. Ziegler, J. Dalmasson, L. Kaufman, T. Richards, J. Albert, G. Anton, et al. "Deep neural networks for energy and position reconstruction in EXO-200". In: *Journal of Instrumentation* 13.08 (2018), P08023. DOI: 10.1088/1748-0221/13/08/P08023.

[30] N. Geneva and N. Zabaras. "Transformers for modeling physical systems". In: *Neural Networks* 146 (2022), pp. 272–289. DOI: 10.1016/j.neunet.2021.11.022.

[31] K. Cao, T. Zhang, and J. Huang. "Advanced hybrid LSTM-transformer architecture for real-time multi-task prediction in engineering systems". In: *Scientific Reports* 14.1 (2024), p. 4890. DOI: 10.1038/s41598-024-55483-x.

[32] F. Andayani, L. B. Theng, M. T. Tsun, and C. Chua. "Hybrid LSTM-transformer model for emotion recognition from speech audio files". In: *IEEE Access* 10 (2022), pp. 36018–36027. DOI: 10.1109/ACCESS.2022.3163856.

[33] E. Grossi and M. Buscema. "Introduction to artificial neural networks". In: *European journal of gastroenterology & hepatology* 19.12 (2007), pp. 1046–1054. DOI: 10.1097/MEG.0b013e3282f198a0.

[34] Y. Bai. "RELU-function and derived function review". In: *SHS web of conferences*. Vol. 144. EDP Sciences. 2022, p. 02006. DOI: 10.1051/shsconf/202214402006.

[35] P. S. Foundation. *Python Language Reference, version 3.8*. Available at https://docs.python.org/3/. 2023. DOI: 10.5281/zenodo.7553275.

[36] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. DOI: 10.48550/arXiv.1912.01703.

[37] Y. Yu, X. Si, C. Hu, and J. Zhang. "A review of recurrent neural networks: LSTM cells and network architectures". In: *Neural computation* 31.7 (2019), pp. 1235–1270. DOI: 10.1162/neco_a_01199.

[38] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. "Dropout: a simple way to prevent neural networks from overfitting". In: *The journal of machine learning research* 15.1 (2014), pp. 1929–1958. DOI: 10.5555/2627435.2670313.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017). DOI: arXiv:1706.03762.

[40] M.-C. Popescu, V. E. Balas, L. Perescu-Popescu, and N. Mastorakis. "Multilayer perceptron and neural networks". In: *WSEAS Transactions on Circuits and Systems* 8.7 (2009), pp. 579–588. DOI: 10.5555/1639537.1639542.

[41] P. Domingos. "A few useful things to know about machine learning". In: *Communications of the ACM* 55.10 (2012), pp. 78–87. DOI: 10.1145/2347736.2347755.

[42] P. Cunningham, M. Cord, and S. J. Delany. "Supervised learning". In: *Machine learning techniques for multimedia: case studies on organization and retrieval*. Springer, 2008, pp. 21–49. DOI: 10.1007/978-3-540-75171-7_2.

[43] F. He, T. Liu, and D. Tao. "Control batch size and learning rate to generalize well: Theoretical and empirical evidence". In: *Advances in neural information processing systems* 32 (2019). DOI: 10.5555/3454287.3454390.

[44] P. Zhou, X. Xie, Z. Lin, and S. Yan. "Towards understanding convergence and generalization of AdamW". In: *IEEE transactions on pattern analysis and machine intelligence* (2024). DOI: 10.1109/TPAMI.2024.3382294.

[45] L. N. Smith. "Cyclical learning rates for training neural networks". In: *2017 IEEE winter conference on applications of computer vision (WACV)*. IEEE. 2017, pp. 464–472. DOI: 10.1109/WACV.2017.58.

[46] X. Chen, S. Z. Wu, and M. Hong. "Understanding gradient clipping in private sgd: A geometric perspective". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 13773–13782. DOI: 10.5555/3495724.3496879.

[47] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, et al. "Mixed precision training". In: *arXiv preprint arXiv:1710.03740* (2017). DOI: 10.48550/arXiv.1710.03740.

[48] C. Ferri, J. Hernández-Orallo, and P. A. Flach. "A coherent interpretation of AUC as a measure of aggregated classification performance". In: *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. 2011, pp. 657–664. DOI: 10.5555/3104482.3104565.

[49] R. Kemker, M. McClure, A. Abitino, T. Hayes, and C. Kanan. "Measuring catastrophic forgetting in neural networks". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1. 2018. DOI: 10.1609/aaai.v32i1.11651.