Data Analysis Report (Re-submission)

Aditya Gautam
Department of Physics, University of Toronto
PHY324: Practical Physics II
October 7, 2024

1 Introduction

Background

Most physics experiments require devices to measure data from phenomenon around us. Any device has multiple components at play, all of which are subject to flaws and uncertainties. So when we use any device, it is extremely important to understand its mechanisms and use that to calibrate the raw output and obtain the type of data we want. The end goal is to convert phenomenon around us to information we can interpret and manipulate in order to understand or verify our theories better.

The author of this paper attempts to calibrate a particle detector which converts incident energy from a particle to an electrical signal. With the help of the detector's response to a fixed source, the paper interprets the behaviour of the detector on a higher level and use the most reliable method for calibration. Additionally, we use our calibration on data from background noise and the data from a signal source to try to make sense of it.

Fitting functions

We used three main functions for fitting as below. These were found by first plotting the data and then analyzing the shape to fit them to known distributions:

- Gaussian for the histograms for the calibrated detector readings:

$$g(x) = A \ * \ exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Log-normal for the calibrated noise data:

$$l_n(x) = A \ * \ exp\left(-\frac{(ln(x)-\mu)^2}{2\sigma^2}\right)$$

- Exponential distribution for the signal:

$$m(x) = \lambda x \ * \ e^{-\lambda x}$$

  (As the fit parameters were not normalized so the mean and sigma value become a little less relevant for interpretation. For analysis we instead used the peaks using calculus and where they were achieved)

Experimental Setup

The setup involves an unnamed particle detector of a read rate of at least 1 MHz. We assume the detector to have only one species of energy deposition: electron-recoils in the sensor arising from incident high energy photons. The detector is connected to a readout circuit (and some device for storing the measurements), a trigger with extreme precision that connects the sources to the detector for timing the onset, an unnamed source that emits 10 keV photons, and a source that emits a random signal in the ROI (0-20 keV). (The author did not do this experiment, and was only given the calibration, noise and signal source datasets.)

For the data analysis, we used Python 3.11.9 running on an x86 windows machine.

2 Methodology

Data Collection

The detector and its readout circuit is set up in a way to obtain a specific pulse, which has a 20 μs rise time (τ_R) and a 80 μs fall time (τ_F).

The following function is used as a normalizing factor so that the term without the scale-able amplitude (A) of the incident particle has an amplitude of unity:

$$y(t) = A \ * \ \left(\frac{\tau_F}{\tau_R}\right)^{-\frac{\tau_R}{\tau_F - \tau_R}} \ * \ \left(\frac{\tau_R - \tau_F}{\tau_F}\right) \ * \ (e^{-t/\tau_R} - e^{-t/\tau_F})$$

Our detector measured he output at a rat of 1 MHz, or once every microsecond. The system is set up to store 4096 samples for each trial and has a near-ideal trigger that positions the start of the pulse at the 1000th sample of the voltage trace, or very close to 1 ms. To compensate for the noise, we first took a noise data set of 1000 samples with no intentional signals on the setup, to obtain a rough idea of the background radiation and signals in the ROI (0-20keV).

For quantifying the detector response, a set of calibration dataset of 1000 samples was taken by emitting 10 keV photons at the detector, assumed to deposit its energy at the detector. Consequentially we also set the Region of Interest to be 0-20 keV to avoid as much random noise from the surroundings.

Lastly, we exposed the detector to our signal source. We again limited the data to the ROI of 0-20 keV, and took 1000 samples. This is the target from which we want to measure and extract the energy spectrum.

(Note: The experimental setup for data collection was done beforehand and we were only provided the information about it in a handout and the data from the calibration, signal, and noise trials recorded in .pkl files, which were then analyzed to build upon this experiment in an attempt to generate the calibration calculations and interpret the signal data.)

Data Analysis: Energy Estimation

After the data had been collected, we created a python code to try different ways to calibrate the detector. As we expected a Gaussian distribution centred at 10keV, from each of the samples, we took a different measure to calculate net energy deposited at the detector, and plotted them on a histogram.

We used the following 6 methods for estimating the energy of the incident radiation:

1. The difference of the maximum and the minimum values recorded by the detector.
2. The difference between the maximum value recorded by the detector and a baseline calculated as the average reading of the detector before the initial pulse (from 0 to 1000 μs). (One adjustment, see appendix)
3. An integral calculated as the sum across the whole trace.
4. An integral calculated as the sum across the whole trace minus the baseline calculated as before, which is multiplied across the temporal range (4096 samples).
5. An integral which is the localized sum of the readings during the pulse onset. As the readings decay near-exponentially after the peak, this was taken to be between the 995^{th} sample and the 1995^{th} sample. (This range was chosen by visual inspection of the calibration data and then optimizing using a function to check where the pulse after the final time was similar to the baseline from before)
6. Fitted pulse amplitude: As we know the rough shape of the pulse to be an exponential curve with a $20\mu s$ rise and a $80\mu s$ fall time, we can fit a curve to this and use its amplitude as the metric for estimating the energy.

For each of the energy estimators, we fit a Gaussian curve to the histograms of the calculation to obtain the mean, standard deviation, chi-squared and chi-squared probability values to assess the goodness of the fit. This follows as we expect the most likely response to correlate to a full 10keV reading. We also took the residuals from the error-bars of the fit to notice any obvious patterns to discredit the estimators. The data was plotted and the values were set in a table to compare the reliability of each fit (see Results).

We then used this to generate an energy calibration factor by relating the calculated mean response to correspond to a reading of 10 keV for each energy estimator. Then the entire dataset is linearly adjusted by this factor to estimate the energy of each response of the detector. These were also fit with a Gaussian to calculate reliability of the fit, and obtain a more general response. Using a python script, we optimized the initial guess-parameters for scipy.optimize.curve_fit to give us the best result for each of the energy estimators.

For choosing the best estimator, we mainly used two things: goodness of fit, and the precision of the detector given to us by that fit. Assuming and observing a normal distribution, the width of the Gaussian gives us the resolution of the detector. For goodness of fit, we used residuals from the fit, reduced chi-squared value, and chi-squared probability.

Using this particular energy estimator and scaling factor for 10 keV, we tuned the noise and signal datasets with the calibration factor to give us meaningful information about the environment and the source signal respectively. We used more appropriate functions to estimate the responses depending on the plot itself. The noise seemed to be a near Gamma/Log-Normal distribution, and the signal source to be a sum/mix of multiple exponential functions. This was done by plotting just the histograms and then evaluating visually according to known distributions.

3 Results

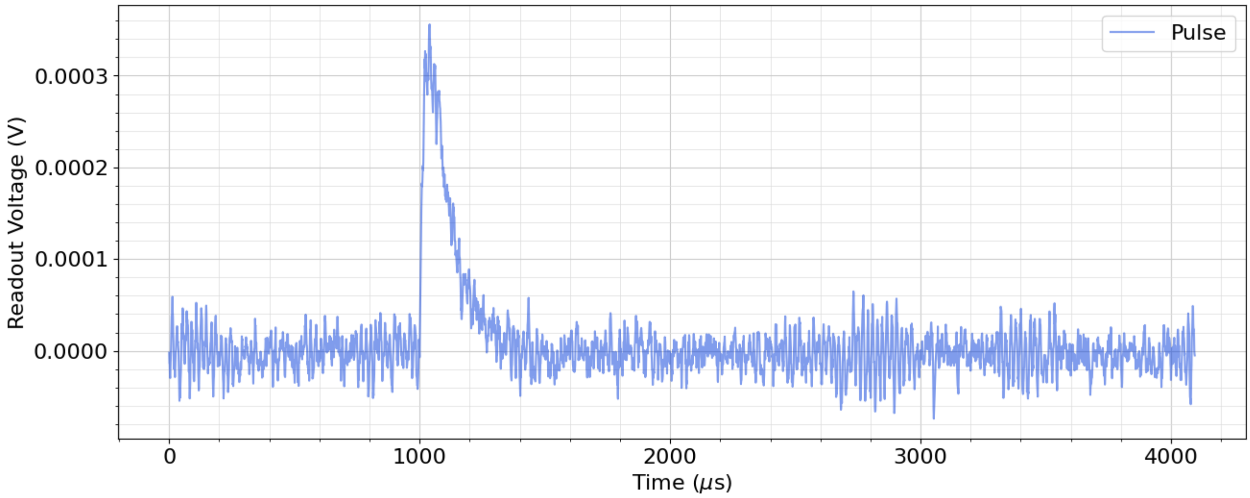
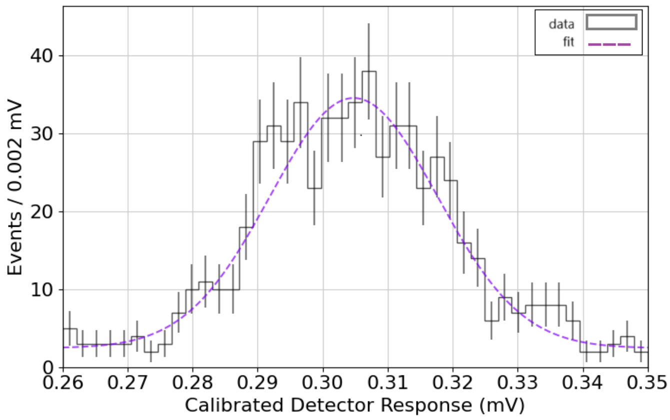


Figure 1: A random pulse taken from one of the thousand samples in the calibration data. The plot represents the voltage reading across time. The detector took 4096 samples at a frequency of 1MHz. The pulse onset is precisely at the 1ms ($= 1000\mu s$) mark due to the trigger (assumed to have an uncertainty of $1\mu s$). The pulse follows a $20\mu s$ rise and a $80\mu s$ fall time.



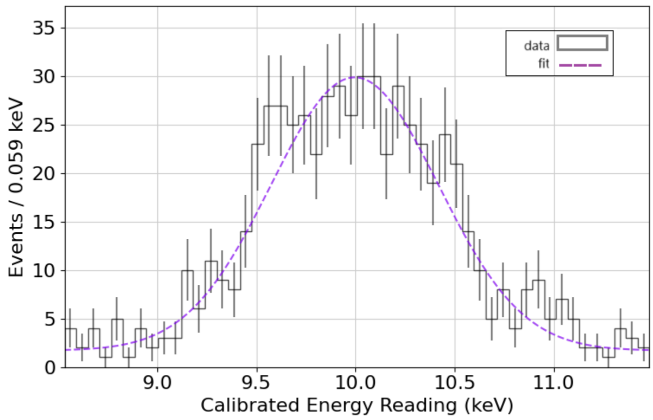


Figure 2: Graphs for the **first energy estimator: max - min reading** of the detector. Both graphs are histograms representing frequency of a certain response at the detector, and purple lines of best fits are *Gaussian* distributions with: the first graph having mean $= 0.305 \pm 0.001$ mV, $\sigma = 0.013 \pm 0.001$ mV, χ^2/DOF $= 0.82$ and χ^2 prob.$= 0.8$. After calibration, the right graph follows with mean $= 9.996 \pm 0.014$ keV, $\sigma = 0.422 \pm 0.052$ keV, χ^2/DOF $= 0.91$ and χ^2 prob.$= 0.6$. The residuals were pattern-less and evenly distributed but not perfect, so overall it is a decent fit with a good resolution for the sensor.

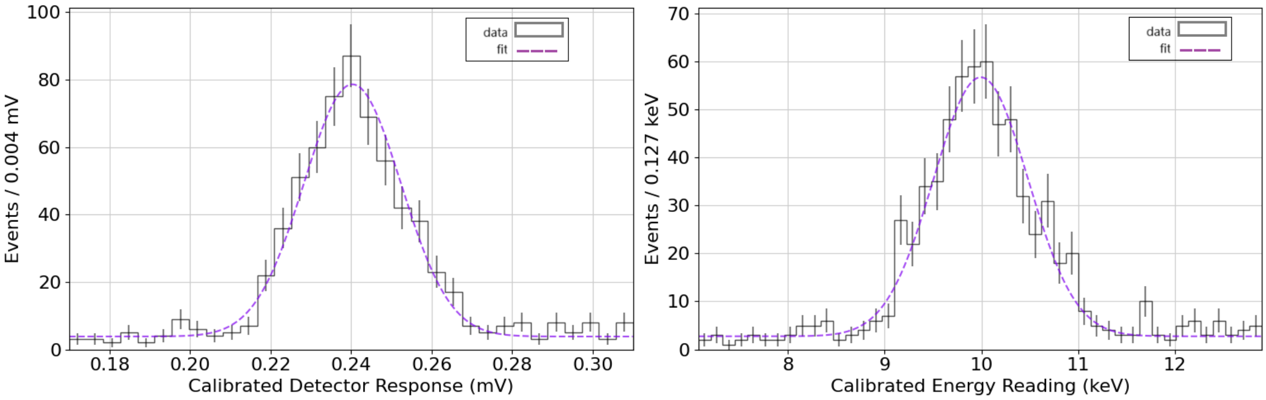


Figure 3: Graph for the **second energy estimator: max - baseline reading** of the detector. Both graphs are histograms representing frequency of a certain response at the detector, and purple lines of best fits are *Gaussian* distributions with: first graph with mean $= 0.240 \pm 0.001$ mV, $\sigma = 0.012 \pm 0.001$ mV, χ^2/DOF $= 1.00$ and χ^2 prob.$=$ 0.5. After calibration, the right graph follows with mean $= 9.994 \pm 0.016$ keV, $\sigma = 0.489 \pm 0.058$ keV, χ^2/DOF $= 0.95$ and χ^2 prob.$=$ 0.6. The residuals were extremely random and evenly distributed. So overall it is a very good fit from the chi-squared values and residuals with a good resolution value.

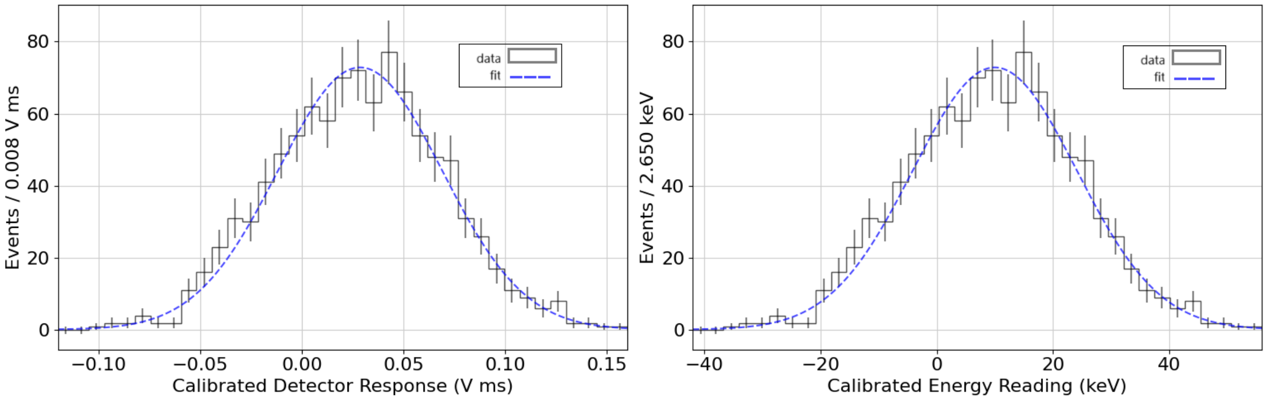


Figure 4: Graph for the **third energy estimator: whole-trace sum integral** for each pulse. Both graphs are histograms representing frequency of a certain response at the detector, and blue lines of best fits are *Gaussian* distributions: first graph with mean $= 0.029 \pm 0.001$ V ms, $\sigma = 0.040 \pm 0.001$ V ms, χ^2/DOF $= 0.71$ and χ^2 prob.$=$ 0.9. After calibration to energy response, the right graph follows with mean $= 10.000 \pm 0.456$ keV, $\sigma = 14.138 \pm 0.645$ keV, χ^2/DOF $= 0.71$ and χ^2 prob.$=$ 0.9. The residuals were pattern-less and evenly distributed, so overall it is an okay fit with bad resolution for the sensor as σ is high.

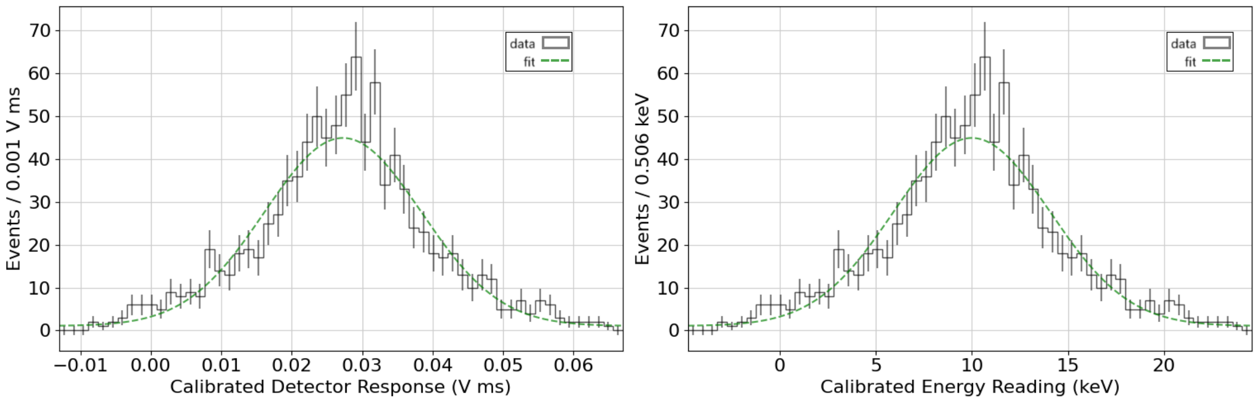


Figure 5: Graph for the **fourth energy estimator: whole-trace sum integral minus baseline** for each pulse. Both graphs are histograms representing frequency of a certain response at the detector, and green lines of best fits are *Gaussian* distributions: The first graph has mean $= 0.027 \pm 0.001$ V ms, $\sigma = 0.011 \pm 0.001$ V ms, χ^2/DOF $= 1.02$ and χ^2 prob.$=$ 0.4. After calibration, the right graph follows with mean $= 10.000 \pm 0.131$ keV, $\sigma = 4.075 \pm 0.054$ keV, χ^2/DOF $= 1.02$ and χ^2 prob.$=$ 0.4. The residuals showed a pattern (see Appendix), so despite good chi-squared values it is a bad fit with an okay resolution for the sensor.

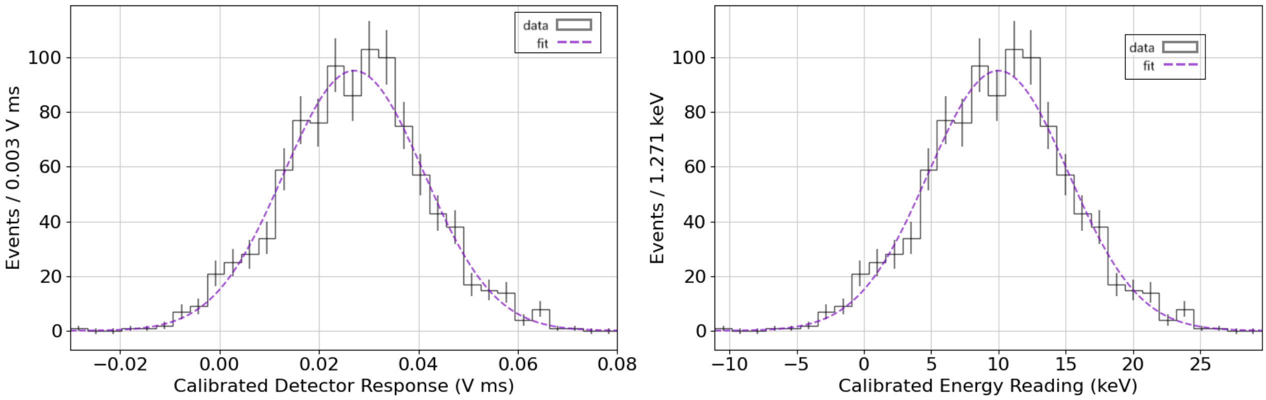


Figure 6: Graph for the **fifth energy estimator: localized sum integral** for each pulse (1ms - 2ms). Both graphs are histograms representing frequency of a certain response at the detector, and the purple lines of best fits are *Gaussian* distributions with: left graph with mean = 0.027 ± 0.001 V ms, $\sigma = 0.014 \pm 0.001$ V ms, χ^2/DOF = 0.83 and χ^2 prob.= 0.7. After calibration, right graph with mean = 10.000 ± 0.168 keV, $\sigma = 5.203 \pm 0.088$ keV, χ^2/DOF = 0.83 and χ^2 prob.= 0.7. The residuals were evenly distributed and pattern-less, so overall it is a decent fit with an okay resolution for the sensor.

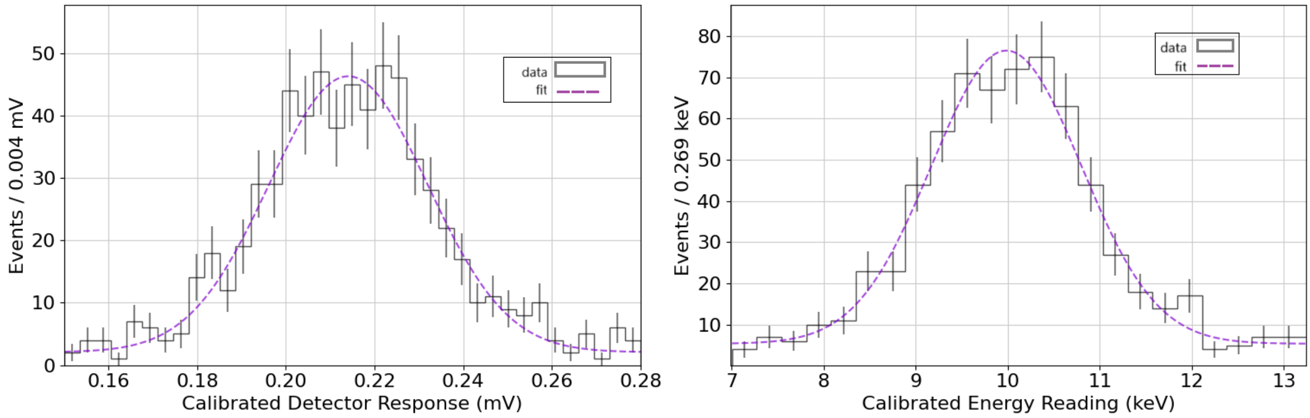


Figure 7: Graph for the **sixth energy estimator: amplitude of the chi-squared fit** of each pulse. Both graphs are histograms representing frequency of a certain response at the detector, and purple lines of best fits as *Gaussian* distributions with: The first graph has mean = 0.214 ± 0.001 mV, $\sigma = 0.018 \pm 0.001$ mV, χ^2/DOF = 0.96 and χ^2 prob.= 0.5. After calibration, right graph with mean = 9.986 ± 0.026 keV, $\sigma = 0.815 \pm 0.021$ keV, χ^2/DOF = 0.83 and χ^2 prob.= 0.9. The residuals were pattern-less and evenly distributed but not perfect, so overall with the chi-squared values it is a decent fit with a good resolution for the sensor.

After the six trials, we chose the Max - baseline amplitude to be the best choice for energy estimation. As seen in figure 3, it has very good chi-squared values, balanced residuals which are also the smallest, and gives us a small resolution for the detector. We used this to interpret and plot the noise data and the signal data to understand their functional form:

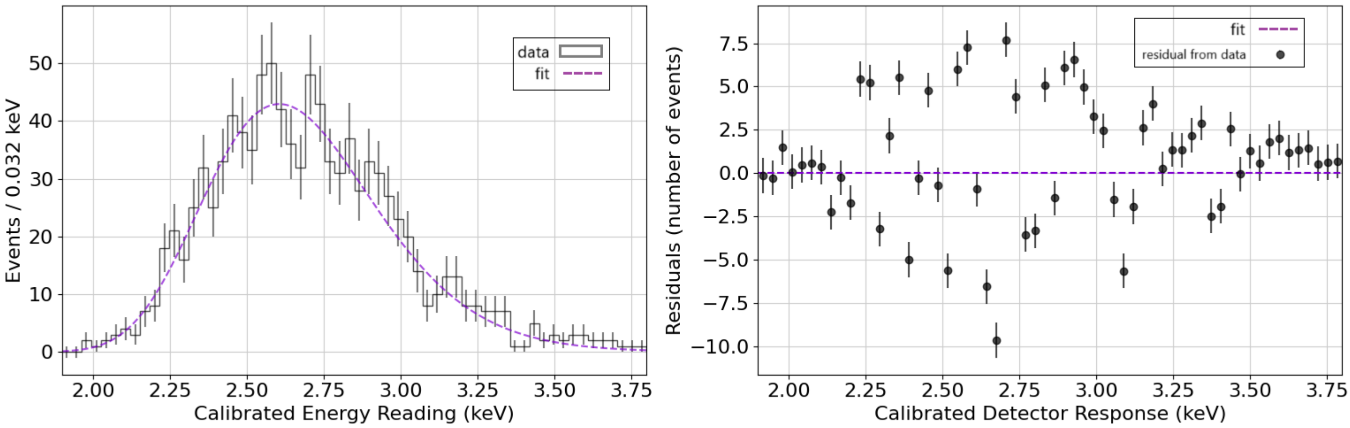


Figure 8: Energy level histogram & line of best-fit with residual for the noise dataset after calibration from estimator 2. Instead of a Gaussian, we modelled this to a *Log-normal* distribution which is used for the fit (similar to gamma distribution but easier to calculate). The optimized function has mean 2.609 ± 0.036 keV and $\sigma = -1.131 \pm 0.017$ keV with χ^2/DOF $= 0.91$ and χ^2 prob.$= 0.7$. As the chi-squared vales are good, and the residuals are randomly distributed (with a slight bias towards positive, which we attribute to too many bins), this is a pretty good fit.

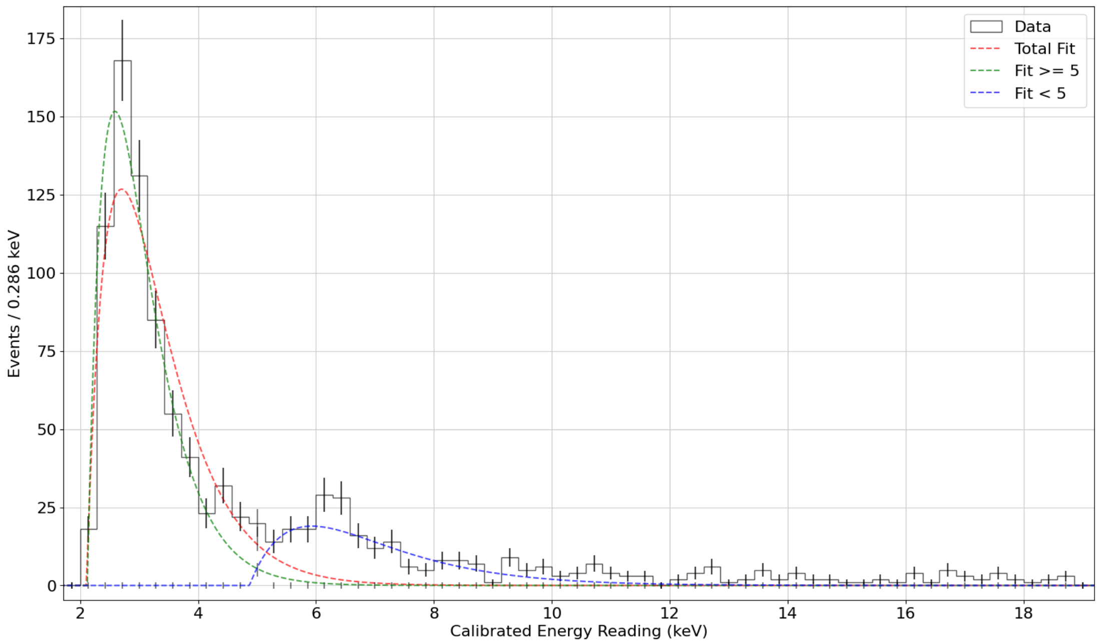


Figure 9: Calibrated signal dataset plot with an overall line of best-fit and two piece-wise lines of best fit: one for readings below 5 keV and one for above 5keV. (Residuals plotted in 11, see appendix). The function used was an exponential function with an offset. Peaks: overall at 2.71 ± 0.02 keV, less than 5 keV at 2.58 ± 0.02 keV, greater than 5 at 5.93 ± 0.05 keV. The χ^2/DOF and χ^2 probability values were: (4.22, 0.0) for overall, (0.73, 1.0) for less than 5, and (1.55, 0.0) for more than 5 respectively. This tells us that the overall fit was not the right function to fit as the probability is zero. Same goes for the fit for greater than 5keV, but with a more reasonable reduced chi-squared value. The fit for less than 5keV has a better reduced chi-squared value, but the high probability tells us that the function was most likely over-fit to the data.This tells us we need more data points, which makes sense considering we only took a partial "favourable portion of the data".

4 Discussion

Overall the energy estimator for amplitude as max minus baseline gave us the best results, which was probably due to a linear measurement and accounting for the noise.

Other energy estimators seemed to be less reliable, especially the area ones. As the integral was taken as the sum of the trace, there was a larger impact of the noise. Another thing of note was that the chi-squared values for area estimators were much less altered by the calibration which we expect was due to the linear scaling on the whole rather than just the max reading.

The noise being a lognormal fit indicates its probably close to the 2.6 keV mean as calculated, but is more likely to be more than not. But we question the reliability of this due to it being just the peak of the noise, than the overall noise across the period, which would impact the detector reading. Nonetheless it seems to be somewhat reliable.

The signal was the most confusing part as the output was a decaying exponential peak with much smaller peaks, which is why we used piece-wise distribution. This is in line with what is commonly seen in detectors with high energy particles as a result of the detector material being ionized from radiation. The fit was bad, but the histogram plot still gives us considerable amounts of information about the signal.

5 References

Provided Lab Manual titled "PHY324 Data Analysis Project, Winter 2024"

Log-normal distribution (2024) Wikipedia. Retrieved on October 7, 2024 from: https://en.wikipedia.org/wiki/Log-normal_distribution

6 Acknowledgement of AI Usage

The author of this paper did not use any AI tools for data analysis, writing code, generating figures or for writing the report.

7 Appendix

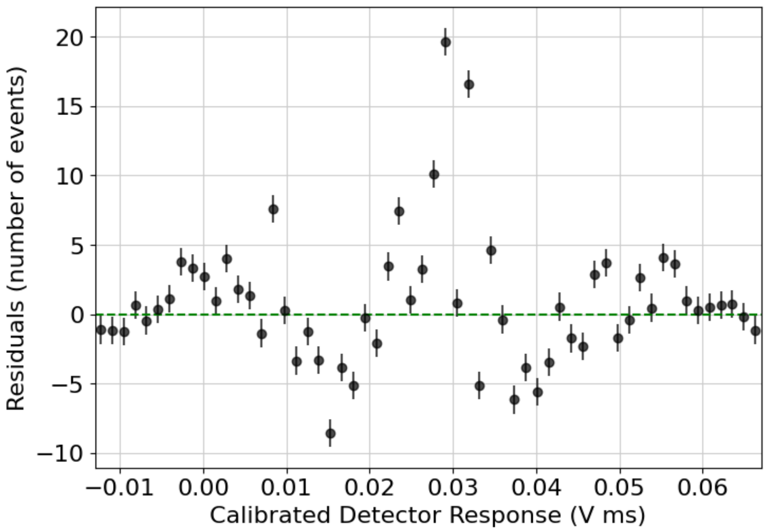


Figure 10: Residual plot for energy estimator 4, which seems to have a sinusoidal pattern. Patterns indicate the fit is not good and has a bad model.

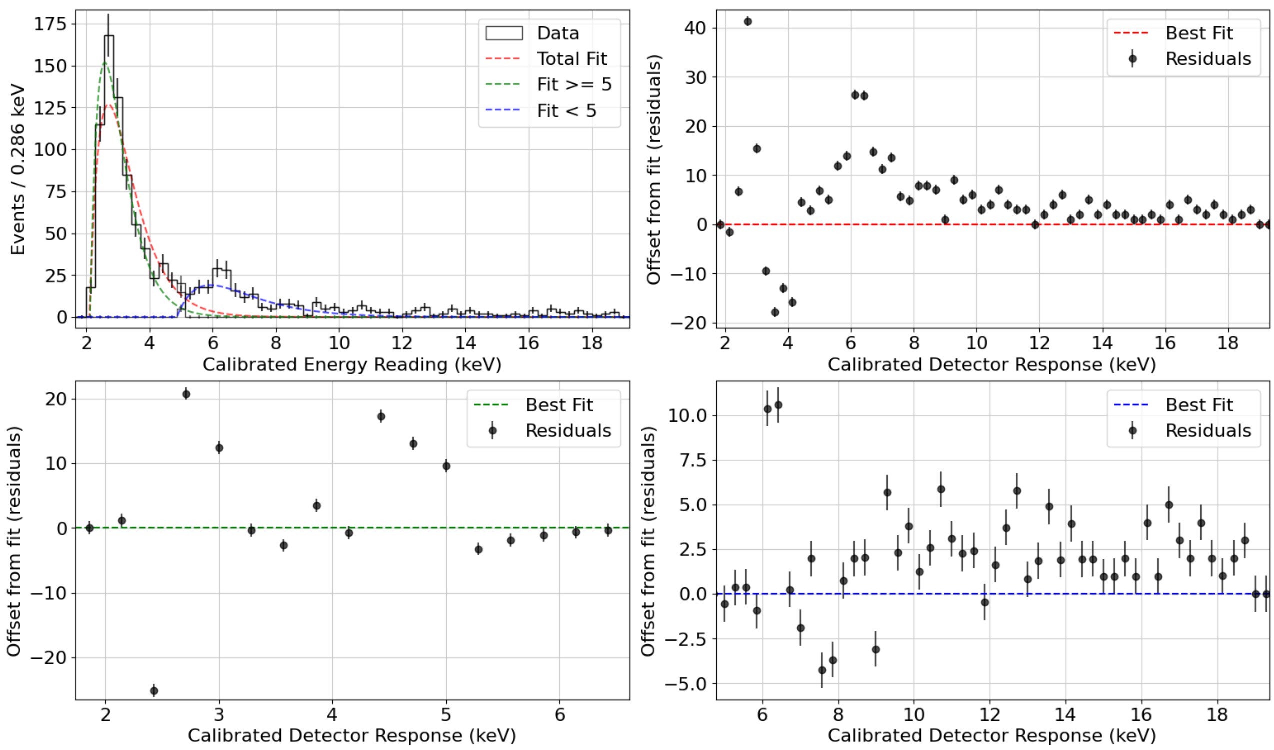


Figure 11: Residual plots for each of the full-spectrum model of the signal.