Data Analysis Project

Shumail Farooqi
PHY324

February 20, 2025

Abstract

To figure out an unknown particle from some signal data, an algorithm was created to estimate the energy of the particle using a particle detector. Data from a particle of known energy (10keV) was used to create some energy estimators. The method that yielded the best result of the amplitude was obtained by subtracting the baseline average from the maximum of the pulse with the conversion factor $C = 41.6keV/mV$. The data was then calibrated and converted back into energy spectrum. Then the procedure was applied to the unknown particle to analyze the energy spectrum of the particle. The method yields result with high resolution for the amplitude of pulses of $0.0117mV \pm 5 \cdot 10^{-4}mV$ and a chi-squared value of $\chi^2 = 1.27$. When the calibration factor is used to determine the energy of unknown particle with the histogram fitted with a sum of skewed and a normal gaussian we find get the goodness of fit of $\chi^2 = 3.87$.

Introduction

We are given three sets of data, Calibration data, noise data, and signal data. All of which are collected using a particle detector by converting the incoming energy of the particle into electrical signals recorded as pulses overtime[1]. The device is set to record a pulse with rise time of $\tau_{rise} = 20\mu s$ and decay time of $\tau_{fall} = 80\mu s$ the form of the resulting voltage readout is given by eqn 1

$$V = A \cdot C \left(e^{-t/\tau_{rise}} - e^{-t/\tau_{fall}} \right) \tag{1}$$

where t is some instance of time during the recording of the pulse, A gives the amplitude of the pulse, and C is the normalization factor [1]. C is given by

$$C = \left(\frac{\tau_{fall}}{\tau_{rise}} \right)^{-t/(\tau_{fall} - \tau_{rise})} \left(\frac{\tau_{rise} - \tau_{fall}}{\tau_{fall}} \right) \tag{2}$$

An idealized pulse data collected from this data acquisition device without any background noise is visualized in figure(1),

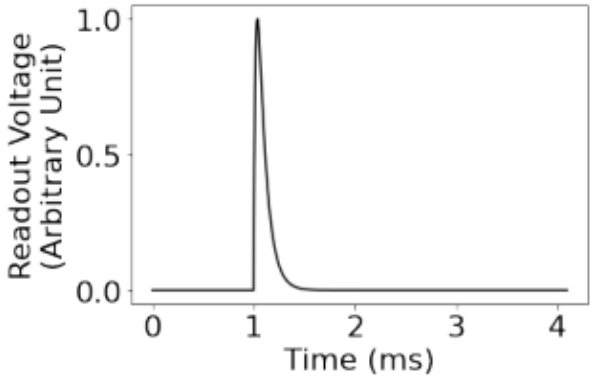

Figure 1: Example of an ideal pulse [1] without any background noise. Pulse is stored as 4096 samples over 4 ms but the pulse is set to start at 1 ms

On the other hand of course in real life scenarios there is background noise which does not give a pulse like the figure above. A more realistic pulse is given in figure 2. The voltage is recorded such that the pulse begins after the 1ms mark (1000 samples). The first 1000 sample data is the baseline which consists of the

background noise in the process of particle detection. Knowing the average behavior of this baseline will be essential in some of the methods used to estimate the energy of the particle.

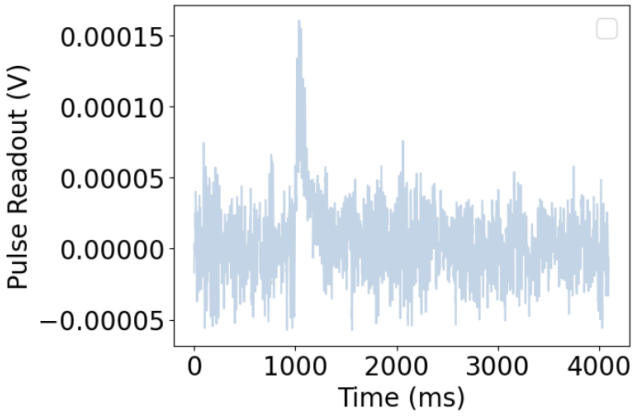

Figure 2: Realistic Pulse voltage readout in voltage, first file from calibration-pk data set given. The pulse is from a source emitting particles with 10keV energy. The pulse has been set to start at 1000 (ms) mark. The whole pulse is recorded as 4096 samples

The goal of this project is to build an algorithm to determine the particles being fired in signal data. Calibration data set contains voltage readouts from a particle detector being fired by a particle with energy 10keV, this data will be used to calibrate the detector, we will use the calibration data to build energy estimators and convert the readout voltage to the energy spectrum of the particle with energy $10keV$ [1]. The energy spectrum will be in the form of histograms (energy vs. events) and using the Gaussian function given below we will fit to the histogram.

$$f(x) = Ae^{-\left(\frac{x-\mu}{\sigma}\right)^2} + B \tag{3}$$

where μ is the mean of the Gaussian which is also the peak, σ the resolution of the Gaussian, A the amplitude and, B the baseline.

Using the calibrated detector and the best calibration energy estimator we will analyze the signal data, which is the data set from a particle we would like to identify. The last data set if of background noise, which will be used to reduced the uncertainties and the background noise from the signal data and give us a better result on the energy of the unknown particle.

Methods

Methods of Energy Estimators

Using the 100 traces of pulses from the calibration data we will create a histogram of readout voltage (amplitude) vs. events using various methods that estimate the energy of the pulse. There will be multiple estimations of the energy using different methods and we will calibrate the methods using the known energy of the particle (10 keV)[1].

Creating and Calibrating the Energy Spectrum

Using the methods described above we can create histograms of amplitude vs. events. A Gaussian will be fit through the histograms and this goodness of fit will help determine an energy estimation method that yields the best result. Using the best estimator we can finally analyze the signal data from the unknown particle.

One method to estimate the energy is by taking the max readout of the pulse, this gives us an amplitude of the energy considering the particle detector device is set to keep the baseline average around 0 voltage mark. For an idealized pulse like in figure 1 this works very well, but since we have background noise it is better to take the amplitude to be the maximum voltage - minimum voltage. This will be our first method we use to determine the energy. The background noise deviates the baseline average from the 0 mark, therefore we can obtain the amplitude by taking the max readout voltage - baseline average.

Another approach we used to determine the energy is through summing the values in the trace. Idea here is that we are estimating the energy as the integral over the pulse by taking the integral over the whole trace. This sort of works since the baseline is around 0 readout voltage so the integral over these samples should in theory be 0. To improve this approach we can subtract the baseline average value from the values to be summed over for the integral. This reduces the effect of background noise and helps improve the resolution of the estimation. Alternatively we can also only sum the part of trace which corresponds to the pulse data, that is data recorded after the first 1000 samples, this will be the last method we implement using the approach of the integral.

Lastly, we will estimate the energy by fitting the function 1 to the pulse data. We will fit each of the 1000 traces from calibration data and then store the amplitudes $(A \cdot C)$ as the amplitude for the pulses.

Since we know that the particles being fired is with energy 10 keV, we can allign the energy spectrum to the actual energy value of the particle. We do this by shifting the x-axis such that the most likely point of the Gaussian is at 10keV, this requires us to multiply the amplitudes by a conversion factor. The conversion factor is given by

$$C = \frac{10keV}{\mu} \tag{4}$$

where μ is the mean of the Gaussian (its peak) and it depends on the energy estimate method.

Fitting Signal Data

After finding the best method of estimation and conversion factor, the signal data was analyzed. To fit the energy spectrum of the signal data the function 5 was used

$$f(x) = 2A\phi\left(x-\epsilon\right)\Phi\left(\alpha(x-\epsilon)\right) + B\phi\left(x-\beta\right) \tag{5}$$

where $\phi = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$ Gaussian, $\Phi = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right)\right]$, where erf is the error function. This function is a sum of skewed Gaussian $2\phi(x)\Phi(\alpha x)$ and a Gaussian $\phi(x)$. α is the skew factor, if $\alpha > 0$ the resulting distribution is right skewed and if $\alpha < 0$ then we have a left skewed distribution. ϵ is the most probable point of the skewed Gaussian and β is the most probable mean of the second Gaussian, and lastly A, B are the amplitudes of the two Gaussians. [2]

Results and Analysis

All six method of estimating the amplitude of the pulse were implemented and the detailed results are given in the appendix, with only the method that gave the best results is given in this section. In general the method of taking the max - min of the voltage performed poorer than the method of taking the max - baseline average of the voltage as the amplitude. The former with reduced chi-squared value of $\chi^2 = 2.44$ and latter with $\chi^2 = 1.07$. Which is to be expected as by subtracting the baseline average we are reducing the effects of the background noise.

Although different result in the approach of taking the integral was found, taking the energy as the sum of all trace gave $\chi^2 = 0.64$ and method of subtracting the baseline average from the values to be summed over game $\chi^2 = 1.17$. The former was an over fit of the gaussian most likely due to high uncertainty. As seen in the residual plot 13 we do see a minor pattern thus resulting in poor fit over all.

The second best estimation was given by fitting the pulse data using the function 1 and taking the amplitude from the constants $(A \cdot C)$, chi-squared value of $\chi^2 = 0.72$ residual 27 showed very little pattern, and overall good resolution of $\sigma = 0.83 \pm 0.04$.

Out of all the methods the best in my analysis was given by the method 2, getting the amplitude by subtracting the baseline average from the max of the pulse. The histogram (amplitude vs. events) is given below

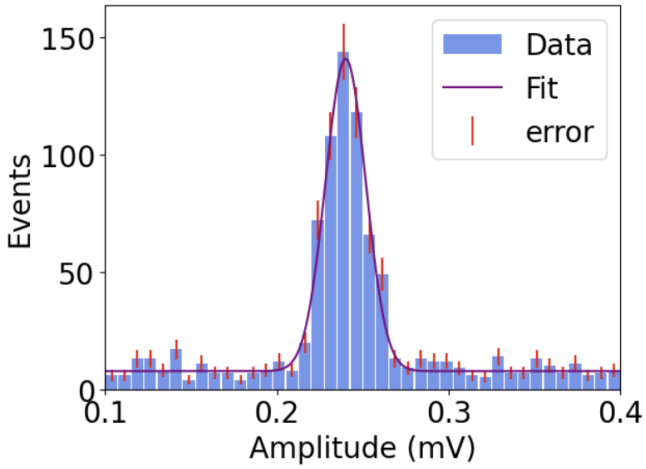

Figure 3: Energy Estimation of pulses from calibration data via max - baseline average method. Fitted using a Gaussian [3] with parameters $A = 133 \pm 8, \mu = 0.24 \pm 6 \cdot 10^{-4}, \sigma = -0.011 \pm 5 \cdot 10^{-4}, B = 7.7 \pm 0.5$ and reduced chi-squared of $\chi^2 = 1.28$. The standard deviation is given by $\sigma = -0.011 \pm 5 \cdot 10^{-4}$

Most of the uncertainty in the parameters comes from the low count bins of far left and right ends of the histogram, which also have low data uncertainty so the fit from curve fit function values fitting these points more than the points from bins with higher count. To address this problem the bins with uncertainty < 2.5 was given an uncertainty of 2.5 so the fit is better for the more important points in the middle.

Now since new know the particle is of energy $10keV$ we can shift the peak of the Gaussian to $10keV$ which requires multiplying the amplitudes by a conversion factor (4) as mentioned before. This gives us a new histogram.

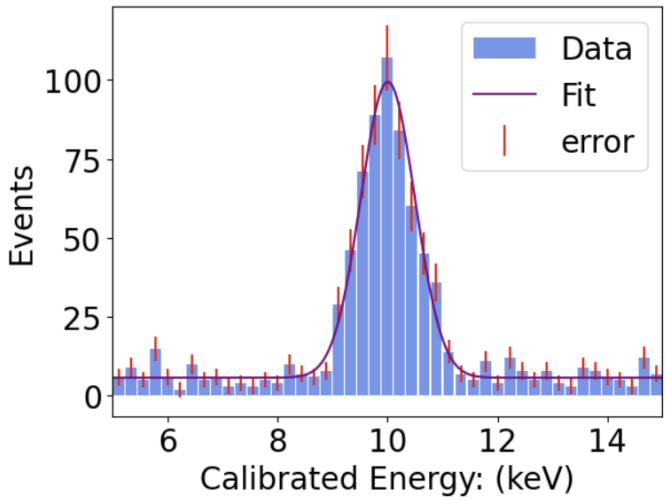

Figure 4: Calibrated Energy Estimation of pulses from calibration data via max - baseline average method. Fitted using a Gaussian [3] with parameters $A = 94 \pm 5, \mu = 10.00 \pm 0.03, \sigma = -0.49 \pm 0.02, B = 5.8 \pm 0.5$ and reduced chi-squared of $\chi^2 = 1.07$. The standard deviation is given by $\sigma = 0.49 \pm 0.02$

Again similar to the previous histogram some of the bins from both far left and right ends were removed as it was effecting the goodness of fit to produce a better fit.

Another thing to consider is the bin width and numbers of bins in the histogram which will affect the shape of the Gaussian and the fit of Gaussian. All the histograms were done in this experiment with around 35 bins, but all of the histograms have some bins removed from each ends, bins with low event count as they affect the goodness of fit. The resolution of this method which is given by the width of the Gaussian is $\sigma = 0.493keV \pm 0.0213$ The residuals of this histogram and the fit also showed no pattern which implies the fit was good.

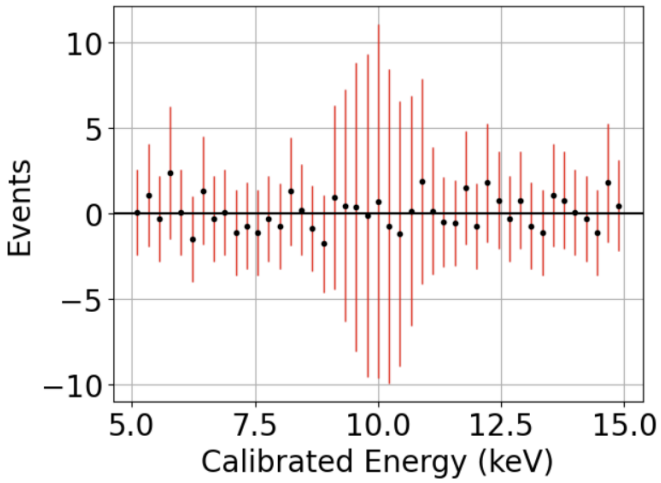

Figure 5: (Energy vs. Events) Residuals of Calibrated Energy Estimation of pulses from calibration data via max - baseline average method. The residuals shows no pattern, high uncertainty for the points in the middle.

Some of the results from other methods is given in the table below, the rest of the information for these methods can be found in the appendix including the histograms generated and the parameters used to fit the histogram.

Table 1: Summary of Energy estimators

Estimator Method	Calibration Factor	reduced χ^2	Resolution
Max - min	33.2 keV/mV	2.26	0.0141 mV $\pm 6 \cdot 10^{-4}$ mV
Max - Baseline	41.6 keV/mV	1.27	0.1170 mV $\pm 5 \cdot 10^{-4}$ mV
Sum of all values	0.358 keV/mV	0.76	40 mV ± 2 mV
Sum - Baseline	0.364 keV/mV	1.25	11.3 mV ± 0.4 mV
Sum of pulse	0.375 keV/mV	1.46	30 mV± 1 mV
Fitting pulse	47.6 keV/mV	1.38	0.0170 mV $\pm 8 \cdot 10^{-4}$ mV

The same method of obtaining the amplitude of the pulse alongside the conversion factor (4) was used on the signal data to calibrate the energy of the unknown particle. The histogram using the max - baseline average method and a calibration factor of $41.6 keV/mV$ is given in the figure **??**

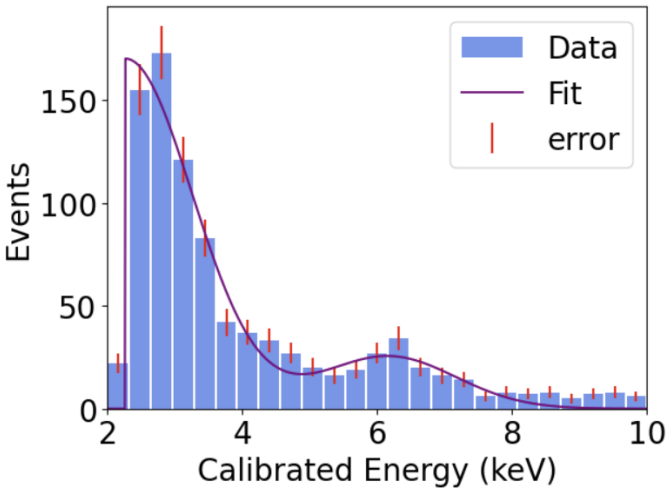

Figure 6: Signal data calibrated using the max - baseline average method. Histogram fitted using a sum of skewed Gaussian and a normal Gaussian as introduced in equation[5] with parameters $A = 213 \pm 3, B = 64 \pm 3, \epsilon = 2.26 \pm 0.07, \beta = 6.2 \pm 0.1, \alpha = 1.2 \cdot 10^5$, and reduced chi-squared of $\chi^2 = 3.87$

Some of bins on the left and right end of the histogram are removed to help the curve fit function fit the distribution better. This fit with $\chi^2 = 3.87$ is good considering the function used in curve fit might not be the optimal function for such distributions. The residuals are given by

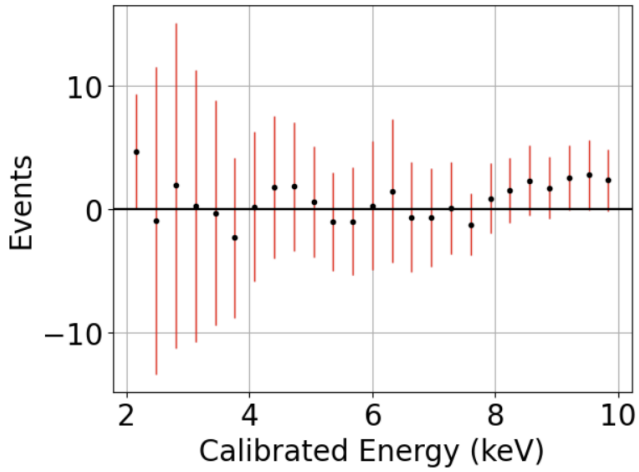

Figure 7: (Energy vs. Events) Signal data calibrated residuals using the max - baseline average method.

The residuals do show some minor pattern but nevertheless the fit is decent as argued above due to it's good resolution of $|\sigma| = 0.49 \pm 0.02$.

Conclusion

Several energy estimation method were developed to determine the correct calibration factor using calibration data from a particle of enery 10 keV.

Energy estimation algorithm that gave the best result out of the estimation methods was by subtracting the baseline average from the max of the pulse, which give the amplitude of the pulse. This method gave the conversion factor of $41.6keV/mV$ between voltage and energy. The method was ultimately chosen due to the goodness of its fit with $\chi^2 = 1.07$, resolution of the Gaussian (histogram of calibration data using method 2) being $0.493keV \pm 0.0213keV$, and the residuals showed no pattern.

Coming in second best method was of fitting the pulse, which gave the calibration factor of 47.6 keV/mV. The histogram was fitted using a gaussian with $\chi^2 = 0.72$ and resolution of $\sigma = 0.83 \pm 0.04$, but ultimately the other method was selected to be used due to its better resolution (lower $\sigma = 0.493 \pm 0.021$).

This method of max - baseline average was used on the signal data to calibrate it's energy. The energy spectrum of signal data was fitted using the sum of skewed Gaussian and a normal Gaussian [5] with parameters $A = 213 \pm 3, B = 64 \pm 3, \epsilon = 2.26 \pm 0.07, \beta = 6 \pm 0.1, \alpha = 1.2 \cdot 10^5$, and reduced chi-squared of $\chi^2 = 3.87$.

Biggest source of uncertainty in this analysis comes from creating the histograms where no algorithms were used to determine the number of bins or the bin width. While most of the histograms had around 35 bins, some of the bins with low event count on the left and right ends of the Gaussian were taken out for a better fit. Much of these adjustment to the histograms is to the discretion of the experimenter.

References

[1] *Data Analysis Project Documentation*. University of Toronto, 2025.

[2] Tom O'hagen A; Leonard. *Bayes estimation subject to uncertainty about the parameter constraints*. Biometrika. 63(1): 201-203, 1976.

Use of AI Statement

No Artificial Intelligence tool was used to make this report

Appendix

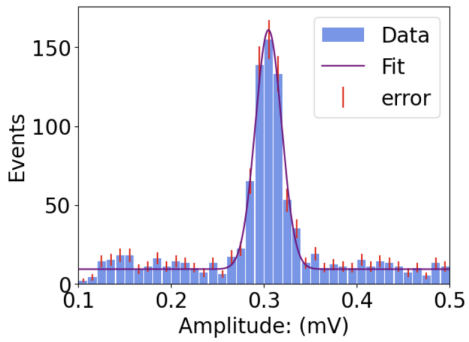

Figure 8: Energy Estimation of pulses from calibration data via max - min method. Fitted using a Gaussian [3] with parameters $A = 152 \pm 9, \mu = 0 \pm 7 \cdot 10^{-4}, \sigma = 0.0141 \pm 5.96 \cdot 10^{-4}, B = 9.2 \pm 0.5$

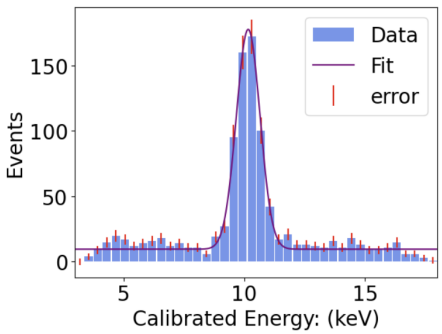

Figure 9: Calibrated Energy Estimation of pulses from calibration data via max - min method. Fitted using a Gaussian [3] with parameters $A = 168 \pm 9, \mu = 10.00 \pm 0.02, \sigma = 0.49 \pm 0.02, B = 9.6 \pm 0.6$ and the reduced chisquared is $\chi^2 = 2.44$

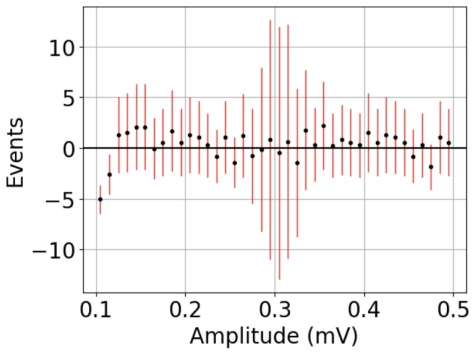

Figure 10: (Amplitude vs. Events) Residual of energy estimation of pulses from calibration data via max - min method.

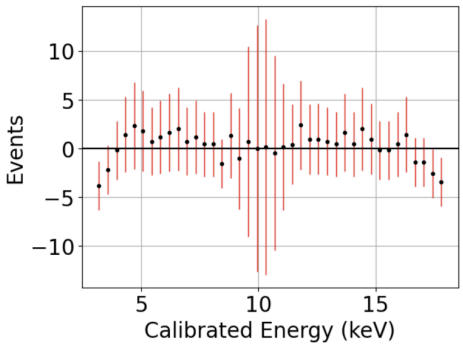

Figure 11: (Energy vs. Events) Residuals of calibrated energy estimation of pulses from calibration data via max - min method.

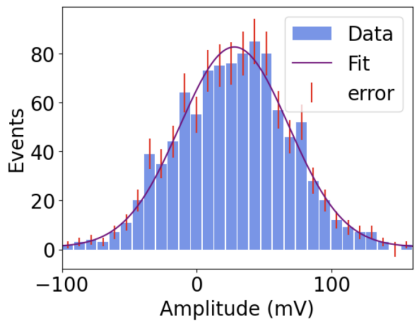

Figure 12: Energy Estimation of pulses from calibration data via sum all method. Fitted using a Gaussian [3] with parameters $A = 82 \pm 3, \mu = 28 \pm 1, \sigma = 40 \pm 2, B = 0.9 \pm 0.9$

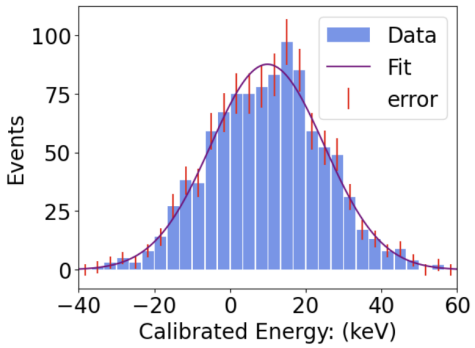

Figure 13: Calibrated Energy Estimation of pulses from calibration data via sum over all method. Fitted using a Gaussian [3] with parameters $A = 88 \pm 4, \mu = 10 \pm 0.5, \sigma = 14.9 \pm 0.6, B = 0 \pm 1$ and the reduced chisquared is $\chi^2 = 0.636$

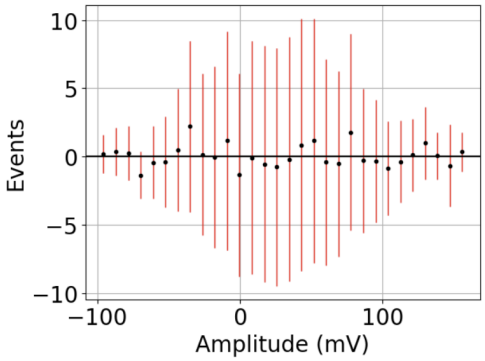

Figure 14: (Amplitude vs. Events) Residual of energy estimation of pulses from calibration data via sum all method.

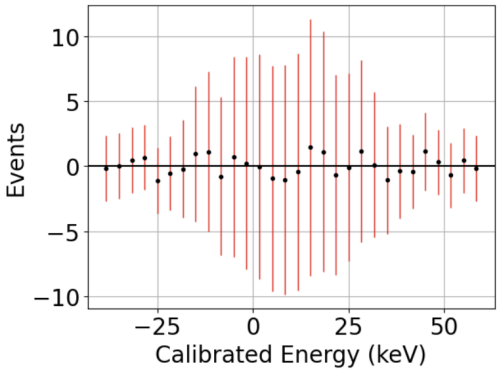

Figure 15: (Energy vs. Events) Residuals of calibrated energy estimation of pulses from calibration data via sum all method.

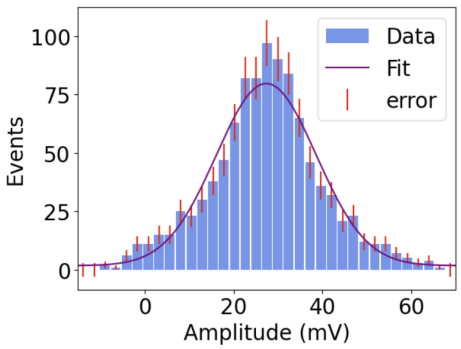

Figure 16: Energy Estimation of pulses from calibration data via sum of all (values - baseline average) method. Fitted using a Gaussian [3] with parameters $A = 78 \pm 3, \mu = 27.4 \pm 0.4, \sigma = 11.3 \pm 0.4, B = 1.7 \pm 0.6$

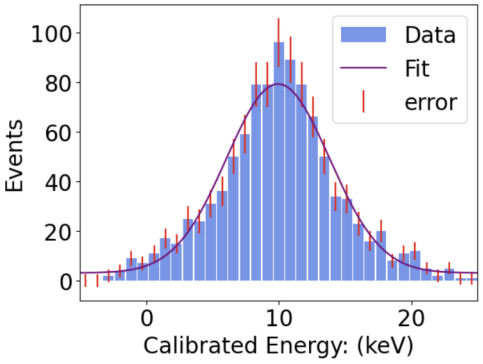

Figure 17: Calibrated Energy Estimation of pulses from calibration data via sum of all (values - baseline average) method. Fitted using a Gaussian [3] with parameters $A = 76 \pm 3, \mu = 10.0 \pm 0.1, \sigma = 3.9 \pm 0.2, B = 3 \pm 1$ and the reduced chisquared is $\chi^2 = 1.17$

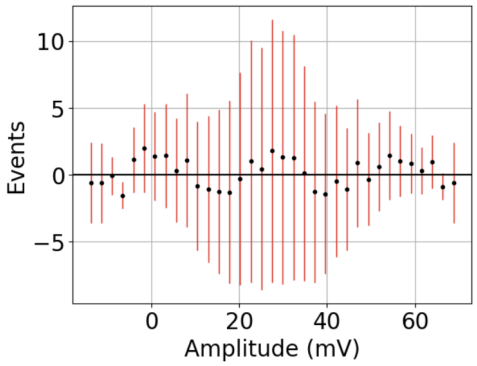

Figure 18: (Amplitude vs. Events) Residual of energy estimation of pulses from calibration data via sum of all (values - baseline average) method.

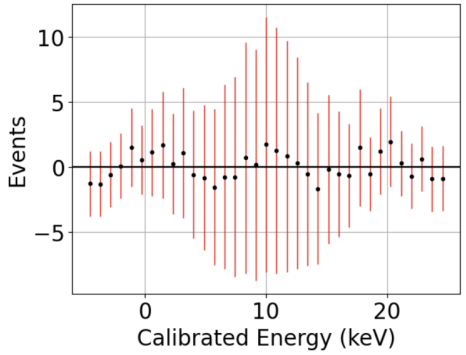

Figure 19: (Energy vs. Events) Residuals of calibrated energy estimation of pulses from calibration data via sum of all (values - baseline average) method.

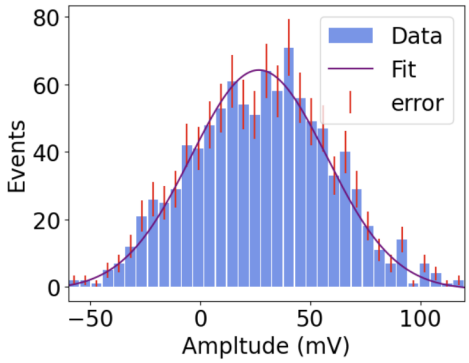

Figure 20: Energy Estimation of pulses from calibration data via sum of only the pulse method. Fitted using a Gaussian [3] with parameters $A = 65 \pm 3, \mu = 27 \pm 1, \sigma = 31 \pm 1, B = -1 \pm 1$

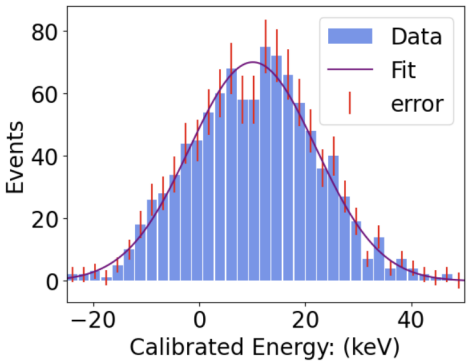

Figure 21: Calibrated Energy Estimation of pulses from calibration data via sum pulse method. Fitted using a Gaussian [3] with parameters $A = 70 \pm 3, \mu = 10.0 \pm 0.4, \sigma = 11.9 \pm 0.5, B = 0 \pm 1$ and the reduced chisquared is $\chi^2 = 0.821$

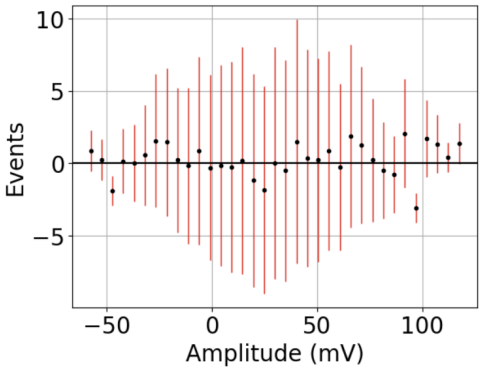

Figure 22: (Amplitude vs. Events) Residual of energy estimation of pulses from calibration data via sum pulse method.

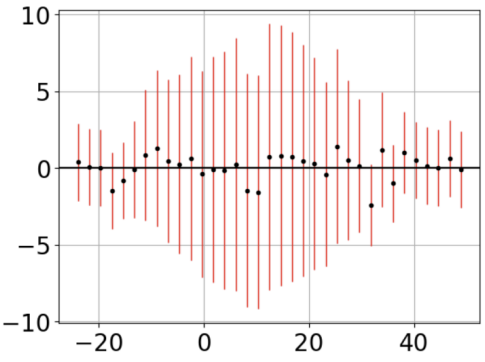

Figure 23: (Energy vs. Events) Residuals of calibrated energy estimation of pulses from calibration data via sum pulse method.

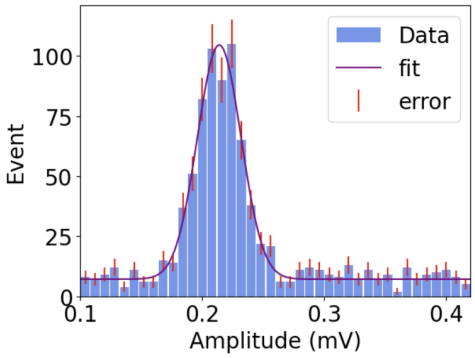

Figure 24: Energy Estimation of pulses from calibration data via fitting pulse method. Fitted using a Gaussian [3] with parameters $A = 97 \pm 6, \mu = 0.213 \pm 9 \cdot 10^{-4}, \sigma = 0.017 \pm 8 \cdot 10^{-4}, B = 7.2 \pm 0.5$

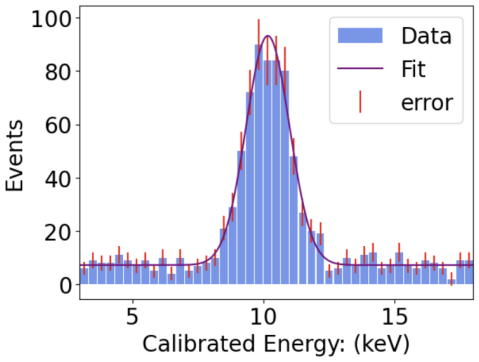

Figure 25: Calibrated Energy Estimation of pulses from calibration data via fitting pulse method. Fitted using a Gaussian [3] with parameters $A = 86 \pm 5, \mu = 10.0 \pm 0.4, \sigma = 0.83 \pm 0.04, B = 7.2 \pm 0.5$ and the reduced chisquared is $\chi^2 = 0.718$

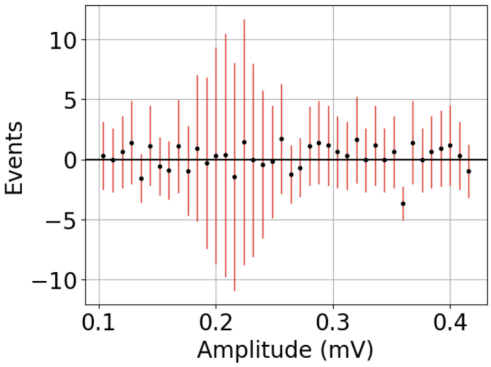

Figure 26: (Amplitude vs. Events) Residual of energy estimation of pulses from calibration data via fitting pulse method.

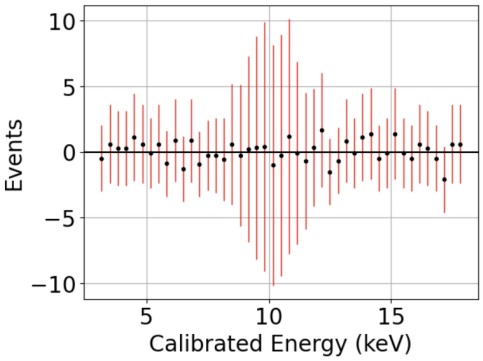

Figure 27: (Energy vs. Events) Residuals of calibrated energy estimation of pulses from calibration data via fitting pulse method.