PHY324 Data Analysis (Resubmission)

Monday, April 10th, 2023 (1 Week Extension)

Jace Alloway - 1006940802

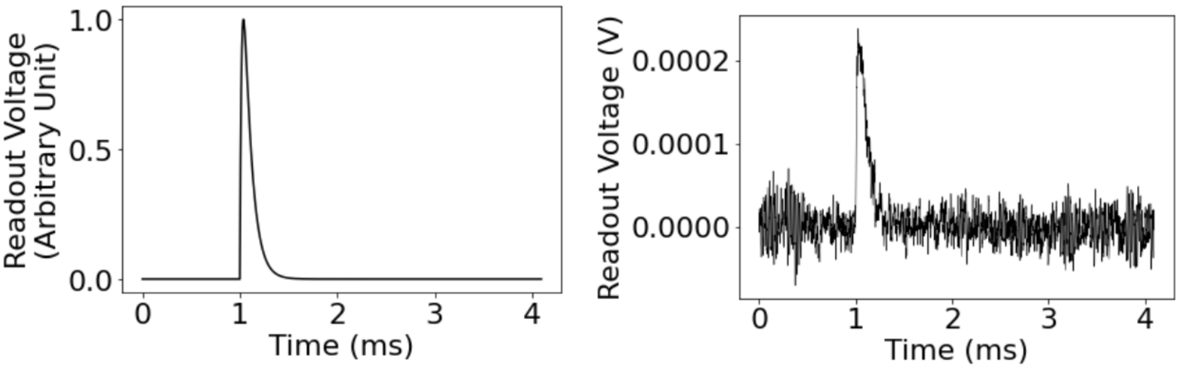
Introduction

In particle physics, scattering experiments often reveal the subatomic makeup of composite particles, revealing key insights of the fundamental building blocks which compose the universe. Scattering experiments are experiments where particles containing high amounts of energy are collided into one another, often resulting in the scattering of other particles due to the energy exchange. The probability of detecting particle remains, after collision, is called the 'cross-section' of the particle which may or may not have been detected.

In most cases, particle detectors operate by detecting the particle's energy which is deposited into the device via electrical signals. These energy detections, called 'pulses', are able to characterize the type of partice which was detected as well as it's momentum, spin, or charge. The goal of this paper is to attempt to characterize the energy of electron recoil scattering initiated by high-energy photon absorption or the Compton scattering of a material[1]. This procedure was motivated by new sensor technology which is able to characterize pulses with a fixed-pulse shape at a 1 MHz rate, whose shape retains the functional form[1]

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

where A is the amplitude of the pulse and V is the readout voltage. In this experiment, due to the nature of electron scattering, it was found that a pulse shape of with a $20\mu s$ rise time and $80\mu s$ fall time was a reasonable model for voltage pulses measured by a sensor[1]. A graphical depiction of a pulse is shown below:



[Figure 1] (Left) The theoretical depiction of an electron-recoil pulse measured with a $20\mu s$ rise time and $80\mu s$ fall time, justified by Equation (1). (Right) A typical pulse recorded by the detector at a 1 MHz rate, including background noise from other environmental decays.

Methodology

Due to the nature of the detector design, each pulse contains a 'pre-pulse' region and an energy measurement (called the 'trace' of the pulse)[1]. Any possible energy perturbation deposited into the detector was designed to trigger the acquisition software to measure the pulse onset as well as the duration of the pulse trace, with a total measurement of 4096 samples per pulse measurement at a 1 MHz[1] rate.

Each pre-pulse region was adjusted to be approximately 0 V in the presence of background voltage, since this determines the appropriate amount of energy deposited into the detector with the 0 volt reference. Since thihs experiment was performed in a controlled lab, background readout voltage is expected, and this was attributed to external decays in the environment (isotopes, etc), temperature fluctations, low-frequency noise, noise in the readout circuit, and other contributions such as random electron motion.

To determine the measured response of energy deposition, a set of calibration data was taken to set a controlled reference upon measuring the energy of an electron recoils. This was completed by exposing the sensor to multiple 10 keV photon-laser pulses, including background contributions[1]. These photon pulses present a group of distribution events located in the region of interest (ROI) of $0 - 20$ keV, the expected ROI for electron-recoil detections. Due to the nature of the experimental setup, it was known that these pulses presented a narrow Gaussian peak around 10 keV[1].

It followed, from the calibration data, to quantify the energy distributions via a group of 'energy estimators': various types of energy-determinations given for all the data trials to determine a given amount of energy deposited into the sensor for each respective trial. Of these energy estimators, are a (i) maximum-baseline (baseline is the average value of the pre-pulse region) amplitude estimate, (ii) maximum-minimum amplitude difference estimate, (iii) integrating across the whole trace of the pulse, (iv) integrating over the trace compared to the baseline value, (v) integrating over the isolated pulse sample range, and (vi) fitting the pulse (with a functional form, such as the one defined in Equation (1)) to estimate the amplitude. Each of these processes are described in data analysis.

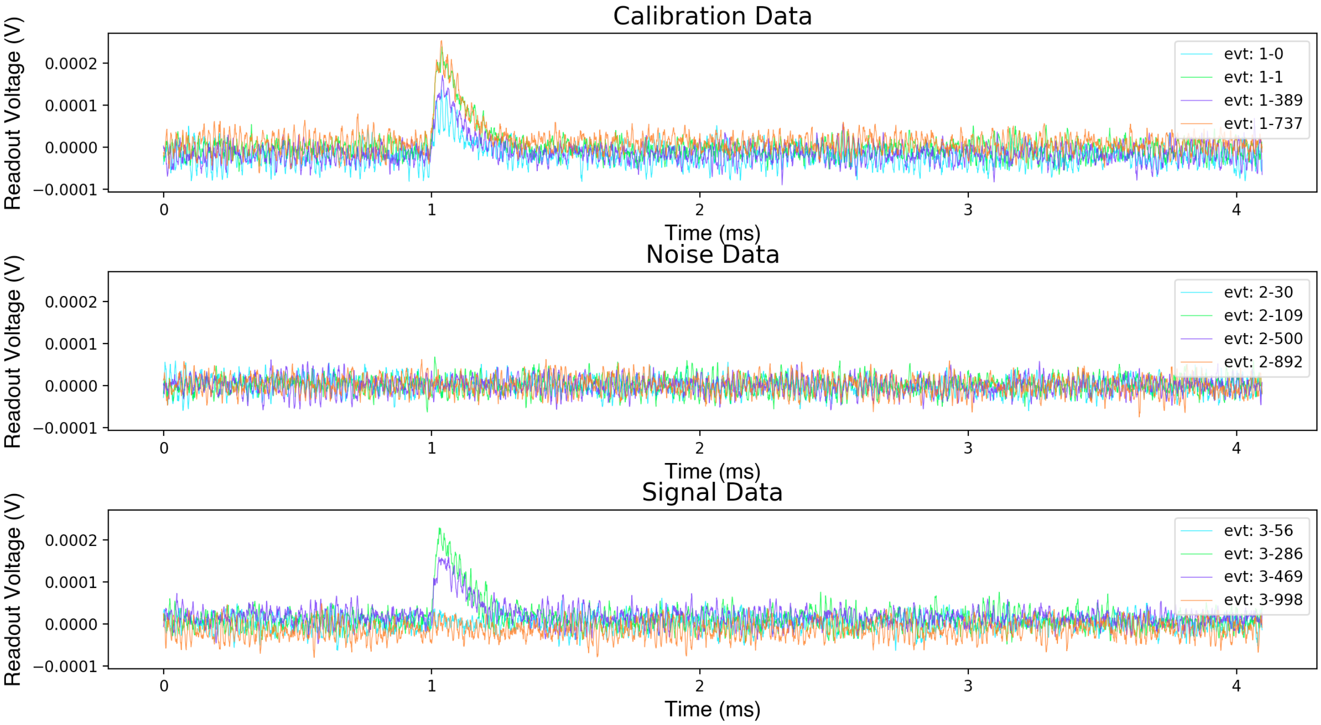
Once each energy estimator was constructed and fitted with a curve. The estimator with the best representation of the calibration data was selected and was calibrated according to the 10 keV measurements. The estimator selected required to have a narrow distribution (ie, a small peak width or 'resolution') with minimal background noise measurements (a narrow peak around 10 keV), since (i) this estimator would appropriately characterize the recoil-electron energy deposition (when applied to the signal data) assuming each electron deposited roughly an equivalent amount of energy into the sensor, and (ii) the background noise amplitudes are small when compared to pulse amplitudes. Then, the selected energy estimator was applied to the acquired data (signal data) from the electron measurements.

Data Analysis

Altogether, the acquisition device obtained 1000 perturbation-triggered pulses from each the electron-recoil measurements, the calibration photon detections, and a controlled set of background data. Each sample contained 4096 samples, with a 1000-sample pre-pulse region. Due to the 1 MHz nature of the detector, each 1000 samples corresponded to 1 ms worth of real-time data.

The first step in data processing was to load and plot sample measurements acquired from the detector. This was completed using `data = pickle.load()` for each of the obtained pickled data files. These pickled files included the calibration data, signal data, and the noise data. Random trials were selected from each of the datasets and was plotted using `matplotlib.pyplot` to show that the data could be loaded, whose plots are shown in Figure 2.

Second, empty arrays of zeros were created for each of the six energy estimators by using `numpy.zeros(len(data))`. The task was to now use multiple 'for loops' for each of the calibration data events to estimate the energy deposited into the detector by the photon beam. First was to determine the energy values of the maximum-baseline estimator. This was done by using a 'for loop' which repeated over each event and followed the process of (i) estimating the baseline value in the pre-pulse region by using `numpy.mean()` on the first 1000-sample slice of each event, and (ii) determining the maximum value of the pulse using `numpy.max()`. These maximum and minimum values were used to calculate the energy difference, and this value was taken to be the estimator array entry. By a similar process, for the maximum-minimum estimator, `numpy.max()` and `numpy.min()` was applied on each of the events in a for loop to determine the energy difference between the two. This value was entered into the maximum-minimum array estimator. Third, by using the addition function += (in a for loop), each sample voltage value was added together, thus resulting in the integration over the whole trace of each event. These integration values were entered into the trace-integration estimator.

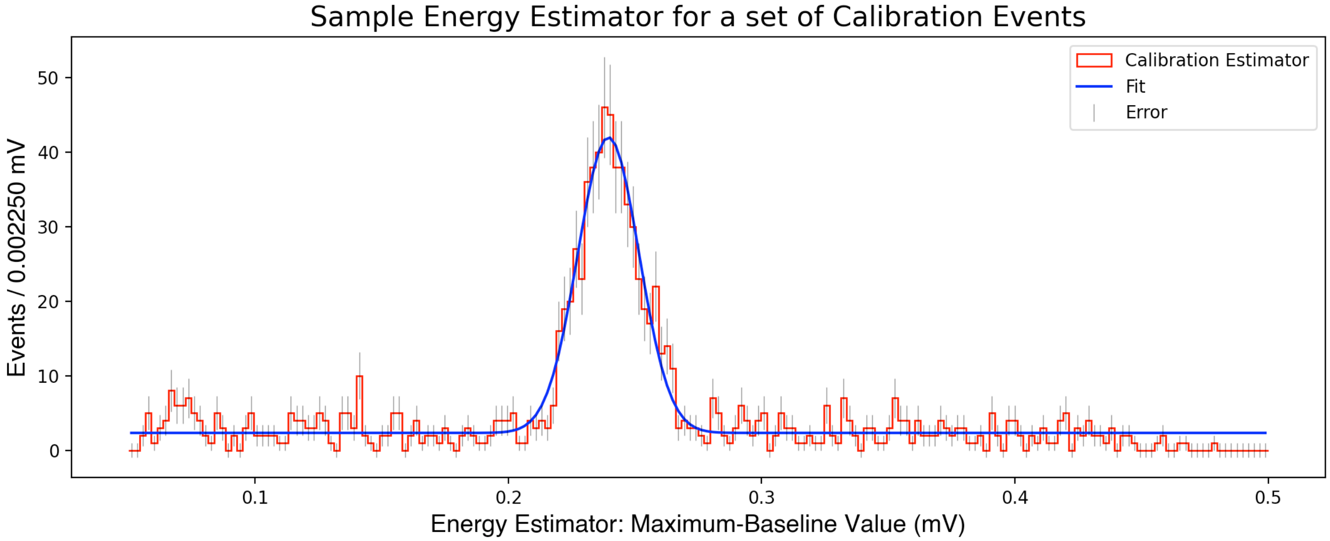


[Figure 2] The selection of sample events plotted from the sensor data. (Upper) Data attained from the calibration event set. (Center) The data acquired from noise perturbations. (Lower) The pulses acquired from measuing electron-recoil events, the signal data.

Fourth, by invoking a similar method to the maximum-baseline estimation, the trace of each pulse was once again integrated over. However, instead of comparing each value to zero, each value was compared to the baseline of each pulse. This process was as previously described, and each value was entered into the baseline-integration energy estimator array. Fifth, by slicing the sample arrays of each of the events, the pulse-region could be isolated. This was completed by comparing the baseline value of each event with the pulse amplitude, and isolating only the values greater than the baseline after the 1 ms pre-pulse region. It was found that, for all samples, a slice of `[1000:1400]` (0.4 ms) appropriately isolated each pulse. Then, by summing in a 'for loop', the sliced array was integrated over. These values were entered into the isolated-pulse integration estimator array. Lastly, by defining a model function in the form of the pulse specified in Equation (1), each pulse was fitted using `scipy.optimize.curve_fit`. For each event, the pulse was fitted with an appropriate amplitude A_i for $i = 0, \ldots, 999$ using the `curve_fit popt` value unique to each pulse. These amplitudes are accurate depictions of the energy deposited into the detector by the photons, and each amplitude was entered into the fit-pulse energy estimator.

The next step was to 'realize' these energy estimators by plotting them in histograms. Histograms were chosen beacuse they allow a visual depiction of which energy-deposits were most prominent, as well as their distributions compared to the other energy measurements. It was observed that changing the bin number (the step-size of the energy spectrum) vastly changed the shape of the data. Increasing the number of bins made the histogram 'bumpy' or 'noisy', and too few bins produces histograms which wouldn't accurately measure the energy distribution. Because of this, it was optimal to select a maximum number of bins which did not produce 'bumpy' histograms whilst not taking away from a distinguished peak. It was noted that approximately 200-300 bins was good for doing this for each trial.

Using `scipy.optimize.curve_fit`, each histogram was fitted with a distribution function. To produce an estimator consistent with the expected narrow-width peak, a Gaussian function was used to fit the data for each trial of the form $Ae^{-(x-p)^2/(2\sigma^2)}$. This implied that wider peak-distributions were already inaccurate depictions of the calibration energy deposited into the detector. An instance of this process is shown in Figure 3 for the maximum-baseline calibration estimator.



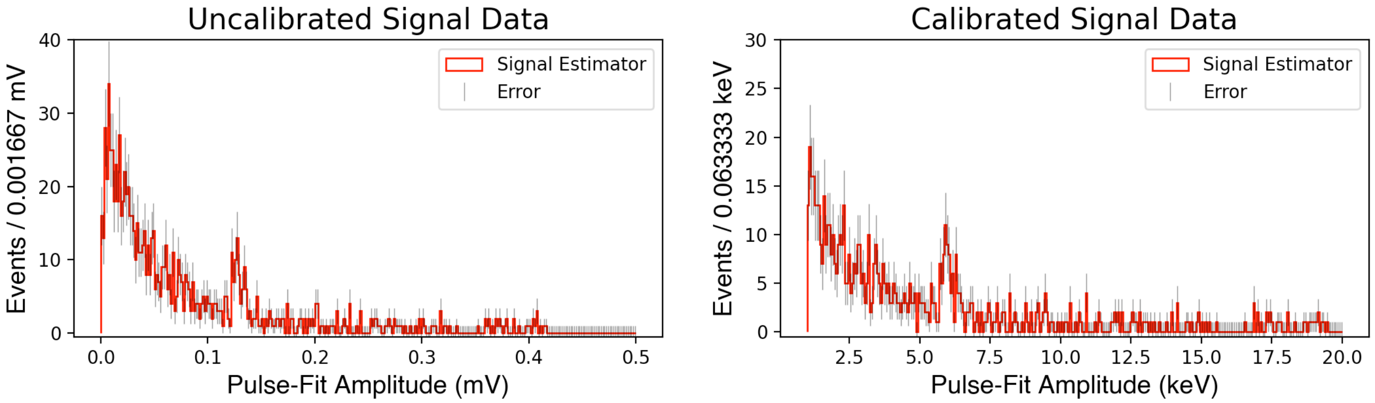
[Figure 3] Sample energy estimator for the maximum-baseline calibration energy estimator. The data, with errorbars, is shown in red and grey, while the respective curve fit taken out is shown in blue. The fit is the shape of a Gaussian distribution.

It followed to execute a χ^2 fit on each of the histograms to determine the quality of each fitting distribution. This was taken out using the χ^2 function $\chi^2 = \frac{1}{\text{dof-params}} \sum_i \frac{(f_i - p_i)^2}{\sigma_i^2}$, where 'dof-params' is the number of degrees of freedom (the bin number) minus the number of fitting parameters, f is the fit function value at i, p_i the i-th data point, and σ_i the associated error, determined by taking the square root value measurement[2] (the standard deviation). The fitting χ^2 percentage was then determined using the `scipy.states.chi2.cdf()` function.

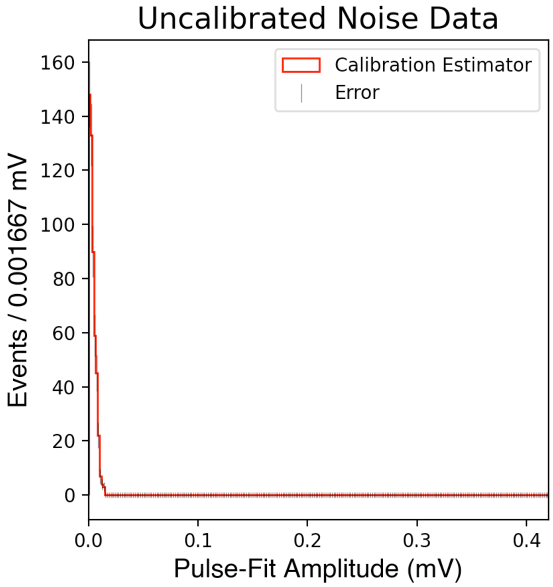
Then, by extracting the `popt` value from the curve fit, the location of the peak value was obtained. The data could then be calibrated by introducing a calibration constant k which 'shifts' the peak to the appropriate value of the peak of 10 keV. Mathematically, this appeared as $k = \frac{10}{f_{\text{peak}}}$ where f_{peak} is the extracted optimal peak value of the Gaussian function. The width of the peak (also called the energy 'resolution') was also extracted from the curve fitting outputs, as this would identify an accurate measure of peak width. This was completed for every event of the calibration data, and each energy calibration histogram is shown in Appendix I.

From the range of the six energy estimators, it was found that the estimator fit which best conformed to the distribution criterion was the pulse-fit estimator. Once again, this estimator contained a narrow peak (low resolution) with minimal external noise found outside the standard deviation range of the peak, and the energy distribution accurately represented a Gaussian distribution of photon energy. This was the selected estimator method and calibration to apply to the measured signal data.

Upon applying the pulse-fit calibration algorithm to the set of signal data, a large amount of lower-energy pulse amplitudes were observed, with a small peak around ~ 0.125 mV. These histograms, along with the calibrations, are shown in Figure 4 prior to fitting. This data was compared with the data acquired from the noise measurements, and it was concluded that due to the nature of the pulse-fit amplitude algorithm, any energy measurement below ~ 0.1 mV is attributed to noise and other external behaviour. The histogram generated via the pulse-amplitude noise estimator algorithm is shown in Figure 5. It was sufficient to filter the data so that a more appropriate curve fit may fit the data measured from the detector.



[Figure 4] The uncalibrated and calibrated sets of signal data processed using the fit pulse-amplitude energy estimator. Notice the large spectrum of noise in the lower region, with a subtle peak located in the ~ 0.13 mV/5.2 keV region.



[Figure 5] The uncalibrated spectrum of energy measurements made to the noise data via the fit pulse-amplitude algorithm. As expected, most of this noise is located is the very low- 0 mV region.

As stated, because of the large amount of noise following the signal data histogram, it was optimal to filter the noisy measurements from the measurements which were above 1 mV, the measurements assumed to represent electron detection. The was taken out after calibration by re-appending the dataset outside of the measured pulse range (5.1 keV - 7.2 keV) with zeros. From here, the sample range may be re-defined with the equivalent bin density prior to filtering. In this case, the bin density was 15 bins/unit keV after calibration.

Once a new distribution was created with the equivalent bin density, a curve fit was carried out in the Gaussian form (like the other fits) with a restricted domain, isolating the filtered pulse range. A respective χ^2 calculation was carried out. The figure representing the depiction of the final calibrated, filtered dataset is shown in Figure 6 in the results discussion.

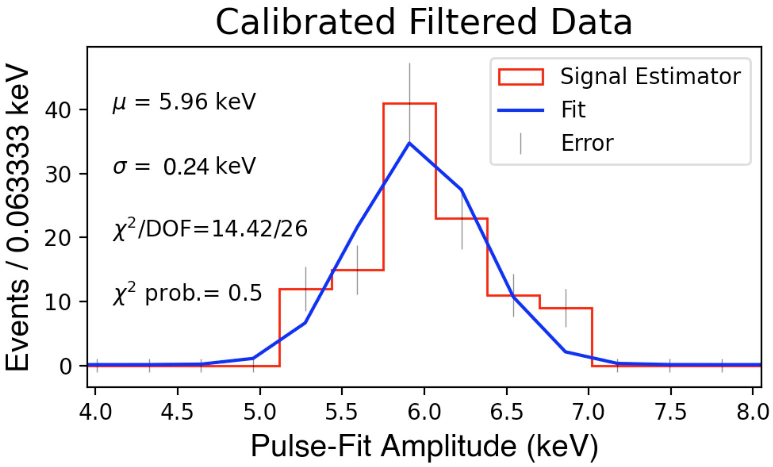
Results and Discussion

The following sets of data were obtained from each energy estimator: calibration factor to 10 keV peak adjustment; energy resolution (peak width); χ^2 fit probability. The selected estimator was the one which had the smallestt resolution (this corresponds to the most narrow peak) and a good χ^2 fit probability. This is shown in Table BB.

As described, the lower data spectra was filtered out and re-appended to zero in the histogram array. Once the data was filtered and re-plotted with the appropriate bin density and energy range, it was fitted. This is shown in Figure 6. The data extracted from the calibration of the filtered data is also shown in Table 1.

Estimator Type (Calibration)	Number of Specified Bins	Calibration Factor (keV/mV)	Energy Resolution (keV)	χ^2 Fit Probability (%)
Amplitude: Maximum-Baseline	300	41.8 ± 0.1	0.5 ± 0.5e-6	0.00
Amplitude: Maximum-Minimum	300	32.8 ± 0.1	0.5 ± 0.3e-6	0.00
Area: Trace Integration Integration	200	0.74 ± 0.04	8.21 ± 0.01	0.00
Area: Trace Baseline Integration	200	0.85 ± 0.05	9.88 ± 0.01	0.00
Area: Isolated Pulse Integration	200	0.94 ± 0.04	9.75 ± 0.02	0.00
Amplitude: Pulse-Fit Optimal Amplitude	300	46.8 ± 0.3	0.23 ± 0.05e-4	0.98
Pulse-Fit Amplitude (Filtered Data)	30	46.8 ± 0.3	0.24 ± 0.05e-4	0.5

[Table 1] Values of the energy estimator evaluation for calibration events. This table includes the width of each estimator fit as well as its peak location, the calibration factor required to calibrated the energy appropriately, the χ^2 fit probability, and the respective number of bins used in each case. Note that as the number of bins increases, so does the quality of the χ^2 fit. It was observed that, regardless the number of bins specified, the values and uncertainties of the peak and resolution did not vary. The uncertainties were taken to be the maximum value of the 'pcov' entry corresponding to each optimal parameter.



[Figure 6] The histogram obtain after filtering the signal data generated after applying the pulse-fit amplitude algorithm. This histogram was generated with 30 bins, equivalent to the same bin distribution density in Figure XX. The respective χ^2 fit of 0.5% implied that the fit was a good model for the data.

Although the new filtered-fitting function seemed to fit the data nicely, there was still a significant loss of data attributed to filtering the lower energy bins. The choice of data filtering was briefly touched on in the data analysis, however it was not fully justified. Recall that, upon assumption, every recoil electron deposited an equivalent amount of energy into the detector. This implies that the range of electron energy depositions should be narrow and (obviously) with more energy than most of the measured background events. The wide range of noise located below this distribution was therefore attributed to greater environmental decays or other anomalies in the environment.

One observation made while processing data was the effect of bin selection of the distribution. As stated earlier, the histograms generated were specified with a set number of bins so that the data measured a smooth energy distribution, while also not inaccurately mis-modelling data (ie, having too many energy measurements in one bin). Since the curve fitting χ^2 value depended on the bin specification, this had to be taken into account. Too many bins and the curve fit would be 'too good' (too many degrees of freedom), and if there were too few bins the fit would not accurarely represent the distribution and the fit would be 'bad'. Overall, as the bin number was changed, the data which was being fitted was changed because the data was being re-distributed according to the number of specified bins. 200-300 bins (over the energy range) was found to be an optimal value for all data sets to have the lowest number of degrees of freedom while still attempting to fit the data. For estimators where the energy distribution did not match a Gaussian distribution at all, no further efforts were taken to specify a 'good' χ^2 value for the fit.

Lastly, all of the extracted data from curve fitting each of the energy estimators is located in Table 2 (Appendix II).

This table includes the uncertainties due the calibration factor, the peak energy value, and the spectrum resolution before and after calibration. It was observed over the course of data analysis that the uncertainties were independent of each other (strictly speaking, the resolution and peak uncertainties). Intuitively, an uncertainty in energy resolution should not impact the uncertainty of the location of the peak, and this is what was observed in Table 2. In rows 1, and 2, while the uncalibrated peak uncertainty was small, the uncalibrated resolution uncertainty varied, as observed in rows 6, and 7. As expected, the peak uncertainty and resolution uncertainty should have no bearing on each other anyways, since these quantities are independently determined by the curve fit optimal parameters. Now consider the uncertainty attributed to the calibration factor. Observe from columns 3 and 6 from Table OO that the uncertainty of the calibration factor significantly contributes to the uncertainty of the peak, raising the order of magnitude by about 10^3 in rows 1, 2, 6, and 7. For smaller uncertainties in the calibration factor (rows 3, 4, and 5), one may further observe that the resultant uncertainties post-calibration did not change the order of magnitude of the calibrated peak uncertainties. As for the resolution of the energy spectra, columns 4 and 7, there was not observed to be a significant change in the resolution uncertainty value post-calibration. Both of these results are as expected, since the calibration factor is strictly 'shifting' the peak of the energy spectrum, and should not contribute to any significant changes in the widths or distribution of the spectrum.

Conclusion

In conclusion, recoil-electron energy detections are characterizable by a Gaussian distribution which correlates to the energy spectrum deposited into the detector for each perturbation. The findings are located in Table 2 (Appendix II) with curve fit χ^2 values located in Table 1, including the calibrated energy values and distribution resolutions.

It was found that the energy estimator algorithm which best modelled the calibration data was the pulse-amplitude fitting estimator, and this was applied to the set of signal data. The signal data was filtered and re-fitted with a Gaussian funciton, and continued to produce an energy distribution which was accurately reflected the data.

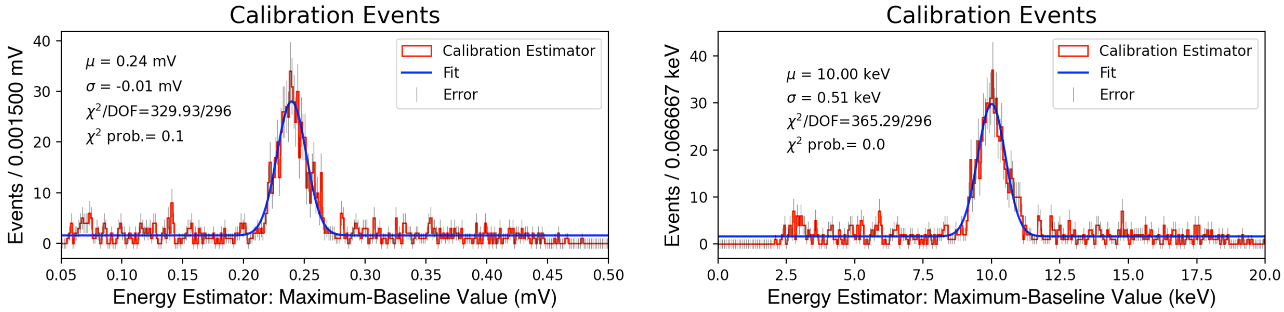
Lastly, the data analysis revealed that the peak and resolution uncertainties of the spectras are not related, however the calibration factor error significantly impacts the order of magnitude of the calibrated peak location uncertainties, and this was assumed to be because the calibration factor works by 'shifting' the data.

Bibliography

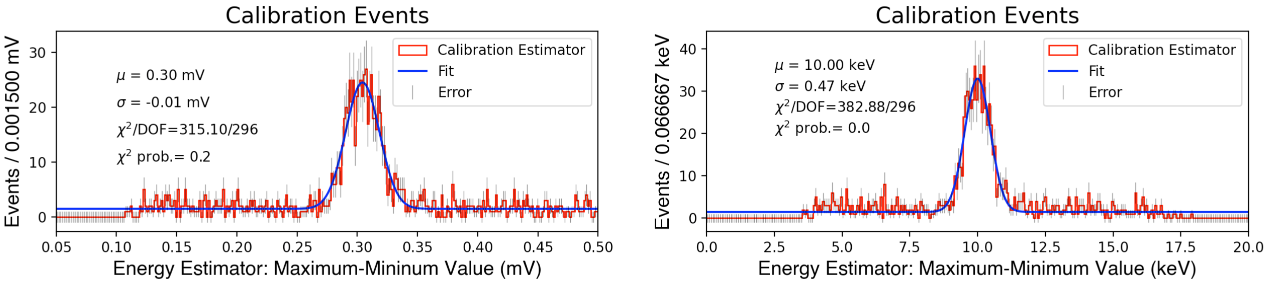
[1] Wilson, Brian; Hong, Ziqing; *PHY324 Data Analysis Assignment*, PDF. January 2023.
`https://q.utoronto.ca/courses/297231/files/24196519?module_item_id=4373276`

[2] Wilson, Brian; Hong, Ziqing; *PHY324 Data Analysis Slides*, PDF. January 2023.
`https://q.utoronto.ca/courses/297231/files/24332635?module_item_id=4428120`

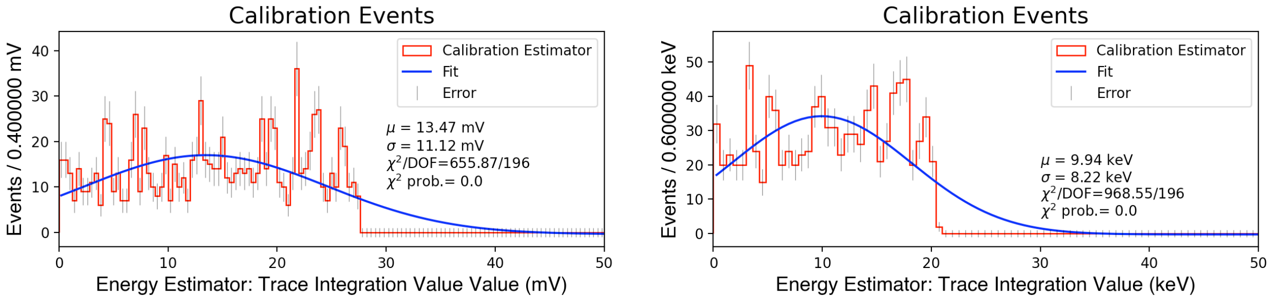
Appendix I: Energy Estimator Calibrations



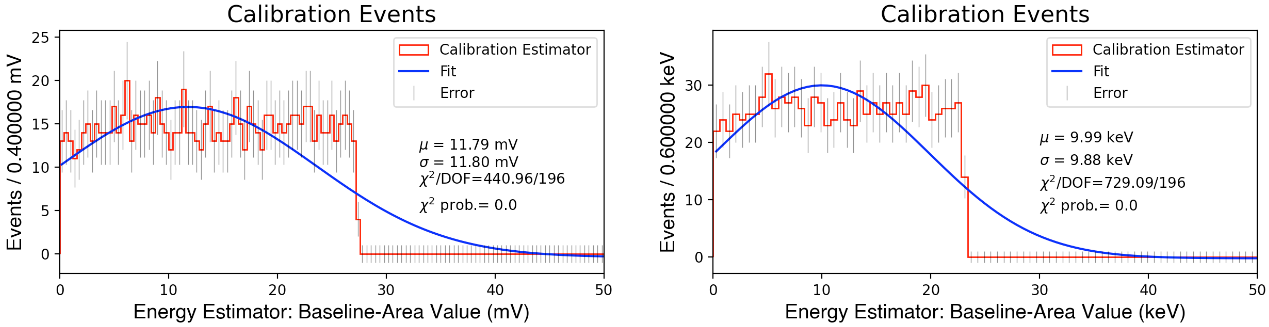
[Figure 7] The maximum-baseline amplitude energy estimator for the calibration data. (Left) The initial histogram. (Right) The spectrum histogram after calibration. 300 specified bins for both the uncalibrated and calibrated spectrums.



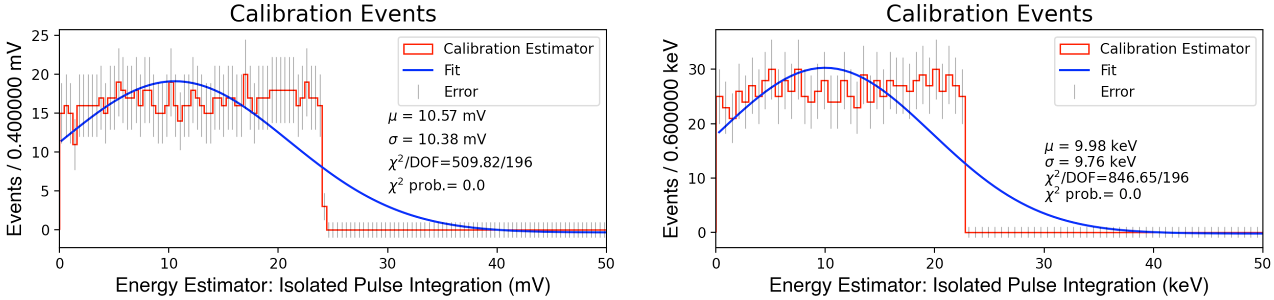
[Figure 8] The maximum-minimum amplitude energy estimator for the calibration data. (Left) The initial histogram. (Right) The spectrum histogram after calibration. 300 specified bins for both the uncalibrated and calibrated spectrums.



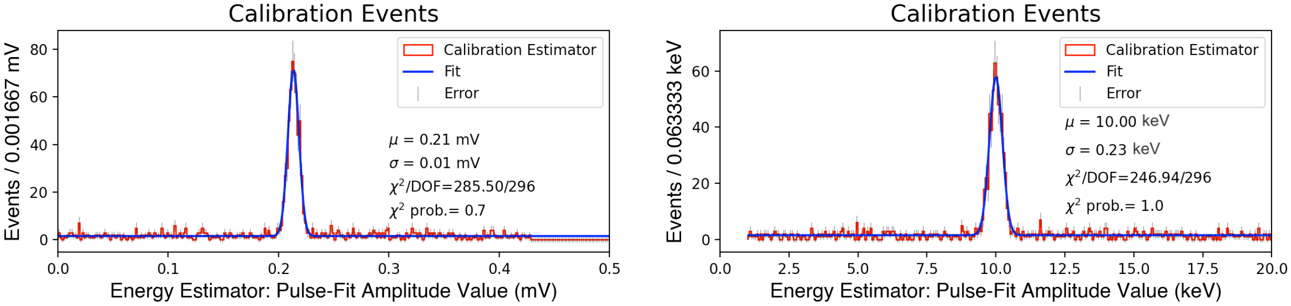
[Figure 9] The area of trace energy estimator. (Left) The initial histogram. (Right) The spectrum histogram after calibration. 200 specified bins for both the uncalibrated and calibrated spectrums.



[Figure 10] The area compared to baseline energy estimator. (Left) The initial histogram. (Right) The spectrum histogram after calibration. 200 specified bins for both the uncalibrated and calibrated spectrums.



[Figure 11] The area of the isolated pulse energy estimator. (Left) The initial histogram. (Right) The spectrum histogram after calibration. 200 specified bins for both the uncalibrated and calibrated spectrums.



[Figure 12] The energy estimator obtained by fitting each pulse with the functional form outlined in equation (1). (Left) The initial histogram. (Right) The spectrum histogram after calibration. 300 specified bins for both the uncalibrated and calibrated spectrums.

Appendix II: Table of Results and Fitting Parameters

Estimator Type	Number of Bins Specified	Pre-Calibration Peak (mV)	Pre-Calibration Resolution (mV)	Calibration Factor (keV/mV)	Post-Calibration Peak (keV)	Post-Calibration Resolution (keV)
Maximum-Baseline Amplitude	300	0.24 ± 0.03e-6	0.01 ± 0.01e-3	41.8 ± 0.1	10 ± 5e-5	0.5 ± 0.5e-6
Maximum-Minimum Amplitude	300	0.3 ± 0.5e-7	0.01 ± 0.03e-3	32.8 ± 0.1	10 ± 4e-5	0.5 ± 0.3e-6
Trace Integration	200	13.47 ± 0.02	11.12 ± 0.05	0.74 ± 0.04	9.94 ± 0.01	8.21 ± 0.01
Trace Baseline Integration	200	11.79 ± 0.04	11.80 ± 0.05	0.85 ± 0.05	9.99 ± 0.01	9.88 ± 0.01
Isolated Pulse Integration	200	10.57 ± 0.02	10.38 ± 0.03	0.94 ± 0.04	9.98 ± 0.01	9.75 ± 0.02
Pulse-Fit Amplitude	300	0.2 ± 0.2e-8	5e-3 ± 2e-9	46.8 ± 0.3	10 ± 5e-6	0.23 ± 0.05e-4
Pulse-Fit Filtered Signal Amplitude	30	0.13 ± 0.02e-5	8e-3 ± 2e-7	46.8 ± 0.3	5.98 ± 0.06e-2	0.24 ± 0.05e-4

[Table 2] The table of calibration results and uncertainty comparison for the different energy estimators. This table includes the number of specified bins per energy distribution, the peak and resolutions of the pre-calibrated and post-calibrated spectrums, and the calibration factor.